The Peilingwijzer as Measurement Model

Michiel Bosma

July 2022

Supervisors: Marcel Boumans and Janneke van Dis
Student Number: 5663067
Name of Programme: History and Philosophy of Science, Utrecht University

$$P_i \sim \mathcal{N}(M_d, F)$$

$$M_d = A_d + H_{bi}$$

$$A_d \sim \mathcal{N}(A_{d-1}, \tau)$$

Contents

1 Introduction

1.1 Motivation and context

There are all kinds of epistemic activities. Counting, narrating, ordering, explaining and much more. Epistemic activities can be made complex through combination. Measuring the political support for political parties is such a complex epistemic activity. Then, the instruments which are commonly used for this epistemic activity are polls. Polls measure the status of public opinion at a certain moment in time and are used in almost all modern, democratic societies. Especially, before elections, they are an important instrument to inform the public (Andersen, 2000). However, these instruments can not be regarded as just neutral instruments. There is evidence that published opinion polls do influence the outcomes of elections. Voters take the polls into account in determining their votes. Dahlgaard et al. (2016, p. 283) state: 'Similar to all other types of information, public opinion polls can influence public opinion'.

An important way in which the polls influence the outcome of elections is the bandwagon effect. The bandwagon effect describes how voters, will be more likely to vote for a party or candidate that is gaining in the polls (Marsh, 1985). They are figuratively speaking jumping on the bandwagon. Evidence for subtle bandwagon effects is also found in the Dutch context, where the description of the support for a political party being in a positive trend made an impact (Van der Meer, Hakhverdian, and Aaldering, 2016). The exact impact is difficult to determine, and the impact may be slight, however, within the Dutch elections even a small difference can already have a large impact on the outcome of the political decision processes since small majorities are common.

Moreover, the polls are used by media parties to invite certain political parties or candidates to political debates (RTL-Nieuws, 2021). When there is only space for a certain amount of political parties in the debate, the polls are consulted to determine who should be invited. For example, only those parties or candidates that are likely to win a seat in parliament are invited. These debates are an important platform to reach potential voters and allow them to learn about their political positions (Van der Meer, Walter and Aelst, 2016). By determining who is invited to a political debate, polls are a factor in our elections and therefore in our political decision-making process.

Therefore, we can conclude that polls are an impactful measuring instrument, which warrants investigation of their functions and assumptions. To narrow down the study, the focus will be on a certain model in the Netherlands that uses multiple polls to arrive at its measurement outcome. This model, called the Peilingwijzer, will be investigated in this thesis.

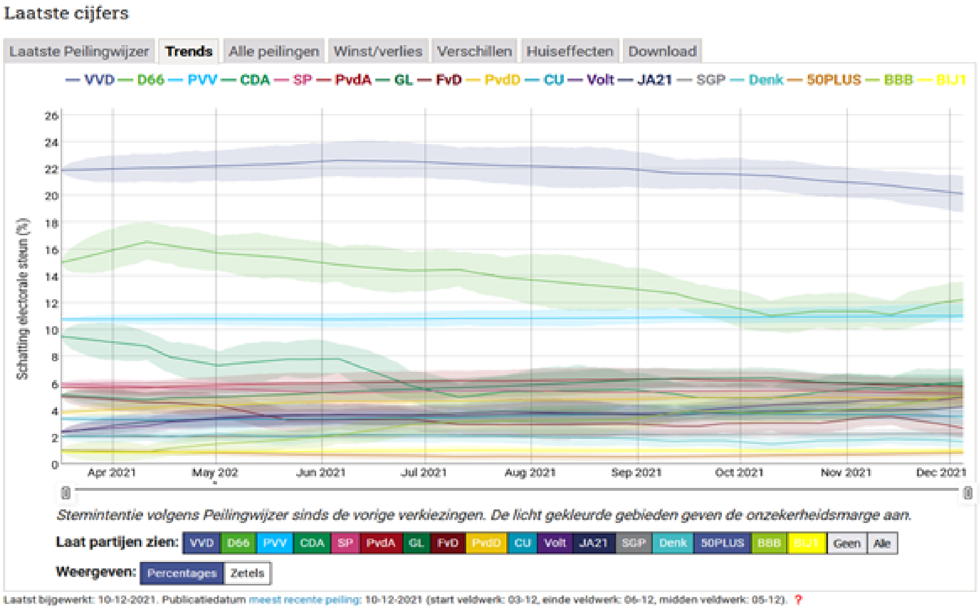
1.2 The Peilingwijzer

The Peilingwijzer is created by Tom Louwerse (2011) as a measurement of electoral support for Dutch political parties. The Peilingwijzer does not collect data but uses the data collected by three different polling stations, which are De Peiling (IPSO/Een Vandaag), IO Research and Kantar, to obtain its estimates (Louwerse, 2011). The goal of the Peilingwijzer is stated by Louwerse (2011) as the following:

> 'The Peilingwijzer makes based on available (House of Representatives) polls an as good as possible estimate of the electoral support for the Dutch political parties. Between the different polling estimates, significant differences are possible and often opposing trends are visible. The Peilingwijzer is an attempt to make sense of this in a responsible, scientific manner.'[1]

The estimates of the Peilingwijzer are updated every time new data is published by one of the polling stations. In Figure 1 the estimates of Peilingwijzer are plotted. The Y-axis signifies the percentages of electoral support for a political party. Then, the X-axis signifies the time variable across which these percentages are plotted. Furthermore, every coloured line represents a specific political party, as indicated by the legend below. Lastly, the area around the lines determines the 95 percent uncertainty intervals around the percentage point estimates.

Figure 1: Estimates of the Peilingwijzer



Stemintentie volgens Peilingwijzer sinds de vorige verkiezingen. De licht gekleurde gebieden geven de onzekerheidsmarge aan.

Laat partijen zien: VVD D66 PVV CDA SP PvdA GL FvD PvdD CU Volt JA21 SGP Denk 50PLUS BBB BIJ1 Geen Alle

Weergeven: Percentages Zetels

Laatst bijgewerkt: 10-12-2021. Publicatiedatum meest recente peiling: 10-12-2021 (start veldwerk: 03-12, einde veldwerk: 06-12, midden veldwerk: 05-12). ?

These estimates are regarded as highly influential in the Dutch political context. In a column in the Dutch news outlet NRC, for example, Tom Jan Meuus (2017) describes Louwerse as follows: 'In a short period of time, he has grown to be one of the most influential people at the Binnenhof'[2] Moreover, estimates of the Peilingwijzer are the only ones that are published in the national news outlet NOS. It is only mentioned that the Peilingwijzer combines the data from the different polling stations, but the estimates of the polling stations themselves are not reported anymore (NOS, 2022). Finally, in the prime minister's debate, organized before the last elections, the estimates of the Peilingwijzer combined with the

[1] The original quote is in Dutch. This quote is a translation by the author

[2] The original quote is in Dutch. This quote is a translation by the author

current seats in parliament were used to select the 6 political parties that could participate (RTL-Nieuws 2021). We can therefore conclude that the Peilingwijzer has developed to be the measurement with the most authority in the Dutch political context.

1.3 Research Question

The authority of the Peilingwijzer comes from the idea that the Peilingwijzer is a more reliable estimate of the electoral support of political parties in the Netherlands than the individual results supplied by the polling stations (Hakhverdian, 2014). However, this reliability is based upon specific assumptions on the measurement process and the phenomenon, as we will see in this thesis. To understand the Peilingwijzer is to understand these assumptions, therefore. Then, we can hopefully become clearer if this position of the Peilingwijzer as the authoritative measurement in the Netherlands is warranted. The research question guiding this research will be, therefore:

Under which conditions does the Peilingwijzer work as a reliable measuring device?

The remaining structure of this thesis is as follows. In chapter 2 the framework developed by Bogen and Woodward (1988) of data and phenomena is explained, which allows us to become precise about the nature of measurement and measurement error. I will introduce the framework in 2.1, the inference from data to phenomena in 2.2, and the resulting view op measurement in 2.3. Chapter 2 therefore mostly sets the stage for the analysis that follows.

In chapter 3 the idea of modelling measurement is introduced to understand what the Peilingwijzer is precisely about. In 3.1 the idea of laboratory procedures for controlling measurement error is discussed, in 3.2 the calculus of observations necessary to replace the laboratory procedures is introduced and in 3.3 I explain the invariance assumption behind the modelling procedure. This chapter is the last theoretical chapter, which introduces the necessary concepts and ideas for the analysis of the Peilingwijzer.

Chapter 4 is the core of this thesis, where the ideas of chapters 2 and 3 are applied to the Peilingwijzer. In 4.1 I conceptualize electoral support as a phenomenon in the sense of Bogen and Woodward (1988) and Massimi (2011). In 4.2 I argue that the Peilingwijzer should be seen as a second-order inference from data to the phenomenon of electoral support. Then, in 4.3 the statistical model of the Peilingwijzer is introduced and its assumptions regarding the nature of the measurement errors and the phenomenon are discussed in 4.4, 4.5 and 4.6. 4.4 discusses the normality assumption, 4.5 the structural bias assumption and 4.6 the stable phenomenon assumption.

