VOLUME 71 | ISSUE 3 | JUNE 2020

BJS

THE BRITISH JOURNAL OF SOCIOLOGY

Special Issue: Put to the Test - The Sociology of Testing

LSE

WILEY

BRITISH JOURNAL OF SOCIOLOGY

EDITOR IN CHIEF: NIGEL DODD

EDITOR: DANIEL LAURISON

BOOK REVIEW EDITOR: REBECCA ELLIOT

All contributions, correspondence and books for review should be addressed to:
The *British Journal of Sociology*, London School of Economics, Houghton Street, London WC2A 2AE, UK.
Tel: +44 (0) 20 7955 7283
Fax: +44 (0) 20 7955 6311
E-mail: n.b.dodd@lse.ac.uk

Publisher: The *British Journal of Sociology* is published by John Wiley & Sons Ltd, 9600 Garsington Road, Oxford OX4 2DQ, UK and 101 Station Landing, Suite 300, Medford, MA 02155, USA, for the London School of Economics and Political Science.

Journal Customer Services: For ordering information, claims and any enquiry concerning your journal subscription please go to www.wileycustomerhelp.com/ask or contact your nearest office: **Americas**: Email: cs-journals@wiley.com; Tel: +1 781 388 8598 or +1 800 835 6770 (Toll free in the USA & Canada); **Europe, Middle East and Africa**: Email: cs-journals@wiley.com; Tel: +44 (0) 1865 778315; **Asia-Pacific**: Email: cs-journals@wiley.com; Tel: +65 6511 8000; **Japan**: For Japanese speaking support, Email: cs-japan@wiley.com. **Visit our Online Customer Help** available in 7 languages at https://hub.wiley.com/community/support/onlinelibrary.

Production Editor: Qingwen Li (email: bjos@wiley.com)

Information for subscribers: The *British Journal of Sociology* is published in four issues per year. Institutional subscription prices for 2020 are: print & online: €872 (Europe), $1241 (The Americas), £684 (UK), $1336 (Rest of World). Prices are exclusive of tax. Asia-Pacific GST, Canadian GST and European VAT will be applied at the appropriate rates. For more information on current tax rates, please go https://onlinelibrary.wiley.com/library-info/products/price-lists/payment. The price includes online access to the current and all online back files to January 1st 2016, where available. For other pricing options, including access information and terms and conditions, please visit https://onlinelibrary.wiley.com/library-info/products/price-lists. Terms of use can be found here: https://onlinelibrary.wiley.com/terms-and-conditions.

Delivery Terms and Legal Title: Where the subscription price includes print issues and delivery is to the recipient's address, delivery terms are Delivered at Place (DAP); the recipient is responsible for paying any import duty or taxes. Title to all issues transfers Free of Board (FOB) our shipping point, freight prepaid. We will endeavour to fulfil claims for missing or damaged copies within six months of publication, within our reasonable discretion and subject to availability.

Members of the British Sociological Association are eligible for a discounted personal subscription rate of the journal. Please apply to British Sociological Association, Unit 3G, Mountjoy Research Centre, Stockton Road, Durham DH1 3UR.

Back issues: Single issues from current and recent volumes are available at the current single issue price from cs-journals@wiley.com. Earlier issues may be obtained from Periodicals Service Company, 351 Fairview Avenue – Ste 300, Hudson, NY 12534, USA. Tel: +1 518 822-9300, Fax: +1 518 822-9305, Email: psc@periodicals.com

Online: This journal is available online at Wiley Online Library. Visit wileyonlinelibrary.com to search the articles and register for table of contents e-mail alerts.

Disclaimer: The Publisher, Editors and the London School of Economics and Political Science (LSE) cannot be held responsible for errors or any consequences arising from the use of information contained in this journal; the views and opinions expressed do not necessarily reflect those of the Publisher, Editors and the LSE, neither does the publication of advertisements constitute any endorsement by the Publisher, Editor and the LSE of the products advertised.

Abstracting and indexing services: The *British Journal of Sociology* is indexed by ABI/Inform Database; Academic Search; America: History and Life; Anthropological Index; Applied Social Sciences Index & Abstracts (ASSIA); ATLA Religion Database; Business ASAP; Criminal Justice Abstracts; CSA Biological Sciences Database; CSA Environmental Sciences & Pollution Management Database; Current Abstracts; Current Contents/Social & Behavioral Sciences; E-Psyche; Ecology Abstracts; Educational Research Abstracts Online; FRANCIS Database; Historical Abstracts; International Bibliographic Information on Dietary Supplements (IBIDS); International Bibliography of the Social Sciences (IBSS); InfoTrac; International Political Science Abstracts (IPSA); Journal Citation Reports/Social Science Edition; Multicultural Education Abstracts; OMNIFILE Full Text Mega Edition; Policy & Politics; ProQuest; PsycINFO; Research Alert; RILM Abstracts of Music Literature; Social Sciences Citation Index; Social Sciences Index/Abstracts; Social SciSearch; Social Service Abstracts; SocINDEX; Sociological Abstracts; Sociology of Education Abstracts; SRM Abstracts; Student Resource Center College; Studies on Women & Gender Abstracts; Worldwide Political Sciences Abstracts.

Periodical ID Statement: The BRITISH JOURNAL OF SOCIOLOGY (ISSN 0007-1315) is published quarterly (March, June, September, December). US mailing agent: Mercury Media Processing, LLC, 1850 Elizabeth Avenue, Suite #C, Rahway, NJ 07065 USA. Periodical postage paid at Rahway, NJ. Postmaster: Send all address changes to The BRITISH JOURNAL OF SOCIOLOGY, John Wiley & Sons Inc., C/O The Sheridan Press, PO Box 465, Hanover, PA 17331 USA.

Wiley's Corporate Citizenship initiative seeks to address the environmental, social, economic, and ethical challenges faced in our business and which are important to our diverse stakeholder groups. We have made a long-term commitment to standardize and improve our efforts around the world to reduce our carbon footprint. Follow our progress at www.wiley.com/go/citizenship

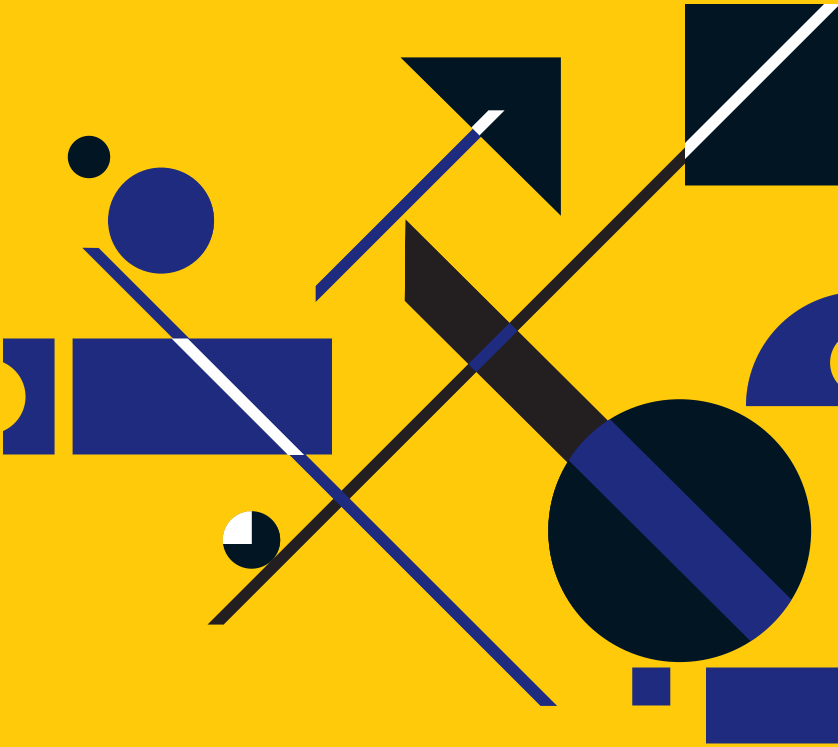
Printed in the UK by Hobbs The Printers Ltd

ISSN 0007-1315 (Print) ISSN 1468-4446 (Online)

View this journal online at www.wileyonlinelibrary.com/journal/bjos

BJS THE BRITISH JOURNAL OF SOCIOLOGY

Special Issue: Put to the Test - The Sociology of Testing



ISSN 007-1315

This journal is available online.
Contact your librarian or visit **wileyonlinelibrary.com**

BJS

The British Journal of Sociology

VOLUME 71 NUMBER 3 JUNE 2020

Contents

Special Issue: Put to the Test - The Sociology of Testing

Received: 20 March 2020 | Accepted: 23 March 2020

DOI: 10.1111/1468-4446.12757

SPECIAL ISSUE

Preface to a special issue on the sociology of testing

Noortje Marres[1] | David Stark[1,2]

[1]Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, UK

[2]Department of Sociology, Columbia University, New York, USA

Correspondence: Noortje Marres, Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, UK. Email: N.Marres@warwick.ac.uk

Funding information
H2020 European Research Council, Grant/Award Number: 695256

A test can be defined as an orchestrated attempt to reveal an entity's potentially unknown properties or capacities. A drug trial, a pregnancy test, and a planetary probe are all procedures designed to ascertain the properties of some entity. However, while tests and testing are well-established social forms, their role in culture, economy, politics, and everyday life seems to be expanding. With smart city experimentation, randomized controlled trials in economic development, and apps to test your personality and the performance of your social networks, the protocols, grammars, and logics of testing are becoming increasingly prominent as ways of intervening in society. In an age defined by computational innovation, testing seems to have become ubiquitous, as tests are routinely deployed as a form of governance, a marketing device, an instrument for political intervention, and an everyday practice to evaluate the self.

As the role of tests is expanding in culture, the economy, politics, and everyday life today, this special issue asks: What are the social and political consequences of increasingly ubiquitous testing in environments in society? What are its implications for relations between innovation, public politics, and everyday life? And what remains of the potential for experimentation as an emancipatory form? These are among the questions addressed in this special issue in which sociologists, anthropologists, and other social scientists put testing to the test.[1]

The lead essay by **Noortje Marres** and **David Stark** offers a programmatic statement for a new sociology of testing.

Luciana Leao's critical examination of randomized control trials in the field of economic development opens up a perspective on testing in society by showing that it is not only the change in location—not the lab, but the field—that makes a significant difference. Rather it is a change in *the object of testing* in the conduct of tests in society that must be worked through in terms of its consequences.

The article by **Joan Robinson** on pregnancy testing shows that taking a pregnancy test does not just test a woman's body, it equally puts at stake her moral attributes/capacities: is she able to take responsibility, is she capable of behaving like an adult? Robinson shows that it is the *socio-material practices* involved in taking this once-defined-as-medical test—from purchasing it over the counter in a supermarket, to the decision to take the test at home or in a public toilet—that are critical to its generative capacities.

Drawing on her extraordinary ethnographic work embedded in the scientific team of a NASA planetary probe, **Janet Vertesi** demonstrates that testing planets does double duty as tests of organizations. The theoretical backdrop to her analysis is the concept of "experimenter's regress" in laboratory testing, well established in Science & Technology Studies. In contrast to the relatively stable institutional context of such prior accounts, Vertesi encounters institutional and organizational uncertainty, leading her to develop the concept of "institutional regress," a previously neglected aspect in the sociology of testing.

The paper by **Jonathan Bach** on the social credit system in China is the first of three contributions in this special issue to explicitly treat testing as a device of governance. Bach shows how the creation of the social credit system entailed the creation of testing environments in society (marked by dynamics of continuous/ubiquitous feedback and intervention) and thus the ubiquitous possibility for all subjects all the time of being experimentally represented and intervened upon.

Martin Tironi is similarly concerned with the deployment of testing in society as an instrument of experimental governance, namely the orchestration of a design experiment in the streets of Santiago de Chile, called "shared streets for a low-carbon district," as part of a wider program advocating the transition to sustainable mobility in the city. Tironi shows how this test was rendered doubly experimental, as it produced effects that were not anticipated by its organizers, inviting—if not altogether compelling them—to reconsider the assumptions that guided the tests' design and implementation.

Nathan Coombs' analysis of stress testing of UK financial institutions provides an exciting alternative perspective on what renders these tests experimental. Reflecting on the sharp reduction in institutions failing the financial stress tests orchestrated by the UK's Financial Authority (FA), Coombs proposes to expand our understanding of what is put to the test in these evaluative exercises. Rather than evaluating the efficacy of stress testing retroactively, in terms of its ability to predict a financial calamity that is still in the future, Coombs proposes to evaluate these in terms of whether they enable the FA to undertake supervisory actions that would not otherwise have been possible.

Noortje Marres' article on street trials of intelligent vehicles considers the extent to which sociological propositions—formulated in the sociology of testing and the sociology of artificial intelligence—are put to the test in these engineering-led tests conducted on the "open road" in three UK cities, London, Coventry, and Milton Keynes. She argues that while these tests are designed to evaluate the performance of individual entities—vehicles, road users—they equally have the capacity to put society to the test.

Our special issue concludes with two papers bearing on identity. The first, by **Willem Schinkel**, analyzes the identities of citizenship. Subjects become citizens, Schinkel demonstrates, along a trajectory of testing practices choreographed by immigation agencies and other agents of the state. Developing a theme that is woven throughout this special issue, he points to social and political practices that operate as tests without being explicitly conceived as such.

Giovanni Formilan and David Stark analyze how artistic identity is a process of ongoing testing by studying the use of aliases among electronic music artists. They show how name-altering practices—varying arrangements of *pseudonymity* (different names), *polyonymy* (multiple names), and *anonymity* (no name, no face)—not only put the artist to the test but also probe the values and audiences that populate this underground music scene.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were generated in this study.

ORCID

Noortje Marres https://orcid.org/0000-0002-8237-6946
David Stark https://orcid.org/0000-0003-2435-9619

NOTE

[1] The papers collected here were first presented at a workshop, "Put to the Test: Critical Evaluations of Testing," Warwick in London, December 10 and 11, 2018, supported by Warwick's Centre for Interdisciplinary Methodologies, the European Research Council (BLINDSPOT project, grant agreement #695256), and a grant from the Sociological Review Association.

How to cite this article: Marres N, Stark D. Preface to a special issue on the sociology of testing. *Br J Sociol*. 2020;71:420–422. https://doi.org/10.1111/1468-4446.12757

Received: 5 February 2020 | Accepted: 11 February 2020

DOI: 10.1111/1468-4446.12746

SPECIAL ISSUE

WILEY

Put to the test: For a new sociology of testing

Noortje Marres[1] | David Stark[1,2]

[1]University of Warwick, Coventry, United Kingdom

[2]Columbia University, New York, NY

Correspondence
Noortje Marres, Centre for Interdisciplinary Methodologies, University of Warwick, Coventry CV4 7AL, United Kingdom.
Email: n.marres@warwick.ac.uk

Funding information
H2020 European Research Council, Grant/Award Number: 695256

Abstract

In an age defined by computational innovation, testing seems to have become ubiquitous, and tests are routinely deployed as a form of governance, a marketing device, an instrument for political intervention, and an everyday practice to evaluate the self. This essay argues that something more radical is happening here than simply attempts to move tests from the laboratory into social settings. The challenge that a new sociology of testing must address is that ubiquitous testing changes the relations between science, engineering, and sociology: Engineering is today in the very stuff of where society happens. It is not that the tests of 21st-century engineering occur within a social context but that it is the very fabric of the social that is being put to the test. To understand how testing and the social relate today, we must investigate how testing operates *on* social life, through the modification of its settings. One way to clarify the difference is to say that the new forms of testing can be captured neither within the logic of the field test nor of the controlled experiment. Whereas tests once happened inside social environments, today's tests directly and deliberately modify the social environment.

> *… anything which can exist at any place and at any time occurs subject to tests imposed upon it by surroundings, which are only in part compatible and reinforcing. These surroundings test its strength and measure its endurance.*
>
> *John Dewey, 1998 (1925),* Experience and Nature, *p. 70*[1]

1 | INTRODUCTION

“Have you been tested?” Context is everything. What counts as a test is completely different, depending on whether you are in a doctor’s office, on the website of the Department of Motor Vehicles, or in one of the 12 Life in the UK Test centers for new citizens in London. “Have you taken the test?” is different from “Were you involved in the test?” for the former suggests that the test was based on a decision, whereas the answer to the latter might well be “I really don’t know.” We are tested. Things are tested. Models and machines are tested, and character, too. There are stress tests for banks, cardiac patients, and nuclear power plants, there are personality and citizenship tests, sound tests and screen tests, and there are tests of strength and tests of faith.

Tests—technological, psychological, pedagogical, medical, juridical, statistical, political, religious, the list goes on—are an important part and familiar form of modern society. They signal that things and people cannot really be known in general or in advance but that their attributes ultimately need to be established on a case-by-case basis. Tests have become such a familiar genre that today almost anything can be a test situation. Sometimes we recognize that testing is underway but sometimes we are unaware. The logics and grammars of testing are so readily at hand that they can be picked up and used in many fields. Furthermore, in an age of computational innovation, more and more environments in society are equipped to facilitate testing, from airport shopping malls to social media platforms.

In recent years, sociologists and scholars in cognate fields like history, anthropology, and philosophy have insisted on the importance of experiments undertaken in social environments “beyond the laboratory,” describing them as sites where new forms of governance, economy and subjectivity are invented (Engels, Wentland, & Pfotenhauer, 2019; Mills & Tkaczyk, in press; Murphy, 2006; Van de Poel et al., 2017). In this essay, we argue that something more radical is happening than simply attempts to move tests from dedicated test sites into social settings like streets and social media platforms. Rather, it is the relations between test sites, on the one hand, and environments in society, on the other, that are changing. As we elaborate in the sections that follow, these recent developments pose a challenge to sociology and, in particular, to the sociology of testing.

We will propose the following: to grasp the significance of the rise to prominence of testing “beyond the laboratory,”[2] testing in society should be studied from the standpoint of their consequences, that is, *on the basis of what tests generate*. Making this argument (presented in greater detail in subsequent sections) means that our approach to testing casts a broad net. Rather than beginning with a restricted definition of tests, we proceeded from a posture of inclusion. It is the case that different fields have laid claim to, and have been defined in terms of, a specific logic of testing (on the natural and the social and political sciences, see Popper, 1961, 2002, respectively; on music and the avant-garde, see Cage, 1973; on democratic politics, see Dewey, 1954). But testing as a way of knowing, valuing, and intervening does not belong to any one social domain in particular, be it science, technology, politics, culture, or religion.

In music, a “demo” is a kind of market test. The same in TV and film where it is known as a “pilot.” Pilot studies in science are preliminary tests of a model or concept. In architecture a model is a test, a kind of experiment (Yaneva, 2005). The draft of a manuscript is an internal test. Circulating a manuscript as a work in progress invites users to test for bugs, a form of academic “beta testing” (Neff & Stark, 2004). A rehearsal is a kind of test (Buchmann, Lafer, & Ruhm, 2016). Conversely, tests of emergency practices and of their management are stagings that can be thought of as a kind of rehearsal (Davis, 2007). In jazz, a jam session can operate to test along multiple dimensions (Hennion, 2003). It can be a test of new melodic, thematic, rhythmic, or choral structures. In this sense, a jam session is experimental and exploratory. It can also be a test to find a new musical partner. In that sense, it is like a try-out and can vary according to whether it tests the skilled performance of the new candidate or tests the compatibility of the existing ensemble with a new candidate (DeVeaux, 1997).

In the field of romantic attachments, a date is a test—of the invited partner but perhaps also of self. New forms of online dating provide more opportunities for population-level testing to see which presentations of self attract more (or different kinds of) responses (Ellison, Heino, & Gibbs, 2006). In 2016 Volkswagen was caught cheating on diesel emissions tests. And in 2018, the British Home Office revoked visas for tens of thousands of students who were alleged to have cheated on an English language test, allegations based on spurious analysis of test results by the US-based private test provider (Bulman, 2019). Most tests are deliberate; but they can also be incidental, as some situations can be grasped, after the event, as, in fact, a test. Others can be a ruse: by reframing a prior action as a test, they offer a disclaimer for an earlier misstep: “It was just a test.”

This recognition that testing does not belong to any particular domain or even era[3] sets us apart from other approaches in the social study of testing. In Science and Technology Studies (STS), tests and testing have long been considered to be what makes science and innovation “special”: it is from the capacity to test in a laboratory that science and engineering have been said to derive their exceptional power to render invisible or distant natures observable (Latour, 1993; Mody & Lynch, 2010; Vertesi, 2015) and durably transform socio-technical arrangements in society (Callon, 1986; Laurent & Tironi, 2015; Pinch, 1993). We therefore suspend some assumptions held in these previous social studies of testing, chief among them the idea that testing “naturally” belongs to the sciences, engineering, or even to experts.

The notion that the sociology of testing should move beyond the laboratory or the field experiments of technologists was central to the work of the French sociology of critical capacities (in French: the sociology of *épreuves*, literally meaning trials). In the 1990s, this emerging school proposed that critical moments are ones in which justificatory claims, based on the principles of market, inspiration, efficiency, loyalty, fame, and the civic (Boltanski & Thévenot, 2006), are put to the test (Stark, 2011). We draw on this work as one of several points of departure. Yes, testing occurs in many sites; but it is not only about justificatory claims. And yes, as well, it is important to grasp the test as a critical moment; but it would be mistaken, as we elaborate below, to characterize tests first and foremost in terms of the forms of justification or orders of worth they elicit. Most importantly, whereas Boltanksi and Thevenot analyse social life as giving rise to trials (making comparisons to scientific and juridical trials) that yield a judgment, we argue that tests should be studied not on the basis of what they resolve but by what they generate.

Tests are generative, they stimulate further testing, not only as *experimental or test regress* in the same modality (see for example Collins, 1981, 1985) but involving *diverse modalities* (Stark, 2011; Marres, 2012) of knowing, valuing, and acting. As we shall argue in more detail below, when one studies what results from the test and not simply the “results of the test”—that is, when one examines consequences less as what is resolved than by what is further generated—then research on testing cannot remain focused on a particular testing moment but must study whether and how any given test operates in an *ecology of testing*. This is a situationalism (Knorr-Cetina, 1981) that attends to the structure of the situation (Stark, 2017).

Moreover, our intuition is that contemporary forms of testing (real-world experiments, platform-based testing, RCTs) mark a more distinctive departure than simply a movement from laboratory to field. As a starting point, think of them as testing “in society” (van de Poel, Asveld, & Mehos, 2017). And then, more radically, begin to see that, in many of the *real-world testings* examined here, tests are not just *in society* but are tests *of society*. The challenge that a new sociology of testing must address is that the very relation between science, engineering, and sociology is changing: Engineering is today in the very stuff of where society happens—and not simply because technology is embedded in or infused by the social. Instead, in our era, engineering tests the very fabric of the social. As we shall see, it is not that the tests of 21st -century engineering occur within a social context but that it is the social that is being put to the test.[4]

2 | POINTS OF DEPARTURE: THE SOCIAL STUDIES OF TESTING AND THE SOCIOLOGY OF EXPERIMENTATION

To examine how the role of testing in society is changing, and to figure out what should be the next steps in the sociology of testing, we draw on and move beyond two established sociological approaches: the social studies of testing developed by scholars in STS from the 1980s onwards, and the sociology of experimentation, which can be traced back to American pragmatist sociology of the Chicago School of the 1920s.

2.1 | The social studies of testing: Finding the social inside the test

In science and technology studies, tests have been granted special importance as sites where the inter-relatedness of science and society becomes visible (Callon, 1986; Latour, 1996; Woolgar, 1990). In the late 1980s, as scholars began to extend insights from the sociology of science to the sociology of technology, they attributed central importance to tests, viewing them as the technological equivalents of scientific experiments: Tests are to technology as experiments are to science (Pinch, 1993). Just as the Sociology of Scientific Knowledge (SSK) had done with science before, sociologists of technology studied tests to demonstrate that there is an irreducible social dimension to testing. In "From Kwajalein to Armageddon?"—a study of intercontinental ballistic missile testing between California and the Marshall Islands—Donald MacKenzie presents this now familiar, Wittgensteinian critique of epistemology:

> *… the result of most testing is accepted routinely as fact. … [N]either "logic" nor "reality" are sufficient to explain this. Other elements must be involved, and in them "the technical/scientific" is inextricably interwoven with "the social". (MacKenzie, 1989, p. 415)*

Striking about MacKenzie's article today is how he defines the "social." He views the social primarily in terms of *convention*, where the testability of any claim or proposition is "defined by the consensual practices of groups of accredited practitioners" (MacKenzie, 1989, pp. 416–417).[5] For MacKenzie, as for other SSK scholars before him, these social practices are the determinant that makes up for the indeterminacy of scientific knowledge. The task for the sociology of testing, for him, is to account for the *legitimacy* of technological testing, and the knowledge it produces, in social terms. The idea that expertise has legitimacy can today easily be doubted. In a historical moment in which the public credibility of science and technology are widely questioned, it cannot be the job of the sociology of testing to explain why they are *not* being challenged in our society.

While MacKenzie takes care to foreground the social environment in which ballistic missile testing takes place—the essay opens with a description of the islanders' situation,[6] his conception of the *social dimension of testing* makes no reference to the test environment. Nor, given his objectives, should he have done so. If the goal of the SSK project was to demonstrate the irreducible social dimension of technology testing, then it was necessary to analyze a "hard" case focusing on highly technical and abstract statistical operations performed in dedicated test sites "away from society."[7] To engage conceptually with the consequences of testing "on the ground" would get in the way of testing SSK's theory of testing. But this belief that it was necessary to study technology testing in dedicated environments at a remove from everyday environments in society came with a cost: SSK's *conception of the social*—in terms of the conventions that prop up the credibility of science and scientists—followed (at least partly) from this investment and was then reinforced by its findings. It is this particular treatment of the social in SSK studies of testing that we believe is no longer satisfactory today. Looking back at this literature from a distance, we are particularly struck by the abstract vocabulary—"convention" securing the "legitimacy" of science—relied on to characterize the social aspects of testing.

Actor-network theory (ANT) was different, in that their studies of technology testing, such as Callon's (1986) well-known study of the electric car, valued tests as sites where the power of innovation to transform society could be demonstrated *by different means* (different that is, from the political, legal, and cultural means by which sociology had typically understood social change, and social order, to be accomplished). Consequently, ANT studies of technology testing did not rely on existing sociological vocabulary, as in MacKenzie's case, to account for the social dimension of testing. Instead they advanced novel concepts, such as "enrolment," "translation," and "interest alignment," to highlight the capacity of tests to implicate social actors in the innovation process, or more precisely, to foreground the capacities of technology testing for ontological transformation. For ANT, the test enables the performative specification of a new set of connections between technical elements, social actors, objects, interests, and so on (Callon, 1984, 1986; Muniesa, 2014).[8]

While SSK and ANT had much to disagree about—ANT proposed "pay attention to the non-humans," while SSK and their successors in the social construction of technology (SCOT) program claimed "don't forget about the humans!" (Bijker & Pinch, 2012)—there was in retrospect much that they shared: both developed new kinds of explanations of the power of science and innovation by locating the social inside the test. By contrast, the sociology of testing advocated here makes the reverse move—it shows how testing and experimentalism have been extended into distinctively social environments, so that it is now society and social life itself that is subject to the technological regime of testing.[9,10]

2.2 | The sociology of experimentation: Finding testing in society

As noted, one of the pressing questions for the sociology of testing today is how to make sense of the implementation of tests and the creation of new types of test environments—smart city pilots, for example, or randomized control trials for economic development—in society (in hospitals, streets, villages, media platforms, etc.). To do so, scholars across the social sciences have in recent years turned to a sociological tradition, the Chicago School, and in particular, the classic proposition of Robert Parks to regard social environments like cities as laboratories (Gieryn, 2018; Gross, 2009; Gross & Krohn, 2005; Guggenheim, 2012). This proposition is today celebrated as a relevant precursor of contemporary efforts to create living laboratories, test beds, and experimental cities (Engels et al., 2019).[11]

In these accounts, the Chicago School stands out for its proposition that social life is already experimental in and of itself (for a discussion see also Marres, Guggenheim, & Wilkie, 2018). Building on the pragmatist philosophy of John Dewey—who had extended the notion of experiment to a wide range of activities, including personal conduct, the running of government, and democratic movements—the Chicago School proposed to study the city of Chicago, its neighborhoods and communities as laboratories. As both Thomas Gieryn and Matthias Gross note, the case for Chicago as laboratory was made on the grounds of its diversity, its social problems, and the unprecedented encounters between relative strangers (African Americans, immigrants and "hobos") in this city, who needed to figure out (try, test) how to live together (Thrasher, 1927, p. 488; Tolman, 1902, p. 116). However, others offered general justifications for the approach. As Chicago sociologist Vivien Palmer (1928) argued, "human beings are everywhere so continually performing their own experiments in group life that the investigator can always find social experiments of many kinds in progress: a systematic, contemporary, observation of these yields significant facts" (p. 8).

The intellectual project of the sociology of experimentation is, thus, in many ways the opposite of 1980s' STS studies of testing: whereas the latter finds the social inside testing, the former finds testing inside the social. The key insights of the Chicago School that social practices increasingly present themselves as experiments (Gross & Krohn, 2005, p. 80) and that sociologists should take advantage of this experimentation by everyday social actors in society (Gieryn, 2018, p. 15) was initially ignored by SSK and ANT studies of testing, where the focus was on hard science and innovation in environments at a distance from the familiar settings of everyday life. Conversely, for a contemporary sociologist like Matthias Gross studying the Chicago School, the experimentality of social life thematized by Park and colleagues has very little to do with traditional notions of science. For him, social experimentation is "completely different" from social science experiments; to treat society as an experiment is to adopt a "sociological perspective [that] has got nothing to do with the idea of sociologists as experimenters in white coats" (Gross & Krohn, 2005, p. 80). Most pointedly in their mirrored opposition, whereas SSK scholar Harry Collins regarded the scientific experiment as a template for all forms of cultural production (Collins, 1985, p. 18, 165),[12] Chicago School advocate Martin Gross argues that "experiments in the real world are, in a sense, the real and true experiments, and the laboratory ideal is a special variation of it" (Gross, 2009, p. 90).[13]

Our task here is not to adjudicate among claims about which are the "real and true experiments" upon which other kinds of tests should be modeled. Instead, it is to understand the achievements and limitations of these two approaches for analyzing testing today. If SSK persuasively demonstrated that technological testing is suffused with social processes, its abstract conception of the social restricted its choice of research settings to specific sites away from mundane society—thus, bringing in the "social" but, in effect, avoiding the societal. If the rediscovery of the Chicago School by scholars of testing decidedly moves our attention onto this otherwise neglected terrain, its notion of the self-sufficiency of social experimentation—that social life is already experimental in and of itself—ignores how expert-led experiments deliberately introduce something new into social life. The proliferation of tests across society is today intimately connected to the organizational, technical, material *modifications of society*, often by computational means—modifications that render social life observable, analyzable, and influenceable *by a variety of actors* (Marres, 2017). In this context, the primary question is not whether social practices can be *defined* as experiments, but how social environments—and possibly social life itself—are *modified* so as to enable experimental operations by a range of different kinds of actors, and putting these actors into a determinate, often unequal, relation.

To study these contemporary tests in and of society, we believe it is necessary to let go of a number of assumptions that, despite their differences, the two approaches share, chief among them the underlying assumption that expert-led testing and social experimentation are best understood when studied in isolation from one another. In our view, it is the very division between the social studies of testing and a sociology of experimentation that risks rendering invisible the testing situations that the sociology of testing should elucidate. It is not just that in these situations, *both* engineers and scientists, on the one hand, and social actors, on the other hand, tend to be implicated. As or more important is that *a diversity of modes of testing* and experimentation are detectable in them: scientific, political, aesthetic, commercial. In recent street trials of autonomous vehicles, engineering tests do double duty as public engagement experiments, and these trials are equally enlisted to conduct activist and creative interventions into street environments, as when the artist James Bridle trapped an autonomous car by laying a salt circle on the surrounding road.[14] It is not just that engineering tests and social experimentation happen in the same social space, the street; rather, environments in society are being modified by a variety of test subjects and agents, in such away that diverse forms of experimentation pose a challenge to one another.

This certainly does *not* mean that the differences between technological testing and social experimentation are today being dissolved, but it does follow that if we study these diverse forms of testing in isolation from one another, we are unlikely to understand how society is transformed by means of testing. We therefore argue for a *new sociology of testing*. But to make that case, we need to do more than push back against existing theoretical frameworks. We need materials to think with, and so we offer two brief accounts, examples of testing in two historical periods, the first from the mid-1950s, the second from our current time some 60 years later. In methodological terms these are not *cases*, and although they involve tests, they are emphatically not offered as *tests* of competing theories. We make our argument in steps. The first example provides an opportunity to point to changes in research strategy that would need to be made if the sociology of testing is to account for testing *in society*. The second example points to changes in infrastructures and practices in society that the sociology of testing will need to take into account in a new era of testing.

3 | FOR A NEW SOCIOLOGY OF TESTING I: STUDYING TESTING IN SOCIETY

3.1 | 1956: The fallout of nuclear testing

In March 1959, *Consumer Reports*, a monthly publication of the Consumer Union, a nonprofit consumer advocacy organization, published a report of their product testing that found radioactive isotopes in milk. Readers of the US magazine would typically consult its pages for reports on the testing of products such as cars, tires, or household appliances. But in this report on "fresh clean milk, which looks and tastes just like it always did" they learned, many for the first time, about strontium-90. A "toxic substance known to accumulate in human bone," this "unseen contaminant" was a health hazard caused by radioactive fallout (Consumer Reports, 1959, p. 103). In testing for the adverse effects of radiological food contamination, the product testing of milk traced a line back to an earlier (and, at the time, ongoing) set of tests—atmospheric nuclear testing.

The first test of a nuclear explosive device, named "Trinity" by J. Robert Oppenheimer, took place in the New Mexico desert at 5:30 a.m. on July 16, 1945. Surface level radiological monitoring was limited to 200 miles from ground zero. But the effects of the Trinity test were incidentally measured far beyond the distances accounted for in the army's surveys. In August of that same year, batches of photographic film belonging to the Eastman Kodak Company in Rochester, New York, began to show inexplicable areas of fogging (Stannard, 1988, p. 885). A physicist at Kodak's Research Laboratory traced the phenomenon to the interleaving paper used to separate photographic films in their packaging. It turned out that this interleaving paper was produced from corn husks, which were traced to a farm in Indiana, some 1,300 miles from the Trinity test site (Bruno, 2003, p. 140). The Kodak physicist concluded that the corn husks were contaminated by radioactive fallout, and eventually published these findings in 1949.

With the onset of the Cold War, atmospheric nuclear testing accelerated in both the United States and the Soviet Union, reaching its peak in 1961–1962 when about 250 megatons were detonated in the atmosphere (UNSCEAR, 2000, p. 160). In 1949 the US Atomic Energy Commission (AEC) launched a parallel test—an investigation into the "long term, widespread hazards" of nuclear testing. The project, code-named Gabriel, was followed by another, Sunshine, in 1953. Analyzing how radioactive isotopes such as strontium-90 found their way into "air, water, soil, plants, animals, and humans" (Bruno, 2003, p. 240), Gabriel and Sunshine were, at one level, tests of testing. But because these investigations hinged on the question of "how many bombs could be detonated before reaching doomsday level?" they were framed as an "essential preparation for a nuclear war" (Bruno, 2003, p. 243). In an important sense, Gabriel and Sunshine were *stress tests*. That is, they were not only conducted to test the effects of nuclear testing, but were also conducted as a test about a catastrophic event—all out nuclear war—that had not actually occurred but could be imagined or projected into the future.

For these projects, the AEC collected data on the pathways of contaminants not only in the food chain but also in the upper atmosphere and in the oceans. Incidentally, the Gabriel and Sunshine tests greatly improved the general understanding of circulation on the planet. It was through such studies, for example, that "atmospheric currents such as the jet stream were detected" (Bruno, 2003, p. 245). Every cloud, even a mushroom cloud, it seems, can have a silver lining.

But the Air Force and the AEC were not the only agencies conducting tests. Following the declassification and publication in 1957 of the Project Sunshine findings about the presence of strontium-90 in nuclear fallout, the US Public Health Service began to monitor levels of radioactivity in air, water, and milk samples in 1958 (Watkins, 2001, p. 301). The real challenge for health officials was to find out how much of the radioactive isotope was getting into bodies. Established procedures called for bone samples which were difficult to get in large numbers. The answer was not discovered by AEC commission but by the "Baby Tooth Survey"—one of the most politically consequential test of nuclear testing conducted at the time—announced in December 1958 by the Greater St Louis Committee for Nuclear Information (CNI), a partnership of citizen and scientific activists, including the environmentalist pioneer Barry Commoner.

CNI called on children and parents to send in baby teeth, along with a 3×5 card with basic information about the donors such as birthdate, where they resided, and whether they were breast-fed or bottle-fed. Mobilized by dentists, churches, libraries, and schools, local residents responded eagerly. St Louis was an ideal site for the survey: the Public Health Service tests in nine US cities had found that the Midwestern city topped the list for having the highest levels of strontium-90 in milk, tainted by dairy cows grazing on grass contaminated by nuclear fallout. Under the direction of a local physician, Louise Reiss, and cataloged by volunteers of the Women's Auxiliary of the St. Louis Dental Society, the project collected almost 15,000 baby teeth in its first year (and more than 300,000 by the time the project ended in 1970). Reiss reported the findings of the survey in an article she published in *Science* in 1961: Children born in 1954 had four times as much strontium-90 in their teeth as those born in 1951

(Reiss, 1961, p. 1670). That study received widespread media attention and is credited as an important moment in the efforts to bring about the 1963 Limited Nuclear Test Ban treaty as acknowledged by President John Kennedy in a personal phone call to Louise Reiss (Watkins, 2001, p. 304).

3.2 | Lessons for the sociology of testing, step 1

We draw three major lessons from this account of the societal fallout from nuclear testing.

3.2.1 | Tests are generative and may give rise to an ecology of testing

Tests, we argue, should be evaluated not only on the basis of their validity or by what they resolve but equally by what they generate. For the sociology of testing, more important than the "test results" is *what results from the test*.

Tests are generative, in the first instance, because they provoke further testing. As we saw in the nuclear testing example, what began as the test of a weapon spawned tests of the atmosphere, of ocean currents, of planetary resilience, strontium-90 in baby teeth and the credibility of US federal agencies. More important, as we shall also see elaborated in the research findings in the articles of this special issue, the further tests that result from testing need not be of the same modality nor in the same register. Not just a more refined test of the test, and not a simple contagion like the spread of a given bacteria, the tests that proliferate are more like different species—technological, personal, political, artistic, and so on.

Our proposed perspective directs attention to a different phenomenon from the one that preoccupied 20th-century philosophers and sociologists of science and technology from Pierre Duhem to Harold Garfinkel as well as Donald Mackenzie noted above: The experimenter's regress. According to the latter relativist proposition, *the results of any test are inconclusive*, and therefore, in principle, can always be contested in further tests. We point to something different: A given test tends to unfold amidst other tests, and often these are *different kinds* of tests. It follows that one of the principles of the sociology of testing would be that tests should be examined within an *ecology of testing*.

When viewed in this way, we can understand why tests are not only inconclusive but also why they can be unsettling. A given test often does not resolve, and the more that it takes place within an ecology of tests operating in different modalities, the less the likelihood of clear resolution. It is the possibility for shifts and dissonant connections between different registers of testing that gives rise to *a distinctive type of ambiguity*, or more precisely, undecidability. Within such a context, testing raises the determinate possibility of change in several different directions. In such a situation contestation is not simply proposing an alternative way to measure some attribute but pointing to a different value altogether along a very different dimension and according to another accountability. This is why tests have the capacity to unsettle. And it is for the same reason that they can be productive without producing resolution (Stark, 2011).

When we look this way at the consequences of testing, we see that tests that involve technical devices can give rise to tests of social relations. This is a principal finding of Joan Robinson's (this volume) research on the home pregnancy test, "What the pregnancy test is testing." Robinson's case is a textbook example of testing "in the wild," as the pregnancy test moved out of the laboratory and "came home from the hospital."[15] The question now addressed: What is tested in the home pregnancy test? The pregnancy test is a technology of discovery. With it a woman can discover, without the mediation of medical professionals, whether she is pregnant. Robinson's sociological insight is to situate the home pregnancy test in the context of social relationships. While revealing a new relation between a woman and her body, the home pregnancy test has the potential to requalify relations between the woman and her partner, her friends, her in-laws, her high school swimming coach, her employer, and others.

Robinson demonstrates that, within such an ecology of testing, the pregnancy test (and not the simple yes/no fact about pregnancy) can put social relations to the test.

3.2.2 | Testing in society takes place at multiple sites, on differing scales, provoking emergent entities and sometimes deploying intimate relations

In our example of testing at mid-20th century we saw that the ecology of testing finds tests taking place in many sites and settings. On the American side, the nuclear weapons tests themselves took place primarily in Nevada and in and over the Pacific, including on the Bikini Atoll, a coral reef in Marshall Islands. But the literal fallout of the tests spanned the globe; and the figurative fallout led to tests in an upstate New York commercial laboratory and in a cornfield in Indiana, tests of soil samples from the dairy farms near 10 major American cities and tests of models of nuclear apocalypse. The “Pacific Proving Grounds” (as the nuclear tests site were named by the US Government) might have been far distant from Washington, DC, but they were certainly not isolated. Nor are the different kinds of tests isolated from each other. Yet, as connected, they were not associated as parts of a coordinated, encompassing test.

In addition to being multi-sited, the new sociology of testing must also be alert to processes whereby testing can be conducted by human entities not at all anticipated in the initial tests (employees at Eastman Kodak, from our example) involving non-human mediators at steps removed (e.g., corn husks and baby teeth as silent witness). In dynamic test environments, moreover, testing enables the activation of unexpected entities (in our example, an association bringing together teachers, dentists, and environmental biologists).

A given nuclear explosion can be measured on various scales, perhaps there is one that goes from a stick of dynamite to a solar storm. Any given test is likely to have many data points, but the total number of atmospheric tests as counted by the Arms Control Association (2019) has been 528. As mentioned, the Greater St. Louis Committee for Nuclear Information eventually collected over 300,000 baby teeth. Whether small or large in the number of data points, tests can involve the extraction of intimate stuff from family life. (children giving up the tooth fairy got a button exclaiming “I GAVE MY TOOTH TO SCIENCE”). The cold calculations of the stress tests to estimate the survival of the human species in the aftermath of nuclear war were conducted on Scale 1—the planet.

3.2.3 | The relation of testing agents and tested subjects is unstable and to an extent reversible, sometimes leading to a test ban

American citizens, Soviet citizens, human beings living anywhere on the earth from Nova Scotia to New Zealand, from Beijing to Buenos Aires, all were involuntarily subjected to the fallout of nuclear testing. Although the initial object of the atmospheric nuclear tests was the weapons system, with further testing of the effects of nuclear testing, humans became research subjects. Some citizens resisted nuclear testing and nuclear weapons absolutely, most famously the Campaign for Nuclear Disarmament in the UK, whose semaphore symbol schematically became the international peace symbol. But, as our example shows, in addition to publicly demonstrating, citizens carried out their own tests, the results of which were offered as demonstrations that strontium-90 was being absorbed into infant bodies. That is, instead of the ANT dichotomy of compliant versus unruly, these subjects were *contesting*. The St Louis moms put the Atomic Energy Commission to the test.

No one volunteered to be the subject of nuclear testing. Most significantly, no one could opt out. No single person could choose to be individually exempted from being involved in atmospheric tests. For any person to be protected from the results of nuclear testing, everyone on the planet needed to be protected. The answer: A test ban.

4 | FOR A NEW SOCIOLOGY OF TESTING II: STUDYING CONTINUOUS, UBIQUITOUS TESTING

4.1 | 2016: Fallout from online personality testing

In 2013, scientists at Cambridge University—notably Michal Kosinski and David Stillwell at the Psychometrics Centre—developed a third-party Facebook app, mypersonality.org. For this online personality test, users answer some questions and receive their scores on a five-factor model of personality, the so-called "Big Five," or "OCEAN" model (for Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism; see John & Srivastava, 1998). Various prior studies had shown that personality traits can predict patterns of activity such as consumer purchases, sexual orientation, or political behavior. Of interest to the Cambridge University researchers was the following question: Could these traits be found to correlate with online behavior such as "liking?" If so, one could use data gathered through social media to make inferences about personality (even without the user taking a personality test).

On the basis of their research, Kosinski and colleagues answered this question in the affirmative, publishing a number of scientific articles positing that the analysis of social media data captured by means of online tests could predict private attributes such as sexual orientation and political preference (Bacharach, Grepel, Kohli, Kosinski, & Stillwell, 2014; Kosinski, Stillwell, & Graepel, 2013; Lambiotte & Kosinski, 2012). While the study attracted significant media attention because of its findings, including headlines such as "Gay? Conservative? High IQ? Your Facebook 'likes' can reveal traits" (Boyle, 2013), Kosinski and colleagues themselves addressed publics as testing subjects, as they publicized this research through the website, youarewhatyoulike.com, where "you can sample the method for yourself," as NBC put it (Boyle, 2013). Reporting of related Facebook-based experiments highlighted the experimental potential of their insights: taking advantage of the interactive features of online platforms, it became possible to seamlessly *move from the analysis* of user attributes to *targeted behavioral intervention*, promising to render "real-world behaviour … amenable to online interventions," as computational social science colleagues put it in *Nature* (Bond et al., 2012).

The scaling up of this experiment equally depended on the roll-out of personality testing online. In Spring 2014, Aleksander Kogan created a Facebook app "thisisyourdigitallife," using Amazon's Mechanical Turk to pay users one dollar for downloading its app and completing its incorporated personality questionnaire. Approximately 270,000 individuals downloaded thisisyourdigitallife, and through Facebook's third-party APL's this allowed Kogan to harvest tens of millions (50–87 million) of Facebook profiles. Subsequently it was revealed that the company Cambridge Analytica used Kogan's app to collect and analyze data for political campaign purposes (Cadwalladr & Graham-Harrison, 2018).

While *The Guardian* journalists who broke this story became the poster-persons of the public exposure of what has become known as the Cambridge Analytica scandal, their reporting critically relied on independent counter-expertise, *and counter-testing* of Facebook's platform settings, in the effort to experimentally determine its treatment of personal data. In December 2016, Pierre-Olivier Dehaye, a Belgian mathematician with a long-standing interest in the use of personal data by tech companies, began researching the role of Facebook-based personality testing in political campaigning (Dehaye, 2017). He also reported on the concerns that Kosinski himself has about the efforts of Kogan and associates "which was trying to purchase data obtained via his team's Facebook-linked MyPersonality psychology app for reasons it would not reveal." From early 2017 onwards, Dehaye began methodically sending data access requests to Facebook to investigate Facebook personal data processes and practices, in particular, how it handles information on custom audiences, and Pixel data—that is, web browsing tracking-data (Ausloos, 2018).

In February 2017, Michal Kosinski—by then an associate professor of organizational behaviour at Stanford Business School—published new research demonstrating that it is possible to predict sexual orientation from facial images scraped from data websites using neural networks (Wan & Kosinski, 2017). Dubbed an AI gaydar, this research also attracted significant media attention. This claim was put to the test in an especially well-publicized counter-study of the "AI gaydar," conducted by the Princeton psychologist Alex Todorov, together with research scientists at Google, Margaret Mitchell, and Blaise Agüera y Arcas. They used Mechanical Turk to organize a survey of 8,000 Americans that asked 77 yes/no questions such as "Do you wear eyeshadow?", "Do you wear glasses?", and "Do you have a beard?" as well as "Are you same sex attracted?" Comparing the results of this survey with an analysis of portrait photography on dating websites, they demonstrated that Kosinski's test reveals not sexual orientation but the photographic conventions observed in dating websites, that is, social stereotypes (Agüera y Arcas, Todorov, & Mitchell, 2018).

4.2 | Lessons for the sociology of testing: Step 2

What can we learn from our second account of testing? There is, of course, at least one major way in which the fallout from psychological testing (as an example of new forms of testing in the online setting) is similar to nuclear testing: in each case, millions learned that they were the unwitting subjects in a series of tests. We focus here on three ways in which they might be different.

4.2.1 | Engineers have moved into the social

The mid-century nuclear tests affected intimate relations in society—the radioactive particles released by atmospheric testing entered the bones and teeth of children across the country. But it required the intervention of citizen activists who mobilized intimate relations (by way of baby teeth and mothers' milk) to demonstrate these effects. Today, tests by scientists and engineers are purposefully designed to extract and deploy intimate information to deliberately act on populations and intervene in society.

For the engineers of nuclear testing then, human beings—when considered at all—were simply part of the environment in which their tests were being conducted. For the new engineers of big data testing today, the behavior of human beings is the object to be tested. Even more, for research teams at Google, Microsoft, Airbnb, Uber, Facebook, and other firms, it is less that they are conducting psychological tests *in a social environment* but that the social environment is itself the object of testing. "Uber Engineering" celebrates computational social science, and its counterparts at Facebook, Twitter, and elsewhere do the same. Some of these teams are headed by researchers with PhDs in the social sciences, including those with titles like "Lead Trust Scientist," but these research teams are more likely to be led by a "Data Scientist" with degrees not only in informatics, computer science, or network science but also other disciplines ranging from anthropology to physics. What is a social relation? What is trust? What network structures of communication erode or promote trust? In many of the research teams of platform capitalism, you would not be laughed out of the room for posing such questions and offering ways to operationalize them as formulations to be tested—precisely because these concepts deeply inform the data infrastructures that have been rolled out across environments in society and are instrumental to the value propositions of these companies.

4.2.2 | The sites and the logics of testing have changed

Social engineering, of course, is not new. Social engineers (the term dates from the end of the 19th century) have long attempted to persuade or influence voter and consumer behavior (for an insightful overview see Wu, 2017), and for this they conducted research involving tests. What is new, we suggest, is that the sites and the logics of testing have changed. In the past, the test situation was a moment that was spatially, temporally, and infrastructurally marked off from the ongoing flow of life. Focus groups, for example, brought research subjects into a laboratory setting (Lezaun, Muniesa, & Vikkelsø, 2013). The use of field tests in marketing research (for example, to determine the effects of product placement on lower shelves versus upper shelves in the supermarket aisle, see Cochoy, 2004) entails recognition that the social environment matters, but test data were analyzed away from the field site, and insofar as the analysis informed intervention, this involved a complex chain of translations, passing through various departments, before any intervention could be made on the shopfloor (changes in product placement, for example). Online environments change this because the underlying data architectures provide for data capture, analysis, and feedback across computational networks, enabling continuous modification of user interfaces, at the front end, based on insights gathered at the back end, with it often being opaque to users which of these operations happen in which settings, inside one's device, in data centers, or somewhere else yet.

But the difference, we suggest, is even more radical because the test situation also need not be *temporally* marked off from the flow of everyday life. Online testing, as in the case of A/B testing of search engine interfaces with different segments of users (Seaver, 2019) is ubiquitous, continuous, and seamless. In principle, in the digital setting—and, increasingly, in digitally equipped settings like a mall or a smart street—anything can be a test situation, any data point can be data for a test of some kind. Because digitally traced behavior (including something so quotidian as a keystroke) can be recorded in a database, and because the very environment where behavior is monitored is equipped to influence behavior, there is the ubiquitous possibility for any subject at any point in time of being experimentally represented and intervened upon. In the digital society today, the measurement of social life is performed as part of the conduct of social life, as every click and retweet generates *at the same time* a social action and a data point (Marres, 2017).

Moreover, social life is tested as part of social life: it is not just represented, but methodically intervened upon, by various actors involved, except that some actors—such as those with the capacity to modify interface settings—occupy an infrastructurally speaking more privileged position to undertake such tests than others. Which is to say, even as testing today extends across environments in society and centers of calculation, it remains very much the case that only some actors can move between these different settings. Testing in society puts social actors in a determinate, unequal relation. As we elaborate in the concluding section, in the real-world test environment as curated by engineers, users can tweak the "settings" on their user interface, but they cannot control, nor are they knowledgeable about, the broader systemic settings of the social media platforms and digitally enhanced environments in which they are participating.

The blurring of lines between the test situation and the world to which the test is meant to refer, is what in our view is being invoked by terms like "smart," "real-world," and "living lab." It is not confined to online environments, but can be found in a variety of technology-intensive settings, for example in the medical field. The medical world was once divided into two domains, the biomedical realm of research and the clinical realm of care—in other words, between *testing* and *treatment*. Because of the uptake of new, data-intensive techniques in experimental protocols and routine treatment protocols, for example, in oncology, "care increasingly displays experimental features" (Cambrosio, Keating, Vignola-Gagné, Besle, & Bourret, 2018, p. 210). With the development of "point-of-care" testing for personalized treatment, testing is folded *into* treatment itself (Beitelshees, 2012). This form of testing, like genetic signature tests, for example, creates

> *contexts where the frontiers between clinical and research activities are constantly moving. Indeed, both activities are often carried out in the same arenas, or are interconnected, to the extent that the actors concerned find their boundaries difficult to draw. (Rabeharisoa & Bouret, 2009, p. 697)*

In some clinical trials research protocols can remain open "on an ongoing basis, so to speak *sine die*" even after publication of trial results, leading some researchers to gesture to "an infinite protocol that probably foreshadows a new way of performing clinical trials" (Cambrosio et al., 2018, p. 216).

4.2.3 | Counter-testing is difficult and dominated by experts

The revelations about personal data extraction and opinion manipulation by social media platforms and platform-based social science were in part the result of counter-studies by individual experts like Olivier Dehaye. However, these revelations did not only result in public criticism of the tech sector, they equally had the paradoxical effect of making independent or critical research on social media more difficult. After the scandal broke, Facebook changed the settings of its Application Programming Interface and app review procedures, with the result that third-party applications including several social research apps, like Bernard Rieder's Netvizz, stopped working.[16] Around the same time, a group of academic researchers supported by the not-for-profit Mozilla Foundation, criticized Facebook for not doing enough to support rigorous social science, saying its tools "do … not satisfy basic criteria for validity and reliability."[13] Even if independent Facebook research—including counter tests—helped expose the harmful fallout of social media testing, it equally had a backlash effect: Facebook's subsequent interventions made counter-testing more difficult.

It is too early to say, but this challenging environment for counter-testing may yet change: The General Data Protection Regulation (GDPR) aims to legislate out of existence the original social media business model, which is based on extraction of personal information without the data subject as transaction partner (Bellanova & González Fuster, 2018). This equally threatens the test logic exemplified by personality testing, which depends on an undemanding informed consent regime to be in place. Are we on the road to a new test ban?

Recall that no individual could opt out of nuclear testing and that this elementary fact was one of the factors leading to a test ban. By contrast, the possibility to opt out—or in—is a notorious grey area in digital media environments (Gerlitz & Helmond, 2013): being listed in a user's contact book—or passing through a smart street—may be all that is needed for someone to be enlisted as a test subject. Efforts are being made to restore the possibility of opting out of online experiments, including through the GDPR, which is designed partly for this purpose. However, it is not clear if the rights-based framework of the GDPR is even enforceable: it has not yet been put to the test, as implementation has only just begun. Furthermore, it is not clear that today's online personality tests are met with the level and intensity of public disapproval that would be required to mobilize sufficient support for a test ban. The intended effect of the GDPR appears to be to produce a market for personal data services, and potentially, testing. Is personality testing ban-able? At the moment it does not look likely. Is it more likely that, rather than an outright ban, we will see a zoning of test environments in society, with low and high thresholds of testability for different types of settings, say a premium app versus Facebook; an upmarket eco-resort versus Disneyworld?

5 | FROM TESTS IN SETTINGS TO TESTING SETTINGS

We have argued that it is not enough for sociology to study testing "in society." True, testing has moved from the laboratory into society, insofar as (a) testing methodologies are today prominent in the social environment, whether it is in the form of street trials of intelligent vehicles conducted (Marres, this volume) or citizen tests conducted as part of immigration procedures by the UK home office and the Dutch border agency (Schinkel, this volume), and (b) testing methodologies developed in science and engineering are today applied to social phenomena, as in the development RCTs studied by Luciana de Souza (this volume), and the social credit system in China (Bach, this volume). But, in these and other cases, something even yet more radical than a move "beyond the laboratory" is taking place. We state it now concisely and then elaborate for clarity: Whereas we traditionally think about testing taking place *within a setting*, today's engineers are *testing the settings*. We used to think that, to understand the relation between tests and society, we must attend to the wider contexts in which tests are conducted. But to understand how testing and the social relate today, we must investigate how testing operates *on* social life, through the modification of its settings. One way to clarify the difference is to say that the new forms of testing can be captured neither within the logic of the field test nor of the controlled experiment. Whereas tests once happened inside social environments, today's tests directly and deliberately modify the environment.

Such modifications of the settings in which social life unfold happen most clearly in computational and computationally inflected environments: on social media platforms, or inside transport environments, for example, with systems in place to monitor, analyze, and regulate movement. But settings in society may also be turned into test environments by the insertion of a mere device into existing social environments, say, a Facebook button on a coffee shop's wifi page, a "smart" traffic light onto a street, or an "intelligent vehicle" with a pedestrian detection feature passing in the street that passes its observation to an automotive data center. In this regard, the critical distinction is that between the idea that testing occurs inside the social environment (a claim that still conforms to the logic of the "field test") and the idea that testing today involves the very modification of social environments.

Take the navigation app Waze. This app relies on mobile data that its users continuously send back to determine the properties of current traffic flow, and to this end this app deploys some of its users as probes: when a traffic jam is forming, most users will get routed around the trouble spot, but some will be deliberately sent into the congested area in order to measure the properties of this disruption, effectively serving as test subjects. These "test settings" of Waze do not just imply that individual settings (give me the fastest route) may be overridden by the app, limiting the agency of users, or at least delaying their arrival time. More important for our purpose, is that by using Waze users willfully subject themselves to experimental settings: not only may a user's settings be updated to "test subject" status without the user knowing, any Waze user is in principle subject to test conditions that may be changed at any point depending on the system needs.

These types of experimental devices—which we can think of as generating *total test environments*[17]—differ both from field tests and from classical, controlled laboratory experiments. We address these in turn. In the case of the field test, the aim is to find an appropriate setting which has the required properties so that a given phenomenon can be observed there (Morgan, 2013). The concurrent expectation is that the increase of complexity and authenticity of what is observed in a field test will inevitably be traded off against a decrease in control over the phenomenon to be studied. Real-world experiments do not escape this tradeoff so much as render it irrelevant—because its logic and its goals are different, at least where testing on social phenomena is concerned. Here, the objective is not to observe an existing social phenomenon—to secure its authenticity. In place of *finding* (selecting, choosing) a setting or sampling among several settings, the operations that produce today's total test environment consist of minor modifications in the environments in society so as to render *the setting capable of data capture, analysis, and feed-back*—that is, to equip it as a test environment, to enable representation *and* intervention—even if aspirationally—on a more or less durable basis.[18]

It is not technically correct to say, in this case, that the test happens inside a given environment in society, insofar as the whole point—or at least a major point—of the test is to modify that environment, to establish connections *between* the field and the laboratory, between entities making up the street environment and the data center, so that it becomes possible to dynamically adjust feedback—say, navigation directions, but it could also be the working of a traffic light—based on measurement—in app-to-data cener and car-to-traffic light communication. It is also to say that we can only understand what is going on in that setting, if we take into account its infrastructural extension, spanning field and lab, street and data center.[19] The test situation (and correspondingly the test setting) does not just arise in the street, but across the street and the data centers it connects to.

But such computationally enabled test environments also differ from the classical, controlled experiment. In a real-world test environment, the propensity of the setting to inform, inflect, or influence the social phenomena that unfold within them is considered—by the engineers of the social—a *positive* feature. By contrast, in the philosophy of science pertaining to the classical experimental model, it was considered *a sign of weakness* if you needed to modify environments in order to induce the phenomenon under scrutiny. Such a brittle, poor, inauthentic test was assumed to lack robustness. By contrast, in the setting of the real-world experiment, the more the environment and the entities that constitute it have the proven capacity to influence and modify the behavior of the entities inside them, the more productive, the more successful, from a scientific and engineering point of view, it is considered to be. As Jonathan Bach (this volume) writes in his essay on the Chinese social credit system, it is a "feature not a failure."

For these reasons we argue that new forms of testing are not occurring within settings so much as they are testing the settings. To figure out what modifications of the smart environment (targeted route redirections, changing settings of traffic lights) enable experimental operations upon societal phenomena (traffic flow, mobility, but also more ephemeral phenomena like trust), of producing the desired modification in the phenomenon under scrutiny (less congestion, increased use of alternative means of transport), settings are continuously modified and continously test users or passers-by. The test-ing settings of experimentation beyond the laboratory, then, are emphatically not the settings that you, the user, can adjust and tweak on your user interface, but the macro settings that are specific to the testing environment itself. Which is also to say, the curation of such test environments in society is intimately connected to the development that we captured hereabove in the slogan: "engineering has moved into the social." As tech industries have overseen an unprecedented extension of infrastructures for data capture, analysis, and feedback across environments in society, in the form of social media, smart transport systems, digital payment, and so on, an extensive environment has become available to scientists and engineers who work with the industries not only to research social life—what is trust?—but also for intervening in it—how to enhance trust?

6 | A CRITICAL MOMENT FOR SOCIOLOGY AND THE SOCIOLOGY OF TESTING

These new developments concerning test settings raise fundamental questions that will need to be addressed by the new sociology of testing. Most crucially—but also most challengingly—we must address the implications for our understanding of "the social": as we detect a shift from tests in settings to testing settings, should we move from studying testing in its social contexts to analyzing the ways in which testing puts social life to the test? We believe the contributions to the Special Issue convincingly make clear that a good place to start is to investigate how testing in society gives rise to test-ing situations: tests do not just render phenomena know-able and actionable, they put social relations at stake (Robinson, this volume; Tironi, this volume). In undertaking such investigation, a new sociology of testing can build on work in pragmatist sociology which focuses on the analysis of *critical moments* (Boltanksi & Thévenot, 1999; Hutter & Stark, 2015) and *trials of explicitness* (Muniesa & Linhardt, 2011; see also Guggenheim, 2014; Marres, 2012), the idea that when habitual ways of doing get interrupted in social life, whether by accident or as a consequence of deliberate disruption, social actors are prompted to articulate their attachments and relations. These moments in which social actors are put to the test are distinctive as "problematic situations."[20] However, in today's total test environments, because test settings are infrastructurally configured, and because testing can be taking place without one's knowledge at any time, critical moments come about in a different way: they arise from practices of testing and counter-testing in diverse modalities—research, art, journalism—as part of an ecology of testing. Furthermore, as the creation of test environments in society is often done with the explicit purpose of governing and influencing social life, a sociology of testing must come to terms with the possibility that the fundamental "logics of testing" that pragmatist sociologists are trained to detect in situations do not necessarily present a constitutive property of social life, but are materially, technically, and politically inflected by exogenous, interested agencies, like engineering, or "strategic niche management," or experimental governance.

The second question the new sociology of testing will need to address concerns the concept of *proxy* that has been so important in STS studies of testing. In Trevor Pinch's canonical paper (Pinch, 1993; see also Downer, 2007), testing is centrally concerned with a problem of representation. In his account, engineering tests are a "proxy" that *stands for* something "out there" in the world. But in real-world testing, the focus is not on the creation of dedicated, controllable test environments away from society, but on the modification of "real-world" settings in society. If the test setting (as we argued above) is no longer spatially and temporally separate from the environment in society, does this notion of proxy still apply? It seems clear that the problem of the proxy—the problem of reference—is not going to go away, even as engineering approaches to knowing society are increasingly prominent today. But it might mean that the new forms of testing are neither proxies that are from the outset designed to *stand for* something "in society" nor critical moments that already *stand out* as specific situations as part of these situations. Testing situations will have to be analyzed as unfolding across settings.

Thirdly and finally, it seems likely to us that the study of testing in society will confront sociologists with a choice: what position should we as sociologists take on the new forms of testing and the engineering of the social? We might be tempted to conclude that the very features that make real-world test environments attractive from the standpoint of social engineering—the possibility to impact society through the creation of testing environments—is what renders these environments utterly unsuitable as a site of enquiry for sociology. It signals that insofar as "social life" can be said to occur in test environments, it is inauthentic, an artefact of engineering. It is true that what goes on in test environments is not representative of all of society. But we misunderstand test environments in society if we approach them in "proxy" mode (if we consider them from the standpoint of the field test, and view it as a bastard version of the controlled experiment, and end up saying: "this is not society, society is over there"). Today's test environments are key sites where forms of life, forms of experience are defined, contested, and indeed, tested today.

What, then, is the other option? A sociology of testing could make it its purpose to specify the new methodologies of social engineering from a critical, reconstructive perspective (Entwistle & Slater, 2019; Marres et al., 2018). Such an approach would not go along with narrow framings of the object and objectives written into testing settings: social life does not just unfold inside the test environment (what is trust on Facebook?), the test environment is very much part of it (we should ask: what does trust *become* on Facebook?). Tests do much more than reveal the properties of environments and phenomena unfolding inside them, *tests* operate on social relations, they may reveal capacities and be deployed to hide them. "Testing" is then the name of a fundamental social effect, to be specified by the new sociology of testing. This is what defines situations sociologically speaking—that *they are testing*. But, whatever the answers to our three questions above, one thing seems clear to us: It is in addressing these and similar theoretical and empirical challenges that sociology faces a critical moment in which it is put to the test.

ACKNOWLEDGMENTS

Work for this paper was supported by an Advanced Research Grant from the European Research Council, grant agreement no. 695256. We are grateful to Pablo Boczkowski, Hjalmar Bang Carlsen, James McNally, Elena Esposito, Jonathan Bach, Fabian Muniesa, Trevor Pinch, Joris van Hoboken, and Michael Guggenheim, for comments, criticisms, and suggestions.

DATA AVAILABILITY STATEMENT

This article is based on a dicussion of academic literature and reports available in the public domain.

ORCID

Noortje Marres https://orcid.org/0000-0002-8237-6946
David Stark https://orcid.org/0000-0003-2435-9619

NOTES

[1] With thanks to Willem Schinkel who brought this quote to our attention in his closing remarks at the workshop "Put to the Test: Critical Evaluations of Testing," Warwick in London, December 10 and 11, 2018.

[2] We use the term "laboratory" loosely in this essay, as referring to dedicated, contained sites where controlled conditions have been created for the detection, monitoring and analysis of defined phenomena.

[3] Ronnell (2007), for example, points to tests as the quintessential element of modernity, signaling the drive to move beyond known limitations.

[4] Whether this development is *really new*, or a continuation or re-activation of previous developments, is a question for the new sociology of testing to answer. In this essay we examine the intuition that existing approaches in the sociology of testing leave us under-equipped to analyze testing in society, and formulate questions that a new sociology of testing will need to address.

[5] MacKenzie (1989) draws on the holistic philosophy of science of Pierre Duhem to make the point that tests are "always open to challenge" (p. 430). In his account, it is the job of the sociology of testing to explain why they are not always challenged in practice. "Beliefs about what constitutes adequate testing are conventional, but that does not make them any less significant" (p. 430). Here, as in the French convention school, the preoccupation is with how tests produce legitimate results.

[6] His article begins with an evocative account of the consequences of ballistic missile testing in situ, citing at length a journalistic description of Kwajalein, one of the Marshall Islands that "[the Pentagon] cleared the inhabitants off … and they are now crowded on a tiny speck about two miles north of Kwajalein Island. … without a reliable water supply, without proper medical care, and even without sufficient room to bury their dead…" (p. 409).

[7] ICBM testing—the production of accuracy estimates for intercontinental ballistic missile trajectory predictions at the Vandenberg Air Force base in California—fit both meanings of "hard." Estimates and equations were the hard stuff of engineering, and their isolation from "society" made them a hard (i.e., difficult) case in which to discover the social.

[8] Although a different terminology from that of SSK, this ANT vocabulary nonetheless places limits on our understanding of the role of testing in society today. Concepts like enrollment (Callon, 1984), as well as script (Akrich, 1992), led actor-network theorists and researchers to analyze the role of social actors in technology testing primarily in terms of compliance and resistance. However, as we will discuss below, testing in society today raises different questions: the question is not whether or not social actors are enrolled by means of the test, but which modes of knowing, valuing, and acting get activated as social actors engage with the test: do they, for example, create art works, deploy counter-expertise, or initiate counter-testing, as we will see later on.

[9] Actor-network theory had of course already demonstrated that tests and testing offer powerful instruments for transforming society, by virtue of their displacement—and transportability—from dedicated test settings to environments in society: for example, the displacement of a "smell test" from the perfume laboratory into focus group settings, and/or perfume shops, provides a powerful instrument for creating new products, new experiences, new preferences, and new markets (Latour, 2004; Muniesa, 2014). However, whereas for actor-network theorists, the transformative capacity of test critically depends on the re-production of contained laboratory conditions in social environments (the "laboratization thesis"), we make a different argument: it is through the introduction of tests into distinctively social environments (the street, the city square, social media conversations)—precisely not laboratories, but societal spaces—that science and engineering are today gaining the capacity to extend engineering logics into distinctively social phenomena—trust, interaction in public space, collective behaviour, well-being, and so on.

[10] With thanks to Trevor Pinch for suggesting this formulation.

[11] Michael Guggenheim (2012) argues that the Chicago's School definition of the city as laboratory should be understood metaphorically, as a rhetorical strategy design to make space for social science. We agree here below that the Chicago School approach to the city as laboratory stands out for its lack of interventions in urban environments (they did not intervene to make the city more like a laboratory). However, we go on to develop a different perspective on experiments conducted in urban environments: in our view these experiments precisely do not fit the definition of laboratory that Guggenheim relies on ("a procedure that often results in a space with the properties to separate controlled inside from uncontrolled outside"—p. 101), which defines experiments in terms of the ability to contain phenomena in a controlled environment. This condition of containment, in our view, is lifted, or at least relaxed, in contemporary instances of real-world experimentation.

[12] Sociologist Harry Collins viewed science as an "examplar" and analogy for "all other forms of social and conceptual innovation" (1985, p. 165) where he proposed "science is a representative example of cultural activity …" expressing the hope that "social and political scientists will be able to use his and other modern studies of science to illuminate more general problems" (p. 166).

[13] Similarly, Gross's idea that social experimentation is "different" (Gross & Krohn, 2005) from experimental social science reminds us of Latour's (1988) exclamation in the Pasteurization of society: to transform society with the aid of the vaccine—with the aid of a laboratory—in his account "is completely different" from all other attempts to change society by other means (the law, political mobilization, fashion).

[14] https://jamesbridle.com/works/autonomous-trap-001.

[15] Robinson (2016) discusses a US court ruling that "pregnancy" was not a disease.

[16] Bernhard Rieder, creator of the app for Facebook network analysis and visualization Netvizz, submitted his app for review and was refused access to Facebook data on the grounds that "permissions data must be visibly used within your app" (see Rieder, 2018).

[17] With thanks to Chris Anderson (Leeds University) who suggested this term to us during the December 2018 Put to the Test workshop in London.

[18] They are in some ways similar to what Morgan (2013) calls "society's experiments"—"situations [which] present themselves as if an experimenter has designed a laboratory experiment within the world" (p. 346). Except that in real-world tests, experimenters do in fact design at least some aspects of the test environment in society.

[19] This operation is invoked by Gieryn (2006) in his account of the Chicago School (see the quotation below). However, where continuity between field and lab in Gieryn's account is a methodological accomplishment, in real-world experimentation this operation takes the form of socio-technical and material modifications of connected environments in society so as to create a continuous test environment (for a discussion, also see Marres, 2017, p. 53). "Authors of the Chicago School oscillate between making Chicago (the city) into a laboratory and a field-site. On some occasions, the city assumes the qualities of a lab: a restricting and controlling environment, whose placelessness enables generalizations to 'anywhere,' and which demands from analysts an unfeeling detachment. On other occasions, the same city becomes a field-site, and assumes different qualities: a pre-existing reality discovered by intrepid ethnographers who develop keen personal sensitivities to the uniquely revealing features of this particular place" (Gieryn, 2006).

