20
Accurate But Not Reproducible? The Possible Worlds of Public Opinion Research
Felix Keller

Abstract. In a paper titled “Do the social sciences create phenomena? The example of public opinion research,” Osborne and Rose (1999) argued that there are no substantial differences between the construction of scientific objects in the social sciences and in the natural sciences. The existing differences are only gradual, not categorical. Consequently, the same criteria of reproducibility would be applicable to the findings of scientific research and public opinion research.

The thesis presented in this paper is that parallelizing the scientific objects produced by the social sciences and the natural sciences is valid only if the focus is the material techniques of the sciences: the scientific laboratory in the narrow, restricted sense. However, if one recognizes that the context and environment of the scientific objects in the sciences is completely different from those in the social sciences, the parallelization quickly ends. This fact limits the possibility of reproducibility, and alters the definition of the objects themselves.

20.1 Introduction

The criterion of reproducibility shows an ambivalent status in discussions about science: on the one hand, it is regarded as a “general norm of science” (Mittelstrass 1992, p. 61), a “gold standard of sciences”;¹ on the other hand, it is seen as difficult to meet even for the hard sciences (Giles 2006).

The question arises whether reproducibility can be a meaningful criterion for public opinion research at all, especially, since their status as a science is debatable if not doubtful compared to “exact” sciences like physics or chemistry (Collins 1994, Kuhn 2012, p. 15). And within the social sciences public opinion research is one of the most controversial areas. It is hard to see how such a strong norm as reproducibility is applicable to public opinion research without the obvious result that the criterion is never met and public opinion research and the social sciences in general are once again seen as the “impossible sciences” (Turner and Turner 1990).

¹See the introductory section to this volume for a general account of the concept, and Steine’s article (this volume) on the historical genesis of norms in the natural sciences.

© 2016 John Wiley & Sons, Inc. Published 2016 by John Wiley & Sons, Inc.
However, a careful look at the problems of reproducibility in public opinion polls and the strategies used to develop similar criteria shows that the epistemological situation of public opinion research is so specific that it can enlighten the broader preconditions of reproducibility, at least where social processes are involved. First, I want to discuss the intrinsic idea of public opinion research, and the value given to reproducibility and similar criteria within the discipline. Then I will discuss the status of the knowledge that the discipline creates, with reference to some epistemological reflections, and lastly, I will go back to the scientific status of reproducibility in public opinion research and in the social sciences themselves.

20.2 Reproducibility: A Missing Criterion in Public Opinion Research?

The public is flooded with the tables, numbers, and graphics used in gathering and presenting people’s opinions on almost every issue, whether of high or mediocre importance. Public opinion research has perhaps the highest visibility and greatest public presence of all the domains of the social sciences. Moreover, it professes to be just the scientific observation of opinions, a science with its own methods and techniques. However, if reproducibility is a core value of the sciences, then it may be surprising that in the important textbooks representing the current status and core ideas of public opinion research, the question of reproducibility is simply lacking.

In the SAGE Handbook of Public Opinion Research (Donsbach and Traugott 2008), the concept is completely missing. The terms “reproducibility” and “reproduction” do not appear in the standard textbook Polling America: An Encyclopedia of Public Opinion (Best and Radcliff 2005). It is the same situation in the often-used Public Opinion: Measuring the American Mind (Bardes and Oldendick 2012) – here too, the search for a discussion of reproducibility shows no results.

In Public Opinion Quarterly, the most influential public opinion research journal, existing since 1937, we find only one article whose main subject is reproducibility: “House Effects and the Reproducibility of Survey Measurements” (Smith 1982). All 76 other articles containing the term “reproducibility” refer to Guttman’s concept of a “reproducibility coefficient,” used in a specific statistical scaling technique (Guttman 1944), the scalogram analysis, with a different meaning of the word.

Why is the concept of reproducibility missing? Are the observations of public opinion research just not reproducible? Are they, therefore, without scientific
validation or objectivity? Are there alternative concepts? And do they have the same functions? In fact, Smith’s article on “house effects” (effects created by the survey institutions) is the only one that treats the concept of reproducibility directly. But Smith (1982, p. 55) articulated the same significance of reproducibility as the natural sciences do:

If a measurement of a particular phenomenon can be replicated by independent observation under like conditions, there is presumptive evidence that the measurement is, within a given degree of exactitude, accurate and reliable.

To investigate the reproducibility of public opinion research, Smith started an experiment asking four survey houses the same questions in “representative” surveys: two university institutes (University of Michigan and University of Chicago) and two of the most important commercial survey agencies (Roper and Gallup). To standardize the experiment as much as possible the questions were worded identically and the researchers tried to “minimize differences in timing and context” (Smith 1982, p. 55).