Then, in chapter 5, the assumptions in the Peilingwijzer are discussed to see whether they are reasonable under certain conditions. A critical judgment of the assumptions is performed in 5.1 and in 5.2 possibilities for improvement following the analysis are put forward. Lastly, in chapter 6, the thesis is concluded in 6.1 and some reflections are presented in 6.2

2 Data and Phenonema

2.1 Bogen and Woodward: Data and Phenonema

The framework that is used to analyse this question is the framework supplied by Bogen and Woodward (1988) and Woodward (1989) in their seminal papers about data and phenomena. In this framework, a distinction is made between data and phenomena.

Data are that which register on some sort of recording device. They are in principle accessible to human perception and are public records, meaning that they are not limited to a single access point (Woodward, 1989). Data are caused by all sorts of factors which are idiosyncratic to the specific way in which the data are collected (Woodward, 1989). These idiosyncratic factors are for example a specific experimental design, equipment that is used, the human that is registering the data and the culture in which the data are collected.

Phenomena, on the other hand, are features of the world which are relatively stable. Contrasting phenomena with data, we see that phenomena by definition do display those stable patterns that data are lacking. They are less closely tied to a specific detection or recording device through which data are produced (Woodward, 1989). The idea here is that phenomena exist to some extent independent of the specific way through which they are detected, and should therefore display some stability across different detection procedures. As Woodward (1989, p.395) states: 'One expects that a genuine phenomenon should be replicable or reproducible - it should be such that it recurs or can be made to recur in different situations or contexts'. This robustness of phenomena in contrast to data is a crucial difference in this distinction.

2.2 Inference from Data to Phenomena

In the framework of data and phenomena, data serve as a means of providing evidence for the existence of facts about phenomena. Here, data play the role of observables in the scientific process by being visible for direct human perception or through the use of scientific instruments (Bogen and Woodward, 1988). In any collection of data, many local causal factors, idiosyncratic to the specific data collection process, influence the specific data that is produced. In scientific practice, it is almost impossible to completely and fully understand all the local factors that have influenced the data production process, since these factors are idiosyncratic to the specific situation in which the data was collected. Consequently, these factors are not of interest to the scientist if the goal is to learn about stable features of reality. It is therefore then that the scientific process is not about collecting facts about data, but about phenomena (Bogen and Woodward, 1988). This highlights the necessity to infer facts about the phenomena from the data.

Then, data collection is often performed in what is described as a noisy environment. In the framework of data and phenomena, noise consists of those local factors which are specific for this data collection but do not form a part of the phenomenon. In attempting to filter out the noise, we attempt to infer the facts about the phenomena from all the local factors influencing the data. This signal which is present in the data is of interest to us, not the complete data itself. As Woodward (1989, p.396 - 397) beautifully summarizes:

> 'The problem of detecting a phenomenon is the problem of detecting a signal in this sea of noise, of identifying a relatively stable and invariant pattern of some simplicity and generality with recurrent features - a pattern which is not just an artefact of the particular detection techniques we employ or the local environment in which we operate'

The line of reasoning here is that of an inductive inference, in which we go above just the information that is present in the data (Woodward, 2011). As we make an inductive argument, we extend the description of certain characteristics of observed events to the corresponding characteristics of other events which have not been observed (Keynes, 1909). From the data, which we have fully observed, we make a statement about some phenomenon, which influence we assume to be present in the data, but that we have not directly observed. Whenever we make this inferential argument, we must take knowledge of the phenomenon and the nature of our observations into account (Keynes, 1909). This background knowledge is necessary to make the claim plausible that the nature of the things that we have observed and the nature of the things that we have not observed are such that we can make an inductive analogy between the two.

Through these additional assumptions about the nature of our measurement and the phenomenon that we are studying, the inductive inference can be performed and made plausible. Inferring facts about the phenomena from the data is not an easy task, however. All kinds of errors can be thought to occur which can warrant doubt about the validity of our claims about the phenomena. We could for example fail to control for background or confounding factors in our data or make mistakes in our statistical analysis (Woodward, 2011). What matters here is therefore not that we can provide a full causal explanation of the data, but that we can make the claim plausible that the data is reliable evidence for our claim about the phenomena (Woodward, 1989).

$$d_i = f(P, u_i)$$

In the equation above, we see how we can conceptualize the data production process in the presence of both the phenomenon P and additional causal factors or noise u_i. Individual data points d_i are conceptualized as being caused partly by the phenomenon P and the idiosyncratic factors u_i, which are different for each data point (Woodward, 2010). In the scientific process, we only observe a collection of data points d_i. The problem that arises is to make an inference from d_i to P in the presence of these additional factors u_i, which we are unable to control and understand fully.

2.3 Measurement

When measuring a phenomenon, therefore, based on data, we attempt to establish a quantitative fact about this phenomenon. This quantitative fact is of the form P_1 is true, where P_1 is a statement where we equate a phenomenon with a number (Woodward, 2000). Different quantitative statements of the form P_1 are possible and in our measurement procedure, the validity of these different claims is tested such that hopefully a fact about this phenomenon can be established. To establish this quantitative fact, therefore, an inference from the data to the phenomenon has to be made.

Moreover, in the description by Hasok Chang (2004) of the problem of nomic measurement, it becomes especially clear that what is actually at stake is precisely an inference from something observable, the data, to something unobservable, the phenomenon. The problem of nomic measurement is described as follows: When we attempt to measure a quantity, often we can not measure this quantity immediately, it is not immediately apparent to our senses. Therefore, we have to infer this quantity from another quantity. If that quantity is in a sense directly observable we can stop here, but often we need to infer this quantity from again another quantity. The question becomes then how we should infer our quantity from our directly observed quantity. We can use the idea of a functional mapping here to formulate the structure between these different quantities. However, we can never empirically directly observe the form that this function should take since that needs the actual values of our unknown variable, which was our unknown starting point, to begin with (Chang, 2004).

This idea of nomic measurement is complementary to the framework of data and phenomena and highlights that what is at stake is an inferential argument in which we attempt to learn a fact about something unobservable from something observable. It also highlights that this is not an easy task, since we need to make assumptions about the relationship between our observable quantity and our unobservable quantity. When modelling a measurement, therefore, we make a functional mapping from the data to the phenomenon in a modelling framework.

If we return to the equation describing the data production process, we can see that these additional factors u_i can be conceptualized as the measurement error in the data (Woodward, 2010). They form the difference between our data and our phenomenon and are consequently a description of the fact of how much we are off in measuring the phenomenon. Therefore, we need assumptions about the additional factors u_i, which comes down to assumptions about our measurement errors, to make a successful inference from the data d_i to our phenomenon P. These assumptions always have to be about something that we can not directly observe and can therefore not be empirically demonstrated, but only made credible.

$$d_i = f(P, u_i)$$

Since these additional factors u_i have a wide variety of different causes and attributes, a general theory of error that deals with these factors will always come short in addressing this (Bogen and Woodward, 1988). Nevertheless, there have been a wide variety of attempts at taming errors (Boumans et al., 2015). In the next sections, we will discuss strategies that scientists employ to do this and make this inference plausible.

3 Modelling Measurement

3.1 Laboratory Procedures

Woodward (1989) describes multiple procedures to address the reliability of the inference from the data to the phenomena. Much of these strategies can be grouped under what Boumans (2015) describes as laboratory procedures. Laboratories, in contrast to doing science in the field, are places where phenomena can be isolated from their environment and where they can be investigated through manipulation or intervention. We can for example control the influence of our confounding factors through the use of physical isolation to reduce the amount of noise u_i that influences the data apart from the influence of the phenomenon. Or we design the experiment in a laboratory in such a way that we have reason to believe that confounding or background factors operate uniformly and therefore can easily be filtered out (Woodward, 1989). Or we could control the influence of these confounding factors in a way such that we have reason to believe that we understand their influence on the data production process.

Laboratory procedures can therefore be understood as procedures of control of the additional factors u_i in the data production process, which culminates in data d_i. Through these procedures, that clean the environment, we can make plausible that only limited additional factors u_i are present and that they operate in a way that we understand. Creating the right data even in a laboratory that can serve as evidence for facts about phenomena is not straightforward, however. It requires elaborate planning by experimenters or data collectors to make sure that only the right background factors play a role in causing the data, that the data can be replicated to a certain extent and that the data can be reduced so that statistical control is possible (Woodward, 1989). The more control we have over the production of our data in a laboratory, the more we can influence the nature of our observations and the more reliable our measurement could be.

However, for much of the sciences, such control is in practice completely unfeasible. Boumans (2015) therefore defines field sciences as such sciences that study phenomena that cannot be studied in a laboratory. The reasons for this can be related to the nature of the phenomenon, as in the case of the social sciences, where phenomena may be too large and complex to ever be studied in isolation. Or it can be related to ethical reasons, such as in the case of the medical sciences, where not every experiment can ethically be performed and lessons need to be learned from studying the phenomena outside the experiment as well. This distinction cuts through the natural sciences, social sciences and humanities divide. Some of the natural sciences, such as ecology, may be better grouped under the field sciences, while for example, psychology as a social science to a certain extent studies phenomena in a laboratory.

3.2 Calculus of Observations

In a field science, where in contrast to a laboratory science, control of the data production is not possible, the laboratory procedures are replaced with a calculus of observations. A calculus of observations is defined through the following: 'A calculus of observations is a numerical combination of the values of the observations, such that the estimate is as accurate as possible' (Boumans, 2015, p.60). The measurement value is accurate when our estimate $\hat{P}$ is close to the true value representing the phenomenon P. So by making assumptions about the nature of our observations in relation to the phenomenon, we can make a numerical combination of the values of our observations, which supposedly bring our measurement closer to the measurand, the true value of the phenomenon (Boumans, 2015).