[20] The notion can be found in John Dewey's emphasis that inquiry—discovery—happens in "troubled, perplexing, trying situations" (Dewey, 1998, p. 140). For Mische and White (1998), a "situation" is a special kind of setting, defined by them as "problematic, … episodes that cast our prescribed roles and trajectories into question" (p. 697). For more on Dewey's notion of inquiry see Stark (2011, pp. 2–9); on methodological situationalism see Stark (2017, 2011, pp. 32, 185–186). Formilan and Stark (this volume) examine critical moments in which electronic music artists probe and test identities.

REFERENCES

Agüera y Arcas, B., Todorov, A., & Mitchell, M. (2018, January 11). Do algorithms reveal sexual orientation or just expose our stereotypes? *medium.com*. Retrieved from https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477

Akrich, M. (1992). 'The de-scription of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society: Studies in sociotechnical change* (pp. 205–224). Cambridge, MA: MIT Press.

Arms Control Association. (2019, February). The nuclear testing tally. *armscontrol.org*. Retrieved from https://www.armscontrol.org/factsheets/nucleartesttally

Ausloos, J. (2018, April 10). Paul-Olivier Dehaye and the raiders of the lost data. *law.kuleuven.be*. Retrieved from https://www.law.kuleuven.be/citip/blog/paul-olivier-dehaye-and-the-raiders-of-the-lost-data/

Bacharach, Y., Grepel, T., Kohli, P., Kosinski, M., & Stillwell, D. (2014). Your digital image: Factors behind demographic and psychometric predictions fro social network profiles. In *AAAMAS '14: Proceedings of the 2014 International Conference on Autonomous Agents and Multi-Agent Systems*. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

Bellanova, R., & González Fuster, G. (2018). No (Big) Data, no fiction? Thinking surveillance with/against Netflix. *The Politics and Policies of Big Data: Big Data Big Brother*.

Beitelshees, A. L. (2012). Personalised antiplatelet treatment: A RAPIDly moving target. *The Lancet*, *379*(9827), 1680–1682.

Bijker, W. E., & Pinch, T. (2012).Preface to the anniversary edition. In W. E. Bijker, T. Hughes, & T. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. xi–xxxiv). Cambridge, MA: MIT Press.

Boltanski, L., & Thévenot, L. (1999). The sociology of critical capacity. *European Journal of Social Theory*, *2*(3), 359–377.

Boltanski, L., & Thévenot, L. (2006). *On justification*. Princeton, NJ: Princeton University Press.

Bond, R. M., Fariss, C. J., Jones, J. J., Kramer, A. D. I., Marlow, C., Settle, J. E., & Fowler, J. H. (2012). A 61-million-person experiment in social influence and political mobilization. *Nature*, *489*, 295–298.

Boyle, A. (2013, March 12) Gay? Conservative? High IQ? Your Facebook "likes" can reveal traits. *nbcnews.com*. Retrieved from https://www.nbcnews.com/sciencemain/gay-conservative-high-iq-your-facebook-likes-can-reveal-traits-1C8805606

Bruno, L. A. (2003). The bequest of the nuclear battlefield: Science, nature and the atom bomb during the first decade of the Cold War. *Historical Studies in the Physical and Biological Sciences*, *33*(2), 237–260.

Buchmann, S., Lafer, I., & Ruhm, C. (2016). *Putting rehearsals to the test practices of rehearsal in fine arts, film, theater, theory, and politics*. Berlin, Germany: Sternberg Press.

Bulman, M. (2019, April 27). Home Office under investigation after 1,000 suddenly deported over English test cheating claims. *independent.co.uk*. Retrieved from https://www.independent.co.uk/news/uk/home-news/home-office-student-visa-english-language-test-cheat-a8888926.html

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *theguardian.com*. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Cage, J. (1973). Experimental music: Doctrine. *Silence*. Middletown, CT: Wesleyan University Press.

Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. InJ. Law (Ed.), *Power, action, and belief: A new sociology of knowledge?* (pp. 196–233). London, UK: Routledge.

Callon, M. (1986).The sociology of an actor-network: The case of the electric vehicle. In M. Callon, J. Law, & A. Rip (Eds.), *Mapping the dynamics of science and technology* (pp. 19–34). London, UK: Palgrave Macmillan.

Cambrosio, A., Keating, P., Vignola-Gagné, E., Besle, S., & Bourret, P. (2018). Extending experimentation: Oncology's fading boundary between research and care. *New Genetics and Society*, *37*(3), 207–226.

Cochoy, F. (2004).Is the modern consumer a Buridan's donkey? Product packaging and consumer choice.In K. M. Ekström, & H. Brembeck (Eds.), *Elusive consumption* (pp. 205–228). London, UK: Bloomsbury.

Collins, H. M. (1981). Son of seven sexes: The social destruction of a physical phenomenon. *Social Studies of Science*, *11*(1), 33–62.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London, UK: SAGE.

Consumer Reports. (1959, March). The milk all of us drink and fallout. *Consumer Reports*, *24*.

Davis, T. C. (2007). *Stages of emergency: Cold war nuclear civil defence*. Durham, NC: Duke University Press.

Dehaye, P.-O. (2017, March 3). Cambridge Analytica demonstrably non-compliant with data protection law. *medium.com*. Retrieved from https://medium.com/personaldata-io/cambridge-analytica-demonstrably-non-compliant-with-data-protection-law-95ec5712b61

DeVeaux, S. (1997). *The birth of bebop: A social and musical history*. Berkeley, CA: University of California Press.

Dewey, J. (1954). *The public and its problems*. Chicago, IL: Swallow Press.

Dewey, J. (1998[1933]). Analysis of reflective thinking. *The Essential Dewey*, *2*, 137–144.

Dewey, J. (1998[1925]). *Experience and nature*. New York, NY: Dover Publications.

Downer, J. (2007). When the chick hits the fan: Representativeness and reproducibility in technological tests. *Social Studies of Science*, *37*(1), 7–26.

Ellison, N., Heino, R., & Gibbs, J. (2006). Managing impressions online: Self-presentation processes in the online dating environment. *Journal of Computer-Mediated Communication*, *11*(2), 415–441.

Engels, F., Wentland, A., & Pfotenhauer, S. M. (2019). Testing future societies? Developing a framework for test beds and living labs as instruments of innovation governance. *Research Policy*, *48*(9), 103826.

Entwistle, J., & Slater, D. (2019). Making space for "the social": Connecting sociology and professional practices in urban lighting design. *British Journal of Sociology*, *70*(5), 2020–2041.

Gerlitz, C., & Helmond, A. (2013). The like economy: Social buttons and the data-intensive web. *New Media & Society*, *15*(8), 1248–1365.

Gieryn, T. F. (2006). City as truth-spot: Laboratories and field-sites in urban studies. *Social Studies of Science*, *36*(1), 5–38.

Gieryn, T. F. (2018). *Truth-spots: How places make people believe*. Chicago, IL: University of Chicago Press.

Gross, M. (2009). Collaborative experiments: Jane Addams, Hull House and experimental social work. *Social Science Information*, *48*(1), 81–95.

Gross, M., & Krohn, W. (2005). Society as experiment: Sociological foundations for a self-experimental society. *History of the Human Sciences*, *18*(2), 63–86.

Guggenheim, M. (2012). Laboratizing and de-laboratizing the world: changing sociological concepts for places of knowledge production. *History of the Human Sciences*, *25*(1), 99–118.

Guggenheim, M. (2014). Introduction: Disasters as politics—Politics as disasters. *The Sociological Review*, *62*(Suppl. 1), 1–16.

Hennion, A. (2003). Music and mediation: Towards a new sociology of music. In M. Clayton, T. Herbert, & R. Middleton (Eds.), *The cultural study of music: A critical introduction* (pp. 80–91). London, UK: Routledge.

Hutter, M., & Stark, D. (2015).Pragmatist perspectives on valuation: An introduction. In A. B. Antal, M. Hutter, & D. Stark (Eds.), *Moments of valuation: Exploring sites of dissonance* (pp. 4–16). Oxford, UK: Oxford University Press.

John, O. P., & Srivastava, S. (1999). The Big Five Trait taxonomy: History, measurement, and theoretical perspectives. In L. A. Pervin, & O. P. John (Eds.), *Handbook of personality: Theory and research* (pp. 102–138). Guilford Press.

Knorr-Cetina, K. D. (1981).The micro-sociological challenge of macro-sociology: Towards a reconstruction of social theory and methodology. In A. V. Cicourel & K. D. Knorr-Cetina (Eds.), *Advances in social theory and methodology: Toward an integration of micro-and macro-sociologies* (pp. 1–47). Boston, MA: Routledge & Kegan Paul.

Kosinski, M., Stillwell, D. J., & Graepel, T. (2013). Private traits and attributes are predictable from digital records of human behavior. *Proceedings of the National Academy of Sciences of USA: US National Academy of Science*, *110*(15), 5802–5805.

Lambiotte, R., & Kosinski, M. (2012). Tracking the digital footprints of personality. *Proceedings of the IEEE*, *102*(12), 1934–1939.

Latour, B. (1993). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

Latour, B. (1996). *Aramis, or the love of technology*. Cambridge, MA: Harvard University Press.

Latour, B. (2004). How to talk about the body? The normative dimension of science studies. *Body & Society*, *10*(2–3), 205–229.

Laurent, B., & Tironi, M. (2015). A field test and its displacements. Accounting for an experimental mode of industrial innovation. *CoDesign*, *11*(3/4), 208–221.

Lezaun, J., Muniesa, F., & Vikkelsø, S. (2013). Provocative containment and the drift of social-scientific realism. *Journal of Cultural Economy*, *6*, 278–293.

MacKenzie, D. (1989). From Kwajalein to Armageddon? Testing and the social construction of missile accuracy. In D. Gooding, T. Pinch, & S. Schaffer (Eds.), *The uses of experiment: Studies in the natural sciences* (pp. 409–436). Cambridge, UK: Cambridge University Press.

Marres, N. (2012). *Material participation: technology, the environment and everyday publics*. Basingstoke, UK: Palgrave.

Marres, N. (2017). *Digital sociology*. Cambridge, UK: Polity Press.

Marres, N., Guggenheim, M., & Wilkie, A. (2018).Introduction: From performance to inventing the social. In N. Marres, M. Guggenheim, & A. Wilkie (Eds.), *Inventing the social* (pp. 17–37). Manchester, UK: Mattering Press.

Mills, M. A. H., & Tkaczyk, V. (in press). *Testing hearing: The making of modern aurality*. Oxford, UK: Oxford University Press.

Mische, A., & White, H. (1998). Between conversation and situation: Public switching dynamics across network domains. *Social Research*, *65*(3), 695–724.

Mody, C. C., & Lynch, M. (2010). Test objects and other epistemic things: A history of a nanoscale object. *The British Journal for the History of Science*, *43*(3), 423–458.

Morgan, M. S. (2013). Nature's experiments and natural experiments in the social sciences. *Philosophy of the Social Sciences*, *43*(3), 341–357.

Muniesa, F. (2014). *The provoked economy: Economic reality and the performative turn*. Routledge.

Muniesa, F., & Linhardt, D. (2011). Trials of explicitness in the implementation of public management reform. *Critical Perspectives on Accounting*, *22*(6), 550–566.

Murphy, M. (2006). *Sick building syndrome and the problem of uncertainty: Environmental politics, technoscience, and women workers*. Durham, NC: Duke University Press.

Neff, G., & Stark, D. (2004). Permanently beta: Responsive organization in the internet era. In P. N. Howard & S. Jones (Eds.), *Society online: The internet in context* (pp. 173–188). Thousand Oaks, CA: SAGE.

Palmer, V. M. (1928). *Field studies in sociology: A student's manual*. Chicago, IL: University of Chicago Press.

Pinch, T. (1993). "Testing—One, Two, Three... Testing!": Toward a sociology of testing. *Science, Technology, & Human Values*, *18*(1), 25–41.

Popper, K. (1961). *The poverty of historicism* (3rd ed.). New York, NY: Harper & Row.

Popper, K. (2002). *The logic of scientific discovery*. London, UK: Routledge.

Rabeharisoa, V., & Bourret, P. (2009). Staging and weighting evidence in biomedicine: Comparing clinical practices in cancer genetics and psychiatric genetics. *Social Studies of Science*, *39*(5), 691–715.

Reiss, L. Z. (1961). Strontium-90 absorption by deciduous teeth. *Science*, *134*(3491), 1669–1673.

Rieder, B. (2018, August 11). Facebook's app review and how independent research just got a lot harder. *thepoliticsofsystems.net*. Retrieved from http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/

Robinson, J. H. (2016). Bringing the pregnancy test home from the hospital. *Social Studies of Science*, *46*(5), 649–674.

Ronnell, A. (2007). *The test drive*. Champaign, IL: University of Illinois Press.

Seaver, N. (2019). Knowing algorithms. In J. Vertesi & D. Ribes (Eds.), *Digital STS* (pp. 412–422). Princeton, NJ: Princeton University Press.

Stannard, J. N. (1988). *Radioactivity and public health: A history*. Washington, DC: Office of Scientific and Technical Information.

Stark, D. (2011). *The sense of dissonance: Accounts of worth in economic life*. Princeton, NJ: Princeton University Press.

Stark, D. (2017). For what it's worth. *Research in the Sociology of Organizations*, *52*, 383–397.

Thrasher, F. (1927). *The gang: A study of 1313 gangs in Chicago*. Chicago, IL: University of Chicago Press.

Tolman, F. (1902). The study of sociology in institutions of learning in the United States. II. *American Journal of Sociology*, *8*(1), 85–121.

UNSCEAR. (2000). *Report to the general assembly (Annex C: Exposures from man-made sources of radiation*. Retrieved from http://www.unscear.org/docs/reports/annexc.pdf

van de Poel, I., L. Asveld, & D. C. Mehos (Eds.). (2017). *New perspectives on technology in society: Experimentation beyond the laboratory*. London, UK: Routledge.

Vertesi, J. (2015). *Seeing like a rover: How robots, teams, and images craft knowledge of mars*. Chicago, IL: University of Chicago Press.

Wan, Y., & Kosinski, M. (2017). Deep neutal networks are more accurate than humans at detecting sexual orientation from facial images. *PsyArXiv* preprint. Retrieved from https://osf.io/zn79k/

Watkins, E. (2001). Radioactive fallout and emerging environmentalism: Cold war fears and public health concerns, 1954–1963. In G. E. Allen & R. M. McLeod (Eds.), *Science, history and social activism: A tribute to everett mendelsohn* (pp. 291–306). Alphen aan den Rijn: Kluwer.

Woolgar, S. (1990). Configuring the user: The case of usability trials. *The Sociological Review*, *38*(Suppl. 1), 58–99.

Wu, T. (2017). *The attention merchants: The epic scramble to get inside our heads*. New York, NY: Penguin Random House.

Yaneva, A. (2005). Scaling up and down: Extraction trials in architectural design. *Social Studies of Science*, *35*(6), 867–894.

How to cite this article: Marres N, Stark D. Put to the test: For a new sociology of testing. *Br J Sociol*. 2020;71:423–443. https://doi.org/10.1111/1468-4446.12746

Received: 12 March 2019 | Revised: 4 October 2019 | Accepted: 7 November 2019

DOI: 10.1111/1468-4446.12723

SPECIAL ISSUE

WILEY

What's on trial? The making of field experiments in international development

Luciana de Souza Leão

Department of Sociology, College of Literature, Science and the Arts, University of Michigan, Ann Arbor, MI

Correspondence
Luciana de Souza Leão, Department of Sociology, College of Literature, Science and the Arts, University of Michigan, 500 South State Street, Room 3224, Ann Arbor 48109, MI, USA.
Email: lsleao@umich.edu

Abstract

In the last 20 years, the drive for evidence-based policymaking has been coupled with a concurrent push for the use of randomized controlled trials (RCTs) as the "gold-standard" for generating rigorous evidence on whether or not development interventions work. Drawing on content analysis of 63 development RCTs and 4 years of participant observation, I provide a rich description of the diverse set of actors and the transnational organizational effort required to implement development RCTs and maintain their "scientific status." Particularly, I investigate the boundary work that proponents of RCTs—also known as *randomistas*—do to differentiate the purposes and merits of testing development projects from doing them, as a way to bypass the political and ethical problems presented by adopting the experimental method with foreign aid beneficiaries in poor countries. Although *randomistas* have been mostly successful in differentiating RCTs from the projects evaluated, I also examine cases where they were not able to do so, as a means to highlight the controversies associated with implementing RCTs in international development.

KEYWORDS

economics, field experiments, global poverty, international development, NGOs

1 | INTRODUCTION

In the last 20 years, the drive for evidence-based policymaking has been coupled with a concurrent push for the use of randomized controlled trials (RCTs) as the "gold-standard" for generating rigorous evidence for whether or not development interventions work. Initially restricted to a handful of researchers located in Poverty Labs within economics departments in the US, RCTs are now being used to test almost everything from strategies to reduce electoral corruption in Sierra Leone to microcredit programs in Peru (Banerjee & Duflo, 2011). In the development community, if researchers, governmental officials, and donors want to know "what really works", it is widely accepted that RCTs must be implemented to avoid a naïve or biased answer (Deaton, 2010; Harrison, 2011). Furthermore, the growing institutionalization of the method is evidenced by wide coverage in the press and the conferment of several of the most prestigious awards for contributions to economics on Esther Duflo, a leading development economist and RCT advocate (Ogden, 2016).

Development RCTs institute experiments in everyday life to measure the impact of different poverty-alleviation policies by comparing the results of treatment and control groups. A 1997 primary school deworming RCT is representative of this type of experimentation (Miguel & Kremer, 2004). In this RCT, the goal was to estimate the effect of intestinal diseases on educational outcomes for young children. Researchers selected 50 schools in rural Kenya to receive deworming medicine for free, and 25 schools were selected as controls that initially did not participate in the program. Researchers compared educational outcomes in the control and treatment groups and found considerable improvements in test scores and school attendance not only for students in treatment schools, but also spillover effects for kids that did not receive treatment. This finding led to a policy recommendation of distributing school-based deworming pills throughout the developing world, and similar programs had reached over 285 million children by 2017 (JPAL, 2017). Experiments like this are now implemented in distinct policy areas trying to answer a variety of development questions. In common, they use the comparison between control and treatment groups to test the effectiveness of development projects in countries from the Global South.

How are US-based researchers able to implement field experiments in developing countries and to persuade others about their "scientific status"? Similar to other forms of field experimentation, the implementation of development RCTs requires ongoing negotiations between the need to control the "messiness" of the field for scientific purposes, while guaranteeing the cooperation, access, and buy-in of local populations so that the experiment can happen (Henke, 2000; Kohler & Vetter, 2016). As field-sites, however, developing countries present particular challenges for the successful balance between control and cooperation. These are places where any type of foreign intervention is inevitably entangled in the controversial politics of the foreign aid industry (Escobar, 2012; Ferguson, 1990), making the distinction between development practice and research particularly blurry (Rayzberg, 2019a). How do development economists bypass the ethical and political controversies associated with foreign aid and convince others that developing countries can serve as appropriate sites to test economic theories?

To answer these questions, I adopt concepts from science and technology studies to describe the network that needs to be in place to implement development RCTs and to make them scientifically and politically plausible, reproducible, and disseminated (Eyal, 2013). Particularly, I highlight the multiple actors involved in implementing these experiments and the complex chains of transcriptions required to generate data from the messy reality of the field and turn it into something useful for both academics and policy officials (Latour, 1999). I demonstrate that proponents of development RCTs, or *randomistas* as they became known (Deaton, 2010), build on the ambiguity of what is being tested—is it an economic theory, a development project, or both?—to navigate the political, practical, and ethical problems associated with development aid and with randomly assigning social policy beneficiaries to treatment and controls groups. In doing so, I also argue that the scientific success of development RCTs is contingent on their ability to construct the image of "the field" as a place that is free of politics and bureaucratic interference. For the most part, *randomistas* have been successful in creating a boundary between the purposes and merits of testing development projects and doing them. However, I also examine cases where they were not successful in order to highlight the controversies associated with implementing RCTs in international development.

In recent years, a number of social scientists have started to problematize development RCTs, mostly by highlighting the strategies that *randomistas* adopt to transform contested development questions into seemingly technical problems (e.g., Berndt, 2015; Deaton & Cartwright, 2016; Rayzberg, 2019a). This paper contributes to this scholarly conversation about development RCTs in two key ways. First, by combining content analysis of 63 RCTs and ethnographic data collected during 4 years of fieldwork, it provides a systematic empirical account of how these experiments operate on the ground. This account highlights the diversity of actors involved in RCTs, while so far attention has been given exclusively to the *randomistas* themselves (exceptions are Kabeer, 2019; Rayzberg, 2019b). Second, this paper unveils the processes through which development RCTs can lose their scientific status, helping to elucidate the problematic nature of RCTs from the viewpoint of the actors most affected by these experiments. In doing so, it contributes to a broader understanding of the politics of testing in international development and the inequalities inherent to the diffusion of global evaluation standards.

The article is organized as follows. The first section explains what development RCTs are and illustrates how they differ from previous experiments done in economics. Second, I present my methods and data. Relying on ethnographic data, in the third section, I describe the organizational network that needs to be in place to make development RCTs work and how actors involved in this network deal with controversies associated with RCTs to make them credible to academic and policy audiences. Fourth, I explain the strategies that *randomistas* adopt to bypass the problematic nature of using the experimental method in developing countries, and I address the main ways that development RCTs are contested. Finally, I conclude by drawing implications from my case for a critical evaluation of testing.

2 | PREVIOUS EXPERIMENTATION IN ECONOMICS AND THE NOVELTIES OF DEVELOPMENT RCTS

Experimental trials in economics have a long tradition. In the 1960s and 1970s, the Negative Income Tax Trials used the experimental method to test the effectiveness of different social policies and had long-term effects on welfare debates in the United States (Rogers-Dillon, 2004). Likewise, during the 1980s and 1990s, economic research took the form of laboratory experiments, testing different hypotheses on behavioral economics that were highly influential for the conceptualization of electronic markets and behavioral theory (Guala, 2007). The development RCTs and Poverty Labs that are the object of this study, even if they build on the prestige of these previous experiments, differ from early economics experiments in two ways.

First, in contrast to most behavioral economics research that takes the form of laboratory experiments (in which volunteers enter a research lab to make decisions in a controlled environment), development RCTs function as field experiments: they are not only performed in the field, but also the division between control and treatment groups is done with real people, schools, and communities living their everyday lives. As *randomistas* themselves point out: “There may be more to learn about human behavior from the choices made by Kenyan farmers confronted with a real choice than from those made by American undergraduates in laboratory conditions” (Duflo, 2003, p. 8). This means that development RCTs, at least in their design, do not have to deal with criticism about “mock settings” or “stage action”: the experiments take place in situ (MacKenzie, Muniesa, & Siu, 2007). Instead, similar to other field sciences, they face a different type of credibility challenge, namely, the need to continuously demonstrate that experiments in developing countries could retain certain characteristics of lab sciences, such as generating generalizable, “placeless knowledge and being inconsequential” (Guggenheim, 2012, p. 102; Kohler & Vetter, 2016).

Second, while development RCTs and the social experiments that took place in the United States in the 1960s similarly happen in the field, they differ in scale, objectives, and geographical reach. In a recent publication from the main Poverty Lab (JPAL), researchers differentiate their experiments from the past ones on the following basis:

> *Unlike the early “social experiments” conducted in the United States…many of the RCTs that have been conducted in recent years in developing countries have had fairly small budgets, making them affordable*

> *for development economists. Working with local partners on a smaller scale has also given more flexibility to researchers, who can often influence program design. As a result, RCTs have become a powerful research tool.* (Duflo et al., 2007, p. 3)

"Researchers who can often influence program design." This comment points to the key difference between development RCTs in the 2000s and earlier "social experiments" in the United States: while the latter partnered with US government agencies, current RCTs are implemented together with non-governmental organizations (NGOs) and foreign aid agencies in developing countries, allowing for much more flexibility to use the experimental method in field conditions than before (de Souza Leao & Eyal, 2019). Hence, while the 1960s' social experiments could not assign participants randomly to a "no-treatment" control group and used the non-sampled population as their implicit control for comparisons (Riecken & Boruch, 1975), development RCTs seek to randomly assign beneficiaries to control groups, attempting to portray in this way an image of greater scientificity than previous experiments. Yet, since development RCTs are implemented in remote areas of developing countries, this also means that these experiments lack administrative data that is available in developed countries; they have to deal with language and cultural barriers, besides having to cope with greater levels of uncertainty and risk (Teele, 2014).

Furthermore, because development RCTs are implemented by mostly US-based researchers in developing countries, the distinctions between economic knowledge-making and development policy governance is particularly tricky to manage. As we will see in the third section, this is because development RCTs are more than simple hypothesis-testing instruments for economic theory, they are also used to redistribute social resources, to measure the impact of development projects, and to propose directions for future foreign aid interventions (Rottenburg et al., 2015). In this process, the line between what counts as experimentation and what counts as a development project is constantly in flux, posing similar ethical questions as processes used for recruiting human subjects for global clinical trials (Petryna, 2009; Rottenburg, 2009), related to whether development RCTs exploit or aim to help the global poor.

In sum, development RCTs combine aspects of both forms of previous experimentation in economics, but applied to a novel territory of intervention, that is, developing countries. On the one hand, development RCTs retain some characteristics of lab science, such as control groups and the aim to identify the causal effect of interventions. On the other hand, they are decisively in the field, as they are conducted "in the messiness" of everyday life of developing countries, "where borders cannot be effectively policed" (Henke, 2000, p. 484). Building on Gieryn (2006, p. 32), therefore, similar to other field sciences, development RCTs gain legitimacy by "preserv[ing] and draw[ing] simultaneously—and in a complementary way—the assumed distinctive virtues of both lab and field." How do *randomistas* manage to do so?

3 | METHODS AND DATA

To answer this question, I use two analytical strategies. First, I build on a dataset constructed as part of a larger project, in which I compiled a random sample of 63 RCTs done by the Abdul Latif Jameel Poverty Action Lab at MIT, or JPAL, led by the charismatic scholar-activist Esther Duflo. My analytical sample was defined on January 13, 2016. Of the 625 RCTs that were listed in J-PAL's library that day, I excluded 100 studies that were not conducted in developing countries. From 525 RCTs, I randomly selected 100 RCTs to be analysed. I then excluded all RCTs that were still ongoing or for which I could not identify a corresponding publication, arriving at a final sample of 63 RCTs. For each RCT publication, I coded information regarding their study design, the authors, and relevant information about implementing and funding partners (see the Appendix for descriptive statistics of the sample).

Second, my account relies on 4 years of participant observation with *randomistas*. During this period, I participated in two RCTs related to microfinance in Peru (2007), and one RCT in the field of financial education in

Brazil (2010–2012). For the purposes of this paper, I complemented this ethnographic data with an analysis of the controversy regarding development RCTs that appears in academic and policy debates. The publications analysed include academic articles, blog posts, and public interviews given by *randomistas* and international policy actors. It is to the analysis of this data that I now turn.

4 | THE MAKING OF DEVELOPMENT RCTS: ACTORS, PROCESSES, AND CONTROVERSIES

In her TED Talk, Esther Duflo (2010) explained how *randomistas* would revolutionize the international development field:

> *It's not the Middle Ages anymore, it's the 21st century. And in the 20th century, RCTs have revolutionized medicine by allowing us to distinguish between drugs that work and drugs that don't work. You can do the same randomized controlled trial for social policy. You can put social innovation to the same rigorous, scientific tests that we use for drugs.*

As has been extensively documented, however, the "revolution" that RCTs brought to the medical field depended on a contentious process that required political and organizational efforts to convince the multiple actors involved of the possibilities of adopting the experimental method with human subjects (Carpenter, 2010). The same is true for development RCTs. Implementing the experimental method in field conditions to answer international development questions is a huge task that involves multiple actors and resources.

In this section, I build on both my RCT sample and on my ethnographic data to describe how economists implement field experiments in developing countries; how they enrol relevant actors into field experiments; and the chain of transcriptions required to generate data from poor individuals' behavior in the field that will then be published in academic journals. The findings in this section unveil the type of stakeholder enrolment and organizational efforts that enable *randomistas* to establish the idea that developing countries are appropriate field-sites for scientific analysis—that is, sites that are not contaminated by geopolitical interests or bureaucratic politics. Through this assessment, I also demonstrate how the boundary work done to differentiate development research and practice allows *randomistas* to productively dismiss failures in field experiments as ideological or bureaucratic problems.

4.1 | The RCT network: A diverse set of actors

While much attention has been paid to *randomistas*, who are the public faces of RCTs and arguably the most influential actors in the network (Ogden, 2016), implementing any given development RCT requires collaboration among a number of actors: the research team, the fieldwork team, survey firms, policy beneficiaries, funding agencies, and implementing partners. Below, I briefly describe each of these in order to highlight the diverse set of actors involved in development RCTs, as well as the internal negotiations that happen among these actors to make these experiments possible.

1. Research Team: The implementation of an RCT starts with the Research Team, based in Poverty Labs within economic departments, such as JPAL at MIT. In my analytical sample, 93% of these researchers had received their PhDs in Economics, and 7% graduated in Political Science. *Randomistas* have academic and administrative roles: they formulate the research question, design the experiment, analyse the data, and publish the academic papers, but they also have a key role in convincing the policy community of

the relevance of their work, in building the reputation of their labs, and in negotiating with funding agencies.

2. Field Team: Although the Research Team formulates the experimental design, implementing an RCT requires that an extended group of professionals work directly in developing countries, close to the project location. These individuals have a different profile than *randomistas*. From the 123 local staff that appear in my sample, 45% had graduate degrees in Economics and Public Policy, but many instead had graduate degrees in Development Studies (20%) and even in the Humanities (12%). The Field Team has a diverse range of tasks—from explaining the experimental methodology to partners implementing the policy and asserting the quality of the experiment, to hiring survey firms and solving any unexpected problems in the field. As a *randomista* explained to me, they are considered "the voice and eyes of the researchers" in the project site, and play a key role in controlling the quality of the experiment. Below, I will show how these actors also have great discretionary power in the transcription process involved in RCTs.
3. Implementing Partners: These are the organizations whose policies will be evaluated by the Research Team. Partner organizations can be divided into high-level decision-makers and their staff (which I treat as a separate actor). The vast majority of RCTs are conducted with NGOs: in my analytical sample, for example, 75% of implementing partners are NGOs or for-profit organizations involved in microfinance projects together with local NGOs. In my ethnographic work, I observed that these NGOs are led by a highly educated group with extensive international experience, which came either from their relationship with foreign aid agencies or professional training. The role of these policy managers in RCT implementation is to establish the institutional partnership with Poverty Labs, provide financial and infrastructural resources, and disseminate results together with *randomistas*.
4. NGO Staff: This group is responsible for the day-to-day work of development organizations. Staff members include teachers, nurses, and micro-credit agents, among others. While I did not have information about these actors in my RCT sample, during my fieldwork, I could examine the key features of their profile: they have either secondary education or a BA degree in a less prestigious field of study; they usually do not speak English, but they have a great deal of tacit knowledge regarding the policy being implemented. NGOs' staff are the ones that have their daily activities most affected by the experiment: they modify their practices to respect the treatment and control group division, report back to the Field Team and the high-level staff about their activities, and make the connection between the survey firm and policy beneficiaries, while also being directed evaluated.
5. Policy Beneficiaries: These are the individuals, or the local population, who are affected by the policy being tested and will respond to the survey questionnaires. *Randomistas* refer to them as "The Poor" (a broad category that involves a multiplicity of groups) whose behavior they are trying to understand and shape. Results from field experiments depend on their willingness to answer survey questionnaires properly, but they have no influence over the design of the RCTs. Yet, all administrative data available of their behavior and socioeconomic background is closely monitored to assess the impact of development policies.
6. Survey Firms: In order to obtain data from development RCTs, a crucial part of the process is hiring and training either a local survey firm or independent surveyors. In my fieldwork, I noticed that surveyors had a university degree and experience with survey implementation, but no international work experience or English proficiency. Surveyors are trained to implement the questionnaires with the policy beneficiaries, are taught the basics of the experimental method, and have their activities closely supervised by the Field Team.
7. Funding Agencies: Poverty Labs count on a diverse portfolio of agencies to fund their activities. In my sample, 34% of funding came from multi-lateral institutions (e.g., World Bank and USAID), 40% from philanthropic foundations (e.g., Bill and Melinda Gates Foundation), and the rest from a mix of local sources and country donors. Their role in implementation is to provide the financial and infra-structural resources, but Funding Agencies can also have an impact on the type of projects that will be evaluated. Individuals in these agencies have a similar profile to the policy-makers: elite academic training and vast international experience.

4.2 | Chains of transcriptions: The process of implementing an RCT

The actors in the RCT network are situated in distinct geographical locations, and they influence field experiments with different weights and strategies. Perhaps the most powerful actors in the network are located in academic departments in the US or Europe, where the Research Team and Funding Agencies make decisions regarding the design and costs of RCTs and decide which NGOs to partner with at the local level. The immense infrastructure required to implement RCTs, however, is situated in remote areas of developing countries. Although NGOs, Surveyors, and Policy Beneficiaries have little influence on the design of the experiments, without their active cooperation, the RCT could not be implemented.

Contrary to the transcription of natural objects (Latour, 1999), transcribing the behavior of poor individuals depends greatly on the ability of researchers to govern the web of social relations that exist in the field. To do so, the transcription process starts with an intense communication exchange between the Research and Field Teams to decide on the best experimental design possible to evaluate the NGO's policy. Both teams use available administrative data to estimate the minimum size and design of the experiment that can simultaneously guarantee statistical power (hence securing the robustness of the results), while sounding politically feasible.

After a decision is reached about the experimental design at the Poverty Lab level, another round of analysis and negotiation happen between the Field Team, high-level officials, and staff from NGOs. This is when the preferred experimental design by the researchers gets a reality check from policy managers, in what can be a very contentious negotiation process. In my fieldwork in Brazil, for example, one teachers' union threatened to strike against what they considered an overly ambitious RCT size. In response, the Research Team substantially reduced the experimental design to secure teachers' buy-in and the continuity of the experiment. Similar to this case, it is common for the RCT strategy to be simplified based on constraints of funding, data, ethical and political concerns, infrastructure, and operations. For a development RCT to happen, policy and academic sides have to compromise and agree on a final experiment design.

With an experiment strategy in hand, the Field Team is then responsible for hiring a survey firm that will apply the questionnaires and do the data-entry of the surveys, as well as training the street-level bureaucracies implementing the policies to adapt their practices to fit the experiment design. At this stage, again, local bureaucracies often react against attempts to substantially change their daily practices in the name of the experiment, prompting the Field Team to continuously remind them about the importance of securing the quality of the RCT and the surveillance mechanisms they will use to ensure that the NGO staff does so. The training period is thus key to avoid that "NGOs boycott the RCT" and that they "appreciate the value of our experiment," as pointed out by two Field Team members in Peru.

Following this training period, an initial questionnaire is implemented with all the beneficiaries from control and treatment groups before the new policies being tested are put into practice. It is during the implementation of this baseline survey that the journey of the RCT data from the field to publications starts. A first, non-trivial, step is finding the individuals that will be interviewed—the ones from the treatment group are relatively easier to find because surveyors can count on the NGOs' administrative apparatus to schedule the interviews. Interviewing individuals in the control group, however, is harder because these individuals will not benefit from the policy, they commonly live in places without formal addresses, and they have to voluntarily agree to serve as research subjects (Rayzberg, 2019a).

After individuals from control and treatment groups are found, the actual paper survey can then be implemented. Usually, understanding how "The Poor" behaves means asking individuals to wait in line to be interviewed and to answer a 10- to 50-page-long questionnaire. Interviews last from 30 min to 2 hr. They are done in the local language of the village where the experiment will take place, and the interview schedule consists mostly of behavioral questions, which are frequently hypothetical questions, such as the examples below:

Determining financial literacy	*Determining aversion to risk*
Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, would you be able to buy more than, exactly the same as, or less than today with the money in this account?	*Suppose you had to choose between the following two options:* *A. You receive 30 dollars with certainty.* *B. A coin toss: if the outcome is heads, you receive 100 dollars; if it is tails, you receive nothing. Do you understand the two options? [Check for understanding]. Which would you prefer?*
a) More than today;	*a) 30 dollars with certainty;*
b) Exactly the same as today;	*b) 100/0 dollars coin toss;*
c) Less than today.	*c) Don't know.*

Through answers to one or two questions like the examples above, researchers will then create an index measuring, for example, "Financial Literacy Levels" of respondents, as well as "Aversion to Risk" indexes. Besides hypothetical behavioral questions, questionnaires have socioeconomic and policy implementation questions to guarantee that the control and treatment group divisions were respected.

After the baseline survey is finished, the development intervention begins. The time distance between this first survey and the final survey, when the results of the policies will be assessed, varies, ranging from 3 months to up to 2 years. During this period, NGO staff are expected to respect the division between control and treatment groups, and the fieldwork team is constantly monitoring if they do so. The forms of surveillance of the quality of the experiment are multiple and key to the success of the RCTs: they range from surprise visits to different project sites and monitoring of administrative data of the NGO, to meetings with beneficiaries and policy officials, and a great deal of face-to-face conversations to remind street-level staff of the importance of respecting the design of RCTs.

Finally, after the policy being tested ends, a second survey is administered with the same questionnaire, and the data-entry of all the paper surveys begins. This data is intensely "cleaned" by the Field Team, then sent to universities where the analysis and eventual publications will be carried out. At the end of this data-cleaning process, *randomistas* receive a database file containing the behavior of "The Poor." Even at this point, when the data has already been completely decontextualized and is ready to be analyzed, there will still be many communications between *randomistas* and the Field Team to explain non-intuitive results and signs of control group contamination before researchers agree on a final analysis. Even if individuals in the Field Team are the ones with the lower credentials in the academic side of the RCT network, they have a lot of discretionary power to explain what happened in the field for the authors of the final papers.

Hopefully at this point it is clear that the process of implementing development RCTs is long and demands coordination efforts between actors based in different continents who speak different languages and have different cultural practices, in addition to demanding a great deal of financial and material resources. In each of the steps that I described, adaptations and pragmatic decisions to deal with shortcomings are made. Moreover, considering the long time that RCTs take from their initial formulation until the data from the final survey is collected, the experiments have to deal with a lot of discontinuity and turnover from NGOs' staff, survey firms, and even in the Field Team, which contributes to making the chains of transcriptions from the field to universities even more complex and attempts to control this process even more ardent. Similar to what Rosengarten and Savransky (2019) describe about medical RCTs, these adaptations render evidence produced by development RCTs incredibly situated and dependent on the specific relational dynamics that characterize their implementation.

4.3 | Controversies: Controlling the control group and randomizing social benefits

For all the apparatus described above to function, the biggest challenges for both researchers and implementing partners are randomization and the quality of "no-treatment" control groups. It is only by solving these two issues that *randomistas* guarantee the scientific status of their work and are thus able to portray an image of their field experiments as free from politics or bureaucratic interference—that is, as different from the image of development work as inherently corrupt and inefficient, as suggested by some prominent economists (Easterly, 2007). Yet, considering that they are working with development projects, there is nothing trivial about solving these challenges.

It is no surprise that the creation of a randomly selected "no-treatment" control group with foreign aid beneficiaries is politically and ethically controversial (de Souza Leao & Eyal, 2019). Whenever the development intervention involves the distribution of resources and services, assigning individuals to control groups inevitably raises strong resistance from the development community. In fact, randomized assignment can easily be considered illegal (Glennerster, 2015), unless the development intervention can be framed as a non-entitlement, or if it can be framed as merely a "nudge" meant to overcome behavioral obstacles rather than a form of assistance (Berndt, 2015). How can bureaucrats, NGO staff, or politicians justify giving financial resources and social services to some people but not to others who need just as much? For this reason, *randomistas* have to show that their random assignment was not affected by political or bureaucratic considerations, and thus have to deal with the resulting constraints associated with providing benefits solely for one part of the target population (and withholding benefits from others equally in need) in the name of the scientificity of their experiments.

Moreover, in cases when randomization does happen, welfare regulations and political pressures create the risk of "substitution bias"—when individuals in the control group have good substitutes for the tested policy—which could contaminate the quality of experimental and control groups, and hence undermine the trust in the RCT findings (de Souza Leao & Eyal, 2019; Heckman, Hohmann, Smith, & Khoo, 2000). This means that the scientific success of the RCT network described above is contingent on *randomistas*' ability to show that participants in the experiment did not benefit from other social policies that could interfere with the results—something that would be impossible in richer countries, but that is feasible in remote areas of developing countries.

The tasks of randomizing social benefits and controlling the control group are made both more manageable and more challenging if we consider the normative and regulatory environment that characterizes the current international development field (Watkins, Swidler, & Hannan, 2012). First, in response to the widespread perception that foreign aid was ineffective and corrupt, the 2000s have been characterized as a period of traditional "aid fatigue," and by the entry of a new set of actors, mainly private foundations and NGOs, to the development field (Easterly, 2007; Krause, 2014). As mentioned above, the fact that researchers are now working with NGOs, rather than local governments as was done before, means that they can minimize the political and ethical problems posed by randomization (de Souza Leao & Eyal, 2019).

Second, the privatization of foreign aid has also altered the *type* of aid that is disbursed and under what conditions, and has created new accountability struggles for these new actors (Krause, 2014). Considering that aid typically flows from foreign donors to global NGOs, who are the ones responsible for selecting local partners who will then implement projects in small villages where "development" is supposed to happen, many authors have pointed to the principal–agent problem that characterizes the current aid chain (Swidler & Watkins, 2009). On the one hand, this long funding chain, combined with the fact that NGOs operate in unfamiliar cultural and political terrains marked by "the loss of hope in development" (Krause, 2014, p. 42), results in an aid environment characterized by strong attempts at control by donors and by a focus on measuring aid effectiveness. On the other hand, there is great suspicion and criticism of whom the new private donors and *randomistas* are accountable to, turning the rhetoric around experimenting on the poor of the Global South into a target of criticism:

> *Donors increasingly want to see more impact for their money… Some go so far as to insist that development interventions should be subjected to the same kind of randomised control trials used in medicine,*

> *with "treatment" groups assessed against control groups... But truly random sampling with blinded subjects is almost impossible in human communities without creating scenarios so abstract as to tell us little about the real world. And trials are expensive to carry out, and fraught with ethical challenge... People of the south deserve better. (Op-ed signed by 15 leading economists in* The Guardian[1]*)*

In sum, *randomistas* deal with two types of controversies to successfully portray development RCTs as scientific and free from politics or other ideological struggles that characterize the foreign aid field. On the academic side, researchers have to convincingly show that assignment to treatment and control groups was random, without political interference, and that throughout the experimental period the control group was not contaminated. However, the same characteristics that make development RCTs scientifically rigorous are the ones that make them so politically controversial. This is because, on the policy side, *randomistas* have to convince other actors of the importance of randomizing the distribution of social benefits and controlling the control group, while dealing with the fear of corruption of development projects and with criticisms of the advisability of foreign aid. The latter issue is particularly salient in debates about post-colonialism and the role that foreign funders have in developing countries (Li, 2007; Tilley, 2011).

5 | BOUNDARY WORK BETWEEN TESTING AND DOING DEVELOPMENT

To deal with these controversies, *randomistas* build on the strategic ambiguity of what is on trial to distance themselves from the politics of foreign aid, while being close enough to the field to dictate what works in international development and for other development actors to find partnering with Poverty Labs attractive. Hence, while development RCTs are conceptualized as a means of finding the causal impact of a policy, they are simultaneously touted as serving to test economic theories, making it purposively unclear what is being tested—is it a development intervention, an economic theory, or both? The excerpt below is typical of the strategic ambiguity adopted by *randomistas*:

> Can a RCT tell us not just whether an intervention worked, but also how and why? *When designed and implemented correctly, RCTs can not only tell us whether an intervention was effective, but also answer a number of other policy-relevant questions... However, as with any single study, a RCT is just one piece in a larger puzzle. By combining the results of one or more RCTs with economic theory, descriptive evidence, and local knowledge, we can gain a richer understanding of an intervention's impact. (JPAL, 2017)*

Furthermore, *randomistas* avoid the political debate about development and foreign aid by testing development economic theories that are quite simple, such as in the RCT example introduced earlier: "If you give deworming pills to school children, they get sick less frequently, and go to school more often" (see also Abdelghafour, 2017). Even when *randomistas* limit themselves to these small, short-term questions, they deem each trial as answering one part of broader questions, which can guarantee their contribution to development economic theory. In this way, *randomistas* connect themselves back to the discipline of economics, not to program implementation or the more complicated politics of foreign aid (de Souza Leao & Eyal, 2019).

To put it more generally, *randomistas* resolved the political problem posed by randomization with foreign aid beneficiaries by maintaining a productive ambiguity so that RCTs mean different things to different constituencies at different times. Not only do field experiments equivocate between exercises in theory-building and solutions to policy problems, but as Rayzberg (2019a) shows, *randomistas* employ multiple devices to frame the RCT as an ambiguous object, partly a development intervention and partly a test, in order to overcome political and practical obstacles to randomization. If the RCT is an intervention, local inhabitants and NGO staff would be keen to participate, but would not consider it legitimate to arbitrarily deny the benefits to some participants in the control group.

If the RCT is testing an economic theory, assignment to the control group would be legitimate, but people would be much less eager to participate. The various different designs of development RCTs, Rayzberg (2019a) suggests, can be understood as framings meant to contain and manage this problem, to entangle intervention and test together so as to be able to recruit participants, but then also to disentangle them, so randomization is possible.

As a result, it is common for development RCTs to have a design in which the control group receives some kind of treatment—either a reduced version or the status quo. The lesser treatment is a boundary object. The researchers consider it the null condition of the theory-building test; the participants consider it an intervention. This is also why experimental phase-in design is so popular.[2] Phase-in is a temporal framing device (Rayzberg, 2019a). During the first-time frame, the control group receives no treatment; but during the second time frame it receives the same treatment as the experimental group. So intervention and testing are disentangled in the present, but entangled in the future. Phase-in makes the perceived unfairness of being in a control group much more manageable, since participants are promised to receive social benefits in the near future.

5.1 | When the boundary work fails

The mechanisms illustrated by Rayzberg (2019a) are important for understanding the boundary work involved in making development RCTs successful. Similar to what Gieryn (1983) explained about the continuous need to demarcate boundaries between science and varieties of non-sciences in order to establish scientific authority, *randomistas* construct boundaries between their economic research trials and simple program evaluation in order to secure their space in the two worlds. This boundary work, however, is ongoing and historically changing, and many times *randomistas* are not able to maintain the productive ambiguity between research and development intervention. My analysis found two main reasons why the boundary work typically fails.

The first reason why *randomistas* are unable to maintain the boundary between research and intervention is that policy beneficiaries sometimes discover that they are in control groups and revolt against the experiment. By adopting the phase-in design mentioned above, *randomistas* are usually able to bypass this problem by promising that individuals will benefit from the development project in the future, but it is common for participants to confront researchers about their status as members of the control group, even if they are promised future gains. As I often heard in fieldwork: “Being in the control group rarely goes unnoticed.” For example, in my observations of microcredit RCTs in Peru, individuals in the control group complained multiple times to the NGO that surveyors asked questions about their financial behavior in order to limit their chances of receiving credit in the future. While surveyors were employed by the Poverty Lab, control group individuals conflated their work with the work of the microcredit NGO.

In cases similar to this one, participants blurred the boundaries between research and intervention by holding *randomistas* responsible for both their own and the NGO’s actions, while the Field Team continuously insisted on their distinctiveness. In this case, the Field Team demarcated the boundaries between their role as researchers and the role of the NGO. They were asking questions about the financial behavior of families to learn how “The Poor” made financial decisions, not to limit or facilitate access to microcredit. Yet, while the distinction between research and intervention was important for the Field Team, for “research participants this difference is often arbitrary, irrelevant, or entirely meaningless”, as pointed out by Rayzberg (2019a, p. 389).

Second, and relatedly, there are cases in which high-level policy officials push back against the interests of *randomistas*. This opposition may be related to the perception that Poverty Labs’ research “draws its motivation from academic concerns that overlap imperfectly with the issues that matter to development practitioners” (Ravallion, 2009, p. 48); or to the specific politics of the programs that *randomistas* choose to evaluate. When this happens, *randomistas* are rarely able to maintain the productive ambiguity between research and intervention and face backlash regarding the political goals of their tests and how connected they are to broader ideological and political agendas happening in the foreign aid field. In other words, their field becomes politicized or contaminated, diminishing their ability to claim its scientific status. A recent battle between UK-based Action Aid and *randomistas* can serve to illustrate this type of resistance.

In this RCT, three US-based economists partnered with Liberia's federal government to test the effectiveness of a public-private partnership to introduce what they referred to as "American-style charter schools or the UK's academies to Liberia's underperforming education system" (Romero, Sandefur, & Sandhotz, 2017). In order to implement the RCT, however, *randomistas* insisted that the Partnership Schools for Liberia (PSL) program be first tested as a pilot in order for experimental evidence to be generated. As a reaction, Action Aid released a Request for Proposal for qualitative researchers to closely observe the PSL pilot (Edwards, 2017). Action Aid saw many problems with this RCT: (a) it involved an unrealistic financial investment in schools that would not be reproducible at a larger scale; (b) it aimed to weaken public education in favor of *specific* private partners; and (c) it did not answer the right questions:

> *[We have] concerns about whether it is feasible for the RCT, however rigorous, to isolate the added value that arises from the involvement of private providers. How will the government of Liberia or anyone else learn anything, other than the already evident truth that if you spend more money per student and have smaller class sizes you will get better results? (Archer, 2016)*

While defending the pilot, *randomistas* got entangled in the politics of post-war Liberia, choosing to publicly defend Liberia's president and minister of education for their commitment to piloting the new educational policy. Similar to the Field Team strategy presented above, *randomistas* insisted on the strong boundaries and differences between their research and the politics of the PSL:

> *We also agree with Action Aid and Education International on more substantive matters… The key difference is that we do not read these points as criticisms of our study. Rather, many of the points below are concerns over the program itself, many of which we share… We would reframe these concerns as hypotheses that we have explicitly designed the randomized evaluation to test. (Romero et al., 2017)*

Yet, even with their attempt to separate their experiment from the politics of the pilot, the public backlash has been huge, and researchers are now accused of having a hidden agenda of privatization of educational systems around Africa. Indeed, Education International—a partner organization of Action Aid—included the PSL pilot in its global campaign to "oppose privatisation and commercialisation of public education" (EI, 2017), making the controversy visible to an even wider audience. As aptly put by a World Bank official[3]: "The new Liberia school experiment is destined to be a Rorschach test of which side of the private education debate you sit on." Whether this was the intention of researchers or not, their scientific neutrality has been put into question by parts of the development community.

By characterizing their critics as doing "advocacy not research," however, the three authors did not see their test as politicized, claiming instead that they were collaborating for an "ongoing debate that [was] increasingly disciplined by facts" (Romero et al., 2017). Put differently, when the boundary work failed, *randomistas* quickly reinforced the boundaries between research and intervention, and put on their "scientist hats," dismissing development actors as driven by "conviction" or, as Esther Duflo, has claimed multiple times, as being driven by the "three I's—Ideology, Ignorance, and Inertia" (Duflo, 2017, p. 9).

6 | CONCLUSION

In this paper, I have argued that the key element for understanding the current success of development RCTs is unveiling how *randomistas* both draw and blur the boundaries between development practice and economic research in order to implement field experiments. To do so, I presented the massive organizational effort and the diverse set of actors involved in the making of RCTs and the continuous boundary work that *randomistas* do to differentiate the purposes and merits of testing development projects from doing them, as a way to bypass the problems presented by adopting the experimental method with foreign aid beneficiaries in poor countries. In conclusion, I address two implications of this research for the study of testing and tests.

First, continuing a tradition in science and technology studies, my research draws attention to the ongoing negotiations between the desire to control and the need for cooperation that characterize the implementation of field experiments (Henke, 2000; Kohler & Vetter, 2016). To this end, the case of development RCTs demonstrates that successful field experimentation depends on the researcher's ability to productively manage the ambiguity of what is being tested. Yet, contrary to many field sciences, in which scientific authority results from an intense familiarity with certain places (Gieryn, 2006), in development RCTs, managing this ambiguity means that *randomistas* distance themselves from the broader history, politics, and particularities of their object of study, purposively dissociating developing countries from the broader dynamics of the foreign aid industry. As a consequence, development RCTs differ greatly from the "heroic interventions of modernity, based on the narrative of progress" that often characterized foreign aid, focusing instead on diffuse, "evidence-based melioristic interventions" that are portrayed as behavioral in nature (Berndt, 2015; Rottenburg et al., 2015, p. 10).

Second, my findings connect to debates about the performativity of tests (MacKenzie et al., 2007). One of the arguments that I put forward in this paper is that the credibility of RCTs to the scientific community is contingent on the belief that field experiments were not influenced by material, ideological, or political interests. Yet, similar to other types of tests, my research shows that the attempts to control developing countries make development RCTs extremely performative, since their success depends on development projects adapting their operations to fit experimental procedures. Indeed, as the dissemination of RCTs takes place, research and policy designs increasingly start to look like textbook RCTs. The UK government, for instance, recently published a policy paper—"Test, Learn, Adapt: Developing Public Policy with Randomised Controlled Trials" (Haynes, Service, Goldacre, & Torgerson, 2012), suggesting that all government policies be designed with an RCT evaluation in mind.

At a time when testing and global policy standards are diffusing with higher speed than before, my findings are also relevant to assess the potentially exclusionary nature of top-down control over acceptable forms of evidence. As the case of development RCTs rests clear, the imposition of evidence hierarchies in international development and its consequences for the distribution of global aid are based on the premise that development projects can and should be modified for testing purposes. As I have demonstrated, however, transforming developing countries into appropriate "fields" for experimentation prompts resistance, misunderstandings, and uneven negotiations that often exclude the voice and interests of local beneficiary populations. Moving the conversation forward, more qualitative research on the views of these actors is needed to envision the possibilities of creating alternative models of evaluation, as well as more democratic methods of transferring resources to countries from the Global South.

ACKNOWLEDGEMENTS

The author wants to thank Diego Povoas and XingJian Li for their work in this project, as well as Gil Eyal, Margarita Rayzberg, and participants of the Put to the Test workshop at Warwick University for their helpful comments and suggestions.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

NOTES

[1] "Buzzwords and Tortuous Impact Studies Won't Fix a Broken Aid System," *The Guardian online*, July 16, 2018. Available at https://www.theguardian.com/global-development/2018/jul/16/buzzwords-crazes-broken-aid-system-poverty [Accessed on November 2, 2018].

[2] In my sample, in 35% of studies the control group received either the status quo (i.e., the development project) or a slightly reduced version of it. In 40% of studies, researchers adopted the phase-in design, in which the control group receives the development policy in a later period. In the remaining 25% of studies, the control group did not receive any type of treatment.

[3] Matt Collin in Twitter, September 8, 2017. Available at https://twitter.com/aidthoughts/status/906181438390902784 [Accessed on November 2, 2018].

REFERENCES

Abdelghafour, N. (2017). Randomized controlled experiments to end poverty. *Anthropologie & Développement*, *46–47*, 235–262.

Archer, D. (2016). *The challenges of education reform and privatization in Liberia*. Retrieved from http://www.actionaid.org/2016/12/challenges-education-reform-and-privatisation-liberia.

Banerjee, A., & Duflo, E. (2011). *Poor economics: A radical rethinking of the way to fight global poverty*. New York: Public Affairs.

Berndt, C. (2015). Behavioral economics, experimentalism and the marketization of development. *Economy and Society*, *44*(4), 567–591.

Carpenter, D. (2010). *Reputation and power—Organizational image and pharmaceutical regulation at the FDA*. Princeton, NJ: Princeton University Press.

De Souza Leão, L., & Eyal, G. (2019). The rise of Randomized Controlled Trials (RCTs) in international development in historical perspective. *Theory and Society*, *48*(3), 383–418.

Deaton, A. (2010). Instruments, randomization, and learning about development. *Journal of Economic Literature*, *48*(2), 424–455.

Deaton, A., & Cartwright, N. (2016). Understanding and misunderstanding randomized controlled trials. *NBER Working Paper Series*, No 22595.

Duflo, E. (2003). Poor, but Rational? *MIT Working Paper*, *747*.

Duflo, E. (2010). Social experiments to fight poverty. *TED Conference*, Retrieved from https://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty.

Duflo, E. (2017). The economist as plumber. *American Economics Review: Papers & Proceedings*, *107*(5), 1–26.

Duflo, E., Glennerster, R., & Kremer, M. (2007). Using randomization in development economics research: A toolkit. CEPR Discussion Papers, No 6059.

Easterly, W. (2007). *The white's men burden: Why the west's efforts to aid the rest have done so much ill and so little good*. New York: Penguin USA.

Edwards, S. (2017). *Funding struggle: Can Liberia's controversial privately run schools pilot continue?* Retrieved from https://www.devex.com/news/funding-struggle-can-liberia-s-controversial-privately-run-schools-pilot-continue-91061.

EI. (2017). *"Development cooperation" and "unite for quality education campaign"*. Retrieved from https://ei-ie.org/en/detail_page/4641/development-cooperation.

Escobar, A. (2012). *Encountering development: The making and unmaking of the third world*. Princeton, NJ: Princeton University Press.

Eyal, G. (2013). For a sociology of expertise: The social origins of the autism epidemic. *American Journal of Sociology*, *118*(4), 863–907.

Ferguson, J. (1990). *The anti-politics machine: Development, depoliticization, and bureaucratic power in Lesotho*. Cambridge: Cambridge University Press.

Gieryn, T. (1983). Boundary-work and the demarcation of science from non-science: Strains and interests in professional ideologies of scientists. *American Sociological Review*, *48*(6), 781–795.

Gieryn, T. (2006). City as truth-spot: Laboratories and field-sites in urban studies. *Social Studies of Science*, *36*(1), 5–38.

Glennerster, R. (2015). *So you want to do an RCT with a government: Things you should know*. Retrieved from http://runningres.com/blog/2015/12/9/so-you-want-to-do-an-rct-with-a-government-things-you-should-know.

Guala, F. (2007). How to do things with experimental economics. In D. MacKenzie, F. Muniesa, & L. Siu (Eds.), *Do economists make markets?* Princeton, NJ: Princeton University Press.

Guggenheim, M. (2012). Laboratizing and de-laboratizing the world: Changing sociological concepts for places of knowledge production. *History of the Human Sciences*, *25*(1), 99–118.

Harrison, G. (2011). Randomization and its discontents. *Journal of African Economies*, *20*(4), 626–652.

Haynes, L., Service, O., Goldacre, B, & Torgerson, D. (2012). *Test, learn, adapt: Developing public policy with randomised controlled trials*. Retrieved from https://www.gov.uk/government/publications/test-learn-adapt-developing-public-policy-with-randomised-controlled-trials.

Heckman, J., Hohmann, N., Smith, J., & Khoo, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *Quarterly Journal of Economics*, *115*(2), 651–694.

Henke, C. (2000). Making a place for science: The field trial. *Social Studies of Science*, *30*(4), 483–511.

JPAL. (2017). *Jameel poverty action lab website*. Retrieved from www.poverty-action.org.

Kabeer, N. (2019). Randomized control trials and qualitative evaluations of a multifaceted programme for women in extreme poverty: Empirical findings and methodological reflections. *Journal of Human Development and Capabilities*, *20*(2), 197–217.

Kohler, R., & Vetter, J. (2016). The field. In B. Lightman (Ed.), *A companion to the history of science*. Chichester: John Wiley & Sons.

Krause, M. (2014). *The good project: Humanitarian relief NGOs and the fragmentation of reason*. Chicago, IL: Chicago University Press.

Latour, B. (1999). *Pandora's hope: Essays on the reality of scientific studies*. Cambridge, MA: Harvard University Press.

Li, T. (2007). *The will to improve: Governmentality, development, and the practice of politics*. Durham, NC: Duke University Press.

MacKenzie, D., Muniesa, F., & Siu, L. (2007). *Do economists make markets? On the performativity of economics*. Princeton, NJ: Princeton University Press.

Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. *Econometrica*, *72*(1), 159–217.

Ogden, T. (2016). *Experimental conversations: Perspectives on randomized trials in development economics*. Cambridge, MA: MIT Press.

Petryna, A. (2009). *When experiments travel: Clinical trials and the global search for human subjects*. Princeton, NJ: Princeton University Press.

Ravallion, M. (2009). Evaluation in the practice of development. *The World Bank Research Observer*, *24*(1), 29–53.

Rayzberg, M. (2019a). Fairness in the field: The ethics of resource allocation in randomized controlled field experiments. *Science, Technology and Human Values*, *44*(3), 371–398.

Rayzberg, M. (2019b). *Controlling the field: Experimental social science and politics of evidence in international development* (PhD dissertation). Northwestern University, Sociology Department, Chicago, IL.

Riecken, H., & Boruch, R. (1975). *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press.

Rogers-Dillon, R. (2004). *The welfare experiments: Politics and policy evaluation*. Stanford, CA: Stanford University Press.

Romero, M., Sandefur, J., & Sandholtz, W. (2017). *Will an RCT change anyone's mind? Should it?* Retrieved from https://www.cgdev.org/blog/will-rct-change-anyones-mind-should-it.

Rosengarten, M., & Savransky, M. (2019). A careful biomedicine? Generalization and abstraction in RCTs. *Critical Public Health*, *29*(2), 181–191.

Rottenburg, R. (2009). Social and public experiments and new figurations of science and politics in postcolonial Africa. *Postcolonial Studies*, *12*(4), 423–440.

Rottenburg, R., Merry, S., Park, S., & Mugler, J. (Eds). (2015). *The world of indicators: The making of governmental knowledge through quantification*. Cambridge: Cambridge University Press.

Swidler, A., & Watkins, S. (2009). "Teach a man to fish": The sustainability doctrine and its social consequences. *World Development*, *37*(7), 1182–1196.

Teele, D. (2014). *Field experiments and their critics: Essays on the uses and abuses of experimentation in the social sciences*. New Haven, CT: Yale University Press.

Tilley, H. (2011). *Africa as a living laboratory: Empire, development, and the problem of scientific knowledge, 1870–1950*. Chicago, IL: Chicago University Press.

Watkins, S., Swidler, A., & Hannan, T. (2012). Outsourcing social transformation: Development NGOs as organizations. *Annual Review of Sociology*, *38*, 285–315.

How to cite this article: de Souza Leão L. What's on trial? The making of field experiments in international development. *Br J Sociol*. 2020;71:444–459. https://doi.org/10.1111/1468-4446.12723

APPENDIX

Analytical Sample

TABLE A1 Descriptive statistics

Total (*n* = 63)	
Countries	23 different countries
Average size of RCT	8,660 individuals
Median size of RCT	2,156 individuals
Average duration	12 months
RCT topic: Finance	38%
RCT topic: Education	24%
RCT topic: Health	22%

Received: 12 March 2019 | Revised: 27 January 2020 | Accepted: 23 March 2020

DOI: 10.1111/1468-4446.12758

SPECIAL ISSUE

WILEY

What the pregnancy test is testing

Joan H. Robinson

Department of Interdisciplinary Arts and Sciences, The City College, CUNY, New York, NY, USA

Correspondence
Joan H. Robinson, JD, PhD, The City College of New York, 25 Broadway,
7th Floor, New York, NY, 10004, USA.
Email: jrobinson1@ccny.cuny.edu

Funding information
Directorate for Social, Behavioral and Economic Sciences, Grant/Award Number: GG009686

Abstract

Is the test result positive or negative? Tests that occur in labs and doctors' offices pose specific questions to try to obtain specific information. But what happens in the social world when these tests never see the inside of a lab or doctor's office, and instead they are used in a house, in a Walmart bathroom, or in a dormitory bathroom stall? Putting the diagnosis aside, what does the presence of these tests do to social life? This paper examines one such test, the home pregnancy test, and specifically, its use in contemporary intimate life of people who do not want to be pregnant. Pregnancy tests test for pregnancy. But what else is the pregnancy test putting to the test? To investigate this, I spent 8 years studying American pregnancy tests using a qualitative mixed methods approach. This paper draws on some of my research materials, specifically, 85 life history interviews. Each participant was asked to recall, in full, all of their experiences with home pregnancy tests throughout their lives, resulting in well over 300 narratives of home pregnancy test usage which I qualitatively analyzed. I find that more than just a test for a pregnancy, the use of the home pregnancy test is a test of roles, relationships, and responsibilities in social life. These findings suggest implications for social life as more biomedical tests move out of the purview of the medical establishment.

KEYWORDS
gender, pregnancy, reproduction, test, women

1 | INTRODUCTION

"I touched my stomach and said, 'Test' and they understood." Paulette was abroad in Poland. It was a family trip with her elderly grandmother to a homeland Nana had not seen since she fled the Nazis. Paulette described her role in her big, hectic family trip, and the multiple ironies of discovering a possible pregnancy in, of all places, Poland, were not lost on Paulette. They had tried to exterminate her family and lineage, but had narrowly failed, and here she was possibly defying them once again. But front and center in her mind were her intimate relationships.

She considered her new boyfriend. "Oh, my God, what if I'm pregnant? We've only been together a short time, like, I don't know what I would do." How would the outcomes of the test affect this new relationship?

> *I had told him before because he was the one who wanted to not use condoms and I was a little apprehensive about it. And I had told him, like, honestly, I don't think I would have an abortion if I got pregnant. Like, I am totally 100 percent in favor of abortion. . . but I'm in an age where I really want children and, like, I just don't think I could do it.*

Testing herself in the context of her family's trip brought to the fore her relationships with her mother and her mother's mother, Nana:

> *I'm with my entire family—like, Nana, my mom, my two uncles . . . we drove from Warsaw to this small town where my grandmother was born in Poland and then my family erupted into a huge fight that lasted, like, three hours. . . Of course, I'm trying to find a pharmacy on the way and come up with some sort of excuse why I need to go into a pharmacy. I wouldn't tell them. Like, I just don't have that kind of relationship with my mom [to tell her] that I was getting a pregnancy test. She doesn't know I have a boyfriend, because it's new. And it was, like, a really emotional experience going to where Nana was born.*

After many failed attempts, Paulette eventually slipped away and was able, through the common human gestures of pregnancy like putting her hand on her abdomen, to purchase a pregnancy test. She soon learned that she was not pregnant. But Paulette's pregnancy test did not enter a barren social landscape. Who is privy to a woman's[1] sexual and reproductive behavior and outcomes, and how does she manage those intimate boundaries? As this paper will show, pregnancy tests do not just test for pregnancy. Pregnancy tests in the hands of women test a variety of other social roles, relationships, and intimacies, which show, for instance, the stigmas about women's sexuality and the pressures of contraceptive responsibility that women like Paulette face and tackle regularly (see, e.g., Riessman, 2000).[2]

Forty years ago, when an American woman wanted to know if she was pregnant, she could make an appointment with a medical professional who could conduct a pregnancy test and deliver the results (Robinson, 2016). Propelled by the American medical establishment's control, surveillance, and neglect of women's health, the women's health movement of the 1970s sought to put women's health "into their own hands," and the home pregnancy test went into the hands of American women in this social and regulatory landscape (Robinson, 2016).

Twentieth-century feminists argued that the Pill and other forms of contraception would liberate women from fear of endless pregnancies and huge families, and even, in some imaginations, women could have sexual equality to men. Women using various forms of contraception today can indeed take steps to control their fertility more than ever in human history. Nevertheless, short of permanent sterilization as a means of contraception (see, e.g., Denbow, 2015), this study finds that sexually active heterosexual women still worry about and plan for possible unwanted pregnancies far more than their male counterparts. The physical and emotional toll of pregnancy risk is still almost entirely borne by women.

Since the 1970s, home pregnancy tests have become part of our American cultural and consumer landscape (Leavitt, 2006; Robinson, 2016). While other over-the-counter biological tests are just beginning to enter the marketplace, pregnancy tests can now be found at nearly every American corner drug store and gas station, online, and in bulk, and are usable by potentially half of the population multiple times throughout their lives. Their boxes promise in bright pinks and purples, “Be the first to know!” and “Over 99% accurate!” Women from their teenage years through menopause who have had contact with sperm can go to the drug store and take a pregnancy test instead of going to the doctor to find out whether they are pregnant.

Nevertheless, as a widespread and potentially bellwether technology, the social landscape of home pregnancy tests can inform the changing landscape of medical care. Today's patients check their health data on their Apple Watch, take their temperatures with an app in their smart phones, mail their own medical tests to a lab, and search WebMD with their symptoms, whereas previously they had to have an appointment with a doctor. These changes have happened due to the changing nature of the medical economy and increasing access by lay people to the information and tools formerly only available to or through doctors. The pregnancy test is likely at the forefront of a massive amount of care work being transitioned from hospital to home, including very skilled medical work. This shift will affect all of us, and it demands more study by sociologists as it happens (Oudshoorn, 2011). This paper examines a subgroup of pregnancy test users, those who do not want to be pregnant, and examines how they test just one area of their lives, their intimate lives.

2 | THEORY

Social scientists and social theorists who study reproduction have shown that pregnancy is unique among human conditions in a multitude of social and biological ways (see, e.g., Almeling, 2015; Balasubramanian, 2018; de Beauvoir, [1949] 2011; Casper, 1998; Duden, 1999; Fessler, 2006; Franklin & Ragone, 1997; Gálvez, 2011; Gutiérrez, 2008; Lappé, 2016; Lopez, 2008; Luker, 1995; Mamo, 2007; Markens, 2007; Morris, 2013; Morris & Robinson, 2017; Murphy, 2013; Rapp, 2000; Roberts, 1998; Rock-Singer, 2018; Rothman, 1985; Stevens, 2015; Waggoner, 2015). Prior research on the home pregnancy test has been primarily historical and cultural and has begun to examine contemporary social dimensions (Leavitt, 2006; Layne, 2009; Layne, 2003; Childerhose & MacDonald, 2013; Robinson, 2016). Scholars agree that women are compelled “to know” whether they are pregnant as early as possible for a variety of social reasons, but this early knowledge is acquired during a period in which miscarriage is most likely (Leavitt, 2006; Layne, 2009; Layne, 2003). Prior sociological studies of home pregnancy tests, similarly with historical and cultural focus, have shown the home pregnancy test’s history is deeply intertwined with the women’s health movement, regulatory law, and gender and age inequalities (Robinson, 2016). This study of the home pregnancy test contributes to this growing literature in the sociology of reproduction, a sub-field which has studied everything from pre-pregnancy care to international gestational surrogacy in India, but has paid little attention to pregnancy tests.

But my analysis of the home pregnancy test also makes an important contribution to the sociology of testing, in particular to the growing area of study of biological tests that are conducted at home. Social science has shown that diagnosis is a social creation and is never value-neutral, no matter who is delivering or receiving it (Jutel, 2009, 2011, 2015), and the same is true for diagnostic tests that are conducted at home. In what follows, the home pregnancy test is theoretically positioned in relation to two such tests that have been previously studied: telemedicine tests and DNA tests (Oudshoorn, 2008, 2011; Nelson, 2008, 2016, 2018). This prior research shows that biological tests conducted at home enter a complex landscape of social relations.

Research on home testing interacts with a broader literature in feminist science and technology studies that holds that human bodies are thoroughly intertwined with a material world—what Donna Haraway famously labeled “cyborg” (Haraway, 1990; see also Mol, 2002; Suchman, 2007). This science and technology studies literature argues that knowledge is situated, gendered, and political (see, e.g., Haraway, 1988; Oudshoorn, 1990; Wajcman, 1991). Similarly, the present study follows a material object, the home pregnancy test, that interacts with and brings outside of the body that which is thought to be inside a person’s body. The information and knowledge contained in the pregnancy test, as with other biological tests, is gathered, hidden, and revealed through complex formulas of non-human objects, biology, and social roles, relationships, and responsibilities. Regardless of the result, which is often monumental in itself, the test illuminates a variety of other tests in social life. Telemedicine and genealogy DNA tests entail similar complex testing formulas that the study of the home pregnancy test brings to the surface.