Differences were found, due to the concept of sampling and due to probability distributions, and that is no surprise (see Stahel, this volume, for a more detailed discussion). However, the differences among the survey houses were quite substantial for all of the answers. Let us look at a typical polling question:² Selected individuals were asked if the government should spend more or less money on “halting the rising crime rate.” In the four survey waves, the range of agreement that “too much” was spent on crime prevention was between 24.4% and 18.2% in each wave – a difference of more than 6%.³ Between 72% and 59.7% of respondents wanted the government to spend more money on halting the crime rate – a difference of 12%.⁴

²The exact wording of the question was: “First I would like to talk with you about some things people think about today. We are faced with many problems in this country, none of which can be solved easily or inexpensively. I’m going to name some of these problems, and for each one I’d like you to tell me whether you think we’re spending too much money on it, too little money, or about the right amount. First (READ ITEM A) … are we spending too much, too little, or about the right amount on (ITEM)? READ EACH ITEM; CODE ONE FOR EACH.” Capitalized letters signify interviewer instructions. The answer categories were: “Too Much,” “Too Little,” “About Right,” and “Don’t Know.”

³However, the scales were different: Two waves allowed the answer “Don’t Know,” two did not. At least in this case, these additional categories do not change the answer patterns and the relative differences systematically (in others they do).

⁴Smith used classical statistical tests to assess whether the differences were hazardous or significant, assuming simple random sampling – although there is probably no modern social survey using a clean random sampling (including this one). Nevertheless, Smith claimed that
So the identical question produced an incoherent pattern of answers among the survey waves, and the differences between the answers were statistically highly significant. Neither Smith, nor a later paper by two important authors in the field, Converse and Traugott (1986), referring to Smith’s experiment could explain why this was so. An unknown number and quality of relevant external factors that were not measured may be responsible for the differences. Converse and Traugott summarized them in a “total survey error” which does not focus only on sample errors, but on any source of error.

Smith showed the importance of the reproducibility criterion in public opinion research not only using methodological reasoning, but also based on empirical investigation. There is no argument why reproducibility should not be a criterion as in the hard sciences, so it is very surprising that the scientific community discusses the problems of reproducibility in public opinion research so marginally. At most, reproducibility is discussed in the sense of the related concept “reliability”, and almost always in regard either to panel studies or to the stability of scaling techniques. The whole experimental situation of the “house effects” and of what an independent reproduction of the experiment would require (Mittelstrass 1992) are barely ever discussed.

Public opinion research basically seems to ignore this general norm of science but still claims to be a scientific method. This leads to the question whether alternative techniques are used for producing scientific-community and public trust in its numbers (Porter 1995), and if so, how similar to reproducibility these techniques are.

Since no science is planned on the paper desks of methodologists and then executed in scientific reality, the best example of a development of alternative techniques has a probability of $p < .0001$ of being accidental (if the “Don’t Know” category is excluded, $p = 0.047$; Smith 1982, p. 67).

Recently Nate Silver also estimated house effects of survey institutions, regarding pre-election polls. However, his estimations of the effects are not explained and can, therefore, not be discussed here – see fivethirtyeight.blogs.nytimes.com/2012/06/22/calculating-house-effects-of-polling-firms/.

Despite their significance, if reproducibility is in fact a core value of science, the mentioned articles are rarely cited by the community: Smith’s paper 23 times, Converse and Traugott’s paper 27 times (see isiknowledge.com/wos). Other articles discussing reproducibility of the whole survey dispositive (not panel surveys where a person is repeatedly interviewed) could not be found.

More recently, a specific kind of problem with reliability in public opinion research surfaced in so-called question order effects, where the order in which successive questions are asked in a questionnaire or survey makes a difference for the outcomes. A pertinent study of these effects is due to Moore (2002). For approaches to understand them see Wang and Busemeyer (this volume).
concepts for creating trust in numbers is the paradigmatic breakthrough, in Kuhn’s sense, by which the technique of polling entered the scientific field. Textbook science focuses on this breakthrough to celebrate the advent of modern scientific polling. But it is not that easy: Parts of the story are not told in the textbooks, though they are of interest concerning the reproducibility and credibility of public opinion research.

20.3 Big Data versus Science: The Breakthrough of Modern Polling

In the United States, where modern polling originated, predictions of election results were part of the political folklore at least since the beginning of the 19th century (Rényi 1998, p. 342; Smith 1990, p. 21f). Almost every occasion where people met – army meetings, festivals, river crossings – was used to gather votes for presidential candidates. The results of these “straw polls,” as they were called, were submitted to the local newspapers. Soon the newspapers themselves initiated what they called “probe votes,” since the public and the political representatives were curious about the findings (Herbst 2003).

Straw polls were spectacular media performances. Between 1916 and 1926, over 60 straw-poll enterprises were founded. To be competitive with the other forecasters, they tried to optimize their samples by increasing the number of participants in their polls. The obvious philosophy behind this strategy was: the more answers gathered, the greater the validity and credibility of the results. A big data industry ante nomine flourished: soon straw polls were collecting tens of thousands of ballots (Bradburn and Sudman 1988, p. 13).