To be clear, standardized procedures can still be part of a calculus of observations, however, in a calculus of observations these standardized procedures will never be sufficient and a numerical combination of the values of the observations will be required. So, due to the incompleteness of the standardized procedures in the calculus of observations, we always have to make certain assumptions about the nature of the errors that remain and the relationship between the data and the phenomena. These assumptions are made in a modelling framework. Consequently, when a measurement model is constructed, the measurement errors are tamed by binding them in a certain framework through assumptions (Hacking, 1990).

To conclude, the control that is impossible in field sciences in the laboratory is replaced with methods of computation through a measurement model. Modelling a measurement can then theoretically be captured through the equation below (Boumans, 2015).

$$\hat{P} = M[(d_i; \alpha)]$$

In this model, $\hat{P}$ is the estimation of the quantitative value of the phenomenon. Through the estimation of $\hat{P}$ we make an inference from the data to the fact about the phenomenon that $\hat{P}$ is true. M is the operator in which we describe that we model $\hat{P}$. This modelling is done by combining our data d_i with model parameters α. We can compare this equation with the original data and phenomena equation. There, in the data and phenomena framework, we conceptualize the data d_i as being produced partly by the phenomenon P and partly by the variety of other factors u_i, which are idiosyncratic to the individual data. However, we are unsure about these other factors u_i, so we can never directly observe the phenomenon P. In a calculus of observations, we model the phenomenon as a function of our data d_i and additional parameters α which replaces the control in the laboratory. The additional parameters α reflect the assumptions that we make about the phenomenon and the measurement errors u_i.

Much of the theory of errors that have been put forward in historical accounts are assumptions about the nature of u_i, which is assumed to inhibit certain characteristics such that we can theoretically model them in parameters α (Boumans et al., 2015). Whenever these assumptions are reasonable can naturally be debated. In this step towards the modelling framework in a calculus of observations, a crucial assumption is made, however. This assumption will be discussed in the next section.

3.3 Invariance Assumption

The assumption that is necessary to move to a modelling framework is the assumption that the additional parameters α are invariant under the domain in which we model, and estimate our phenomenon. Invariance describes the idea that a property of a system remains unchanged under certain transformations (Neuber, 2012). These transformations are the transformations across time and space in our domain. If we assume that our parameters α are different at each time point at which we collected data, the modelling process would never be able to start. By assuming invariance in our parameters α and therefore assuming some structure in the variety of causal factors u_i, we can model the phenomenon.

As Kaila (1979, p. 131) states: 'There is knowledge only when some similarity, sameness, uniformity, analogy, in brief, some "invariance" is found and given a name. In knowledge, we are always concerned with "invariances" alone'. In constructing a measurement model we assume there to be some stability in the phenomenon P, but also in the variety of other factors u_i, which we model across space and time. This does not mean that this variety of other factors u_i have to be equal across space and time, but only that we can tame them in a certain mathematical formulation, such as a frequency distribution. This assumption of invariance makes the calculus of observations possible, consequently, since we assume that our calculus of observations remains stable across our data points d_i. This assumption is not present in our calculus of observations but is important nevertheless.

To conclude, we have seen how in field sciences, where laboratory control is unfeasible, this control is achieved through a calculus of observations. This calculus, based on the invariance assumption, enables us to model the phenomena, estimate its parameters and consequently perform a measurement. In the next chapter, we will apply this thinking to the Peilingwijzer. I argue that the Peilingwijzer can be best understood as a calculus of observations to achieve a data to phenomenon inference. The specific calculus that the Peilingwijzer employs will moreover be analysed.

4 The Peilingwijzer as an Inference from Data to Phenomena

4.1 Electoral Support as a Phenomenon

In our analysis of the Peilingwijzer as an inference from data to a phenomenon, it is important to first become clear about the nature of the phenomenon that the Peilingwijzer aims to measure. The phenomenon is described in the goal of the Peilingwijzer quoted in section 1.2, where it is described as the electoral support for Dutch political parties. The concept of electoral support is closely related to that of public opinion but has some narrowing factors.

The concept of public opinion emerged during the Enlightenment, however both concepts of public and opinion date back much longer (Price, 2008). The idea of a public dates back to at least Roman times (publicus) meaning people or the public interest. The term opinion is referred to a particular way of knowing or believing something to be contrasted with knowing something based upon facts or the truth. However, at the end of the eighteenth century, the concept evolved into something like the will of the people, the common will or the public conscience influenced by the work of Rousseau (Price, 2008). In his social contract, people bind themselves to a contract, but they are not subject to any authority except the general will, which is the first formulation of public opinion. The general will is the will towards which all citizens will come if they reason as social and not as individual people. The general will is therefore different from the summation of individual preferences since it is common to all rational, social citizens (Rousseau et al., 2002).

Our current usage of the concept of public opinion, however, follows more the liberal revisions to the concept by Bentham and Mill (Price, 2008). They stated that the only public opinion which was worthy was that of an indeterminate public opinion, which reflected all diverging interests of all individuals equally. These ideas were closely related to the liberal ideas where the preferences of all individuals were in principle equal. The general will of Rousseau could according to them not respect these equal but diverging interests. Public opinion, therefore, came to reflect the individual preferences of every member of the public, which need to be summed. Voting procedures can in this way transform these individual preferences into a public opinion (Price, 2008). Electoral support for Dutch political parties as a phenomenon is then a summation of those individual preferences of every member of the public that is eligible to vote. This eligibility criterion makes this phenomenon slightly different from the liberal conception of public opinion but follows a similar logic. The focus is on the individual preferences, which are summed to form electoral support.

From the previous discussion, it should be clear the phenomenon of electoral support can not be captured in a laboratory. It is a social phenomenon, existing in an entire country, and laboratory control is therefore unfeasible. This is of crucial importance for the rest of the discussion, where we see the Peilingwijzer, in this field science, as being an elaborate calculus of observations. This calculus of observations establishes control of the phenomenon, which is necessary since control in the laboratory is impossible.

Moreover, the description of phenomena following the account by Bogen and Woodward (1988) describes phenomena as features of the world that exist independent of human observers. In this sense, the framework paints a realist picture of the scientific process (Massimi, 2011). However, in the case of the Peilingwijzer, the phenomenon is a social phenomenon which as a concept has developed over time (Price, 2008). Therefore, it is difficult to conceptualize the phenomenon of electoral support as being found and ready-made in nature. Moreover, since it is a social phenomenon, and therefore fundamentally interlinked with the human world and human observers, the realist account of it being independent of human observers is problematic.

Fortunately, Massimi (2011), while discussing the realist aspirations of Bogen and Woodward (1988), provides us with an alternative. Phenomena, in her account, do not have to be conceptualized as existing out there and being found in nature. They can, as Massimi (2011) demonstrates, also be conceptualized in a Kantian setting. Phenomena can then be thought of as conceptualized appearances. A conceptualized appearance is an appearance that has been brought under some categories of understanding. Scientific knowledge is then gained by subsuming appearances visible to our human perception under concepts of understanding in the form of theories, concepts or models of the world (Massimi, 2011).

In a Kantian stance behind phenomena, we do not have to subscribe to a specific realist metaphysics, where we believe in phenomena existing out there. Instead, the operation of the scientific process together with principles of reason constitute phenomena as concepts through which we order reality. The aspects that constitute phenomena as stable objects in contrast to unstable data are still applicable. Only, it is now possible that in the scientific process when we learn about the phenomena, we reconceptualize it to improve our structuring of reality. Through this schema, we can also understand phenomena, that needed conceptual construction. Then, the phenomenon of electoral support does not have to be conceptualized as found in nature but it can be conceptualized as a stable concept through which we order appearances and learn about reality. For our purposes, the phenomenon in the Peilingwijzer will be conceptualized in this Kantian setting.

For electoral support to be a phenomenon in this sense, however, it should exhibit some stability, such that it can be an object of scientific theorizing and knowledge. Woodward (1989) explicitly discusses the use of surveys as a source of information about political attitudes in an electorate, which is exactly the phenomenon that we are analysing in the Peilingwijzer. The collection of data may indicate very little about features of stable attitudinal phenomena if we have doubts about the stability of the phenomenon itself. If individuals do not possess these stable preferences, the entire inference from data to phenomena may be unwarranted, since there is no phenomenon that our measurement model estimates.

The assumption that there is a stable phenomenon is further discussed in 4.6 and criticisms of this idea in 5.1. Currently, we only have to be clear that electoral support is the phenomenon in question that the Peilingwijzer is dealing with and from now on I will discuss the practical case of the phenomenon of electoral support in the Peilingwijzer.

4.2 The Peilingwijzer as a Second-Order Inference

In the case of the Peilingwijzer not one, but two, inferences from the data to facts about electoral support are made. The estimates of the Peilingwijzer are consequently the result of an elaborate calculus of observations. First, the polling stations collect the data, where the observable quantities are the responses that individuals provide to the polling stations. The procedures established to collect the data are attempts to control the idiosyncratic factors u_i influencing the data d_i. These procedures consist, for example, of standardization of the data collection process to make sure that the variety of other factors influencing the data is as low as possible (IO-Research, n.d.). Nevertheless, these procedures are insufficient since the phenomenon is present not in the laboratory but in the field. Hence, the necessity for a modelling framework in the calculus of observations. From the collected data, they make an inference towards Dutch electoral support using a sample-to-population inference. In this inference, assumptions about the nature of the variety of causes, u_i are already present and assumed to be of a certain degree such that it can be modelled in parameters α.