2.1 | Telemedicine

Telemedicine, the practice of monitoring, diagnosing, and treating patients at a temporal and spatial distance from a clinician, has major implications for healthcare systems globally (Oudshoorn, 2008, 2011). As communication technologies are now widespread in wealthy parts of the world, and diagnostic technologies can be mobile and easily used, telemedicine is a means to potentially dramatically cut healthcare costs and makes some form of medical care possible when in-person visits are not possible. Today, carework falls in large part to women, people of color, and lower income people, and these shifts in the healthcare system could change the distribution of carework in ways that make more work for those groups. Ruth Schwartz Cowan's extensive study of the boom of home appliances, *More Work for Mother* (1983), found that appliances ironically made women's lives harder. The growth of telemedicine could produce similar results. In an examination of electrocardiograms (ECGs) within the context of telemedicine, Nelly Oudshoorn (2011) showed how ECGs used at home similarly changed the distribution of work, responsibility, and surveillance of patients, involving new actors such as specially trained nurses to read the results.

Like home pregnancy tests, the administration of telemedicine tests occurs remotely from a medical professional and may involve family member participation.[3] In my study, most home pregnancy tests were used with no medical advice whatsoever. Moreover, rather than involving family members, women who do not want to be pregnant often do extensive work to hide their use of a pregnancy test from everyone, including partners, roommates, and relatives.

Finally, while the results of telemedicine tests can implicate the family members in their capacity of carers of the primary patient, telemedicine tests do not typically implicate family members to the extent that a pregnancy test does, particularly in the case of paternity. A pregnancy test almost always implicates another biological and/or social parent.[4] The gravity of the test results for others may be great and can put the woman who is taking the test at risk of harm, both social and physical.[5]

2.2 | Genealogy DNA tests

Telemedicine tests, like pregnancy tests, raise issues of remote medical care work and sometimes participation of others—they are social tests of present social circumstances. Home pregnancy tests also raise issues of biology and genealogy, and like DNA genealogy tests, home pregnancy tests are also social tests of biological family ties.

The use of commercial DNA tests like 23andMe and AncestryDNA has become common in the US as a way to deliver risk profiles for disease as well as purportedly "racial" makeup (Nelson, 2008, 2016, 2018). A new facet of these tests that has become apparent in the last few years is their ability to reveal genealogical relationships to other living people, what I will call for the purposes of this illustration "genealogy DNA tests." These genealogy DNA tests have been used to find criminals, sperm donors, and long-lost relatives, and the tests are worthy of serious study in their own regard.

Genealogy DNA tests have more in common with pregnancy tests than meets the eye. They promise to reveal critical information about the self and show biological and genetic connection to others—they reveal kinship. Testing for kinship changes the way people understand and relate to their families, housemates, and distant relations. Focusing on the social and behavioral aspect of the test rather than the "biological," the differences between DNA genealogy tests and pregnancy tests can be primarily organized in two ways: temporally and gendered.

Temporally, pregnancy tests test the recent social past (i.e., the sexual intercourse) and a possible and imminent social future (i.e., pregnancy and its sequelae), whereas DNA genealogy tests primarily test the past. DNA tests may reveal sexual decisions of our parents or more distant relatives, but pregnancy tests reveal sexual decisions that were made very recently. With pregnancy tests, sexual behaviors and actions led to test results—no sex, usually no pregnancy. Another temporal dimension is that pregnancies can be terminated and the history can be largely erased, whereas the relations in DNA cannot be severed. In the current state of commercial genetic testing, such as 23andMe and AncestryDNA, a tester is implicating not only known family members but distant relations and generations to come, and they cannot ever undo it (Nelson & Robinson, 2014). DNA is immutable through time and generations, whereas pregnancy is mutable and has an implicit timeline. Pregnancy tests are tests of recent social relations, contemporary intimate life, and social future, and usually implicate people we know. DNA genealogy tests are tests of social history that may have an impact on contemporary and future social life, of people the tester may never even know.

Critically, pregnancy affects some members of the population and not others—it is not universal. It becomes most salient to the person who bears the most consequences, the woman.[6] There is an unavoidable timeline, and decisions need to be made—potentially very stigmatizing decisions.

The temporalities, mutability, and all of the associated stigmas make the individual responsibility of testing greater for women doing home pregnancy tests today, a time in which pregnancy tests are able to detect pregnancy within the legal abortion timeline, which was not always the case. The tests also render apparent the physical differences between otherwise egalitarian different sex couples.

Like Oudshoorn and Nelson, this paper examines social life through the lens of a pedestrian but critical object in many people's lives, the home pregnancy test. Rather than assume the pregnancy test is following its moniker, by only testing for pregnancy, this study instead takes a standpoint from science and technology studies that things do not always do what they are supposed to do (Akrich, 1992; Cowan, 1987; Mol, 2002). When social scientists follow those technical objects around with a careful ethnographic approach, they discover expected and unexpected roles and uses in social life (Akrich, 1992; Nelson, 2008, 2016, 2018). Sometimes, social scientists discover entirely different uses and even different users whose lives might not be obvious to social scientists (Star, 1991; Wyatt, 2003). By taking this approach, social scientists can also reveal the "ideological mechanisms that have made so many of those lives invisible" (DeVault, 1999, p. 30).

3 | METHODS AND DATA

This paper is drawn from a larger 8-year study of pregnancy testing that used mixed methods including archival research, historical interviews, ethnography, legal research, semi-structured interviews with users, partners, bystanders, and doctors. The data from this paper are drawn from this larger study. This paper examines the social relations of home pregnancy testing among different sex couples who are at risk of unwanted pregnancies, and as such, the paper excludes all same-sex couples and different-sex couples who are trying to conceive. Those couples do not have accidental pregnancies. In different sex relationships, it is women who are at greatest risk of unwanted pregnancy, and as such, different sex couples are a good case to examine because they bring to light issues of stigma around women's sexuality and the pressures of responsibility. Different-sex and same-sex couples who are never at risk of unwanted pregnancy, and who may be *trying to conceive*, bring forth different issues, and those I examine elsewhere. Differently put, this paper examines contemporary intimate life through the ethnographic lens of the pregnancy test. Aside from the immediate result, what is the pregnancy test testing?[7]

First, I recruited home pregnancy test users and partners through listserves, online posts, and paper fliers. Many women[1] who volunteered had used home pregnancy tests starting in their teenage years through adult years and included a variety of different hopes, expectations, and outcomes within one woman's life. Male partners were, understandably, less likely to volunteer for a pregnancy test study. When they did volunteer, it was typically to discuss their recent tests trying to conceive with their adult wife. I then interviewed all of these men about their whole life story, including pasts with unwanted pregnancy scares. To recruit more male partners who were at risk of unwanted pregnancies, particularly younger men, I tried many times to recruit college-aged men through the same pathways of fliers and online recruitment that I had used to recruit women, all to no avail. To recruit younger men, I resorted to giving a talk in a packed, beer-smelling college fraternity house to discuss (and educate on) “pregnancy scares." The session was lively, informative, and as one man said, “sad.” The thirst for the conversation was palpable. After the talk, I distributed my recruitment fliers and recruited some of the fraternity members to participate in life-history interviews.

Each of the 85 interviewed participants—women and men—was asked to recall, in full, all of their experiences with home pregnancy tests throughout their lives, totalling well over 300 narratives of home pregnancy test usage. All of the interviews included handwritten notes and ethnographic memoing. The interviews were nearly all taped, though two participants opted out of taping due to sensitive information that they revealed in their interviews. These refusals can be seen as both a methodological challenge, informing us about the design of studies of sexuality like the one here, as well as providing additional data about some of the topics in question, in particular, risks of stigma and privacy.

The 83 taped interviews were transcribed by a professional transcription service. I qualitatively analyzed all of the transcripts and my own interview notes using traditional sociological methods of open coding and patterning (Charmaz, 2006). I began the analysis using NVIVO software but ultimately re-coded all of the interviews on paper copies of interview transcripts.

My participants resided in the United States and ranged from their late teens to sixties, from high school educations to PhDs, and were about 30% people of color. I assigned pseudonyms to protect participants’ privacy. Qualitative work in the sociology of reproduction typically skews white and middle class, and while this study was the same, its participants were more diverse racially, geographically, and religiously than the leading studies in this subfield. The sample for this paper, with its focus on the possibility of unwanted pregnancy, included some interviews in full and some partial interviews. For instance, many couples in same-sex relationships were primarily interviewed about trying to conceive, and that data was excluded from the sample for this paper. However, many of those same individuals also reported experiences with the risk of unwanted pregnancies, from past relationships with different sex partners, friends and family who were users, and extramarital affairs with different sex partners. These latter cases were included in the sample for this paper.

4 | FINDINGS

This study finds that for women who do not want to be pregnant, home pregnancy tests have become socially and morally mandatory. In my data, the cultural script produces an assumption, impetus, and moral obligation that women must test. Additionally, women want to know if there is something inside of them that has the potential to change every aspect of their lives, including their health, their relationships, their living situation, their work, and the lives of their descendants. Women may feel pressure to home test for pregnancy more than other health conditions because of the time constraints associated with abortion and the unavoidable progression timeline of pregnancy.[8] Women also feel more pressure to test because of the availability of the over-the-counter home pregnancy test. In addition to testing for the condition of pregnancy, the pregnancy test tests a variety of areas of their social life including their roles, their relationships, and their responsibilities.

4.1 | Testing roles

Of all age groups in this study, the possibility of an unwanted pregnancy is the most unsettling to teenagers who do not want to conceive. In the case of teenagers, home pregnancy tests are testing whether the birth control methods failed and/or was used properly. Additionally, from a sociology of testing perspective, the pregnancy test tests the performance of adulthood. Social adulthood, in this context, imagines able-bodied and able-minded adults making independent and private decisions about their sexual and medical bodies, including sexual choices and consumer choices. I found that teenagers taking home pregnancy tests were testing out this role. When young women, who live with their parents or guardians, go to a store to buy a test, they test their performance as an adult. They test their composure. They try to take the test in secret, and thus test their privacy. More broadly, they are testing management of their own reproduction. The performances and trials are experiments with their reproductive bodies and their social roles. These social tests are tests that can be failed.

Though all of the interviewees were 18 or older and living apart from their parents at the time of the interview, many of them first encountered a home pregnancy test when they were a younger teenager living in their childhood home. Most were in intimate relationships of short or uncertain duration—a high school boyfriend, "a casual thing," or "not serious." Several women described being sexually active with someone who was "not a good guy" (or worse), and thus someone who was not a reliable partner with whom to share information, decisions, or support, let alone a child. Most of the narratives were marked by shame about being a sexually active teenager.

The purchase of the test tested the geographical boundaries of their privacy. Several women described living in small towns or close-knit communities and the risks of buying a test there. Their social subterfuge included traveling to different stores or even towns to purchase the tests: "I went to like, a town over to buy the test," or "a drug store a couple of towns over," or they were "driving to another town."

Nearly all of the women who purchased a test in their teen years recalled the test of interaction at the store's check-out. The women skirted the truth and even lied to prevent an embarrassing breach of status. Helene recalled a "pregnancy scare" when she was sixteen: "And I—I remember saying to the person at check-out—'Oh, these are for biology tests that I'm doing for class'—you know, like, lying about what they were for, and I'm sure he knew exactly what they were for." Bonnie described how she and her best friend, both age 16, would buy them for each other to avoid the shame and fear of purchasing their own tests, "[I]t just felt like this isn't for me, so it's just—I can buy this." Another woman explained, "At the time, you know, when you're sixteen, there's no mystery about what you hope the outcome of that pregnancy test will be. So I was just so, so embarrassed. I think I probably just was really like, also bought a magazine and some tights to take attention away from the pregnancy test so it wasn't the only thing glaring on the cash register." Yet another recalled, "And I just was so ridiculously embarrassed, I hid that sucker under a pile of crap I didn't need in my handcart. And I considered stealing it, because that would have been less shameful to me, you know?"

Many women reported they felt "scared," including in their own home. Women tried to go about their ordinary business and pretend nothing was amiss, thereby testing their composure to avoid more embarrassment. They also went to great lengths to "hide the evidence," knowing what they were doing would be revealed as a transgression of their role as an asexual child. Women "threw them away in the corner trash can," "stuffed it at the bottom of our garbage outside," wrapped the test in toilet paper and "stuff[ed] it down in the kitchen garbage pail," tried to "be cool about it," and "felt very stealth about it." Olga wrapped the test in toilet paper and took it out "to the big garbage that would sit in the driveway. . . just so nobody would—you know, see what I had been up to." Women hid the evidence because even the possibility of pregnancy threatened to reveal their nascent sexuality to adults who (in an American context) generally tried to prevent or ignore it.

4.2 | Testing relationships: Sexuality and privacy

When women had moved out of their childhood home, not only were they testing their independence, but also their new social relationships. Can my information be handled in a respectful way? Can my romantic relationship be one that lasts? Is this a person with whom I want to conceive and possibly raise a child? For many people in my study, the answer to one or more of these questions was "no."

Interviewees who lived on college campuses were often afraid about the stigma and shame from their college community related to their sexual behavior and potential to be perceived as promiscuous. Though the reasons were somewhat different, their strategies were similar to younger women. Florence described going to the pharmacy near her college to buy a test: "[it] was actually the most awkward and scariest thing. . . someone's gonna figure it out, like see me buy it and start whispering around campus. And so I just made sure I bought a whole bunch of junk food and everything just to hide it underneath it all." Shary described the lengths to which she went to hide buying the test in college:

> *So I assumed that they had them on campus but then I was like, "Someone's going to see me." So I was—I have to go off campus to find it. I had to take the bus so I had to look at a bus map and sort of figure out the nearest, closest drug store—cause I didn't want it to be too close. I didn't want to run into anybody. It was like this big secret—like—a spy event I had to do. Like I had to—I put on a hoodie—and I went to the—across the river. It's across the river I didn't think I was going to see anybody. Across the river at that time was a pretty rough neighborhood. I went to the drug store and bought it.*

Women linked taking the test to a test of their promiscuity and possible social exclusion from their college communities. Rose took one in the Walmart bathroom, the store at which she had purchased the test. Telling me filled her with even more shame, acknowledging the conflict of what is perceived as cheap and dirty (a public Walmart restroom), and what is perceived as value-laden and pure (pregnancy). Expressing how she felt to even tell me the story, Rose exclaimed, "So embarrassing!"

Wendy recalled how she and her casual boyfriend Kyle were talking about becoming exclusive, so she used the next month as a window of opportunity to "hook up" with different men every weekend, thinking it might be the last chance in her life. Thinking back on taking the pregnancy test, Wendy recalled, "My rationale, so I went a little crazy the last month of being single and I just remember sitting in my room, like, this is karma. Why is this karma for me being slutty this last month?" She volunteered that she would not know who the father was if the test came back positive, and that she was embarrassed about that. Following what Wendy called "the pregnancy scare," she immediately became exclusive with Kyle and has remained with him since. The pregnancy test, which had tested for pregnancy with other men, put her relationship with Kyle to the test.

Though most of the narratives were filled with shame, when taking a pregnancy test was reflected on as the beginning of a positive relationship, the test-taking was viewed more positively. Gert described her home pregnancy testing in college, closely linking the memories to the beginning of her relationship with her current husband, "I was a little embarrassed, I guess, you know, sort of similar to buying condoms and was a little worried. But you know, it was a little bit exciting, too, because it was almost like it proves you are a sexual being in the world, that you're doing—something to get you pregnant." Gert viewed her sexual activity as an accomplishment and a coming-of-age that made her proud, and the test was a symbol of that.

Friendships were also tested, with mixed results. While she was in college, Whitney took a pregnancy test in a house that she shared with friends. After she took the test, she recalled, "I thought I threw it away, but I guess one of my friends got a hold of it and put it on top of the Wall of Shame that night." In their house, the women had a "wall of shame" where they made fun of each other and themselves. By placing the pregnancy test on the wall of shame, her roommates reminded her that her "hook up" had been with the wrong type of partner, and that they were entitled to surveil, judge, and shame her. Whitney's home pregnancy test tested her friendships, her privacy, her promiscuity, and her choice of men.

Mindy, who later became a doctor, faced a test of her own composure during someone else's health crisis. Mindy recalled her college roommate taking a home pregnancy test while she was present and connected it to her medical role now: "We were in the dorm room, she went to the bathroom and did the whole pee on a stick thing, and I think she waited with it, and came out and told me it was negative. I was there for support, the whole freak out session, 'I can't have this baby, I have to finish school,' having her process that stuff." Mindy was one of many women who saw themselves as safety nets for friends who needed support. Grace remembered:

> *And she took the first one. Positive. Took the second one. Positive. Took the third one. Positive. Took the fourth one. Positive. And we were—well, she peed on them all at the same time. And then they—you know,*

> *as they were coming, you know, across and she's like, bing, ding—And we're standing there in the kitchen like us three and as they like, you know, show their sign, all of our eyes are getting bigger and we're like, Uh. And she starts to cry, the girl who took them. And we were like, Uhh, oh my gosh. And so then it was just like this somber, almost like someone had told us that someone we knew died.*

She proceeded to accompany her friend to multiple doctors to ultimately get mifepristone (oral medication) to terminate the pregnancy.

Home pregnancy tests also tested women's relationships with their sexual partner. Whether the partner was asked to be involved with the test, or simply told the result after, it was used as a test of not just their relationship, but of his maturity and responsibility and whether he can handle major life moments. As Victor, 21, explained, involvement in the test can be a measure of the relationship. He described it from what he imagined to be the woman's perspective:

> *But I think that would differ a lot for who the person was you were taking it for. If it's someone you're dating, I think that's good [to be involved in pregnancy testing]. If it's someone you kind of had a one-night stand with, or someone you just kind of like had a fling with, it might make more sense to take it yourself and then discuss it with them.*

For some male partners, the home pregnancy tests felt like a test of their future. Teddy explained how he wished his serious college girlfriend, now his wife, had not even told him about her regular pregnancy testing: “By involving with me was like somebody having, you know, kind of a gun to my head and constantly pulling the trigger but there's no bullet in it.” To Teddy it felt like his life was on the line in a game of Russian roulette. He feared the social and economic repercussions of a pregnancy and his emotional response to an abortion, and by repeatedly testing, he felt she was testing his patience with repeated threats to his future. He wished she had the fortitude to keep it to herself. In retrospect, he explained, his youth and lack of knowledge about women's biology made him act badly, and he now knows that she was testing his ability to be a supportive partner.

Like Teddy, many men and boys in this study lacked basic biological information about their sex lives and had poor information about sex and reproduction. This finding is similar to that of other social scientists (Almeling, 2018). According to both men and women in this study, the possibility of pregnancy was not on boys' and men's minds two weeks after they had sex. In this way, the unsettling nature of the pregnancy test is different for the girl or woman who must physically bear the consequences. While men's sense of pregnancy risk could be characterized as unaware or indifferent, in some instances the women's sense of pregnancy risk may have been exaggerated due to the stigma that they would personally suffer. In this way, although men came from a place of lack of knowledge, the men's sense of pregnancy risk might in some cases be characterized as more realistic than their woman partner's, which could also be from the women's inadequate sex education in addition to the stigma and other consequences.

4.3 | Testing relationships: Assessing support and inclusion

In my interviews with women, women recounted countless stories of taking pregnancy tests without their male sexual partner's knowledge, which explains why men who I interviewed knew of far fewer tests. Though it takes two to tango, the interviews revealed rather starkly that women are the gatekeepers of their own health, well-being, and bodies, and men are often kept in the dark about possible pregnancies.

Because the locus of the pregnancy is in women's bodies, women make decisions to include or exclude partners (and others) by sorting people into different categories based on their expected levels of emotional support for the women's condition and the decisions she (or they) might face. When a friend, family member, or sexual partner was expected to require more emotional work than the woman would receive back in the form of support, that individual would be kept in the dark, and a woman would take care of herself as best she knew how and often alone. Fiona explained: "I just didn't feel like he was my support person." When others were needed to provide care and support, women typically relied on existing support networks.

Whether and how men are included in women's pregnancy testing is determined largely by the nature of the relationship between the two people. The most blatant distinction is whether the relationship is for sexual purposes only (what participants called "a hook-up" or "an encounter"), in which case he is most easily ruled out of the picture, or whether the relationship includes an interpersonal or emotional component.

Women interviewees recalled more hook-ups in their teens and twenties, so keeping pregnancy testing from one's sexual partner occurs more among younger women generally. Similarly, women who do not want to be pregnant, and might terminate the pregnancy if there was one, were also less likely to tell the male partner. Hook-ups aside, women in committed, long-term, monogamous relationships often made the decision to exclude or delay the involvement of their sexual and romantic partner from pregnancy testing. The reasoning behind that decision-making process is not immediately apparent and often kept to oneself.

The frequency of women taking pregnancy tests alone—in particular, excluding their partner—is staggering, and it can teach us something about women's reproductive station today. In early pregnancy, women are individual and autonomous, and they grant permission to others to be included. Women have the right and the responsibility to be alone, and they often are. By staying alone, women do not invite the myriad concerns of others, whatever those might be, and they maintain more control than they might otherwise. The division of reproductive labor, which already includes for women the embodied gestational labor, also includes the emotional labor of caring for themselves and anyone they tell about potential pregnancies and realized pregnancies. Women frequently choose to leave men in the dark because, even if they are cared for, they feel that it is in their best interest to be with others or alone. Tony explained his relationship with an ex-girlfriend:

> *We both knew that we were not right for each other. We both knew that we were not ready to have a child. And so she had an abortion. I do know it was home pregnancy test, but only because the phone call that had come from her was actually from her roommate. We'd broken up probably three weeks prior and hadn't talked since.*
>
> *And I think she was just feeling so overwhelmed and—and didn't want to talk to me unless she had to. So her roommate had called me and clearly the roommate was the one who had been with her when she'd taken the home pregnancy test.*

Women in this study viewed partner inclusion in pregnancy testing as an equalizing-horizontal relationship. They were sharing the reproductive consequences of their shared sexual decisions.

For men, in some cases the inclusion was welcome, for instance because they understood that she still bore the greater responsibility and decisional power, and in some cases it was unwelcome because they wished to remain aloof of the reproductive reality. Inclusion can also be a loss of status. Men also described women's roles as central and their own as peripheral, which resonates with prior research (see Almeling & Waggoner, 2013). She gets to decide when to take the test, and she has more knowledge about her reproductive cycles. Especially in the case where they don't want to conceive but will keep the child, a pregnancy can alter both of their futures.

In her relationship with others, the pregnancy test tests whether: (1) the partner is a one-night stand or hook-up or whether there is an emotional component that the women want to/ are able to share; (2) they are very serious and decide to have a child; and (3) he is a "good guy" and someone she wants through life's tests.

4.4 | Testing responsibility

In the case of teenagers living in their childhood homes, the women were primarily testing their performance of adulthood. In the case of women who had recently moved out of their childhood home, they were testing both their independence as well as their new relationships with friends and sexual partners. But what about the situation where none of the aforementioned are in question? In the case of women whose adulthood and relationships are stable, one test is of her capacity to take responsibility.

Mindy, a white, married mother of two and a doctor in a prestigious hospital emergency room in Manhattan, New York City, explained her decision to not tell her husband about recent pregnancy testing:

> *[S]o ever since I had my second one, we kind of have decided that we're done—and so, I'm on the Pill. But every once in awhile, I'll take it late or I'll forget to take it and I'll have to double up. And then, I always worry that I'll get pregnant again. . . first of all, it's totally irresponsible of me to miss a pill—but I mean, I work overnights and some nights I'll work until, like, 11:00 at night and then I get home at 1:00 in the morning and I just forget—which is not a great excuse, because I have to be a responsible adult, but it doesn't happen that often that I miss a pill. But anyway, I don't want to tell him, because he'll be like, "Oh, my gosh, how could you be, like, a grown woman and a professional and, like, forget to take a pill?" So I actually don't want to tell him because I think he'll be mad that I forgot.*

The reason for Mindy's decision not to tell her husband about the pregnancy test was to avoid being judged by him and having a fight about it. After the test came back negative, she walked out of her Manhattan apartment and threw the test in the street garbage can so it would not be seen by her husband in their home garbage. The municipal public street garbage assisted Mindy in keeping her secret, as it probably has many others for many other reasons.

Mindy was supposed to have it all in her control—her marriage, her two children, her prestigious career, and her regular management of contraception. Her inability to be a "responsible adult" made her feel badly, despite her hectic schedule and working hours. She used the pregnancy test to hide the fact that she missed the Pill because her husband might be mad and she would be embarrassed and even belittled. Interestingly, even Mindy, the woman who appears to have it all (including, of course, domestic workers who make her and her husband's careers and child-raising possible (see, e.g., Hondagneu-Sotelo, 2001; Rosenbaum, 2017)), does significant amounts of work to maintain the appearance of complete control and responsibility, even to her husband.

5 | CONCLUSION

Unless a couple is trying to conceive, the "two week wait"—the period between ovulation and knowledge about a pregnancy—is still almost exclusively women's business. The immediate circumstances of the pregnancy test and the gravity of the result make it seem like the only thing being tested is whether a woman is pregnant, but it is not.

The pregnancy test not only tests for pregnancy, it also tests a wide range of other social roles, relationships, and responsibilities. Even just the acquisition, taking, and disposal of the test is a social test, regardless of the result. If she does not want to be pregnant, the outcomes of being pregnant, or even the possible pregnancy, can be sources of shame. At these moments, it can be a test of her responsibilities—to practice safe sex, be chaste, be a good daughter or student, or just stay on the "right path." The potentially pregnant woman bears responsibility for (and the rights to) finding the relevant information, making decisions about who to include and exclude, and think about all possible outcomes. Her partner and others may or may not be expected to be part of it, and the decision to include or exclude them is up to her. The pregnancy test tests accountability (particularly of fatherhood and parenthood), partnership and support, women's role as mother, worker, and partner, and asks in which order does she organize these values.

When I asked women whether they'd take a home pregnancy test or go to a doctor, nearly every woman I have spoken with about this topic formally and informally over the last 8 years expressed preference for a home pregnancy test.

In the course of years of research and casual conversations about home pregnancy tests, only one woman (so far) has told me that she did not take a home pregnancy test when she was trying to conceive. She opted to be a single parent by choice and had undergone intrauterine insemination (IUI) with a doctor—it "did not occur to [her]" to have a moment alone with a home pregnancy test by herself, nor did she have an opportunity to have a home pregnancy test moment with a partner. This single anecdotal case reinforces the finding that the home pregnancy test is used for a variety of social reasons. In what she referred to as certain "medicalized" circumstances in which a woman is choosing to become a single parent, a myriad of social reasons to use a home pregnancy test are non-existent.

This study of the home pregnancy test suggests that tests for biomedical information are not just that. They bear witness to boundaries in social life, trust and care, friendships and partnerships, longevity and dissolvability of relationships, shame and sexual stigma, performances of adult status, and responsibility. They are tests of past and future actions and tests of past and future relations. In short, they are tests of social life—positive, negative, and all the complex areas in between.

DATA AVAILABILITY STATEMENT

The data are only available to study investigators approved by the Institutional Review Board in compliance with American laws governing human subjects research as well as HIPAA.

NOTES

[1] Recruitment for this study was aimed at pregnancy test users and included outreach to LGBT and specifically trans parenting groups, as of course some transmen can become pregnant intentionally or unintentionally. Each of the potentially pregnant participant volunteers for this study self-identified as a "woman," so that term is used in this paper. This paper does not make claims about the experience of transmen or all people with uteruses.

[2] In this paper I adopt the understanding of women's sexual and reproductive stigma developed by Riessman (2000). This understanding of stigma posits that information that can discredit a person is deeply contextual and is not one-sided. This understanding of stigma does not assume respect of individual rights of privacy, these stigmas may be understood differently by the individual herself, and these stigmas vary by race, class, gender, culture, and over the reproductive life course. Finally, this understanding of stigma allows more room for agency, correction, and resistance.

[3] Telemedicine tests like ECGs are more medically institutionalized, whereas over-the-counter tests may not involve a clinician at all. An interpretation and diagnosis, if needed, is organized by a medical professional, who is available via communication devices, and the associated costs may be borne by insurance or a government healthcare system. Moreover, the medical provider still has a significant role in determining the proper course of care and treatment.

[4] The exception is the case of single mothers by choice who use an anonymous sperm donor, discussed below in the Conclusion.

[5] Additionally, unlike telemedicine which can be distributed through the life course and is more common among those considered "ill," home pregnancy tests are used during mid-life when women are considered most "healthy" and are able to be most "independent" in their care, both aspects contributing to their ability to conduct them alone.

[6] This study focuses on self-identified ciswomen who had sex with self-identified cismen and are therefore at risk of unwanted pregnancy. People who engage in same-sex intercourse, cismen, and transwomen, are not at risk of unwanted pregnancy, and were not included in this study. Like ciswomen, some transmen who have sex with cismen or transwomen may also be at risk of unwanted pregnancy, and recruitment included LGBT and trans-parenting groups, but none of the recruited participants identified as such.

[7] Science and technology studies show us that things do things that they aren't intended to do, thus revealing unexpected facets of social life (see, e.g., Akrich, 1992).

[8] In the US, it is much easier to get an abortion during the first trimester (13 weeks), which in practice is 11 weeks from a missed period. Medication abortions (mifepristone and misoprostol) are typically available during the first 10 weeks, which in practice is 8 weeks from a missed period. Women know that an abortion is easier legally, logistically, biologically, and to some people ethically, the fewer weeks the pregnancy has progressed. Women and their support systems sometimes refer to the fetus as "cells," differentiating it from a human (Rapp, 2000).

REFERENCES

Akrich, M. (1992). The de-scription of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society* (pp. 205–224). Cambridge, MA: MIT Press.

Almeling, R. (2015). Reproduction. *Annual Review of Sociology*, *41*, 423–442.

Almeling, R. (2018). Presentation, “Whither GUYnecology? The missing science of men’s health and how it matters for reproduction”. *ASA Regular Sessions: Considering Men and Partners*, August 12.

Almeling, R., & Waggoner, M. R. (2013). More and less than equal: How men factor in the reproductive equation. *Gender and Society*, *27*(6), 821–842.

Balasubramanian, S. (2018). Motivating men: Social science and the regulation of men’s reproduction in postwar India. *Gender and Society*, *32*(1), 34–58.

Casper, M. (1998). *The making of the unborn patient: A social anatomy of fetal surgery*. New Brunswick, NJ: Rutgers University Press.

Charmaz, K. (2006). *Constructing grounded theory: A practical guide through qualitative analysis*. Thousand Oaks, CA: Sage Publications.

Childerhose, J. E., & MacDonald, M. E. (2013). Health consumption as work: The home pregnancy test as a domesticated health tool. *Social Science and Medicine*, *86*, 1–8.

Cowan, R. S. (1983). *More work for mother: The ironies of household technology from the Open Hearth to the Microwave*. New York, NY: Basic Books.

Cowan, R. S. (1987). The consumption junction: A proposal for research strategies in the sociology of technology. In W. E. Bijker, T. P. Hughes, & T. Pinch (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 261–280). Cambridge, MA: MIT Press.

de Beauvoir, S. ([1949] 2011). *The second sex*. New York, NY: Vintage.

Denbow, J. (2015). *Governed through choice: Autonomy, technology, and the politics of reproduction*. New York, NY: NYU Press.

DeVault, M. L. (1999). *Liberating method: Feminism and social research*. Philadelphia, PA: Temple University Press.

Duden, B. (1999). The fetus on the “farther shore”: Toward a history of the unborn. In L. M. Morgan & M. W. Michaels (Eds.), *Fetal subjects, feminist positions* (pp. 13–25). Philadelphia, PA: University of Pennsylvania Press.

Fessler, A. (2006). *The girls who went away: The hidden story of women who surrendered children for adoption in the decades before Roe v. Wade*. New York, NY: Penguin Press.

Franklin, S., & H. Ragone (Eds.) (1997). *Reproducing reproduction: Kinship, power, and technological innovation*. Philadelphia, PA: University of Pennsylvania Press.

Gálvez, A. (2011). *Patient citizens, immigrant mothers: Mexican women, public prenatal care, and the birth-weight paradox*. New Brunswick, NJ: Rutgers University Press.

Gutiérrez, E. R. (2008). *Fertile matters: The politics of Mexican-origin women’s reproduction*. Austin, TX: University of Texas Press.

Haraway, D. J. (1988). Situated knowledges: The science question in feminism as a site of discourse on the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Haraway, D. J. (1990). *Simians, cyborgs, and women: The reinvention of nature*. New York, NY: Routledge.

Hondagneu-Sotelo, P. (2001). *Domestica: Immigrant workers cleaning and caring in the shadows of affluence*. Oakland, CA: University of California Press.

Jutel, A. (2009). Sociology of diagnosis: A preliminary review. *Sociology of Health and Illness*, *31*(2), 278–299.

Jutel, A. (2011). *Putting a name to it: Diagnosis in contemporary society*. Baltimore, MD: Johns Hopkins University Press.

Jutel, A. (2015). Beyond the sociology of diagnosis. *Sociology Compass*, *9*(9), 841–852.

Lappé, M. (2016). The maternal body as environment in autism science. *Social Studies of Science*, *46*(5), 675–700.

Layne, L. (2003). *Motherhood lost: A feminist account of pregnancy loss in America*. New York, NY: Routledge.

Layne, L. L. (2009). The home pregnancy test: A feminist technology? *Women’s Studies Quarterly*, *37*(1 & 2), 61–79.

Leavitt, S. A. (2006). “A private little revolution”: The home pregnancy test in American culture. *Bulletin of the History of Medicine*, *80*(2), 317–345.

Lopez, I. (2008). *Matters of choice: Puerto Rican women’s struggle for reproductive freedom*. New Brunswick, NJ: Rutgers University Press.

Luker, K. (1995). *Dubious conceptions: The politics of the teenage pregnancy crisis*. Cambridge, MA: Harvard University Press.

Mamo, L. (2007). *Queering reproduction: Achieving pregnancy in an age of technoscience*. Durham, NC: Duke University Press.

Markens, S. (2007). *Surrogate motherhood and the politics of reproduction*. Berkeley, CA: University of California Press.

Mol, A. (2002). *The Body Multiple: Ontology in Medical Practice*. Durham, NC: Duke University Press.

Morris, T. (2013). *Cut it out: The C-section epidemic in America*. New York, NY: New York University Press.

Morris, T., & Robinson, J. H. (2017). Forced and coerced C-sections in the United States. *Contexts*, *16*(2), 24–29.

Murphy, M. (2013). *Seizing the means of reproduction: Entanglements of feminism, health, and technoscience*. Durham, NC: Duke University Press.

Nelson, A. (2008). Bio science: Genetic genealogy testing and the pursuit of African ancestry. *Social Studies of Science*, *38*(5), 759–783.

Nelson, A. (2016). *The social life of DNA: Race, reparations, and reconciliation after the genome*. Boston, MA: Beacon Press.

Nelson, A. (2018). The social life of DNA: Racial reconciliation and institutional morality after the genome. *British Journal of Sociology*, *69*(3), 522–537.

Nelson, A., & Robinson, J. H. (2014). The social life of DTC genetics: The case of 23andMe. In D. L. Kleinman & K. Moore (Eds.), *Routledge handbook of science, technology, and society* (pp. 108–123). New York, NY: Routledge.

Oudshoorn, N. (1990). On the making of sex hormones: Research materials and the production of knowledge. *Social Studies of Science*, *20*(1), 5–33.

Oudshoorn, N. (2008). Diagnosis at a distance: The invisible work of patients and health-care professionals in cardiac telemonitoring technology. *Sociology of Health and Illness*, *30*(2), 272–295.

Oudshoorn, N. (2011). *Telecare technologies and the transformation of healthcare* (1st ed.). London, UK: Palgrave Macmillan.

Rapp, R. (2000). *Testing women, testing the fetus: The social impact of amniocentesis in America* (1st ed.). New York, NY: Routledge.

Riessman, C. K. (2000). Stigma and everyday resistance practices: Childless women in South India. *Gender and Society*, *14*(1), 111–135.

Roberts, D. (1998). *Killing the black body: Race, reproduction, and the meaning of liberty*. New York, NY: Vintage.

Robinson, J. H. (2016). Bringing the pregnancy test home from the hospital. *Social Studies of Science*, *46*(5), 649–674.

Rock-Singer, C. (2018). *Prophetesses of the body: American Jewish women and the politics of embodied knowledge* (Doctoral Dissertation). Columbia University.

Rosenbaum, S. (2017). *Domestic economies: Women, work, and the American dream in Los Angeles*. Durham, NC: Duke University Press.

Rothman, B. K. (1985). The products of conception: The social context of reproductive choices. *Journal of Medical Ethics*, *11*, 188–192.

Star, S. L. (1991). Power, technology and the phenomenology of conventions: On being allergic to onions. In J. Law (Ed.), *A sociology of monsters: Essays on power, technology, and domination* (pp. 26–56). New York, NY: Routledge.

Stevens, L. M. (2015). Planning parenthood: Health care providers' perspectives on pregnancy intention, readiness, and family planning. *Social Science and Medicine*, *139*, 44–52.

Suchman, L. A. (2007). *Human-Machine Reconfigurations: Plans and Situated Actions*. Cambridge, UK: Cambridge University Press.

Waggoner, M. (2015). Cultivating the maternal future: Public health and the prepregnant self. *Signs: Journal of Women in Culture and Society*, *40*(4), 939–962.

Wajcman, J. (1991). *Feminism confronts technology*. Oxford, UK: Polity Press.

Wyatt, S. (2003). Non-users also matter: The construction of users and non-users of the internet. In N. Oudshoorn, & T. Pinch (Eds.), *How users matter: The co-construction of users and technology* (pp. 67–79). Cambridge, MA: MIT Press.

How to cite this article: Robinson J. What the pregnancy test is testing. *Br J Sociol*. 2020;71:460–473. https://doi.org/10.1111/1468-4446.12758

Received: 12 March 2019 | Revised: 12 July 2019 | Accepted: 17 November 2019

DOI: 10.1111/1468-4446.12725

SPECIAL ISSUE

WILEY

Testing planets: Institutions tested in an era of uncertainty

Janet Vertesi

Sociology Department, Princeton University, Princeton, NJ, USA

Correspondence
Janet Vertesi, Sociology Department, Princeton University, 122 Wallace Hall, Princeton, NJ 08544, USA.
Email: jvertesi@princeton.edu

Funding information
This study was supported by the US National Science Foundation Grant #1552469

Abstract

Prior accounts of the experimenter's regress in laboratory testing are set against the background of a relatively stable institutional context. Even if the tools are new or the object of investigation is unknown, participating entities are named, a certain degree of funding is presumed, and an organization exists to conduct the test. In this paper, I argue that this background assumption obscures the importance of institutional and organizational context to the sociology of testing. I analyze ethnographic data gathered among a NASA team whose funding is uncertain, whose mission organization is not yet established, and whose object of investigation is inaccessible. In what I characterize as "ontological flexibility," I reveal how scientists shift their accounts of object agency in response to changes in their institutional environment. As they describe the moon as "uncooperative" or "multiple" while they make appeals to institutions at various stages of support in their exploration projects, this reveals the presence of an "institutional regress": a previously overlooked aspect of the sociology of testing.

KEYWORDS

institutions, testing, uncertainty

1 | INTRODUCTION

Testing has been a topic of analysis in science studies since the early days of the Sociology of Scientific Knowledge programme. Formative accounts examine the required knowledge of the outcome before a test begins (Collins, 1985), the challenges of making "similarity and difference" judgments between experimental results and scientific conclusions (Barnes, 1982; Pinch, 1986, 1993), and the work of aligning the world outside the laboratory with an experimental apparatus and findings within (Latour, 1987). While these concerns address ambiguity in the process of science at the bench, they do not account for testing under the considerable conditions of uncertainty that characterize contemporary techno-scientific work.

Such uncertainties arise on multiple fronts. One is the instability of public institutions that fund scientific work in the wake of the recent global financial crisis. Contemporary scientific institutions contract their spending and compete in a resource-limited environment, while benchmarks for success and expectations for outcomes are high. While the question of scientific patronage is hardly new (Ashworth, 1991; Berman, 2012; Biagioli, 1993, 2007), scientists arguably face renewed anxieties as they aim to satisfy investors, politicians, and publics in a period of austerity.

Uncertainty also reigns in the emerging organizations that support *innovation*—the very word which civic leaders hope will address contemporary economic woes. Innovation not only requires exploring novel domains, but doing so amid emerging configurations of actors and organizational porousness (DiTomaso, 2001; Neff, 2012; Powell, Koput, & Smith-Doerr, 1996; Powell, Packalen, & Whittington, 2012; de Vaan, Vedres, & Stark, 2015; Stark & Vedres, 2009). There is no established group of experts next door with "magic hands" to assist, draw upon, or groom; nor are there weighty regulatory bodies to invoke. In many cases, the organization is so nascent that it exists only as a group of well-intentioned individuals, perhaps supported by a trickle of funding: not a robust constellation of experts with well-defined roles, clear authority relations, and steady paychecks. Such environments—which I will here generalize as beset with "institutional uncertainties"—are not the sorts of places wherein an experimental system just needs a tweak to produce hoped-for results.

So how do actors determine whether or not a test is successful when the object of investigation, the tools to investigate it, and the organization and funding to support the test are all, together, uncertain and unknown? Further, and a key question for the sociology of scientific knowledge: what do actors do to claim that they can satisfy testing requirements and produce successful tests when there is no certain future for their team? This paper examines such problems in the context of planetary scientists' prospective exploration of Jupiter's moon, Europa. At this stage, there are no probes and no built instruments, only enigmatic clues from prior probes and telescopic observations. Support for any mission to Europa must come from NASA, an established scientific institution. But during a critical period of the mission's development, the space agency's budget was repeatedly slashed, US Congress and the presidency changed hands, and there was no agency administrator.[1] The scientists struggled to enter their mission's development phase where they would need an appointed staff, an organizational structure, and a defined path toward launch.

Under these conditions of institutional uncertainty, I observed planetary scientists deploy a form of talk that animates their object of investigation, describing it as "uncooperative." As the agency of their proposed object of investigation shifted from passive to unruly, what I describe as a form of "ontological flexibility," I argue that observing such shifts in scientists' accounts can help the analyst to determine changes in the wider context of the object's investigation, including the central yet previously overlooked role of institutional stability in exiting the tester's regress. I thereby advance a claim for moving beyond instrumentation, similarity and difference, replication and evidential significance in the sociology of testing to include a proposed *institutional regress*.

1.1 | Methods and site

This paper draws upon ethnographic immersion with NASA's mission to Jupiter's moon Europa, originally planned to launch in 2023. A group of planetary scientists has agitated for two decades for funds to build an orbiter to characterize the moon's ice shell and to map it from space. At the same time, an overlapping community of scientists involving astrobiologists has argued for a lander to visit the moon's surface (on astrobiology see Helmreich, 2009; Helmreich, Roosth, & Friedner, 2015). NASA, the United States' space agency, approved a Europa orbiter and initiated a stream of developmental funding in 2014, selecting instruments and establishing a project team for the craft in 2015. In 2016, a team was assembled to conduct early pre-development studies for a lander as well.

One of the original “Medicean Stars” charted by the astronomer Galileo, Europa is a small moon that orbits the planet Jupiter. Prior spacecraft visits showed Europa to be covered in ice and criss-crossed with fissures somewhat like those seen in Antarctic regions on Earth. Scientists hypothesize that a global ocean feeds this ice sheath beneath its surface. Given its proximity to Jupiter and ensuing radiation and tidal heating, Europa hypothetically holds the key ingredients—water, the appropriate chemistry, and energy—to foster the development of life. This makes it a compelling target for investigation, but also produces a situation of ontological uncertainty. After all, experimental outcomes are unknown—and unknowable until a craft is built and sent to investigate. At present, only a series of prior NASA spacecraft have visited Europa, with a European probe planning to conduct observations in the Jupiter system in the 2020s.

I began following NASA’s Europa missions in 2009, including observing the orbiter’s science definition team meetings until 2015, at which point that mission began its development phase in earnest. I attended meetings regularly for two years, both remotely and while immersed at a key site of mission work, the Jet Propulsion Laboratory in Pasadena, California. My fieldwork coincided with the orbiter’s “Phase A” period: when a mission with a launch date is announced, personnel are staffed, mission technologies are articulated, and a viable plan for implementation, costing, scope, and delivery are assembled. With no spacecraft yet built, this phase largely consists of meetings and teleconferenced calls to determine what the spacecraft will look like, what it will do, and what its price tag might be. During this period, the lander mission was considered “pre-Phase A”: funded with a skeleton staff to produce a mission concept only, with no further promise of certain institutional support.

Over 2 years, I regularly attended meetings and events with the newly appointed orbiter team as they began working on the design of their probe and instrumentation, and I attended the intensive workshops associated with the lander’s feasibility study. Because missions of this scale involve individuals and institutions all over the country, I made every effort to achieve a balance between “co-presence” in online meetings, forums, and teleconference calls (Beaulieu, 2010) as well as in-person meetings in key institutions in California, Michigan, Colorado, and Maryland. I also interviewed project personnel, in many cases with my recorder off in order to enable my participants to speak freely. At the end of each day I recorded detailed fieldnotes, later transcribed, entered into, and coded in nVivo12, with codes developing in situ and later expanding as they were rediscovered across the site. I also produced memos on many of these codes to develop the insights I describe here.

2 | AN “UNCOOPERATIVE” MOON

Over more than a decade of attending planetary science conferences, mission operations events, science team meetings, and social gatherings as an ethnographer, I had never heard planets personified. Investigators sometimes anthropomorphized their exploring robots, but planets and moons were *places* (Messeri, 2016): sometimes dangerous or mysterious, but otherwise there for the exploring. Entering a project in its early planning phase, I discovered a different side of planets altogether. Suddenly, I heard scientists say that the success or failure of a future test “all depends on how Europa is behaving, and we can’t control that.” The moon’s properties were likely to “confound” or “confuse” prospective instrumentation. The question was not whether their theories were correct, their tests would work, or their spacecraft would operate as intended: it was whether Europa would “cooperate” with the scientists exploring it.

For instance, during the lander’s science definition team meeting, I observed a conversation between a small group of 18 scientists and their funding representative at NASA’s headquarters. The scientists had various areas of expertise—geochemistry, ice properties, planetary imaging, extremophiles—and many of them had not met each other before. They assembled in a windowless conference room at the NASA Jet Propulsion Laboratory around a heavy wooden table, with one or two of them phoning in over WebEx teleconference lines. Over 2 days of

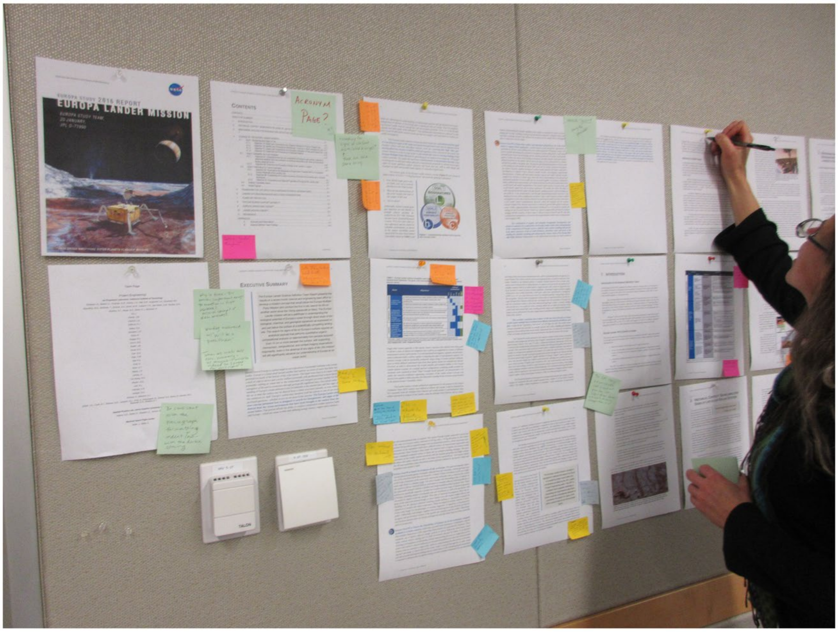


FIGURE 1 The lander team finalizes their report, posted around three walls of a meeting room for ease of editing [Colour figure can be viewed at wileyonlinelibrary.com]

intensive discussion they agreed upon the premise of an extensive report to lay out the lander mission's parameters (Figure 1). The NASA representative opened by commenting that mission success explicitly relied upon "a bug hunt": that is, "something that will find life":

> *What I'm asking you is, are you going to come up with a science concept as a mission where on the twenty-sixth day of this [twenty-five day] mission we can have a press conference at Headquarters and announce to the world that we have found life on Europa based on the measurements we have acquired… You guys will be on the stage. You will be hosting that press conference announcing your results, you or one of your colleagues.*
>
> *(Lander SDT Meeting; August 9, 2016)*

Referencing a few controversial examples which were familiar touchstones to everyone in the room, he followed up: "I don't want to put you guys up on stage and three seconds later the entire science community's throwing tomatoes at you, so it's gotta be a robust finding." At the same time, he admitted that "we don't have any instruments that can detect life; we don't even have consensus [on] what measurements we should take." Then he cautioned that another parameter for mission success was the ability to stay within a narrowly constrained cost cap.

The project—find life, do so robustly in 25 days, and do it with limited resources—proposed nothing less than a Kennedy-esque moonshot, yet without the decade of step-wise work or the expansive budgets of the lunar landing program. In the ensuing discussion, the scientists did not question the feasibility of this charter within constrained costs, the presence of detectable life off-world, or the precariousness of mission support. Instead, they offered the following responses:

Astrobiologist: How much up-front assumption *about the Europa that we will visit* can we place on our measurement? By which I mean, if we say we can definitively detect life *if Europa provides us* the following three things—it's basically a gangbusters biosphere and *it can't help itself* and *it's putting* biosignatures on the surface … then yes, this instrument or these instruments can detect it definitively. Short of that assumption, we can't make that guarantee.

Geologist: I feel like it's difficult to understand if we're in Bon-Jovi-land [a reference to the song, "Living on a prayer"] or actual academic understanding of the answer [of life or no life]. Because it's difficult to imagine a way that you could defend absolutely everything without a lot more [instrumentation] than we're going to be able to carry. Maybe that's the pessimist's view.

Scientist-turned-administrator: Would it be alright if we built a couple of, as we talked about yesterday, of end member scenarios of *the Europa that cooperates with our mission* and said in these cases, in that cooperation as it's going, we would be willing to sign up, if not, *if it's a different Europa*, how can we predict--?

Geologist: Just as we need to be clear about what a positive signature is, *we need to protect against a false negative*. Failing to make a definitive detection with this payload, *we can clearly conclude that we do not have the Europa that we assumed*. (Lander SDT Meeting; August 9, 2016; emphasis mine)

In response to compounded uncertainties yet under the keen awareness that a promise needed to be made, the scientists spoke of Europa as if it were an active, slippery, multiple world. The stakes are high for those in the room, but not just because "extraordinary claims require extraordinary evidence," in Carl Sagan's oft-referenced phrase about detecting extra-terrestrial life. As a senior scientist explained, "[If we] fail to produce something that is very focused that is what our agency and our key supporter in Congress wants, we are going to be looking at another twenty years before we go to Europa[.]" Arguing for relaxed requirements was therefore not a solution: this would not guarantee further opportunities for exploration and a steady stream of funding. The NASA representative agreed. "If there's no life on the sample, that does not prove that there's no life on Europa," he said, "but in reality, it makes it a lot more difficult to get the next mission off the ground."

An especially wicked element of this tester's regress was of concern to the lander scientists: the problem of the false negative. That is, Europa might foster life, but their instruments might not detect it. Alternatively, the moon could be primed for life, but life might not be present. As one scientist put it, "As excited as many of us are about finding signs of life elsewhere, *Europa doesn't care about that*. Europa could be entirely well habitable, but entirely uninhabited" (March 19, 2017, emphasis mine). As the precise properties of the object they are investigating are unknown or even unknowable—such as life on Europa—the outcome of their future tests was precarious.

This sensibility toward planetary cooperation and false negatives was also projected into mission histories. The lander team took its cues from the Viking mission, a NASA project that landed two probes on Mars in 1976 to look for signs of life. Uncertainty over Mars' surface features in the seventies famously led that team to change their landing site at the last minute due to hazardous rocks, invisible in the blurry orbital images captured by prior scouting missions (Ezell & Ezell, 1984).[2] Europa lander participants judged these issues to be parallel to their own lack of knowledge of Europa's surface features 40 years later. They referred to their lander as "Viking on ice," and their chief scientist interviewed Viking members and read its mission reports.

I had always heard planetary scientists describe the Viking life-detection experiments as a failure, either because there was no life on Mars to detect or (more generously) because the instrumentation was mismatched to the then-unknown Martian environment. The scientists leading the Europa lander meetings recast this historical experience as: "Viking worked beautifully but *Mars just did not cooperate*." In discussion behind the scenes and in front of their senior review panel of engineers and scientists evaluating the mission proposal, the lead scientist repeated this interpretation, addressing "the fear of Viking" (i.e., the fear of failure) directly:

> *… let's not make the same mistakes that Viking purportedly made when it did not find life on Mars and, some say, set the Mars Program back decades. I think that's a false read of history … I think that Viking was tremendously successful.* Mars didn't, did not cooperate. *If Pathfinder [20 years later, had] landed on a nice little green golf course, some might say that Viking went before its time and we should have done*

> *more mapping [to identify a greener place to land]. But that was not the case. Viking was very successful.* *(Lander mission review, June 17, 2017; my emphasis)*

The lead lander scientist recast Viking's failure as based on a limited understanding of life: one based on metabolic indicators or the detection of organics, as opposed to microbial understandings that populated later literature. Viking's life detection tests therefore proposed indicators that Mars could not—or perhaps refused to—meet. When the Europa lander team compiled their "recommendations that have come from Viking," this included the need to "assume contrasting definitions for life" including biochemical, geological, and chemical contexts, to avoid the problem of the false negative that had purportedly plagued Mars. As Viking went from a failure to a success in their talk, Mars also went from a passive planet to an *uncooperative* world: a place that might have had life on it, after all—and we might have known it! —if only it had cooperated.

3 | THE INSTITUTIONAL REGRESS

At first glance, this situation appears to be a rather wicked case of what Harry Collins calls the "experimenters' regress" (Collins, 1985), later adopted in sites of technological testing (MacKenzie, 1990). The regress states that in order to know whether or not the test worked properly, one must know the outcome in advance. For instance, detecting gravitational waves—if they are there for the detection—requires building a gravitational wave detector. But we will only know if the gravitational wave detector is working if it detects gravitational waves; and we will only know if we have detected gravitational waves if we build a working detector—and so on. Put differently, if an experiment fails, it is unclear who is to blame: a faulty instrument, a clumsy investigator, an incorrect theory, or a mischaracterized object. As Collins explains, "the competence of experimenters and the integrity of experiments can only be ascertained by examining results, but the appropriate results can only be known from competently performed experiments, and so on" (Collins, 1985, p. 130). Inconclusive results simply raise too many questions about the quality of instrumentation or the trustworthiness of the testers to stand alone.

When a hoped-for result is not found, is undetermined or uncertain going into the test situation, scientists must deploy considerable social resources to solve the problem, from debating what constitutes a good—or bad—detector, to managing reputations. But the resources that testers typically deploy to resolve this regress only partially apply to the Europa case. Like gravitational waves or solar neutrinos, the moon is undetectable without instrumentation; yet these instruments must travel hundreds of millions of miles away from any "magic hands" that might fix, modify, or tweak an investigation in progress (Collins, 1985; Polanyi, 1966). The form and qualities of these prospective probes are also under discussion, so there is no calibration routine that could develop trust in instrumentation (Vertesi, 2015), and no similarity and difference judgments to adjudicate (Pinch, 1993). Further, a regulatory appeal or a simulation tweak cannot change the parameters or interpretations of a prospective test when implementation is over a decade away (Downer, 2007).

This regress is not only beset with ontological and methodological uncertainty: such as, questions about the moon or about prospective instrumentation. It is also plagued with *institutional uncertainty*. In the wake of the 2009 financial crisis, the Europa orbiter faced no less than *two* intensive redesigns to downsize before it was even allowed to begin preparations toward Phase A. It was only when a Congressman from Texas who was enthusiastic about the prospect of finding life on Europa was put in charge of the relevant appropriations committee, that federal legislation was put in place to fund both the orbiter mission and studies for a prospective lander. Funding for the Europa orbiter and lander is therefore tied up in political machinery subject to congressional turnover, government shutdowns, and budgetary holdups, such that there is no guarantee that money for such a mission, its science, and its personnel will flow from the space agency to the science community (as in Aviles, 2018; Berman, 2012; Block & Keller, 2009).

There are micro-institutional issues at play too. Without an established mission there are no organizational practices, workplace routines, cultures or other technics that might stabilize a shared future (Beckert & Bronk, 2018;

Feldman, 2004; Fine, 2010; Vaughan, 1997). As the prospective missions involve scientists from planetary science, astrobiology, and space engineering, the cluster of institutions, patrons, or objects that can stabilize their knowledge claims (Callon, 1984; Latour, 1983, 1987, 1993) or adjudicate on the validity of results (Collins, 2004; MacKenzie, 1990) may diverge on key issues. In this way, an organizational context is essential to managing the "creative frictions" that could generate interdisciplinary insight into Europa's properties (Girard & Stark, 2003). Thus both missions moved forward under conditions of institutional uncertainty as to whether or not they would be fully funded, institutionally staffed, and supported throughout their Phase A and beyond.

One scientist I interviewed brought these elements together as he explained the process of proposing a successful mission to Mercury:

> *I mean, when you put in the proposal*, this is the sort of thing you will use to argue for the mission, you'll argue for the science. *[… Y]ou hope that you'll be able to make that measurement and you make your best estimate of what you'll get going in. Sometimes circumstances don't work as predicted.* You're also counting on the planet to cooperate. If the planet doesn't do what you're expecting it to because this is a discovery mission. *We've never orbited Mercury. You can't fault yourself for hoping that you can do it and then not being able to.*
>
> *(Interview, May 19, 2015, emphasis mine)*

The nature of the scientists' mission is such that they do not know what they will find, even though they must lay out strong characterizations and proposals in advance in order to instantiate the organizational and institutional environment under which such questions could even be answered. As such, these cases speak to the *institutional conditions* under which Europa, Mars, or Mercury become "uncooperative" in testing: that is, it points to an "institutional regress." In the proposal phase, there is no organizational order or structural stability in the investigation to speak of. The assembled scientists are not selected to join the mission nor fully funded to pursue their investigations. Many are working on "soft money" and face difficulties in assembling full salaries at their institutional homes, while others donated their time. One senior scientist at the lander meetings repeatedly requested that the headquarters representative at least identify a mission *class* they should aim for in their planning. This would help to set expectations for their resources, reporting structure, and social relations that they might be able to rely upon in future.

Alongside their attempts to frame uncertain, future-oriented action and stabilize an organization for their craft, then, the moon's ontological flexibility reveals a fundamentally unstable institutional setting. Under these conditions, when "sometimes circumstances don't work as predicted" it is not the scientists who are at fault, or the fact that they did not know enough about the planet to produce an appropriate test. Instead, their form of talk places the *planet* at fault for not cooperating with the scientists' plans, and for not being the object that scientists thought it was.[3]

4 | ONTOLOGICAL FLEXIBILITY AND ANALYTIC OPPORTUNITY

If institutional context is central for addressing the experimenter's regress, then we should expect to see the language of "cooperation" fade away when institutional conditions become more certain. I observed this effect on the Europa orbiter team. Unlike the lander, the orbiter mission began its Phase A development period in 2015. NASA selected a team, a suite of instruments that this group of scientists and engineers would eventually build, and provided them with startup funding, an organizational chart, and clear lines of authority. While they still faced uncertainty as to whether or not their mission would pass the various review phases prior to construction and launch, Europa scientists' work on the orbiter took place under a greater degree of institutional stability than their work on the lander.

It was against this backdrop that orbiter scientists began to investigate the potential for "plumes" on Europa, an effect of geysers of water shooting into space from crevices in its frozen shell. Similar plumes had been spotted on another icy moon, Enceladus, by the Cassini mission 10 years prior. The orbiter team therefore hosted a joint meeting with Cassini scientists to discuss these interconnections. At this meeting, a Europa orbiter scientist presented a new paper she had co-authored with the lead lander scientist and a few other teammates, claiming to spot the plumes. The group had applied for time on existing telescopes, including the Hubble space telescope and ground-based observatories: all well-established entities. They collected 17 images over 3 months. When the first part of their experiment failed to detect the plumes, she did not explain this as a failure of her own competence or as a null-detection of the purported plumes. Instead, she stated: "Unfortunately Europa is not that cooperative." Because Voyager spacecraft images had proven that Io, Europa's neighboring moon, was covered in active volcanos, she joked, "I like Io. Way more cooperative!" The audience laughed.[4] Once Europa aligned in the right place on the telescopic detector, however, significant features appeared in the data. "Maybe Europa was a little cooperative, I don't know," she shrugged (Europa-Enceladus workshop, October 15, 2016).

This scientist's language will sound familiar to science studies scholars as one of conscripting a recalcitrant object into a robust network of human/non-human relations. Indeed, listening to these scientists talk, they appeared not to be adherents of Harry Collins' Empirical Programme of Relativism, but instead to be lay Actor-Network Theorists! As they discussed definitions for life or detection thresholds for plumes and the possible future instrumentation and organizational arrangements that they could reasonably construct, launch, and operate to detect these features, they appeared to perform a prospective "heterogeneous engineering" that aimed to stabilize the preconditions of their eventual knowledge production (Law, 1987). They hoped to "muster allies" such as equipment, funders, publics, and objects of study in order to produce laboratory-sanctioned truth claims that can move in the world (Latour, 1983, 1987, 1993), and to establish performative circuits within which Europa might cooperate, once it was conscripted into a network of instrumentation, institutes, and willing scientists (Akrich, 1992; Callon, 1984; MacKenzie, 2006; Muniesa & Callon, 2007). This fieldsite therefore presents an analytical puzzle: how literally should we take actors' own, local accounts of agency in the field? Is language about "uncooperative" planets a lay sociological term for the struggles of heterogeneous engineering, the unruliness of material objects, and the isolation of "agential cuts" in the actor-network (Barad, 2003; Barad, 2007; Strathern, 1996)?

Certainly we could *analytically* describe the future result of such tests as a question of conscripting a non-human planet into a network of devices, objects, and scientists that would stabilize it, give it meaning, and allow it to perform as intended. This does not mean, however, that we must mirror scientists' agential associations and disassociations in our own analytic discussion. It is one thing to elevate an observed and reported phenomenon to a theoretical category (Glaser & Strauss, 2009), but another altogether to associate empirical phenomena directly with theoretical categories without risking, at the very least, the problem of residuals (Bowker & Star, 1999) or the difficulties of formal analysis (Garfinkel, 2002).

Instead, I suggest we return to Michael Lynch's question of *ontography*: that is, the empirical investigation of practical ontologies: "how particular identities and differences are negotiated and instantiated in specific circumstances" (Lynch, 2013). I propose that the sociologist of scientific knowledge treat actors' *ontological flexibility*—these linguistic shifts from passive to "uncooperative" worlds, for instance—as a form of talk that reveals the shifting conditions of material action and accountability. Describing the moon as a passive or an active object neither violates a stable category nor reveals scientists' frustrated attempts to enact it or to conscript it to their cause. Instead, the shift demarcates a change in the institutional circumstances of the moon's investigation. The missing link in the telescopic observation of the plumes, for instance, was not only the moon's ready cooperation or the assembly of human and instrumental actors, each of which had worked together before. The mission also possessed new-found institutional stability: an organization, funding for its investigations, legitimate interactions with telescopic observatories, and the promise of continued support.[5]

This is not only visible in talk of "cooperation," but also in material practices in the laboratories supported by the mission under development. Orbiter laboratories featured planetary simulators and analog environments, staffed by scientists who sometimes spent months working in Antarctica or on oceanographic expeditions to better understand the Europa they might visit. Prior work on planetary analogs has described them as sites that turn planets into places, as constraints upon digital manipulation, and as training grounds for exploration (Messeri, 2014; Vertesi, 2015; see also Messeri, 2016; Shindell, 2010). Observing them this time around, I looked to see if these prior analyses would hold, or if laboratory conditions stabilized Europa's participation amid ontological complexity (Mol, 2002; Ribes & Jackson, 2013). Neither was the case: instead, these simulations *prepared for the radically different moons* that the craft might eventually encounter. That is, as institutional support for the missions coalesced, scientists' material practices in the laboratory *made multiple Europas* cooperatively present for investigation—in order to demonstrate that the real Europa would eventually cooperate with their future probe.[6]

For example, one scientist I visited grew ice crystals that approximated different hypothesized values for Europa's ice shell, and applied diffractive optics to qualify whether or not different flight spectrometers would be able to tell those scenarios apart (Figure 2). Another bombarded a test plate with small particles to characterize how well different particles emerging from differently composed Europas would be detected as part of his eventual instrument (Figure 3). Unlike certain Mars simulators, which attempt to imitate atmospheric or ground conditions on a better-known planet, these Europa laboratories were outfitted with pluralities: many different compounds, variants, compositions, and qualities.

These multiple Europas involved in prospective testing were not incommensurate or associated with different expert visions (Mol, 2002). Instead, they were put to work as a practical form of institutional appeal. As the orbiter team neared their critical design review to exit Phase B and begin building their spacecraft, I observed a meeting where members of an orbiter instrument team described their process of "science verification and validation." Without such a "validation," funding would not be released for their instrument's construction. Working from the literature, the instrument's Principal Investigator (PI) first compiled a series of published conjectures about the properties of Europa's ice shell that their instrument was designed to investigate: thick or thin crust, rough or smooth surface, pockets of water or solid ice, and more. He identified measurements that his instrument would be able to take that could perceive a Europa with such properties. His team then ran a series of simulations, in which the computer assembled a variety of possible ice shells with a mixture of these potential characteristics. Next, he explained: "We throw multiple Europas at the instrument, hundreds and thousands of them." The goal was to demonstrate that his instrument would be able to make successful observations regardless of the configuration of Europa's ice shell. The review board would only be satisfied, "as long as more than half of the Europas

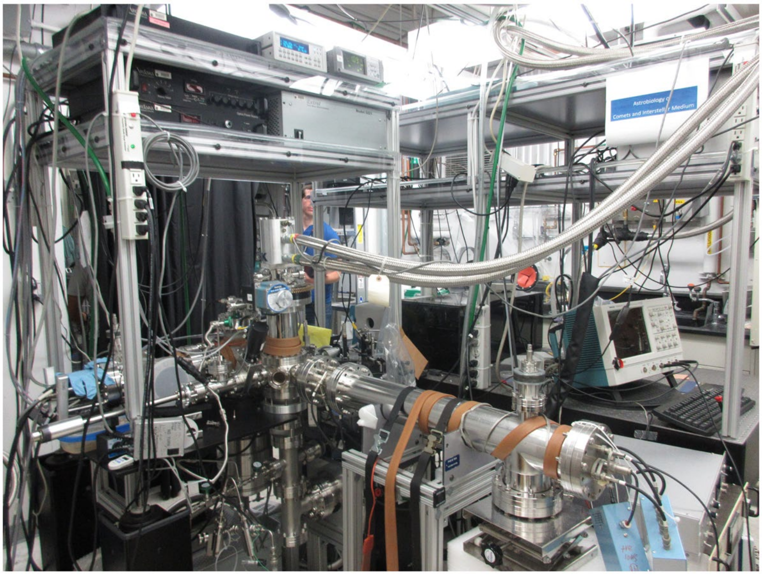


FIGURE 2 A cryo-laboratory for spectral investigation of Europa's ice properties [Colour figure can be viewed at wileyonlinelibrary.com]

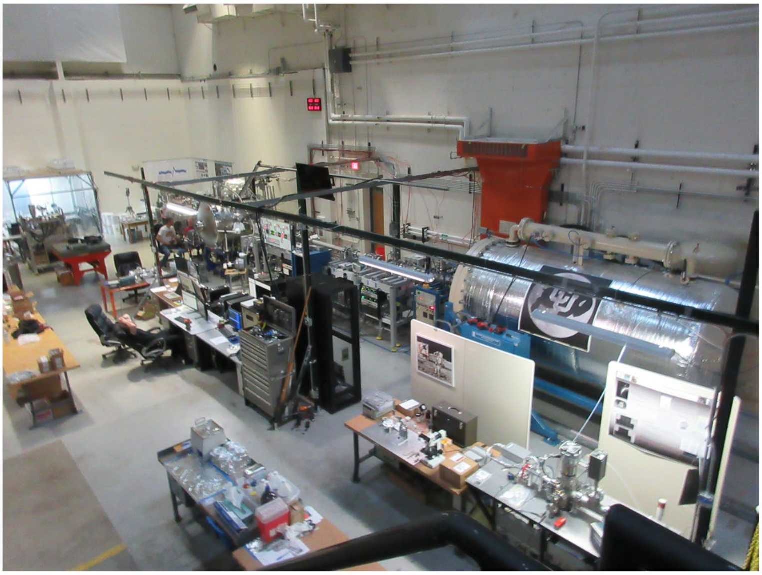


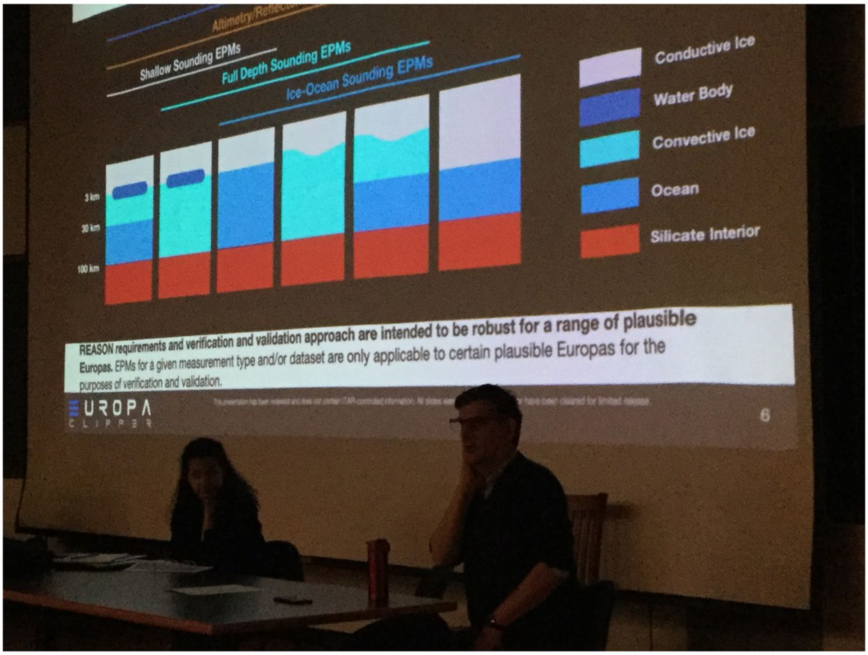
FIGURE 3 A facility dedicated to testing how different particles will interact with a prospective flight instrument [Colour figure can be viewed at wileyonlinelibrary.com]

in the literature make it through" this rigorous test. The PI's presentation slides stated that his "requirements and verification and validation approach are intended to be robust for a range of plausible Europas," and indicated that certain measurements "are only applicable to certain plausible Europas for the purposes of verification and validation" (Figure 4; June 12, 2018). Multiple Europas were manufactured and cooperatively tested *in order to pass the upcoming review.*

Multiplicity, here, was an observable, practical, material approach to securing continued institutional support for the instrument team. The transition I observed from an "uncooperative" Europa to "a range of plausible Europas" did not speak to a material ontological or methodological change. The instruments are still not yet built or tested, and only the eventual encounter with the moon in situ will provide the ultimate "test of the test" (Collins, 1985; Pinch, 1993) to ascertain whether these early judgements were valid.[7] Instead, such ontological flexibility speaks to changing forms of institutional certainty.