One of the biggest players in this big data game was the New York magazine The Literary Digest. In 1936 its database contained 10 million voters (1936a). In the elections of 1920, 1924, 1928, and 1932, its polls were right. “Not only right in the sense that they showed the winner,” the magazine wrote (1936b), they forecast the actual popular vote, with such a small percentage of error [less than 1% in 1932] that newspapers and individuals everywhere heaped such phrases as “uncannily accurate” and “amazingly right” upon us.

A genuine professional experience in building databases with heterogeneous sources was the foundation and basic substance of these polls (Robinson 1932, Chap. IV): gather addresses from almost every source, ask people to describe the different aspects of their political life, summarize the results, and present them as the volonté générale of the people.
The practice of linking the results of different polls with elections is exemplary for the discussion of the scientific value of the techniques: a correspondence between a poll and an election means the poll was "unbiased." *Literary Digest* ascribed this quality to its own polling without hesitation: "It has always previously been correct," they boasted, "even its critics admit its value as an index of popular sentiment" (1936a). In 1928 *Literary Digest* took a poll with 18,000,000 ballots. Given the enormous organizational resources needed, even without the help of modern computing techniques, *Literary Digest*’s straw polling was big data analysis in the modern sense. However, whether the forecast would be correct was still in the “lap of the gods” (1936a).

In its last straw poll before the election, *Literary Digest* took a poll of ten million ballots and failed. After receiving more than 2,300,000 ballots, it counted 57% votes for Landon and 43% for Roosevelt. The *American Institute of Public Opinion* predicted 55.7% for Roosevelt and 44.3% for Landon. In the election, Roosevelt received 60.8% and Landon 36.5%.

George Gallup, based on a sample of only a few thousand probe votes, predicted the right candidate: Roosevelt. The *Literary Digest* promised to wear sackcloth and ashes ("Is our face red!" (1936b)) and in the end had to close its doors. This event has been celebrated as the breakthrough to modern scientific polling. A new star in the heaven of the social sciences had risen (Field 1990, p. 35):

Even a cursory historical review of the use of the questionnaire survey method in US public opinion research clearly reveals that the so-called “modern era” of polling began in 1936.

By now straw polls are banned in textbooks as examples of wrong, non-scientific sampling. What seemed to be a breakthrough is just a “statistical myth.”

### 20.4 The Birth of a Statistical Myth

Since history is written by the winners, the most popular explanation of the obvious success of the new techniques was what Gallup and his *American Institute of Public Opinion* postulated: the *Digest*’s samples were biased by the criterion of social class. They focused on the upper classes of society, and the workers were underrepresented (Gallup and Rae 1940, p. 42). Surely the election of 1936, in the time of the New Deal, was a special event that redefined the traditional conflict lines and party preferences.

A new “struggle between the haves and the have-nots” replaced the traditional conflict lines. It was believed that the unemployed voted for Roosevelt
with a quota of about 80% (Kaufmann 1942, p. 46). In general, the votes for the Democrats originated (Gallup and Rae 1940, p. 48)

from the ranks of the “ill-fed, ill-clothed, and ill-housed” lower third of the population, the group to which the Digest’s sampling plan failed to give full weight.

And why did the Digest sample get it wrong? Because the Digest, in contrast to other straw-poll institutions, started its sampling process with the purchasers of its own books and magazines, a database containing the names of the reading public: lawyers, architects, physicians, engineers, and members of high-society clubs. Then they extended it with addresses from the telephone directories and the automobile registration files (Robinson 1932, p. 56). So the conclusion was that because the telephone users and automobile owners were seldom in the strata of the “have-nots,” the Digest database had oversampled the middle and upper classes, and the lower classes were poorly represented.

But it is not that simple. The Digest, knowing the rumors about the bias of their samples and aware of the problem, initiated strategies to avoid a sampling bias. And after the fail, the magazine made a great effort to show that its database had indeed included the have-nots, that the idea they were “... not reaching certain strata simply will not hold water” (1936b).

However, there is a factor of the Digest’s method (an important difference to Gallup’s method) that is of great significance in terms of reproducibility: the Digest received its answers anonymously. The use of raw numbers is presented as a quality, an indication of the great authenticity of the polls, that they are not manipulated, that they represent the pure reality (1936a, p. 5):

These figures are exactly as received from more than one out of any five voters polled in our country – they are neither weighted, adjusted nor interpreted.

But, as a consequence, there was no possibility for a so-called post-stratification: no way to trace back to the respondents and verify which persons and with which characteristics had answered and to check whether there were segments of voters whose answers were inadequate. In the words of Literary Digest: “We don’t know what proportion went to persons who had voted for Roosevelt in 1932 or what proportion went to persons who had voted for Hoover, because our polls are always secret, and the ballots come back with no signatures, no identifying characteristics of any sort except the post-marks” (1936b, p. 8).
impossibility of tracing them meant the *Literary Digest* could not reproduce the problem that it was certainly aware of: Perhaps the bias was not in the sample but in the returns, and this was probably an important cause of the fail (Bryson 1976).