The assumption behind this sample-to-population inference is that the sample is random. For a sample to be random every member of the population has to have an equal chance to be included in the sample. In practice, this is unlikely, however. Therefore, the polling stations make use of population categories and weight factors. Observations are then given a weight in the sample using a weight factor which is based upon relevant characteristics. If we, for example, find it important that our observations are equal across Dutch provinces, then the polling station weighs those observations out of provinces which were underrepresented higher than the observations from overrepresented provinces. The weight is taken as the proportion of over or under-representation. The idea is that this creates an equal weight in the population estimate (IO-Research, n.d.). However, we need to assume that we have established the correct categories to categorize the population and that we have consequently weighted the sample correctly. In practice, there will be an almost infinite amount of categories, so the assumption comes down to the idea that we have not missed a category which we in our measurement deem important. This has to be judged by the experts in the domain of measurement.

So, in this inference from data to electoral support, assumptions about the error distributions in our measurements and the representativeness of our observations for the general population are necessary to make the data to electoral support inference possible. Since, this inference is performed first, relating observables to unobservables, I propose to call this a First-Order Inference. The Peilingwzijzer takes the inferences from the polling stations as data in its measurement model (Louwerse, 2011). From this data again a modelling framework is constructed in the calculus of observations, which infers from the data provided by the polling stations, a fact about the phenomenon. I propose to call the Peilingwijzer a Second-Order Inference since it takes as inputs already the outcomes of a data to phenomenon inference instead of primarily relating observed data to electoral support.

Two separate inferences are consequently employed in this calculus of observations, one done by the different polling stations between their observations and their population estimates, and one done by the Peilingwijzer between the inputs provided by the polling stations and the overall measurement outcome. This reflects the separation between the First-order and Second-order Inference, as highlighted above.

4.3 The Statistical model of the Peilingwijzer

So, The Peilingwijzer as a statistical model is part of a calculus of observations. By numerically combining the values of the observations, the measurement can be as accurate as possible if this is done successfully. The statistical model is displayed below (Louwerse, 2011).

$$P_i \sim \mathcal{N}(M_d, F)$$

$$M_d = A_d + H_{bi}$$

$$A_d \sim \mathcal{N}(A_{d-1}, \tau)$$

In this measurement model, the inference from data to measurements of electoral support is performed by stating invariant relations between the data d_i and additional parameters α. The three equations demonstrate different assumptions about our measurement errors and the phenomenon of electoral support and will in turn be discussed in 4.4, 4.5 and 4.6. The definition of all the parameters is presented in table 1.

Parameter	Definition
P_i	The percentage of electoral support for a political party in a certain poll i
M_d	The average of the normal distribution $\mathcal{N}$ on day d
F	The error margin in the poll
A_d	The 'real' percentage of electoral support for a political party in the population on a certain day d
H_{bi}	The house effect of polling station b in poll i
A_{d-1}	The 'real' percentage of electoral support for a political party in the population on the day before $d-1$
τ	The random walk prior (the error margin)

Table 1: Legend of the Peilingwijzer

All the parameters in this model are estimated using Bayesian Modelling, specifically the Markov Chain Monte Carlo (MCMC) method (Louwerse, 2011). This method relies upon the following general principle: 'Anything we want to know about a random variable x, we can learn by sampling from g(x), the probability density function of x' (Jackman, 2000, p. 307). MCMC methods seek to characterize the posterior distribution of the unknown parameters.

They do this not by optimizing the parameter, but by sampling through an iterative process (Jackman, 2000b). In this way, the MCMC method searches through the space of possibilities for each of the parameters and produces reports of these searches as samples of the density of the posterior distribution of the parameters. Summary descriptions of these samples are reported for further inference and description.

In the Peilingwijzer first, there are 10.000 burn-in iterations. Burn-in iterations are the first iterations which are performed. The start of the iterative process is random and these burn-in iterations are dependent on the starting point. Consequently, the burn-in iterations in the MCMC method are done and then thrown away to alleviate the dependency on the starting point Then, the MCMC method is run for 200.000 iterations. However, there may be a relation between subsequent iterations, which we would like to remove from our final output. Therefore, a thinning factor of 80 is used, meaning that one out of every 80 iterations is taken to create an a posteriori sample distribution of 2500 (Louwerse, 2011). This means that for every parameter 2500 estimates are produced and summary statistics of these 2500 estimates are used as outputs for the model.

The output of the Peilingwijzer as seen in figure 1 follows this posterior distribution of the MCMC method. The mode of this distribution is reported as the most likely outcome for this variable. Moreover, 95 percent credible intervals are created by looking at the posterior distribution as the output of the model (Louwerse, 2011). However, it can not be stressed enough that the validity of these outcomes has to follow the validity of the measurement model. The Monte Carlo method can give outcomes consistent with the measurement model. If our measurement model is unreliable, so will the outputs produced by the Monte Carlo Method (Elishakoff, 2003). To answer the question under which conditions the Peilingwijzer is reliable, therefore, we need to investigate the specific assumptions that are made to tame the noise in the data u_i and which makes the inference from the data to facts about electoral support possible.

4.4 The Normality Assumption

The first assumption is presented in the first equation of the Peilingwijzer:

$$P_i \sim \mathcal{N}(M_d, F)$$

The equation represents the idea that the electoral support for a political party in a poll P_i is drawn from a normal distribution. P_i is not estimated in the model but is data coming from the polling stations. P_i represents the estimated electoral support for a single political party in a certain poll i. The average of this normal distribution on a certain day d is represented by M_d. The standard deviation of this distribution is represented by F. F is also not estimated in the measurement model but is the reported error margin for that specific political party in a certain poll i. The normal distribution describes a family of continuous probability distributions, which have the same shape but differ depending on their location and parameters of scale (Ahsanullah et al., 2014). The entire distribution is then defined through the parameters of the mean and the standard deviation.

The equations in the normality assumption together with the structural bias assumption reflect the difference between random and systemic errors, which is used to tame the uncertainty about the idiosyncratic factors u_i (Hacking, 1990). The distinction between random and systemic errors is a conceptual innovation in the theory of errors due to Bernoulli (Sheynin, 1979), where he distinguished between 'abberationes chronicae' (systemic or chronic errors) and 'abberationes momentaneae' (random errors) (Fischer, 2011). Modelling the poll outcomes as being drawn from a normal distribution attempts to tame the random error present in our measurements. In this section, we will attempt to understand how this is done and which assumptions are made in this process.

As Stahl (2006) explains, the development of the normal distribution was guided by the search for an error curve where measurement errors could be modelled. Across measurements, differences are always found, and the question arose how these different observations can be converted into a single number, the data representative. Discussions were had on whether the median, the average or possibly another method would provide the best answer. The first scientist to provide a systematic treatment of the random errors in measurement was Galileo in his treatment of errors in observations in astronomy (Stahl, 2006). He outlined the following principles, which are adapted and improved upon, but not completely rejected, in later developments. First, he stated that we should believe that there is only one number, which reflects the true value. Then, we assume that all observations do inhibit some error and that small errors are more likely than large errors. Finally, it is assumed that the observations are distributed symmetrically around the true value. This means that the errors are distributed symmetrically around zero.

However, this was not yet the idea of a probability distribution. Des Moivre in 1708 obtained the normal distribution as a distribution that peaks around the mean. He obtained this by studying the limit of a coin toss or a binomial distribution (Hacking, 1990). Each experiment in this sense consists of doing a certain amount of coin tosses n. We can then count how many times out of n in each experiment we get head. The distribution of the number of occasions that we get head will approach this normal distribution when we increase n (Hacking, 1990). Then, in 1809, Gauss published his work on the method of least squares estimation (Sowey and Petocz, 2017). The normal distribution in this work was proposed as a theoretical model. This theoretical distribution could function as a model for the probability distribution of real-world random measurement errors. In this distribution, the mean functions as the most probable estimation of the true value. The assumptions that were made to create the normal distribution were the following (Stahl, 2016, p.104):

> ' 1. Small errors are more likely than large errors 2. For any real number ϵ the likelihood of errors of magnitudesϵ and ϵ are equal 3. In the presence of several measurements of the same quantity; the most likely value of the quantity being measured is their average'

If we use the normal distribution in this sense, however, there is a danger of circularity here. Gauss had deduced his error law or distribution from the principle that the arithmetic mean is the most probable estimation, but this is an assumption which also describes a fact, which we attempt to demonstrate. The question that arises is whether these random errors in observations actually obeyed a normal distribution or that it was a convenient hypothesis, allowing for easier control and calculation (Fischer, 2011).