This resonates with prior studies of spacecraft teams during the operations phase of their missions, when funding is secured and roles are routinized. Sources of uncertainty here are externalized but explained in routine ways through other ontological shifts. For instance, Vertesi notes that scientists on the Mars Exploration Rover mission engaged in "linguistically drawing and redrawing the boundaries between human and machines" (Vertesi, 2015, p. 189) by switching from the first person pronoun to the third person when the robots did not obey commands. Mazmanian, Cohn and Dourish (2014) describe "dynamic reconfiguration" as an organizational process that ascertained, under conditions of a flight anomaly, which social or material agents were to blame. For mission members, neither case of ontological flexibility revealed a crisis of configuration among humans and machines (Haraway, 1991; Suchman, 2006). Rather, these were tools that were ready-to-hand: a locally acceptable switch deployed in the complexity of changing object relations in which all members knowingly took part. Examining how and when actors themselves figure and reconfigure their objects of investigation reveals that object properties are not so much stake, as they are embedded in and arising from organizational practices on Earth. In the case of the Europa missions, the moon's "cooperation" depended less on its objective properties than upon a cascade of institutional uncertainties on Earth.

Amid these uncertainties, scientists are at a loss to establish the performative loops that guarantee trustworthy experiments—and results. This was certainly the case on the lander science definition team, where scientists spoke of being "in ambiguity land," and expended considerable effort attempting to stabilize different



REASON requirements and verification and validation approach are intended to be robust for a range of plausible Europas. EPMs for a given measurement type and/or dataset are only applicable to certain plausible Europas for the purposes of verification and validation.

FIGURE 4 “A range of plausible Europas” cooperate in helping a science team pass their instrument’s review [Colour figure can be viewed at wileyonlinelibrary.com]

aspects of their prospective mission. They elaborated in detail what they would need to detect in order to satisfy a scientific objective, as well as how to prove where a detected microbe came from. They appealed to the public by considering what questions they would be asked at the eventual press conference; and to peer review as one scientist wondered aloud how to satisfy “the senior editors of *Science* magazine” 20 years from now. They proposed varieties of prior instruments that had flown on other missions in an effort to ground prospective claims in similarity judgements, known calibration issues, and trustworthy reports. They imported a pre-existing lander design from the Phoenix mission and the landing system from the Curiosity rover. They also appealed to earth-bound analogs of icy terrain in the Antarctic to stabilize their sense of Europa’s ice sheath. Each of these elements offered a form of grounding by appealing to known quantities, limiting ambiguity. Yet each was also up for interpretation. After all, a single tweak to instrumentation would demand a change in the entire probe’s design, and vice versa; similarity and difference judgments were at stake in the Antarctic exercises; and the process of peer review in the 2030s was anyone’s guess. Returning to the lander case from the orbiter’s embrace of cooperative multiplicities in a later phase of its development, we are reminded that the question of whether or not scientific investigations can proceed with confidence is not only one of object uncertainties or the participation of instrumentation. When the organization is not yet established and institutional support is contingent upon insurmountable conditions, there is no available exit to the tester’s regress.

5 | CONCLUSION: TESTING INSTITUTIONS

Scientific institutions are at a crossroads. The decade following the crash of 2009 has seen agencies with limited funds to transfer to public projects, operating under new political trajectories. As the scientists I observed faced an experimenter’s regress associated with distant life detection on a moon that is largely unknown, they did not blame the disinterest from NASA headquarters in funding large projects, the change in federal administration, or the needs of the institutional players involved—all of which might comprise (or compromise) their eventual project. Instead, they simply described their object of investigation as potentially uncooperative.

Tracing ontological flexibility in scientists' talk reveals the local anxieties associated with organizational and institutional instability as they attempt to plan for an unpredictable future during a tumultuous present: that is, it points to an institutional component inherent to the experimenter's regress. This form of talk does valuable work as a local way of accounting for uncertainties while facing the threats to competence posed by the experimenter's regress. The profundity of this regress is not only due to epistemic uncertainties about Europa and its ability to harbor life, but to institutional uncertainties on Earth and the inability to stabilize prospective investigations.

Actors hedge against their unknowable future by shifting local organizational deficiencies onto the moon itself. The object of investigation thus becomes a shape-shifting trickster who can transform one or another future path into a reality (Haraway, 1988), upon which blame may be locally cast, hope might be projected, and accountings for future action might hinge. Should the moon not turn out to be quite what was intended, neither Earth-bound humans, their political supporters, nor their expensive instrumentation are at fault.

Importantly, then, scientists have their own local ways of describing the multiplicity, dualities, and agency of the objects in their purview as much as do sociologists of science. As object agencies and other qualities emerge and recede in scientists' testing talk, these same shifts are something of a test for the sociologist. They require us to return to our initial investigations of testing—and to other features of the sociology of scientific knowledge—with fresh eyes, to see what invisible role organizations and institutions might always have played in the background of early sociological accounts. Our core analytics were developed at a time when Western scientists at least enjoyed robust institutions for their work, relatively high levels of sustained funding, and powerful advisory relationships with government agencies. As the path forward becomes less certain, we must return to questions of patronage not simply as a question of pitches or appeals (Biagioli, 2007), of twisting scientific concerns to satisfy funders (Berman, 2012; Oreskes & Conway, 2010), or of the effects of various institutions upon epistemic work (Biagioli, 1993; Collins, 1998; Fine, 2010; Shapin & Schaffer, 1985; Vertesi, 2020). Stable institutional contexts are a precondition for practical epistemic and ontological accomplishments. Without them, object ontologies dissolve in scientists' talk and evade their interaction.

As one of the few ways in which a nascent group can address the considerable tensions in which contemporary scientific work takes place, ontological flexibility deserves both our investigation and our empathy. As the practices and sites of testing expand to a variety of social problems realms—with stakes raised, expectations high, and public funds limited—we should expect to see many more such "uncooperative" objects accompany the application of testing to public life.

ACKNOWLEDGEMENTS

The author thanks David Stark, Noortje Marres, Trevor Pinch, Michael Lynch, Jordan Bimm, David Reinecke, two anonymous reviewers, and the members of the Put to the Test workshop for their assistance in improving this article. Sincere thanks also to the Europa Clipper science team and the Europa Lander science definition team, especially Bob Pappalardo, Kevin Hand, and Cynthia Phillips.

CONFLICT OF INTEREST

The author reports no conflict of interest.

DATA AVAILABILITY STATEMENT

Author elects to not share data (due to privacy/ethical restrictions).

NOTES

[1] At time of writing, there was also no selected launch vehicle to transport the probe to space.

[2] It also led mission scientist Carl Sagan to suggest placing a motion detector on Viking's cameras in case a Martian should walk past (see Sagan, 1969).

[3] For a related, meaningful grammatical substitution, see Ochs, Gonzales, and Jacoby (1996).

[4] Some of these scientists had planned to take images of Io's limb precisely to look for its famous volcanoes, which they had hypothesized to be on its surface. Laughter in the audience thus came from those who knew what it was like to have an object of investigation eventually cooperate and perform as intended. Further, thanks to subsequent missions, Io's properties are better known than Europa's, potentially making it "more cooperative." On performativity in experiment see MacKenzie (2006 and Muniesa and Callon (2007).

[5] From the actor-network perspective, we might see the material practices I describe here enacting a cooperative, conscripted Europa, or to consider this study as a moment in a stage of assembling a functioning network that already includes the conscripted moon. Yet the moon is still far from reach. Each of these materials—laboratory equipment, the printed pages on the wall—are conscripted instead into building a stabilized institutional context through which money is transferred, equipment is purchased, and an organizational form is established. This network will ultimately collapse the distance between Earth and Europa and conscript the planet. In other words, scientists must first attain local organizational stability in order to successfully enact Europa. Without this functioning, funded laboratory environment they cannot hope to "raise the world" (Latour, 1983).

[6] According to space historian Jordan Bimm, simulated Mars environments by the Air Force and by NASA in the 1960s also moved away from fully approximating Mars' atmosphere to creating 'Mars-like' or 'Mars-lite' environments (see Siegal et al., 1963).

[7] In other words, this was a test (in silico), for a test (the design review), for a test (the eventual investigation).

REFERENCES

Akrich, M. (1992). The description of technical objects. In W. E. Bijker & J. Law (Eds.), *Shaping technology/building society* (pp. 225–258). Cambridge, MA: MIT Press.

Ashworth, W. B. J. (1991). The Habsburg circle. In B. T. Moran (Ed.), *Patronage and institutions: Science, technology, and medicine at the European court, 1500–1750* (pp. 137–167). Rochester, NY: Boydell Press.

Aviles, N. B. (2018). Situated practice and the emergence of ethical research: HPV vaccine development and organizational cultures of translation at the National Cancer Institute. *Science, Technology, & Human Values*, *43*(5), 810–833.

Barad, K. (2003). Posthumanist performativity: Toward an understanding of how matter comes to matter. *Signs*, *28*(3), 801–831.

Barad, K. M. (2007). *Meeting the universe halfway: Quantum physics and the entanglement of matter and meaning*. Durham, NC: Duke University Press.

Barnes, B. (1982). *T.S. Kuhn and social science*. New York, NY: Columbia University Press.

Beaulieu, A. (2010). Research note: From co-location to co-presence: Shifts in the use of ethnography for the study of knowledge. *Social Studies of Science*, *40*(3), 453–470.

Beckert, J., & R. Bronk (Eds.). (2018). *Uncertain futures: Imaginaries, narratives, and calculation in the economy* (New product edition). New York, NY: Oxford University Press.

Berman, E. P. (2012). *Creating the market university: How academic science became an economic engine*. Princeton, NJ: Princeton University Press.

Biagioli, M. (1993). *Galileo, courtier: The practice of science in the culture of absolutism*. Chicago, IL: University of Chicago Press.

Biagioli, M. (2007). *Galileo's instruments of credit: Telescopes, images, secrecy*. Chicago, IL: University of Chicago Press.

Block, F., & Keller, M. R. (2009). Where do innovations come from? Transformations in the US economy, 1970–2006. *Socio-Economic Review*, *7*(3), 459–483.

Bowker, G., & Star, S. L. (1999). *Sorting things out: Classification and its consequences*. Cambridge, MA: MIT Press.

Callon, M. (1984). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review*, *32*(1_suppl), 196–233.

Collins, H. M. (1985). *Changing order: Replication and induction in scientific practice*. London: Sage Publications.

Collins, H. M. (1998). The meaning of data: Open and closed evidential cultures in the search for gravitational waves. *American Journal of Sociology*, *104*(2), 293–338.

Collins, H. M. (2004). *Gravity's shadow: The search for gravitational waves*. Chicago, IL: University of Chicago Press.

de Vaan, M., Vedres, B., & Stark, D. (2015). Game changer: The topology of creativity. *American Journal of Sociology*, *120*(4), 1144–1194.

DiTomaso, N. (2001). The loose coupling of jobs: The subcontracting of everyone? In I. Berg & A. Kalleberg (Eds.), *Sourcebook of labor markets: Evolving structures and processes*. New York, NY: Kluwer Academic/Plenum Publishers.

Downer, J. (2007). When the chick hits the fan: Representativeness and reproducibility in technological tests. *Social Studies of Science*, *37*(1), 7–26.

Ezell, E., & Ezell, L. (1984). *On Mars: Exploration of the Red Planet, 1958–1978. NASA history series*. Washington, DC: National Aeronautics and Space Administration.

Feldman, S. (2004). The culture of objectivity: Quantification, uncertainty, and the evaluation of risk at NASA. *Human Relations*, *57*(6), 6910718.

Fine, G. A. (2010). *Authors of the storm: Meteorologists and the culture of prediction* (Paperback ed.). Chicago, IL: University of Chicago Press.

Garfinkel, H. (2002). *Ethnomethodology's program: Working out Durkheim's aphorism*. Lanham, MD: Rowman & Littlefield.

Girard, M., & Stark, D. (2003). Heterarchies of value in Manhattan-based new media firms. *Theory, Culture & Society*, *20*(3), 77–105.

Glaser, B. G. & Strauss, A. L. (2009). *The discovery of grounded theory: Strategies for qualitative research* (4. paperback printing). New Brunswick, NJ: Aldine.

Haraway, D. J. (1988). Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist Studies*, *14*(3), 575–599.

Haraway, D. J. (1991). *Simians, cyborgs, and women: The reinvention of nature*. New York, NY: Routledge.

Helmreich, S. (2009). *Alien ocean*. Cambridge, MA: MIT Press.

Helmreich, S., Roosth, S., & Friedner, M. I. (2015). *Sounding the limits of life: Essays in the anthropology of biology and beyond*. Princeton, NJ: Princeton University Press.

Latour, B. (1983). Give me a laboratory and I will raise the world. In K. Knorr Cetina & M. Mulkay (Eds.), *Science observed: Perspectives on the social study of science* (pp. 141–170). London: Sage.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Latour, B. (1993). *The pasteurization of France* (A. Sheridan & J. Law, Trans.). Cambridge, MA: Harvard University Press.

Law, J. (1987). Technology and heterogeneous engineering: The case of Portuguese expansion. In T. Pinch, W. E. Bijker, & T. P. Hughes (Eds.), *The social construction of technological systems: New directions in the sociology and history of technology* (pp. 111–134). Cambridge, MA: MIT Press.

Lynch, M. (2013). Ontography: Investigating the production of things, deflating ontology. *Social Studies of Science*, *43*(3), 444–462.

MacKenzie, D. (1990). *Inventing accuracy: A historical sociology of nuclear missile guidance*. Cambridge, MA: MIT Press.

MacKenzie, D. A. (2006). *An engine, not a camera: How financial models shape markets*. Cambridge, MA: MIT Press.

Mazmanian, M., Cohn, M., & Dourish, P. (2014). Dynamic reconfiguration in planetary exploration: A sociomaterial ethnography. *MIS Quarterly*, *38*(3), 831–848.

Messeri, L. (2014). Earth as analog: The disciplinary debate and astronaut training that took geology to the moon. *Astropolitics*, *12*(2–3), 196–209.

Messeri, L. (2016). *Placing outer space: An Earthly ethnography of other worlds*. Durham, NC: Duke University Press.

Mol, A. (2002). *The body multiple: Ontology in medical practice*. Durham, NC: Duke University Press.

Muniesa, F., & Callon, M. (2007). Economic experiments and the construction of markets. In D. MacKenzie & L. Siu (Eds.), *Do economists make markets? On the performativity of economics* (pp. 163–189). Princeton, NJ: Princeton University Press.

Neff, G. (2012). *Venture labor: Work and the burden of risk in innovative industries*. Cambridge, MA: MIT Press.

Ochs, E., Gonzales, P., & Jacoby, S. (1996). "When I come down I'm in the domain state": Grammar and graphic representation in the interpretive activity of physicists. In E. Ochs, E. A. Schegloff, & S. Thompson (Eds.), *Interaction and grammar* (pp. 328–370). New York, NY: Cambridge University Press.

Oreskes, N., & Conway, E. M. (2010). *Merchants of doubt: How a handful of scientists obscured the truth on issues from tobacco smoke to global warming*. London: Bloomsbury Press.

Pinch, T. (1986). *Confronting nature: The sociology of solar-neutrino detection*. Dordrecht, the Netherlands: D. Reidal.

Pinch, T. (1993). Testing, one, two, three-testing: towards a sociology of testing. *Science, Technology & Human Values*, *18*, 25–41.

Polanyi, M. (1966). *The tacit dimension*. Gloucester, MA: P. Smith.

Powell, W. W., Koput, K. W., & Smith-Doerr, L. (1996). Interorganizational collaboration and the locus of innovation: Networks of learning in biotechnology. *Administrative Science Quarterly*, *41*(1), 116–145.

Powell, W. W., Packalen, K., & Whittington, K. (2012). Organizational and institutional genesis: The emergence of high-tech clusters in the life sciences. In J. F. Padgett & W. W. Powell (Eds.), *The emergence of organizations and markets*, Princeton, NJ: Princeton University Press.

Ribes, D., & Jackson, S. J. (2013). Data bite man: The work of sustaining a long-term study. In L. Gitelman (Ed.), *Raw data is an oxymoron* (pp. 147–166). Cambridge MA: MIT Press.

Sagan, C. (1969). *Draft work statement for motion detection capability of Viking 1973 Mars camera*. Retrieved from NASA Ames History Office, NASA Ames Research Center, Moffett Field, CA (Levinthal Papers, Box 14, Folder 20).

Shapin, S., & Schaffer, S. (1985). *Leviathan and the air pump: Hobbes, Boyle and the experimental life*. Princeton, NJ: Princeton University Press.

Shindell, M. B. (2010). Domesticating the planets: Instruments and practices in the development of planetary geology. *Spontaneous Generations: A Journal for the History and Philosophy of. Science*, *4*(1), 191–230.

Siegel, S. M., Halpern, L. A., Giumarro, C., Renwick, G., & Davis, G. (1963). Martian biology: The experimentalist's approach. *Nature*, *197*(4865), 329-331.

Stark, D., & Vedres, B. (2009). *Structural folds: Generative disruption in overlapping groups*. MPIfG discussion paper. Retrieved from http://www.econstor.eu/handle/10419/36525

Strathern, M. (1996). Cutting the network. *Journal of the Royal Anthropological Institute*, *2*(3), 517.

Suchman, L. (2006). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Retrieved from http://ebooks.cambridge.org/ref/id/CBO9780511808418

Vaughan, D. (1997). *The Challenger launch decision: Risky technology, culture, and deviance at NASA* (1st ed.). Chicago, IL: University of Chicago Press.

Vertesi, J. (2020). *Shaping science: Organizations, decisions, and culture on NASA's teams*. Chicago, IL: University of Chicago Press.

Vertesi, J. (2015). *Seeing like a Rover: How robots, teams, and images craft knowledge of Mars*. Chicago, IL: University of Chicago Press.

Mazmanian, M., Cohn, M., & Dourish, P. (2014). Dynamic Reconfiguration in Planetary Exploration: A Sociomaterial Ethnography. MIS Quarterly, 38(3), 831–848.

How to cite this article: Vertesi J. Testing planets: Institutions tested in an era of uncertainty. *Br J Sociol*. 2020;71:474–488. https://doi.org/10.1111/1468-4446.12725

Received: 12 March 2019 | Revised: 14 January 2020 | Accepted: 19 February 2020

DOI: 10.1111/1468-4446.12748

SPECIAL ISSUE

WILEY

The red and the black: China's social credit experiment as a total test environment

Jonathan Bach

The New School, New York, NY, USA

Correspondence
Jonathan Bach, The New School, New York, NY, USA.
Email: bachj@newschool.edu

Abstract

China's social credit system is an unusually explicit case where technology is used by multiple actors to turn human behavior into a test object on behalf of the state's goal of modifying the larger social environment, making it an intriguing setting for thinking about the new sociology of testing. This article considers how China's search for a usable "credit" score to both allocate financial resources and explicitly measure a citizen's trustworthiness creates an emergent experimental system of governance similar to, yet not quite captured by, the kinds of experimental processes observed in literature on the platform as a form of market-based governance. As a site where "seeing like a state" and "seeing like a market" converge, the social credit system is a vantage point for observing the changing relationship between moral and economic domains in an era of digital platforms. The article highlights the experimental quality of the system and its emerging system of governance structured around reward and punishment and argues that strategic ambiguity, institutionalized through the affordances of digital platforms, is an important part of the design of this large-scale social experiment.

KEYWORDS

ambiguity, China, experimentation, platform, social credit, value

1 | INTRODUCTION

“Experiment is simply observation under controlled conditions” flatly declared a 1917 article on sociology and the experimental method (Chapin, 1917, p. 133). Just over a century later, as the articles in this special issue attest, sociologists would hardly make such a sweeping claim: today neither observation, nor control (in its multiple senses), nor even agreement on conditions can be taken for granted in the sociology of experimentation. As Trevor Pinch noted already in 1993, “we live in the age of the test” (p. 27) in which testing and testing discourses have created the conditions for what Marres and Stark (this issue) call the “total test environment.” There is hardly a better example of this today than China's social credit system—an audacious national endeavor to create a “trustworthy” society by developing a system that collects, digitizes, and shares data on hundreds of millions of people and corporate entities in order to generate an automated national system of rewards and punishment.[1] Can the social credit system platform be considered a testing system? If so, what, and who, is being tested? This article argues that the social credit system is an intriguing case for thinking about the new sociology of testing because it is an unusually explicit setting where technology is being used by a variety of actors to turn human behavior into a test object in the service of a state goal of modifying the larger social environment.[2] More specifically, it considers how an emergent experimental system of governance can be discerned in China's search for a usable “credit” score as a means for both allocating financial resources and explicitly measuring the trustworthiness of citizens, a system that is both similar to, and yet not quite captured by, the kinds of experimental processes observed in literature on the platform as a form of market-based governance. The article suggests that through this massive social experiment we can observe the changing relationship between moral and economic domains in an era of digital platforms.

The Chinese social credit system is technologically unprecedented, logistically challenging, and promises a new metric for connecting economic and moral behavior, that of “sincerity.” The stated goal of the social credit system is encapsulated in an oft-quoted phrase from its original planning document, which envisions a future where “the trustworthy roam everywhere under heaven while making it hard for the discredited to take a single step” (Planning Outline, 2014). To achieve this requires a digital architecture upon which rests a nationwide system of incentives and disincentives. These, in turn, seek to shape a sincerity society by bundling together diverse goals, from exposing deadbeats and miscreants in order to strengthen the integrity of banking and safety systems, to encouraging people to keep their dogs on leashes, cross the street properly at crosswalks, and visit one's elderly parents. These incentives and disincentives are produced through an “always on” system that continuously gathers data from a wide and expanding array of behavioral traces, both first order (e.g., online purchases, social media, facial recognition, location tracking) and second order (e.g., reports generated from court or municipal records systems), and feeds them into algorithmic systems which generate the rewards or punishments meant to modify the social environment. Since both data collection and punishment for infractions can take place in the everyday world, everyone can be in principle subject to some kind of test all of the time. To use the banal example of walking in the city, in one of Shenzhen's widely reported pilot experiments, jay walkers will have their face recognized by a camera pointed at the crosswalk, displayed on a large screen on the side of the road for public shaming, have a fine sent directly to their phone with a picture of their indiscretion, and negatively influence their municipal social credit score (Tao, 2018).

Social control through modern management techniques is hardly new, and the social credit system experiment fits the modernist desire to make value objectively visible through numerical scores. This impetus lies at the heart (or heartlessness) of bureaucratic rationalization and its attempts to create legible subjects of governance out of the messy ambiguity of human behavior. The Chinese system thus draws on and shares similarities with other market-based systems that also use big data to score, sort, reward, and punish consumers (if not in such explicit ways). In this sense it is part of an emerging global regime of classification that Fourcade and Healy (2017) argue is now operative. Drawing on examples mostly from the United States, Fourcade and Healy argue that society has entered a phase where “seeing like a market” has replaced the administrative gaze of the state (cf. Scott, 1999), creating classification situations which are as totalizing as anything the state could come up with. Unlike the classic administrative state, however, "[a]s digital traces of individual behaviors are aggregated, stored, and analyzed, markets see people through a lens of deserving and undeservingness, and classification situations become moral projects" that hide their moral judgement behind seemingly impartial and objective methods of scoring (2017, p. 24).

The Chinese case sits squarely within the classifying, sorting, and stratifying logics of information that Fourcade and Healy lay out, but with one important distinction: rather than mystifying moral judgement behind the veil of a market economy, the Chinese state could not be more explicit about connecting moral judgement with economic behavior. What happens when "seeing like a state" and "seeing like a market" converge? Rather than a top-down hierarchy, one possibility this article considers is that this convergence creates an ecosystem with many moving parts that functions in practice through strategic ambiguity. This includes plausible deniability about whether its emergent variations are part of the plan, or trial balloons, unwelcome by-products, or rogue actions. Unlike the images of an unerringly omniscient system that the social credit system often seems to provoke, this article argues that it is more analytically interesting, and even more unsettling, to see it as a complex system of experimentation drawing on legacies from both platform capitalism and socialist model making.

The article begins by presenting the social credit system as a way to align moral and economic domains in the rapidly changing society of post-Mao China. It then highlights the experimental quality of the system, showing how its complex ecology of multiple actors connect computational technological experiments with a history of socialist policy experimentation. The result, and focus of the subsequent section, is an emerging system of experimental governance structured around the reward/punishment system, continuously observing, evaluating, and acting upon behavior which itself is ambiguously located at the intersection of moral and economic performance. The article concludes by arguing that ambiguity is part of the design of this large-scale social experiment, a feature, not a failure, and discusses how it is institutionalized through the affordances of digital platforms.

2 | SUTURING ECONOMIC AND MORAL DOMAINS

In English the word credit originally referred to honor and virtue, but has become more narrowly construed as one's ability to repay a debt and thus increasingly connected to finance (the word "worth" retains more of this multiplicity of meanings, see Stark, 2009). The term "credit" in Chinese (*xinyong* 信用) draws on, and reclaims, the wider meaning of the term, while also referring to financial credit. *Xinyong* is more active than its English equivalent, since the second character *yong* (用) means to use, mobilize, or deploy, and *xin* (信) means belief, trust, or trustworthiness. Thus, one could render it as the deployment of trust, or trust in action. This makes the concept of "credit" well positioned to suture economic and moral domains, allowing it to function as a value across technical, economic, moral, and political registers (Ortiz, 2013, p. 76), as well as traditional and modern utopian visions of meritocratic societies. It is no accident, then, that the concept of social credit emerged to address three kinds of instability in the domains of productivity, trust, and population control alongside China's surging market reforms in the late 20th and early 21st centuries.

One of the most urgent tasks of China's technocratic regimes continues to be to contain, consolidate, and redirect the massive economic and social energy released by Deng Xiaoping's market reforms (begun in 1979). Growth, however, requires immense capital, especially for construction and land development, and capital requires huge amounts of loans. The lack of any credit history for the first generation after Mao made taking risks particularly risky—it was hard to tell who might default, or take out multiple loans for the same project, or have a history of fraud. If the economy was to continue to grow, extending and expanding credit as the basis for a productive market society was important. Maintaining and controlling economic growth also became a vital source of legitimacy for the Communist Party, especially after the political crisis and crackdown of 1989. This modernization of the market is the context in which Jiang Zemin first articulated the need for a social credit system, though as

Creemers (2018, p. 9) notes, it took 5 years before the first Interministerial Joint Conference was convened in 2007 to begin development of the plan, which would be announced as part of legal reforms in 2014.

A major concern then (as now) was the rise of shadow banking to make up for the inadequacies of the formal banking sector, which spurred growth but hid bad loans in the dark. When the 2008 financial crisis struck in the US and spread around the world, it was like a bad preview of what might lurk in the debt that was driving the Chinese economy. Developing criteria for being creditworthy, and driving people out of the shadows, became a major priority. A further problem, and an even bigger one, was how to expand credit to the majority of the population who used cash, had no loan history, and often did not even have a bank account. For the economy to keep growing, people needed access to credit if they were to purchase not only cars, washing machines, and smart phones, but also apartments, travel, and education, and take out loans to start businesses. To enable this there needed to be both proxies for creditworthiness, so that loans could be made even with a lack of credit history, and more opportunities to build a credit history through various kinds of smaller loans. Microcredit and rural financing schemes were important for this, but the challenge went even further to reach all sectors of society, and the Internet emerged as a key vehicle for financial inclusion, allowing for the creation of entirely new opportunities for extending credit and gathering data (Loubere, 2017).

The second challenge was in the area of social trust. While some researchers have argued that, contrary to expectations, a generalized sense of trust is strong in China (Steinhardt, 2012), this is not the common perception and there is no doubt that the government is extremely nervous about the destabilizing consequences of public unrest. A long litany of corruption, scandals, and preventable spectacular disasters is habitually seen as undermining the social fabric. Indeed, the 2014 Social Credit System Planning Outline goes on at length about the depressing regularity and seeming inevitability with which these kinds of things happen, listing "grave production safety accidents, food and drug security incidents…commercial swindles…counterfeit products, tax evasion, fraudulent financial claims…and other such phenomena that *cannot be stopped* in spite of repeated bans" (Planning Outline, 2014, p. 1, emphasis added).

This sober assessment and accompanying concern about the consequences of insufficient trust for continued market development reflects a kind of cognitive dissonance between Deng Xiaoping's famous pronouncement that there was no contradiction between socialism and the market economy and the communist critique of markets as enabling greed, undermining loyalty to any cause other than money, and seducing people to lose their moral bearings and become untrustworthy.[3] This dissonance appears in the well-established trope of the reform era that links the rise of the market economy with a generalized crisis of morality. For example, the year before the Social Credit Score plan was unveiled, Chinese news reported that "Trust Among Chinese 'Drops to Record Low,'" citing an annual survey of "social mentality" (China Daily, 2013). This survey followed lingering public outrage from the watershed 2008 powdered milk scandal that sickened 300,000 children after suppliers added the chemical melamine to increase the protein content. The cause of this scandal lay in the combination of reduced profits for dairy farmers due to market reforms and the lack of regulatory oversight, and was just one of many examples where the market was seen to have either driven people to desperation or destroyed their moral compass, leading to widespread corruption, fraud, and cheating (Bloomberg News, 2019). These sometimes resulted in massive industrial accidents, but people seemed to be even more outraged by the kinds of indifference that led to viral news stories about bystander apathy as children were run over by wealthy drivers in sports cars (e.g., "Little Yue Yue") and other similar incidents (Notar, 2017). As the economy continues to grow, moral decline continues to dominate Chinese concerns about what they find most worrying about the state of their country (Ipsos, 2019).

The third challenge which the social credit system sought to address was the loosening in the reform era of relatively effective population management mechanisms, especially the household registration system (*hukou*) that placed strict limits on where one could legally reside and thus access schools, health care, and (legal) work, and the file system (*dang'an*) which contained an individual's work and personal history in the form of reports written by teachers, employers, official records, and documents from one's life (including schoolwork and self-criticism). This file followed the person as they moved throughout life, making it, as Jie Yang (2011, p. 508) put it, "the basic socialist database authorized by the communist party, the only legitimate hermeneutic authority and a central domain for the production and reproduction of state power. It was the site at which life and society were fused in politics."

While both the hukou and dang'an continue to exist, they are vestiges of their former selves. Paper files exist outside of the sphere of digital governance and are harder to keep track of, to access, and to preserve, and most workers no longer have their lives shaped by their work units.[4] Mobility has increased tremendously, rendering the constraints of the hukou system onerous, both to individuals and to municipalities who confront millions of "illegal" residents who have flocked to the cities to fill labor needs from construction to factories to service jobs. The city of Shenzhen's attempt to deal with approximately six million unregistered migrants resulted in 2010 in a foreshadowing of reforms to the hukou system and the future social credit system. They included a point scheme to incentivize migrants to apply for Shenzhen hukou, something unthinkable in the Mao era. Migrants could earn points for years worked, skill level, on-time insurance, and tax payments, and volunteering. At the same time, temporary migrants would receive a biometric municipal "green card" which improves access to city services and could be tracked electronically, its data including marital status, family planning, employment, insurance, criminal, credit, purchase and address records (Bach, 2013). In the lead up to the social credit system, figuring out new forms of population management that could direct, sort, track, and make people productive became a policy priority.

These three challenges—increasing access to credit, decreasing mistrust in the market, and using digital technology to replace older systems of population management—became the key components of the social credit system, held together by the overarching policy goal to create a "sincerity culture." The word "sincerity" (*chengxin*诚信, also translatable as social integrity, honesty, or credibility), appears 135 times in the plan. A properly functioning "credit economy" is the precondition for sincerity as the core social value, for example, by presenting the "lawful application of credit information and a credit services system" as necessary for "establishing the idea of a sincerity culture, and carrying forward sincerity and traditional virtues" (Planning Outline, 2014).[5] As Chorzempa, Triolo, and Sacks (2018, p. 11) explain it, "sincerity" functions as a "top level design concept," while "credit" is "the platform" upon which sincerity runs.

3 | EXPERIMENTATION

Rather than the implementation of a top-down system, the past decade can be understood as the effective launch of a massive experiment in which an ever-greater number of actors are involved, from local to provincial governments, over 30 national ministries, and the central bank and its various components, to e-commerce, private credit reporting companies, and bike and ride share programs and hotels (cf. Creemers, 2018; Meissner, 2017; Ohlberg, Ahmed, & Lang, 2017, p. 11). There is great variety and innovation across and between the actors, even as they all feed information into a data backbone (the National Credit Information Sharing Platform, operational as of 2015). Through this mechanism, data can be efficiently shared with anyone who can access the information by looking up a person or business by their unique 18 digit identity number (the Unified Social Credit Number System, which eventually should be given to all legal persons, whether an individual or a corporate entity) (Ohlberg et. al., 2017, p. 10). Any office needing to do a credit check—and this could be a vastly expanding category, since it can include everything from applying for a loan to booking a train ticket or hotel room—would theoretically simply enter the number into a computer hooked up to the database and see the relevant information. This is the crux of the social credit system, although the various experimental components of its complex ecology effectively expand its functions, limits, and contours, as the following section will suggest.[6]

The kind of score, what kind of data it should consider, how it should circulate, and how it can work to shape behavior by providing benchmarks, rewards, and punishments is far from straightforward. Consider what constitutes relevant information. There are multiple types of credit ratings accessible and living together on the national platform, generated from different sources, and most likely arrived at using different methods (Meissner, 2017). The data itself come from such diverse sources as banks and schools, consumer transactions, tracking data from mobile phones, facial recognition, behavioral data from online searches and social media comments (Chen & Cheung, 2017). Different rules govern corporate and individual data, and in both cases only some of the data is public, with access to non-public data presumably restricted to relevant government agencies. In principle there should be a match between the need of the situation, whether extending a loan, hiring for a government position, or accessing services, and the information available (for example, whether a business has been in compliance with government regulations in the past), yet it is difficult to say precisely how the data are integrated.

To make matters more complicated, there is no one-size-fits-all data collection method for sectors as diverse as tourism, health care, insurance, and construction, which are just four of the 14 sectors which have developed social credit plans.[7] The 30 ministries feed the national data sharing platform with over 400 data sets, and each ministry is responsible for designing the systems to collect and share the sets (Meissner, 2017). Similarly, local governments themselves can include numerous agencies, the systems that connect the data from local governments have to be built locally, and what counts as data is often subject to local judgment.

Creating a common scoring method across the vastly different domains from which data is drawn, then, is where experimentation enters in earnest. Because the goal is both to extend financial credit *and* modify behavior (to create law-abiding citizens in a context of widespread corruption, fraud, and non-compliance), the state has to find a way to not only act as a broker of existing information, but also somehow integrate this information into a new product—a measure of one's sincerity that can be used across domains (cf. Stark, 2009, p. 17). To figure out how to do this, the Chinese state has turned to its historically time-honored form of policy experimentation: running many experiments simultaneously, iterating various versions, and learning from trial and error (Heilmann, 2008, 2009). Schneider (2018) shows how this model of governance extends to the digital economy, where the state prioritizes (within political parameters) a diversity of actors working collaboratively to develop new technologies.

These actors include the central government, the People's Bank of China (PBoC), major corporations such as Alibaba and Tencent, and a vast network of smaller actors that interlock with them, such as provincial and municipal governments, app developers, third party hardware and software companies, businesses, and private (or semi-private) companies developing rating, scoring, and risk management tools, all in a symbiotic relationship, watching and learning from each other. “Let a hundred social credit schemes bloom” might be one way to characterize this phase. By encouraging experimentation with social credit across ministries, municipalities, and private enterprises, the government distributes the task of technological and bureaucratic innovation and thus blurs the lines between public and private and between sanctioned and unsanctioned schemes.

These experiments are not only distributed across sectors and organizations, but also spatially, echoing the special economic zone policy that solidified China's export manufacturing rise by testing policies across the country (Bach, 2017). “Pilot” experimental sites for social credit systems include over 43 cities, provinces, plus new Free Trade Zones (Zhang & Zhang, 2016). These diverse localities are given substantial autonomy in trying out versions of ranking and rating systems on enterprises and individuals. Such local experiments have been the source of much of the Western news reporting on the social credit system, such as Rongcheng, in Shandong province where a Social Credit Management Office gives all residents 1,000 points and they have to either keep from losing points (e.g., minus 5 points for traffic infractions) or find ways to balance out the losses by scoring points, for example by volunteering or earning awards for heroic deeds as defined by the local government (one can also go over 1,000) (Daum, 2019a; Mistreanu, 2018). Rongcheng takes the point system one step further than most by giving residents letter grades based on their score bracket, with perks in the form of waived deposits, heating bill discounts, lower interest rates, and preferential treatment, which mimics some of the perks associated with purely commercial reward systems.

Rongcheng's grading system experiment (which is ongoing) is a second iteration after a famously failed earlier version in Suining (Jiangsu province) from 2010, which predated the 2014 Planning Outline, and which citizens and media criticized as too arbitrary, top-down, and invasive (Daum, 2019b; von Blomberg, 2018). Rongcheng, so far, appears not only to have encountered less resistance, but also to have spawned a small copy-cat industry in "microsystems" such as neighborhoods, villages, or hospitals, acting on their own initiative to develop their own scoring systems, as Simina Mistreanu (2018) reports, noting how in a nearby village "the [scoring] criteria boil down to whether you take care of your parents and treat your neighbors nicely." The regional capital Jinan made the news for including dog ownership behavior in a social credit program where, following the third infraction, demerits (e.g., for not having a leash or not cleaning up afterwards) could result in surrendering ownership of the dog (China Pet Market, 2018; Soo, 2018).[8]

4 | EXTENDING CREDIT

What we see in these stories is a kind of purposeful mission creep, where "credit" is extended in both its forms, financial and moral, through corresponding types of behavior (paying bills on time and being nice to your neighbor). The boundaries of the experimental system of governance is blurry—indeed, as Bing Song (2019, p. 34) puts it, "the point of pilot programs is to explore boundaries." Thus, Creemers (2018, p. 11) notes how in Zhejiang province, the provincial government also took it on themselves to "include the sincerity of civil servants" (as measured by local criteria) in their scoring system, even though the central government did not explicitly mandate this. Shanghai launched its "Honest Shanghai" app in 2016, developed by a third party company (Zhengxin Fangsheng) that allows users to check the scores of businesses as well as their own public credit score (Schmitz, 2017). A local court in Jiangsu ordered the telephones of debtors to shame them by forcing callers to first hear the message "The user of the number you've dialed has been listed as someone who is avoiding debt repayments ordered by the Guanyun County People's Court" before it will ring (China Daily, 2017), while Hebei province launched a "deadbeat debtors map" that will tell people via WeChat when they are within 500 meters of someone who defaulted on their loans (Ma, 2019). Beijing is reportedly planning to sort underground train riders into separate lines based on facial recognition matching of their credit scores (Linder, 2019).

Ministries have been experimenting with scoring systems for improving the efficiency of their respective industries in bidding, contracts, and sales, and have accordingly developed attendant reward and punishment schemes. For consumer-facing industries, such as travel, this meant developing reward and punishment systems for users of their services as well as vendors, resulting in some of the more notorious blacklists, such as automatic temporary travel bans for inappropriate behavior on trains or planes. Private sector platforms have developed their own extensive reward schemes that intersect in uncertain ways with the public sector social credit ecosystem. Like the ministries and local governments, private sector companies were encouraged to experiment, and eight companies were chosen to compete for the contract to further develop the platform for data scoring and sharing after 2020, increasing their incentive (though none were selected).[9]

While the companies have insisted on their independence from the government and stress they only concern themselves with rewards not punishments, consumers can be forgiven for not knowing where the boundaries are. For example, in Rongcheng, residents can use the popular online payment system Alipay to pay utility bills, but more to the point, they can use their Alipay-linked customer score known as Sesame Credit (aka *Zhima* Credit, with the motto "Trust Makes it Simple") to gain rewards from the municipal government (Mistreanu, 2018). In other words, a local government recognizes a private company's internal customer loyalty score as a proxy for good behavior, which in turn can boost a resident's municipal social credit score. Apparently this does not work the other way around (i.e., a higher grade in the Rongcheng municipal system does not seem to boost one's Sesame Score), but since Alibaba's scoring algorithms are proprietary, and thus secret, exactly what goes into them is a black box. This blurs the line between private and public rewards and punishments: if someone is blacklisted, say, from staying in luxury hotels or reserving a first class seat on a train, the *private* e-commerce sites Taobao and Tmall (both owned by Alibaba) will not allow that person to purchase luxury goods. Further, while Alibaba will not share customer data with the government without their consent, that consent is mandatory to participate in the online ecosystem.

Similar to how in the US, "creditworthiness predictors use information about the size and strength of a person's social network, exchanged messages, tagged photos, browsing habits, education, searches, and geo-spatial data from mobile phones" (Geslevich & Lev-Aretz, 2016, pp. 343–344) to establish a composite, in the Chinese social credit ecosystem Sesame Credit uses "online behavior and interpersonal relationships" as part of a five-part formula for calculating one's Sesame Score.[10] It seems clear that a person's score is impacted by the actions of others with whom they are associated, even if exactly how is not. This kind of information, combined with all the other aggregate data, could reasonably be used to predict a person's activity, though how accurately and exactly to what end is a matter of speculation. Political control, however, is a high priority for the government and would doubtless play a role in the development of any such predictive system. Accordingly, the interoperability of the social credit system with China's extensive sensor infrastructure of facial recognition and location tracking and their accompanying surveillance affordances is what gives rise to fears of Orwellian implications, especially given the experimental use of technology as part of the large-scale internment and "re-education" of Uighur populations in the Western province of Xinjiang (Vanderklippe, 2018).

5 | ALGORITHMIC AMBIGUITY

The reward/punishment system connects the environments where algorithms are secretly developed (analogous to the laboratory) with the society that this "laboratory" seeks to modify, and from which it draws its data. For the individual operating within this computationally imbricated society, rewards and punishments are generated and enforced by systems that operate on algorithmic principles beyond their knowledge. The algorithms work to effectively adjust the settings on what actions constitute rewardable or punishable behavior, as well as the settings on the technological guardians who watch over an individual's movements in order to insert their data into the system (e.g., the facial recognition software that catches jaywalkers) or enforce an exclusionary punishment (e.g., the software that prevents the purchase of a luxury good).

The blacklist is in some respects little different from the consequences elsewhere of having bad credit and/or a criminal record—difficulty getting loans, government support, buying property, getting a government job or a job that requires a level of trust to insure public safety, and so on.[11] The particular Chinese twist on this is to add blacklists for socially high status activities, such as consumption at luxury hotels, restaurants, vacations, nightclubs, golf courses, home renovation, or buying cars, in addition to the more well-known bans on high speed train travel or flying (Daum, 2017, 2018; Supreme People's Court, 2015). While there are rules about blacklists (e.g., there must be evidence of non-compliance), one can easily imagine a wide variety of localized variations. Similar to the mission creep described in the last section, blacklists have metastasized into what we might call blacklist fever, where infractions of administrative rules (not only court orders) become the reason to place someone on a blacklist, and all kinds of organizations can institute blacklists (Creemers, 2018).

This has not seemed to impact apparent high levels of popular support for the social credit system, especially among those who stand the most to gain from the rewards (e.g., well-off, educated, urban males), but also more generally from those who see the system as a good faith effort to improve people's quality of life (Kostka, 2019; Ruengrangskul & Wenze, 2019). Yet there is no way to know exactly how all the various transgressions one commits will add up. While people are supposed to always be notified in advance of being put on a blacklist, and given the chance to appeal or to remove themselves through compliance, this sometimes does not seem to happen. The oft-repeated dictum from the Planning Outline (2014) that whoever violates the rules somewhere will be restricted everywhere seems to impose maximum inconvenience and reputational damage without, it seems, much regard for proportionality (though most blacklists are limited in time). And since the government does not (yet) keep one score that aggregates all the various scores from this city, that ministry, this company, or that association, there are stories of blacklisted persons finding ways around bans (by using different kinds of IDs, for example, which rely on different databases, or did not both check the blacklist). Furthermore, even though public data generally has a sunset clause of 5 years, there is little control over how third parties might harvest or re-use disclosed data (Chen & Cheung, 2017, p. 26), never mind malicious operators who might hack into the system.

Thus, while in principle black and red (trustworthy) lists are nationally accessible by anyone with access to the sharing platform database and the 18 digit number, it is not clear which information will make it all the way up to this list, or what the threshold is for being put on a blacklist. Is one late notice for an overdue bill enough? Ten notices? How many overdue library books can be a problem? As long as you pay the fine for jaywalking you will only lose some municipal points, so that should not put you on a blacklist, but it is hard to shake the feeling that no good will come from having a D rating in Rongcheng. The blacklists are only supposed to be for those who violate the law or rules, but what goes into a redlist?

The answer, it turns out, is intriguingly unclear. Englemann, Chen, Fischer, Kao, and Grossklags (2019) have rigorously identified an interesting information asymmetry between government blacklists and redlists. If blacklists are supposed to discourage "bad" behavior, and redlists to encourage "good" behavior, it would be important for citizens to know what constitutes good or bad in the social credit system, since, as they put it (p. 9), "to negotiate a norm one must have the necessary epistemic resources to do so." Yet they found that there is far more transparency in the social credit system about what constitutes bad behavior than good. Blacklist categories cite legal or regulatory chapter and verse to justify the punishment, yet the justifications for appearing on a government redlist were vague: people received honorable titles and "decorations" with no clear explanation about the criteria for these. Furthermore, the outcome was significantly different—blacklists resulted in material punishments, while redlists resulted in symbolic reputational rewards.[12] Fear of being on a blacklist was a clear motivating factor for compliance, yet they found no evidence that being on a redlist was motivation for any behavior.

Their explanation for this information asymmetry (2019, pp. 9–10) relates directly to the question of how strategic ambiguity and policy experimentation work together to produce a total test environment. If the social credit system were entirely non-transparent and nobody knew the reasons for being on a black or red list, this would defeat the stated purpose of the system to encourage trustworthy behavior, as learning from it would be impossible. The other extreme, importantly, is equally problematic: if the system were completely transparent, then, it would open itself to gaming at a large scale and norm compliance could become more like market transactions, also defeating the stated purpose of the system to reconcile morality and the market. Maintaining the system as semi-transparent allows, they conclude (p. 10), for the social credit system to guard against the "transformation of moral behavior into market transactions," a risk that appears as the unwanted but seemingly inevitable by-product of a scoring system that adapts market-based governance techniques.

6 | TESTING DIALECTICS

By fulfilling one of the stated aims of the socialist market economy—to serve as a stage for a socialist future yet to come—the social credit system tries to dialectically resolve two of capitalism's core tensions: between the market as both an enabling and a corrupting influence on society, and between the market and the state as two separate spheres in conflict with each other. This article has argued that it does this in large part through the cultivation of ambiguity about what, exactly, should be considered part of *the* social credit system—how exactly, data are collected, shared, and used. This allows plausible deniability to work as a buffer in case of unpopular or unsuccessful developments, enabling the central government to come in and shut down or revamp experiments (such as Suining's grading system or Tencent's credit scoring service) while also saving face. This ambiguity insures that, although from 2020 onwards everyone will be subject to a "credit check" and scores will be made public, nobody will be able to say for certain what is being tracked, collected, shared, or considered in a score. This makes the social credit system similar in some ways to the old dang'an file system where the files, as Yang (2011, p. 512) writes, “became an object of constant speculation, rumor, and fear, both hallucinatory and spectral.” The rumor function is thus productive, but also a balancing act: If the emerging system is too inconsistent it loses the very trust and effectiveness that it claims to be calling into being. Yet if it is too rigid, it can also lose its legitimacy and become too easily gamed or manipulated and subject to the very corruption it is supposed to forestall. Thus it has to be a dynamic, adaptive system, not only to make the system resilient, but also to allow it to take the form of a total test environment where everyone is subject to a battery of tests every time they go online, move in certain locations, pay bills, attend class, respond to administrative issues, check into a hotel, apply for a job, walk dogs, visit parents, and so on.

A degree of what we might call strategic ambiguity allows the social credit system to present itself to multiple audiences without calling undue attention to its potential contradictions.[13] In this concluding section I suggest that the institutionalization of this ambiguity is made possible by the affordances of the digital platforms upon which the system rests. This is because platforms actively shape, rather than merely facilitate, the relationship between economic, moral, and political domains (Andersson Schwarz, 2017, van Dijk, 2013), making them integral to the kind of computational infrastructures that make social life observable to big data (Marres, 2017). For this reason, the notion of the platform has increasingly occupied a central place in explaining the digital economy, but it also has proved to be an elusive and somewhat mystical phenomenon (de Kloet, Poell, Guohua, & Yiu, 2019; Gillespie, 2010; van Dijck, Poell, & de Waal, 2018). In the context of digital economies, a platform usually refers to the combination of software and hardware needed to run specific applications (e.g. the Apple operating system). Its importance to the changing nature of capitalism is the way in which these “platforms” enable the direct extraction of rents from information generated when users and providers enter into exchanges of various sorts on the respective platform. If information is the newest fictional commodity (in Karl Polanyi's sense), platforms are to the extraction of surplus value in the digital economy what factories were in the industrial economy (cf. Cohen, 2017, p. 135; Andersson Schwarz, 2017, p. 384). More than facilitating exchange, though, platforms have an almost alchemical function as the locus where social relations are translated into computer code only to be translated back into social codes, thus effectively working, as Andersson Schwarz (2017, p. 380) puts it, to refashion social order and “truth and knowledge.”

This refashioning happens in an ecology where platforms beget new platforms in an ongoing cycle. As Langley and Leyshon (2017, p. 19) put it, platforms are actants that “curate connectivity” by actively mobilizing participation. The more people use platforms, the more data are generated, which in turn generates new platforms to perform new functions “situated in ever-widening ecologies of mutual interplay, co-dependence, and productivity” (Andersson Schwarz, 2017, p. 383). The irony is that while this proliferation of platforms has a horizontal, seemingly democratic element to it, Andersson Schwarz shows how it unfolds within an ever-more monopolistic environment (Google, Amazon, Facebook, Apple in the US-dominated digital ecologies; Tencent, Alibaba, and Baidu in the Chinese). This charges platforms with a “paradoxical tension between the logic of generative and democratic innovations and the logic of infrastructural control” (Eaton et al. in Andersson Schwarz, 2017, pp. 278–279). This paradox seems to apply even in the context of a powerful and controlling authoritarian state such as China, where, as de Kloet et al. (2019, pp. 253–254) argue, “platformization also opens up or affords possibilities for empowerment, labor and play,” complicating “the idea of the alleged totalizing control of platform algorithms” through “the simultaneous dynamics of exploitation and empowerment.”

In this way, because the social credit system explicitly seeks to not only channel behavior into compliant forms, but also to actively modify society into a new “sincerity culture,” it might open up new ways for understanding how platforms enable governance, including the conditions for ambiguity and modification of the system itself. For example, Caprotti and Liu (2020) show how in China the datafication resulting from the platform economy has placed data “at the centre of state-corporate, and state-market relations” making data “a key asset that requires negotiation and control.” This, they argue, allows platform economies to recalibrate the relationship between the state, corporations, and citizens and, as a result, transform how cities are experienced, regulated, governed, and measured.

As a complex platform ecology, then, the social credit system could be seen as connecting and transforming statist and market logics, aiming to reconcile the residual fear of the market as inherently amoral with the imperative to consume, invest, and grow the economy.[14] For states to “see” they need to delineate or create disciplinary sites where bodies could be administratively acted upon (schools, prisons, etc.), while for markets to “see” they need to decipher sites of barter and exchange (marketplaces, stores, trading floors, etc.). If the platform, as Julia Cohen argues (2017, pp. 135–136), replaces and rematerializes markets rather than merely entering or expanding them, perhaps the platform becomes the privileged site for a new hybrid of state and market in the axial shift from disciplinary to control societies (Deleuze, 1992), something that the social credit system seems to intuitively grasp.

7 | CONCLUSION

What might the social credit system tell us about the form and function of social experiments in the digital age? As a system it contains multiple tests: of loyalty (to the state, but also to specific institutions and to each other), of trustworthiness, of obedience, of compliance, of one's ability to modify behavior, as well as technical tests of the various infrastructures and platforms upon which the system runs, and tests of algorithmic abilities and programming skills. This is the inverse of the kind of controlled experiment championed by Chapin in the early 20th century, and even from the kind of proxy tests analyzed by Pinch at the century's end: it is not about observing the results of a test to see if the experiment is successful or not, but about modifying behavior and using the feedback from these modifications to adjust and calibrate the system as a whole (Marres and Stark, this issue). And in this case the system as a whole is the party-controlled state, which is experimenting with how best to entice, compel, and convince people to follow its ideas of proper behavior across a diversity of domains in a dynamically evolving social and economic environment.

This kind of total test environment thus requires not a method for calculating probability of outcomes, but an architecture for modifying behavior under conditions of uncertainty. The social credit system platform is arguably such an architecture. By bringing elements of platform capitalism together with the legacies of socialist model making, it strategically uses ambiguity to suture the gaze of the state and market. This allows, on the one hand, for the Chinese government to maintain power in the face of significant centrifugal forces by bringing contradictions inside its purview rather than externalizing them (as it did by allowing capitalists to become members of the communist party), to extend its biopower through datafication and surveillance, and to benefit from platform capitalism's depoliticizing effects (cf. Fourcade & Healy, 2017). On the other hand, despite the goal of control, the same ambiguity also pushes and pulls at the limits and the contours of the system, working the edges of platforms around their “complex interplay between datafication and affordance, between money and meaning, and between surveillance and security” (de Kloet et al. 2019, p. 253). The result is what seems like a permanently beta platform system, where automated systems of monitoring, analysis, and feedback create a scoring infrastructure that bring the moral and the economic together, but also leaves them in an ambiguous relationship, as Englemann, Chen, Fischer, Kao, and Grossklags (2019) indicated for the punishment/reward system, de Kloet et al. (2019) for digital infrastructures, and Chen and Cheung (2017, p. 3) for legal “grey spaces” that make China the “ideal laboratory” for experiments with social credit. This raises the question in this experimental context about who is testing whom.

Perched at the intersection of the state's disciplinary power and the market's modular muscle, the social credit system aspires to be the ultimate dialectical technology. It seeks to hold contradictions in play without collapsing under their weight and, if successful, to be transparent yet opaque, economic and moral, designed by humans yet automated (so as to reduce the possibilities for corruption), able to make moral acts and financial actions commensurable, to prevent “discredited” persons from dragging down the system while providing incentives for them to become credible and creditable. This is a tall order and hardly preordained, but as an ongoing experiment it opens new, if sometimes unsettling, research horizons for understanding how ecologies of testing are transforming the social systems in which they are embedded.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

NOTES

[1] The system includes, as of mid-2019, 990 million individuals and 25.91 million enterprises (Xinhua, 2019).

[2] In this article I focus primarily on the emergence of the social credit system as relates to individuals. The system also applies to corporations, and applies to any foreign company doing business in China. The corporate dimension thus moves the scope of the system to a global level in intriguing ways, calling for a separate analytical treatment.

[3] Of course, forms of this critique are endemic to capitalist societies as well.

[4] Though see Chen and Cheung (2017) for an interesting discussion of how the legal parameters of one‘s *dang‘an* file today (which they translate as a "personal archive") intersect with privacy claims surrounding access to one‘s social credit data.

[5] On the prehistory of sincerity in Chinese culture see Mueller and D'Ambrosio (2017).

[6] For detailed overviews of how the social credit system functions, including expanded technical discussion of the details mentioned in this article, see Trivium China (2019), Creemers (2018), Meissner (2017), Ohlberg et al. (2017). See also the chart created by Daum (2019b) and related materials on his China Law Translate project website.

[7] The 14 sectors and industries that were mandated to establish social credit plans by 2017 are Agriculture, Automobiles, Construction, E-Commerce, Energy, Food and Drugs, Health, Insurance, Logistics, Media, Sports, Steel, Tourism, and Transportation. See Meissner (2017, p. 5).

[8] It should be noted that pet dogs played a role in Maoist anti-bourgeois discourse as a symbol of decadence, and were hunted and killed during the Cultural Revolution.

[9] The eight were Sesame Credit (Ant Financial/Alibaba), Tencent Credit, Koala Credit, Pengyuan Credit, Sinoway Credit, Qianhai Credit, China Chengxin (Sincerity) Credit, and Intellicredit. Ultimately the government chose none of them. According to reports they were too focused on their own business interests and faced data quality problems, though one wonders whether the companies did not also resist sharing everything with the government. Instead, the government asked the companies to collaborate together in developing the infrastructure for a national level private-public credit bureau called Baihang Credit (Koetse, 2018), something which may yet bring the government one step closer to an integrated, centralized, credit score. These companies had been given special licenses to develop credit rating systems, and after the decision was made not to select a winner, the licenses for these companies were not revoked, but allowed to continue in a kind of legal gray space.

[10] The other four cover conventional credit histories, contractual compliance, verifiable information, and, most subjectively, behavior during consumption transactions.

[11] At first, punishments were restricted to people who did not comply with a court order (e.g., failing to pay a fine or compensation ordered by a court, or operating a business after the license has been revoked). Later punishments were expanded for various kinds of misbehavior, such as being disruptive on trains.

[12] Note that we are talking about people's appearance on government lists, not rewards within private sector or municipal point systems.

[13] It might be instructive here to think about ambiguity as a property of innovation across organizational forms, from states to firms (Stark, 2001) and NGOs (Bach and Stark, 2002).

[14] De Kloet et al. (2019) note how the Chinese term for platform "evokes images of a podium, a stage, or an elevated structure that invites people to gather, to act, to work, and to express opinions and ideas" which is similar to Gillespie's (2010) discussion of how the different meanings of the term work together to produce its power by bringing "discourses into alignment," though perhaps their real power comes from not aligning them too perfectly.

REFERENCES

Andersson Schwarz, J. (2017). Platform logic: An interdisciplinary approach to the platform-based economy. *Policy & Internet*, *9*(4), 374–394.

Bach, J. (2013, March). Shenzhen: Constructing the city, reconstructing subjects. *Open Democracy.*

Bach, J. (2017). Shenzhen: From exception to rule. In M. O'Donnell, W. Wong, & J. Bach (Eds.), *Learning from Shenzhen: China's post-Mao experiment from special zone to model city* (pp. 23–28). Chicago, IL: University of Chicago Press.

Bach, J., & Stark, D. (2002). Innovative ambiguities: NGOs' use of interactive technology in Eastern Europe. *Studies in Comparative and International Development*, *37*(2), 3–23.

Bloomberg News. (2019, January 22). China's lethal milk scandal reverberates a decade later.

Caprotti, F., & Liu, D. (2020). Emerging platform urbanism in China: Reconfigurations of data, citizenship and materialities. *Technological Forecasting and Social Change*, *151*.

Chapin, S. (1917). The experimental method and sociology. *The Scientific Monthly*, *4*(2), 133–144.

Chen, Y. C., & Cheung, A. S. Y. (2017). The transparent self under big data profiling: Privacy and Chinese legislation on the social credit system. *Journal of Comparative Law*, *12*(2), 356–378.

China Daily. (2013, February 18). Trust among Chinese "drops to record low".

China Daily. (2017, July 26). Court in Jiangsu authorizes "ringtone shaming" for defaulters.

China Pet Market. (2018, October 30). Social credit system for pet owners.

Chorzempa, M., Triolo, P., & Sacks, S. (2018, June 14). *China's social credit system: A mark of progress or a threat to privacy?* Washington, DC: Peterson Institute for Economics Policy Brief. Retrieved from https://www.piie.com/publications/policy-briefs/chinas-social-credit-system-mark-progress-or-threat-privacy

Cohen, J. (2017). Law for the platform economy. *UC Davis Law Review*, *51*, 133–204.

Creemers, R. (2018, May). China's social credit system: An evolving practice of control. *Social Science Research Network*.

Daum, J. (2017). Map of the 2014–2020 social credit plan. *China Law Translate*.

Daum, J. (2018, March 30). The redlists are coming! The blacklists are coming! *China Law Translate*.

Daum, J. (2019a, November 29). Getting rongcheng right. *China Law Translate*.

Daum, J. (2019b, February 17). Social Credit Joint-Enforcement MOU breakdown. *China Law Translate*.

de Kloet, J., Poell, T., Guohua, Z., & Yiu, F. C. (2019). The platformization of Chinese Society: Infrastructure, governance, and practice. *Chinese Journal of Communication*, *12*(3), 249–256.

Deleuze, G. (1992). Postscript on the societies of control. *October*, *59*, 3–7.

Englemann, S., Chen, M., Fischer, F., Kao, C. Y., & Grossklags, J. (2019). Clear sanctions, vague rewards: How China's social credit system currently defines "good" and "bad" behavior. In *Conference on Fairness, Accountability, and Transparency (FAT* '19)*, January 29–31, Atlanta, GA.

Fourcade, M., & Healy, K. (2017). Seeing like a market. *Socio-Economic Review*, *15*(1), 9–29.

Geslevich, P. N., & Lev-Aretz, Y. (2016). On social credit and the right to be unnetworked. *Columbia Business Law Review*, *2016*, 339–425.

Gillespie, T. (2010). The politics of "platforms". *New Media & Society*, *12*(3), 347–364.

Heilmann, S. (2008, January). From local experiments to national policy: The origins of China's distinctive policy process. *The China Journal*, *59*, 1–30.

Heilmann, S. (2009). Maximum tinkering under uncertainty: Unorthodox lessons from China. *Modern China*, *35*(4), 450–462.

Ipsos. (2019). *What Worries the World* (p. 17). Paris: Ipsos Report.

Koetse, M. (2018, June 10). Baihang and the eight personal credit programmes: A credit leap forward. *What's on Weibo*.

Kostka, G. (2019). China's social credit systems and public opinion: Explaining high levels of approval. *New Media and Society*, *21*(7), 1565–1593.

Langley, P., & Leyshon, A. (2017). Platform capitalism: The intermediation and capitalization of digital economic circulation. *Finance and Society*, *3*(1), 11–31.

Linder, A. (2019, October 31). Beijing to sort metro riders. *Shanghaist*.

Loubere, N. (2017). China's internet finance boom and tyrannies of inclusion. *China Perspectives*, *4*(112), 9–18.

Ma, A. (2019, January 22). China app shows map of people in debt for social credit system. *Business Insider*.

Marres, N. (2017). *Digital sociology*. Cambridge: Polity Press.

Meissner, M. (2017, May). China's social credit system: A big-data enabled approach to market regulation with broad implications for doing business in China. *China Monitor*.

Mistreanu, S. (2018, April 3). Life inside China's social credit laboratory. *Foreign Policy*.

Mueller, H.-G., & D'Ambrosio, P. (2017). *Genuine pretending: On the philosophy of the Zhuangzi*. New York, NY: Columbia University Press.

Notar, B. (2017). "My dad is Li Gang!" or seeing the state: Transgressive mobility, collective visibility, and playful corruption in contemporary urban China. *Asian Anthropology*, *16*(1), 35–53.

Ohlberg, M., Ahmed, S., & Lang, B. (2017, December). Central planning, local experiments. The complex implementation of China's social credit system. *China Monitor*.

Ortiz, H. (2013). Financial value: Economic, moral, political, global. *HAU: Journal of Ethnographic Theory*, *3*(1), 64–79.

Pinch, T. (1993). Testing—One two three testing! Towards a sociology of testing. *Science, Technology & Human Values*, *18*(1), 25–41.

Planning Outline for the Construction of a Social Credit System (2014–2020). (2014, June 14). State Council, Peoples Republic of China. Translation at https://chinacopyrightandmedia.wordpress.com/2014/06/14/planning-outline-for-the-construction-of-a-social-credit-system-2014-2020/

Ruengrangskul, N., & Wenze, M. F. (2019). China's social credit system as a stimulant of donation behavior: Assessment of student opinions. *International Journal of Organizational Innovation*, *11*(4), 165–178.

Schmitz, R. (2017, January). What's your 'public credit score? The Shanghai government can tell you. *npr.org*.

Schneider, F. (2018). *China's digital nationalism*. New York, NY: Oxford University Press.

Scott, J. (1999). *Seeing like a state: How certain schemes to improve the human condition have failed*. New Haven, CT: Yale University Press.

Song, B. (2019). The west may be wrong about China's social credit system. *NPQ: New Perspectives Quarterly*, *36*(1), 33–34.

Soo, Z. (2018, October 28). Forget to leash your dog? Chinese credit scoring system for owners means you could lose your pet. *South China Morning Post*.

Stark, D. (2001). Ambiguous assets for uncertain environments: Heterarchy in postsocialist firms. In P. DiMaggio, (Ed.), *The twentieth century firm* (pp. 69–104). Princeton, NJ: Princeton University Press.

Stark, D. (2009). *The sense of dissonance: Accounts of worth in economic life*. Princeton, NJ: Princeton University Press.

Steinhardt, C. (2012). How is high trust in China possible? Comparing the origins of generalized trust in three Chinese societies. *Political Studies*, *60*(2), 434–454.

Supreme People's Court. (2015). Several provisions on restricting high-spending and related spending by persons subject to enforcement. *China Law Translate*.

Tao, L. (2018, March 27). Jaywalkers under surveillance in Shenzhen soon to be punished via text messages. *South China Morning Post*.

Trivium China. (2019, September 23). Understanding China's social credit system. *Social Credit Watch*.

van Dijck, J. (2013). *The culture of connectivity: A critical history of social media*. New York, NY: Oxford University Press.

van Dijck, J., Poell, T., & de Waal, M. (2018). *The platform society: Public values in a connective world*. New York, NY: Oxford University Press.

Vanderklippe, N. (2018, February 27). China using big data to detain people before crime is committed. *The Globe and Mail*.

Von Blomberg, M. (2018). The social credit system and China's rule of law. *Mapping China Journal*, 2, 77–113.

Xinhua. (2019, June 14). China boasts world's largest social credit system: Official.

Yang, J. (2011). The politics of the Dang'An: Spectralization, spatialization, and neoliberal governmentality in China. *Anthropological Quarterly*, *84*(2), 507–533.

Zhang, K., & Zhang, F. (2016). Report on the construction of the social credit system in China's special economic zones. In T. Tao, & Y. Yuan (Eds.), *Annual report on the development of China's special economic zones* (pp. 153–171). Singapore: Springer.

How to cite this article: Bach J. The red and the black: China's social credit experiment as a total test environment. *Br J Sociol*. 2020;71:489–502. https://doi.org/10.1111/1468-4446.12748

Received: 12 March 2019 | Revised: 14 November 2019 | Accepted: 15 November 2019

DOI: 10.1111/1468-4446.12718

SPECIAL ISSUE

WILEY

Prototyping public friction: Exploring the political effects of design testing in urban space

Martín Tironi

School of Design, Pontificia Universidad Católica de Chile, Santiago, Chile

Correspondence
Martín Tironi, School of Design, Pontificia Universidad Católica de Chile, Av. Libertador Bernardo O'Higgins 340, Santiago, Chile.
Email: martin.tironi@uc.cl

Abstract

The use of prototypes as testing instruments has become a common strategy in the innovation of services and products and increasingly in the implementation of "smart" urban policies through living labs or pilots. As a technique for validating hypotheses about the future performance of products or policies, prototyping is based on the idea of generating original knowledge through the failures produced during the testing process. Through the study of an experimentation and prototyping project developed in Santiago de Chile called "Shared Streets for a Low-Carbon District," I analyse the technique of prototyping as a political device that can make visible (or invisible) certain entities and issues, determining what the experimental entities can *do* and say. I will show how the technique of prototyping defines modes of participation, what is visible and thinkable, what can be spoken and what is unspeakable. In this sense, I examine two ambivalent capacities of prototyping: as a mechanism of management and enrolment that seeks to prescribe normativities (problem-validating prototype) and as an event that can make frictions tangible, articulating matters of concern and ways to open up alternative scenarios (problem-making prototype).

KEYWORDS

curatorial interventions, public frictions, smart city, urban laboratory, urban prototyping

1 | INTRODUCTION

Over the past several years, a special interest has developed among public and private agencies in understanding and governing social processes through experimentation and testing. The proliferation of experimental practice has even led some to refer to ours as an experimental society (Gross, 2016) and to "collective experiments" (Latour, 2001). One of the distinctive aspects of in-laboratory experimentation is the drawing of boundaries between a controlled interior and an uncontrolled exterior (Guggenheim, 2012), but the idea of confined space has begun to be diluted through the multiple uses and meanings that the notion of experiment has opened up in innovation, governance, and public politics. Though they have done so in regard to different concerns, recent work in Science and Technology Studies (STS) has explored how experimentation and experimentalism ceased to be exclusively located in the sanitized spaces of scientific laboratories and infiltrated spaces of social life to varying degrees (Callon, Lascoumes, & Barthe, 2009; Corsín Jiménez, 2014; Gross & Krohn, 2005; Lezaun & Millo, 2006; Marres, 2012).

The language of experimentation has left the laboratory to become a strategy for testing innovations in real contexts (Laurent & Tironi, 2015; Marres, 2012). It also has become an alternative method of governance (Caprotti & Cowley, 2017; Evans, Karvonen, & Raven, 2016; Laurent, 2017) based on the coordination of diverse interests in which local communities are called on to co-design solutions (Voytenko, McCormick, Evans, & Schliwa, 2016). Moreover, the idea that everything should be "tested" has become one of the maxims of the culture of innovation, and is mobilized in spaces of technological design, services design, urban spaces, computing, and in the world of business and management.

In this article, however, my purpose is not to trace the full range of spaces, practices, and discourses that are mobilized in these *experimental geographies* (Kullman, 2013). Rather, I will focus on a particular space in which the terms experimentation and testing have been taken up in an especially forceful way: I am referring to the space of smart urbanism. Through terms such as "urban labs," "urban experiments," "pilot projects," "hackathons," "future labs," and "urban prototyping," the promoters of Smart Cities projects recognize in the language and practice of experimentation alternatives to traditional forms of handling and managing social and environmental problems, and seem to be becoming a privileged strategy for grappling with increasingly complex environments (Evans et al., 2016). Many stakeholders are interested in incorporating a rhetoric of "urban experiments" into the management of organizations and the city (Laurent & Pontille, 2018). One distinctive element of the experiments undertaken by civic authorities and urban agencies is that the design, dramaturgy and *mise en scène* of these interventions are just as important as the results of the experience or even more so. Furthermore, the logic of testing that proliferates in these smart initiatives tends to feed off of promises of "learning by doing" and "bottom-up innovation" where the goal is to provide experiential, living, scalable results and conditions that will allow experts to work with common citizens to define solutions to the problems in question (Kimbell & Bailey, 2017).

This article contributes to the sociological understanding of urban experiments by offering an analysis of the deployment of a specific experimental technique in this context: prototyping. This procedure consists of creating an experimental, preliminary (*proto*) version (or type) of a certain product, service, or project in order to test certain qualities, characteristics, or reactions in the "real" environment before it is finalized (Kimbell & Bailey, 2017; Sanders & Stappers, 2014; Suchman, Trigg & Blomberg, 2002). Simply put, prototyping can be defined as an effort to materialize and verify an idea through an object, event, or experience. While prototypes can take on many forms (models, sketches, drawings, interventions, storytelling, etc.), a common characteristic is materialization—the principle of making tangible ideas (hypotheses, suppositions, intentions, concepts, etc.) through the crafting of provisional objects which enable experiential and empirical verification. From this perspective, the use of prototypes does not only serve to validate an idea quickly and economically. It also facilitates the identification of problems, needs, and possible opportunities while allowing for dynamics of elicitation and anticipation among the various stakeholders involved (Suchman et al., 2002).

In an effort to explore these elements, I will analyse the case of "Shared Streets for a Low-Carbon District," an urban experiment conducted in downtown Santiago de Chile to test the attitude towards and possible adoption of more sustainable and smart habits in the city. Explicitly inspired by a logic of "living labs" and using the slogan "Live the experiment of a new city," the project sought to develop a "laboratory" of prototypes in which to assess Santiago residents' willingness to adopt smarter and more ecological habits, and the possible demands for infrastructures that would be appropriate for sustainable modes of mobility (walking and cycling).