Gallup’s technique only counted several thousand ballots. His institute called this a *representative cross section of voters, or group sampling*, today known as a quota sample. With this technique the population, the whole country, was segmented along criteria that were *a priori* estimated to be relevant for political opinion making: income level, political preferences, gender, education, occupation, race, and religion – so that “the right proportion of citizens from every walk of life” would be represented in the sample.8

Is this the key to the success of modern polling? As mentioned, *Literary Digest*, too, used a form of quota sampling to avoid over-represented or under-represented samples. The difference is that Gallup’s new method of sampling established a loop back to the society the sample was drawn from, so the social structure of the sample could be compared with the structure of the population – but compared using only the criteria *already* defined as relevant.

To investigate the causes of *Digest’s* failure, the *American Institute of Public Opinion* asked the respondents a series of questions about their previous voting behavior, whether they owned telephones and cars, and whether they had received a *Literary Digest* ballot and sent it back. They were also asked if they had changed their minds after sending the ballot in (compare Squire 1988). The *American Institute of Public Opinion* data could clearly show that the *Literary Digest* polls were inadequate compared to its measurement system: Gallup’s survey signalized that 79% of respondents owning no car or phone voted for Roosevelt, whereas 55% of car and phone owners voted for Roosevelt (Squire 1988, p. 130).

These results are the origin of the textbook version of the story. But it is interesting that Gallup’s institute did not publish the basis of its calculations “explaining” the fail of *Literary Digest* (Squire 1988, p. 129), even in its compendium of all its questions and findings since its beginnings (Gallup 1972). Why? Simple calculations show that Gallup’s data is itself biased: if its replication survey had been as representative as claimed, Roosevelt would have received 66% of the votes instead of 61%, a difference equal to the *Literary Digest* er-

---

8 “When the method of group sampling is employed ... care must be taken to see that the national cross section includes the right proportion of citizens from every walk of life. A careful study must first be made of the population structure. By and large, the most important factors which have to be considered are occupational groups, income levels, political preferences, age, education, racial and religious groups” (Gallup 1944, p. 26).
ror (in the other direction). The reason for the bias in Gallup’s survey remains unknown. Perhaps, as Squire estimates, it was the psychological effect of respondents wishing they had supported the winners and not wanting to be on the loser’s side (Squire 1988, p. 131). But this insight is hypothetical, and other explanations remain speculations.

What is clear is that Literary Digest’s survey, and its problems, could not be reproduced even by the new method Gallup provided. Times had changed, the election was over, the voters were either losers or winners, and human memory keeps interacting with historical events. The object of the research had been modified irretrievably: it would take a time machine to reconstruct the failure. So the reasons why Literary Digest failed, after years of almost perfect predictions, are not so clear as the textbooks and their celebration of scientific sampling suggest.

There is evidence that the answers must have been biased by a mix of factors, where the very high non-response rates of the big sample also played a role (and hazard, also). And there is evidence that the question why Literary Digest failed could never have been answered (Squire 1988). That class-biased sampling led to the fail remains just “a statistical myth,” as Bryson (1976) wrote in the American Statistician – a myth generated to promote the new sampling techniques.9

Furthermore, the American Institute of Public Opinion, along with other institutions, experienced a similar disaster some years later in 1948 (Bradburn and Sudman 1988, p. 29). It estimated that Thomas Dewey would receive 49.5% of the vote and Harold Truman would lose with 44.5%. It turned out that Truman received 49.5% and Dewey received 45.15%. Surprisingly, that disaster was not only seen as bad luck, as Literary Digest had seen its disaster, but rather as a moment to reflect upon and calibrate the new instruments for measuring political life.

Finally, the discussion of this fail brought in a new scientific norm (and not only as a success in terms of public attention) by which to evaluate survey techniques: the criterion of the accuracy of election predictions. A special analysis of the poll error by the American Institute of Public Opinion explained, on the basis of the second questioning of earlier respondents, that the institute had stopped its pre-election interviews too early, at a moment when many people had not yet

---

9 The strategies behind the disqualification of Literary Digest’s big data method as inadequate can also be regarded as a kind of “boundary work” (a concept that Thomas Gieryn (1983) introduced some time ago: strategies to establish a demarcation line between science and non-science in the emerging field of the quantitative social sciences). See Reinhart’s article in this volume for a discussion of boundary work in relation to reproducibility in social science.
decided how they were going to vote; and that the campaign was still going on
and people were changing their minds (Gallup 1972, p. 769ff.).

Since the same instruments were used to explain the failure as had produced
the failure (sample surveys), the explanation remains a hypothesis. We find the
same situation as in the Literary Digest disaster. Since the political society as
it existed before the election could not be reproduced after the election, the
explanation of the fail had to remain logically unclear. The question persists in
every attempt to replicate a sample: Did the society and its discourses change,
or were the samplings inadequate?