Särndal (1971) explains that the explicit theoretical formulation of the assumptions behind the normal distribution, which adresses the circularity present in Gauss, comes a while later in the work of Hagen. In 1837 Hagen, in his *Grundzüge der Wahrscheinlichkeits-Rechnung*, formulated what he named the Hypothesis of Elementary Error (Hagen, 1837). The assumption by Hagen was the following (Fischer, 2011, p. 81):

> ' ... the error in the result of any measurement is the algebraic sum of an infinitely large number of elementary errors ["elementäre Fehler"], which are all equally large, and of which each single one can be just as positive as negatives'

So Hagen states that the observational error is the summation of an almost infinite number of elementary errors, which all have the same value and the same probability of being positive or negative (Särndal 1971). So, if we can assume that individual errors are caused by effects which are independent, equal, and large in number, then the error curve of the sum of several observations will follow a normal distribution. For full understanding, causes being independent means that they are unable to affect each other. Knowledge about one cause, will not help us in predicting the outcome of another.

The idea of elementary error or "elementäre Fehler" was closely related to a specific model of binomial probability. These elementary errors or random causes can be conceptualized through a black-and-white urn model. Here, black and white balls are drawn from an urn with replacement. The black and white balls concern a positive and negative value of the elementary error, and in this situation, the error sums up to be a large number of elementary errors (Fischer, 2011). This highlights that in Hagen's initial hypothesis the elementary error can only take two values α en - α with equal probabilities (Särndal 1971). Hagen proved that based on these assumptions, the normal distribution could be derived.

The normal distribution arose to deal with random errors in observation but was later due to the work of Quetelet also used to describe individual differences in objects belonging to the same group (Hacking, 1990). So, to describe, for example, the distribution of height across a population. In our case, the old usage of the normal distribution as dealing with measurement errors is of importance. The Hypotheses of Elementary Errors is beautifully summarized by Galton (Stahl, 2016, p.111). Interesting here is that the conceptual innovation describing that the hypothesis can also be applied to individual differences in a population is mentioned, which was not present in the original formulation of Hagen.

> ' It will be remembered that these are to the effect that individual errors of observation, or individual differences in objects belonging to the same generic group, are entirely due to the aggregate action of variable influences in different combinations, and that these influences must be (1) all independent in their effects, (2) all equal, (3) all admitting of being treated as simple alternatives "above average" or "below average; (4) the usual Tables are calculated on the further supposition that the variable influences are infinitely numerous'

The question however is still in which situations this assumption can be warranted. In 1870 Crofton, an Irish mathematician, in his *On the proof of the law of errors of observations* provides a strong argument in defence of this assumption. Moreover, he describes the situation in which it can be credible to assume it. He provides the argument for the general version of the Hypothesis of Errors, in which the errors are formed through a summation of a large number of smaller errors, but in which the errors themselves do not have to be equal as in Hagen's formulation (Crofton, 1870). Crofton (1870) describes the case of Astronomy as the field in which the reasonableness of this assumption is taken to be most credible since he describes Astronomy as the science of observation, where we have learnt the most from experience. In Astronomy, in earlier times, since observations were still quite coarse, errors were caused by a small number of principal causes all having a large impact. Examples of these were observations with only the naked eye and the cause of refraction in observations. Crofton (1870) describes that through innovations and understanding these causes were found and attenuated or accounted for.

In these previous times, the Hypothesis of Errors was quite unlikely, and we could not expect the errors to be tamed through a normal distribution. However, when these sources of the principal causes of errors were taken into account, astronomers searched for remaining sources of errors. They found not again a few numbers of errors, but a great number of sources of error, all being small in impact. It is beautifully described by Crofton (1870) as the idea that it was as if a small number of forest trees had been cut down, but leaving an innumerable growth of shrubs and brushwood at their feet remaining to be cleared. This working is also described for other cases, such as artillery and machinery, where if the few great sources of imperfection are understood and remedied, then the number of minor sources of error increases rapidly. He moreover also compares this to the working of diseases in causing the deaths of humans. The more diseases as a cause of great sources of death for humans are remedied, the more the causes of death for humans will be smaller and smaller, with each having a fewer impact on the whole (Crofton, 1870).

In situations therefore where we can believe that the great sources of error are remedied, then we can understand the sources of error more and more as consisting of a small number of errors with a small impact on the whole. If this is the case then the Hypothesis of Errors is credible and we can describe our errors through a normal distribution. Currently, we call this requirement the requirement for identically and independently distributed causes (iid). Personally, I prefer the terminology of the Hypotheses of Errors, since it immediately brings to the foreground that that is what the assumption is all about.

The Peilingwijzer therefore, by describing the random errors through a normal distribution, has assumed the Hypothesis of Errors. Only now, the Peilingwijzer does not possess a large number of observations but simulates these extra hypothetical measurements by assuming that the result of our polling station is drawn from this normal distribution. The question of whether this is warranted comes down to the question of whether we can assume the Hypothesis of Errors to hold. The normality assumption consequently tames the random errors present by assuming something about the nature of these random errors in u_i. That is, they individually consist of a large number of errors, with all having a small impact on the whole. This calculus of observations allows the Peilingwijzer to supplement control of these errors.

When critically analysing the argument by Crofton (1870) however, there is a threat of a circular argument. We assume our calculus of observations to be more accurate since we have made an assumption about the nature of the individual errors. This Hypothesis of Errors however as Crofton (1870) explains is based upon the idea that our measurements already are quite accurate and that we have removed the principal sources of errors for every individual observation, which leaves as a remainder a large number of errors all being small in impact. But with our measurements, we attempt to make credible the idea that they are reasonably accurate. It seems like we are assuming, at the start of the process, that which we exactly attempt to demonstrate at the end.

Currently, we have to contend ourselves with the understanding that the Hypothesis of Errors is the assumption which is present in this equation and that this assumption is the foundation of this part of the calculus of observations. And this calculus of observation, as we remember, allows the inference from the data d_i to facts about electoral support in the presence of these unknowable factors u_i by modelling it as a function of the data d_i and additional parameters α. In section 6.2 it will be further discussed how the Peilingwijzer could still make progress even in the threat of this circularity.

4.5 The Structural Bias Assumption

The second assumption in the Peilingwijzer is displayed below:

$$M_d = A_d + H_{bi}$$

This equation describes that the estimated average of the normal distribution from the first equation M_d consists of two parts. The first part is the actual percentage of electoral support for a political party on a certain day d and is represented by A_d. The second part is the house effect of a certain polling station b for a specific poll i. This variable is represented by H_{bi}.

This equation reflects the idea that apart from random errors (abberationes momentaneae), systemic errors (abberationes chronicae) are also present. While random errors are allowed to fluctuate across measurements, the systemic errors are conceptualized as those errors which are stable across measurements (Fischer, 2011). Gauss indicated that he would only concentrate on what he called the irregulares seu fortuiti (random errors) and not deal with the constantes seu regulares (systemic errors) (Sheynin, 1979). The normal distribution, therefore, attempts to tame the random errors only. Systemic errors are in theory easier to conceptualize and remedy than random errors. That is, since if we have found that there is a systemic error in our observations with a certain amount, then we only have to apply that constant negatively over all our observations, and we have remedied the systemic error.

Louwerse (2016) describes that even as the random error can be assumed to have an expected value of zero over a large number of samples, a systemic error can still be present. The structural error is conceptualized as the structural bias of a polling station H_b. The idea is that it may be the case that a certain polling station consistently over or underestimates a certain political party. If we can estimate this structural bias in our model, we can adjust for it and estimate the unbiased estimate A_d.

The assumption that there is something like a structural error in the Peilingwijzer is made credible by the empirical fact that every polling station has standardized procedures for collecting data, which all differ slightly from each other. These standardized procedures are equal across measurements and therefore a structural error that is present in these procedures can be present across time. So, the methodological choices made by a polling station are assumed to produce a systemic bias or error, which can be estimated through the variable of house effects H_b in the model (Fisher et al., 2011).

The crucial assumption which is implicit in the equation is that as a whole the house effects sum up to zero. As is stated in the Peilingwijzer (Louwerse, 2011): The model assumes that the house effects sum up to zero: the average polling station 'is correct', is the assumption [3]. This assumption is crucial since it allows the information coming from different polling stations to become comparable. We can pool different estimates of the same political party if we assume that we have removed the systemic error from the observations (Jackson, 2005). And in estimating the systemic error, we must assume something about the nature of the systemic error. In the Peilingwijzer that assumption is that the average polling station will possess no systemic error. In this way, we can estimate how much each polling station deviates from this standard and estimate the structural errors as house effects H_b.

This assumption is hard to empirically sustain, however, since presuming that we have knowledge of the systemic error across polling stations already presumes that we know the population quantity that is estimated (Jackson, 2005). After an election, however, we could possibly see whether this assumption is warranted by comparing the average error with the election result to see whether there is any structural bias across the polling stations (Jackson, 2005). This strategy of calibration, where our measurements are compared with a standard, which in this case is the election result, could allow this assumption to become more or less credible.

Even in this learning of past elections, there is no guarantee that the same approach will work again. If we have calibrated the structural bias across the polling stations against the elections, this could also change. This argument of induction, where we learn from the past, presumes a uniformity of which we are uncertain. To be clear, the Peilingwijzer is not calibrated against the election results. However, a comparison of the estimated house effects with the previous election results can inform us whether this assumption is more or less credible. The Peilingwijzer in this way only demonstrates the differences between different polling stations (Louwerse, 2016).