But rather than trying to define what a prototype is, in this article I will use this empircal case to explore what prototyping *does* to social processes and the issues that it is supposed to solve. The growing adoption of experiments involving prototypes as an urban intervention strategy must be examined in terms of what it encapsulates and generates, what it makes invisible and reveals. Different authors have emphasized how prototyping can enable open and iterative processes of deliberation that deepen formal democratic logics (Corsín Jiménez, 2014; Hillgren, Seravalli, & Emilson, 2011; Kimbell & Bailey, 2017; Tironi, 2018), but little attention has been paid to prototypes' capacity to contain or discipline certain realities. In this article, I seek to interrogate the "democratizing" and "open" dimension of the prototype, showing that its capacities to elicit new realities depend on curatorial interventions and narrative options. I will show the ambivalent activation enabled by prototypes. On the one hand, the intervention mobilizes a robust logic of "problem validating prototype," enfolding and configuring certain visions of reality. On the other hand, the intervention helps to unfold new realities and relationships, activating what I will call a "problem-making prototype."

In this sense, I analyse the technique of prototyping in urban space as a political device that can make visible or invisible certain entities and issues, determining what the experimental entities can *do* and *say*. I seek to show how the prototyping technique can manage what Rancière (2000) called *partage du sensible*, determining the degrees of epistemic and ontological openness of experimental settings. The purpose is to critically interrogate the practices of experimentation through prototyping in the context of policy-making. I will argue that rather than simply serving as a means to define a future original, prototyping is a political process in at least two ways: (a) as material and narrative technology for demonstrating and justifying certain options, and (b) as a mechanism of explicitation that can make visible the unexpected friction. Following Tsing (2011), frictions are defined as uncomfortable, unequal, and unstable moments that can "make worlds" and generate inventive forms of interaction.

The case is based on an ethnographic study that consisted of in situ observations and interviews conducted during the various phases of the intervention and prototyping processes. Furthermore, eight in-depth interviews were conducted during and after the experiment with key stakeholders from the ONG Ciudad Emergente (Emergent City, or CE) as well as organizations that collaborated on the project including the Municipality of Santiago and Fab Lab Santiago. The article is organized as follows: First, I review recent discussions of urban experimentation and prototyping techniques and the notion of *partage du sensible* developed by Rancière to understand the opportunities afforded by prototyping. Next, I present the case of "Shared Streets for a Low-Carbon District," analysing the testing and prototyping operations developed by the people who promoted the intervention and the curatorial logics of the publication and circulation of its results. Finally, I reflect on the limits and capacities of prototyping logics in the context of smart urbanism experiments. I argue that urban prototyping situations like the one I analyse in which a series of complex socio-spatial conditions and reactions are tested should not only be approached as laboratories of validation and foreshadowing, but also as privileged spaces for deploying forms of open exploration of frictions and the unknown.

2 | SMART EXPERIMENTATION IN URBAN SPACE

The language of smartness is reorganizing discussions of the city (Marvin, Luque-Ayala & McFarlane, 2016) and impacting how we think about protocols and techniques of intervening in it. The Smart City paradigm has transformed experimental protocols into a new form of urban governance, redefining repertoires of contemporary urban development (Evans et al., 2016; Gabrys, 2014; Halpern, LeCavalier, Calvillo, & Pietsch, 2013). While the idea of city as testing laboratory does not emerge with the smart imaginary and has its roots in the utopic and modernist development experiments of new cities in the 1960s (Caprotti & Cowley, 2017; Gieryn, 2006; Guggenheim, 2012), the term experimentation is being deployed particularly strongly by smart urbanism proponents and in various participatory urbanism movements (De Lange & De Waal, 2013; Gabrys, 2016). Based on the assessment of the need to develop new ways to confront the future challenges of large cities, the rhetoric of experimentation provides an opportunity to promote and direct these changes in order to move towards more sustainable and intelligent cities. Smart solutions proponents tend to present their services, products or solutions as experimental activities or pilots open to learning with the "real city." They also materialize a concept of the city as an entity that can be intervened in, tested, and calculated under the logic of a controlled laboratory (Karvonen & van Heur, 2014; Marres, 2018).

This experimental regime of the Smart City covers extremely varied topics including tests for promoting more sustainable urban environments linked to electromobility or the decarbonization of the city (Evans & Karvonen, 2014), environmental pollution monitoring initiatives (Gabrys, 2016), urban cycling routes (Tironi & Valderrama, 2018) and transition policies for addressing climate change (Bulkeley & Castán Broto, 2013). In an effort to replicate the European and US trend of moving from fossil fuel cities to sustainable ones, China and India are promoting extremely ambitious policies based on the development of Smart City pilot projects and experiments in which governments and transnational corporations work together (Karvonen, Cugurullo, & Caprotti, 2018). Over the past few years, Latin America also has become a showcase for experiments and interventions in the field of services and policies oriented towards promoting more intelligent and sustainable cities. Projects that follow the living lab and pilot projects with unprecedented enthusiasm have proliferated (Broto & Calvet, 2016; Tironi & Criado, 2015).

These initiatives—many of which are spectacular from the perspective of their staging—are not conceived of as closed solutions, but as "learning" laboratories in which there is an effort to show the need for certain changes and the possibility of anticipating new scenarios. Some authors have called this mode of intervention and urban governance in which cities—and those who live in them—are subjected to controlled and provisional test "test-bed urbanism" (Halpern et al., 2013). Marres (2012) has explored these practices of the production of knowledge beyond the laboratory (in domestic spaces, for example) as "living experiments," emphasizing that they are sites that exhibit certain socio-technical reconfigurations, introducing new relationships and entities.

These urban experiments are characterized by involving citizens or users in order to demonstrate the openness of this type of project to feedback from "normal" citizens (Harrison & Donnelly, 2011). This celebrated aspect of such efforts tends to invoke the forms of production of knowledge of Smart City experiments as added value given that their forms of knowledge production are no longer based on abstract models that are removed from "reality" (Tironi & Valderrama, 2018). Furthermore, urban experiments have been analysed as demonstration exercises undertaken to thematize and make public certain innovations through the participation of certain audiences (Laurent & Tironi, 2015; Marres, 2012). Experimentation has emerged as a tool for facilitating certain changes that "push" actions of transition (Bulkeley & Castán Broto, 2013). As such, smart projects present the use of the term "experimentation" as part of an attempt to innovate in forms of government and planning. For example, one distinctive characteristic of this form of "urban innovation" is the interest in connecting private companies and public institutions to everyday urban experiences (Evans et al., 2016). This is related to the emergence of the "open government" agenda, a perspective that is increasingly being adopted by local governments to promote the use of new methods to engage stakeholders, gather data, and experiment before implementing policy (Kimbell & Bailey, 2017).

3 | PROTOTYPING AS A TECHNIQUE FOR MAKING THINGS VISIBLE

In this context of proliferation of "smart experiments," prototyping techniques have become particularly important tools for developing these interventions in the urban space. As Alberto Corsín Jiménez explains, the languages of prototyping and open-endedness, of provisionality and experimentation, are taking hold as models for cultural practice in the city (Corsín Jiménez, 2014, p. 382).

The purpose of prototyping is to test the behavior of certain elements prior to stabilizing a final product or model (Sanders & Stappers, 2014). The creation of prototypes is also a strategy for recognizing users' needs. As such, this artefact generally promotes "a material deliberation with concepts and ideas that might not otherwise be apparent" (Forlano, 2016). The use of this testing strategy is thus often aimed at reducing risks of failure of certain services or products, turning the prototype into a learning and forecasting strategy that is increasingly valued by public administration entities (Galey & Ruecker, 2010; Kimbell & Bailey, 2017).

From a sensibility that is close to science and technology studies (STS), the notion of prototyping finds various meanings that go beyond being merely a validation technique. Thus, the prototype has been understood as an ongoing process that allows for the coordination of socio-material realities that go beyond their users and designers (Suchman et al., 2002). As Suchman and colleagues put it, the prototype allows for more or less lasting alignment of the diverse socio-material interests inscribed in the device (2002, p. 168). From a perspective inspired by the sociology of expectations, Wilkie (2014) emphasizes the mediating and mainly performative nature of the prototype, which produces knowledge and artefacts as well as users, futures, and realities. Hillgren and his colleagues (2011), conceive of the process of prototyping as "thinging," that is, not only as a thing (an object) but as a socio-material relationship in which issues can be dealt with. From a different perspective that adopts speculative design methodologies (Dunne & Raby, 2013) and inspired by the work of Stengers (2005), Michael (2012) analyses inventive events provoked by prototypes. These test devices not only introduce forms of exploration that are open to multiple types of rationality, but also produce a productive idiocy, expanding the range of the possible and materializing modes of sociological research. The event of prototyping offers an opportunity to slow down (Stengers, 2005) the processes, generating what Michael calls "inventive problem making," activating more provocative and ambiguous questions (Wilkie, Michael, & Plummer-Fernandez, 2015). DiSalvo (2014) suggests that prototyping, understood as a process of "critical making," constitutes a space in which to generate collaborative practices of exchange, inquiry, and elucidation of the conditions for participation in design. From this viewpoint, the design process allows us to learn about users or entities and to explore how those users or entities constitute and modify each other. That is, instead of seeking out the generation of entirely transparent and familiar objects that are available for a specific solution, the idea would be to think about design as a place of reflection and prospective confrontation through objects, materialities, and experiences that challenge consensus and that are taken as a given (Domínguez Rubio & Fogué, 2015).

In dialogue with this literature, in this article I will argue that there is a need to agnostically analyse prototyping capacities, recognizing local contexts and their political uses. The generative qualities of prototyping associated with its provisional and ongoing condition, which are open to learning through iteration, do not guarantee the generation of deliberative and democratic spaces. On the contrary, and as I will demonstrate, the technique of prototypes does not always happen openly and can be used following a prescriptive protocol that justifies and validates certain normativities, hiding the moments of overflow and dissensus. In this sense, prototyping processes contain a curatorial policy based on a scripting operation (Huybrechts, Hendriks, & Martens, 2017) that determines what is considered to be part of the testing process and what is not. When evaluating the inventive capacity of prototyping processes, it is thus necessary to address the curatorial narrative mobilized to present the results of testing artefacts. This curatorial work that prototypes deploy is linked to a logic of "problem-validating prototype" that involves narratives, materializations, and justifications. I will show that this logic, which is usually mobilized through specific definitions of innovation and driven by experts, is an attempt to generate a scenario of corroboration and attestation "from above" through tests with preconfigured functions. As I will argue, the prototyping techniques were used to prescriptively solve problems defined ex ante rather than using them for a situated exploration of current and potential troubles (Haraway, 2016).

Following Rancière (2000), I argue that the technique of prototyping mobilizes a curatorial politics that defines/discriminates against modes of participation, what is visible and thinkable, what can be spoken and what is unspeakable. In other words, I propose considering prototyping as an event that introduces a process of *partage du sensible* to distribution of space, that is: forms of activity that determine the very manner in which something in common lends itself to participation and the way in which individuals play a part on this participation. For Rancière, the *partage du sensible* refers to an ontological policy that defines modes of participation, what is visible and thinkable, what can be speakable and unspeakable. In contemporary society, this is divided between those who can determine what can be perceived and that which is to be excluded (2000, p. 12). This arbitrariness entails separation between the disadvantaged individuals and those with the power to decide for others (legitimate and illegitimate persons) and forces arbitrary forms of democratic consensus, which performs a prior identification of interests, assuming and projecting the hopes of the people. One of the challenges that Rancière identifies involves generating the conditions for a political practice that is carried out on the basis of the frictions, disruptions, and differences of the participants involved. That is, instead of imposing what can be represented and perceived (assigning persons' static positions), or prefiguring solutions out of context, the process should propitiate possibilities—always contingent and ephemeral—so that those who are "invisible" can emancipate themselves from the conditions imposed by others. This implies a re-articulation of the *partage du sensible*, that is, a shift towards dissensus or forms of counter-participation, questioning the "ontological politics" behind the processes of design intervention (Boano & Kelling, 2013; Keshavarz & Mazé, 2013).

Returning to the case examined in this article, I describe how the urban prototyping process precipitated a series of disruptive situations and dissensus, generating forms of recalcitrance and discussion in response to the experiment conditions. However, the vitality and openness that the prototyping experience generated is subject to a curatorial policy that framed what was validated or not validated within the experimental setting and public discussion. It is precisely this ambivalence of prototyping's capacity to problematize and evoke unexpected realities (problem-making prototype) and its capacity to inscribe and validate certain visions (problem-validating prototype) that I seek to demonstrate using the urban experiment "Shared Streets for a Low-Carbon District."

4 | TESTING A SMART AND ECOLOGICAL SENSIBILITY

An unprecedented "urban laboratory" was introduced in the early morning hours of Friday, September 2, 2016 in the Santiago neighborhood of Lastarria. It was entitled the "Shared Streets for a Low-Carbon District." The intervention, which was led by the NGO Ciudad Emergente (CE), sought to create an experimental scenario that would allow the willingness to transition towards more sustainable neighborhoods to be tested and help identify the benefits of doing so.

The project received the support of various Chilean and international entities. The latter included the consulting firm ARUP, the Eden Project (experts on the development of "community" lunches) and the London School of Economics Cities Programme. However, the main source of funding was the UK Foreign and Commonwealth Office through its "Smart Cities/Infrastructure" and "Climate Change and Low-Carbon Transition" programmes. At the local level, the intervention received the support of the Smart Cities Unit of the Ministry of Transportation, the Ministry of the Environment, Fab Lab Santiago and the Municipality of Santiago. The municipality intended to make an important investment in the area in order to develop infrastructure for sustainable mobility (cycling and pedestrians) as part of the Comprehensive Mobility Plan and the creation of "quiet areas." Authorizing the construction of an experimental bike lane in the area was thus politically convenient because it would help to identify the community's "demand" and raise awareness about the issue. In this sense, the ecological agenda in the area of sustainable mobility offered a favorable context for implementing this experiment.

The NGO, which was created by architects, designers, and social scientists, bases its intervention strategies on urban experiments through "tactical actions." CE is Chile's first and leading tactical urbanism organization, and it was founded to address the challenge of involving citizens in decision-making processes through localized micro-actions and interventions that can trigger long-term changes (interview with CE's director conducted on August 2, 2016). In this sense, the level of experimentation of the NGO must manage the interests of its clients and funders as well as the choices that are made when showing the results.

The organization is described as a "Laboratory for Citizen Urbanism Tactics and Tools" that conducts experimental interventions to promote changes in habits, to enhance citizen participation and to build capacity and relationships between public officials and civil society. The co-founder and Executive Director of CE stated that these actions are based on tactical urbanism and are "light, quick, cheap and involve people in the construction or improvement of a public space." One of the suppositions of CE is that the urban fabric includes "emergent" forms of community building that are commonly invisible to the bureaucratic planning gaze. The organisation's objective is thus to activate and strengthen these emerging communities through "citizen activation tactics" and "social intercommunication 2.0" tools (Figure 1).

In this context, the project "Shared Streets for a Low-Carbon District" justified the need to conduct this experiment and the deployment of a series of prototypes in the urban space based on three narrative operations.

Crises and more sustainable futures: First, both the officials who financed the experience and CE as the entity that executed it explored the need to use a narrative informed by the idea of environmental crisis. Starting from the premise that Santiago is one of the most polluted cities in Latin America, the entities that promoted the initiative constantly alluded to the possibility of turning this experience into a "model" space for creating "more sustainable futures." For the Director of the Smart Cities Unit, one of the agencies that promoted the project, Shared Streets, was a good "laboratory" for showing that it is possible to develop more ecological habits and thus mitigate the environmental crisis that the capital is experiencing. The Director of CE argued that, "Shared Streets seeks to address the urgent issue of climate change through changing habits" (interview conducted on August 2, 2016). As

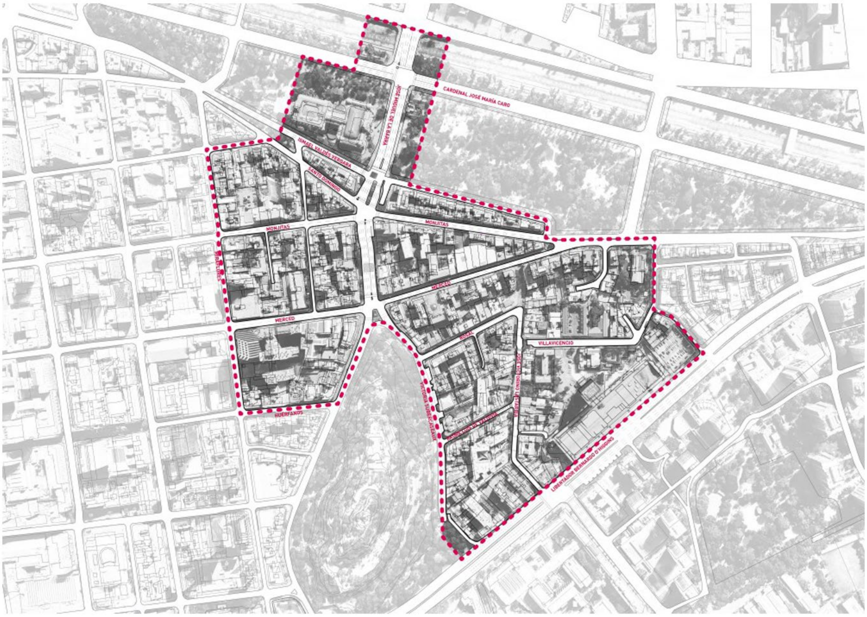


FIGURE 1 Map showing the experiment area [Colour figure can be viewed at wileyonlinelibrary.com]

such, the experiment was initially configured on the basis of a political interest in including Santiago on the path of sustainable urbanism, assuming that the need for greener infrastructure was part of the cause.

Learning and scaling: Second, the experiment's organizers constantly referred to the narrative of learning locally before scaling up. As one of the CE members put it, the importance of this type of intervention is not only creating citizens with greater "ecological awareness" and co-creating a more sustainable city. It is also a matter of being able to scale the results of the experiment to other places. As such, the identification of a "testing" space was a requirement for the development of the experience because it allowed changes to be introduced in a controlled way and for participants to learn from the experience and then extrapolate to other places. For the Ministry of Transportation Smart Cities Unit, this experiment allowed them to "show the people that a Smart City is not only the implementation of technology within the city. It is also about how this technology is introduced through community participation." The experiment was thus promoted as a space of learning-with-the-community in order to then be able to extend them to other territories.

Citizen participation and activation: The third narrative used for the deployment of the Shared Streets intervention consisted of the need to create an opportunity for citizen participation and dialogue regarding expectations about more sustainable urban spaces. As one of the CE experts interviewed stated, the goal was for people to participate in the processes of transforming the city, and in order to achieve this, mechanisms had to be generated that would allow the people who inhabit the territory to be heard. For this reason, the project sought to gather and explore the perceptions of individuals regarding the idea of sharing the space with non-motorized types of mobility (cycling or pedestrian mobility), "and with that reducing CO_2 emissions and combating climate change through changing attitudes." The assessment of this willingness to share spaces and use non-motorized means of mobility had as a condition the creation of prototypes for listening to citizens' voices. As the director of CE suggests, the experimental and emerging nature of this type of urbanism resides in the fact that it does not impose specific barriers to entry. Instead, its vocation is to create temporary, defined spaces of co-production of knowledge, involving both experts and non-experts in the search for solutions to the city's problems.

In the pages that follow, I will describe the problematic trajectory of a series of prototypes that were deployed simultaneously by CE during experiments in the Lastarria neighborhood of Santiago, one of the capital's densest and busiest areas. This "prototype system," as it was called, was to offer a welcoming testing space for more sustainable practices. At the same time, as we will see, the experiment was meant to validate pre-existing interests using the experimentation as problem-validating prototype, a strategy that tends to negate or "package" complexity.

4.1 | Experimental bike lanes

One of the most important interventions developed by CE was the construction of bike lanes in the experimentation field that were open from 7 a.m. to 7 p.m. each of the three days of the experiment. The hypothesis was that this experimental prototype would make tangible a latent need in the community: the need for cycling infrastructure. As one of the project directors put it, "The idea of this prototype is to test how the city could change in order to leave a certain amount of installed capacity behind." As such, this experimental bike lane was not meant to be the finished product, but a temporary "solution" for testing and validating people's willingness to use bicycles and the positive effects that this generated, increasing levels of neighborhood sociability and reducing local pollution levels (Figure 2).

The prototyping process began two days before the official opening of the "urban laboratory" (Friday, September 2) when 30 CE volunteers (most of them from design and architecture schools located in Santiago) began to prepare the experiment site, installing the equipment and tools that would be used to develop the experience. One of the preparatory actions that people noticed most, and which marked the distinctive aesthetic of the “laboratory” was the set of blue circles that was painted on the ground. The eye-catching circles were fundamental to the deployment of the system of prototypes that comprised the experiment: in addition to showing that there was no separation between the sidewalk and the street (and the idea of a shared space), the aesthetic intervention marked a celebratory atmosphere of the figures of cyclists and pedestrians that would be reinforced by the “big lunch” intervention.

The experimental bike lane circuit—which connected the two ends of the perimeter of the testing area—was mapped out in collaboration with the municipality. The local government facilitated the work of installing the prototype, provided temporary infrastructure that would lend the experiment legitimacy and “tested” the level of acceptance prior to the installation of a permanent bike lane.

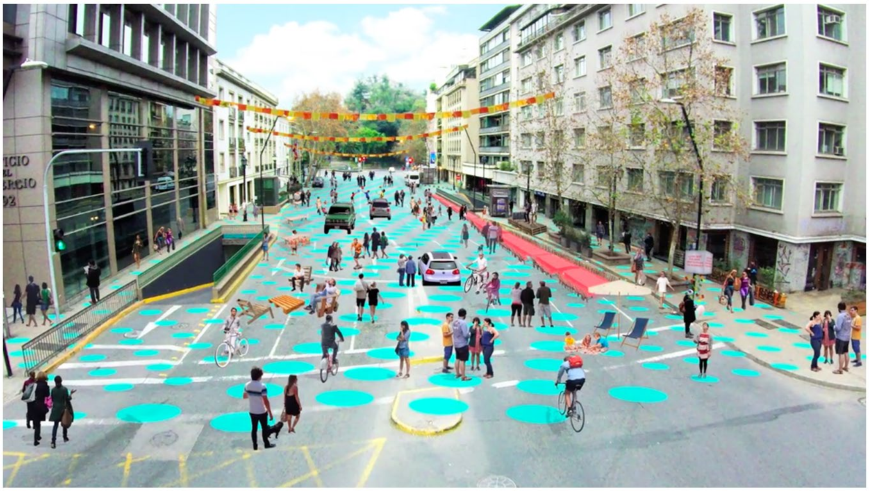


FIGURE 2 Rendering of the experiment [Colour figure can be viewed at wileyonlinelibrary.com]

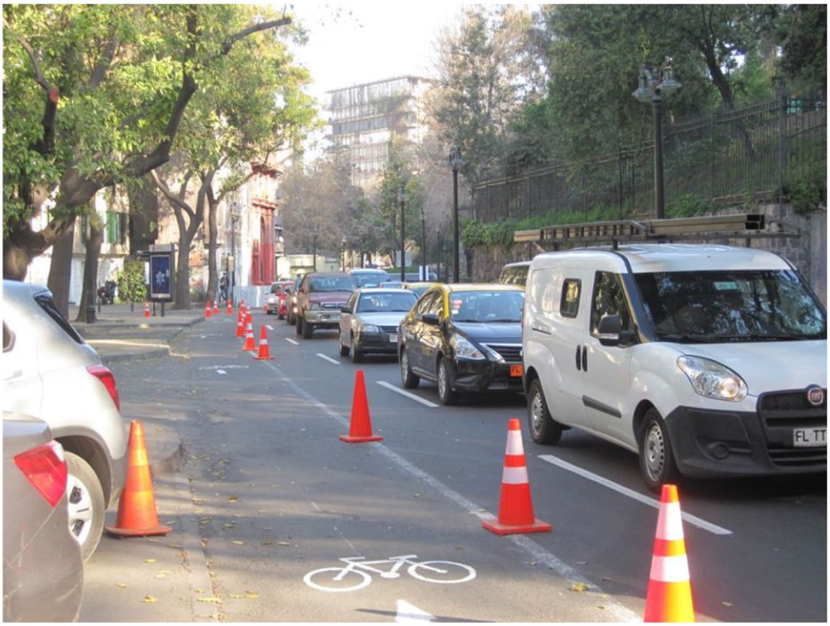


FIGURE 3 Experimental bike lanes and idea trees [Colour figure can be viewed at wileyonlinelibrary.com]

The bike lanes were built using orange cones to mark the areas that could be used for cycling. In addition, vinyl markers were used on the ground in certain parts of the circuit to mark intersections and signs with arrows. In order to provide more organization and safety (which was one of the main concerns that officials raised in regard to the experiment), the bike lane had “human traffic signals,” volunteers who indicated the “correct” use of the circuit and ensured that other vehicles did not enter it (Figure 3).

In addition, “idea trees” were placed throughout the area. These “design probes” composed of overlapping post-it notes were used to foster interaction with the audiences, eliciting reactions and involvement with the experience.

4.2 | Malón urbano (big lunch)

Another prototype developed by CE involved inviting groups of people to take part in a “big lunch” or *malón urbano*. Based on earlier experiences in the UK and traditions in Chile, the purpose of this activity was to activate the participation of local residents by inviting to them share a meal and discuss urban and ecological problems. Designed to engage organizations linked to the focus of the intervention (artists, cyclists’ organizations, neighborhood groups, etc.), the *malón urbano* was held during the evening on the last day of the experiment (Sunday). A special area was set up with long tables and chairs where residents and passers-by could sit and engage in open conversations. The organizers hired bands to perform and set up a series of pop-up stores offering t-shirts, caps, accessories, and bicycle repair services.

The tables, which were built especially for this occasion, were carefully designed to encourage conversation and generate “an atmosphere of camaraderie.” Each table was equipped with chairs, post-it notes and pens to encourage participants to engage in the “spontaneous generation of ideas and record the discussion.” Each table also had a monitor who was responsible for organizing the conversation and, where possible, for encouraging dialogue around the topics of the experiment, such as climate change, pollution, and willingness to adopt more sustainable habits. While specific or pre-set roles were not assigned to the participants, during our observation one could appreciate the presence of certain implicit understandings of how things should be as well as a sense of “community” in terms of preferable values and habits. Far from providing an opportunity to identify disagreements or differences regarding the type of city that one wants, the encounters during the meal took place in a context of consensus without much dissident or controversial voices (Ethnographic notes, Sunday, September 4, 2016).

4.3 | Smart Citizen Kit

Parallel to the urban prototypes, a series of sensors were installed in order to gather metrics and indicators to demonstrate the impacts of the experiment on bicycle use and reduction of air pollution in the area. These sensors were the Smart City component of experimentation, as one CE member told us. In the spirit of open-source technologies and the social innovations that emerged from other urban laboratories, the Smart Citizen Kit (SCK) environmental sensor was distributed to some residents in the experimentation area to measure variables such as temperature, humidity, light intensity, noise levels, nitrogen, and CO_2. The SCK[1] is a low-cost hardware device created by Fab Lab Barcelona to democratize environmental monitoring and empower people to produce their own cities. One of the qualities of the device highlighted by its creators is that it does not operate as a “black box” but as an “open box” that is compatible with non-experts and free experimentation.

The idea was to include residents, non-experts and those affected by motorists’ pollution in the measurement of a series of variables to evaluate the impacts of using greener forms of transportation. The experiment thus made them true intelligent sensors of their own neighborhood. The SCK increased residents’ awareness of environmental issues and trained them to get involved in the work of gathering information about urban pollution. The residents chosen to take part in this activity were given an information sheet that outlined the project objectives. An engineer later installed and activated the SCK.

5 | COUNTER-PARTICIPATION AND MAKING TANGIBLE THE UNEXPECTED

The installation of this "system of prototypes" came up against various problems that highlighted aspects that we had not anticipated. The efforts to verify willingness to use the bike lanes and share the space were exceeded by a series of events and disruptions that exposed entities and voices that had not been considered in the testing. Although those responsible for the project tried to develop a controlled experiment—that is, the idea was to limit the endeavor to testing the hypotheses inscribed in the prototypes—the experience led to events that challenged the suppositions of the experience. While the differences and dissensus generated by the experiment led to a rearticulation of the sensible (Rancière, 2000) in which new relationships emerged, these developments were silenced by the curatorial policy that framed the experiment.

5.1 | Testing the drivers rage?

From the very first day, the intervention generated a high level of traffic, unleashing unpleasant situations for certain audiences. Many drivers who were unhappy with or indifferent to the goals of the intervention constantly honked their horns to show their rejection of the experiment. Instead of "sharing the street," drivers seemed to "suffer" through the experiment, showing their unhappiness with the mechanisms used to "test" the idea of a sustainable city. As one of the volunteers responsible for the experimental bike lanes said, "Today there is a pretty tense environment in the streets. People come up to us to complain about and criticise what we are doing. Drivers have no interest in participating."

Heated discussions took place between pedestrians, cyclists, and drivers. During rush hour (between 6 p.m. and 7 p.m.), the Municipality of Santiago ordered the experimental bike lane removed. While it was subsequently reintroduced in a way that would have less of an impact over the next two days, particularly in the morning, there was a sustained feeling of tension and chaos (Ethnographic notes, Saturday, September 3, 2016). Many of the bike lanes had to be modified to give space back to drivers because of the traffic generated by the intervention.

Drivers infiltrated the experiment and were viewed as "pollution" by the organizers, as an element that wasn't included in the experiment protocol. The prototypes had been designed to verify the willingness of cyclists and pedestrians to participate, not to be problematized by external agents that were not considered in the setting. As such, the forms of appropriation that the drivers developed in response to the CE design were considered anecdotal. In the words of one of the project organizers, they were viewed "as irrelevant to the experiment's objectives."

It is important to note that despite this decision to minimize drivers' reaction to the prototypes, from the beginning they were key stakeholders in the experimentation and its purpose. One of the main assessments that inspired the "Shared Streets" experiment was the exponential growth of the number of vehicles in Santiago and the need to test willingness to co-exist with different modes of transportation. As the Director of EC said:

> *Santiago has been declared an "environmental saturation zone" due to the particulate matter. This is a problem, and the number of vehicles increases exponentially every year. Chile and Russia are the OECD member states that have the largest number of traffic accidents and deaths. There is an issue of coexistence. As such, this "Shared Streets" project addresses an issue of conflict with drivers…. Today we compete for space in the street…. They are streets that basically were not designed for people, but for cars. (Interview with CE's director conducted on August 2, 2016)*

The category of car drivers clearly played a significant role in the experiment. The prototype system was designed with this "audience" in mind and, at least theoretically, its presence inspired the various interventions. There was an effort to measure the willingness to coexist among the different modes of transportation and environmental effects of decreasing the use of motor vehicles. However, when the experiment was executed, the critical reaction and interactivity of drivers with the intervention methodology was not considered part of what was to be measured. Their form of participation clashed with the normative and methodological premises of the testing process, so the decision was made to consider these troubles and manifestations of dissensus noise that could be omitted. The forms of agonism produced by the prototyping experience (a process that I call "problem-making prototype") were not translated as meaningful data for rethinking the suppositions of the intervention and deploying new futures or questions, instead imposing a logic of reduction of the problem definition and their testing.

5.2 | The hipster "invasion"

The atmosphere of camaraderie and dialogue that organizers sought to prototype through the urban lunch was questioned by the audiences that were supposedly "external" to the experiment and thus invisible to the organizers of the experience. Ethnographic observations revealed the emergence of a resistant or at least indifferent response to the Shared Streets laboratory. The people interviewed during the experience described the experiment as an "invasion" by elites and hipsters who are removed or disconnected from the neighborhood and the "true" needs of the people. Given the neighborhood's proximity to local tourist attractions, the development of the system of prototypes deployed seemed to attract more tourists and passers-by than residents. In fact, several people were angered by the experiment's purpose because it ignored other more urgent needs such as security. Some felt that it was an error for the government to authorize "freezing" a part of the city given that it generated unnecessary traffic in the area.

Some criticized the utopic or unrealistic nature of the experience and the idea that CE would want to "change the city in five minutes," showing that they were sceptical of these "playful" and "rapid" modes of promoting new urban habits. Others even criticized the color of the circles painted in the street because they thought it made the neighborhood less attractive. One Facebook user said that the idea of sharing the street seemed "downright stupid," which led CE to answer:

> *Just as people thought that women's suffrage was stupid one hundred years ago and now see it as a matter of common sense, we want to promote a city with a common sense that involves streets that allow for slow vehicular passage and pedestrian flow. I hope it doesn't take you one hundred years to understand this.*

This confrontation between the parties responsible for the experiment and divergent forms of participation revealed the generative capacity of the intervention (making visible critiques and justifications) as well as the lack of tolerance of the experimental setting in regard to accepting recalcitrant reactions and those that went against the reactions that had been imagined.

5.3 | The failure of the Smart Citizen Kit

The use of the SCK was subjected to interference and failures. As one of the NGO's experts explained, "It wasn't difficult to find people who wanted to install the kit. It was hard to find people who have the technical capacity to manage the kit."

Some users began to report "calibrating failures" related to the SCK requirements, which impacted the sensor's measurements. Many of the sensors experienced failures due to poor WiFi connections, poor placement, blackouts or residents' negligence.

However, the main reason the sensors failed was a lack of care and proper handling. For example, some people disconnected the sensor when they had to plug in a different device or when the sensor got in the way of domestic activities such as cleaning.

All of these failures and discontinuities caused the data obtained to include noise or errors or for data collection to be interrupted for several hours or days. The director of the NGO stated that the SCK proved to be "more rigid than expected" and was difficult to maintain and align with the domestic ecosystem.

6 | CURATION BY PROTOTYPE

CE's reaction to the situations generated by these prototypes was to leave them out of the narrative and official results. They published the expected reactions to the prototypes, but not the unanticipated, recalcitrant events that infiltrated the intervention.

They never alluded to the conflicts that the deployment of the various urban tactics caused in the reports and presentations on the lessons learned from the experiment. The frictions that emerged from the experiment were not considered worthy of being addressed, and were instead seen as noise that had to be eliminated.

From this perspective, a stakeholder from the Municipality of Santiago said that "the data provided by the intervention support the construction of more sustainable urban environments." Along these same lines, CE's director said that CO_2 concentration "decreases five-fold when the city is shared" during a presentation on the results of the experiment.

Rather than including the creation of audiences who were resistant to ecological sensibility or the multiple failures produced in the installation of the SCK in the design, the focus was on prototyping the emergence of an "eco-friendly" audience.

In this sense, the validating—not exploratory or open to non-knowledge—use of the prototypes prevailed. The experimental prototypes were mobilized to highlight certain preconceived occurrences and realities, privileging the creation of certain audiences and sensibilities over others. In other words, the testing technique sought to inscribe a particular regime of *partage du sensible*, making the frictions and forms of counter-participation (Tironi, 2018) that occurred during the experiment disappear.

Meanwhile, the officials were satisfied with the results and spirit of the experience, particularly in terms of the visibility and turnout. They recognize that it was mainly demonstrative and, as a member of the Municipality of Santiago put it, "This CE project is tied to projects that we hope to develop, and it helps us to monitor what we are going to do." However, the experiment did not in any way inform specific or strategic decisions regarding the future project.

6.1 | Final discussion: Between the "problem-validating prototype" and the "problem-making prototype"

Through the case of experimentation, I have shown how testing experiences act politically on the world. The prototyping process involves ontological politics (Mol, 1999), that is, certain versions of making worlds. In the case of the prototyping of sustainable habits in a neighborhood in Santiago, we observed specific modes of translating the world, of designing, distributing, and qualifying the agencies at play. The series of temporary tests that a prototype undergoes during this process reveals possible events and narratives upon which agencies can make and diagram interests, materialities, and forces. I have thus shown that testing with prototypes introduces a singular political language: its function cannot be reduced to a simple evaluation of a preconceived idea. Rather, throughout the process one comes up against situations of uncertainty and frictions that produce opportunities for transformation and redefinition.

Prototype tests can be understood as modes of *partage du sensible* (Rancière, 2000): they make possible ways of making visible/invisible certain matters of concern. To put it differently, prototypes make certain facts or interpretations of reality exist, curating what can be seen or said. The prototype's fragile materiality has the potential to take up the unstable and uncertain, bringing us closer to realities that cannot be fixed or that require new definitions and repertoires of approaches (Tironi & Hermansen, 2018). This is precisely what happened in the "Shared Streets" project. Various ways of understanding the concept of the sustainable city collided and the failures and reactions that emerged during the experiment called into question the very design of the intervention and its ways of "convening" audiences.

In this sense, one could say that the narrative and material events produced by the urban experiment prototyped the "reality" of Santiago under two different modalities.

The first is a "problem validating prototype" logic in which the results of the test align themselves under a scheme of *partage du sensible* that determines what is valid, possible, and valuable for the ecological transition. In this case, the experiment rested on a strong dichotomy between normal and abnormal, expelling from the discussion of the "sustainable city" all of the excesses and frictions that did not fit within the defined ontological framework. The politics of this form of testing derive their capacity to manage and nudge from the order of the sensitive in order to prescribe ideas and colonize spaces, reducing multiplicity and differences (Rancière, 2000) to a verifying and consensual narrative about a sustainable "feeling." In fact, as the director of CE states:

> *The municipality of Santiago will invest 150 million pesos to implement more permanent cycling infrastructure. It is saying that it will do so ahead of time. So instead of seeking to determine whether or not a bike lane is good or bad, our prototypes are looking to address the issue of changing habits and making people aware of an important issue. (CE director, interviewed on August 2, 2016)*

In a way, the prototyping and testing processes that the NGO CE implemented were oriented towards a validating curatorial meaning, representing preconceived institutional interests. This process was "supported" by the reputation gained from having European (UK) funding sources and the technical support of the London School of Economics. This determines the incentives for making invisible the failures and frictions generated during the intervention to a great extent, imposing a "politically correct" narrative on the need to validate sustainable urban policies.

Here the urban reality is conceived of as something to intervene-in-to-change. Urban prototyping is a tool for issuing a call and producing a type of participation that could validate (Marres, 2012), comment on, or complement desired changes. From this perspective, social issues are something to be revealed and confronted or affected or redesigned through the technologies of nudge and co-design. In contrast to closed laboratories in which failure and recalcitrance are a form of refuting hypotheses, this open-air prototyping allows designs to be deliberately arranged to avoid failure and uncertainty and generate results that can be activated and scaled for governance (Karvonen, Cugurullo, & Caprotti, 2018).

On the other hand, the Shared Streets experiment led to a significant process of problem-making prototyping in which the different issues and discussions were unexpectedly created based on the testing artefacts deployed. The most interesting aspect of the urban experiment produced by CE was not its capacity to validate a "demand" for more ecological surroundings through prototypes or its effort to critique the motorized city through the intervention. Its generative capacity was instead found in its problem-making capacity to deploy and invent (Marres, Guggenheim, & Wilkie, 2018), a scenario in which recalcitrant realities will be manifested for the "sustainable consensus." Despite the curating adopted by the parties responsible for the experiment in order to present the urban testing results, the installation of the prototypes in the streets generated counterfactual narratives with a series of interactions, failures and disruptions that I have tried to depict here. While the forms of counter-participation that the testing process triggered were not incorporated by those responsible for the intervention in the final report, these disruptions cannot be understood nor would have happened without the intromission/provocation produced by the prototypes, which served as agitators for making the drivers' voice explicit.

Here prototypes are not a strategy for finding a given reality or for transforming it under a predefined imperative. Rather, the prototypes acted here by evoking and provoking issues that did not fall within the parameters set by the experimenters. Here the urban reality ceases to be a problem and becomes a field of as yet unknown possibilities in an effort to expand the ways of understanding the problems at hand. This form of operating of prototypes forces entities and visions that had been ignored or undervalued to be recognized, promoting a commitment to the unexpected and recalcitrant and generating scenarios of intervention that go beyond the logic of the dominant solutionism.

The goal of this dichotomy—problem-validating prototype/problem-making prototype—is not, of course, to exhaust the capacities of the prototyping process in the urban space or to arrive at an apologia for indetermination and the speculative. Rather, at a time in which experimentalness and turning to urban laboratories begin to proliferate in our "smart" cities as a tool for democratic debate, it seems necessary to critically analyse the uses and scopes of these testing strategies. As we have discussed here, the culture of prototypes is ambivalent in terms of its political effects, and its designs do not always ensure the openness necessary to learn from failure and the unexpected.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author (Martín Tironi), upon reasonable request.

NOTE

[1] The SCK contains various sensors, a data processing board, a battery and a cover. The data are automatically uploaded when it connects to a WiFi signal.

REFERENCES

Boano, C., & Kelling, E. (2013). Toward an architecture of dissensus: Participatory urbanism in South-East Asia. *Footprint*, *7*(2), 41–62.

Broto, V. C., & Calvet, M. S. (2016). Green enclaves, neoliberalism and the constitution of the experimental city in Santiago de Chile. In J. Evan, A. Karvonen, & R. Raven (Eds.), *The experimental city* (pp. 107–121). New York, NY: Routledge.

Bulkeley, H., & Castán Broto, V. (2013). Government by experiment? Global cities and the governing of climate change. *Transactions of the Institute of British Geographers*, *38*(3), 361–375.

Callon, M., Lascoumes, P., & Barthe, Y. (2009). *Acting in an uncertain world*. Cambridge, MA: MIT Press.

Caprotti, F., & Cowley, R. (2017). Interrogating urban experiments. *Urban Geography*, *38*(9), 1441–1450.

Corsín Jiménez, A. (2014). Introduction. The prototype: More than many and less than one. *Journal of Cultural Economy*, *7*(4), 381–398.

De Lange, M., & De Waal, M. (2013). Owning the city: New media and citizen engagement in urban design. In *Urban land use* (pp. 109–130). Apple Academic Press.

DiSalvo, C. (2014). Critical making as materializing the politics of design. *The Information Society: An International Journal*, *30*(2), 96–105.

Domínguez Rubio, F., & Fogué, U. (2015). Unfolding the political capacities of design. In A. Albena & A. Zaera-Polo (Eds.), *What is cosmopolitical design?* (pp. 143–160). London, UK: Ashgate.

Dunne, A., & Raby, F. (2013). *Speculative everything: Design, fiction, and social dreaming*. London, UK: MIT Press.

Evans, J., & Karvonen, A. (2014). 'Give me a laboratory and I will lower your carbon footprint!'—Urban laboratories and the governance of low-carbon futures. *International Journal of Urban and Regional Research*, *38*(2), 413–430.

Evans, J., Karvonen, A., & Raven, R. (Eds.). (2016). *The experimental city*. Routledge.

Forlano, L. (2016). Decentering the human in the design of collaborative cities. *Design Issues*, *32*(3), 42–54.

Gabrys, J. (2014). Programming environments: Environmentality and citizen sensing in the smart city. *Environment and Planning D: Society and Space*, *32*(1), 30–48.

Gabrys, J. (2016). *Program earth: Environmental sensing technology and the making of a computational planet* (Vol. 49). Minneapolis, MN: University of Minnesota Press.

Galey, A., & Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, *25*(4), 405–424.

Gieryn, T. F. (2006). City as truth-spot: Laboratories and field-sites in urban studies. *Social Studies of Science*, *36*(1), 5–38.

Gross, M. (2016). Give me an experiment and I will raise a laboratory. *Science*, *Technology*, *& Human Values*, *41*(4), 613–634.

Gross, M., & Krohn, W. (2005). Society as experiment: Sociological foundations for a self-experimental society. *History of the Human Sciences*, *18*(2), 63–86.

Guggenheim, M. (2012). Laboratizing and de-laboratizing the world changing sociological concepts for places of knowledge production. *History of the Human Sciences*, *25*(1), 99–118.

Halpern, O., LeCavalier, J., Calvillo, N., & Pietsch, W. (2013). Test-bed urbanism. *Public Culture*, *25*(2), 272–306.

Haraway, D. J. (2016). *Staying with the trouble: Making kin in the Chthulucene*. Durham, NC: Duke University Press.

Harrison, C., & Donnelly, I. A. (2011). A theory of smart cities. In *Proceedings of the 55th Annual Meeting of the ISSS-2011* (Vol. 55, No. 1). Hull, UK. Retrieved from https://bit.ly/1knh7bW

Hillgren, P. A., Seravalli, A., & Emilson, A. (2011). Prototyping and infrastructuring in design for social innovation. *CoDesign*, *7*(3–4), 169–183.

Huybrechts, L., Hendriks, N., & Martens, S. (2017). Counterfactual scripting: Challenging the temporality of participation. *CoDesign*, *13*(2), 96–109.

Karvonen, A., Cugurullo, F., & Caprotti, F. (Eds.) (2018). *Inside smart cities: Place*, *politics and urban innovation*. Routledge.

Karvonen, A., & van Heur, B. (2014). Urban laboratories: Experiments in reworking cities. *International Journal of Urban and Regional Research*, *38*(2), 379–392.

Keshavarz, M., & Mazé, R. (2013). Design and dissensus: Framing and staging participation in design research. *Design Philosophy Papers*, *11*(1), 7–29.

Kimbell, L., & Bailey, J. (2017). Prototyping and the new spirit of policy-making. *CoDesign*, *13*(3), 214–226.

Kullman, K. (2013). Geographies of experiment/experimental geographies: A rough guide. *Geography Compass*, *7*(12), 879–894.

Latour, B. (2001). What rules of method for the new socio-scientific experiments? In *Experimental cultures: Configurations between science*, *art*, *and technology*, *1830–1950* (pp. 123–135). Berlin, Germany: Max-Planck-Institut für Wissenschaftsgeschichte. Retrieved from http://www.bruno-latour.fr/node/372

Laurent, B. (2017). *Democratic experiments: Problematizing nanotechnology and democracy in Europe and the United States*. Cambridge, MA: MIT Press.

Laurent, B., & Pontille, D. (2018). Towards a study of city experiments. In C. Coletta, L. Evans, L. Heaphy, & R. Kitchin (Eds.), *Creating smart cities* (pp. 90–103).

Laurent, B., & Tironi, M. (2015). A field test and its displacements. Accounting for an experimental mode of industrial innovation. *CoDesign*, *11*(3–4), 208–221.

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

Marres, N. (2012). *Material participation. Technology*, *the environment and everyday publics*. London, UK: Palgrave Macmillan.

Marres, N. (2018). What if nothing happens? Street trials of intelligent cars as experiments in participation. In *TechnoScience in society*, *sociology of knowledge yearbook*. Nijmegen, the Netherlands: Springer/Kluwer.

Marres, N., Guggenheim, M., & Wilkie, A. (2018). *Inventing the social*. Manchester, UK: Mattering Press.

Marvin, S., A. Luque-Ayala, & C. McFarlane (Eds.) (2016). *Smart urbanism: Utopian vision or false dawn?* Abingdon, UK: Routledge.

Michael, M. (2012). De-signing the object of sociology: Toward an "idiotic" methodology. *The Sociological Review*, *60*(1_suppl), 166–183.

Mol, A. (1999). Ontological politics: A word and some questions. *The Sociological Review*, *47*(S1), 74–89.

Rancière, J. (2000). *La partage du sensible: esthétique et politique*. Paris: La fabrique éditions.

Sanders, E. B. N., & Stappers, P. J. (2014). Probes, toolkits and prototypes: Three approaches to making in codesigning. *CoDesign*, *10*(1), 5–14.

Stengers, I. (2005). The cosmopolitical proposal. In B. Latour & P. Weibel (Eds.), *Making things public: Atmospheres of democracy* (pp. 994–1003). Cambridge,IL: Center for Art and Media in Karlsruhe, MIT Press.

Suchman, L., Trigg, R., & Blomberg, J. (2002). Working artifacts: Ethnomethods of the prototype. *British Journal of Sociology*, *53*(2), 163–179.

Tironi, M. (2018). Speculative prototyping, frictions and counter-participation: A civic intervention with homeless individuals. *Design Studies*, *59*, 117–138.

Tironi, M., & Criado, T. S. (2015). Of sensors and sensitivities. Towards a cosmopolitics of "smart cities"? *Tecnoscienza: Italian Journal of Science & Technology Studies*, *6*(1), 89–108. Retrieved from http://www.tecnoscienza.net/index.php/tsj/article/view/217

Tironi, M., & Hermansen, P. (2018). Cosmopolitical Implications in the Prototyping Process: Ethnographic design practice at the National Zoo in Santiago, Chile. *Journal of Cultural Economy*, *11*(4).

Tironi, M., & Valderrama, M. (2018). Unpacking a citizen self-tracking device: Smartness and idiocy in the accumulation of cycling mobility data. *Environment and Planning D: Society and Space*, *36*(2), 294–312.

Tsing, A. L. (2011). *Friction: An ethnography of global connection*. Princeton, NJ: Princeton University Press.

Voytenko, Y., McCormick, K., Evans, J., & Schliwa, G. (2016). Urban living labs for sustainability and low carbon cities in Europe: Towards a research agenda. *Journal of Cleaner Production*, *123*, 45–54.

Wilkie, A. (2014). Prototyping as event: Designing the future of obesity. *Journal of Cultural Economy*, *7*(4), 476–492.

Wilkie, A., Michael, M., & Plummer-Fernandez, M. (2015). Speculative method and twitter: Bots, energy and three conceptual characters. *The Sociological Review*, *63*(1), 79–101.

How to cite this article: Tironi M. Prototyping public friction: Exploring the political effects of design testing in urban space. *Br J Sociol*. 2020;71:503–519. https://doi.org/10.1111/1468-4446.12718

Received: 12 March 2019 | Revised: 8 July 2019 | Accepted: 8 December 2019

DOI: 10.1111/1468-4446.12739

SPECIAL ISSUE

WILEY

What do stress tests test? Experimentation, demonstration, and the sociotechnical performance of regulatory science

Nathan Coombs

School of Social and Political Science, University of Edinburgh, Edinburgh, United Kingdom

Correspondence
Nathan Coombs, School of Social and Political Science, University of Edinburgh, 15a George Square, Edinburgh, United Kingdom.
Email: nathan.coombs@ed.ac.uk

Funding information
Leverhulme Trust, Grant/Award Number: ECF-2014-304

Abstract

After their successful introduction during the 2007–2009 financial crisis, central bank stress tests were adopted as a fixture of international banking supervision. However, in recent years a new normal has emerged where banks are expected to pass the tests, raising questions about the tests' usefulness and legitimacy. Combining a dramaturgical interpretation of regulatory science with the idea of performativity in the sociology of finance, this article understands stress tests as a sociotechnical Goffmanian performance. With a focus on the Bank of England's program, the paper argues that the Bank's decision to make their tests "predictable" is an attempt to shore up central bank legitimacy by constraining regulatory discretion. This is accomplished through the use of calculative and procedural stage management techniques which allow the Bank to control the contingency of the testing process while demonstrating its objectivity. Nevertheless, the conclusion suggests that in the context of low levels of trust in central banks, routine declarations of "all clear" may undermine public confidence in the tests' credibility and necessity. The study draws on 20 interviews with high-level regulators, financial practitioners and other stakeholders in the Bank of England's stress tests.

KEYWORDS

central banks, finance, legitimacy, performance, regulation, stress testing

1 | INTRODUCTION: WHERE'S THE DRAMA GONE?

When in 2016 the Bank of England announced the results of their annual stress test, the UK's Channel 4 News visualized the story by showing the Royal Bank of Scotland's (RBS) logo straining as bricks piled on top. While the testing of building materials has little in common with the complicated calculative procedures involved in financial stress tests, the image helped to dramatize the results. It was not just that RBS's capital would fail to stay above a relatively arbitrary regulatory minimum in a hypothetical scenario where unemployment shoots up to 12% and house prices fall by 35% (Bank of England, 2016). The Scottish bank would implode, fracture, shatter irrecoverably. The image amplified the Bank's message: RBS should raise more capital, urgently.

Such visualizations have become rarer in recent years as reports on stress tests have retreated from televised news programs to the pages of the specialist financial press. For although in their early years a significant number of global banks were failed in stress testing programs, both the Federal Reserve's and Bank of England's tests have since seen a steady decline in the number of failures, settling at zero at the time of writing (Figure 1). This does not mean that the announcement of a test's results has become entirely drama free: a few banks continue to skim close to their regulatory capital minimum on some measures, or be criticized on other grounds. But with so few institutions being failed outright, the occasion has become predictable. In the language of the financial sector, the tests are now BAU: Business as Usual. All of which has led commentators to ask whether the tests have outlived their usefulness (Tett, 2015).

What explains the transformation of stress tests from a show of regulatory force to a relatively uneventful ritual? Drawing on 20 interviews with high-level regulators, financial practitioners and other stakeholders in the Bank of England's stress tests, this article argues that the Bank's decision to frame their tests as "predictable" was motivated by the need to manage financial sector concerns about regulatory discretion while reassuring the public that the financial system is secure. To understand that dual messaging, the paper combines a performative interpretation of "regulatory science" (Irwin, Rothstein, Yearley, & McCarthy, 1997; Jasanoff, 1987, 1990, 1995; Rushefsky, 1986; Salter, Leiss, & Levy, 1988) with the notion of performativity in the sociology of finance (Callon, 1998; Callon, Millo, & Muniesa, 2007; Coombs, 2016; MacKenzie & Millo, 2003; Millo & MacKenzie, 2009). Hilgartner's (2000) Goffmanian understanding of regulatory performances is particularly relevant here,

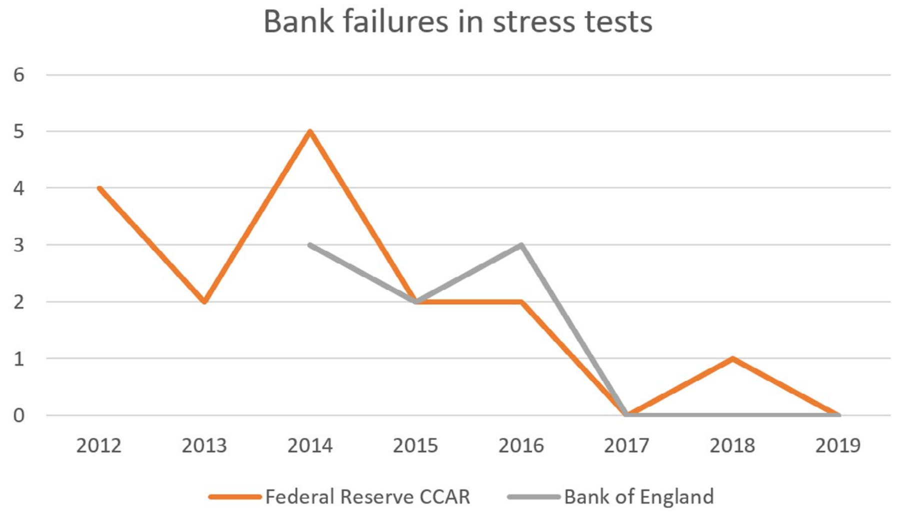


FIGURE 1 Number of banks failing the stress tests (2012–2019)
Source: Author's own based on Federal Reserve and Bank of England data. Failure in the Federal Reserve's Comprehensive Capital Analysis and Review (CCAR) is defined by 'objections to the capital plan'. For the Bank of England's tests, failure is defined as a bank resubmitting their capital plan during the testing process. [Colour figure can be viewed at wileyonlinelibrary.com]

since it points to how the authority and credibility of policy-relevant science is not given but must be actively cultivated through persuasive frontstage demonstrations. Where the paper goes further is in showing how the experimental core of the testing process is also shaped by the communicative needs of the central bank. This involves the deployment of calculative and procedural stage management techniques to control the contingency of the testing process and demonstrate its objectivity. In short, the paper shows that the predictability and low/no failure rate of recent tests is as much a sociotechnical accomplishment as the earlier use of stress testing to legitimize the recapitalization of the banking sector.

At the same time, the paper's critical coda contends that these technocratic procedures might have unintended and counterproductive consequences. Following commentators who argue that confidence in central banks has still not been restored after the financial crisis (Braun, 2016; Dietsch, Claveau, & Fontan, 2018; Riles, 2018; Tucker, 2018), it is possible to conjecture that repetitive judgments of "all clear" may be received incredulously by their public audience or simply contribute to the tests losing their "hold on the collective imagination" (Jasanoff, 2005, p. 248). That has implications for their political economy: without an attentive audience the tests are more likely to be progressively watered down to accommodate financial interests, or dispensed with altogether. Ultimately, then, the wider message is that studies of high-profile public tests should be about "the sort of social and political relationships embedded within society as whole" (Pinch, 1993, p. 38). With testing becoming an increasingly ubiquitous governmental technology, researchers should place the technical details of tests in a large-scale interpretive frame to understand how, when, and under what conditions they are taken as legitimate forms of knowledge production on which to base policymaking.

The rest of the paper is organized as follows. The first section revisits key concepts in the sociology of testing and brings them into dialog with work on regulatory science and central banking. The second section presents the study's data sources. The third section interrogates how financial actors perceive the tests' accuracy and the performative function of pass/fail judgments. The fourth section shows that the tests also serve as a vehicle for experimental "macroprudential" policies and how they have elicited concern about regulatory discretion. The fifth section details three stage management techniques employed by the central bank to control the contingency of stress tests and their reception. The sixth section reflects on the potential implications of the low/no failure rate of recent tests for the legitimacy of the practice. The conclusion follows.

2 | REGULATORY SCIENCE IN CENTRAL BANKING

Testing may not at first seem the most sociological of subjects. But as the nascent sociology of testing showed in the 1980s and early 1990s, once realist preoccupations are dispensed with it is possible to examine the different frames of meaning communities of practitioners bring to experimental results (Collins, 1987, 1988; Collins & Pinch, 1979; MacKenzie, 1990; Pinch, 1993). From this perspective, testing is not a merely technical exercise but a practice invested with interpretive and political saliency where the accuracy of results can be contested by social actors (MacKenzie, 1990). Addressing how this dynamic plays out on the public stage, Collins (1988) makes an important distinction between "experiments" and "demonstrations." Experiments are *pre-closure* attempts to infer the properties and behaviors of technologies among the "core set" of specialists. Demonstrations, on the other hand, are *post-closure* attempts by specialists to exhibit experimental results so that the public can see them "with their own eyes." Collins's examples are tests intended to demonstrate the safety of flasks of nuclear waste and a new type of jet fuel. Despite both tests having "apparently clear and convincing outcomes" (Collins, 1988, p. 733), these were interpretations by journalists and onlookers which did not align with the judgments of the tests' designers.

Coming at a similar set of issues but from an institutional level of analysis, political sociologists have grappled with the differences between "regulatory science" and regular research science (Irwin et al., 1997; Jasanoff, 1987, 1990, 1995; Rushefsky, 1986; Salter et al., 1988). Building on the observation that in the latter half of the twentieth century regulatory agencies were tasked with increasing responsibility for risk assessment that forced them to take an active role in knowledge production (Jasanoff, 1990), the concept of regulatory science (and equivalents such as "mandated science," "trans-science," and "policy-relevant science") draws attention to a sociologically distinctive character of scientific research when entangled with policymaking. Scholarship has observed how the expectation that regulatory agencies respond rapidly to emerging risks pushes them to lower the evidentiary bar compared to academic research science, while inviting adversarial challenge from businesses and civil society actors concerned that science is being manipulated to support an ideological agenda. Meanwhile, another strand of the literature has pointed to the problem of democratic accountability when governmental decision making becomes the responsibility of opaque technocratic committees, alienating publics from the agencies entrusted with representing their interests (Bijker, Bal, & Hendriks, 2009). Hilgartner's (2000) interpretation of regulatory science as a Goffmanian performance attempts to make sense of how regulatory actors navigate these dilemmas. On their back stage, regulators are well aware of the precariousness of the knowledge on which they have to make decisions. That is why the authority of science is mobilized so forcefully on their frontstage: to supress uncertainty and reassure the public that regulatory decision making is objective and credible.

Interestingly, central bank operations have not previously been analyzed through the lens of regulatory science despite researchers drawing attention to the "scientization" of central banking and its immersion in "technical rationality" since the 1980s (Abolafia, 2012; Marcussen, 2009; Mudge & Vauchez, 2016). Nonetheless, similar conclusions have been reached independently. Although central banks are preeminent producers of macroeconomic knowledge in certain academic fields (Dietsch et al., 2018; Mudge & Vauchez, 2016), scholars have noted that this body of technical knowledge has been continually tested and contested. Goffman's dramaturgical analytic has therefore been invoked to make sense of how central banks sustain public trust in their operations (Abolafia, 2012; Braun, 2016). For example, Abolafia notes that during the "stagflation" crisis of the 1970s, which threw into question the use of the "Phillips Curve" to steer interest rate policy and left rudderless the Federal Open Markets Committee's private deliberations, in public the U.S. Federal Reserve continued to communicate complete confidence in its policy decisions (Abolafia, 2012). Others have also noted that macroeconomic forecasts play a frontstage role in bolstering the credibility of the central banks' policy commitments despite having a poor track record in predicting economic developments (Beckert, 2016; Evans, 1997).

This paper argues that the Goffmanian notion of performance can also shed light on central banks stress tests but that their performative aspects should be located in the sociotechnical operations at the core of the experimental process. For as a large body of science studies inspired economic sociology has demonstrated, models and market devices play a vital evaluative and coordinating role in financial markets (e.g., Callon, 1998; Callon et al., 2007; MacKenzie & Millo, 2003; Millo & MacKenzie, 2009). That is because they do not simply represent financial "objects" but constitute them in the act of measurement. Less has been written on how regulatory authorities make use of models and devices (Coombs, 2016, 2017, 2020; Langley, 2015; Lenglet, 2011; Williams, 2009), but the point carries across: because the meaning of categories central to regulatory practice such as "risk" and "value" are constructed on the basis of radically reflexive and uncertain market dynamics (Beckert & Bronk, 2018), these categories are highly reactive to acts of measurement. Regulators are aware of these performative dynamics and stage manage their stress tests accordingly.

The article also offers insights about how regulatory performances confer legitimacy on the decision making of governing authorities. Taking seriously the role played by publics in constructing the legitimacy of regulatory science, work on public reason has proposed the concept of "civic epistemologies" (Jasanoff, 2005, 2011b; Miller, 2005, 2008) to make sense of the culturally specific "shared understandings about what credible claims should look like and how they ought to be articulated, represented and defended" (Jasanoff, 2005, p. 249). The major insights of this work stem from a comparison between the United States, United Kingdom, and Germany, which demonstrate a surprising diversity of "social and epistemic arrangements" (Miller, 2008, p. 1898). For example, in the United States quantitative analysis is preferred and the legitimacy of regulatory science is established adversarially in the legal system, whereas in the United Kingdom trust is placed in public reports by respected civil servants. Yet the attachment of the concept of civic epistemology to national cultural comparison (Jasanoff, 2005,

p. 270) has arguably played a role in closing off other research avenues, such as how historical and political context shapes how publics receive and interpret regulatory performances. With a focus on the role played by stress tests in the post-financial crisis political conjuncture, this article provides an example of the idiosyncratic ways in which publics may be enlisted (or fail to be enlisted) in the construction of regulatory legitimacy.

3 | DATA SOURCES

The paper draws on 20 semi-structured interviews (each lasting 45–150 minutes) with high-level regulators, financial practitioners, and other stakeholders in the Bank of England's (Bank's) stress tests (22 interviewees in total). Launched in 2014, the Bank's public testing program is a relative newcomer compared to its international counterparts. The European Banking Authority's (EBA) program stretches back to the crisis-era Eurozone tests in 2009, and the Federal Reserve's (Fed's) most challenging program (the Comprehensive Capital Analysis and Review—CCAR) has run annually since 2011. Perhaps because it was able to learn from these institutions' experiences, the Bank's program is generally considered a compromise between their approaches. The Fed's CCAR is a "top-down," labor-intensive process in which banks submit their balance sheet data to the Fed, who then runs the stress simulations. In contrast, while the EBA crafts the scenario its "bottom-up" approach delegates responsibility for running simulations to the banks themselves and their national authorities.[1]

Because of the Bank's evenly weighted combination of top-down and bottom-up testing—in which the Bank sets the scenario, the commercial banks simulate it themselves, and the Bank checks their results—it was necessary to speak to both public and private actors involved in the process. These interviewees were contacted using a snowball method beginning with contacts established by attending three industry conferences on stress testing in London in 2016. The initial aim of the research was to discover how the stress scenarios are designed, interpreted, and modeled. Over time, the investigation broadened to explore how the policy space has changed over the years, and comparisons between different central banks' approaches to stress testing. The mix of public and private sector interviewees provided a balanced sample of interpretive and normative expert opinion on the motivations for, quality of, and potential problems with the Bank's stress tests.

Given that this research concerns a confidential policy space, all interviews were conducted on the promise of anonymity. However, Table 1 breaks down the interviewees by professional category. In a number of cases, where individuals had moved between positions and across the public-private divide over preceding years, they are classified according to their most relevant professional position. Most of the key individuals at the Bank involved in the designing and implementing of its financial stability architecture and stress testing program were consulted. So too were senior personnel involved in managing risk, stress testing, and capital allocation at the UK's commercial banks. It should also be noted that some regulatory interviewees requested to read the manuscript prior to publication. They used this opportunity to correct technical errata but did not comment on its arguments or substantive claims.

Further sources which inform this discussion include national and transnational policy documents on central bank stress testing, commentaries in the financial press, and reports by industry groups and think tanks. Finally, a review of the archive (1997–2018) of the banking sector's premier trade journal, *Risk*, was conducted to understand the historical background of the technique.

TABLE 1 Study's 22 interviewees (2016–19) listed by professional category

Current and former regulators at Bank of England and European regulatory institutions	8
Current or former risk, stress testing and treasury managers at UK banks	8
Financial "quants" and software engineers	3
Other stakeholders (consultant, lobbyist, academic)	3

4 | PERFORMATIVE MEASUREMENT

As the sociology of testing has observed, while scientists and the public at large see testing as experimental if it improves the accuracy of knowledge, it is possible to understand testing as socially productive without abiding by realist commitments—testing can be seen as producing useful knowledge even if one suspends judgment on whether the "truth" is being approached asymptotically. In his analysis of the US Treasury's 2009 Supervisory Capital Assessment Programme, Langley (2013) makes just that move: he interprets the test's success in delivering a "positive affective charge" to markets during the turmoil of the financial crisis as a function of its perceived "precision" rather than its supposed accuracy. This section goes further by introducing a Goffmanian dimension to the analysis (Goffman, 1959). It argues that while the central bank's pass/fail judgments play a performative role in potentiating supervisory interventions they also supress uncertainty about the accuracy of the measurement on the experimental back stage of the process.

Although the origins of stress testing tend to be associated with state-directed interventions during the financial crisis (Geithner, 2014), it has for a long time been a specialist, subaltern risk management technique in the financial sector. For example, in the absence of recent experiences of nationwide mortgage defaults in the United States, in the 1980s mortgage-backed securitizations were stress tested against the macroeconomic conditions of the Great Depression to estimate their risk (MacKenzie, 2011). Stress testing became more prominent in banking in the 1990s and 2000s when transnational capital regulations insisted that banks' use the method alongside price-based statistical techniques such as "Value-at-Risk" (Izquierdo, 2001; Quagliariello, 2009). However, it was not until the 2007–2009 financial crisis, and the breakdown of banks' risk management systems, that regulators took control of the process in their supervisory interventions. Rather than letting each bank create its own scenarios, regulators crafted a common scenario and evaluated the results comparatively across the banking system, allowing them to direct state recapitalization efforts and release useful information to market participants.

Given the perceived success of these exercises, the procedure was rapidly instituted as a regularized fixture of post-crisis international banking supervision. There are procedural differences between different countries' testing regimes, but a test invariably begins when the central bank releases its annual scenario—a spreadsheet of severely adverse variables projected 3–5 years into the future (these typically include macroeconomic indicators such as GDP and unemployment, as well as financial variables such as inflation and interest rates). In the Bank's program, responsibility then passes to the commercial banks to expand these variables into a more detailed macroeconomic projection, estimate the probability that their loans will default and the losses incurred, and calculate the impact on their capital. The results are then returned to the central bank along with a detailed narrative account justifying their assumptions, choice of models and governance process (the "qualitative" component of the test). In the event that a bank dips below the regulatory capital minimum in the scenario simulation, the central bank will likely require it to raise more capital or impose restrictions on its dividend payments and share buybacks (what the Bank euphemistically calls "resubmitting the capital plan").

The stated aim of the tests is to assess banks' "capital adequacy" by exposing them to "severe but plausible" hypothetical crisis scenarios and bringing their capitalization levels up to a sufficiently resilient standard (Bank of England, 2015). While the results of a decade of central bank led testing have yet to be put to the test in an actual crisis, the history of the technique does not impart great confidence in the accuracy of the tests' results. Critics have noted that there is no shortage of examples of where stress tests painted a flattering picture of the strength of financial institutions just prior to their collapse (Dowd, 2015a, 2015b). Indeed, like all forms of financial risk management there are serious questions to be asked about their accuracy (Millo & MacKenzie, 2009). The insight holds particularly in respect to commercial banks' own pre-crisis stress tests, which by all accounts were remarkably untesting and failed to leave a meaningful impression on their risk management or conduct (Interview, September 13, 2016; Risk, 2006). Yet if the notion of "experimentation" is unmoored from concerns with accuracy and attached to a performative understanding of acts of measurement as productive of new realities (Callon, 1998), then the relatively large numbers of banks that have failed the regulatory tests over the years[2] suggests that the procedure can be considered experimental—they led to supervisory actions that would not otherwise have been possible. There is also evidence that the tests have furnished knowledge enabling major interventions by policymakers. For example, according to the testimony of Andrew Bailey, then Deputy Governor of the Bank of England, the plug was pulled on a planned deal between the Co-operative Bank and Lloyds in 2013 only after the results of the stress test came in (Interview, November 22, 2016).

Another important consideration is how regulators and financial practitioners understand the knowledge produced by the tests. While there is still a tendency in public discourse and certain sociological quarters to see financiers as "model dopes" (MacKenzie & Spears, 2014, p. 419) convinced that they can transform uncertainty into tidy quanta of risk, heightened concern with "model risk" in the aftermath of the financial crisis has meant that many regulators and practitioners have internalized a more pragmatic stance. As a stress testing manager at a major UK bank put it: "Whether something is 10 or 20, it doesn't really matter… the one thing about stress testing and forecasting, the answer is wrong. It won't ever be that" (Interview, November 22, 2016). Stress tests may not provide the "truth" of how a bank would really fare in a doomsday scenario, but they do "tell you the direction where things move, how things are interacting, where risk may come about. And that's the important bit" (Interview, November 22, 2016).

From the regulatory standpoint, the tests are valuable because they "create a lot of information for the banks themselves and for their supervisors" (Interview, December 14, 2016). Or as another regulatory interviewee put it: "None of the supervisors care that much about the headline number that comes out of the exercise. It's getting the firms to do the analysis which is interesting" (Interview, November 22, 2016). That might be an exaggeration given that it is the "headline number" which determines whether banks pass or fail the tests. But it does point to how, for the actors involved in the testing process, belief in the accuracy of the results is not a prerequisite for seeing the exercises as productive experiments. In addition to serving a performative role in legitimating supervisory interventions, the pass/fail judgments provide the frontstage "expressive equipment" (Goffman, 1959, p. 22) of the experimental process: imparting a sense of accuracy and certainty that specialists, aware of the complex backstage apparatus of supervisory requirements and calculations, tend to be skeptical about.

5 | MACROPRUDENTIAL EXPERIMENTATION

We have seen that the measurement of capital adequacy is interpreted more critically on backstage than it is presented on the public frontstage. However, there is an even deeper and hard-to-penetrate backstage of the stress tests reserved for a "core set" (Collins, 1988) of technical experts. As the Bank states when describing the purposes of their program, it "is not solely about calculating estimates of bank capital in the adverse scenario. Rather, it represents a set of tools that allows policymakers to explore and better understand the vulnerabilities of the financial system" (Bank of England, 2015, p. 9). In that exploratory capacity, stress tests conform closely to the understanding of regulatory science as pushing beyond the cognitive scope of regular research science (Irwin et al., 1997; Jasanoff, 2011a). What distinguishes "macroprudential" applications of stress tests are just how far to the outer edge of regulatory knowledge they skirt. With these functions lending a potentially high level of discretionary power to the central bank, this section argues that the Bank's decision to frame their stress tests as "predictable" can be understood as a performance of accountability intended to help establish the tests' legitimacy in the eyes of financial practitioners.