20.5 Generating Trust

As a consequence, the prestigious Social Science Research (SSR) Council, founded
in 1923 to promote social science research, observed with consternation the “wide
confusion and misgivings about the reliability of the polls” (SSR 1948, p. 799).
The council reclaimed rigid robustness tests of the findings (SSR 1948, p. 622,
my italics):

Good tests of the accuracy of polling methods are important in the develop-
ment of polling and survey methodology. Without suitable tests there
is danger that users of those methods will drift into an unjustifiably high
degree of confidence in the accuracy of their current methods.

While the council recognized the advent of public opinion research as a ubiquitous
tool for producing representations of political opinions, it demanded that the
findings be grounded in “real” political life and that a comparison of the poll
results with the election results must remain the central criterion of robustness
and reliability.

This is called “accuracy.” A different test of accuracy, or even a different
replication method, is hard to find: the SSR postulated the election forecast as
the only way to prove the worth of any polling institution (SSR 1948, p. 602).

10The title of this chapter is a reference to Theodore Porter’s work Trust in Numbers (Porter
1995), which investigates the broader history of statistics to search for the roots of credibility
of statistical numbers. See also Porter, this volume.

11It is remarkable to observe how the new science was searching for a criterion of stability,
called reliability or accuracy. Of special interest is an article by Lucien Warner. He denies the
criteria of strict accuracy of election predictions (Warner 1939, pp. 389f.): “But even elections
are a far from satisfactory check upon public opinion surveys, for they do not reflect public
opinion with complete accuracy. Many factors distort electoral results.” Even the weather, he
believes, has “a selective influence upon rural versus urban votes.” De facto, only a census of
Election returns are not a direct or even a good test of the adequacy of polling on issues like the Marshall plan, civil rights, etc. But elections are useful for testing the adequacy of polling methods for estimating the percentage of the vote going to each candidate from various groups in the population. No better test is now known.

Perhaps it is this document that should be associated with the birth of modern public opinion research, rather than the myth of the breakthrough in 1936.

The argument is that the accuracy of polls in predicting election results is a criterion for judging the reliability and validity of public opinion research. Elections figure, ironically enough, as the main test of the whole system of public opinion research: the researchers observe whether the elections reproduce the survey results. This criterion obviously serves the same function in public opinion research as reproducibility does in the experimental sciences: proof that the phenomenon being measured is not an artifact of the instruments and the experimental devices themselves, but "exists" in the real world, so a result can be "verified with an external source" (Traugott 2008b, p. 408). So far, no other agreed measures or standards exist for evaluating the accuracy of polls (Traugott 2008b, p. 408; Traugott 2011, p. 316).

Due to the rarity of such verifications comparing polls with elections, "accuracy" remains the hard criterion of public opinion research to this day (Gosnell 1937, Sheatsly and Hyman 1953, pp. 480ff, Traugott 2011). Even the publication itself of the differences between predictions and election results has the character of a ritual. For example, the American National Council of Public Polls has published its analyses of presidential election polls starting from the origins of modern polling. Every new method of questioning is critically reviewed and measured against other methods. But what does this evidence mean in the broader sense of the reproducibility of the findings that public opinion research presents to the public? What is the point of a science predicting a phenomenon that surely and soon will define itself (and without any doubt): election results?

The answer to the second question is easy. Public opinion research is not primarily interested in predicting elections, probably even in financial terms. Elect...
tion surveys are only a small part of all the public opinion research being done. Gallup (1939, p. 10) has already formulated the reason why the scientific and public focus is on the election forecasts, in the loaded words:

The chief usefulness of the sampling survey, however, does not lie in forecasting elections. While such forecasts provide an interesting and legitimate activity, they probably serve no great social purpose. They do, however, offer the public a convincing demonstration of the accuracy of sampling procedures.

This says that if the sampling procedure is correct or accurate in terms of an election forecast, people can be asked almost anything and their answers will accurately represent the whole, the nation. This transfer of “trust in numbers” over to virtually any question that may be asked is the genuine concept of public opinion research to this day (Gallup 1944, p. 26).

There is no room here to seriously discuss all the problems of public opinion research methods. The validity of the answers (do people understand the answers in the same manner? do they answer what they actually think? are the answers the effect of the questions?) and the sampling problems (is the statistical urn well mixed? did all respondents get equal chances to participate?), to name some examples, are all hard problems that have been debated since the beginning of modern polling.\textsuperscript{13}

But under the fictional assumption of an ideal world where all methodological problems have been solved, i.e., where sampling bias and validity of questions and answers pose no problems anymore, we can raise some general questions about the reproducibility of scientific and social knowledge where public opinion polls build an instructive case. This is a way to answer our first question: What are the strategies for establishing accuracy in the broader context of scientific reproducibility?