The Normality Assumption with as its foundation the Hypotheses of Errors and the Structural Bias assumption create fixed points A_d out of which the evolution of electoral support over time is graduated. As Chang (2004) in his work on the measurement of temperature states, fixed points form the justificatory standards in our measurement scheme. This does not mean that those fixed points are fixed with absolute certainty, however, it means that

[3] The original quote is in Dutch. This quote is a translation by the author

we are more certain about their validity or their link to the empirical world than the rest of the measurement that we are attempting to perform. As we will see in the next section, the Peilingwijzer, through the stable phenomenon assumption, deviates in a small but significant way from the above description. The fixed points do not specifically determine the evolution of the electoral support across time, but merely provide guiding points that guide the direction of the phenomenon.

4.6 The Autonomous Phenomenon Assumption

The third assumption in the Peilingwijzer is displayed below:

$$A_d \sim \mathcal{N}(A_{d-1}, \tau)$$

Lastly, this equation states that the real amount of electoral support on a certain day d is drawn from a normal distribution. This variable is represented by A_d. The mean of this normal distribution is represented by the real percentage of electoral support on the day before A_{d-1}. The standard deviation of this normal distribution is represented by τ. So, the crucial assumption in guiding the flow of our quantitative statements about the phenomenon is that the measurement of a certain day is only dependent on the measurement of the day before, but not on any history before that.

This assumption reflects the idea that the phenomenon of electoral support possesses autonomous behaviour. Irrespective of our measurements, electoral support has a relationship to the electoral support of the day before which is captured in a normal distribution, which describes a statistical law. Describing this relationship through a normal distribution indicates that most of the steps will be small and close to the average of the day before A_{d-1}. Large deviations become less and less likely as the size of the step grows. This assumption consequently reflects the idea that major shocks to the system are quite unlikely. Of importance here is that the normal distribution does not reflect assumptions about the measurement errors but reflects assumptions about the phenomenon itself.

Hacking (1990) describes the conceptual innovations by Quetelet as crucial in this regard. Before Quetelet, the normal distribution was a theory of error which only described deviation from true values. It did not describe or capture objects in the world but only the imperfection in our measurements (Porter, 1985). Quetelet applied the same normal distribution to natural variation and as Porter (1985, p.67) states: 'The error curve became for the first time an attribute of nature itself'.

Quetelet consequently makes the conceptual step to equate error and variation. Because, he argued, measuring the height of an individual a large number of times or measuring the height of a large number of individuals from a homogenous population once amounts to the same statistical techniques (Hacking, 1990). As Hacking (1990) argues, this move is fundamental, since we move from a description of measurement errors to a description of phenomena in nature which can be described by statistical laws.

So in this assumption of the Peilingwijzer, there is the phenomenon of electoral support to explain and describe and this is done through a crucial statistical assumption, namely the assumption that the behaviour of the phenomenon across time will follow a normal distribution. That is, not the entire phenomenon follows a normal distribution, but the differential curve between days (the curve of differences) is assumed to follow this normal distribution. The conceptual innovation necessary for this assumption is that phenomena can be described and explained through an autonomous statistical law. The normal distribution in this assumption as a description of natural variation of electoral support is crucially conceptually quite different from the normal distribution as a description of measurement errors in the normality assumption.

The description of electoral support through an autonomous statistical law is reflective of the idea that a phenomenon should have stable characteristics, as outlined in the framework of Bogen and Woodward (1988). Across the time and space of the measurements, the phenomenon is assumed to possess stability. The stability of the normal distribution indicates that the average difference between days will be the difference which is assumed to occur the most often. Deviations from this average, positive and negative, will be fewer and fewer present. The stable phenomenon assumption has important implications, moreover. Because of the description of electoral support through an autonomous statistical law, the results obtained from the normality and structural bias assumption are not taken simply as our best information, but merely as the direction towards which the phenomenon should direct itself. The evolution of electoral support is therefore not interpolated across the measurement points obtained through the normality and the structural bias assumptions, but these measurement points are merely used as guiding points. This assumption is used to protect the Peilingwijzer against the possibility that a result from the polling station is an outlier (Louwerse, 2011).

The Peilingwijzer consequently assumes that radical shocks of the system, where the electoral support for a political party jumps from one position to another, are not possible by stating that it follows an autonomous statistical law. This is not an assumption about the measurement errors as in the normality and structural bias assumption but an assumption about the phenomenon itself. This assumption allows the Peilingwijzer to determine the electoral support on days on which no results are reported without having to result in interpolation. The question however is whether this assumption is warranted and empirically credible. Because again there is a threat of circularity here. We need the Peilingwijzer to determine whether the phenomenon of electoral support can be described through this distribution but this is already an assumption in the Peilingwijzer itself.

To conclude, we have seen that the Peilingwijzer is part of a calculus of observations. This calculus of observations is created and performed to make an inference from the data d_i to the phenomenon P in the presence of unknowable local causes u_i possible. Due to the normality assumption which tames the random errors, the structural bias assumption which tames the structural errors and the stable phenomenon assumption, which binds the phenomenon in an autonomous statistical law, the inference is performed and the measurement is done. We have seen hopefully how many conceptual steps are necessary to make this inference possible, and that measurements of this kind are not an easy feat. The additional assumptions are a necessary feat of human ingenuity and judgment. However, this necessity for human judgment may also threaten the objectivity and foundation of the measurement procedure. This aspect is reflected on in part 6.2, and the question is asked how measurements could still be reliable in the presence of these assumptions.

5 The Peilingwijzer as a Reliable Measurement?

5.1 Critical Judgment of Assumptions

In this project, we have seen that the Peilingwijzer can be conceptualized as a calculus of observations. This calculus of observations is performed to make an inference from the data to a measurement of electoral support. Crucial in this calculus are the assumptions discussed in 4.4, 4.5 and 4.6. As a short recap, the normality assumption to deal with random errors in section 4.4 is based upon the Hypothesis of Errors, which states that individual measurements are caused by a large number of small independent causes. And in the argument by Crofton (1870) we have seen that in situations where we can believe that the large sources of errors are eradicated, this Hypothesis may be reasonable. Moreover, in the structural bias assumption to tame structural errors, the crucial assumption is that the systemic errors across polling stations amount to zero. This assumption allows the information from different polling stations to be pooled. Lastly, the autonomous phenomenon assumption, assumes that electoral support as a phenomenon can be described by an autonomous statistical law, where the average difference between days will occur most and deviations from this average will occur less and less.

To allow for a further critical discussion of the Peilingwijzer and to understand under what conditions the Peilingwijzer is a reliable measurement, this understanding is already paramount, I argue. This clarity is hopefully already gained. Nevertheless, in this section, I will attempt a critical discussion of the strengths and weaknesses of the assumptions in the Peilingwijzer. This section is more tentative than previous sections, however. Still, to keep making progress critical judgment is necessary and this section is an attempt at exactly that. Then, in section 5.2 possibilities for improvement are suggested.

First, the assumption that electoral support is such a stable phenomenon as described by Bogen and Woodward (1988) and Massimi (2011) can be questioned. As Woodward (1989) himself outlined in some areas of investigation we can collect plenty of data, but fail to identify a clear phenomenon. In the case of the Peilingwijzer, it is assumed that there is such a thing as stable individual preferences which culminate in electoral support. However, as Grant and Patterson (1975) outline, the problem of non-attitudes has been the focus of a debate in the political sciences.

The problem is that there may be individuals without stable preferences, conceptualized as non-attitudes. These individuals can have attitudes which change dramatically across a short period of time or inhibit great variation across a short change in the measuring instrument, for example in the question asked. Converse (1970) outlines that there is reason to believe that there are two groups in the population. One group with stable preferences, which displays considerable stability, and another group with considerable instability or non-preferences. Only the former can be considered a true phenomenon, as Woodward (1990) states. The question is how large the latter group is to consider the Peilingwijzer a true data to phenomenon inference.

Moreover, even when we can assume this, the assumption that we can describe the evolution of electoral support by an autonomous statistical law is quite strong. Electoral support is here modelled as a natural science phenomenon, where measurement results can not influence its behaviour. However, as we have seen in the bandwagon effect, the measurement results of the Peilingwijzer may influence the evolution of electoral support itself. If this is the case, then the assumption of an autonomous statistical law can be too strong. The Peilingwijzer itself can be the shock to the system, which is assumed to not be possible to occur. As Marsch (1985, p.51) states:

> 'Processes of this kind are of theoretical interest because they affect the possibility of stable prediction in the social sciences; if the very act of predicting that one party will win an election can be a self-fulfilling prophecy then the natural scientific model of the social sciences may be compromised'

Then, the Hypothesis of Errors, which is the foundation of the normal distribution, is guided by the idea that our measurements are already quite accurate such that the remaining errors are large in number and small in size. Doubts raised by the previous considerations may consequently inform us of the validity of this assumption and judgment. If we have doubts about the stability of our phenomenon as outlined above, we can question whether our measurements are already quite accurate. Consequently, the assumption that the random errors are normally distributed can be problematic.

The description of structural errors could help provide credibility for this assumption, however. If the Structural Bias assumption is able to correctly filter the structural and therefore large errors out of the calculus of observations, the remaining random errors could be normally distributed. The assumption that the structural errors across disciplines amount to zero is problematic, however. In the work of Kahneman et al. (1982) discussing the nature of judgment under uncertainty, which is exactly what we are concerning ourselves with, structural biases are identified. These structural biases are most likely also at work in the practices of the polling stations. Consequently, it is, I argue, unlikely that there is no structural bias across the polling stations. If structural biases across the polling stations would be present, then the assumption that the structural error will be zero across polling stations would also be problematic.