For almost a decade before the crisis, thinkers at the Bank for International Settlements (BIS) in Basel, Switzerland (the central bank for central banks) had been sounding the alarm about the limits of "microprudential" supervisory approaches (Borio, 2003; Clement, 2010). Instead of focusing regulatory efforts on ensuring the sound management of individual banking institutions, they instead proposed a "macroprudential" approach for safeguarding financial stability. In the aftermath of the crisis, with policy entrepreneurs from the BIS exerting greater influence, this led to an "ideational shift" (Baker, 2013) that invited central banks to adopt a "god's eye view" of finance, encouraging them to assume a more "commanding stance" (Dorn, 2016, p. 1). How this ambition should be put into practice was, however, left open-ended and indeterminate (Stellinga, 2019). While initially articulated in bold anti-cyclical terms, where regulators would seek to smooth the credit cycle and deflate speculative bubbles, in many jurisdictions this gave way to the less ambitious resilience agenda focused on just increasing the quality and quantity of bank's loss absorbing capital (Thiemann, 2019). Uniquely, the Bank of England was able to combine both agendas. After receiving enhanced statutory powers from Parliament, the Bank employed "strategic ambiguity" (Best, 2012; Van Gunten, 2017) and subordinated "anti-cyclical policies to the goal of resilience" (Thiemann, 2019, p. 570). At the same time, the architect of the Bank's new approach to financial stability governance, Paul Tucker, insisted that resilience was non-quantifiable and would be a moving target based solely on the judgment of the newly formed Financial Policy Committee (FPC) (Kynaston, 2017, p. 775). On that basis the Bank was able to present its anti-cyclical policies as motivated by and contributing to the overarching resilience agenda.

The way that stress tests sit at the "junction of macro and micro prudential supervision" (Interview, November 22, 2016) means that they are well placed to perform this goal. One of the ironies of the macroprudential "paradigm shift" is that, while often framed by authorities and understood by scholars as an all-encompassing reorientation towards the systemic level, most of the tools at regulators' disposal remain rather traditional. The increasing space the shift has afforded for blue skies approaches to financial system modeling (such as agent-based and network approaches) has not been matched by broader channels for regulatory intervention (which continue to be operationalized mainly through microprudential supervisory channels). From the central bank's perspective, the appeal of stress testing is that it has "two functions" (Anderson, 2016, p. 11): assisting in the project of macroprudential knowledge creation, while being connected to supervisory interventions for governing individual banks. Hence, from 2016 onwards, the Bank infused macroprudential anti-cyclical rationales into its essentially microprudential stress tests with the introduction of the "Annual Cyclical Scenario" (ACS). The scenarios' severity would now reflect the judgment of the FPC about the current position in the financial cycle (Bank of England, 2015, p. 5). Further, the tests could "inform" the setting of countercyclical capital buffers, in which minimum capital requirements would be notched up and down (from 0% to 2.5%). So, while from 2016 the tests as a whole were given an anti-cyclical justification, the countercyclical buffer would allow the FPC, who meet on a quarterly basis, to make immediate interventions in response to emerging threats.

Given the novelty of the countercyclical buffer, private sector interviewees were unsure about its merits. At one institution, though, an individual responsible for a large UK bank feared that rather than reflecting objectively the position in the financial cycle the buffer could just reflect whatever the Financial Policy Committee "has got a bee in its bonnet about" (Interview, December 14, 2016). This interviewee was concerned that the buffer is used "in the way that it's presented in its policy statement to Parliament… and not to just have unlimited power to impose huge capital surcharges on banks at the drop of a hat" (Interview, December 14, 2016). These remarks reflect a more widespread unease, expressed in feedback to the Bank's discussion paper on its stress testing program (Bank of England, 2014), that the Committee's power to judge the position in the financial cycle might stray into unacceptably discretionary territory, providing regulators with an "unfettered ability to control individual and system-wide bank capital ratios" (Bank of England, 2014, p. 7). That reflects unease at the political level, expressed as far back as 2011, with the discretion granted to the FPC with its macroprudential mandate (Baker, 2013).

The Bank's decision to frame its stress tests from 2016 onwards as "predictable" (Bank of England, 2015, p. 25) should be understood in this context. The notion of predictability is well established in modern monetary policy and is integral to the idea of governing interest rates by shaping market expectations (Blinder, 2004; Holmes, 2014; Wansleben, 2018). In much the same way that market participants are supposed to be able to anticipate movements in interest rates by attending closely to policymakers' signals, the idea is that market participants can predict developments in the stress scenarios' severity and decisions about the use of the countercyclical buffer by consulting the FPC's financial stability reports (Brazier, 2016, p. 76). In the words of a regulatory interviewee: "We know that it takes time for firms to build capital. So to the extent that we can make the scenarios more systematic and predictable, it can help firms with that" (Interview, August 23, 2016). Yet given the common association of stress testing with “thinking the unthinkable” and criticism that the tests have become predictable over the years (Glasserman & Tangirala, 2015), it is surprising that the Bank adopted this language. To be clear, the predictability that the Bank publicly endorses does not necessarily imply a zero failure rate regime. However, as the next section will show, this framing is not unrelated to the sociotechnical stage management of the tests that produces these results.

What explains the Bank’s adoption of the potentially confusing notion of “predictable” stress tests? This paper cannot present a definitive account, but the sociology of public knowledge provides clues when it emphasizes how sensitive regulators and regulated are to perceived transgressions of objectivity and accountability (Jasanoff, 2011a). To be accused of abusing their discretionary license and pulled into public controversy is a point of acute sensitivity for central banks who cherish their operational independence and justify it based on their value-neutral expertise (Conti-Brown, 2016). Those fears are heightened in the case of the Bank’s macroprudential regulation, given how far it pushes beyond the economic knowledge sanctioned by academic research communities (Thiemann, 2019; Thiemann, Aldegwy, & Ibrocevic, 2017). The financial cycle that the Bank’s scenarios and countercyclical buffer are anchored to remains an uncertain theoretical postulate, contingent to a large extent upon the measurement device applied to financial and macroeconomic data (Kranke & Yarrow, 2019; Stellinga, 2019). A regulator at the Bank conceded: “there’s never going to be a single financial cycle, it depends on what approach you take to try and estimate it” (Interview, March 15, 2019). Given the degree of uncertainty and the level of regulatory discretion permitted by the macroprudential policy mandate, it is therefore credible to see the Bank’s embrace of predictability as intended to reassure financial market participants that their discretionary license will not be exercised liberally. Accountability is performed to their private sector audience in a rhetorical framing of stress tests at odds with common understandings of the practice.

6 | THREE ACTS OF STAGE MANAGEMENT

The previous sections detailed two experimental applications of stress tests. First, as a performative measurement technique potentiating supervisory interventions. Second, as a vehicle for pursuing macroprudential policies at the outer limits of the central bank’s regulatory science. This section addresses the stage management techniques employed by the central bank so that these experiments enhance rather than undermine “public confidence in the banking system” (Bank of England, 2015, p. 9). While the goal of boosting confidence seems in tension with the idea of an authentic experiment—even implying a propagandistic imperative which undermines the tests’ objectivity (Dowd, 2015b)—the ambiguity of the measurement standards creates room for manouver. For as the Bank insists, there is no objective standard of resilience than can be fixed and quantified. Neither is there a “correct” level at which stress scenario severity would be sufficiently testing of banks’ resilience: “severe but plausible” is vague enough to encompass stress events from the mundane to the extreme. The central bank thus has considerable latitude to employ calculative and procedural techniques to manage the testing process and to attempt to reconcile its experimental and demonstrative goals.

6.1 | Scenario modelling and impact anticipation

A striking aspect of the Bank’s tests of capital adequacy is the extent to which the results fall within a narrow range. Some banks come close to failing, or fail marginally, but there are very few instances of banks failing spectacularly or cruising through the test unscathed. This is surprising: one might expect that tests exploring “severe but plausible” doomsday scenarios would produce volatile results. Even with the embrace of “predictability,” the consistent performance of commercial banks in the exercises requires explanation. To understand this regularity requires stepping back stage. Years before the launch of the Bank’s public stress testing programme the Bank developed their own “top down” stress testing model called the Risk Assessment Model of Systemic Institutions (RAMSI). In contrast to “bottom up” tests, where banks use their own models and utilize their detailed balance sheet data when simulating the scenario, RAMSI makes use of banks’ publicly available balance sheet data (Burrows, Learmonth, & McKeown, 2012). While reliance on this public data reduces the granularity and precision of the “top down” stress test, it does not blunt RAMSI’s usefulness for shaping the testing procedure. First, at the preparatory stage, when experimenting with potential scenarios. Second, to anticipate the chosen scenario’s impact on individual banks’ balance sheets. Third, to check that banks’ results do not divert too far from what the Bank expected them to be.

RAMSI is therefore a vital technology for the Bank to fine tune the choice and “calibration” of its scenarios in meeting its policy objectives. One of these objectives involves setting the scenarios’ severity at a level that reflects the FPC’s judgment about the current position in the financial cycle. Another objective revealed by this research involves pushing banks to the edge of failure without actually failing them. A regulatory interviewee close to the scenario design process put it this way:

> *So I think there is a genuine emphasis on, sort of, almost trying to get the right scenario, which ultimately is a scenario, in the [Financial Policy] Committee’s judgement, they’d be uncomfortable if banks failed with it… They are sort of saying, if there is a scenario much worse than this, you know, you might expect some banks to fail. We are not a zero-failure regime, but they are trying to get to that level. (Interview, August 23, 2016)*

Since the differences between the projections produced by RAMSI and banks’ own simulations are of a degree rather than of a kind, the Bank can attach a probability to a bank failing the test prior to a bank conducting its own test. An interviewee at a large UK institution revealed that one of their primary interests prior to running a stress test is in the Bank informing them of the likelihood of a “capital shortfall actually materialising” (Interview, December 14, 2016). By this, the interviewee does not mean the probability of an extreme event actually occurring, but the probability of their bank failing the stress test and being asked to raise more capital. These comments suggest that the trend of banks passing the stress test is a policy choice deliberately enacted through the central bank’s modelling technologies. A stress testing manager at a large UK bank confirmed this intuition: “frankly no one is going to fail the stress test from a quantitative perspective… [with] the type of scenario the PRA [Bank of England’s Prudential Regulation Authority] is using” (Interview, November 22, 2016).

6.2 | Foreclosing weakness and demonstrating strength

For obvious reasons, banks are not keen on failing the stress test: it can result in them having to raise more capital and carries “reputational risks” that the market can punish them for (in share price devaluations). Central banks also have an interest in a low number of banks failing their stress tests; they are attuned to the reactive nature of markets, where an authoritative declaration of a bank’s weakness can turn into a self-fulfilling prophesy if depositors withdraw their money or funding markets turn away their business (Interview, October 30, 2016). Central banks are sensitive to the ways in which bank failures might reflect poorly upon their supervision: if a bank fails a test, questions might be asked about why they passed it in previous years and what has gone wrong in the interim. Finally, regulators are receptive to the opinions of the banks themselves, who do not like the pass-fail judgments and lobbied against them.

One option for the central bank to prevent failures is just to stop issuing these judgments (as did the European Banking Authority in 2017). The Bank’s approach, in keeping with the preference of the British Bankers’ Association (BBA and BSA, 2014, p. 14), has instead been to imply failure but to avoid using the word explicitly. Whereas in the Federal Reserve’s reports “objections to the capital plan” is a proxy for “failure,” there is no clear equivalent in the Bank's reports, which only identify which banks were required to resubmit their capital plan or raise more capital. Further ambiguity derives from the lag between the deadline for banks returning their stress test results to the central bank in the spring and the central bank's publication of the results in the winter. In the interim, there is ample time, if a bank performed poorly, for it to raise more capital and for the central bank to spin the results into a more positive narrative. One regulatory interviewee insisted that "you want to be in control of the messaging to the media" (Interview, October 30, 2016). Another described this as "supervisory actions behind the scenes", citing the case of Standard Chartered's poor performance in the 2015 test:

> *So Standard Chartered last year [2015] were a bit too close to the edge… So in the gap between July and December they went and raised capital. When the report came out, you could get the story: Standard Chartered didn't have quite enough capital at the start of the process, but they have now, which is a nice outcome for the central bank… you don't really want people going: "oh my god, there's a problem with a big bank". You know that's kind of the exact opposite of what you're meant to be doing as a regulator. (Interview, November 22, 2016)*

An interviewee at another institution agreed that these supervisory actions should be the story rather than talk of passing or failing the test. Going further, they interpreted these actions as meaning that such outcomes have already been superseded within the Bank's framework.

> *We don't like the fail word; it's not about failure or not failure; it's about what information the stress test tells us and therefore what action the individual bank is taking with the supervisor… so in that sense you could say that we don't have a hurdle rate [pass/fail threshold] … we have, you know, being asked to resubmit the capital plan. (Interview, December 14, 2016)*

This resistance to the pass/fail judgments was echoed by a banking executive who complained that "as soon as you start to label pass or fail, that's a bit of a red flag…how can you pass or fail a theoretical question about a theoretical scenario about something that may or may not happen" (Interview, March 15, 2019). The trend of increasingly few banks failing the Bank of England's stress tests is therefore not just because the scenarios are designed to deliver that result. It is also a product of a retroactive procedural technique in which the central bank can emphasize the results from the perspective of the supervisory actions that have since improved a bank's capitalization.

6.3 | "A little extra loop": Applying a gloss of objectivity

Within the Bank's stress testing program the countercyclical buffer operates as the rapid response vehicle for the tests' general anti-cyclical framework. One of its distinguishing features is that unlike the stress scenario's severity (which is determined by the judgment of the FPC), the decision on whether to apply the buffer is supposed to be "informed" by the results of the stress test. Given that the scenarios are being designed in such a way that few banks fail the test, how can the tests furnish knowledge which helps policymakers decide at what level to set the countercyclical buffer? An individual responsible for a large UK bank doubted they could:

> *It still seems a little bit circular to me in that it seems to start from a perspective of saying: well we think that given where we are in the financial cycle we're about a third of the way up and so banks should be holding about 1% of the counter-cyclical buffer, let's design the scenario that stresses the banks to the point where the expected result is that they will have to hold a 1% counter-cyclical buffer. (Interview, December 14, 2016)*

Reflecting on the circularity and seeming redundancy of the procedure, the interviewee further argued that "inserting this stress test piece, it's like a little extra loop… the same judgement would be there but instead of going straight from a to b, you a – stress test loop – b" (Interview, December 14, 2016). This raises the question of whether "the stress test is actually telling you anything new, or is it telling you that yes it was possible for you to make the data fit what you originally believed to be the truth" (Interview, December 14, 2016). Another banking executive saw the reasons given for the application of the countercyclical buffer as cynically ambiguous. The result is that the regulator can say about the stress test's role in setting the countercyclical buffer: "'oh it's just informing us, we're just using this as a judgement, oh it's okay'. They just hide behind it [the ambiguity] because that gives them carte blanche to do anything they want to do" (Interview, March 15, 2019).

If these insinuations are true, they suggest that in setting the countercyclical buffer the stress test simply provides a gloss of objectivity to the subjective judgment of the Financial Policy Committee. Or in the terms offered by the sociology of organizations, the testing device, on this account, serves a ceremonial function is sustaining the myth that the central bank's decision making is grounded in fully rationalized technoscientific knowledge (Meyer & Rowan, 1977). As scholars have noted, such performative scientism is common in central banking (Abolafia, 2012; Braun, 2015; Marcussen, 2009; Mudge & Vauchez, 2016). While in their operations "custom, compromise, and pragmatism mix with expert judgement," the resulting confusion and uncertainty is "strategically obscured from public view" (Abolafia, 2012, p. 94). At the same time, this insight does not mean that the central bank is harboring ulterior rationales for applying the countercyclical buffer. If the financial cycle is as an artefact of measurement devices and expert judgment, then the stress test can be understood as a performative frontstage device which legitimates the central bank's interventions.

7 | A PREDICTABLE ACT OF PUBLIC THEATRE?

The previous sections showed that the Bank has ample tools at its disposal to stage manage the testing process, and this is responsible, at least in part, for enacting the policy decision to have few banks fail the test. How are these increasingly predictable demonstrations of financial stability being received by the public? It is not possible to ascertain empirically what the public thinks of stress tests given that there have been no surveys of public opinion. Hence, the arguments which follow are somewhat speculative in nature. But by building on public commentaries that criticize the tests as a manipulative performance, it is possible to conjecture reasonably that recent developments may threaten the tests' legitimacy in the public's eyes.

Andrew Haldane, Chief Economist at the Bank of England, argues: "Trust is the lifeblood of all things monetary and financial, including central banks. And incredulity is Kryptonite" (Haldane, 2017, p. 2). For that reason, the years following the 2007–2009 financial crisis have been testing for central banks. Central banks were widely held to have been either complicit in the financial excesses of the preceding decade or guilty of fatal complacency (Johnson & Kwak, 2010). That had led commentators to argue that public trust in central banks has collapsed, with these institutions finding themselves on the receiving end of populist provocations, challenges to their independence, and skepticism about their policy commitments (Goodhart & Lastra, 2018; Riles, 2018; Tucker, 2018). In this context, does a zero failure rate in recent stress tests, and the implication that the financial system is resilient to any plausible scenario, risk undermining the credibility of the procedure?

Critical commentaries suggest ambient skepticism about the tests among experts in the public sphere. For example, a report by the US Office for Financial Research on the Federal Reserve's testing program found a remarkable degree of consistency in bank losses year on year, suggesting "increasing predictability over time" (Glasserman & Tangirala, 2015, p. 16). The authors worry that this could lower the informational value of the results and "may lead to pressures to weaken the process, given the costs involved in its implementation" (Glasserman & Tangirala, 2015, p. 19). Picking up on the report, the *Financial Times* columnist Gillian Tett goes further, using the occasion to ask whether the tests have "outlived their usefulness" (Tett, 2015, n.p.). Tett proposes a "cynical view" and asks her readership to "accept that the real value of the tests is as a piece of public theatre" (Tett, 2015, n.p.). In a series of influential reports for the Adam Smith Institute, Kevin Dowd casts doubt on the Bank of England's stress testing program for similar reasons: "even if the Bank had severe doubts about the strength of the banking system, it cannot admit to them… The stress tests cannot then be credible, because only a reassuring answer can ever be allowed" (Dowd, 2015b, p. 27). Finally, the public interest advocacy group, Finance Watch, has accused the Eurozone stress tests of "trying too hard to reassure the public" and argued that this undermined their intention to "dispel lingering doubts and restore trust" (Stiefmüller, 2018, n.p.).

Even if these expert concerns do not diffuse to the wider public sphere there is still the risk that the tests' lack of drama and surprise will cause the tests to recede from public view. As Jasanoff argues, in order to maintain their hold on the collective imagination "the utility of the state's knowledge producing endeavours must repeatedly be brought home" (Jasanoff, 2005, p. 248). Given the high-stakes political economy of banking regulation, that observation has particular saliency. It may be the case that we have reached the point at which central banks wish to disassociate stress testing with crisis governance and establish a new form of "peace time" stress testing (Schuermann, 2016). It may also be the case that banks are now as capitalized as regulators wish to see them and so continued failures would lack an obvious supervisory response. Nevertheless, it is possible to imagine reconfigurations of the practice that would not see it converted into an uneventful ritual. For example, banks could have their response to the scenarios awarded a positional grade on a rank from strongest to weakest. Or banks could be failed on a broader set of criteria than just how well their capital stands up in the scenario.

In the absence of such creative reconfigurations, central bank stress tests that just involve a routine confirmation of the banking sector's health seem unlikely to maintain public interest. A consultant for the banking industry argues: "There becomes a danger that these things become routine and they become boring and dull and they're just done for the sake of doing it" (Interview, March 15, 2019). From this perspective, predictability imperils the communicative function of the tests:

> *Behind the scenes, all of those rules, it's all quite technical. In actual fact, the man in the street doesn't really get that, doesn't really understand that… so what the regulators have been doing is that stress testing for me is a more political animal which they can more clearly articulate and by then putting a pass or failure to it, they can really turn the screw. (Interview, March 15, 2019).*

Given that capital regulation has a large effect on banks' share prices (Admati & Hellwig, 2013; Goodhart, 2015), it is no small matter for the regulator to "turn the screw." What happens when the screw is released? One potential future is for stress tests to be progressively watered down to accommodate the preferences of financial sector actors. Indeed, such changes are already afoot. In 2019 the Federal Reserve announced that fewer banks will be subject to the qualitative aspects of its testing programme and that it will be dropping pass/fail judgments on this component of the tests. The Bank's stress tests are framed more ambiguously, making it harder to identify specific changes which reveal a loosening of supervisory stringency. But the decision in 2018 to repeat the same scenario as in 2017 confirms the increasing emphasis placed on predictability rather than testing banks with new and challenging crisis scenarios (Bank of England, 2018).

Albeit inconclusive, these closing reflections point to the need for the sociology of testing to incorporate large-scale interpretive frames into their analyses of regulatory performances. This paper only speculates about how the historical and political context in which stress tests are performed shape their reception and legitimacy. But in a time in which public trials and experiments are becoming an increasingly ubiquitous governmental technology (Adkins & Ylöstalo, 2018), there is an opportunity for sociologists to add a further layer to their research by studying under what conditions tests become a legitimate and credible basis for policymaking. To some extent this recommendation is anticipated by Pinch (1993). He makes a programmatic call for sociology of testing to not just document the interpretive flexibility of different social actors, or the epistemic cleavages between experts and publics, but to situate tests within their broader social and political contexts. If this study has shown one thing it is that there are analytically challenging but revealing entanglements between the technical details of tests and the audiences to which they are performed.

8 | CONCLUSION

Combining Goffman's notion of social performances with an understanding of the performativity of economic measurement devices, this article finds that central bank stress tests are sociotechnically stage managed to control their contingency and demonstrate their objectivity. The Bank of England's decision to make its annual tests "predictable" relies on an existing apparatus of calculative and procedural techniques that were already utilized when the central bank was more willing to fail commercial banks and make high profile supervisory interventions. At the same time, these observations are not meant to imply that stress tests are nothing more than a performance, in the colloquial sense of the term. Even though the regulatory stress tests have become uneventful on their frontstage, it is possible that the data collected in the exercises is proving useful for backstage macroprudential experimentation.

And yet, tests of this scale do not exist in a social vacuum. This paper concludes by asking whether stress tests, which are immensely time consuming and expensive to administer, can survive their transformation into an uneventful ritual. Already the regulatory pendulum is beginning to swing back in a deregulatory direction. Does the lack of drama and surprise in recent stress tests risk losing the level of public attention necessary to uphold supervisory stringency in the face of a pushback by powerful and well-funded financial actors? These dynamics might seem closer to the concerns of political economists than to the preoccupations of the science studies scholars associated with the sociology of testing. But if the sociology of testing is to grapple with the proliferation of trials and experiments across the governmental spectrum then tackling such questions is unavoidable. If scholars can go one step further and provide compelling answers, then a revived sociology of testing might move from being a sub-field of the sociology of scientific knowledge to the forefront of economic and political sociology.

ACKNOWLEDGEMENTS

This article has benefited from feedback by Benjamin Braun, Dylan Cassar, David Stark, Bart Stellinga, Matthias Thiemann, Arjen van der Heide, and Timo Walter as well as the participants of the workshop "Put to the Test" workshop held in London, December 2018 and the panel "Financial Governance II" at Intersections of Finance and Society, London, December 2019. I also thank the anonymous reviewers, whose incisive comments helped to improve the manuscript. Some of the interviews this paper draws from were funded by a grant from the Leverhulme Trust [ECF-2014-304].

DATA AVAILABILITY STATEMENT

Author elects to not share data. Research data are not shared.

NOTES

[1] The European Central Bank has taken a greater role in the process since introduction of the Single Supervisory Mechanism in 2014.

[2] Banks that have failed stress tests in the United States and United Kingdom include some of the countries' largest institutions: Barclays (2016), Citigroup (2012, 2014), the Co-operative Bank (2014), Deutsche Bank (2015, 2016, 2018), HSBC (2014), and the Royal Bank of Scotland (2014, 2015, 2016), among others.

REFERENCES

Abolafia, M. Y. (2012). Central banking and the triumph of technical rationality. In K. Knorr Cetina & A. Preda (Eds.), *The Oxford handbook of the sociology of finance* (pp. 94–112). Oxford: Oxford University Press.

Adkins, L., & Ylöstalo, H. (2018). Experimental policy, price and the provocative state. *Distinktion: Journal of Social Theory*, *19*(2), 152–169.

Admati, A., & Hellwig, M. (2013). *The bankers' new clothes*. Princeton and Oxford: Princeton University Press.

Anderson, R. W. (2016). Stress testing and macroprudential regulation: A transatlantic assessment. In R. W. Anderson (Ed.), *Stress testing and macroprudential regulation: A transatlantic assessment* (pp. 1–30). London: CEPR Press.

Baker, A. (2013). The new political economy of the macroprudential ideational shift. *New Political Economy*, *18*(1), 112–139.

Bank of England. (2014). *Summary of feedback received on the stress testing discussion paper*. London: Bank of England.

Bank of England. (2015). *The Bank of England's approach to stress testing the UK banking system*. London: Bank of England.

Bank of England. (2016). *Stress testing the UK banking system: 2016 results*. London: Bank of England.

Bank of England. (2018). *EU withdrawal scenarios and monetary and financial stability: A response to the House of Commons Treasury Committee*. London: Bank of England.

BBA (British Bankers' Association), and BSA (Building Societies Association). (2014). Response to: "A framework for stress testing the UK banking system". A discussion paper - Bank of England (BoE) - October 2013. London.

Beckert, J. (2016). *Imagined futures: Fictional expectations and capitalist dynamics*. Cambridge, MA and London: Harvard University Press.

Beckert, J., & Bronk, R. (Eds.) (2018). *Uncertain futures: Imaginaries, narratives, and calculation in the economy*. Oxford: Oxford University Press.

Best, J. (2012). Bureaucratic ambiguity. *Economy and Society*, *41*(1), 84–106.

Bijker, W. E., Bal, R., & Hendriks, R. (2009). *The paradox of scientific authority: The role of scientific advice in democracies*. Cambridge, MA and London: MIT Press.

Blinder, A. (2004). *The quiet revolution: Central banking goes modern*. New Haven and London: Yale University Press.

Borio, C. (2003). *Towards a macroprudential framework for financial supervision and regulation? BIS working papers 128*. Basel: Bank for International Settlements.

Braun, B. (2015). Governing the future: The European Central Bank's expectation management during the great moderation. *Economy and Society*, *44*(3), 367–391.

Braun, B. (2016). Speaking to the people? Money, trust, and central bank legitimacy in the age of quantitative easing. *Review of International Political Economy*, *23*(6), 1064–1092.

Brazier, A. (2016). The Bank of England's approach to stress testing the UK banking system. In R. W. Anderson (Ed.), *Stress testing and macroprudential regulation: A transatlantic assessment* (pp. 69–84). London: CEPR Press.

Burrows, O., Learmonth, D., & McKeown, J. (2012). *RAMSI: A top-down stress-testing model*. Financial Stability Paper 17. London: Bank of England.

Callon, M. (1998). Introduction: The embeddedness of economic markets within economic theory. In M. Callon (Ed.), *The laws of the markets* (pp. 1–57). Oxford and Malden, MA: Blackwell Publishers.

Callon, M., Millo, Y., & Muniesa, F. (Eds.). (2007). *Market devices*. Oxford: Blackwell.

Clement, P. (2010, March). The term 'Macroprudential': Origins and evolution. *BIS Quarterly Review*, 59–67. Retrieved from: https://www.bis.org/publ/qtrpdf/r_qt1003h.pdf

Collins, H. M. (1987). Certainty and the public understanding of science: Science on television. *Social Studies of Science*, *17*(4), 689–713.

Collins, H. M. (1988). Public experiments and displays of virtuosity: The core-set revisited. *Social Studies of Science*, *18*(4), 725–748.

Collins, H. M., & Pinch, T. J. (1979). The construction of the paranormal: Nothing unscientific is happening. *The Sociological Review*, *27*(1_suppl), 237–270.

Conti-Brown, P. (2016). *The power and independence of the Federal Reserve*. Princeton, NJ: Princeton University Press.

Coombs, N. (2016). What is an algorithm? Financial regulation in the era of high-frequency trading. *Economy and Society*, *42*(5), 1–25.

Coombs, N. (2017). Macroprudential versus monetary blueprints for financial reform. *Journal of Cultural Economy*, *10*(2), 207–216.

Coombs, N. (2020). Financial regulation. In C. Borch & R. Wosnitzer (Eds.), *Routledge handbook to critical finance studies*. Abingdon: Routledge.

Dietsch, P., Claveau, F., & Fontan, C. (2018). *Do central banks serve the people?* Cambridge: Polity Press.

Dorn, N. (2016). Introduction: Questions asked. In N. Dorn (Ed.), *Controlling capital: Public and private regulation of financial markets* (pp. 1–18). Abingdon: Routledge.

Dowd, K. (2015a). Central bank stress tests: Mad, bad, and dangerous. *Cato Journal*, *35*(3), 507–524.

Dowd, K. (2015b). *No stress I: The flaws in the bank of England's stress testing programme*. London: Adam Smith Research Trust.

Evans, R. (1997). Soothsaying or science?: Falsification, uncertainty and social change in macroeconomic modelling. *Social Studies of Science*, *27*(3), 395–438.

Geithner, T. F. (2014). *Stress test: Reflections on financial crises*. London: Random House Business Books.

Glasserman, P., & Tangirala, G. (2015). *Are the Federal Reserve's stress test results predictable?* Working Paper Series 15–02. Washington, DC: Office of Financial Research.

Goffman, E. (1959). *The presentation of the self in everyday life*. New York: Anchor.

Goodhart, C., & Lastra, R. (2018). Populism and central bank independence. *Open Economies Review*, *29*(1), 49–68.

Goodhart, L. M. (2015). Brave new world? Macro-prudential policy and the new political economy of the federal reserve. *Review of International Political Economy*, *22*(2), 280–310.

Haldane, A. G. (2017). *A little more conversation, a little less action: Speech given at Federal Reserve Bank of San Francisco*. London: Bank of England.

Hilgartner, S. (2000). *Science on stage: Expert advice as public drama*. Writing Science. Stanford, CA: Stanford University Press.

Holmes, D. R. (2014). *Economy of words: Communicative imperatives in central banks*. Chicago: University of Chicago Press.

Irwin, A., Rothstein, H., Yearley, S., & McCarthy, E. (1997). Regulatory science: Towards a sociological framework. *Futures*, *29*(1), 17–31.

Izquierdo, A. J. (2001). Reliability at risk: The supervision of financial models as a case study for reflexive economic sociology. *European Societies*, *3*(1), 69–90.

Jasanoff, S. (1987). Contested boundaries in policy-relevant science. *Social Studies of Science*, *17*(2), 195–230.

Jasanoff, S. (1990). *The fifth branch: Science advisors as policymakers*. Cambridge, MA and London: Harvard University Press.

Jasanoff, S. (1995). Procedural choices in regulatory science. *Technology in Society*, *17*(3), 279–293.

Jasanoff, S. (2005). *Designs on nature: Science and democracy in Europe and the United States*. Princeton, NJ: Princeton University Press.

Jasanoff, S. (2011a). The practices of objectivity in regulatory science. In C. Camic, N. Gross, & M. Lamont (Eds.), *Social knowledge in the making* (pp. 307–338). Chicago and London: University of Chicago Press.

Jasanoff, S. (2011b). Cosmopolitan knowledge: Climate science and global civic epistemology. In J. S. Dryzek, R. B. Norgaard, & D. Schlosberg (Eds.), *The Oxford handbook of climate change and society* (pp. 129–143). Oxford: Oxford University Press.

Johnson, S., & Kwak, J. (2010). *13 bankers: The wall street takeover and the next financial meltdown* (1st ed.). New York: Pantheon Books.

Kranke, M., & Yarrow, D. (2019). The global governance of systemic risk: How measurement practices tame macroprudential politics. *New Political Economy*, *24*(6), 816–832.

Kynaston, D. (2017). *Till time's last sand: A history of the Bank of England 1694–2013*. London and New York: Bloomsbury.

Langley, P. (2013). Anticipating uncertainty, reviving risk? On the stress testing of finance in crisis. *Economy and Society*, *42*(1), 51–73.

Langley, P. (2015). *Liquidity lost: The governance of the global financial crisis*. Oxford: Oxford University Press.

Lenglet, M. (2011). Conflicting codes and codings: How algorithmic trading is reshaping financial regulation. *Theory, Culture & Society*, *28*(6), 44–66.

MacKenzie, D. (1990). *Inventing accuracy: A historical sociology of nuclear missile guidance*. Cambridge, MA and London: MIT Press.

MacKenzie, D. (2011). The credit crisis as a problem in the sociology of knowledge. *American Journal of Sociology*, *116*(6), 1778–1841.

MacKenzie, D., & Millo, Y. (2003). Constructing a market, performing theory: The historical sociology of a financial derivatives exchange. *American Journal of Sociology*, *109*(1), 107–145.

MacKenzie, D., & Spears, T. (2014). 'A device for being able to book P&L': The organizational embedding of the Gaussian copula. *Social Studies of Science*, *44*(3), 418–440.

Marcussen, M. (2009). Scientization of central banking: The politics of a-politicization. In K. Dyson & M. Marcussen (Eds.), *Central banks in the age of the euro* (pp. 373–401). New York: Oxford University Press.

Meyer, J. W., & Rowan, B. (1977). Institutionalized organizations: Formal structure as myth and ceremony. *American Journal of Sociology*, *83*(2), 340–363.

Miller, C. A. (2005). New civic epistemologies of quantification: Making sense of indicators of local and global sustainability. *Science, Technology, & Human Values*, *30*(3), 403–432.

Miller, C. A. (2008). Civic epistemologies: Constituting knowledge and order in political communities. *Sociology Compass*, *2*(6), 1896–1919.

Millo, Y., & MacKenzie, D. (2009). The usefulness of inaccurate models: Towards an understanding of the emergence of financial risk management. *Accounting, Organizations and Society*, *34*(5), 638–653.

Mudge, S. L., & Vauchez, A. (2016). Fielding supranationalism: The European Central Bank as a field effect. *The Sociological Review*, *64*(2_suppl), 146–169.

Pinch, T. (1993). "Testing - one, two, three … testing!": Toward a sociology of testing. *Science, Technology, & Human Values*, *18*(1), 25–41.

Quagliariello, M. (2009). Macroeconomic stress-testing: Definitions and main components. In M. Quagliariello (Ed.), *Stress-testing the banking system: Methodologies and applications* (pp. 18–23). Cambridge: Cambridge University Press.

Riles, A. (2018). *Financial citizenship: Experts, publics, and the politics of central banking*. Ithaca and London: Cornell University Press.

Risk. (2006, November). Banks should improve stress testing, says UK FSA. *Risk*.

Rushefsky, M. E. (1986). *Making cancer policy*. New York: SUNY Press.

Salter, L., Leiss, W., & Levy, E. (1988). *Mandated science: Science and scientists in the making of standards*. Dordrecht, the Netherlands: Springer.

Schuermann, T. (2016). Stress testing in wartime and in peacetime. In R. W. Anderson (Ed.), *Stress testing and macroprudential regulation: A transatlantic assessment* (pp. 125–140). London: CEPR Press.

Stellinga, B. (2019). The open-endedness of macroprudential policy. Endogenous risks as an obstacle to countercyclical financial regulation. *Business and Politics*, 1–28. https://doi.org/10.1017/bap.2019.14

Stiefmüller, C. (2018). Banks stress-tests 2018: Trying too hard to reassure. *Finance Watch*. Retrieved from https://www.finance-watch.org/banks-stress-tests-2018-trying-too-hard-to-reassure/

Tett, G. (2015). Stress tests for banks are a predictable act of public theatre. *Financial Times*. Retrieved from https://www.ft.com/content/e6548700-c1d1-11e4-bd24-00144feab7de

Thiemann, M. (2019). Is resilience enough? The macro-prudential reform agenda and the lacking smoothing of the cycle. *Public Administration*, *97*(3), 561–575.

Thiemann, M., Aldegwy, M., & Ibrocevic, E. (2017). Understanding the shift from micro- to macro-prudential thinking: A discursive network analysis. *Cambridge Journal of Economics*, *42*(4), 935–962.

Tucker, P. (2018). *Unelected power: The quest for legitimacy in central banking and the regulatory state*. Princeton and Oxford: Princeton University Press.

Van Gunten, T. (2017). Washington dissensus: Ambiguity and conflict at the International Monetary Fund. *Socio-Economic Review*, *15*(1), 65–84.

Wansleben, L. (2018). How expectations became governable: Institutional change and the performative power of central banks. *Theory and Society*, *47*, 773–803.

Williams, J. W. (2009). Envisioning financial disorder: Financial surveillance and the securities industry. *Economy and Society*, *38*(3), 460–491.

How to cite this article: Coombs N. What do stress tests test? Experimentation, demonstration, and the sociotechnical performance of regulatory science. *Br J Sociol*. 2020;71:520–536. https://doi.org/10.1111/1468-4446.12739

Received: 13 March 2019 | Revised: 14 August 2019 | Accepted: 14 November 2019

DOI: 10.1111/1468-4446.12730

SPECIAL ISSUE

WILEY

Co-existence or displacement: Do street trials of intelligent vehicles test society?

Noortje Marres

Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, United Kingdom

Correspondence
Noortje Marres, Centre for Interdisciplinary Methodologies, University of Warwick, Coventry, CV4 7AL, United Kingdom.
Email: N.Marres@warwick.ac.uk

Abstract

This paper examines recent street tests of autonomous vehicles (AVs) in the UK and makes the case for an experimental approach in the sociology of intelligent technology. In recent years intelligent vehicle testing has moved from the laboratory to the street, raising the question of whether technology trials equally constitute tests of society. To adequately address this question, I argue, we need to move beyond analytic frameworks developed in 1990s Science and Technology Studies, which stipulated "a social deficit" of both intelligent technology and technology testing. This diagnosis no longer provides an effective starting point for sociological analysis, as real-world tests of intelligent technology explicitly seek to bring social phenomena within the remit of technology testing. I propose that we examine instead whether and how the introduction of intelligent vehicles into the street involves the qualification and re-qualification of relations and dynamics between social actors. I develop this proposal through a discussion of a field study of AV street trials in three cities in the UK—London, Milton Keynes, and Coventry. These urban trials were accompanied by the claim that automotive testing on the open road will enable cars to operate in tune with the social environment, and I show how iterations of street testing undo this proposition and compel its reformulation. Current test designs are limited by their narrow conception of sociality in terms of interaction between cars and other road users. They exclude from consideration the relational capacities of vehicles and human road users alike—their ability to co-exist on the open road. I conclude by making the case for methodological innovation in social studies of intelligent technology: by combining social research and design methods, we can re-purpose real-world test environments in order to elucidate social issues and dynamics raised by intelligent vehicles in society by experimental means, and, possibly, test society.

KEYWORDS

automobility, autonomous vehicles, real-world testing, social studies of testing, sociology of AI, STS

O Public Road
You express me better than I can express myself
You shall be more to me than my poem
Walt Whitman, The Open Road

1 | INTRODUCTION

At least since early 2016, so-called autonomous vehicles, or driverless cars, have been tested on urban roads[1] across the UK, in London, Milton Keynes, Bristol, and Coventry.[2] The most well known of these tests are funded by the UK government, by way of its Centre for Connected and Autonomous Vehicles (CCAV), and among their principal aims is to demonstrate the capacity of intelligent vehicle technologies to operate successfully amidst social complexity, on the open road. As the Department for Transport explained the approach in its 2015 code of practice for such testing:

> *Manufacturers have a responsibility to ensure that highly and fully automated vehicle technologies undergo thorough testing and development before being brought to market. Much of this development can be done in test laboratories or on dedicated test tracks and proving grounds.* However to help ensure that these technologies are capable of safely handing the many varied situations that they may encounter throughout their service life, *it is expected that controlled "real world" testing will also be necessary. Testing of automated vehicle technologies on public roads or in other places should be facilitated while ensuring that this testing is carried out with the minimum practicable risk. (Department for Transport, 2015, my italics)*[3]

As the lead engineer for one of the UK trial projects, Dr. Simon Tong of the Greenwich Automated Transport Environment, or GATEway, project put it more briefly to the *Financial Times* newspaper in the summer of 2017: the aim of the trials is to get "driverless vehicles *to learn how to get along* with city transport" (Wright, 2017, italics mine). Use of such language is revealing, in that it highlights the ambition of these technical projects to attribute social capacities to machines, like "learning to get along," to "socialise" intelligent machines, at the very least, the inclination to invoke such ambitions as part of the public legitimation of these projects. What is more, the encounter between automated vehicles and other road users is not just assumed to be one-way. UK driverless car trials are also poised as an occasion *for road users* to become familiar with these relatively new technologies. As Ian Forbes, Head of UK's

Centre for Connected and Autonomous Vehicles (CCAV) stated: “what is important is that [tests] are taking place in the real world. A crucial part of the development of this technology is allowing people to experience it” (Parliament, House of Lords, 2017).

How should we understand and assess these diverse justifications for testing of intelligent vehicle technology on public roads, in terms of the technical requirements for performance testing of intelligent machines (which must be able to handle “many varied situations during their service life”) and the purportedly public commitment of creating “engaging experiences” for people? Previous studies have noted how on-the-road testing of autonomous vehicles (a) requires new forms of governance in support of “social learning” in addition to “machine learning” (Stilgoe, 2018); (b) gives rise to new types of actor-relations such as that between driverless vehicles and third-party road users (Tennant, Howard, Franks, Bauer, & Stares, 2016); and (c) enlists publics in innovation processes in potentially new, situational ways (Marres, 2019). In this paper, I would like to focus on a more general, or even fundamental, question, namely, *whether and how street trials of intelligent vehicles bring social phenomena within the remit of automotive innovation*. Scholars in the philosophy and sociology of technology have recently proposed that real-world testing of technology “beyond the laboratory” can be characterized as a form of social experimentation (Van der Poel, Asveld, & Mehos, 2017) and as “tests on and in society” (Engels, Wentland, & Pfotenhauer, 2019): these studies view “the introduction of new technology into society … as a learning process in which the consequences of it emerge only gradually” (Van der Poel et al., 2017, p. 1) and as involving “the enrolment of (more or less) well-defined populations as subjects of scientific inquiry and technological testing” (Engels et al., 2019, p. 10). In this paper, I evaluate the proposition that real-world technology trials test society through an empirical analysis of street tests of intelligent vehicles in the UK. I make the case for a re-constructive approach: I propose that the trials in question do not in their current design qualify as societal tests, but have the potential to do so, which, if it is to be realized, requires a modification of test protocols.

To date, UK trials have mostly been engineering-led, and this raises the question of whether and how the test design, methodology, and implementation are capable of operationalizing the aforementioned publicly stated commitments to bring social phenomena—such as situational complexity and the co-existence of diverse users in the mundane environment of the street—within the frame of intelligent technology research and development. Indeed, from a sociological point of view, this would be a highly unexpected, and a truly remarkable feat, as sociological and anthropological studies of artificial intelligence (AI) have long argued that the technologies in question suffer from a “social deficit,” that they are incapable of situated engagement, and attunement to social complexity (Button & Dourish, 1996; Suchman, 1987).

The question of whether and how street trials of autonomous vehicles are capable of operationalizing the stated ambition to test and develop the social capacities of machines, then, has wider relevance for two areas of contemporary sociological enquiry: the sociology of artificial intelligence, and social studies of testing. Regarding the first, it has recently been argued that the current proliferation of automated agents and intelligent machines across society, in the form of fully scripted social media accounts, home assistants and “deep learning” applications in social domains from medicine and policing, has transformed the conditions for sociological engagement with machine intelligence (Castelle, 2018; see also Hildebrandt, 2019). In the 1980s and 1990s, sociologists used to criticize prevalent scientific approaches to the design of artificial intelligence for their limited and/or reductive treatment of social and societal aspects of cognition, behavior, interaction, and life in general (Joerges, 1989; Suchman, 1987; Woolgar, 1985). However, in view of the de facto proliferation of AI applications across society, it has been proposed that the time has come for sociologists to move beyond critiquing the “social deficits” of the scientific representations of AI *and to analyze communication among heterogeneous actors in actually existing, partly automated environments from a sociological perspective* (Bialski, Brunton, & Bunz, 2019; Esposito 2017). This paper takes up this invitation, but with a notable modification: instead of studying communication with machines as if it is happening “in the wild,” as “naturally occurring,” I will approach artificial situations—like intelligent vehicle technology testing—in social environments, as a key object and resource for the sociology of AI: not only do real-world tests serve as a device for the introduction of intelligent machines to society (Van der Poel et al., 2017); they also present sociology with an empirical occasion, where it becomes possible to study the introduction of AI to society as an unfolding event, and to specify its consequences, including possible transformations of the capacities of social actors, the relations between them, and wider social dynamics.

Turning to social studies of technology testing, it should be noted that a similar assumption about the disregard for social aspects of technology in engineering has been operative in these studies, which have been undertaken in Science and Technology Studies and related fields since the 1980s. In the article "Testing—One, Two, Three…Testing! Toward a Sociology of Testing," Pinch made the telling point that "test data are usually thought of as providing access to a purely technical realm" (1993, p. 25). The above characterizations of street tests of intelligent vehicles, however, could be taken to suggest that this limitation to the technical in technology testing is being surmounted in contemporary trials of intelligent technologies, as test representatives repeatedly express their commitment to bring social phenomena like interaction with pedestrians and the experience of people within the frame of the test. Taking seriously this possibility—which is different from confirming the suggestion—I want to propose, means that a sociology of testing must move beyond the binary question of *whether or not* the remit of technology testing can include social phenomena, to engage with the far more open-ended, processual question of *how* social aspects are rendered visible, qualified, surfaced and/or obfuscated at the occasion of the test.

With this broader aim in mind, this paper begins by asking whether the commitments to test and develop the social capacities of intelligent machines was operationalized during the implementation of the CCAV-funded street trials of autonomous vehicles. Based on an analysis of public documents and participant observation conducted in trial sites in London, Coventry, and Milton Keynes, between February 2016 and November 2018, I will conclude that, alas, social phenomena still elude the autonomous vehicle tests in question. However, at the same time, the implementation of autonomous vehicles tests in the UK streets does surface social consequences of intelligent technology. By studying the testing of intelligent machines in social environments as generative events, sociology can make a key contribution to interdisciplinary efforts to elucidate these consequences. I propose that to do this well, we need to engage in methodological innovation, and to illustrate this I end by reporting on a recent interdisciplinary experiment that I undertook with colleagues in the "driver-in-the-loop" simulator at the University of Warwick in 2016. Taking up the design research method of prototyping, this experiment re-purposed street tests of intelligent vehicles to serve the ends of sociological enquiry, demonstrating how the introduction of autonomous machines into the street elicited distinctively social dynamics, such as stigmatization.

2 | THE CASE FOR REAL-WORLD TESTING OF AUTOMOTIVE TECHNOLOGY: "LEARNING FROM UNEXPECTED SITUATIONS"

The stated commitment to bring social aspects of the functioning of intelligent vehicles within the remit of technology testing, during the CCAV-funded trials in the UK between early 2016 and late 2018, was not limited to public statements. It was not just about publicity. Each of the publicly funded trials of autonomous and connected vehicles in Greenwich, Milton Keynes, and Coventry had an explicit focus on assessing the interaction between vehicles, environment, and road users, and indeed, with the wider social environment. Thus, the stated purpose of the Autodrive trials in Milton Keynes and Coventry, which a press release termed "the UK's largest trial to date of connected and autonomous vehicle technology on public roads," was to:

> *explor[e] the benefits of having cars that can "talk" to each other and their surroundings—with connected traffic lights, emergency vehicle warnings and emergency braking alerts. The vehicles rely on sensors to detect traffic, pedestrians and signals but have a human on board to react to emergencies. The trials are testing a number of features and most importantly seeking to investigate how self-driving vehicles interact with other road users. (Tute, 2017)*

This focus on testing the interactional capacities of intelligent vehicles on roads, in turn, is often justified on methodological grounds. The capacity of machines to operate in a social environment is central to the understanding of "intelligence" in intelligent vehicle research and development, where autonomous operation requires vehicles to negotiate unexpected encounters and, as quoted above, "many varied situations." The testing of these vehicles on the street rather than in the laboratory, or in dedicated automotive test sites, tends to justified in reference to precisely these requirements of intelligence. Large-scale street trialling is said to be *the only way* in which the interactional capacities of these machines can be fully assessed and developed. As the *Financial Times* put it succinctly, "Testing in real world conditions is essential for driverless cars to learn from unexpected situations that would be difficult to simulate, such as how humans react to a driverless vehicle" (Campbell & Yang, 2018). One of the defining features of intelligent vehicle technology is its alleged capacity to respond more or less spontaneously to dynamic occurrences in the road environment—we might paraphrase: to operate in a testing social environment. In a promotional video released by Jaguar Land Rover on the occasion of the already mentioned Coventry street trials of self-driving vehicles, a Jaguar Land Rover (JLR) engineer explained:

> *The car is navigating in the urban environment, interacting with other traffic. This will always be the ultimate test for this type of vehicle. We have always had control over the environment and the urban environment is far more unpredictable. There are many more dynamic elements for the car to sense and react to. But we have been using all of this data to refine our systems and make sure that they do deal with them in the correct way. We found a massive challenge in predicting how pedestrians are going to react.*

The focus on interactional capacities of intelligent vehicles has a technological justification: environmentally situated interaction and communication is the next frontier in automotive innovation insofar as this is the next big thing that data-intensive, "learning" computational technology renders cars capable of.[4] Testing intelligent machines in the street is the methodological corollary of this technological claim.

However, the socialization of machines through testing does not just serve the technological optimization of their intelligence, it is equally presented as enabling the repositioning of cars, and the automotive system, more generally, in its relation to society. At the closing event of the Autodrive project in October 2018 in Milton Keynes, where project results were presented by the Autodrive director, the JLR project lead and others, many if not most speakers made reference to the societal benefits that intelligent vehicles would bring: increasing road safety, reducing congestion, mobility as a service, regeneration of regional economies.[5] In an earlier, informal but in some ways more spectacular announcement on Twitter, engineers of Jaguar Land Rover, a partner of the Autodrive consortium, noted that the wider objective in intelligent vehicle development is to make cars "relevant to all demographics" (Figure 1). These stated benefits, to be sure, present the type of justification one would expect from publicly funded engineering projects, but some of them disrupt more customary framings of social actors in the automotive sector. As Lochlann Jain (2004) reminded us, at the start of the previous century the societal introduction of automotive technologies was accompanied by the establishment of individualistic, driver-centric frameworks for accountability, with traffic regulations, insurance policies, and safety procedures biased towards drivers, at the expense of pedestrians in particular. By contrast, recent publicity around intelligent vehicles prominently feature socially defined agents—like busy mums and disabled persons—among the principal beneficiaries of the transition to self-driving cars.

Public presentations of the CCAV trials, and public outreach undertaken as part of the trials, often place social scenarios and narratives in the foreground. The intended effect of situating abstract automotive technologies in real-world social contexts, seems to be to make them feel real, but also to dramatize the transformation of automotive infrastructure by invoking a transformed car culture. Thus, the Greenwich Gateway trials, which ran on the Greenwich peninsula in South London between 2016 and 2018, were accompanied by an exhibition in the

Insider
@insiderwest

Panellists at #InsiderMadeintheUK discussing the importance of making driverless cars relevant to all demographics

FIGURE 1 @Insiderwest, Twitter, September 20, 2016

London Transport Museum developed in collaboration with the Royal College of Art, in which visitors were invited to join in the imagination of "driverless futures," producing visual mappings that feature wishful scenarios such as "I would no longer need to be the driver when my mates go to the pub," "safer streets, even for pet animals," or, "more time to socialise, do fun or useful things on journeys."[6]

To be sure, in the documentation of the GATEway trial there certainly are traces of an individualistic, objectified, non-social framing of on-the-road activity. This project has tested a variety of intelligent mobility technologies including Oxbotica driverless pods equipped with computer vision, pedestrian detection technology, and autonomous steering capacity:

> *The Greenwich experiment is exploring a fundamental question about how autonomous vehicles will fit into city streets. Oxbotica, the Oxford company that developed the vehicle's software, is trying to improve the pods' ability to track people, cyclists and other non-vehicular objects. Performance on city streets will depend on how well they can* navigate around non-mechanised obstacles. *(Wright, 2017, italics mine)*

However, investigation of the human-machine encounter in the Greenwich streets was *not* limited to the machine's steering abilities. A central component of the project was the "observation and surveys of pedestrian interactions" (Fernández-Medina, Delmonte, Jenkins, Holcome, & Kinnear, 2018; McDowell-Naylor, 2018) and sentiment mappings of local attitudes and of the trial itself, labelled "Rate my drive" and "Rate my ride" (Commonplace, 2018), which collected, analyzed, and visualized "real-world" comments on the arrival of driverless cars, mentioning the already operative driverless DLR and the difficulty of negotiating fog and rain in London. The trial protocols were explicitly designed to facilitate encounters with the public, as the end of project report explains:

> *One of the main objectives of the GATEway public trials was to provide open service-like operations where members of the public would be free to "walk up" to the pod stops and use the service. This ensured that the project was engaging different groups of users instead of being limited to only including groups with a particular interest in the technology.*[7]

As socially defined scenarios, agents and locations are so prominently invoked in CCAV trial designs and public presentations, should we infer that society is successfully brought within the frame of intelligent technology testing in these instances? The regimes of justification activated by the CCAV trials suggest that street testing is deployed to locate automotive technology in a social environment not just spatially but ontologically: the tests' focus on the capacity of intelligent vehicles to "co-exist" with others on the road implies a departure from the individualistic ontologies classically associated with driving (Denis & Urry, 2009), towards a more "socially aware" approach to automotive innovation. The "interactional framing" of the UK street tests of intelligent vehicles should then be understood not just in relation to a next stage of technological—data-intensive, AI-led—development, but in relation to wider efforts to redefine the relation between automotive systems and society, culturally and economically speaking. However, as I will go on to show, while this promise is consistently invoked in the public presentation of street tests of intelligent technology, it is not followed through—operationalized—in the methodology and trial design of street tests of intelligent vehicles.

3 | FROM THE LAB TO THE STREET: BEYOND THE "SOCIAL DEFICIT" OF INTELLIGENT TECHNOLOGY?

Taking one step back, it should be noted that the commitment to testing intelligent vehicle technology in the social environment, on public roads, can be understood as moving us beyond historical tendencies in technology testing in the automotive sector. Until recently, laboratory-based testing was considered the established paradigm in the automotive sector. In *From the Road to the Lab to Math* (2010), the organization studies scholar Paul Leonardi, who has conducted extensive fieldwork in car companies, shows how, over the course of the 20th century, prominent forms of performance assessment such as crash testing—with its iconic plastic dummies getting shaken and crushed inside a car on an indoor test track—and other forms of automotive testing, like societal impact modelling, have been increasingly confined to dedicated test sites and lab-based computer simulations. In his account, the defining development in automotive testing has been the move away from testing on "the open road" in the early 20th century, to controlled experimentation in the closed spaces of the lab and then to modelling-driven "simulation" at the start of the 21st. Today's automotive testing in everyday environments like the street arguably take us beyond this narrative.

Street tests of intelligent vehicles can be taken to exemplify a wider paradigm shift in technology testing, and the management of relations between innovation and society, through testing. A variety of scholars have commented on the rise of "real-world" experimentation, whereby technologies are increasingly tested in sites "beyond the laboratory" (Van der Poel et al., 2017; see also Gross & Hoffmann-Riem, 2005). Contrasting such experimentation to controlled laboratory experiments, Van der Poel and colleagues note that real-world testing entails affirmation of the fact that "only after its implementation will we gradually learn about the impacts of a technology on society, the normative and moral issues raised by such processes, and the best way to embed it in society." They give the example of Autopilot, Tesla's driver-assist feature which has been advertised as enabling autonomous or driverless driving, and was rolled-out in Tesla cars at an early stage of its development, and presents it as an exemplar of experimentation in society: "because Tesla explicitly recognizes the technology as experimental, but also because the experience of using it may lead to further improvements in the system … along the way" (Van der Poel et al., 2017, pp. 5–6). In a similar vein, Jackson, Gillespie, and Payette (2014) have discussed the "beta-testing model" for the introduction of technology to society (see also Neff & Stark,2004).

In line with the logic of data-intensive machine learning invoked above, real-world testing of intelligent vehicles is here understood as implementing an approach to technology testing that derives from software development, which Jackson et al. (2014) label "beta-testing." They argue that, in the tech industry, it has become customary to release experimental products and services to users at an early stage in their development, as companies release relatively untested, unstable devices into everyday environments, relying on user trials and field tests to identify not only technical problems with the applications in question, but also ethical, social, and legal issues with their functioning in society (on this point, see also Marres, 2018; Neff & Stark, 2004). Finally, Laurent and Tironi (2015) have suggested that street trials of smart vehicles, in Saint Denis near Paris, implement a new, emerging paradigm in automotive innovation which they call "experimental innovation." They contrast this model with an older, industrial approach to the introduction of automotive technology to society in France: whereas the former involved the construction of "complete socio-technical systems in-house," car companies today increasingly enter into partnerships with a diverse set of agencies in government, business, and society in order not just to implement a new form of transport, but to configure "a whole ecosystem" of mobility, in which social actors become involved as stakeholders, and in which the very role of these agencies in the transport system is put at stake (Laurent & Tironi, 2015, p. 211). In street trials of intelligent vehicles in Greenwich, Milton Keynes, Coventry, can we observe a similar approach to the one identified by Van der Poel, Jackson et al. and Laurent and Tironi? Could we even observe *an expansion or radicalization of the experimental approach to the introduction of technology to society in these cases*, *as these tests do not just locate testing in the social environment*, *and enrol social actors as stakeholders and/or test subjects*, *but define the very object of technological innovation in social terms*?

It is necessary to pose this question, partly because it helps to bring into view current limitations of the sociology of intelligent technology as formulated in the 1980s. Real-world testing of intelligent technology can be taken to amount to a partial falsification of assumptions formulated in the sociology of AI in that period, insofar as sociologists of technology claimed then that artificial intelligence research, "lacks a social theory" (Woolgar, 1985). This claim continues to reverberate today: in a recent article on the sociology of robots, Esposito (2017) posits that "the sociological perspective is not involved in designing algorithms, which are programmed without adequate consideration of social and communicative aspects" (see also Sloane & Moss, 2019). Meister (2014) states that to date sociology has had limited influence on the field of social robotics, and notes artificial intelligence's reliance on narrow interactional framings of sociality (see also Alač, Movellan, & Tanaka, 2011). These critiques build earlier work in the sociology and anthropology of technology, which examined the blind spot for social aspects in technical fields, positing a "social deficit" of computational systems (Joerges, 1989; Suchman, 1987).[8] These authors claimed that AI and robotics disregard or ignore the situated, contextual, and generative character of human-machine interaction (Suchman, 2007; see also Suchman & Weber, 2016). Button and Dourish (1996) neatly sum up this critique in observing that many problems with computational systems derive "not so much from their technological limitations, but more from their insensitivity to the organisation of work and communication in real work environments." The question is: are these criticisms still valid and/or effective ways of engaging sociologically with intelligent machines in a context defined by real-world testing?

One of the problems with the 1980s social deficit thesis, in my view, is that it prevents us from analysing how social phenomena emerge and/or are articulated experimentally—as proto-type—in real-world technology testing in society, in ways that may or may not be recognized by the designers of these tests, but which could help to surface societal aspects and issues that otherwise tend to remain un- or under-analyzed. Briefly put, positing the social deficit often means that the sociological potential of technology tests remains out of view. Here it is relevant to note that sociologists have applied the notion of a “social deficit” not just to intelligent technology, but also to technology testing, even if they did not explicitly use this label. In “Testing—One, Two, Three... Testing!,” Pinch (1993) sets himself the task to disprove the notion that testing is “providing access to a purely technical realm” (p. 25). In relation to computer systems, he notes that “tests [of computer systems] can be construed to be as much about testing the user as they are about testing the machine. ... any technology that requires the user to act in new sorts of ways (such as when a new technology is first introduced) will involve some in vivo testing” (p. 36). However, Pinch goes on to note that user testing tends to be highly constrained in the process of technology development, as the smooth functioning of technology is widely understood to depend on predictable and disciplined user behaviour. In Pinch’s account, sociologists may be sensitive to the social dimensions of technology testing, prevailing approaches to the development of computational technology are not:

> *the machine has embedded within it assumptions about us whereby our future interaction with it can be projected. This “embedded projection” gives the appearance of a kind of volition—it seems that the machines are training us to use them properly. ... The possibility of negotiating with and persuading the machine ... are extremely curtailed. ... For technological systems with interchangeable parts, it is highly desirable to have the potentially capricious user black boxed. (pp. 37–38)*

This focus in early 1990s social studies of testing is still traceable in well-known STS concepts like the socio-technical script, which lead us to understand the role of users in technology testing first and foremost in terms of compliance and resistance (Woolgar, 1990), the enrolment or not of “humans” (Callon, 1986). This approach may be less suitable when empirical sociology is confronted with artificial situations in which specific forms of social action—interact with the machine!—are being framed and promoted, and actors are qualified in terms of social attributes (“busy mum,” “disabled person”).

In what follows I will argue that it is, however, not enough for sociology to observe and describe how social phenomena feature in intelligent technology testing. We equally need to attend to how such phenomena are bracketed and/or disavowed in the test. Most importantly, to realize the potential of real-world testing for social enquiry, sociology must move beyond treating tests as objects of enquiry, as has been the norm in social studies of testing, and engage methodologically with real-world testing. If such tests are to qualify as sociological tests, we will need to modify test protocols so as to attune them to sociological phenomena. On this point, it is relevant to remember that, even in the 1990s, not everyone in the social studies of testing subscribed to the idea that AI “lacks a social theory.” In that same period, Collins (1990) proposed that intelligent technology tests can be approached as tests of sociological propositions. His study of artificial systems (1990) argued that “the artificial intelligence experiment, is not just a problem of engineering and psychology, but an empirical test of deep theses in the philosophy of the social sciences” (p. 8). Collins proposed that AI tests could be treated as experimental operationalizations of fundamental questions such as: is there a distinction between social action and behavior? Does methodological individualism obtain? Can knowledge be acquired without participation in a social community? Latour’s (1996) slogan, that technology is sociology “by other means” (p. 210) expresses a similar confidence in the equivalence of engineering-based and sociological approaches. Whereas Pinch, then, diagnosed a kind of “social deficit” in relation to prevalent protocols in technology testing, namely their indifference to a recalcitrant, testing user, Collins and Latour suggested that engineering paradigms were already attuned to societal phenomena, in and of themselves, *without any modification of the test protocol required to achieve this.*

Today’s real-world testing of intelligent technologies in social environments requires a different approach. Even as the “social deficit” of intelligent technology can no longer be assumed, neither is it possible, today, to express confidence in a spontaneously given, instead of hard won, equivalence between engineering and sociology.

As real-world tests explicitly bring social phenomena ("interaction," "experience," "public engagement") within the frame of engineering-based research designers, it becomes obvious that engineering is different—methodologically speaking—from sociology. As I would like to show, there is also a range of sociologically relevant phenomena which are patently *not* included within engineering-led real-world tests, or at least not in the UK street trials of intelligent vehicles under discussion here. When considering the implementation of intelligent vehicles tests in the UK streets, real-world testing can for the most part *not* be said to bring society within the experimental remit. We can nevertheless establish a significant difference with previous sociological accounts of the "social deficit" of intelligent technology and technology testing: in the trials under scrutiny, the social deficit primarily presents *a methodological problem.*

4 | TESTING TECHNOLOGY BUT NOT SOCIETY: "LEARNING TO GET ALONG" AS AN INDIVIDUAL, NOT A RELATIONAL, CHALLENGE

While accounts of UK street trials of intelligent vehicles published in news, online, and governmental media deploy social frames to characterize the trials, a different picture emerges from our fieldwork. Together with colleagues, I attended trials and visited test sites in Greenwich (London), Milton Keynes, Coventry, and the Horiba Mira Test site in Nuneaton between 2016 and 2018, and in many cases, we found that opportunities for interaction between intelligent vehicles and other road users were significantly curtailed. To start with, in all cases the trials design and implementation included significant provisions for the containment and management of the machine's encounter with "social complexity," to the point that the trial situation could not be said to qualify as an operationalization of that notion from a sociological perspective. These efforts at containment take various forms, but they include the use of media embargos to prevent the public being notified of tests until after the tests in question were concluded, as was the case during the Coventry trials in November 2017 and the Milton Keynes trials in October 2016 and October 2018. This does not mean the trials aren't noticed or recorded by the public: in November 2017, the *Coventry Telegraph* published a video of Autodrive and JLR vehicles on the streets of Coventry city center, showing a fenced-off bit of road, with test vehicles entering and leaving a car park to enter a stretch of road turned into the test track. A user commented: "It needed people every 10 s to stop anyone crossing the roads lol. So they can't cope with someone stepping out."[9] Furthermore, social media offer myriads of reports of intelligent vehicle sightings on urban streets, as in Milton Keynes where a Twitter user noted in the autumn of 2016 that they "had to drive out of the way of one of Milton Keynes' autonomous cars on the sidewalk yesterday. By the time my phone booted it was gone."

While the ostensible aim of the tests is then to demonstrate the capacity of technology to co-exist with other road users, the settings in which they take place tell a more complex story. In all of the trials we observed, the test involved material, organizational, and regulatory operations upon the street environment, which rendered it more passive, less open, and more compliant with the machine's needs. In Milton Keynes, cyclists were not permitted on the pedestrian pathways on which driverless pods were tested (Figure 2). Probed on this point during a public panel discussion, a city counselor clarified the legal background to this state of affairs: technically "you can't run a pod on a pedestrian pathway because it was passenger carrying so we know we need to change the regulation to change pedestrian paths into roads."[10] In several cases, pedestrians themselves were prohibited from using pedestrian pathways where intelligent vehicles were being trialed. On encountering a driverless pod in Milton Keynes in October 2018, I was politely but firmly asked to get out of the way. When I enquired whether pedestrian detection wasn't part of the vehicle's technical features, a street marshal wearing the customary Hi-Viz jacket pointed out to me that the ride in question was part of a public demo, and the test vehicle had notable guests inside.[11] Preparations of the vehicle's path, finally, do not just take the form of temporary provisions like marshals and temporary fencing along the test route, but also involve material intervention in the street environment. Along the Thames Path on the Greenwich Peninsula, where the Gateway trial took place, a high blue fence, which looks like

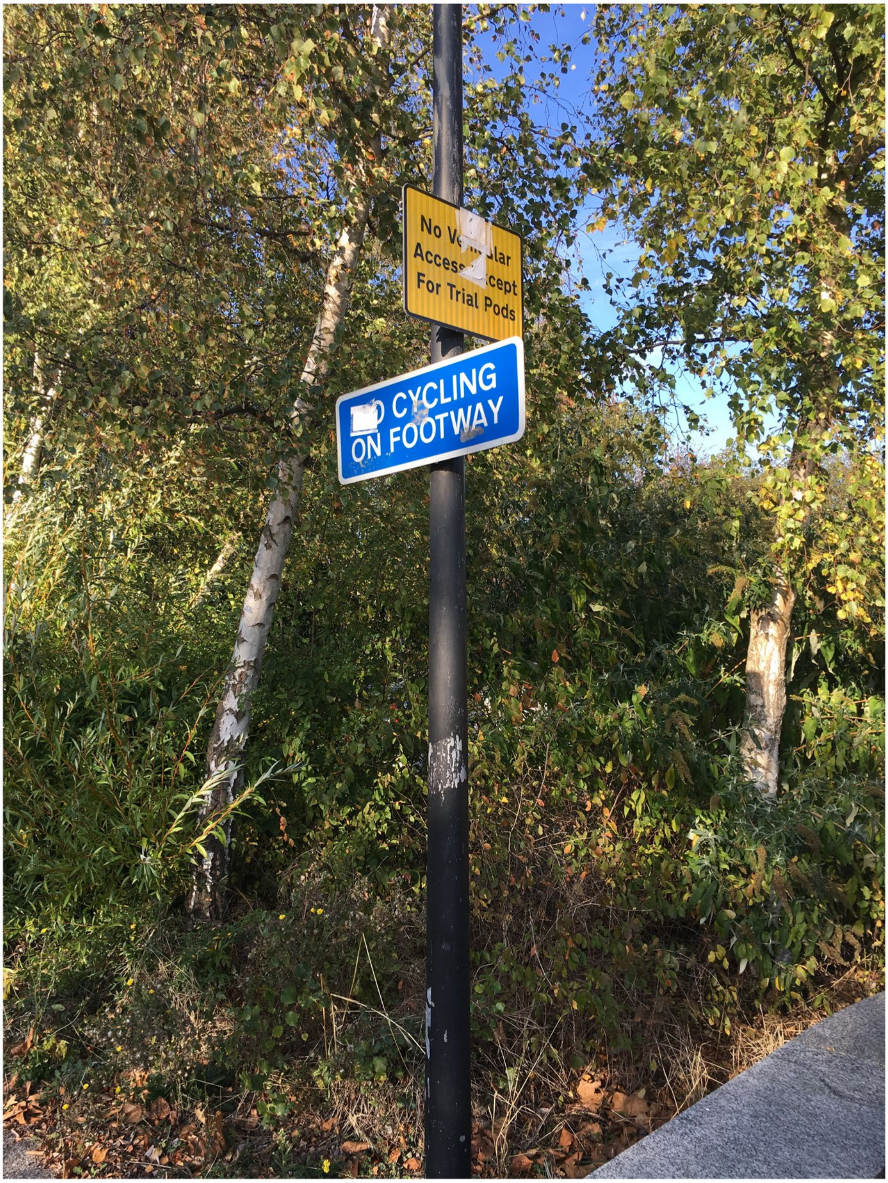


FIGURE 2 Driverless pods test route, Milton Keynes [photograph by the author], October 2018 [Colour figure can be viewed at wileyonlinelibrary.com]

it is there to stay, separates the shuttle path from the neighboring conference center. As the trial organizers noted during a public presentation, we "added a distinct path for the shuttle, a 'shuttle route' so that pedestrians have an expectation that a shuttle is operating, with a logo on the floor of a pod, and we have improved surfacing for the pods."[12] An extensive CCTV system allows for the monitoring of this pathway. Apparently, when there were too many people the shuttle wouldn't run.

While public accounts of intelligent vehicle trials emphasize the commitment to investigate and enable the co-existence of machines and road users in the street, observation of the tests' implementation bring into focus a number of limitations that call this ambition into question, when approached from a societal perspective. To be sure, many of these limitations make sense from a technical point of view—the machine after all is *still learning* to get along with city transport, they severely limit the ability of social actors to even engage with the trial. Importantly, the limitations in question are not only technical, but also methodological: while the capacity to "share the road with others" is the stated objective of the trials, the trial protocols specify this objective as an individual not a relational challenge. This was apparent during a demonstration of Autodrive driverless technology at Horiba Mira, the UK automotive test site. Here, stakeholders were invited into test vehicles to experience various features of connected and autonomous vehicle technology, including this Emergency Vehicle Warning (EVW) demo. As I put it in my fieldnotes at the time:

> *The driver—German?—tells us about the "blind spot system"—a system for automated vehicle detection—but he says: "no bikes, no horses, and no pedestrians." However, it* can *detect roundabouts and stop lines. I ask about information asymmetry, talk about how making space for a vehicle involves coordination work, which is hard when we don't know who has and hasn't received the "smart" signal [from the passing emergence vehicle]. He says nothing for a few seconds. Then: yes, we assume this will be mandatory for all cars. Another passenger takes a more constructive approach: he asks whether the people in the emergency vehicle need to do anything to get this signal out. They say, yes, they will have a switch in their fire truck—which they can switch on the same time as their siren. A fire truck comes by.*

In an exchange like this, the constitutive role of situated interaction and the need for the mutual coordination of action among multiple actors is clearly missed. Indeed, the test design seems to render its participants insensitive to the following relational question raised in and by on-the-road interaction: how, in an encounter between diverse entities and agents—in society, in public—can we find a language or register of communication that diverse actors are able to share? The experimental design for establishing communication between machines and human entities in a social environment, then, *misses the central challenge of co-existence, a challenge that is constitutive of social life*: how to negotiate difference, how to relate across chasms that separate cultures, genders, classes, experiences? Street trials of intelligent vehicles may be presented, in media outlets like the *Financial Times*, as directly concerned with social phenomena like interaction and engagement, the implemented trial designs have technology, not social life, as their object. That is also to say, an analysis of the methodology and implementation of street trials of autonomous vehicles from a sociological perspective suggests that what we need is a reconstructive approach to intelligent technology testing in society: if society is to be brought within the frame of street trials of intelligent technology, prevalent test protocols will need to be modified.