\textbf{20.6 The Possible Worlds of Public Opinion Research}

In the sense of an ideal typical or pure type as postulated by Max Weber (which means: a situation cleared of reality and reduced to its principal idea or logic, so its pure principle and consequences can be studied), we parenthesize all the methodological problems of polling. We assume an ideal world of the pollsters in which to make our reflections on accuracy and reproducibility. Following this postulate, every poll and every poll company creates what analytic philosophers call “a possible world.”

\textsuperscript{13}See the reconstruction of the debates in my archeology of opinion research (Keller 2001).
For Kripke (1980, pp. 15ff) a possible world is nothing like an exotic world on a distant planet – it is more like a possible state of a situation. Thinking with possible worlds has a lot to do with the probability and possibility of events, with chance and, therefore, with statistics, as Hacking (1975b, Chap. 14) showed in his investigation of the emergence of probability.

Kripke illustrated a possible world with a well-known probability experiment: throwing dice. The dice on a table have 36 possible states, sums of the numbers of dots on each die. Ignoring everything about the world around them, they build 36 possible mini-worlds. Only one of these mini-worlds is actualized by a throw (e.g., 4 and 2 dots), the other 35 worlds remain only possible. But they can become an actual world at the next throw of the dice.

Defining a world with possible states, abstracted from all other aspects, is a very good approach to the idea behind polling: the sides of a throw and number of dots are equivalent to predefined answer categories (“you will vote for whom ...”) and distributions of answers. In this sense, it is not surprising that elections and election results were omnipresent in the possible-world discussion. For instance, Kripke chose the example of the 1968 election (Kripke 1980, pp. 40–53 and 115–116) which Nixon won. Kripke discussed a possible world in which Nixon does not win the election and asked (a question of minor interest to us here) if in such a world Nixon is still Nixon (Kripke 1980, p. 44).

Why does the 1968 election figure as the preferred example of a possible world? Because a different election outcome was easily imaginable: the 1968 US elections turned out to be extremely close. Richard Nixon won with 43.4% and was basically equal with Hubert Humphrey, who was supported by 42.7% of the popular vote, while the party-independent George Wallace reached only 13.5%.

A possible world in which Nixon might have lost the elections is of great plausibility, and such possible worlds were explored by the polls, each following the same categories and referring to the same population: before the elections the Gallup Poll measured 43% for Nixon, 42% for Humphrey, and 15% for Wallace. In contrast, the Harris Poll, another big organization, showed Humphrey as winner with 45%, Nixon with only 41%, and Wallace 14%. Given the probability errors inherent to sampling, each of the possible worlds could have been the actual one realized in the elections.

For a pollster, elections, except at their political constitutive moment, are indeed nothing but sample surveys. As Archibald Crossley (1937, p. 25), one of the founding fathers of modern polling, wrote about the Digest disaster:

---

The greatest difficulty of all is the fact that the election itself is not a census, but an application of the sampling principle. Every poll is therefore a sample of a sample. There is no guarantee whatever that the 1940 election will be the same kind of sample that the 1936 election was. And the 1936 election probably was a different cross-section of all voters from the cross-section of the 1932 election.

In other words, for pollsters elections and sample surveys both represent possible worlds in an epistemic sense (this is correct, when election results and results of polls both use descriptions like “the Americans,” an imagined community never counted as a whole, neither in polls nor in elections). It is the power of state, not the pollsters’ methodology, that lets the possible world of elections become the “real” one.

Here, a conceptual differentiation is necessary on what “possible” means. As Hacking (1975b, p. 123) insisted, grammatically the term “possible” has two aspects: “possible for” and “possible that.” There is an important difference between these two kinds of possibility: polls showed that it was possible that Nixon wins and, at the same time, that it was possible for Nixon to win. The latter is the possibility in the sense of the Latin potentia (power, ability): it signalizes an occasion realizable in the real world. The “possibility that” is called epistemic probability (Hacking 1967, p. 148) and concerns the sphere of chance and calculations, a sphere that does not necessarily affect the real world. The range of error ritually mentioned in the published results of polls (usually 3%) is a signifier of this kind of epistemic possibility (even this range is seldom used statistically correctly).

The latter aspect shows, moreover, the political potential: the results of the Harris Poll signalized that it was still possible for Humphrey to win the elections (with 45% of supporting answers). In other words, the concept of “possible worlds” includes two aspects (Hacking 1975a, p. 336): the possibility of a state in the sense of its probability, and the possibility of a state regarding its potential to realize as an “actual world.”

Crossley’s statement above can be reformulated as follows: Since an election is not a census, it is – at an epistemic level – one of a number of conquering possible worlds each containing information about whom the American population will prefer as president. In the case of elections and polls, it is evident that the scientific prestige of the polling institutions is based on the epistemic aspect: the possible worlds generated by polls are structurally similar to the world realized by elections.