Nevertheless, even regarding the critical analysis of the assumptions in the Peilingwijzer, we can see the Peilingwijzer as a step in a process of improvement, I argue. The aim of this work is to make further improvements, guided by a critical analysis of its assumptions, possible. The Peilingwijzer in a sense was also an improvement upon the isolated results by the different polling stations by making explicit that a calculus of observations based upon specific assumptions was needed to measure such a complex phenomenon as electoral support present in a field science. Whether the Peilingwijzer is a more accurate measurement is difficult to determine since it is clear, I argue, that most of the assumptions are to a certain extent problematic and do not hold fully in practice. But we can see it as the first stage in an epistemic process in which we improve and enrich our measurement models. So, in the next section, some suggestions for improvement are made.

5.2 Possibilities for Improvement

In this section, possibilities for improvement are suggested. I do not claim to have all the answers but provide merely guiding points for research and conceptual development. Firstly, I argue, that it is crucial that the stability of the individual preferences in a population is further researched to see for which individuals a stable phenomenon exists. An example of this would be to follow individuals for a longer time and to track to what extent their preferences are stable. This research could then be incorporated into the Peilingwijzer to create a division between those preferences that are stable and those that we should put less trust in.

Moreover, research is also needed in the distribution of polling outcomes. What curve do the results from polling stations represent when multiple polls are performed at the same moment? This research could also study this using a large sample and a resampling method, which artificially creates samples. Studying the distribution of polling outcomes here could inform the Peilingwijzer what the distribution of random errors could be. Another point is to calibrate the industry bias with previous elections. We could study for every political party how much the industry has erred in previous elections and adjust the assumed structural error to be higher or lower compared to these outcomes. This could allow for more accurate structural errors instead of assuming that they cancel out.

Moreover, the autonomous statistical law could also be improved by allowing for shocks in the evolution of electoral support. An extra parameter could be incorporated which represents this shock and allows for more or less deviation across time periods. The question of whether it is likely that such a shock occurs is a question of human judgment that should be debated by experts in the field. The day of the outcome of the Peilingwijzer itself could in this way also receive a higher shock factor, allowing the effect of the measurement model itself to be incorporated. These shocks are a reflection of the idea that the normal distribution may also be unfitting for this phenomenon.

Lastly, I argue, that triangulation would also improve the measurement process. Triangulation can be seen as the idea that multiple methods or models are used to describe or measure the same phenomenon. If they agree with each other, then this provides evidence that the models are more accurate. In the case of the Peilingwijzer, we could create multiple measurement models, all with slightly different assumptions, and investigate whether these differences allow for radically different outcomes. In moments where the models are agreeing we have reason to believe that the measurements are accurate, however, when large disagreement is found, there is a signal that our measurements are less accurate. In this way, we can assess to what extent we should trust the outcomes of the measurement models.

These suggestions show that the Peilingwijzer should not be the final measurement model but should be seen as a stage in the measurement process. Moreover, since assumptions are crucial in a data to phenomenon inference in a field science, rational and critical debate about these assumptions across experts is necessary to keep facilitating improvements in the measurement model. Hopefully, the above suggestions provide a first step in fostering this debate.

6 Conclusions and Reflections

6.1 Conclusions

This project started with the following research question: *under which conditions does the Peilingwijzer work as a reliable measuring device?*. To answer this question, the framework of data and phenomena was introduced. Data d_i are caused by a multitude of local causes u_i, idiosyncratic to the specific data production process, which can also be conceptualized as the measurement errors, and the phenomenon itself. In this way, the data can in principle be used as evidence for facts about phenomena. This is done by making an inference from the data to facts about the phenomena. Measurements then are performed by making such inferences towards quantitative statements about phenomena.

In a laboratory, this is done by cleaning the environment, creating as much control of u_i as possible. In a field science, where full control is infeasible, this is done by creating a calculus of observations. This calculus of observations is obtained by modelling the phenomena as a function of the data d_i and additional parameters α, which reflect assumptions about the measurement errors and the phenomenon. To use this modelling framework, however, it needs to be assumed that this function is invariant across the domain of our measurement. The Peilingwijzer should then, I argue, be seen as an inference from data to the phenomenon of electoral support. Electoral support is conceptualized as the summation of individual preferences of those individuals that are eligible to vote. The Peilingwijzer is consequently a calculus of observations. Consequently, the statistical model of the Peilingwijzer reflects assumptions on the measurement errors and the phenomenon to establish this calculus observations.

The normality assumption tames the random errors and relies at its core on the Hypothesis of Errors, which is based upon the idea that our measurement errors are large in number and small in size. Moreover, the structural bias assumptions allow for the information from different polling stations to be pooled by assuming that the structural error of the industry amounts to zero. Lastly, the stable phenomenon assumption describes the evolution of electoral support through an autonomous statistical law, where the theory of errors is transformed into a description of nature itself. Differences between days are modelled through a normal distribution, assuming that shocks to the system are more unlikely than the average difference.

At this part, we have an answer to the research question. The Peilingwijzer is a reliable measurement when the assumptions as explicated above hold. However, many of the assumptions are at least questionable, as is discussed in section 5.1. Consequently, this warrants doubt about the reliability of the Peilingwijzer as a whole. Nevertheless, doubts about some of its reliability does not make the Peilingwijzer immediately useless. We could put more trust in its outcomes in situations where we have more confidence in the reasonability of its assumptions and be more critical at moments where we deem the assumptions to be fundamentally wrong. Having a clear and deep understanding of the assumptions present in the Peilingwijzer is paramount for this and hopefully this project has provided that. Being clear about its assumptions can consequently also guide us in its use.

6.2 Reflections

After these conclusions, the question can be asked how improvements in a field science are ever possible since assumptions will possibly always to a certain extent be problematic. If assumptions are always necessary how do we progress? Because the problem is clear. We have seen in the assumptions discussed that at their core these assumptions have to be made credible from knowledge about the measurements and the phenomenon that is studied. However, to obtain this knowledge measurements need to be credible and reliable, which is based on the assumptions made. As Chang (2004, p. 221) states:

> 'The basic problem is clear: empirical science requires observations based on theories, but empiricist philosophy demands that those theories should be justified by observations. And it is in the context of quantitative measurement, where the justification needs to be made most precisely, that the problem of circularity emerges with utmost and unequivocal clarity'

To make measurements in the Peilingwijzer we need assumptions based upon human judgment. However, to base the validity of our human judgments, we need measurements. This is the problem of circularity that Chang describes. In this section, I will outline what in principle is a process that scientists or the scientific community can take to remedy this problem.

As Chang (2004) argues, the circularity of justification can be remedied through a self-improving process that can be best seen as a spiral upward. This process, which is described as a process of epistemic iteration, is the idea that successive stages of knowledge building are performed, where each stage builds on the previous stage. At each step, justification is found in a combination of the measurements at previous stages, theory and concept building warranted by those measurements and the assumptions that accompany them. Moreover, the idea is that each step is somewhat less perfect than the previous, however, a clear foundation or a perfect description can not be expected. The idea of iterations allows us to understand how progress may be possible without having a clear foundation.

Crucial aspects of this idea of epistemic iteration are that successive steps display features of enrichment and self-correction. Enrichment is a process in which the previous step is not criticized, but refined, making it more nuanced and complex. The model in the next step is extended in, for example, the number of variables, scope, or complexity of the assumptions involved. Self-correction on the other hand is the process where each successive critically analyses and alters aspects of the previous step (Chang, 2004). In this aspect of iterative progress, assumptions that have been found to be less credible or unwarranted by observational results are altered. So, by making small iterative improvements, we can understand how progress is in principle possible. However, it has to be stressed, that an ultimate foundation is not to be found. There is always a chance that we at some point will return to the first step and start over. Nevertheless, through this process of epistemic iteration, progress can be made.

After a measurement, it is possible for the measurement outcomes to be critically discussed and hereafter inform theory and concept building. This theory can in turn inform the assumptions that we have made in the previous measurement stage, and we can enrich and self-correct these assumptions. Due to these modifications, new measurements can be made, which are hopefully more precise and accurate. And the epistemic iterative process continues. The increasing precision in this progressive cycle is due to the increase of complexity in the model because of enrichment. Moreover, we hope that the model becomes more and more accurate at later stages and becomes closer and closer to the true value that we attempt to measure. However, I argue, that this is dependent on the starting point of the epistemic process and the successive steps. If due to measurement or information from other disciplines or fields, we lose trust in our first most basic assumptions, we must conclude that, even though we believed the model to be quite precise and accurate, this was actually not the case.

The assumptions are however often hidden in the quantification of the assumptions in the measurement model. As Porter (1995) states, quantification is a technology of distance, where quantitation has the implicit trust of making knowledge independent of the people who make it. In the Peilingwijzer the assumptions are construed in mathematical equations and therefore create the illusion that no assumptions were necessary for its construction. We have seen however that this is not the case but that the Peilingwijzer as a calculus of observations explicitly relies on these necessary assumptions. This illusion could make the process of epistemic iteration difficult, I argue, by preventing rational debate about the assumptions involved. We should not deal with a lack of trust in human judgment by hiding the assumptions which will be always present in the measurement model in quantitative statements.