5 | TESTING ON THE SOCIAL ROAD: THE EXPLICATION OF MUTUAL CONSTRAINTS

To sum this up in a straightforward manner, the test protocols implemented in the observed street trials display a technological bias. They are organized to put intelligent technology prototypes to the test, to challenge *their* capacities to interact with other road users—so that these capacities can be qualified, strengthened and developed further. But this experimental approach is not extended to social actors present in the situation, or at least not intentionally so. Pedestrians, cyclists, other road users and passengers are not put to the test in this same way, their capacities to relate—to learn to get along with others—is not subject to deliberate experimental qualification (although their attitudes towards the trials and intelligent vehicles are documented and reported upon. However, the (re-)qualification of social actors implicated in the trials—their capacities and relations—*does* occur, as an *effect* or *consequence* of the trials' implementation (even if it is not its intended object). In the case of the Greenwich test, a variety of experimental effects did and do arise: for example, some actors have explicitly taken issue with its stated commitment to machines "learning to get along" with others, objecting that the trial's design in fact prevents this hypothesis from being put to the test. As *E&T Magazine* reported this summer, "[c]yclists were horrified when a dedicated riverside bike lane was commandeered for use by an autonomous shuttle bus in the Greenwich Peninsula district of London earlier this year as part of another government-backed trial" (Loeb, 2017). Not co-existence of machine and human, but displacement of other road users by machines is here explicated as an empirical consequence of street tests of driverless cars. An organization called the Pedestrian Liberation Front has been considering lodging a formal complaint against an Ocado trial with driverless delivery vans in Woolwich: it alleged that pedestrians are discriminated against by the trials as they take over streets previously marked as "pedestrian-friendly" (Loeb, 2017). Such interventions surface possible requalifications of the capacities of social actors—*in casu* pedestrians—as a consequence of the introduction of intelligent vehicles into the street. When approached from the standpoint of their effects, it then appears that street testing of intelligent vehicles *can* do double work—they can facilitate technological as well as social experimentation. While we can clearly detect sociological limitations of engineering-led designs of real-world technology tests, they may give rise to social experimentation all the same.

Rather than continuing to insist on the shortcomings—the social deficits—of the current designs of intelligent vehicle street trials—their less-than-relational, less-than-experimental envisioning of the encounter between social actors and intelligent machines in UK streets—I would like to emphasize that street trials have *the potential* to surface and frame social aspects of intelligent vehicle technology. Crucial in this respect is a particular disjuncture between the trials' design and the effects of their implementation: While the tests I observed mostly rely on an *interactional* definition of the encounter between machines and other road users, *street testing in practice pushes the commitment to co-existence beyond this narrow "interactional frame."* The question of the possible co-existence of diverse road users—cars, cyclists, pedestrians—is not just a question of the physical encounter of individuated actors in the street. It includes the question of how their existence in the street is environmentally enabled—cycling, for example, is associated with geographic proximity (living in the city), while automobility implies societal inter-dependence (Latimer & Munro, 2006).[13] From this perspective, the challenge of co-existence inevitably exceeds the question of interaction in the street, but includes the wider challenge of how to negotiate difference—"what currently separates" different road users in terms of their commitments and dependencies, capacities, their life forms, and so on, and what threatens to constitute their diverse life forms as mutually exclusive, and to render their co-existence impossible: How *can* cyclists, pedestrians, drivers and the driven co-exist peacefully on the UKs open roads?[14] While it is not currently addressed within CCAV trial designs, the question of an agent's capacities for "co-existence" may equally be posed of pedestrians and cyclists, and cycling advocates: are *they* prepared to share the street with others? And: are *they* able to recognize the societal constraints and interdependencies expressed in the car system? If we are to "learn to get along," the encounter among mutually contested forms of mobility must be explored, and indeed, tested.

The above criticisms of UK street trials with intelligent vehicles as excluding sociological phenomena like mutual coordination in situ and the negotiation of difference among diverse actors from the trial design, can then be reformulated as a methodological challenge: is it possible to design a street trial that enables the exploration of these relational challenges raised—and highlighted—by the appearance of intelligent vehicles in the UK streets? Can street trials of intelligent vehicle technologies serve as experimental occasions for the explication of the varied mutual constraints that today limit, and indeed, render impossible, co-existence among diverse entities on the open road? Together with my colleagues [Rebecca Cain, Ana Gross, Lucy Kimbell, and Nerea Calvillo] we undertook an interdisciplinary experiment in 2016 that took up these broad questions. The pilot project brought together sociologists, design researchers, and vehicle engineers at the University of Warwick to examine possibilities of conducting social research in the "driver-in-the-loop' simulator, a new simulator that was:

> *designed specifically to test real-world robustness and usability of smart, connected and autonomous vehicle technology. … Using 30 miles of photorealistic, real world driving routes presented via a 360-degree high definition visuals, accompanied by 3D surround sound and real vehicle motion, they will deliver an immersive experience for driver-in-the-loop technology evaluations.*[15]

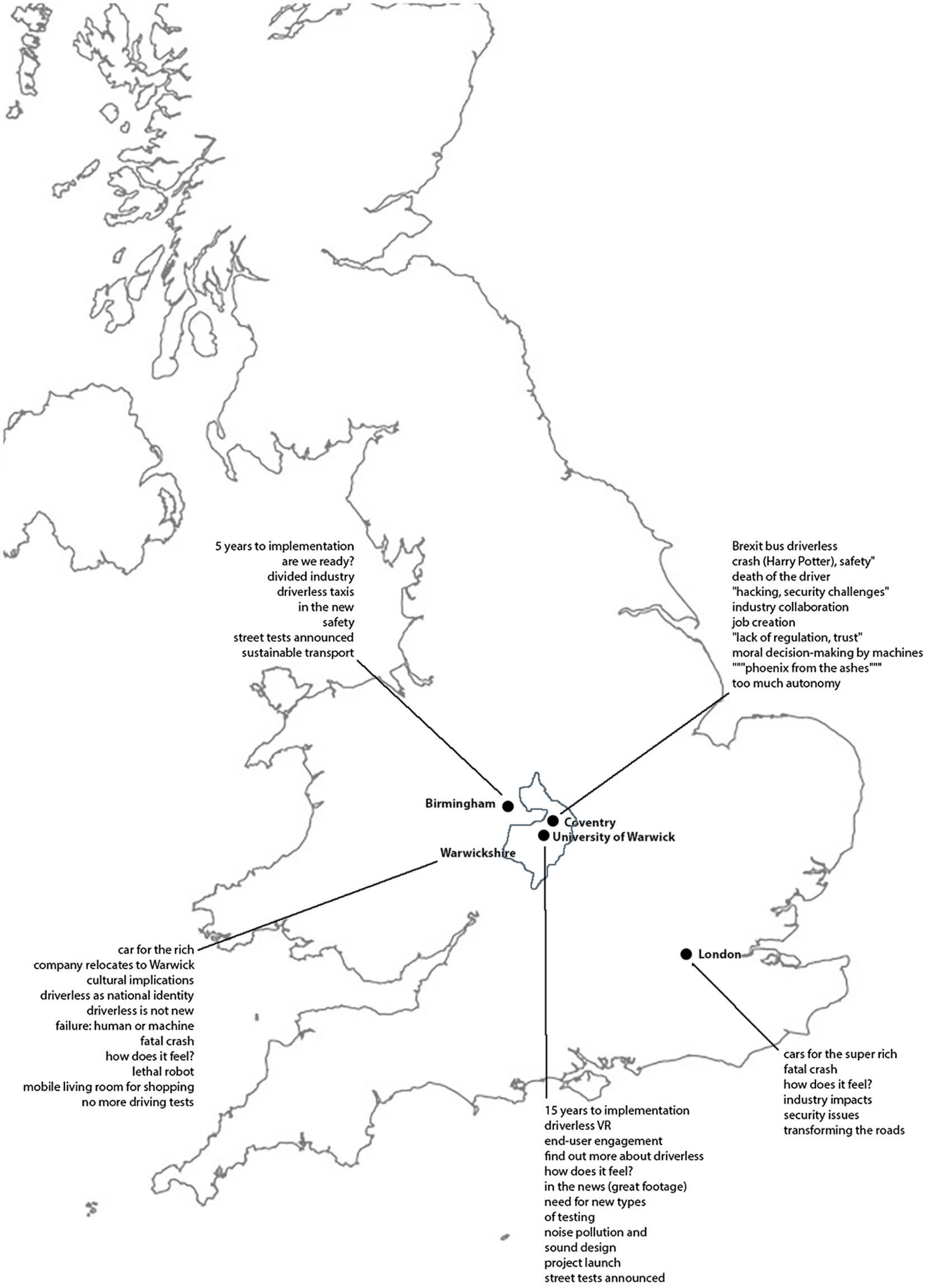


FIGURE 3 Issues raised in relation to driverless cars in the West Midlands on Twitter (2016, Figure designed by N. Calvillo) [Colour figure can be viewed at wileyonlinelibrary.com]

We asked: can this test environment, which was designed for engineering-led research on intelligent vehicles, including the development of "next generation communication protocols … and approaches to validate sensing technologies like Radar, LiDAR camera and ultra-sonic,"[16] be adapted to investigate social aspects of autonomous vehicles? Elsewhere I report on the test methodology in more detail (Marres, Cain, Gross, Kimbell, & Ulahannan, 2017), but for the purposes of this article it is important to note it consisted of two stages. In a first step, we conducted a public debate mapping, in which we used digital methods to identify issues raised on the social media platform Twitter in relation to autonomous vehicles and driverless cars in the West Midlands. The aim of this exercise was not to arrive at a representative overview of public debates on our topic, but rather, to build up a list of issues raised by driverless cars in this region. We used a query-based Twitter data-set, including tweets containing the terms driverless, CAV, and intelligent vehicle.[17] For the purposes of this particular debate-mapping exercise, we analyzed only the tweets of which the account description listed relevant locations (Coventry, Birmingham, Warwick, Warwickshire). In order to identify the issues raised on Twitter in relation to the Coventry/Warwick trials, we then manually coded our Twitter data using a loose interpretative framework, which we visualized using different criteria (such as uniqueness and frequency in Figure 3). This initial issue visualization suggests, we found, that UK street trials of driverless cars enable the explication of wider societal constraint, as they mobilized terms such as "mobile living room for shopping," "job creation," "Brexit bus driverless," and "car for the rich." We invited students to visit public places to seek responses to this and similar visualizations, on the Warwick campus, and in Coventry city center, in an effort to deploy our issue maps as "devices of elicitation," assuming that displaying concerns offers a way of inviting everyday actors—in the street—to make sense of issue formations, and to elaborate on them. Results were mixed, however, in that public responses focused as much on the representation (why Twitter? How did you get to this picture?) as on the issues surfaced by our research.

In a next step, we conducted a participatory exercise in the driver-in-the-loop simulator, inviting social researchers, designers, engineers, and policy makers, to explore autonomous vehicle issues in this environment. Presenting the simulator as an environment for the exploration of "issue-scapes," we invited participants to annotate the simulator, using issue terms featuring on our Twitter maps. Participants received instructions as to how to produce an issue scape, namely by using sticky notes to attach issues to objects present in the simulator—like the car itself, or the stretch of Coventry road projected on the 360-screen surrounding them. There were also cardboard figures available for annotation, representing human actors and cardboard boxes representing non-humans (machines, technologies, institutions, etc.). While participants were clearly fascinated by the simulator—the technology—itself, they also made significant efforts to locate some of the issues raised on Twitter inside the driverless simulator environment, in doing so elaborating and generating further issue articulations. Thus, "the elderly" were introduced in the setting in the form of an "old lady" cardboard figure, which was settled into the back seat of the car, with a note on the window stating "a dashboard that says 'old lady on board': stigma." Someone attached a "basket for collecting road kill" to the front of the car, arguably signaling the lack of provisions for animal presence in the simulator environment. In the relative darkness of the simulator, one of the vehicle engineers observed that Coventry as a transport environment has one crucial feature that he believes his profession needs assistance in negotiating: "city-ness" (Marres et al., 2017).

6 | CONCLUSION

The location of intelligent automotive technology testing "in" society remains an unfinished project, from the standpoint of the sociology of technology testing. The STS proposition that engineering can be regarded as "sociology by other means" (Latour, 1996) cannot, at present, be extended to real-world testing of intelligent vehicles in the UK streets: the transformation of intelligent technology tests into sociological tests, if this is feasible at all, will require significant modifications of test protocols, at the very least. However, neither does it suffice today to diagnose a "social deficit" of intelligent technology testing, and leave it at that. The exclusion of social phenomena from technology tests does not only present a substantive limitation of these tests. It can equally be approached as an experimental, methodological challenge: how *could* street trials of intelligent technology serve the elucidation of societal issues? In the driver-in-the loop simulator, we tried out ways of introducing society into the test environments—can "issue maps" do this job?—as well as ways of dramatizing societal concerns in this setting—can we use cardboard figures to explicate dynamics of stigmatization? In taking this approach, we treat society not as a "model community" to be demonstrated—and promoted—in a test environment, as is the prevalent, rather unexperimental approach in test bed design (Engels et al., 2019), but as composed of an open-ended set of actors, relations, issues, and dynamics that a test environment may render explorable, and perhaps indeed, test-able. In this sense, environments designed for the real-world testing of intelligent technology do have the potential to "test society."

It may then be time to redefine what is put to the test in technology testing in society—not just the capacities of machines, but also their relations to social actors, and the relational capacities of all involved. And also, to redefine the job description of the sociology of testing, and what it adds to engineering in these cases. STS scholars have long argued that technological tests inevitably also put social actors and arrangements to the test, yet they too—not only engineers—have applied different criteria to both. Whereas technology testing was defined in terms of the *qualification and development* of machinic capacities, the test of human capacities was often described by sociologists of technology testing in much flatter terms—most notably in terms of the enrolment, compliance, and alignment of users and stakeholders. The challenge of how tests can qualify and activate relational capacities was consequently not really broached by social studies of technology in the 1980s. Intelligent technology testing in social environments today brings this challenge to the fore. Conceivably, the "co-existence" of social actors and machines on the road could be achieved by compelling humans to comply with machinic requirements. But surely realization of the ambition to achieve intelligence, in the street environment, would require something different, something like relational attunement between diverse road users, and this will require the design of different tests than those currently being rolled out on UK streets.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the author. Some data are not publicly available due to ethical restrictions including their containing information that could compromise the privacy of research participants.

ORCID

Noortje Marres https://orcid.org/0000-0002-8237-6946

NOTES

[1] This includes pedestrian paths and cycling lanes, city streets, and ring roads, in many cases these streets and roads are closed off for the duration of the test. In other cases, CCTV is relied upon to monitor the test environment.

[2] Different terms are used to describe the prototype vehicles in question: popular media tend to refer to "driverless cars," while government has opted for "connected and autonomous vehicles," and the field of engineering tends to refer to intelligent vehicles, as in the Intelligent Vehicles group at the University of Warwick. In this article I opt for the latter term, as it usefully highlights the issue I am concerned with here: the ability of these vehicles to orientate their actions towards others, and to invoke a sociological understanding of intelligence (for an example of such interactions, see Belcher, 2017).

[3] The same report stresses that "[p]articular consideration should be given to the concerns of more vulnerable road users including disabled people, those with visual or hearing impairments, pedestrians, cyclists, motorcyclists, children and horse riders" (Department for Transport, 2015).

[4] This way of justifying the release of intelligent vehicles onto urban streets invokes a popular narrative about the roll out of autonomous cars on public roads which was publicized widely in the wake of the Silicon Valley autonomous vehicle releases in the shape of Google's Self-Driving Car and Tesla's Autopilot, the claim namely that only by clocking up very large numbers of "test miles"—which for tech-based automotive companies like Tesla are reported to be in the

billions (see Lambert, 2018)—will the computational systems implemented in these vehicles for environmental sensing, decision-making, and navigation be able to "learn" how to operate in the unpredictable environments of the "open road" (see Stilgoe, 2018). To my knowledge, this argument about scale was not made by representatives of the CCAV trials.

[5] UK Autodrive International CAV Conference, Transport Catapult, Milton Keynes, October 11, 2018.

[6] Driverless Futures, Designology Pop-Up Studio, Transport Museum, London, April 2017.

[7] "The three main participant groups were as follows: Participants who signed up in advance to receive information about GATEway and to participate in shuttle trials; 'Walk up' participants who were in the area and were interested in trying the shuttle as part of their journey; Participants from the local area who saw publicity about the shuttles and wished to experience them" (Fernández-Medina et al., 2018). Incidentally, efforts by myself and a collaborator to sign up for participation in the trial were unsuccessful, we never received a reply.

[8] These arguments must be distinguished from philosophical critiques of AI such as John Searle's (1980) "Chinese Room" thought experiment, which argues that computers lack subjective features such as intentionality.

[9] Another user commented: "shouldn't the public no what goes on, news to me", *Coventry Telegraph*, Facebook live page, November 15, 2017. https://www.facebook.com/coventrytelegraph/videos/1638060592917755/?q=Coventry%20Telegraph%20driverless

[10] UK Autodrive International CAV Conference, Transport Catapult, Milton Keynes, October 11, 2018.

[11] One of the functions of the marshals in CCAV trials is to ensure members of the public had a positive experience of AVs, according to McDowell-Naylor, who studied the GATEway trial in Greenwich. He quotes one of the trial organizers: "we don't want any dramas because that means that something has probably gone wrong" (McDowell-Naylor, 2018, p. 174). This type of curation of "experience" by means of public trials, I am arguing here, is not compatible with a conception of tests as a form of "social experimentation."

[12] GATEway presentation, Driverless Technology Conference DTC16, Milton Keynes, November 22, 2016.

[13] Indeed, this latter circumstance is usefully highlighted in the intelligent vehicle discourses that foreground of *care and caring subjects*—busy mums, the visually impaired, the elderly.

[14] I build here on earlier work on so-called "issue-publics," which I defined in terms of diverse actors being jointly and antagonistically implicated in a matter of concern, "[such as] environmental NGOs and leading international banks [involved in fossil fuels], [who] are bound together by mutual exclusivities between their attachments to the matter at hand" (Marres, 2005, p. 129).

[15] https://warwick.ac.uk/newsandevents/news/world146s_most_adaptable/

[16] https://warwick.ac.uk/newsandevents/news/world146s_most_adaptable/

[17] We collected our Twitter data with the aid of T-CAT—the Toolset for the Capture and Analysis of Twitter data (Borra and Rieder, 2014), between June 10 and September 9, 2016. The data set of region-specific tweets was small, containing 662 tweets. We identified a total of 138 issue terms.

REFERENCES

Alač, M., Movellan, J., & Tanaka, F. (2011). When a robot is social: Spatial arrangements and multimodal semiotic engagement in the practice of social robotics. *Social Studies of Science*, *41*(6), 893–926.

Belcher, A. (2017, November 17). Driverless cars on the city's roads: How the people of Coventry have reacted. *Coventry Telegraph*. Retrieved from https://www.coventrytelegraph.net/news/coventry-news/driverless-cars-13916143

Bialski, P., Brunton, F., & Bunz, M. (2019). *Communication*. Minneapolis, MN: Minnesota University Press.

Borra, E., & Rieder, B. (2014). Programmed Method: Developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262–278.

Button, G., & Dourish, P. (1996). Technomethodology: Paradoxes and possibilities. In *CHI '96 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Vancouver, Canada: ACM.

Callon, M. (1986). The sociology of an actor-network: The case of the electric vehicle. In M. Callon, A. Rip, & J. Law (Eds), *Mapping the dynamics of science and technology* (pp. 19–34). London, UK: Palgrave Macmillan.

Campbell, P., & Yang, Y. (2018, February 11). Didi Chuxung tests self-driving taxis on public roads. *Financial Times*.

Castelle, M. (2018, September 28). Social theory for generative networks (and vice versa) [Blog post]. Retrieved from https://castelle.org/pages/social-theory-for-generative-networks-and-vice-versa.html

Collins, H. M. (1990). *Artificial experts: Social knowledge and intelligent machines (inside technology)*. Cambridge, MA: MIT Press.

Commonplace. (2018). *GATEway project Sentiment mapping analysis*. TRL Limited. Retrieved from https://gateway-project.org.uk/wp-content/uploads/2018/06/D3.6-GATEway-Sentiment-mapping-Summary-report.pdf

Dennis, K., & Urry, J. (2009). *After the car*. Cambridge: Polity Press.

Department for Transport (UK). (2015). The pathway to driverless cars: A code of practice for testing. Retrieved from https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/446316/pathway-driverless-cars.pdf

Engels, F., Wentland, A., & Pfotenhauer, S. M. (2019). Testing future societies? Developing a framework for test beds and living labs as instruments of innovation governance. *Research Policy*, *48*(9), 103826.

Esposito, E. (2017). Artificial communication? The production of contingency by algorithms. *Zeitschrift für Soziologie*, *46*(4), 249–265.

Fernández-Medina, K., Delmonte, R., Jenkins, R., Holcome, A., & Kinnear, N. (2018). *102200 GATEway Trial 1: Deployment of a micro-transit vehicle in a real-world environment*. TRL Limited. Retrieved from https://gateway-project.org.uk/wp-content/uploads/2018/06/D5.3a_TRL-Trial-1-Project-Report_PPR858.pdf

Gross, M., & Hoffmann-Riem, H. (2005). Ecological restoration as a real-world experiment: Designing robust implementation strategies in an urban environment. *Public Understanding of Science*, *14*(3), 269–284.

Hildebrandt, M. (2019). Privacy as protection of the incomputable self: From agnostic to agonistic machine learning. *Theoretical Inquiries in Law*, *20*(1), 83–121.

Jackson, S. J., Gillespie, T., & Payette, S. (2014, February 15–19). The policy knot: Re-integrating policy, practice and design in CSCW studies of social computing. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. Baltimore, MA: ACM.

Lochlann Jain, S. S. (2004). "Dangerous instrumentality": The bystander as subject in automobility. *Cultural Anthropology*, *19*(1), 61–94.

Joerges, B. (1989). Romancing the machine: Reflections on the social scientific construction of computer reality. *International Studies of Management and Organization*, *19*(4), 24–50.

Lambert, F. (2018, July 17). Tesla's fleet has accumulated over 1.2 billion miles on Autopilot and even more in "shadow mode", report says. *Electrek*. Retrieved from https://electrek.co/2018/07/17/tesla-autopilot-miles-shadow-mode-report/

Latimer, J., & Munro, R. (2006). Driving the social. *The Sociological Review*, *54*(1_suppl), 32–53.

Latour, B. (1996). *Aramis, or the love of technology*. Cambridge, MA: Harvard University Press.

Laurent, B., & Tironi, M. (2015). A field test and its displacements. Accounting for an experimental mode of industrial innovation. *CoDesign*, *11*(3–4), 208–221.

Leonardi, P. M. (2010). From road to lab to math: The co-evolution of technological, regulatory, and organizational innovations for automotive crash testing. *Social Studies of Science*, *40*(2), 243–274.

Loeb, J. (2017, July 6). Pedestrians rage at autonomous pods and delivery bots on pavements. *Engineering and Technology*. Retrieved from https://eandt.theiet.org/content/articles/2017/07/pedestrians-rage-at-autonomous-pods-and-delivery-bots-on-pavements/

Marres, N. (2005). *No issue, no public* (doctoral dissertation). University of Amsterdam. Retrieved from https://pure.uva.nl/ws/files/3890776/38026_thesis_nm_final.pdf

Marres, N. (2019). What if nothing happens? On street trials of driverless cars as experiments in participation. In S. Maassen, C. Schneider, & S. Dickel (Eds.), *TechnoScience in society*, *Sociology of Knowledge Yearbook*. Nijmegen: Springer/Kluwer, forthcoming.

Marres, N., Cain, R., Gross, A., Kimbell, L., & Ulahannan, A. (2017, April). *Surfacing social aspects of driverless cars with creative methods* (workshop report). University of Warwick. Retrieved from http://www2.warwick.ac.uk/fac/cross_fac/cim/events/driverlesscarswithcreativemethods/

McDowell-Naylor, D. (2018). *The participatory, communicative, and organisational dimensions of public-making: Public engagement and the development of autonomous vehicles in the United Kingdom* (Unpublished PhD thesis). Royal Holloway, University of London, London, UK.

Meister, M. (2014). When is a robot really social? An outline of the robot sociologicus. *Science, Technology and Innovation Studies*, *10*(1), 107–134.

Neff, G., & Stark, D. (2004). Permanently beta: Responsive organization in the internet era. In P. N. Howard & S. Jones (Eds.), *Society online: The internet in context* (pp. 173–188). Thousand Oaks, CA: SAGE Publications.

Parliament, House of Lords. (2017). Connected and autonomous vehicles: The future? 2nd Report, HL 2016-7 (115). Retrieved from https://publications.parliament.uk/pa/ld201617/ldselect/ldsctech/115/11502.htm

Pinch, T. (1993). "Testing—one, two, three... testing!": Toward a sociology of testing. *Science, Technology, and Human Values*, *18*(1), 25–41.

Van de Poel, I., L. Asveld, & D. C. Mehos (Eds.) (2017). *New perspectives on technology in society: Experimentation beyond the laboratory*. London, UK: Routledge.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424.

Sloane, M., & Moss, E. (2019). AI's social sciences deficit. *Nature Machine Intelligence*, *1*(8), 330–331.

Stilgoe, J. (2018). Machine learning, social learning and the governance of self-driving cars. *Social Studies of Science*, *48*(1), 25–56.

Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions*. Cambridge University Press.

Suchman, L. A. (1987). *Plans and situated actions: The problem of human-machine communication*. Cambridge: Cambridge University Press.

Suchman, L. A., & Weber, J. (2016). *Human-machine autonomies* (Working paper). Lancaster, UK: Lancaster University.

Tennant, C., Howard, S., Franks, B., Bauer, M. W., & Stares, S. (2016). Autonomous vehicles—Negotiating a place on the road: A study on how drivers feel about interacting with autonomous vehicles on the road. Retrieved from https://www.lse.ac.uk/website-archive/newsAndMedia/PDF/AVs-negociating-a-place-on-the-road-1110.pdf

Tute, R. (2017, November 24). Driverless vehicle testing on public roads hailed as landmark moment. *Infrastructure Intelligence*. Retrieved from http://www.infrastructure-intelligence.com/article/nov-2017/driverless-vehicle-testing-public-roads-hailed-landmark-moment

Woolgar, S. (1985). Why not a sociology of machines? The case of sociology and artificial intelligence. *Sociology*, *19*(4), 557–572.

Woolgar, S. (1990). Configuring the user: The case of usability trials. *The Sociological Review*, *38*(1_suppl) 58–99.

Wright, R. (2017, June 7). Driverless vehicles learn to get along with city transport. *Financial Times*. Retrieved from https://www.ft.com/content/8fbba39a-34db-11e7-99bd-13beb0903fa3

How to cite this article: Marres N. Co-existence or displacement: Do street trials of intelligent vehicles test society?. *Br J Sociol*. 2020;71:537–555. https://doi.org/10.1111/1468-4446.12730

Received: 12 March 2019 | Revised: 17 January 2020 | Accepted: 21 January 2020

DOI: 10.1111/1468-4446.12743

SPECIAL ISSUE

WILEY

State work and the testing concours of citizenship

Willem Schinkel

Department of Public Administration and Sociology, Erasmus University Rotterdam, Rotterdam, Netherlands

Correspondence
Willem Schinkel, Erasmus University Rotterdam, Burgemeester Oudlaan 50, 3062 PA, Rotterdam, Netherlands.
Email: schinkel@essb.eur.nl

Abstract

Anyone trying to be a citizen has to pass through a set of practices trying to be a state. This paper investigates some of the ways testing practices calibrate citizens, and in doing so, perform "the state." The paper focuses on three forms of citizenship testing, which it considers exemplary forms of "state work," and which all, in various ways, concern "migration." First, the constitution of a "border crossing," which requires an identity test configured by deceptibility. Second, the Dutch asylum process, in which "being gay" can, in certain cases, be reason for being granted asylum, but where "being gay" is also the outcome of an examination organized by suspicion. And third, the Dutch measurement of immigrants' "integration," which is comprised of a testing process in which such factishes as "being a member of society" and "being modern" surface. Citizenship is analyzed in this paper as accrued and (re)configured along a migration trajectory that takes shape as a *testing concours*, meaning that subjects become citizens along a trajectory of testing practices. In contributing both to work on states and citizenship, and to work on testing, this paper thus puts forward the concept of citizenship testing as state work, where "state work is the term for that kind of labor that most knows itself as comparison, equivalency, and exchange in the social realm" (Harney, 2002, pp. 10–11). Throughout the testing practices discussed here, comparison, equivalency, and exchange figure prominently as the practical achievements of crafting states and citizens.

KEYWORDS
borders, citizenship, integration, migration, state, testing

1 | INTRODUCTION: STATE WORK AND/AS TESTING PRACTICES

Anyone trying to be a citizen has to pass through a set of practices trying to be a state. Those who become visible as "migrants," in particular, are subject to a variety of practices aimed at ascertaining their identity, provenance, and formal citizenship status. Such practices often have the character of tests, and this paper discusses three kinds of tests as a way of analyzing the co-constitutive effects of "states" and "migrants." These tests can be tests of identity, as when people's passports are checked at the border. They can be tests of veracity, as when the accounts by asylum seekers are scrutinized by state officials. And they can also be tests of belonging, as when social scientific measurement apparatuses are deployed to measure "immigrant integration." In this paper, I discuss these three cases as ways to shed light on the coproduction of the state and citizenship. I conceive of citizenship as a *testing concours*. That means citizenship is here considered not as an individuated bundle of rights and duties (cf. Marshall, 1963), nor in terms of "acts of citizenship" by subjects themselves (Isin & Nielsen, 2008), nor exclusively as "migrant citizenships" (Nyers, 2015). These are valuable ways of speaking of citizenship, but here I adopt the complementary approach of considering specific trajectories of citizenship as inextricably bound up with state technologies for the visualization of migration.

An increasing literature argues that migration is not an object readily available for observation, but exists by rendering the movement of certain people visible as "migration." Migration, in other words, does not consist of people travelling from one country to another, but of the registration of the movement of some, but not of others. On the one hand, this perspective builds on work in the "autonomy of migration" literature (Bojadžijev & Karakayali, 2010; Mezzadra, 2010; Moulier-Boutang, 1998; Papadopoulos & Tsianos, 2013; Rodriguez, 1996), which considers migration as a phenomenon characterized by moments of relative autonomy from either socio-economic circumstances or state control. On the other hand, it builds on work that emphasizes the visualization of migration. Not only does migration become a spectacle in and through border control (De Genova, 2013), the very "thing" called migration does not become available for public consideration prior to forms of visualization that are constitutive of it. In other words, migration does not precede visualization (Dijstelbloem & Broeders, 2014; Dijstelbloem, van Reekum, & Schinkel, 2017; Scheel, 2013; van Reekum & Schinkel, 2017). What is called "autonomy of migration" might be more suitably termed "autonomy of mobility." Mobility only becomes "migration" once state officials record data, check passports, grant visa, refuse entry, statistically assess movement, and out of the meshwork of these practices comes the object called "migration," which remains unstable because it can never be fully traced back to data.

At the same time, visualization, registration, and surveillance should not be seen as emanating from sovereign state bodies with the capacities to visualize. Rather, what this paper discusses as citizenship testing allows an understanding of sovereignty by way of mobility (cf. Papadopoulos, 2018). As Annalisa Pelizza has said, "migrant registration and identification" can be seen as the "co-construction of individuals and polities" (Pelizza, 2019, p. 1). Seen in this light, migration control and the calibration of people's citizenship enable what Timothy Mitchell has called "state effects" to exist and persist (Mitchell, 1991, 1999). In visualizing migration, there is something to be *decided*, but that is just another way of saying that the specter of sovereignty requires the possibility of performing ritualized operations on the movement of people that cannot be contained, and never were. The visualization and public enactment of what we call "migration," over against the mere movement of bodies across the earth, thus presents "the state" with ways to prop up itself, to give credibility to the existence of state effects. Sovereignty cannot exist without its ritualized confirmations (Hansen & Stepputat, 2005). To say that migration and the state are co-constitutive of each other is reductive of the generally asymmetrical ways in which particular practices and associations (the ones we tend to call "state") attach to other practices and associations (the ones we end up calling "migration") in order to enact "migration" as, itself, a particular effect of the state. As Mitchell has said, what is called "state" comes out of "methods of organization, arrangement and representation that operate with the social practices they govern, yet create the effect of an enduring structure apparently external to those practices" (Mitchell, 1999, pp. 77–78).

This paper looks at such effects by considering different moments and spaces of "migration" in terms of particular kinds of testing practices. These testing practices constitute the work, which can be properly called "state work" (Harney, 2002), that renders "migration" and "citizenship" not only visible but also more broadly available for practical engagement and for public assessment and concern. State work is to be seen not as the work enacted by the state, but the ongoing process of state formation through particular kinds of labor (i.e., the production of state effects) in which the visualization, calibration, and administration of "migration" and "citizenship" serves as one of several occasions to produce state effects. In the midst of a presumed ongoing "migration crisis" in Europe, this is all the more relevant because, as Nicholas de Genova has said, there is a "plurality of contenders for sovereign power" (De Genova, 2017, p. 5). Likewise, such a perspective contributes to a growing awareness of "de-naturalizing the national" in migration studies (Amelina & Faist, 2012).

Rogier van Reekum has productively differentiated between the performative effects that pertain to the visualization and calibration of "migration" and what he calls the demonstrative effects that always accompany such performativity but exceed it by drawing attention to the specificity of the association that produces it (van Reekum, 2018). By enacting the public life of "migration," "the state" demonstrates the contingency and limitations of its own entanglement in the world of territories, materials, signs, and moving bodies. As van Reekum says:

> *migration is, through its varied and contradictory visualizations, also a test in which the capacities and tendencies of an association—"Europe", for instance—are demonstrated. (van Reekum 2018, p. 4)*

In close proximity to this conception of migration, this paper considers "citizenship" as consisting of a variety of tests. Its main aim is to highlight the different ways in which "the state" is enacted through a state work that takes the form of tests performed on "migrants." In the same process, the citizenship of the subjects of these tests is enacted in line with an increasing emphasis in Western countries on "earned citizenship" (Van Houdt, Suvarierol, & Schinkel, 2011).

The next section discusses the concept of "citizenship concours." The following sections discuss three different cases—border crossings, asylum decisions, and immigrant integration monitoring—as sites along a citizenship testing concours. The first of these cases is largely conceptual and theoretical in nature. The second is based on ethnographic data collected by Hertoghs (2019), and analyzed by her and myself (Hertoghs & Schinkel, 2018), and the third is based on a discourse analyses I conducted (cf. Schinkel, 2017).

2 | CITIZENSHIP AS TESTING CONCOURS

Testing has become central to the configuration of citizenship in Western countries. The spread of police work, as well as of securitization and surveillance, and the multiplication of borders all entail rhythmic setups in which statuses are confirmed or denied, bodies are included or not, and borders are crossed or not. To speak of a "testing concours" in the context of migration and citizenship is meant to evoke the etymological sense of "concours" as "moving, running, or flowing together," and as concerning both a gathering or accumulation and the tracing of a trajectory. Deriving from *concurrere* (to run together), it at the same time signifies competition and flight. In the context of testing, it recalls notions such as a *concours d'élégance*, a "parade of vehicles in which the entrants are judged according to the elegance of their appearance" (OED, 1923). Something quite close to this happens when physical appearance is compared across face and document at the border (my first case study in this paper). And it occurs when asylum applicants in the Netherlands are tested in terms of their sexual identity (second case). Immigrant integration surveys (third case) likewise have an element of "beauty contest," of being interpellated to present the self in a morally ideal form that mirrors as closely as possible a benchmarked national norm.

To consider citizenship as a trajectory of tests also means to see it in terms of indetermination, as practices for the calibration of potentiality that, as outcome, steer people in certain directions, enabling or preventing their circulation and their options. Citizenship as test so to speak resides in the impasse between test entry and test result, if only because the outcome of a test tends to give rise to another test. State work here consists of practices and sites of indetermination, of not-yet, and this waiting mode is indeed typical of common experiences of, and "at" the border. This is true for those who wait to have their passports checked at sites that stand in for "territorial borders" (which most of the time are not "at" the territorial border). But it is also true for those who, in a literal sense, have long since arrived but whose mode of living is nonetheless observed, by means of immigrant integration monitoring, as arrested, as not yet up to speed with modernity, not yet fully "arrived" (Boersma, 2019; Boersma & Schinkel, 2018). This monitoring amounts to a test of the degree to which they have actually "arrived," the term for which tends to be immigrants' "integration," which since 1994 is interchangeable in policy language with "citizenship" in the Netherlands (Schinkel & van Houdt, 2010).

The time of the test, then, condemns one to impasse, to a life on hold. However briefly in some cases (as in the passport check), in others this can be an extended modality of living (as in immigrants monitored in their "catching up" with modernity). Those who reside "in the hold" of such tests experience state work. Their lives are impending, pending the results of tests. But as is so often the case with experiments, tests, or pilots, one outcome of a test is that more tests are needed. Even if one passes a test (one is granted asylum), one is enrolled in yet another series of tests (citizenship or language tests, immigrant integration tests), and so those who emerge from the object of "migration" are not readily cut loose from it, but are rather enrolled in a testing concours that renders their citizenship available for public scrutiny.

A life in the hold of citizenship tests does not mean the total blocking of mobility. Rather, as Tazzioli has shown, it decelerates, obstructs and troubles the trajectories of migrants (Tazzioli, 2017). Precisely through this kind of "migration control," the state itself emerges as an actor with an address. Because I am interested in the ways testing configures the tester (in this case: "the state"), I will not pay attention here to something that, at first sight, would have been obvious to include: citizenship tests designed by governments to explicitly "test" the degree to which immigrants are "assimilated" to "host societies." In recent years, much research has appeared on these tests (Bauböck & Joppke, 2010; Etzioni, 2007; Michalowski, 2011; Monforte, Bassel, & Khan, 2019; van Houdt, 2014; van Oers, 2013). In the Netherlands, from which I will draw two of my three cases here, such tests have been subject to critique because of their unrealistic character: hardly any "native Dutch" would pass such a test (who knows what to call the sticks that separate your groceries from the next person at the supermarket?). In the Netherlands, citizenship tests are also conducted prior to immigration, at consulates in "sending countries." As naturalization tests, they can be quite consequential. Furthermore, as recently shown in the UK, they have real effects in terms of sorting inclusion and exclusion (Monforte et al., 2019). I will not discuss them extensively here for two reasons. The first is that they are part of, or tend to function as, a spectacle, and their attention in both critical and affirmative social scientific work always threatens to contribute to their spectacular mode of operating. The second reason not to include them here is more pertinent as it concerns the work that I would like the concept of a test to do here. This paper proposes thinking about citizenship and migration as trajectories interlaced with, and in part coordinated by, various testing practices that are not necessarily explicitly, deliberately, and consciously planned *as* tests. In particular in the context of this special issue, this is a way of drawing attention to something that tends to be little emphasized: it points at social practices that operate as tests without being explicitly conceived as such. There are practices and situations that function as tests without being imagined as tests. And there are demonstrative effects without demonstrations as the explicit aim. This paper therefore also seeks to push our understanding of social practices as tests, and to do so it is most fruitful to apply the concept of test to cases where that word is not necessarily used.

The following sections contain the three case studies to develop this conception of citizenship as testing concours: (1) identification and border crossing; (2) the establishment of applicants' sexual identity in the Dutch asylum procedure; (3) the monitoring of immigrant integration. These cases constitute three stops, and three configurations of the hold, in an ongoing concours of citizenship.

3 | FIRST STOP: CROSSING A BORDER AND THE TESTS OF IDENTIFICATION

In the prevailing vernacular of borders, the figure of the cross is omnipresent. Borders are crossed, and to be able to cross a border is a key feature of citizenship. Movement proceeds orthogonally to territorial lines. Yet an elaborate infrastructure is in place to provide this vernacular with its everyday assumed naturalcy. Infrastructure, by definition, is distributed, and so what gets to be abstractly considered as a line is rendered possible by a concrete but dispersed set of tools and practices. First of all, the line needs stand-ins in the form of gates, fences, walls, and signs. This spatial semiology cannot be collapsed into a single "line" but it can enact what is understood as the "line" of the border. It also indicates zones where this enactment takes place, and these zones are zones of the hold: people are held up, both in the sense of being slowed down and having to wait "at the border," and in the sense of potentially being held in rooms or cells, for interrogation or internment. So while on the one hand, the spatial semiology of the border is a gesturing that signals the waning of state power "at the border" (Brown, 2010), on the other hand it enacts zones where that power is displayed in one of its most raw forms. The zone of the hold, however, is a membrane. It exists to permeate, to sift and sort, and therefore it is a zone of testing. Testing results in speeding up again, in the continuation of circulation, or in an extension of the hold. The zone of the hold is structured in a binary way of "go" and "no go." It is important to note that this account conforms to the semiological gesturing that enacts the border-as-line. There are many other ways in which, ex post facto, the border is said to have been crossed, for instance when migrants turn up that have crossed the border "irregularly." The point here is precisely that the "regular" way of crossing a border requires a testing setup consisting of state work enacting both the border as a line, and the state as a sovereign body.

That binary setup is structured by a seemingly simple test: the match between document and person. Usually the document is a passport, and often it needs to additionally contain a visa or equivalent, which may or may not be a digital registration. The test thus consists of detecting a match between person and document/registration. The determination of a match, or a lack of one, is what is usually called "identification." Identification is a specific kind of testing procedure in that it involves the correlation between *two separate visual tests* (van Reekum & Schinkel, 2017). Each of these tests involves the attribution of identities to a body that, if identification succeeds, shows up twice and with a sufficiently high degree of resemblance. Identification thus involves two separate series that are correlated so that the same body yields similar results in the two series. For instance, facial attributes of a body are matched to the attributes of a picture in a passport, names pronounced are correlated with names registered, height observed is correlated with officially reported height, and reported place of birth is checked with registered place of birth. The simplest way of stating this, is that a border guard checks certain attributes of a person's passport with certain attributes of the person in front of them. But this is far from a trivial matter and in fact concerns the double construction of, on the one hand, a body assembled from attributes inferred from the passport, and, on the other, a body made up of observed attributes from the person "at" the border.

This conception of identification has two consequences. First, it means that identification is not a process of memory, an assessment of what is already familiar. Rather, it works towards the singular by means of the individuation or instantiation of identities (of race, ethnicity, gender, nationality, etc.) that are translocal—in fact, their translocal character is the very reason the test of "crossing the border" can be performed at all. And, second, it means identification is based on the comparison of one body with a multitude of other bodies, since this is the only way to individuate attributes to a body. Identification, in other words, entails the attribution of the singular by means of comparison from within the plural. It locates the individual within a multitude of comparable individuals that each participate differently in a set of shared attributes. Why does identification involve the correlation of two separate visual tests? The main reason is that it is governed by what Valentin Groebner (2007) has called *deceptibility*. Deceptibility exists in personal identification because identification works by attaching certain markers to bodies that are, of necessity, detachable. I cannot use my body as proof that I am who I say I am, not even in the case of biometrics. Biometrics is merely a way to—potentially—simplify the tests that border work performs.

Biometric data from my body need to be matched with their counterpart in the passport. The body cannot stand in for the image of itself. Historically, of course, the possible exception to this is the tradition of the "real" or "authentic image" of Christ (*vera icon*), as in the shroud of Turin, where the image is an index or imprint of the original and as such participates in it. But that tradition, too, is riddled with the paradox of establishing or contesting the authenticity of the authentic image (cf. Belting, 2005, p. 45). The *vera icon* problematic is repeated in the history of photography, in which William Henry Fox Talbot, for instance, considered "nature's hand" to be present (Bredekamp, 2010, p. 186). But while 19th-century theories of photography often placed themselves in the *vera icon* lineage, later theories contested photography's immediacy (cf. Bate, 2016, pp. 12–13), and with the separation of the photograph from its object, deceptibility re-entered consideration.

And so identification of necessity involves the matching between two separate series, between bodies and documents. A border guard does not *recognize* a person; she or he *identifies* the person by comparing body and document, and by comparing the body with a multitude of other bodies. Faces look like other faces, which is why they can be identified. But because the passport that goes with the face might be faked, stolen, or otherwise a case of deception, deceptibility governs the border test. For any singular identification, this might be a risk, and border guards are well aware of it. But for identification as such, it is a necessary condition. In order to check whether I am who I say I am, I have been given a document that is separable from my body, and this separability of necessity entails deceptibility, the possibility of my passing without the proper passport.

That deceptibility governs the border crossing as test therefore does not mean the test works towards the elimination of deception. On the contrary, it means that deceptibility remains and is the very condition of the test. Deceptive mimicry, deceit, camouflage, must of necessity be possible precisely because of the comparative character of identification. Persons are compared to other persons making up classes defined by nationality, birth, skin color, and so on. But passports, too, are compared to other passports to test their authenticity. Finally, passports and persons are compared, tested for sufficiency of resemblance in the face of a deceptibility that cannot be dispelled. If bodies, faces, and passports wouldn't look like other bodies, faces, and passports, there would be no possibility of identification. The production of the singular "identity," and along with it the production of a "go/no go," is a result of the comparison from within a plurality of bodies, faces, passports. The problem of identification is the problem of linking up bodies and documents, and the necessary separation of the body from the document means that, in each case, that linking up, and thus the individuation of identity, is what is worked towards, not what can be worked with. And the point of identification is to follow through on the consequences of deceptibility, namely the fact that a body can be *in* but not *of* the place it is at.

The go/no go that is the result of this testing procedure is based on the determination where a body belongs, and how it can, accordingly, circulate. But the test here is possible on the basis of a deceptibility it cannot cancel, and that it is not intended to cancel. Testing is possible on the basis of ambiguity, and its aim is a procedurally justified decision. Testing is thoroughly comparative here: it compares both across bodies and between bodies and documents. In order to stabilize such testing setups, state work has required the historical control of the means of deception. This involves commensuration and the establishment of what I have elsewhere called "comparity," a space of equivalence that enables comparison (Schinkel, 2016). The infrastructure of bodies, buildings, documents, and technologies that the "border" gathers requires comparity spaces, and the historical stabilization of such comparity spaces is another name for "state formation." The ability to test brings effects into the world, not the least of which is what Timothy Mitchell has termed "state effects": the effect of an always internal differentiation between "state" and "society" (Mitchell, 1991). The state's ability to establish a "border crossing" by means of a test of identification is possible on the basis of the spreading of carry-on markers of identity (passports) across a population, so that an individual within that population, when attempting to cross the border, is always already commensurated in the sense that the comparative testing needed for identification can be carried out. This is why one strategy among those seeking to cross borders "irregularly" is to get rid of their passports. Without the required comparability, states have a hard time "processing" persons on the move. The absence of a passport tends therefore to result in longer periods of remaining in the hold of the state, during which other tests are performed to establish people's identity and their deservingness to enter the country—this is what my second case will deal with. For now I would underscore that, because comparative testing "at the border" is what configures the identification of citizens, it also configures their potentialities (go/no go), their capacities to circulate, or their restrictions, their "being on hold" and their being *in* the hold that "the border" constitutes for many today.

Citizenship testing thus begins "at the border," with the very determination of being able to do so, or of having done so, legitimately. I focus here on the logic of identification but do so at the cost of abstracting what actually happens "at" the border, where persons and bodies are raced and gendered. It is crucial therefore to develop this analysis in line with a growing body of work on migration and intersectionality (cf. Amelina, 2017; Bastia, 2014; Bastia, Piper, & Carrón, 2011). The role of gender and sexuality in migration is part of my second case in this paper. For a next stop along the testing concours of citizenship ensues for many today who flee their country and attempt to cross a border to build up a new existence in a foreign country by applying for asylum.

4 | SECOND STOP: TESTING DESERVINGNESS IN LGBT ASYLUM APPLICATIONS

For those who apply for asylum, the hold is a very concrete reality. In most, if not all, Western countries, the asylum applicant is interned, placed in a holding facility that is often called "asylum detention." Ethnographic accounts of asylum detention attest to the experience of lives "on hold," in extended periods of passive waiting and boredom (Hertoghs, 2019). But here, too, the hold is a time and place of testing. And here as well, the outcome of the test is a decision: either a refugee is rejected (and then either denied the status of "refugee" or referred to another country for assessment), or a refugee is accepted and granted asylum. The test here revolves around the determination of the "deservingness" of asylum. To illustrate how this takes place, I draw on work done by Maja Hertoghs, who studied the Dutch asylum procedure up close in two asylum detention centers between 2014 and 2016 by means of an extended ethnography of the Dutch Immigration and Naturalization Service (IND), involving observations of hearings of asylum applicants, interviews with IND officials and asylum lawyers, observations of meetings between IND officials and document analysis of hearing reports and asylum decisions (Hertoghs, 2019).

Deservingness has a variety of criteria (Bohmer & Shuman, 2008) that concern the establishment of a match between individual attributes and formal (legal and procedural) reasons for granting asylum. The Netherlands is one of several countries where a person's sexual identity can be reason for granting asylum, specifically in the case of lesbian, gay, bisexual and transgender (LGBT) identities. An asylum seeker ascribing to one such identity and claiming to be prosecuted or otherwise in danger because of it in their country of origin, may be eligible for asylum in the Netherlands. However, the IND needs to establish the veracity of both these claims: it needs to ascertain the plausibility of an LGBT identity, and it needs to ascertain the plausibility of this being dangerous for the asylum applicant. The latter might be the easiest part of "the procedure" (the emic name for the asylum procedure). Lists of countries where LGBT persons are in danger are used, so that, for instance, a gay man from Jamaica is almost certain to be granted asylum. The first, establishing sexual identity, is the tricky part. It is also the testing part of the procedure, since that which the state regards as a person's sexual identity—meaning here: the legitimate affordances of that identity in terms of circulation and citizenship—is an *outcome* of the procedure. As the IND instructs its officers: "When an asylum seeker claims he is LGBT, it is up to the asylum seeker to substantiate that claimed homosexuality" (quoted in Hertoghs & Schinkel, 2018). From the commensuration of LGBT to "homosexuality" in just one sentence, one can surmise that much nuance disappears, and yet the IND is adamantly intent on finding the veracity of accounts of sexual identity.

In LGBT cases, the procedure thus becomes a test of one's sexual identity, and this means the asylum applicant is called upon to provide a true and authentic account of their sexual identity. And as Judith Butler has argued, such account-giving is inherently paradoxical, since it needs to appeal to social norms, from which the terms in which persons give an account of themselves are derived (Butler, 2005, p. 1). The singularity of the self can, therefore, not be expressed other than by abstracting from that singularity, folding it into norms, expectations, and categories in a narrative account that undercuts the singular. We thus encounter a situation similar to the test of identification discussed in the first case: the singular or individuated can only emerge as the outcome of a test that abstracts from the singular and compares with the general, with categories and norms derived from populations, pluralities. And here, too, deceptibility is key in the sense that one's sexual identity cannot be taken at face value, nor can it be trusted as a mere declarative statement. What the IND seeks to establish is precisely the veracity—the plausibility for all practical purposes—of such statements. Deceptibility is very much present in the everyday workings of the procedure in the form of the suspicion that informs everything IND officers do. This suspicion is considered part of their *raison d'être*, as it is the only guarantee of justice in the face of the perceived need to separate the truly "deserving" from the "undeserving" asylum applicants. In the latter case, "state generosity" is considered abused, and so the establishment of the truth of one's sexual identity happens through a skeptical state epistemology.

But how do we go about this? How do we establish a person's sexual identity? Another paradox ensues, because this happens first of all by conceiving of that sexual identity as an innate, deep-seated personal core or infrastructure of selfhood (Hertoghs & Schinkel, 2018). Yet secondly, it happens by the demand of a credible *performance* of that innate sexual essence. The test of deservingness in the case of LGBT asylum seekers comes down to this, to being able to perform a deep-seated identity. Of course this performance remains primarily discursive, though embodied comportment is part of it and it is reported on by IND officers working towards a decision on an asylum case. And part of the performance is the ascription to the essentialist conception of sexuality the IND deploys. This results in questions about, and detailed accounts of, moments of realization of one's sexual identity and, where applicable, coming out or getting caught or recognized as being what one is. In a remarkable inversion of the suspicion power tends to display, the IND's default suspicion is that a person actually belong to the dominant category of "heterosexuality." The LGBT categories are the non-normal ones (which is perhaps why they can just as easily become commensurated to "homosexuality"), and to belong in one of them needs to be proven. Even though here, too, ambiguity remains and deceptibility governs the test, an account is to be crafted in such a way that veracity may be argued for in a procedural sense, informing a decision. Because everything hinges on the credibility of performance here, that performance is rehearsed. Lawyers and volunteers from the Refugee Council prep an applicant for the kinds of questions to expect at the hearings, and try to actively shape responses and comportment.

Applicants pass the test of deservingness when their accounts are "credible," which means they contain enough details, are not frequently "vague" and are coherent overall as elements of an asylum case. While all these are subjective judgments, the IND goes through elaborate efforts to establish the "objectivity" of the decision. This happens through working with different IND officers at different hearings, by having a decision made by officers not present at these hearings, and by allowing for a preliminary decision ("voornemen") to be challenged by the applicant's lawyer, after which yet another IND officer weighs the evidence. At stake in the test is often an embodied understanding of the meaning of "homosexuality" that conforms to the IND's highly heterosexual understanding of this. Most enlightening in this respect is a case where asylum was refused and the arguments for this were given in the written decision:

> *If the applicant were truly homosexual, he would have been able to credibly speak from his inner feelings and observations about what homosexuality contains, certainly now, after he has been in Europe for many years. (quoted in Hertoghs & Schinkel, 2018)*

What becomes apparent here is that the test of one's sexual identity (which is at the same time the test of one's deservingness of—partial—citizenship) centers on a conception of sexuality as inner core and infrastructure of personhood. Sexuality is like a mold into which a person's life unfolds. Because this inner core impresses itself on so much of life, both inner and outer, it is a deep structure that nonetheless has surface traceability. So if an account of (in this case) a homosexual identity does not conform to the very conception of sexuality the IND adopts, it may conclude that the identity if feigned. If it were a deep infrastructure of selfhood, it would surface and be coherently traceable. In the case mentioned here, the IND explicates:

> *His statements involving what homosexuality contains are least of all convincing. In that context he states … that homosexuality means that a man sleeps with a man. Homosexuals do everything that a man normally puts into a woman into a man, they kiss everywhere, and dress up like women. … It is obvious that the applicant presents a very flat and stereotyped image of what homosexuality contains. An image that might be expected of people that mock people with a homosexual nature or who are ignorant of what homosexuality truly is. It is not credible that the applicant, if he really was homosexual, would present such a superficial and shallow image of what homosexuality is. (quoted in Hertoghs & Schinkel, 2018).*

What is striking here is that the IND, which itself espouses a markedly binary, heteronormative and hence stereotypical conception of what it calls "homosexuality," demands of gay applicants a shared non-stereotypical view of homosexuality. The performance of homosexuality here emerges as the presentation of the right image of homosexuality, which, while binary and heteronormative, must be judged to be too stereotypical for it to pass off as evidence of homosexuality. In this case, that also means the IND officer states that the absence of romantic feelings by the applicant for his sexual partner are considered non-credible.

For an LGBT asylum applicant, then, the hold presents itself as a prolonged test in which an image of sexual identity gradually surfaces in written reports of hearings, in a preliminary decision and in a final decision of deservingness or undeservingness. This test is riddled with ambiguity, and ultimately it is clear for all participants that the facts cannot decide the issue, just as they cannot do so in scientific tests (cf. Latour, 1987). As in science, the "facts" are the outcome of testing and deliberating among the testers. But unlike in science, what allows the facts to emerge, at least for all practical purposes, is a delegated form of sovereign power that crystallizes, very classically, in what is throughout the procedure called "the decision." That decision needs to be well founded, and that is why the test exists, and why suspicion—the desire to avoid error—persists throughout the test. But as is the case with all decisions, the decision is ultimately unfounded, and needs to cling to procedural correctness—proper testing conduct—in order to uphold itself in the face of ambiguity and ultimate arbitrariness. This is the case for all asylum requests, not just those based on sexual identity. Any claim to asylum is met with suspicion and refugeeness—and hence deservingness of asylum status—is met with suspicion and put to the test. As in the case of identification at the border, then, the state emerges in, or better yet as, a testing capacity that configures questions of sovereignty prompted by the circulation of bodies as questions of knowledge of the other. And in both cases, ambiguity is not something the state seeks to eradicate. Rather, ambiguity is what allows the state to manifest itself in the first place. As Rogier van Reekum says, if the state were to really eradicate doubt, "it would progressively rid itself of investigative capacities … Government would truly become administration" (van Reekum, 2018, p. 3). Or in other words, the enactment of borders, in these cases, at the same time demonstrates the specific kind of association that makes claims to statehood—that demonstration is the *effect* of state work, but that is a work which can get to be credibly called "state" only *after* the effect.

5 | THIRD STOP: ASSESSING "IMMIGRANT INTEGRATION"

My last case pertains to what is called, throughout Western Europe, "immigrant integration." It pertains to those who crossed the border, who have rights, even to many who never migrated, but who are nonetheless observed as different, as not merely "in need of" something called "immigrant integration," but as simply tested for it, observed *in terms of* "immigrant integration." Generally, "immigrant integration" is seen in terms of "adjustment to society." What immediately becomes clear, then, is that immigrants, as well as their children, tend to be regarded as somehow external to the domain of "society" (Schinkel, 2017). Indeed, this is why it is common, for instance in the Netherlands, to speak of immigrants as residing "outside society" or, as in Germany, as making up a "parallel society" (*Parallelgesellschaft*). Of course they are neither in France or Belgium, and if they were, they would not be objects of problematization and integration testing in the Netherlands or Germany (Boersma & Schinkel, 2018). So the fact that they are, somehow, "here" (in the Netherlands) is constitutive of the assumption that they are really not quite here (in "society"). This is why there is regular (longitudinal) state-sponsored and state-organized integration monitoring, and it is also why this monitoring is a form of testing. Immigrants and their children *are* "here"—yet what integration monitoring seeks to assess is whether they are really *here* or whether they are, in fact, still in the process of "arriving" (Boersma, 2019; Boersma & Schinkel, 2018). Immigrant integration monitoring closely mirrors the concerns of state policies (Favell, 2003), and this is not very surprising given the fact that often states are the explicit clients, sponsors, and even conductors of the work that integration monitoring constitutes. It relies, moreover, on surveys, population registries, and employment data that are frequently state-owned and state-run. In the Netherlands, the Institute for Social Research (SCP) and the Central Bureau of Statistics (CBS) are the main institutions (both are state institutions) who conduct longitudinal integration monitoring surveys and studies, and my analysis here rests on a discourse analysis of work done by these institutions. This analysis proceeds with a Foucaultian emphasis on the production of objects of problematization, and on the stabilization of encompassing onto-epistemic conceptions of "modernity" and "society," to the point where such conceptions are naturalized to such an extent that they are unavoidable presuppositions of talk and action precisely under conditions of ongoing endangerment of their naturalcy and stability. The resulting state work can thus be seen as a contemporary modification of the state-enacted racism Foucault analyzes in *"Il faut defendre la société"* (Foucault, 1997).

Integration monitoring occurs in a variety of ways, centering mostly on "socio-economic integration" and "socio-cultural integration." For brevity's sake, I shall focus on two forms of the latter kind here. The first pertains to the degree of "modernity" that subjects of integration monitoring are tested to display. This happens in various ways, one of which is to consider their "secularity," which is defined (reductively) as being non-religious and/or being non-churchgoing, even though "church" here mostly stands in for "mosque," and those who might go to church are often not subjects of integration monitoring. The Netherlands is then defined as "secular," and the fact that about a million people identify as Muslims is not taken as an indication that the country has changed in terms of its religiosity. Rather, the conclusion of this type of measurement is that a million people are at a greater or lesser remove from "Dutch society." The test that integration monitoring constitutes, operates as a kind of shibboleth here: if a certain feature is present, then one cannot "pass" as an unproblematic "member of society." The conclusion is that migrants that are more "religious" than "society" is assumed to be (though the latter is not measured in the same monitoring practices) are less "modern," and hence show a "lag," as the modernization literature of the 1950s and 1960s often concluded (cf. Lerner, 1958). In a similar way, anthropologist Johannes Fabian (2014) has argued that anthropology has long assumed that difference constituted a temporal lag. Here, the spatial metaphor of "distance from society" is at once coded as temporal lag (as not being with the program of modernity). Such versions of "difference" are the outcome of tests of difference that ratify themselves because they are conducted solely among those who are considered to "differ."

Something similar occurs in the measurement of "contacts with members of other ethnic groups," which is a second way of establishing "socio-cultural integration." Here, people get lumped into "ethnic groups" by means of pure ascription, and these are compared with "autochthonous Dutch" in terms of the number of contacts members of these groups have with members of other groups. Though officially rescinded by several government research bodies in 2016, the chthonic language of "autochthonous" (literally: from this soil) and "allochthonous" (from other soil) has been pervasive in the Netherlands from at least 1989 (when the Scientific Council for Government Policy started using it) until the present (cf. Geschiere, 2009). The fact that many if not most "allochthones" were and are born on Dutch soil indicates the way these concepts always already operated as strategic markers of problematization. In the analysis of "contacts," it turns out that "autochthonous Dutch" have the least contacts with members of other groups, that is, white Dutch citizens, relatively speaking, keep to themselves. But this is not considered a test of their socio-cultural integration or a demonstration of their lack thereof. Rather, what happens is that the tables are turned, quite literally when tables with numbers of contacts are considered: when reporting results, it appears that "contacts with members of other ethnic groups" in fact means "contacts with autochthonous Dutch." Somehow, those with the least inter-ethnic contacts are promoted to the rank of unproblematic members of society, the benchmark for others whose contacts with those autochthonous Dutch constitute the test of their "integration." Of course, it is difficult for members of a variety of ethnic groups to be in contact with autochthonous Dutch, if the latter have little contact with members of other ethnic groups: this is the same phenomenon. So between the introduction of this measurement of socio-cultural integration and the presentation of its results, this measurement has turned into a test of the degree to which non-autochthonous Dutch approach autochthonous Dutch, who have become the benchmark for "society." But this test, which assumes and measures assimilation, can *only* find difference. This is a difference in degrees. One can be more or less different, more or less well integrated, but there will, due to the definition of groups and the elevation of one of them to neutral reference category, of necessity be difference.

Testing here, as often, pertains to an evaluative assessment of how well specific subjects are doing on selected variables. But the evaluation requires a comparative aspect, a benchmark. A comparity space is silently assumed here, and it is called "society" and consists of "autochthonous Dutch" citizens. For non-autochthonous Dutch, to be a citizen is to approach, as much as possible, a benchmark one cannot possible ever achieve. In other words, the entire setup of integration testing tacitly assumes, from the outset, an image of "society." This is never defined, though it figures when "secularity" becomes the benchmark for a test of religiosity as a marker for a degree of "integration." And it figures, in the sense of emerging from the background in a figure/ground reversal, when "contacts with autochthonous Dutch" become coded as the norm for being "integrated" (albeit not for autochthonous Dutch themselves).

Testing can be considered in terms of revealing the unknown qualities of a certain object. At the same time, throughout the history of testing (most significantly psychological and educational testing, but also in much scientific experimentation), the quality involved in the test was considered known in advance. Testing was, in such cases, configured on the basis of its being sensitive to this, but the quantity or degree to which this quality (property, capacity) existed was the unknown and the hoped-for outcome of the test. In this case, the "quality" amounts to a statement of difference, and testing pertains to the quantity of difference, the degree of "distance from society," the "degree of integration." Of necessity, difference, or a "lag," is the outcome of testing, because testing occurs on the basis of a comparison to a reference category over against which comparative categories are a priori defined as different. In other words, the test produces a "lag," and necessarily so, as it "benchmarks" on a category to which the tested subjects or categories are a priori *defined* as different. In fact, their prior calibration as different is the very reason for being tested. Testing is thus a highly tautological affair here. It has all the characteristics of a rhetorical device that enables the state to problematize immigrant groups with good reason, since tests governed by "objectivity" show difference.

And so, in a way, one could say that the "real" test resides in whether one is tested at all. Those who are tested will a priori be assessed as beyond the realm of "society" proper, however "well integrated" they are. Those who aren't tested for their individual or group-level "difference" receive what I have elsewhere called a "dispensation of integration" (Schinkel, 2017).

So while immigrant integration is "measured," it is indeed more accurate to say that it constitutes a test of the degree of difference of those measured. The real conceit here would be to believe that, given existing conceptions of immigrant integration, something might be learned that wasn't always already internal to the monitoring system itself. Similarly, Niklas Luhmann has argued that the surprise that may follow a scientific test does not present a "realist" argument against the self-referentiality of scientific knowledge:

> *The system can gain confirmation by what it already knows: that it operates in an environment … It thereby learns nothing about that environment—except for the fact that it affirms or negates the system's expectations. The system only ever tests its own expectations. (Luhmann, 1990, p. 261)*

In the case of immigrant integration monitoring, at least, such appears the case, since a "self-projected meaning" underlies the test of immigrant integration from the beginning: "integration" is always already conceptualized as that which supposedly (though this isn't tested) characterizes the ideal of whiteness, of native Dutch with jobs, a nonreligious life characterized by tolerance but also by the exclusivity of their social contacts among themselves.

If integration testing is considered as one enactment of public life through state work, it becomes apparent that the state here attains a strongly moral character. Integration monitoring is a moral monitoring (Schinkel, 2017), and it has the effect of a moral purification of the imaginary domain of "society." For it appears that, as soon as problems appear, they concern people not fully "integrated in society." That "society" itself, then, has no problems. If crime is an issue, it is not "society's" issue, but that of individuals external to it, at a remove from it, who are unintegrated and reside "outside society," even though prisons are generally regarded as "in" society and a vital part of it even. Likewise, "socio-economic integration" entails the assumption that those at the bottom of what is a socio-economic hierarchy are really external to the hierarchy. Poverty is not a problem "in society," but a problem of individuals "outside society," and they are outside society for the very reason that they are poor, hence insufficiently "integrated." "Integration" thus becomes applicable in all cases that deviate from prevailing norms, and it thereby becomes a way of morally cleansing the imaginary of "society." Considering that integration monitoring is a form of state work, it is then important to recognize the moral work that the state performs in order to enact the very imaginary of society.

Testing here configures differentiation vis-à-vis a "normal" benchmark, but that means it is relational and necessarily also pertains to the configuration of the benchmark. That is to say that the quality or capacity tested for, and the characteristics of the tester and the infrastructure of testing are not stable objects that precede the test but derive the plausibility of their assumed stability from the test. What "society" is becomes communicable only by way of displaying difference. This requires the a priori inclusion of "society" as a testing benchmark, but that is not how society becomes publicly available as a reference object. Even though the test always and of necessity required a tacit inclusion of "society," the public availability of this social imaginary emerges as effect *of* the test. In order for that to happen, there has to be a hold. People are imagined as in the hold of their tradition, of religion, of their ethnic minority, their neighborhood, their culture, or their sending country. In short, they are in the hold of their "background," which starkly contrasts with the "neutrality" and whiteness of "society." But that imagination itself means that those people, when they are tested for their integration and demonstrated to be in the hold of their "background," remain in the hold of the very act of coding and imagining. And it means they are in the hold of the state conducting ostensibly epistemic tests upon them.

And this hold can persist into the future along the long lines of migratory affect. There are what one can call "nth generation immigrants" (Schinkel, 2017): those who never actually migrated but descend from "migrants" (but only post-World War II migrants, not the politically neutralized hominid dispersal of the predecessors of "native citizens"). These continue to produce progeny that constitute a potential testing reservoir, since culture and tradition can be tenacious, and assimilation can take generations. Testing for "immigrant integration," then, allows what postcolonial literary scholars call "arrival narratives" to be extended (Boersma & Schinkel, 2018; Quayson & Daswani, 2013). Even at this third stop of the citizenship concours, when immigrants are "settling," they are again stirred into motion in a work of imagination for which testing provides the evidence. Those who migrated, and their children, are still, and potentially perpetually, arriving. State officials measuring integration here show their Derridean feathers: *différer* (to differ) also means *déférer* (to defer to, to postpone). In this sense, a hold persists for subjects of "immigrant integration." They resemble the man waiting before the law in Kafka's parable *Vor dem Gezetz*, who forever remains outside to be tested, in a trial before even getting to the law, only to find out that the real test consisted in being tested or not. Perhaps the best strategy for citizenship when confronted with immigrant integration surveys would be what Kafka (1996, p. 380) writes in another one of his short stories, *Die Prüfung* ("The Test"): "He who does not answer the questions has passed the test."

6 | CONCLUSION: ON TESTING

By discussing these cases, this paper seeks to primarily contribute to work on the entanglement between the state and migration. A secondary contribution it makes is to understandings of testing in political settings. In this concluding section, I in particular draw out some points regarding the relationship between the state and testing.

Seeing citizenship testing as state work means considering it as “state labor under capitalist conditions” where, as Stefano Harney has said, “state work is the term for that kind of labor that most knows itself as comparison, equivalency, and exchange in the social realm” (Harney, 2002, pp. 10–11). Throughout the testing practices discussed here, comparison, equivalency, and exchange figure prominently. Testing thus involves comparity work, the work of crafting a *tertium comparationis*, a shared space of comparison (Schinkel, 2016). This means that the tests are always already inscribed in a calibrated conception of, in this case, a national norm. Especially in the last two cases it becomes very apparent that testing occurs within an overall orientation toward norms of heteronormativity, nativity, and whiteness.

This “always already” is possible because the practices described here are forms of state work. Testing, like experimentation, involves attention and experience. It's a very pragmatic thing, and if state work amounts to testing in the cases discussed, that means “the state” gets to be seen in a pragmatist conception as a practical achievement. But even here, something of what Cassirer (1946) called the “myth” of the state remains. In their history of wonders, Daston and Park argue that natural wonders, secrets and experiments are all phenomena of experience, and all were “proven recipes for medical and magical preparations, [which] often drew on the occult properties of natural substances, and they were excluded from natural philosophy for the same reasons” (Daston & Park, 1998, p. 129; cf. Hadot, 2004, pp. 165–166). And upon closer scrutiny, the practices that constitute “migration” as a testing concours all have their “occult properties.” They are strikingly characterized by imaginaries of "objectivity,” and they appear as knowledge practices. But all the while, they are also ways of proceduralizing what necessarily escapes: they identify in the face of deception and ambiguity, they establish sexual identity as if it were a deep-seated infrastructure of being, and they mobilize a mythic “society” in its absence by rendering difference available for attention and further problematization. The kinds of state work discussed here keep a variety of factishes afloat, such as a “society” with “members,” people that “are” (or are not) “modern.” And these factishes, in turn, are assumed as the epistemic foothold for tests of the veracity of identity and citizenship of those enmeshed in the thing called “migration.” They become, in other words, ways of processing variety.

Testing is an adequate practice for processing variety, because it involves differentiation. The necessity of differentiation is the working assumption of all testing; entities with necessarily homogeneous qualities and quantities need not be tested, they can be classified. Because testing involves differentiation it is a practice that figures in the workings of what Deleuze and Guattari (2004) call the “binary machines” that produce the lines of “hard segmentarity” that run through social life: gender, sex, race, class, citizenship, and so on. And these are not mere classifications. As this paper has sought to illustrate, a variety of epistemic practices exist to construe such segmentary lines as “outcomes” of tests. And because these tests are conducted in and through state work, their outcomes are not evenly challengeable. They depend on decisions, that is, on the authority to, as Marylin Strathern has said, “cut the network” (Strathern, 1996). Considering citizenship as a testing concours means to thus situate and relocate what is generally called “state authority.”