Now we are at the root of the strategy of replacing reproducibility with

---

15 See Stahel (this volume) concerning the mathematical background of confidence intervals.
accuracy: The accuracy of poll results to an election does indeed mean the surveys are reproduced by another “experiment” with quite a similar structure. But the reflections above show that there is a second aspect of possible worlds affecting the scientific nature of the results: the findings may show that it is possible for Humphrey to win. This aspect of the pollster’s potential world is inscribed in political life itself, and it transforms the character of the epistemic aspects of the generated knowledge (and its scientific character too).

Here we have a possible difference between the social sciences and the natural sciences. The description of the world changes the world described, but meanwhile, also, the experiment observes the experimenter. And both this description and this observation are using the same techniques: human perception and symbolic communication.

20.7 Looping Effects between Measurement and Measured

Since modern polling began there have been dozens, maybe hundreds, of articles on the effects of public opinion research upon the public, the media, and the political system. These influences are located on at least three levels: the interrogated population, the media system, and the political system and its actors.

On the first level, Gallup and Rae (1940, Chap. 20) already took care of the so-called bandwagon effect: In a circus, the bandwagon carrying the music band leads the caravan of artists and animals. It figures as a metaphor for the leading candidate or party (Marsh 1985) which people tend to applaud because it is the most visible and attractive wagon in the caravan. However, there is also the so-called underdog effect, the tendency to support the losers (Traugott 2008a). Whether results of polls initiate more self-fulfilling or more self-destroying prophecies is difficult to decide (cf. Collins and Zimmerli, both in this volume, concerning similar cases of “Merton-sensible” systems). As a consequence, the impact of polling on individual voting decisions remains unclear (Lang and Lang 1984), although there is strong evidence that polling has effects upon individual perception and behavior (see Moy and Rinke 2012 for a review).

Second, there is also an effect on the media system: creating collective perception and agenda setting (Champagne 1990). Polling results are welcomed by the media as easily consumable “news” (Herbst 2003, pp. 588f), even if the results signalize only a slight difference to a previous state of affairs (sometimes not even exceeding their statistical range of error). By financing studies, parties, and enterprises, all sorts of collective actors can guide the public’s collective focus.
on a real or imagined problem or issue. If people are asked about something, they will answer somehow (Bourdieu 1992), even on completely imaginary issues (Fishkin 1995, p. 82). There is as much research on the effects of public opinion research on the media (Roll and Cantril 1980) as there is on its effects on individuals (Strömbäck 2012).

The close relationship between political actors, pollsters, and media is even described as an “opinion connection” (Cantril and Cantril 1991). Experiencing the strategic uses of public opinion research and techniques that are opaque for outsiders produces a feeling that the media and the public are “being hijacked by campaign strategists, and focused on imagined flaws in survey methods” (Jacobs and Shapiro 2005, p. 638), or even being manipulated. So the judgment of the press about polls is not always friendly (though this does not stop them from publishing the pollsters’ results).

*Bloomberg Businessweek* wrote about the “Dark Art of Polling,” about how nobody knows how the pollsters brew their results in their quasi-alchemistic labors. Allan Rivlin from the *National Journal* is more despairing: “First, kill all the pollsters” he titles an article about what he calls the “junk-food polls,” and asks “Who’s manipulating whom?” Just as the concrete direction of the influences on individual behavior remains unclear, the concrete interlooping effects between the polls and the media also remain unknown, even to the pollsters themselves (Moy and Rinke 2012), and this makes any use of public opinion research a high-risk endeavor, as the following quotation illustrates.

The body of American public opinion is far too sophisticated and subtle a beast to be prodded in this way. And as if to prove it, on Tuesday the beast bit back.

---

16 Polls whose only goal is agenda setting are called “push polls.” See Medvic (2003), and also NCPP’s norm articulated against push polls: ncpp.org/node/4/#16. However, the demarcation line between push polls and regular polls is not always easy to identify.

17 Compare the election of Jacques Chirac as French president (Delpey 1995). However, there seem to be no great national differences in the skeptical attitude toward polls, see an example for Germany by Güllner et al. (2007).

18 “How could a Gallup organization survey published a week before the election show Mitt Romney up by 5 percentage points, while a CBS / New York Times poll from the same period put him 1 point behind President Obama? Even professional poll watchers don’t know.” Quoted from Peter Coy, November 01, 2012 at www.businessweek.com/printer/articles/80092-the-dark-art-of-political-polling.

As the third and final point of this short overview, there are direct influences on the public support of the candidates running in the elections. Rising poll support usually generates a monetary push for the candidate’s campaign, while declining support dries up the money flow and also the candidate’s public presence (Converse and Traugott 1986, p. 1094). In the US system, third party candidates, since they normally receive little support in the early stages of the election campaigns, are given scarce mention in the polls. As a consequence, they disappear from the answer categories step by step, therefore losing public visibility and financial support almost completely (Herbst 2003, p. 589).