This does not mean that quantification always allows for a more shallow debate. Quantification can be a crucial tool in the scientist's toolbox. However, I argue, that problems appear, especially in field sciences, if quantification is seen as a value-neutral procedure without assumptions. As we have seen in the Peilingwijzer the quantitative equations are based upon fundamental assumptions about the nature of us as observers, our measurements and the phenomenon that we are studying. Reflection on these assumptions consequently requires trained expert judgment. However, a quantitive mentality does not have to exclude expert judgment. It can also foster critical thinking and debate. Moreover, if we are clear about the assumptions involved, then quantification can also be a way in which we make these assumptions precise, as is done in the Peilingwijzer (Levy, 2001).

7 Bibliography

Adams, William J. 2009. *The life and times of the central limit theorem.* Providence, R.I: American Mathematical Society

Ahsanullah, Mohammad, BM Kibria, and Mohammad Shakil. 2014. "Normal distribution." In *Normal and Student st Distributions and Their Applications*, 7-50. Springer.

Andersen, Robert. 2000. "Reporting public opinion polls: The media and the 1997 Canadian election." *International Journal of Public Opinion Research* 12 (3).

Bogen, James, and James Woodward. 1988. "Saving the phenomena." *The philosophical review*, 97 (3): 303-352.

Boumans, Marcel. 2012. "Modeling strategies for measuring phenomena in-and outside the laboratory." In *EPSA Philosophy of Science: Amsterdam*, 2009, 1-11. Springer.

- - -. 2015. *Science Outside the Laboratory : Measurement in Field Science and Economics* . Cary, United States: Oxford University Press USA - OSO.

Boumans, Marcel, Giora Hon, and Arthur C Petersen. 2015. *Error and uncertainty in scientific practice.* Routledge.

Chang, Hasok. 2004. *Inventing temperature: measurement and scientific progress* . Oxford studies in philosophy of science. Oxford ; Oxford University Press.

Converse, Philip E. 1970. "Attitudes and non-attitudes: Continuation of a dialogue."

Crofton, Morgan William. 1870. "IX. On the proof of the law of errors of observations." *Philosophical Transactions of the Royal Society of London* (160): 175-187.

Dahlgaard, Jens Olav, Jonas Hedegaard Hansen, Kasper M Hansen, and Martin V Larsen. 2016. "How are voters influenced by opinion polls? The effect of polls on voting behavior and party sympathy." *World Political Science*, 12 (2): 283-300.

Daston, Lorraine, and Peter Galison. 1992. "The image of objectivity." *Representations*, 40: 81-128.

- - -. 2010. *Objectivity.* First paperback edition. ed. New York: Zone Books.

Elishakoff, Isaac. 2003. "Notes on philosophy of the monte carlo method." *International applied mechanics*, 39 (7): 753-762.

Fischer, Hans. 2011. *History of the central limit theorem: from classical to modern probability theory.* New York ;: Springer.

Fisher, Stephen D, Robert Ford, Will Jennings, Mark Pickup, and Christopher Wlezien. 2011. "From polls to votes to seats: Forecasting the 2010 British general election." *Electoral Studies*, 30 (2): 250-257.

Galison, Peter. 1996. "Algorists dream of objectivity." *Psychology, Public Policy, and Law*, 2 (2): 293-323.

- - -. 1998. "Judgment against objectivity." *Picturing science, producing art*: 327-359.

Grant, Lawrence V, and John W Patterson. 1975. "Non-attitudes: The measurement problem and its consequences." *Political Methodology*: 455-481.

Hacking, Ian. 1990. *The taming of chance.* Repr. ed. *Ideas in context* ; 17. Cambridge [etc.]: Cambridge U.P.

Hagen, G. 1868. "Grundzüge der Wahrscheinlichkeits-Rechnung. Berlin, 1837. 8. 2." *Aufl. Das*, 8 (3).

Hakhverdian, Armen. "De Peilingwijzer is hét medicijn tegen luie scorebordjournalistiek." Stuk Rood Vlees, 20-02-2014. https://stukroodvlees.nl/de-peilingwijzer-is-het-medicijn-tegen-luie-scorebordjournalistiek/.

IO-Research. "Veel gestelde vragen bij opinie- en draagvlakonderzoek." Accessed 23-06-2022. https://www.ioresearch.nl/veelgestelde-vragen-bij-opinie-en-draagvlakonderzoek/.

Jackman, Simon. 2000a. "Estimation and inference are missing data problems: Unifying social science statistics via Bayesian simulation." *Political Analysis*, 8 (4): 307-332.

- - -. 2000b. "Estimation and inference via Bayesian simulation: An introduction to Markov chain Monte Carlo." *American journal of political science*: 375-404.

- - -. 2005. "Pooling the polls over an election campaign." *Australian Journal of Political Science*, 40 (4): 499-517.

Kahneman, Daniel, Stewart Paul Slovic, Paul Slovic, and Amos Tversky. 1982. *Judgment under uncertainty: Heuristics and biases.* Cambridge university press.

Kaila, Eino Sakari, Robert S. Cohen, G. H. von Wright, and Ann Kirschenmann. 1979. *Reality and experience : four philosophical essays.* Vienna Circle collection ; vol. 12. Dordrecht: Reidel.

Keynes, John Maynard. 1909. "A treatise on probability." *Diamond*, 3 (2): 12.

Levy, Ed. 2001. "Quantification, mandated science and judgment." *Studies in History and Philosophy of Science Part A* 32 (4): 723-737.

Louwerse, Tom. 2011. "Peilingwijzer ". Accessed 21-06-2022. https://peilingwijzer.tomlouwerse.nl/.

- - -. 2016. "Improving opinion poll reporting: the Irish Polling Indicator." *Irish Political Studies*, 31 (4): 541-566.

Marsh, Catherine. 1985. "Back on the bandwagon: The effect of opinion polls on public opinion." *British Journal of Political Science* 15 (1): 51-74.

Massimi, Michela. 2011. "From data to phenomena: a Kantian stance." *Synthese* 182 (1): 101-116.

Meeus, Tom-Jan. "Een nieuwe Maurice de Hond bedenken." NRC, 18 januari, 2017. https://www.nrc.nl/nieuws/2017/01/18/een-nieuwe-maurice-de-hond-bedenken-6274388-a1541911.

Neuber, Matthias. 2012. "Invariance, structure, measurement–Eino Kaila and the history of logical empiricism." *Theoria* 78 (4): 358-383.

NOS. " Peilingwijzer: D66 valt ver terug, ook VVD en CDA verliezen ", 20-05-2022. //nos.nl/artikel/2429586-peilingwijzer-d66-valt-ver-terug-ook-vvd-en-cda-verliezen.

Porter, Theodore M. 1985. "The mathematics of society: Variation and error in Quetelet's statistics." *The British Journal for the History of Science* 18 (1): 51-69.

Price, Vincent. 2008. "The public and public opinion in political theories." *The SAGE* *handbook of public opinion research*: 11-24.

Rousseau, Jean-Jacques, Susan Dunn, Gita May, Robert N. Bellah, David Bromwich, and Conor Cruise O´Brien. 2002. *The Social Contract and the First and Second Discourses.* New Haven, United States: Yale University Press.

RTL-Nieuws. "VVD, PVV, CDA, D66, SP en GroenLinks nemen deel aan het RTL Verkiezingsdebat." 14-02-2021. https://www.rtlnieuws.nl/nieuws/politiek/artikel/5214354/vvd pvv-cda-d66-sp-en-groenlinks-nemen-deel-aan-het-rtl.

Särndal, Carl-Erik. 1971. "Studies in the History of Probability and Statistics. XXVII: The hypothesis of elementary errors and the Scandinavian school in statistical theory." *Biometrika*, 58 (2): 375-391.

Schlaudt, Oliver, and Lara Huber. 2016. *Standardization in measurement: philosophical, historical and sociological issues.* London ;: Routledge.

Sheynin, Oscar B. 1979. "CF Gauss and the theory of errors." *Archive for history of exact sciences*, 20 (1): 21-72.

Sowey, Eric, and Peter Petocz. 2017. *A panorama of statistics : perspectives, puzzles and paradoxes in statistics.* Chichester, West Sussex: John Wiley and Sons, Ltd.

Stahl, Saul. 2006. "The evolution of the normal distribution." *Mathematics magazine* 79 (2): 96-113.

Tal, Eran. 2021. "Two Myths of Representational Measurement." *Perspectives on Science* 29 (6): 701-741.

Van der Meer, Tom WG, Armen Hakhverdian, and Loes Aaldering. 2016. "Off the fence, onto the bandwagon? A large-scale survey experiment on effect of real-life poll outcomes on subsequent vote intentions." *International Journal of Public Opinion Research*, 28 (1): 46-72.

Van der Meer, Tom WG, Annemarie Walter, and Peter Van Aelst. 2016. "The contingency of voter learning: How election debates influenced voters' ability and accuracy to position parties in the 2010 Dutch election campaign." *Political Communication*, 33 (1): 136-157.

Woodward, James. 1989. "Data and phenomena." *Synthese*: 393-472.

- - -. 2000. "Data, phenomena, and reliability." *Philosophy of Science*, 67: 163-179.

- - -. 2010. "Data, phenomena, signal, and noise."*Philosophy of Science*, 77 (5): 792-803.

- - -. 2011. "Data and phenomena: a restatement and defense." *Synthese*, 182 (1): 165-179.