Key to the relationship between the state and citizenship testing is that the configuration of tests entails the co-configuration of the tester. This also means that tests need not be, as Hanson (1993, pp. 18–19) holds, “representational techniques” in which the information collected is not the goal but a representation of it (as an IQ test, for instance, tests not the ability to answer questions but a more general thing named “intelligence”). Yet considering citizenship as testing concours illustrates how “migration” and “citizenship” are enacted through a state work that co-configures a social “inside” that is presumed as the pre-existing ground of the test, whether it goes by the name of “nation” or of “society.” This is because the citizenship testing has *public* effects: they both render citizenship and migration available as objects of concern to publics, and they render certain publics available to observation by themselves and others.

One potential limitation of a focus on testing is that one tends to consider setups and practices as consciously designed tests. Thereby, a cognitivist bias is smuggled into the analysis of testing practices. And then "testing" threatens to become a form of deliberation by other means—something that is especially consequential when testing and experimentation are proffered as alternatives (one might say functional equivalents) of liberal, deliberative models of politics. This may be due to the legacy of STS and science studies, and their transposition of the vocabulary and practice of the laboratory (cf. Latour, 1988). In science, tests, experiments, demonstrations, and pilots yield surprise (i.e., information), repetition (redundancy), and ambiguity, but they only ever do so in setups controlled and consciously planned as tests. But the productivity of concepts like test, demonstration, and experiment might be maximized if using them can yield fresh perspectives on phenomena not consciously shaped as tests, demonstrations, or experiments, phenomena that are neither evidence of a "test drive" or part of a rhetoric of testing (Ronell, 2005). A final contribution this paper seeks to make to work on testing lies, then, in considering practices as tests that are not necessarily imagined as such by those involved in them. Paying attention to citizenship and migration in terms of testing reveals that testing involves boundary work, normative benchmarks, and factishes that all concern the configuration of the "test"—in this case: the state and the idealized fiction of the "native citizen" it deploys but cannot substantiate (and thus govern) without testing others.

This also draws attention to an ethics of testing. The state work through which citizenship is configured as a testing concours draws people into policing relations to specific bodies exhibiting specific kinds of mobilities. To presume to be able to test, to assume the power to demonstrate, may be disguised in the objectivity of a modest witness, but it enacts sovereign power and configures the relational as a series of asymmetric capacities to extract exposition effects from bodies. In the cases described here, these asymmetric capacities and the normative benchmarking this comes with, shape a particular ethics of testing in which relationality gets done in asymmetric, suspicious, decisionist ways. Yet, as we know, there are other modalities of relationality. To shape relationality as being-in-common, another meaning of "testing" might be more pertinent: testing as displacing or suspending boundaries, as the open and demonstrative commoning of a form-of-life that does not pretend to already know how to live together.

DATA AVAILABILITY STATEMENT

The ethnographic data that support the findings of this study are not publicly available due to privacy or ethical restrictions. The data on immigrant integration surveys are in the public domain.

REFERENCES

Amelina, A. (2017). *Transnationalizing inequalities in Europe: Sociocultural boundaries, assemblages and regimes of intersection*. London, UK: Routledge.

Amelina, A., & Faist, T. (2012). De-naturalizing the national in research methodologies: Key concepts of transnational studies in migration. *Ethnic & Racial Studies*, *35*(10), 1707–1724.

Bastia, T. (2014). Intersectionality, migration and development. *Progress in Development Studies*, *14*(3), 237–248.

Bastia, T., Piper, N., & Carrón, M. P. (2011). Geographies of migration, geographies of injustice? Feminism, intersectionality, and rights. *Environment and Planning A: Economy and Space*, *43*(7), 1492–1498.

Bate, D. (2016). *Photography: The key concepts*. London, UK: Bloomsbury.

Bauböck, R., & Joppke, C. (2010). *How liberal are citizenship tests?* San Domenico di Fiesole, Italy: Robert Schuman Centre for Advanced Studies.

Belting, H. (2005). *Das Echte Bild. Bildfragen als Glaubensfragen*. München, Germany: C.H. Beck.

Boersma, S. (2019). *Dissociating society: Knowledge, affect and performativity in immigrant integration monitoring* (Dissertation). Erasmus University Rotterdam, Rotterdam, Netherlands.

Boersma, S., & Schinkel, W. (2018). Imaginaries of postponed arrival: On seeing "society" and its "immigrants". *Cultural Studies*, *32*(2), 308–325.

Bohmer, C., & Shuman, A. (2008). *Rejecting refugees: Political asylum in the twenty-first century*. New York, NY: Routledge.

Bojadžijev, M., & Karakayali, S. (2010). Recuperating the sideshows of capitalism: The autonomy of migration today. *e-flux* 17. Retrieved from https://www.e-flux.com/journal/17/67379/recuperating-the-sideshows-of-capitalism-the-autonomy-of-migration-today/

Bredekamp, H. (2010). *Theorie des Bildakts*. Berlin, Germany: Suhrkamp.

Brown, W. (2010). *Walled states, waning sovereignty*. New York, NY: Zone Books.

Butler, J. (2005). *Giving an Account of Oneself*. New York: Fordham University Press.

Cassirer, E. (1946). *The myth of the state*. New Haven, CT: Yale University Press.

Daston, L., & Park, K. (1998). *Wonders and the order of nature*. New York, NY: Zone Books.

De Genova, N. (2013). Spectacles of migrant "illegality": The scene of exclusion, the obscene of inclusion. *Ethnic & Racial Studies*, *36*(7), 1180–1198.

De Genova, N. (Ed.) (2017). Introduction. The borders of "Europe" and the European question. In *The borders of "Europe": Autonomy of migration, tactics of bordering* (pp. 1–35). Durham, NC: Duke University Press.

Deleuze, G., & Guattari, F. (2004). *Anti-Oedipus: Capitalism and schizophrenia*. London, UK: Continuum.

Dijstelbloem, H., & Broeders, D. (2014). Border surveillance, mobility management and the shaping of non-publics in Europe. *European Journal of Social Theory*, *18*(1), 21–38.

Dijstelbloem, H., van Reekum, R., & Schinkel, W. (2017). Surveillance at sea: The transactional politics of border control in the Aegean. *Security Dialogue*, *48*(3), 224–240.

Etzioni, A. (2007). Citizenship tests: A comparative, communitarian perspective. *The Political Quarterly*, *78*(3), 353–363.

Fabian, J. (2014). *Time and the other. How anthropology makes its object*. New York, NY: Columbia University Press.

Favell, A. (2003). Integration nations: The nation-state and research on immigrants in Western Europe. *Comparative Social Research*, *22*, 13–42.

Foucault, M. (1997). *"Il faut defendre la société". Cours au Collège de France. 1976*. Paris, France: Gallimard & Seuil.

Geschiere, P. (2009). *The perils of belonging: Autochthony, citizenship, and exclusion in Africa and Europe*. Chicago, IL: University of Chicago Press.

Groebner, V. (2007). *Who are you? Identification, deception, and surveillance in early Modern Europe*. New York, NY: Zone Books.

Hadot, P. (2004). *Le voile d'Isis: Essai sur l'histoire de l'idée de Nature*. Paris, France: Gallimard.

Hansen, T. B., & F. Stepputat (Eds.). (2005). *Sovereign bodies: Citizens, migrants, and states in the postcolonial world*. Princeton, NJ: Princeton University Press.

Hanson, F. A. (1993). *Testing testing: Social consequences of the examined life*. Berkeley, CA: University of California Press.

Harney, S. (2002). *State work: Public administration and mass intellectuality*. Durham, NC: Duke University Press.

Hertoghs, M. (2019). *Affects of suspicion: An ethnography of compassion, objectivity and state power in the Dutch asylum procedure* (Dissertation). Erasmus University Rotterdam, Rotterdam, Netherlands.

Hertoghs, M. A., & Schinkel, W. (2018). The state's sexual desires: The performance of sexuality in the Dutch asylum procedure. *Theory & Society*, *47*(6), 691–716.

Isin, E. F., & G. M. Nielsen (Eds.). (2008). *Acts of citizenship*. London, UK: Palgrave Macmillan.

Kafka, F. (1996). *Die Erzählungen*. Frankfurt/M., Germany: Fischer.

Latour, B. (1987). *Science in action: How to follow scientists and engineers through society*. Cambridge, MA: Harvard University Press.

Latour, B. (1988). *The pasteurization of France*. Cambridge, MA: Harvard University Press.

Lerner, D. (1958). *The passing of traditional society: Modernizing the Middle East*. New York, NY: The Free Press.

Luhmann, N. (1990). *Die Wissenschaft der Gesellschaft*. Frankfurt/M., Germany: Suhrkamp.

Marshall, T. H. (1963). Citizenship and social class. In S. M. Lipset (Ed.), *Class, citizenship, and social development: Essays by T.H. Marshall* (pp. 71–134). New York, NY: Doubleday.

Mezzadra, S. (2010). The gaze of autonomy. Capitalism, migration and social struggles. In V. Squire (Ed.), *The contested politics of mobility: Borderzones and irregularity* (pp. 121–142). New York: Routledge.

Michalowski, I. (2011). Required to assimilate? The content of citizenship tests in five countries. *Citizenship Studies*, *15*(6–7), 749–768.

Mitchell, T. (1991). The limits of the state: Beyond statist approaches and their critics. *American Political Science Review*, *85*(1), 77–96.

Mitchell, T. (1999). Society, economy, and the state effect. In G. Steinmetz (Ed.), *State/culture: State formation after the cultural turn* (pp. 76–97). Ithaca, NY: Cornell University Press.

Monforte, P., Bassel, L., & Khan, K. (2019). Deserving citizenship? Exploring migrants' experiences of the "citizenship test" process in the United Kingdom. *British Journal of Sociology*, *70*(1), 24–43.

Moulier Boutang, Y. (1998). *De l'esclavage au salariat: économie historique du salariat bride*. Paris, France: Presses Universitaires de France.

Nyers, P. (2015). Migrant citizenships and autonomous mobilities. *Migration, Mobility, & Displacement*, *1*(1), 23–39.

OED (1923). *Oxford English Dictionary*. Oxford: Oxford University Press.

Papadopoulos, D. (2018). *Experimental Practice: Technoscience, alterontologies, and more-than-social movements*. Durham, DC: Duke University Press.

Papadopoulos, D., & Tsianos, V. S. (2013). After citizenship: Autonomy of migration, organisational ontology and mobile commons. *Citizenship Studies*, *17*(2), 178–196.

Pelizza, A. (2019). Processing alterity, enacting Europe: Migrant registration and identification as co-construction of individuals and polities. *Science*, *Technology & Human Values*, *45*(2), 262–288.

Quayson, A., & G. Daswani (Eds.). (2013). *A companion to diaspora and transnationalism*. Oxford: Wiley.

Rodriguez, N. (1996). The battle for the border: Notes on autonomous migration, transnational communities, and the state. *Social Justice*, *23*(3), 21–37.

Ronell, A. (2005). *The test drive*. Chicago, IL: University of Illinois Press.

Scheel, S. (2013). Autonomy of migration despite its securitisation? Facing the terms and conditions of biometric rebordering. *Millennium: Journal of International Studies*, *41*(3), 575–600.

Schinkel, W. (2016). Making climates comparable: Comparison in paleoclimatology. *Social Studies of Science*, *46*(3), 374–395.

Schinkel, W. (2017). *Imagined societies. A critique of immigrant integration in Western Europe*. Cambridge: Cambridge University Press.

Schinkel, W., & van Houdt, F. (2010). The double helix of cultural assimilationism and neoliberalism: Citizenship in contemporary governmentality. *British Journal of Sociology*, *61*(4), 696–715.

Strathern, M. (1996). Cutting the network. *Journal of the Royal Anthropological Institute*, *2*(3), 517–535.

Tazzioli, M. (2017). Containment through mobility: Migrants' spatial disobediences and the reshaping of control through the hotspot system. *Journal of Ethnic and Migration Studies*, *44*(16), 2764–2779.

van Houdt, F. (2014). *Governing citizens: The government of citizenship, crime and migration in the Netherlands* (Dissertation). Erasmus University Rotterdam.

van Houdt, F., Suvarierol, S., & Schinkel, W. (2011). Neoliberal communitarian citizenship. Current trends towards 'earned citizenship' in France, the Netherlands and the United Kingdom. *International Sociology*, *26*(3), 408–432.

van Oers, R. (2013). *Deserving citizenship. Citizenship tests in Germany, the Netherlands and the United Kingdom*. The Hague: Brill.

van Reekum, R. (2018). Patrols, records and pictures: Demonstrations of Europe in the midst of migration's crisis. *Environment and Planning D: Society and Space*, *37*(4), 625–643.

van Reekum, R., & Schinkel, W. (2017). Drawing lines, enacting migration: Visual prostheses of bordering Europe. *Public Culture*, *29*(1), 27–51.

How to cite this article: Schinkel W. State work and the testing concours of citizenship. *Br J Sociol*. 2020;71:556–571. https://doi.org/10.1111/1468-4446.12743

Received: 12 March 2019 | Revised: 31 May 2019 | Accepted: 16 October 2019

DOI: 10.1111/1468-4446.12726

SPECIAL ISSUE

WILEY

Underground testing: Name-altering practices as probes in electronic music

Giovanni Formilan[1] | David Stark[2,3]

[1]University of Edinbrugh Business Shool, Edinburgh, UK

[2]University of Warwick, Coventry, UK

[3]Columbia University, New York, NY, USA

Correspondence
David Stark, Centre for Interdisciplinary Methodologies, University of Warwick, Social Science Building, Coventry CV4 7AL, UK.
Email: d.stark@warwick.ac.uk

Funding information
H2020 European Research Council, Grant/Award Number: 695256

Abstract

Name-altering practices are common in many creative fields—pen names in literature, stage names in the performing arts, and aliases in music. More than just reflecting artistic habits or responding to the need for distinctive brands, these practices can also serve as test devices to probe, validate, and guide the artists' active participation in a cultural movement. At the same time, they constitute a powerful probe to negotiate the boundaries of a subculture, especially when its features are threatened by appropriation from the mass-oriented culture. Drawing evidence from electronic music, a field where name-altering practices proliferate, we outline dynamics of pseudonymity, polyonymy, and anonymity that surround the use of aliases. We argue that name-altering practices are both a tool that artists use to probe the creative environment and a device to recursively put one's creative participation to the test. In the context of creative subcultures, name-altering practices constitute a subtle but effective form of underground testing.

KEYWORDS

alias, anonymity, creative identity, electronic music, pseudonymity, subculture, testing

1 | INTRODUCTION

Experiments are vital for science; tests are important for engineering and technology. Trials matter for the legal system but also for race cars, track stars, and new pharmaceuticals. Athletes try out for Olympic teams but so do aspirants for summer theater. We try on clothes, and sometimes roles while online dating. This paper examines tests and probes in electronic music.

Most generally, artistic production can be considered as a matter of *Tttrial and Eror* (to borrow from the title of a 2002 EP by the German artist Apparat; *Discogs*, 2018a). Even more to the point, musicians test themselves and their audiences, and probe artistic environments. On the one hand, artists' music and leadership role are put to the test. Not settled once and for all, their creative production and the ability to perform onstage are tested repeatedly in ongoing trials. On the other hand, as key participants in a subcultural community (Muggleton, 2000; Muggleton & Weinzierl, 2003; Redhead, 1997; Thornton, 1996), artists also put the boundaries of the electronic music community to the test—sometimes in open opposition to corporate-driven pop music and commercial EDM (electronic dance music).

Our subjects are artists who aspire to, or already make, an impact on electronic music. The key probe that we analyze is a set of name-altering practices, most notably the use of aliases by these artists. Electronic music artists are not unique in creating aliases (McCartney, 2017; Milohnić, 2017; Phillips & Kim, 2009), but they are at an extreme in their use: of the more than 30 artists we interviewed, none had fewer than two aliases, many had three or more, and some techno artists have more than a dozen aliases.

Of course, the names under which artists record, release, and perform their music are only one instance among the several practices artists use to discover their role, reinforce their participation in the artistic community, and assess the evolving state of the subcultural scene (Hebdige, 1979; Muggleton, 2000; Redhead, 1997; Thornton, 1996). Practices pertaining to the domain of technology (e.g., forms of production, types of recording equipment, technologies for live performance; Hennion, 1997, 2009; Prior, 2008), of aesthetics (e.g., style and sound choices, modes of communication; Becker, 1984; Born, 2010), and of network relations (e.g., affiliations with recording companies, preference for small or large venues, industry engagement at different levels; Caves, 2000; Prior, 2018) play a decisive role as well. As we will see, however, name-altering practices provide the artists with a flexible probe they can use to test electronic music at multiple levels.

In electronic music, we outline three primary name-altering practices adopted by artists: *pseudonymity*, that is, the adoption of an alias, more or less divergent from the artist's given name; *polyonymy*, that is, the use of multiple aliases, over time or simultaneously, to release and perform music; and *anonymity*, which can take the form of resolute concealment behind an alias, or complete dismissal of names. It is worth noting that these name-altering practices are not mutually exclusive nor strictly sequential. For instance, anonymity can be pursued by combining pseudonymity with constant avoidance of public appearance. Or polyonymy can include not only music published under multiple pseudonyms, but also records released under an artist's given name. At the same time, the adoption of an alias (pseudonymity) does not always precede anonymity, and the latter can appear at any moment in one's career.

Name-altering practices put the artist to the test. Elaborated further with examples below, we observe that the alias is a device that allows the exploration of one's role and potentiality, while keeping responsibility at a distance. Being connected to the artist but, at the same time, distinguished from her, the alias has a double character which fuels the testing dynamics. On the one hand, the alias has enough distance from the artist, making possible its observation from the outside, and its modification, updating, or even denial. On the other hand, the alias is an expressive and expressed part of the artist, and therefore the test outcome of its use can be projected back onto the artist.

In terms of self-exploration, the alias enables the artist to test aspects of artistic identity. In this frame, when we say that aliases can be used to put artistic identities to the test, it is less a test of character than of using a character as a test. As a character, an alias allows the artist to try out aspects of her creativity. In many such trials the relationship of artist to alias occurs within the context of a third—the audience. But it is important to understand that some tests do not simply present the alias or the music in a situation in which audiences (including critics) are meant to be the final judges. Instead, in some cases, rather than testing the artist's creative role, it is the audience that is put to the test. In this sense, name-altering practices do not only test the artist directly, but also probe the artistic community and its boundaries.

Like in most subcultures (Hebdige, 1979), electronic music developed its cultural boundaries around rituals and tacit codes that set the dividing line between itself and mass consumer culture (Kühn, 2015; Lange & Buerkner, 2012; Schüßler & Sydow, 2013; St John, 2006; Till, 2006). Rave parties during the late 1980s and early 1990s were quintessentially rituals that challenged the centrality of masculinity, authenticity, and meaningfulness of pop culture (from classical music to rock, Hennion, 1997; from corporations to private properties) through the illegal occupation of public and private areas and the creation of temporary spaces centered on the body and the dancing experience (Garcia, 2015; Gilbert & Pearson, 1999). Even today, attendees to now-legal EDM festivals try to preserve the so-called PLUR ideology (peace, love, unity, respect) which characterized the early raving culture (Chen, 2014).

In recent decades, electronic music has experienced an impressive growth, becoming a mass-consumption commodity that generates huge profits for its promoters (O'Malley Greenburg, 2013; Rys, 2016). In this situation, when the original traits of electronic music are jeopardized by threats of assimilation into the dominant culture (Marcuse, 1964), rituals and codes become crucial in contention about the grounding values of the community. These dynamics are played out in the language of "underground." Far from having a clear definition, the notion of underground retains fluid and esoteric elements that make what is "underground" inaccessible to those not "in the know" (Muggleton, 2000; Muggleton & Weinzierl, 2003; Redhead, 1997; Thornton, 1996). References to underground abound in the current discourse on electronic music, yet the concept remains contested. The boundaries of what is underground and, in opposition, what is commercial (or "mainstream," in the scene's language) are not only continuously shifting over time, but they are also differently interpreted by the field's participants. Perhaps the most emblematic example is that of Tiësto, an international superstar DJ that performs pop-oriented electronic music in very commercial venues, but who insists he remained an underground artist.

> *I feel like I am a commercial underground DJ because I never had that commercial appeal, have like top-10 hits. All my music is known in subcultures, it's known in colleges, or on dance festivals. Everyone knows Tiësto, but I had never like 20 hits like Beyonce had, for example. (Tiësto, interviewed on ABC-Nightline, 2018)*

The notion of underground is largely used in the scene to oppose the commercial, mass-oriented, corporate-driven music business, where the artist is consecrated as a popular icon, and her music often comes at a secondary position. In place of offering a sharp definition, we argue that the notion of underground is of analytic interest precisely because its meanings and the boundaries to which it refers are highly contested by the participants themselves. It is in such a context that name-altering practices can be seen as a form of "underground testing."

As we shall see, in opposing the commercial consumer culture that dilutes the significance of the "adversarial qualities" (Moore, 2005, p. 229) of electronic music, artists can use diverse name-altering practices to test the state-of-the-art of the subcultural scene, and initiate its renegotiation. Whereas pop culture rewards clear-cut identities that can be branded and distributed more easily to a large audience, electronic music artists embrace polyonymy to oppose easy categorization and commercialization (Hofer, 2006). Whereas pop culture values personality and faces, electronic music artists disappear into anonymity to restate the centrality of the sound and the dance experience (Hennion, 2009). In both situations, the integrity of the audience's membership to the electronic music culture is put to the test. How far can an artist push genre experimentation before the audience gets upset? Do people value the music or do they look to the artist's stardom? Through name-altering practices, the artist can probe the boundaries of the community, and eventually contribute to its reshaping.

In his essay on the sociology of testing, Pinch (1993) highlights *performance* and *negotiation* as key aspects of testing. According to Pinch, these characteristics remain valid across various typologies of testing—and they apply even more deeply to name-altering practices in electronic music.

First, “Many tests are *performances* that can be witnessed by others” (Pinch, 1993, p. 26; italics in the original). Tests usually happen in public, where witnesses can validate the results of the test, the protocols used, and the overall quality. Tests do not happen in the realm of the private, but in the public sphere.

Name-altering practices are performances in many senses. They perform an indexical function in respect to the artist-audience relation, enabling—or impeding—reference, and thereby constituting the basis for communication. In addition, name-altering practices perform a semiotic function, in that a name can be used to convey a message or a specific position (for instance, the name of the Detroit-based DJ group Underground Resistance has a clear political trait). Ultimately, the visible result of name-altering practices, the alias, is the one that actually performs—on stage and on records. On stage, the named character brings its own story, its own style, and its own creative approach. On records, it defines the sound, the music; it performs a classificatory function.

Second, “the outcome of tests can be treated as a matter of politics and social negotiation” (Pinch, 1993, p. 33). The test, witnessed by others, is often a site of social negotiation. Both the technical specifics of the test and its results are not universal truths, but their reality is instead negotiated at the encountering among a variety of stakes and needs put forward by different groups.

Name-altering practices are also sites of negotiation, in a form that partially exceeds Pinch's suggestion. On the one hand, the community-level result of a name-altering practice (for instance, the audience's reaction to a new alias) is negotiated in many ways—in clubs, by critics on specialized magazines, in promoter-organizer relations. In some cases, when the anonymity of an artist needs to be preserved absolutely, negotiation also happens at a legal level, with trademarked pseudonyms, manufactured documents, and anonymous booking (Pite, 2015). On the other hand, when the boundaries of the electronic music culture need to be re-established, the process of social negotiation between “underground” and “mainstream” proceeds itself through testing. It is not uncommon, in electronic music forums and magazines, to find discussions speculating about the reasons that drove an artist to adopt a new alias. We encountered several situations where forum participants motivated the adoption of a new alias as an artist's way to communicate his or her attachment to electronic music's original roots—especially when the artist in question was currently connected to the mass consumer circuit. In some cases, the discussion about an artist's name change becomes even more salient than the music released under that name.

From the perspective of the artist, a new alias is then a visible action that sets up a moment of renegotiation of what it means to be an electronic music artist. The artist's need for such a renegotiation precedes the name-altering practice. Sent out to the scene, the alias demands feedback and prompts a debate that, ultimately, re-establishes or redefines their and their audience's membership to the subcultural movement. While testing is always a site of negotiation, sometimes negotiation itself can become a site of testing—a moment where probes are sent out to collect information to proceed with the negotiation. Name-altering practices, as we will see, function as powerful probes and tests whenever the grounding cultural features of electronic music are put at risk.

We make no claim that our case is typical of aliases and testing in creative fields. But by displaying naming practices in such acute form, our case provides a distinctive laboratory to study dynamics that are important for the production of creativity as departure from established schemata: creative projects as the result of processes of testing, and creative projects as themselves tests to probe the audience, the community, and the subcultural values. In this dynamic, name-altering practices therefore represent the visible surface of a more subtle, impalpable set of ongoing testing.

We build our argument drawing from more than 30 in-depth conversations we had with electronic music artists, primarily in Berlin and New York.[1] Besides conducting formal interviews, we also attended a number of festivals, club events, studio sessions, and rehearsals to extend our comprehension and knowledge of the electronic music culture. We augment our first-hand insights with interviews and commentaries that appeared on dedicated magazines on electronic music (among others, *Resident Advisors*, *DJ TechTool*, *DJmag*, *xlr8r*, *Pitchfork*, *Discogs*, *Rolling Stone*). And we benefitted also from historical and sociological accounts of electronic music, some of them largely considered the most authoritative sources of knowledge on the field (notably, Gilbert & Pearson, 1999; Hesmondhalgh, 1998; Nelson, 2015; Pinch & Trocco, 2009; Reynolds, 1998; Thornton, 1996).

To study the functioning of name-altering practices in electronic music, we pose three broad questions about the use of aliases. First, why do artists take on a different name (the alias) when they enter the electronic music scene? Second, how do artists cope with the dilemma of stylistic experimentation in a market that rewards consistent identities? Third, how do artists navigate the tension between the subcultural logic that values music and the economic logic that rewards visible personalities?

To address these questions, we structure the body of the paper in three empirical sections. In each section, we discuss the dynamics of aliases together with the evolution of electronic music scene. Purposefully, each section's title is posed as juxtaposition, in which a salient characteristic of the culture is addressed in parallel with a salient question confronting the artist at different moments in his or her career.

In the first section, we present the origins of electronic music as a subcultural phenomenon, and outline the nature of alias as a means for preserving anonymity while acquiring visibility. We claim that the alias is the individual counterpart of the field's collective challenge to the dominant culture, a tool to test the extent to which a previously unknown entity can raise its creative voice.

In the second part, we grapple with the multiplication of subgenres in electronic music and the corresponding multiplication of artists' aliases. We highlight how audience's expectations influence the adoption of new aliases, which not only allows artists to test their expressive freedom, but also to test aspects of electronic music culture. Such multiplication of aliases, however, can also become a source of tension, especially when popularity enters the picture. This puts the artist to a further test.

In the third section, we then discuss the commodification of electronic music in the recent years, and show how name-altering practices can serve as a viable way to respond to this change. We argue that artists that aim to return to electronic music as subcultural practice can step back into anonymity. In doing this, they push name-altering practices to the extreme (no-name alteration) in order to restate their "underground" attitude towards electronic music in opposition to the consumer culture. At the same time, they also ultimately test whether the reception of their music depends on the music itself or on the popularity already gained.

Although the structure of the paper may suggest a temporal evolution, the phases discussed in the three sections should not be thought as temporarily consecutive moments in one's career, but rather as critical moments (Boltanski & Thévenot, 1999; Guggenheim & Potthast, 2011; Hutter & Stark, 2015) that can variously happen throughout one's creative journey.

2 | ELECTRONIC MUSIC RISES | CAN I BE AN ARTIST?

2.1 | The subcultural milieu

Electronic music was born as a new form of approaching classical music through technology (Cross, 1968; Nelson, 2015). It did not, however, remain in the music academies but entered pop culture in 1974 when the Düsseldorf-based band Kraftwerk released *Autobahn*, the LP that introduced a general audience to the sound of synthesizers and drum machines. Recalling the noise of factories and machineries of the industrial era, it was not surprising that electronic music found a fertile environment in the Rust Belt of the United States. Techno music developed in Detroit, house music in Chicago (Reynolds, 1998). From there, it re-crossed the Atlantic, to the UK where it shaped the Second Summer of Love during the Thatcher era (Gilbert & Pearson, 1999); and to Berlin where it captured the dissatisfaction of the working class and the imagination of ethnic and sexual minorities to turn abandoned factories and warehouses into places for collective catharsis in the period of economic collapse in the 1990s following the fall of the Berlin Wall (Bader & Scharenberg, 2010).

For many years, electronic music has been uniquely a subterranean scene. In Detroit and Chicago during the early 1980s, techno and house dancing nights were times and places devoted to the abandonment of social rules, roles, and expectations (Gilbert & Pearson, 1999; Reynolds, 1998). People were allowed to get rid of all the constraints imposed on them by social norms, and in doing this were supported by the industrial aesthetics of the location and the perceptual distortions induced by drugs (Sanders, 2005). Black people in Chicago and working-class kids in Detroit could let the burden of social position outside the dancing space, and get lost in the sound texture. In the UK, in London but also—and especially—in remote locations in the countryside (Gilbert & Pearson, 1999), the rave scene of the late 1980s and early 1990s was able to bring together the aesthetics of punk with the suburban minorities and relax the cultural clash existing among distinct—and in some cases opposite—factions (Hesmondhalgh, 1997). Illegally gathered in open spaces and obscure rooms, people coming from social strata as diverse as soccer hooligans and gender minorities danced together at the same pulse. In Berlin, electronic music became the sound of the critical reunification. People from the East and the West merged in obscure warehouses and abandoned bunkers filled with fog-machine smoke and backlight stroboscope flashes.

As members of a movement living on the fringe of legality and dominant moralism, techno ravers needed ways to safeguard their privacy and legal identity. The Berliner Berghain/Panorama Bar club, one of the most iconic venues for techno music, has a no-photo policy still today. Talking with Will Coldwell from *The Guardian*, the ethnomusicologist Luis-Manuel Garcia discussed privacy protection in terms of freedom of exploration.

> *Clubbers don't need to worry about there being a record of their time there, allowing you to explore your identity or adopt a different one altogether without fear of anyone taking you up on it on the outside. (Luis-Manuel Garcia, in Coldwell, 2016)*

The taboo associated with the 1980s' rave parties, and with the contemporary club culture, surely played a role in spreading practices of pseudonymity among the artists—"20 years ago, the term 'rave' was a drug-fuelled warehouse with sweat dripping from the ceilings" (Peros, 2014). But name-altering practices were also related to political engagement, an attempt to refrain from the commercial branding of artists' faces and to categorically refuse the corporate world—like in the case of Detroit-based DJ groups Scan 7 and Underground Resistance, whose members cover their faces with black kerchiefs, caps, and balaclavas (Pite, 2015).

2.2 | Pseudonymity: Testing a role

However, more than just a way to comply with the aesthetics of a subculture and to engage politically, the peculiar anonymity offered by aliases also enabled aspiring artists to test the feasibility—and eventual limitations—of their active participation to the field.

An alias preserved the privacy of the DJ who had a 4–6 a.m. set on Sunday and then went to work as a bank teller on Monday. For the DJ who was a day-time janitor, on the other side, an alias provided a way to create a character that could escape an otherwise ordinary lifestyle. How could a person with a menial job, with an everyday life, be deeply grounded in the logic of the dominant society, surge to the role of charismatic leader that sets the beat for a large night-long dancing crowd?

> *I look for a DJ who doesn't just play what is popular, but takes risks with his music selection, takes the crowd on a musical journey and, most importantly, can read a room. Knowing how to pick up a room that's a little bit down or being able to bring the vibe down in such a way you don't lose the crowd and then being able to take them back on a musical journey is a key skill for me. (Producer and DJ Erik Morillo, in Jenkins, 2017)*

"Read a room." "Don't lose the crowd." "Take them back." As Maxime put it during a conversation we had in Berlin, the DJ needs a "strong character" to take on the audience and bring it to a dancing experience. A "larger than life" character that, most times, requires a detachment from the artist's personal story. In an interview, the German producer Sebastian Kramer stressed this point.

> *If you know where they're born, which school they went to, what they like to eat for breakfast, and then you listen to the same thing … It doesn't have the same power anymore. (Sebastian Kramer, in Pite, 2015)*

Ideally, some names are naturally more suited to larger-than-life characters. Looking at the list of artists that received a nomination as Best Live Act between 2008 and 2016 from the authoritative electronic music community Resident Advisor we find the following aliases:

- Plastikman, Robotman, El Guitaro, Prins Tomas, Dr. Kevorkian, Jack Da Ripper clearly point to a character that, already in its semantic, overcomes the limits of ordinary people;
- Motor City Drum Ensemble, The Panamax Project, Hyperdrive Inc., Black Jazz Consortium, The Underground Crew, and Desert Stormers leverage collective identities to convey authority;
- Acid Test, Floating Points, Creative Violence, Shellshock, Perpetuous Dreamer, and Deep State, Barricade, Wrong Copy, False, Superlova, Graphite evoke technical, emotional, material, or stylistic properties, making suggestions to the audience in the attempt at signaling a role that embodies a take on music, a creative attitude;
- DJ Nobu, DJ Stingray, DJ Antal, DJ Hell, DJ Limiter put the "DJ" tag before the name, setting the artistic role at a good distance from menial or white-collar jobs;
- Many artists adopt idiosyncratic terms as aliases: Barem, Loxodrome, Simitli, Ratcapa, Blawan, Boddika, Jabberjaw are pure sound that give up any semantic while preserving the indexical function.

Being a leader is far from elementary, though. Unsure of their ability to take on the responsibility of leading the dance floor, and live up to that expectation, the artist needs to confront the audience and explore the limits of their role. Developed internally, the alias is projected out in the electronic music scene, an exploratory probe on a discovery mission. Probing the boundaries of the scene, the alias captures feedback signals that inform its own value, music, style. The alias is then a proper test market, where "experimental launchings of new products are intended to expose problems that otherwise would be undetected until full-scale introductions are underway" (Silk & Urban, 1978, p. 171). By actively wandering in the scene, the pseudonymous alias functions as a device to determine the feasibility of the artist's participation in the scene, and the boundaries of that participation.

The probe nature of an alias, however, is not limited to novices. A role, a creative voice, is not found once and forever. Instead, since the goals of individuals develop and evolve over time, one's role as artist can be rediscovered.

In 2014 Aphex Twin, the most famous alias of UK producer Richard D. James, released his long-awaited sixth studio album, *Syro*. Before releasing it, however, James was unsure of whether people still wanted to hear his music. Just some time before, he had already released hours of unpublished material via the streaming platform SoundCloud using a generic username (most probably, user18081971 or user48736353001; *Discogs*, 2018b). As reported by *Rolling Stone* (Blistein, 2015), however, James ended up rediscovering his role as (still) contemporary producer when a later fundraising campaign to access an Aphex Twin's iconic but rare record was very successful.

> *That was really touching, and really sweet… And I'm getting a bit older. It's like, "Okay. People out there really, really want stuff off me, so I can't deny it. Let's put it out." (Richard D. James, in Blistein, 2015)*

By adopting an alias, artists test their connection to the music, their audience, their artistic idea. As time goes by, like in the case of Richard D. James, the tests can be used either to confirm a previous state (for instance, when the artist's role was accepted), or to probe the scene in search for a new type of connection to the music, the audience, the artistic ideas.

While the alias seems similar to the "sign-equipment which large numbers of performers can call their own for a short period of time" to perform a role on stage (Goffman, 1959, p. 14), it should, however, be thought of more precisely as an assemblage of provisional elements that are meant to test a role.

In fact, aliases are easy to adopt and abandon. Benjamin, a Berlin-based artist who performed under several aliases during the early years of his career, told us that his aliases made it possible for him to say "I can be every time someone else." Often, aliases have a provisional and aspirational nature. In search for an answer to the question "Can I be an artist?", the alias serves as a device to send test signals to both the artist and the scene, to probe the artist's position and the audience, to test creative ideas and the values of the subculture.

2.3 | Pseudonymity: Probing the environment

While pseudonymity is primarily a practice for self-discovery, it sometimes also serves as a probe to sound out the creative environment. Since its inception, electronic music challenged the way society was organized around the private and the demarcation of individual differences—private (and oppositional) in terms of economic consumption (poor/rich), job specialization (worker/manager), community membership (black/white, straight/gay), geographical provenance (East/West), gender (female/male).

Reflecting on the techno challenge to gender binarism, for instance, Gilbert and Pearson (1999) noticed that "techno's asexuality might be seen as a deliberate strategy, a pursuit of neuter *jouissance* which seeks not simply to regress to a moment before the regulating discourses of sexuality took hold of our beings, but to go beyond them into an imagined cyborg future, a place where the fluidity of cyberspace is the medium for non/identity and the robot exoskeleton is the site of a constructable, engineerable, alterable androgynous corporeality" (Gilbert & Pearson, 1999; italics in the original).

In setting up this challenge, electronic music substituted the individual with the collective, the body with the dancing crowd (St John, 2006). Resonating with this, artists did not simply adopt name-altering practices of pseudonymity to test themselves—in absolute terms and in respect to their cultural environment. Instead, the creation of an alias was also a way to confront the community members with the emptiness of dominant binary approach grounded on categorization and branding of names, faces, and bodies.

An extreme case—one that reminds of novelists George Sand and George Eliot—is the one of artist Tatiana Alvarez. The American artist took on the name of Matt Muset (performing on stage under the alias Musikillz) to cheat the gender stereotypes she faced at the beginning of her career when promoters were only concerned with her physical appearance and wanted her to dress up seductively. Retrospectively, she sees this move as a social experiment of reinventing herself.

> *I thought, "I need to be a guy, I need to look like a guy, I need to be the opposite of anything that's sexy". So I put on guy clothes and cut my nails. I didn't want to cut my hair, so I used a wig. I am the right size for a guy, other than having hips and boobs. So I taped down my boobs using a sports bra that was too tight: it has to hurt a bit because that's what affects your posture. The only way to really breathe is to keep it shallow. (de Bertodano, 2015)*

Alvarez not only created a character to face the world of commercialization, she also invented a female alter ego to represent her (him?) as an agent. The two characters sustained each other. "I did almost everything by email. Also, when you're dressed like a guy your body feels different. It hurts, which put me in a bitchy mood, which totally helped" (de Bertodano, 2015). After about one year, Alvarez decided the test was over and her point was made. She now performs as Tatiana Alvarez, also temporarily known in the past as Matt Muset—also known as Musikillz.

3 | THE SCENE DEVELOPS | CAN I DO SOMETHING ELSE?

3.1 | Proliferating subgenres

Grounded on machine-produced sounds, electronic music developed since the beginning as a genre-recombinant genre (Formilan & Boari, 2018). While Chicago-house took influences from Black culture-rooted funk, soul, and rhythm-and-blues, UK jungle received the legacy of fast-speed punk music and mixed it with Caribbean elements. The variety of subgenres, or styles, in electronic music is today impressive (McLeod, 2001). House music, for instance, has at least the following recognized subgenres: Acid house, Ambient house, Balearic beat, Chicago house, Bass House, Deep house, Future house, Tropical house, Diva house, Electro house, Big room house, Complextro, Fidget house, Dutch house, Jungle Terror, Moombahton, Moombahcore, French house, Funky house, Garage house, Ghetto house, Ghettotech, Hardbag, Hard house, Hard dance, Hard NRG, Hip house, Italo house, Jazz house, Kidandali, Kwaito, Latin house, Microhouse, Minimal house, New beat, Outsider house, Progressive house, Rara tech, Tech house, Tribal house, Trival, Witch house.[2]

Inevitably, the diffusion of technology and the Internet have exacerbated the multiplication of subgenres and sub-subgenres (Born, 2005). Technology is not the only driver of the emergence of new subgenres, of course, but nonetheless it supports the artists in the development and introduction of unexpected sonic properties that might eventually be codified into a new genre category (McLeod, 2001).

The variety of subgenres is also directly influenced by the uneven nature of creative production. Unconsciously or purposefully, artists might end up producing music that does not fit into the artistic voice they originally conceived. On the artist side, this demands the inauguration of a new project, perhaps under a new alias. During our conversation in Berlin, Maxime clarified on this aspect.

> *Let's say I'm using an alias. And I'll make a pile of tracks. But I look and say:* this *is definitely not* that. *This can't go with that alias. It's not necessarily that I chose a different identity and then made those tracks. More like open to a different influence. With a new alias you can get freedom of expression to get out of the style prison. (Maxime)*

As a response to sometimes unpredictable creative journeys, artists articulate their artistic voice around multiple aliases. From being pseudonymous during the origins of their career—either because of little visibility, or because of conscious concealment—artists often move to a "polyonymous" situation, where multiple aliases can appear sequentially or simultaneously in one's career. As the genre becomes more and more fragmented into multiple subgenres, so the artists can develop multiple names to participate in diverse contexts.

3.2 | Multiple subgenres, multiple aliases

Given the extensive use of aliases in the electronic music scene, one might wonder why in New York's "contemporary music scene" aliases do not play a significant role. New York's new music is grounded on continuous experimentation and recombination of genres, but its artists do not use aliases. In our view, the motivations that prompt New York's contemporary music artists to use only one name for different subgenres, and the motivations that lead electronic music artists to use different aliases for different subgenres are the same: both are grounded on audiences' expectations.

Audiences attending a contemporary music concert have an expectation for discontinuity: they have no problem hearing a chamber choir composition from an artist at one concert, and the same musician banging on cans at the next one. Actually, they would be surprised—and perhaps disappointed—if a John Zorn performance did not surprise. Their expectation is to be surprised, and they appreciate musicians who violate established genre categories.

Similarly, people attending a club whose poster announces a "minimal techno" event hold clear expectations regarding the type of music they are going to hear. "Progressive techno" or "minimal house" would frustrate them. They expect that the music performed will conform to the announced genre, subgenre, even sub-subgenre category. Similarities and differences depend, at least partially, on the different origins and characteristics of the two scenes.

New York's new music scene is promoted and sustained by expert audiences, composers, and performers (many of them conservatory-trained). In some ways, it represents an intellectual elite whose preference for experimental sounds also constitute a political claim. The audience of electronic music, by contrast, originally comprised social minorities and marginalized individuals who found in this electronic sound not only a place for everyday political engagement (Riley, Griffin, & Morey, 2010) but also a new source of in-group identification (Hesmondhalgh, 2008). So entangled with such subgroup identity politics, the electronic music scene in Berlin was more likely to go hand to hand with the fragmentation of sub-subgenres. For example, according to DJ and producer Ekrem with whom we talked in Berlin, the electronic music scene in Berlin is composed of mutually exclusive micro-scenes each gathering a specific social group (e.g., male and female straight, queer movement, male gay, female gay, transgender, LGBTQIs, ethnic minorities). Lacking a political component which in other genres is conveyed via lyrics, iconic characters, or educated audiences (such as in rock, pop music, contemporary music), electronic music is appropriated by different stakeholders who fill it with their idiosyncratic socio-political features.

Audiences' expectations, combined with the unpredictable dynamics of individual creativity, put electronic music artists in a tense situation. Artists tend to avoid being categorized ("I leave the category game to other people", as our informant Javon put it in New York), but they know that genre differentiation is the basis for the economic functioning of the scene—clubs announce genre-based events, targeting very specific audiences and promoting a well-defined aesthetics. At least momentarily, polyonymy can thus resolve this tension, enabling the artist to pursue new sound directions and maintain her presence in micro-scenes.

3.3 | Polyonymy: Testing one's freedom

The multiplicity of aliases can also be read in a different way. While it can be a tool to differentiate one's creative output and target different audiences, a new alias is also a way to test the extent to which the artist can experiment with different genres without violating the expectations of her audience. Developing a recognizable link to a specific music style, the alias can become a box that constrains the means of expression available to the artist. With a new alias, though, the artist makes room for stylistic exploration without irremediably eroding the sound of another alias—and the popularity it eventually gained. In order to avoid the audience's quick dismissal of new sounds, the artist might keep a new alias detached from their legal name. Talking about artists introducing new aliases, Sebastian Kramer pointed out: "That's why, at least at the beginning, they aren't saying, Hey! It's me again! I'm doing something different!" (Sebastian Kramer, in Pite, 2015).

In addition to testing one's freedom of expression, a new alias also presents the audience with a test, questioning not only the listeners, but the functioning of the electronic music industry as a whole. In 2017, the American producer Porter Robinson (previously known also as Ekowraith and Antigon Moore) released his latest work under the new alias Virtual Self. Accompanying the release, he also published a promotional message that questioned the current state of now-commercial electronic music, and put electronic music fans to an awareness test.

> *Finally—and this might be the goal that's dearest to me—[*the introduction of the new alias Virtual Self*] is to push electronic music in a different direction. As electronic music essentially converged with pop in 2016 (for the second time in the last 10 years, the other time being 2011), I think it's pushed a lot of artists away from risk-taking and passion projects. In the last two years, for most artists, all they really had to do was compromise their style by like 30% and add a safe, inoffensive tropical vocal to have a chance at having a hit—and I think for many, that temptation was too much.*

> *In my opinion, electronic music is at its best and its healthiest when new, exciting, unexpected things are happening. This is a genre that thrives on novelty. And to be totally clear, I don't think that Virtual Self, early 2000s trance, or digital abstract art are the solution or the future at all. But!! I DO think this style is something unexpected, and something I'm uniquely poised to make, because I love it. And that's the precedent I want to set, or at least the approach I want to remind other artists of. (Porter Robinson, quoted in Rafter, 2018)*

Creating a new alias, or changing it, is then a way to get rid of the constraints posed by single-alias situations. As an enabling device, the adoption of polyonymous configurations makes it possible to further test the boundaries of one's creative production. A tool for discovery, polyonymy is a way to try on different creative voices, and try out new genres and styles. At the same time, polyonymy is also a device to question the scene and prompt other artists to persevere with genre experimentation.

Additionally, multiple aliases guarantee the exploration of an artist's attitude towards music making. Brian, a New York-based four-alias artist, thinks about his different aliases as different personalities that, together, ensure him a certain distance from branding and constraints to creativity imposed by the logic of commerce.

> *It's my way of pushing back against branding … Because I struggle to identify myself as one thing and not another. I don't see the point in forcing myself, so I have a bunch of things going on and if people connect the dots, that's fine. If they don't that's fine too. (Pearl, 2017)*

Adopting a polyonymous configuration, artists leverage multiple aliases to effectively surf the contradictions of most creative industries (Caves, 2000). The creative logic allows for (and even supports) boundaryless exploration throughout a variety of subgenres. The economic logic requires specialization, recognizability, and authorship. The cultural logic values subcultural attitude and the aesthetics of anonymity. The creation and use of multiple aliases is then the individual response to these contradictions. Through multiple aliases, artists can simultaneously guarantee recognizable specialization, anonymous aesthetics, and creative exploration. Polyonymy ultimately puts to the test the freedom promoted by electronic music: "Can I do something else?"

3.4 | Polyonymy: Alias as prison

Unfortunately, no pro comes without a contra. Artists with multiple aliases that experience critical and commercial acclaim under multiple names can face a challenging condition. When equally successful, multiple aliases can indeed become multiple prisons. As Maxime put it, the alias "is the escape from one prison. But in that a step to be lost again in another prison." This aspect can put the artist in a demanding situation. Our informant Benjamin expressed the intimate difficulties imposed by separated aliases. His daily routine is punctuated by moments where he has to die and wake up Faxxe (one of his aliases). "I just want to do my thing," he said, stressing the difficulties of being trapped into more than one successful character.

Facing success under multiple aliases, the artist loses the freedom of experimentation that was gained through the introduction of a new alias. Further experimentation would require additional aliases but, as Benjamin (aka Ben-J, aka Faxxe) noticed, multiple aliases may require intense efforts to be managed. In fact, from being indexical tags that point to an artist's music, the aliases become a quick tool to classify subgenres, imprisoning the artist into categorical boxes.

> *Every artist has their own identity and sound, which is always evolving. If music is always evolving, how can an artists' sound be classified? An artist shouldn't fit into a genre, for they should be their own genre. This means that instead of classifying an artist by current trends, we should classify them by their*

> *individual sounds and identity. It's that indescribable feeling you get after hearing a track for the first time which causes you to say "this sounds like a Derek May tune" instead of "this sounds like a post modern, Detroit influenced, down beat, up tempo, old school remix." (Peros, 2014)*

As an alias becomes the substitute for a subgenre, the alias itself turns into a prison for the artist. When pseudonymity loses its aura of mystery, and polyonymy further complicates the picture, the practice of naming can then undergo a dramatic path. Popularity turns into a burden, and poses an ultimate challenge to the artist—a final test that, in some cases, becomes a test for the whole electronic music culture: "Is it still a subcultural movement?"

4 | THE GENRE GOES COMMERCIAL | IS IT ABOUT ME, OR ABOUT MY POPULARITY?

4.1 | Still a subculture?

Over the last decade, electronic music has become very popular. Elements of both its aesthetic and sonic properties have been absorbed and popularized by pop culture and mass-oriented business. The number of summer music festivals that progressively included electronic music artists in their line-ups is uncountable, and more and more festivals have been born with a special focus on electronic music. In this circuit, DJs are now the new rock stars (O'Malley Greenburg, 2012). Some of them earn a million dollars a year, perform worldwide at venues that require the audience to pay three-digit covers, collaborate with pop artists on Billboard-awarded tracks, and fly on private jets from one dance floor to another, even during the same night (Millington, 2017).

More local and music-focused scenes are not impermeable to this growing trend. Yet, some artists are not all comfortable with this situation.

> *I was around with people from the first underground last year [*according to Benjamin, there are three levels of the underground scene, plus the mainstream scene*]. [...] It was horrible, really horrible. Staying awake, waiting for ages, not seeing the most important person of my life. So, this is the reason why I know—and I can say honestly—I never want to reach that area of success. Because it's nothing I want. (Benjamin)*

As artists become more and more popular, they are increasingly confronted with mass-audience expectations of visibility, not only on stage but also, inevitably, online—a dimension many artists feel is not contributing to making good music.

> *I spent a lot of time in my early years in the music industry dicking about with my website and stuff like that, because I felt it was important, whereas I would have been better served by just focusing on making music… (unknown artist, in Taylor, 2014)*

> [Before the Internet*] it's not like we were listening to our new records in our flat and saying, uh, can you imagine who this guy is? We didn't ask questions like that! I didn't know who Octave One was. It was Octave One! Then later on came the name Burden Brothers, and even then it's just a name. (Sebastian Kramer, in Pite, 2015)*

The demand for visibility, combined with the constraints imposed by a popular name, poses a very critical question to the artist: is the music acclaimed because of its own worth, or because of its creator's popularity? From the perspective of artistic production, this suspicion puts the artist to the test. While pseudonymity was a practice to discover herself, and polyonymy a way to increase the channels to express creativity, the artist has ended up imprisoned within acclaimed commodities—her branded aliases. Electronic music, from being the result of an effort to leverage the sonic properties of technology, has now turned into a product for mass consumption—a condition many artists might not aspire to (Scott, 1990).

In addition to questioning the reception of one's creative production, the popularization of electronic music also erodes the aesthetics and values of a culture that developed in opposition to pop culture and corporate-driven consumption, where the artist's image is not second to her music. In a subculture hinging on anti-ego ideology and the centrality of sound, the artist then faces an individual-community dilemma: *How can I fulfil my desire for recognition, authorship and distinctiveness, and at the same time abandon my ego to remain loyal to the "underground" culture?*

Together, these questions represent a difficult challenge. How can an artist have a name, and yet have her music speak in her name? How can she be recognizable, distinctive, and invisible at the same time? How can she avoid to sell out, and still keep on selling her music?

Put to the test by commercial requirements and demanding audiences, the artist can denounce the situation—and address the challenge—by putting her relationship to the music and to the audience itself to the test.

4.2 | Anonymity: Testing the relationship to music

In order to remove ego but retain authorial connection to the music, artists can use anonymity as a mask (Pizzorno, 2010; Sassatelli, 2019)—a camouflage that either hides the physical traits of the artist, or that conceals her whole identity. This move goes back to the cultural origins of electronic music, when artists remained indistinct in the flashing darkness of abandoned warehouses, or almost invisible during rave parties in the middle of UK's nowhere.

On October 19, 2017, the American artist DVS-1—real name Zak Khutoretsky—presented his Wall of Sound show at the Warehouse Elementenstraat in Amsterdam. The show featured a giant sound system that occupied the whole stage, while DJs played their records from the opposite, dark end of the room. The sound was the protagonist.

> *Mad Mike Banks said if you put your face in front of the music, you're putting your ego in front of it. We don't want anyone to be paying attention to our ego, we want everyone to be paying attention to the music and the experience. "We shouldn't be on stage," he continues. "We're not a band. We're a vessel for music. Get us out of the way, get rid of all that extra clutter and fill it with speakers!" (DVS-1, interviewed in McCallum, 2018)*

In order to resolve the tension between the centrality of a name and the centrality of a sound, artists keep their names (or their aliases), but physically disappear from the place where electronic music is consumed—the club. This is a case of masking where the corporeality of the artist is completely anonymized—the artist could have any face, any body, and any temperament. Appearance is not a locus of attention anymore. While authorship remains preserved, it is the sound that now has to speak in the artist's name.

Masking anonymity can also take a more radical form. Instead of anonymizing her physical presence, the artist can hide completely behind an alias. In this case, anonymity is not limited to faces and bodies, but extends to the whole identity of the person who goes by the alias. The case of Traumprinz is illustrative in this respect. Officially known also as Prince of Denmark, Dr. Sun, DJ Metatron, Prime Minister of Doom, and DJ Healer, Traumprinz is a German producer; yet the artist who bears all these names remains a mystery for the music scene (*Discogs*, 2019). A similar case, recalled by the magazine *DJbroadcast*, is that of Burial, a UK producer whose identity has been a matter of speculation for long time.

> *Back in 2008, Untrue—the second album from ambient dubstep producer, Burial—was nominated for the prestigious Mercury Music Prize. The press ran into a slight problem though: nobody knew anything about*

> *him. The Sun's then showbiz editor, Gordon Smart, began a campaign to "out" him, claiming the mystery "threatened one of the biggest nights of the showbiz calendar". (Negligible, perhaps, compared with the threat to his column inches.) When Burial became the bookies' favourite to win, what else were the tabloid purveyors of anti-news to write about? The music? (Pite, 2015)*

The effects of commercial popularization, however, are not always forestalled by a mask. In some cases, artists become famous for being anonymous, acclaimed by commercial audiences and booked by popular venues for their masks. While in the case of Daft Punk, Marshmello, or Deadmau5 this outcome was reasonably searched for purposefully, in other cases the choice to remain anonymous becomes the object of commercialization. The mask, originally an expression of loyalty to electronic music as a subculture that puts the music before the person, remains forcedly attached to the artist.

> *Kramer admits it was incredibly frustrating having his persona dismissed as a marketing gimmick, but acknowledges that it became one regardless of his intentions … The mask remains an integral part of the performance, but it's no longer sacred. It's "still there but I couldn't say I'm the same person who invented it. Time changed … and changed me." (Pite, 2015)*

4.3 | Anonymity: Testing the relationship to audience

Instead of anonymizing the person's corporeality, and thereby testing the artist's relationship to the music in the first place, anonymity can target the alias itself. Music is released with no reference to authorship. Anonymity, in this most radical form, reduces the question about the artist-audience relationship to its core: *Can I have an audience without a name?*

As Aydin explained to us during a conversation on his experience in running record stores in Berlin, anonymous records always have a section in the store's boxes.

> *The Berlin-based act* [ItaloJohnson] *don't wear masks, for starters, and don't have any social media to connect with their fans. Maintaining an air of mystery at all times, they let their finely-tuned tracks speak for themselves… The only way to identify their records is by squinting at a little handstamp in colored ink, or the catalog number in the record's runout. The rest is meant to be sorted out on the dancefloor. (Weiss, 2017)*

The technology of music publishing comes to the aid in the form of the *white label*. A white label is a vinyl record with a white label glued on top of it—and sometimes, like in the case of ItaloJohnson's records, a graphic stamp. In most cases, no one can know who its creator is by simply looking at the label. Its whiteness blinds authorship, allowing the creature to be visible by making its creator invisible. In some cases, however, small stamps retain a feeble trace of authorship, one that only authorized personnel can interpret—"A stamp is a stamp and a stamp is cool. And for that, I would keep stamping every fucking record" (DJ act ItaloJohnson, in Weiss, 2017).

Anonymous records have a long tradition in electronic music. By removing any reference to authorship, anonymity also constitutes a practice with high testing properties. It tests whether the music is worthy in itself; it questions whether the artist can have an audience without having a name—and if he is comfortable with that; it probes if the participants in the electronic music scene honestly share its subcultural, "underground" values.

In this latter respect, complete anonymity also opens up room for the renegotiation of the boundaries between the artist and the field (Prior, 2008). It brings music back to the center of the experience, shooing away the commercialization of names and faces, and discouraging the participation of inattentive audiences in the subculture. During our conversation in Berlin, Javier (aka Maxwell) compared the average fan base to assayers of junk food.

> *When you enter the commercial side, the audience grows so you feel what the audience is expecting and actually it becomes part of what you do. But you always try to ignore it or frame it in a way that allows you to be completely open. [...] It is very hard to ignore it completely the fact that so many people are there. But I think that it comes down on what type of fans you have. If you have an average fan base that they will expect something similar to the previous record. But it is like kind of junk food. (Javier)*

Anonymous records are thus a way to probe the environment and reframe the audience "in a way that allows you to be completely open." Through anonymity, the original experience of electronic music can be restored, and the boundaries of the subcultural scene renegotiated between the artist and the audience. In this process, the unnamed alias allows the artist to discover again the extent to which his or her active participation in the field is still worthy, and if the state-of-the-art cultural scene deserves the effort. Embracing anonymity, the artist can establish a mediated negotiation (Born, 2011) with the subcultural milieu. Sent out as a probe, the anonymous artist tests the various interests at stake in the subculture, and probes unexplored directions for artistic intervention.

Anonymity, variously embedded also in pseudonymous and polyonymous practices, is therefore a powerful tool to establish and sustain the adversarial qualities of electronic music culture. Invisible from the surface, anonymity is then the ultimate underground testing.

5 | CONCLUSIONS

Tests come in different forms. Nuclear tests, pharmaceutical and medical tests, pregnancy tests, stress tests, and psychological tests are all situations where an assemblage of technical and social elements gather together in order to confirm expectations, assay previsions, or discover the expected and the unexpected. Tests are sociological in many forms, in that they either directly involve or exert effects on the society—the social structures, the forms of interaction, the construction of shared meaning. As Pinch (1993) outlined, tests are sites of negotiation where multiple stakeholders, with different goals and motivations, intervene to assess the validity of the test.

In this contribution, we explored name-altering practices (the creation, adoption, and dismissal of aliases) as tests and probes that electronic music artists use to put themselves to the test, and to test the cultural scene they participate in. Articulating name-altering practices in three distinct but interconnected configurations—pseudonymity, polyonymy, and anonymity—we outlined the several moments when aliases serve as tests and probes in shaping one's role and one's audience in a cultural scene.

A (quasi-)leadership role in a subcultural movement is not taken once and for all. The possibilities, limits, and responsibilities of a role have to be repeatedly tested in order to define and refine its boundaries. Additionally, the music culture itself needs to be tested, especially whenever its distinctive features are threatened by pop culture's appropriation.

Underground cultures need forms of underground testing. And, in this respect, name-altering practices in the form of pseudonymity, polyonymy, and anonymity represent subtle configurations of testing the sizable dimension of a subterranean probe.

ACKNOWLEDGMENTS

For their comments, criticisms, and suggestions, we are grateful to Jonathan Bach, Paul DiMaggio, Elena Esposito, Michael Guggenheim, Shamus Kahn, Celia Lury, Josh Whitford, and the participants at the Put to the Test Workshop (Warwick in London, December 2018). Our thanks also to the participants at the CODES Seminar (Center on Organizational Innovation, Columbia University) where we first presented this work.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

NOTES

[1] We assigned pseudonyms to all the given names and the aliases of the artists quoted here.

[2] This list has been retrieved from Wikipedia on November 2, 2018. Its presence on the online encyclopedia is a sign of how these subgenres are well established in the scene.

REFERENCES

ABC-Nightline. (2018, May 31). Inside the life of international superstar DJ Tiesto. 1:48–2:10.

Bader, I., & Scharenberg, A. (2010). The sound of Berlin: Subculture and the global music industry. *International Journal of Urban and Regional Research*, *34*(1), 76–91.

Becker, H. S. (1984). *Art worlds*. Berkeley, CA: University of California Press.

Blistein, J. (2015). Aphex twin alter ego AFX preps new EP, releases new song. *Rolling Stone*. Retrieved from https://www.rollingstone.com/music/music-news/aphex-twin-alter-ego-afx-preps-new-ep-releases-new-song-52019/

Boltanski, L., & Thévenot, L. (1999). The sociology of critical capacity. *European Journal of Social Theory*, *2*(3), 359–377.

Born, G. (2005). On musical mediation: Ontology, technology and creativity. *Twentieth-Century Music*, *2*, 7–36.

Born, G. (2010). The social and the aesthetic: For a post-Bourdieuian theory of cultural production. *Cultural Sociology*, *4*(2), 171–208.

Born, G. (2011). Music and the materialization of identities. *Journal of Material Culture*, *16*(4), 376–388.

Caves, R. E. (2000). *Creative industries: Contracts between art and commerce*. Cambridge, MA: Harvard University Press.

Chen, K. (2014). A glimpse into the taboo of massive rave culture—A photographic essay. *Kellychenrws.Wordpress.Com*. Retrieved from https://kellychenrws.wordpress.com/2014/11/21/a-glimpse-into-the-taboo-of-massive-rave-culture-a-photographic-essay/

Coldwell, W. (2016, July 15). Nightlife reports: Clubbing in Berlin. *The Guardian*. Retrieved from https://www.theguardian.com/travel/2016/jul/15/berlinclubs-nightlife-germany-techno

Cross, L. (1968). Electronic music, 1948–1953. *Perspectives of New Music*, *7*(1), 32–65.

de Bertodano, H. (2015). DJ Tatiana Alvarez: Why I had to dress as a man to get ahead. *Telegraph*. Retrieved from https://www.telegraph.co.uk/women/womens-life/11350294/Why-I-had-to-dress-up-as-a-man-to-get-ahead-DJ-Tatiana-Alvarez.html

Discogs. (2018a). Apparat on Discogs. *Discogs.Com*. Retrieved from https://www.discogs.com/artist/50406-Apparat

Discogs. (2018b). Richard D. James on Discogs. *Discogs.Com*. Retrieved from https://www.discogs.com/artist/435132-Richard-D-James

Discogs. (2019). Traumprinz on Discogs. *Discogs.Com*. Retrieved from https://www.discogs.com/artist/2300500-Traumprinz

Formilan, G., & Boari, C. (2018). Do you note me? Social and cognitive dimensions of categorization in the evaluation of category-spanning creative products. In S. Consiglio, G. Mangia, M. Martinez, R. Mercurio, & L. Moschera (Eds.), *Organizing in the shadow of power: Voices from the Italian community of organization studies* (pp. 555–606). Studi MOA. Roma: Minerva Bancaria.

Garcia, L.-M. (2015). Beats, flesh, and grain: Sonic tactility and affect in electronic dance music. *Sound Studies*, *1*(1), 59–76.

Gilbert, J., & Pearson, E. (1999). *Discographies. Dance music, culture, and the politics of sound*. London and New York: Routledge.

Goffman, E. (1959). *The presentation of self in everyday life*. New York: Random House.

Guggenheim, M., & Potthast, J. (2011). Symmetrical twins: On the relationship between actor-network theory and the sociology of critical capacities. *European Journal of Social Theory, 15*(2), 157–178.

Hebdige, D. (1979). *Subculture, the meaning of style*. London, UK: Methuen.

Hennion, A. (1997). Baroque and rock: Music, mediation and musical taste. *Poetics*, *24*, 415–435.

Hennion, A. (2009). Talking music, making music: A comparison between rap and techno. In D. B. Scott (Ed.), *The Ashgate popular musicology research companion* (pp. 535–555). Farnham, UK: Ashgate.

Hesmondhalgh, D. (1997). The cultural politics of dance music. *Soundings*, *5*, 167–178.

Hesmondhalgh, D. (1998). The British dance music: A case study of independent cultural production. *British Journal of Sociology*, *49*(2), 234–251.

Hesmondhalgh, D. (2008). Towards a critical understanding of music, emotion and self-identity. *Consumption, Markets and Culture*, *11*(4), 329–343.

Hofer, S. (2006). I am they. Technological mediation, shifting conceptions of identity and techno music. *Convergence: The International Journal of Research into New Media Technologies*, *12*(3), 307–324.

Hutter, M., & Stark, D. (2015). Pragmatist perspectives on valuation: An introduction. In A. B. Antal, M. Hutter, & D. Stark (Eds.), *Moments of valuation. exploring sites of dissonance*(p. 4:16). Oxford, UK: : Oxford University Press.

Jenkins, D. (2017, November 8). 11 iconic DJs tell us the acts that influenced them most. *DJMag*. Retrieved from https://djmag.com/content/11-iconic-djs-tell-us-acts-influenced-them-most

Kühn, J.-M. (2015). The subcultural scene economy of the Berlin techno scene. In P. Guerra & T. Moreira (Eds.), *Keep it simple, make it fast! An approach to underground music scenes* (Vol. *1*, 281–286). Porto: University of Porto. Faculty of Arts and Humanities.

Lange, B., & Buerkner, H.-J. (2012). Value creation in scene-based music production: The case of electronic club music in Germany. *Economic Geography*, *89*(2), 149–169.

Marcuse, H. (1964). *One-dimensional man: The ideology of advanced industrial society*. Boston: Beacon Press.

McCallum, R. (2018). DVS1: The techno purist rages against the machine. *DJMag*. Retrieved from https://djmag.com/content/dvs1-techno-purist-rages-against-machine

McCartney, N. (2017). Complicating authorship. *Performance Research*, *22*(5), 62–71.

McLeod, K. (2001). Genres, subgenres, sub-subgenres and more: Musical and social differentiation within electronic/dance music communities. *Journal of Popular Music Studies*, *13*(1), 59–75.

Millington, A. (2017). The wild life of Steve Aoki, one of the highest-paid DJs and most-travelled musicians of the planet. *Business Insider UK*. Retrieved from http://uk.businessinsider.com/the-life-of-steve-aoki-2017-8?IR=T/#over-time-he-learned-how-to-be-a-better-dj-and-started-working-on-vinyl-he-began-remixing-big-artists-and-eventually-started-releasing-his-own-music-5

Milohnić, A. (2017). How to do things with names and signatures. *Performance Research*, *22*(5), 85–93.

Moore, R. (2005). Alternative to what? Subcultural capital and the commercialization of a music scene. *Deviant Behavior*, *26*(3), 229–252.

Muggleton, D. (2000). *Inside subculture: The postmodern meaning of style*. Oxford, UK: Berg Publishers.

Muggleton, D., & Weinzierl, R. (2003). *The post-subcultures reader*. New York, NY: Berg.

Nelson, A. J. (2015). *The sound of innovation: Stanford and the computer music revolution*. Cambridge, MA: MIT Press.

O'Malley Greenburg, Z. (2012). DJs are the new rock stars. *Forbes*. Retrieved from http://www.forbes.com/forbes/2012/0820/feature-disc-jockey-skrillex-music-the-new-rock-stars.html

O'Malley Greenburg, Z. (2013). Electronic cash kings 2013: The world's 20 highest-paid DJs. *Forbes*. Retrieved from http://www.forbes.com/sites/zackomalleygreenburg/2013/08/14/electronic-cash-kings-2013-the-worlds-highest-paid-djs/

Pearl, M. (2017). Breaking through: DJ Wey. *Residentadvisor.Net*. Retrieved from https://www.residentadvisor.net/features/2918

Peros, M. (2014). The taboo of the term "Rave". *TRC*. Retrieved from http://trc.daily-beat.com/lifestyle/2014/12/the-taboo-of-the-term-rave/

Phillips, D. J., & Kim, Y.-K. (2009). Why pseudonyms? Deception as identity preservation among jazz record companies, 1920–1929. *Organization Science*, *20*(3), 481–499.

Pinch, T. (1993). "Testing—One, two, three ... testing!": Toward a sociology of testing. *Science, Technology, & Human Values*, *18*(1), 25–41.

Pinch, T., & Trocco, F. (2009). *Analog days: The invention and impact of the moog synthesizer*. Cambridge, MA: Harvard University Press.

Pite, C. (2015). Identity crisis: The secret world of aliases. *DJbroadcast*. Retrieved from https://www.redef.com/author/55c111b90bbadbea0d6ab15e

Pizzorno, A. (2010). The mask: An essay. *International Political Anthropology*, *3*(1), 5–28.

Prior, N. (2008). Putting a glitch in the field: Bourdieu, actor network theory and contemporary music. *Cultural Sociology*, *2*(3), 301–319.

Prior, N. (2018). *Popular music, digital technology and society*. London: Sage.

Rafter, A. (2018). Porter Robinson: "Electronic music's convergence with pop has stopped artists taking risks". *DJMag*. Retrieved from https://djmag.com/news/porter-robinson-electronic-musics-convergence-pop-has-stopped-artists-taking-risks

Redhead, S. (1997). *From subcultures to clubcultures: An introduction to popular cultural studies*. Cambridge, UK: Blackwell.

Reynolds, S. (1998). *Generation ecstasy: Into the world of techno and rave culture*. New York: Routledge.

Riley, S. C. E., Griffin, C., & Morey, Y. (2010). The case for "everyday politics": Evaluating neo-tribal theory as a way to understand alternative forms of political participation, using electronic dance music culture as an example. *Sociology*, *44*(2), 345–363.

Rys, D. (2016). Global electronic music industry, worth $7.1 billion last year, sees growth slow. *Billboard*. Retrieved from https://www.billboard.com/articles/business/7385168/global-electronic-music-industry-growth-slows-still-worth-billions

Sanders, B. (2005). In the club: Ecstasy use and supply in a London nightclub. *Sociology*, *39*(2), 241–258.

Sassatelli, R. (2019). Recognition and reception. On Pizzorno, identity and the mask. *Sociologica*, *13*(2), 39–43.

Schüßler, E., & Sydow, J. (2013). Organizing events for configuring and maintaining creative fields. In C. Jones, M. Lorenze, & J. Sapsed (Eds.), *The Oxford handbook of creative industries* (pp. 1–30). New York: Oxford University Press.

Scott, D. B. (1990). Music and sociology for the 1990s: A changing critical perspective. *The Musical Quarterly*, *74*(3), 385–410.

Silk, A. J., & Urban, G. L. (1978). Pre-test-market evaluation of new packaged goods: A model and measurement methodology. *Journal of Marketing Research*, *15*(2), 171–191.

St John, G. (2006). Electronic dance music culture and religion: An overview. *Culture and Religion*, *7*(1), 1–25.

Taylor, K. (2014). The life cycle of a career DJ: From 1 to 25 years. *DJ TechTools*. Retrieved from https://djtechtools.com/2014/08/26/the-life-cycle-of-a-career-dj-from-1-to-25-years/

Thornton, S. (1996). *Club cultures: Music, media and subcultural capital*. Cambridge UK: Polity Press.

Till, R. (2006). The nine o'clock service: Mixing club culture and postmodern Christianity. *Culture and Religion*, *7*(1), 93–110.

Weiss, J. (2017). For ItaloJohnson, anonymity is more than a gimmick. *VICE*. Retrieved from https://www.vice.com/en_uk/article/vv5gj3/italojohnson-interview-mix-us-debut

How to cite this article: Formilan G, Stark D. Underground testing: Name-altering practices as probes in electronic music. *Br J Sociol*. 2020;71:572–589. https://doi.org/10.1111/1468-4446.12726