In this logic, it is evident that the reputations of the American presidents and candidates for political office and their chances in the elections are not only observed in the polls, but also actively used by the candidates (Holtz-Bacha 2012, p. 272). The use of polls in campaigns is so prominent that in a poll about polls, 42% of the respondents said the pollsters were manipulating their results to show Obama ahead. This poll about election polls is an example of how the possible worlds of polling have established a highly self-referential system, a complex system of looping effects between polls, media, and individuals.

There is no room here to discuss which of the many supposed effects of polling are real or not, and the directions of their influence: the published research about this problem area fills libraries. Since polls generate public knowledge, it is evident that they affect public knowledge and behavior. Here these effects are mentioned only to show the importance of the polls’ potential power to change real worlds – the validity of the measurements aside. But what do these evident effects mean with regard to the question of accuracy or reproducibility in general?

20.8 Swarms of Possible Worlds

The National Council on Public Polls analyzed the final presidential election polls conducted by the national media dating back more than 50 years. Compared with actual election outcomes, the average error of polls for presidential elections between 1956 and 1996 declined. In terms of criteria of quality, the Council suggests that the polls were getting more accurate over time; that the polls were in fact reproduced in the elections.

This evidence can also be used to support the argument that public opinion

---


research has successfully created its own object: public opinion as measured by its surveys (Blondiaux 1998, p. 71, Osborne and Rose 1999, p. 389), and that it is capable of stabilizing and legitimizing its own phenomena using the elections as another independent experimental system. But here some caution is recommended: phenomena that are stable in time and space, obviously describable, are rare in the sciences, says Hacking (1983, p 227) – even in the natural sciences. Though it seems correct to say that public opinion research talks mainly about “public opinion” today, the conclusion is not necessarily correct that the underlying political articulation is sufficiently established by the tables and numbers of the polls.

After 80 years of research, there is still no consensus about how the phenomena of “public opinion” can exactly be defined, self-contained, and distinguished from other similar phenomena (Holtz-Bacha 2012, pp. 268f). In spite of this fact, public opinion research has successfully transformed the question of how to define public opinion into the question of the interplays between the results of polls and those of elections and to this day that question prevails. However, the election results figure as robust knowledge not because they represent the real expression of public opinion, but because the source of this robustness is state power, legal rules, and even the police who regulate and legitimize the election results. Outside their laboratories, using the concept of accuracy to validate their constructs, the pollsters walk on crutches built of pure governmental techniques.

But more important is that the polls and the elections are not independent of each other. Since the election results temporally follow after the election polls, they can be regarded, to a certain but unknown degree, as an effect of the polls themselves, through the many influence-channels mentioned above.22 They are never an independent verification of the polls. The higher correspondence between polls and elections observed by the National Council on Public Polls could, therefore, be a consequence of the fact that the elections are more and more structured, or even influenced, by the dozens of earlier election polls.

The possible worlds of elections and polls, intermediated by mass media and spin doctors, interact strongly, and stabilize each other. It does not make any sense to regard polls as unique phenomena constructed by lonely and autonomous social scientists in their laboratories. In other words, as soon as an observation of these social realities is fixed and communicated, it becomes a component of what it observed and its original precondition has changed. We find the situation

22In this sense, creating, pushing, public opinions in public opinion research is perhaps better understood as an act of agenda setting than as the measurement of an already preformed entity (Holtz-Bacha 2012, pp. 268f). In recent sociological discussion, this is called the performative aspect of polling (Law 2009, Perrin and McFarland 2011, pp. 101f).
of a so-called infinite regress that has no end. So the possible worlds of public opinion surveys may be built correctly in terms of the criteria the survey science sets for itself, but since its findings alter the reality it measures, there is no way to reproduce the status quo ante.

This is the destiny of the social sciences, in general, just expressed more clearly in public opinion research than elsewhere: Social scientists' descriptions and measurements are not outside the realities they describe, they form a part of them. All instances, scientists, interviewers, politicians, and media interact on the same level, using the same form of symbolic communication. The scientific observation of society takes place within the society it describes (Luhmann 1992). No external (scientific) observation is possible,23 there are only observers who observe observers.

Or, regarding epistemic techniques of public opinion research, a poll “measuring” a public opinion generates nothing other than a possible world of peoples’ political articulation: a world in concurrence and interaction with a whole swarm of other worlds, all pretending to express the political opinions of the people. Trying to reproduce one of these possible worlds in order to show its accuracy does nothing but generate another one. Amazing.

References

— (1936a): Landon, 1,293,669; Roosevelt, 972,897. Final Returns in The Digest’s Poll of Ten Million Voters. The Literary Digest, 1936, pp. 5-6.


23Osborne and Rose (1999, p. 389) argued that there are no substantial differences between the construction of scientific objects in the social sciences and in the sciences. Although there are variations, they should be understood as “matters of degree rather than in terms of a rigid line of demarcation.” However, we insist on the fact that in the case of observing society, there is no external observer. There are only observers interconnected by the medium of symbolic communication (as Luhmann argues). See also Collins’ chapter in this volume concerning a broader context of this question.


