De twijfelachtige kwaliteit van de kritiek op mijn werk en Peil.nl

Drs. Maurice de Hond (Peil.nl)

Op 9 oktober 1976 kwam ik via “In de Rooie Haan” voor het eerst naar buiten met een peiling naar de politieke voorkeur in Nederland. Het was gebaseerd op de wekelijkse onderzoeken van Nipo in haar zogenaamde omnibus (vragenlijsten waarin meerdere onderzoeken werden gecombineerd). Op basis van een analyse van de cijfers van Nipo rondom de Tweede Kamerverkiezingen in de 15 jaar ervoor, ontdekte ik een systematische afwijking (zo werd de VVD structureel onderschat en de PvdA overschat). Tevens ontwikkelde ik een manier om de schommelingen van meting tot meting te dempen vanuit de scores van week tot week bij de vraag “wat heeft u bij de laatste Tweede Kamerverkiezing gestemd?”. Er volgde veel publiciteit, mede omdat uit de eerste peiling bleek dat de PvdA er minder goed voor stond en de VVD beter dan tot dat moment werd aangenomen. Deels kwam dat door de correctie van de systematische afwijking en deels door de ontwikkelingen in de september 1976.

Twee dagen erna kwam er kritiek uit universitaire kring. **Wil Foppen**, van de Erasmus Universiteit, die o.a. verantwoordelijk was voor het veldwerk van het Nationale Kiezersonderzoek (NKO), vond het maar niets wat ik aan het doen was. “Als een steekproef niet goed is, dan moet je het weggooien” was één van zijn statements.

Bij het voorbereiden bij mijn artikel voor Acta Politica in dat najaar, waarin ik mijn methode uitvoerig verantwoordde, bekeek ik ook de verantwoording van Wil Foppen zelf bij het NKO van de Tweede Kamerverkiezingen uit 1972. Daar stond o.a. dat er een duidelijke ondervertegenwoordiging was van SGP-ers in de steekproef van het NKO en dat om het effect daarvan te verminderen er een X aantal willekeurig gekozen ponskaarten van SGP-ers werden gedupliceerd. (Begin van de jaren 70 was dit de enige manier om bij computers wegingen toe te passen: dupliceren van ponskaarten of het weglaten van ponskaarten. Vanzelfsprekend via een willekeurig proces). Blijkbaar had hij zelf een niet-goede steekproef niet weggegooid, maar met een -in die tijd geaccepteerde procedure- zo goed mogelijk gecorrigeerd. En was dat de basis voor analyses die door de Nederlandse politicologen van de verkiezingen van 1972 werd uitgevoerd.

In deze links vindt u de uitleg van mijn aanpak uit 1976 en de kritiek van Foppen terug. Het is interessant om terug te lezen omdat het -helaas- een begin was van een 40-jarig durend patroon:

- De methode uitgelegd.
- De kritiek van Wil Foppen
- Mijn inhoudelijke reactie hierop

Keer op keer werd in de laatste 40 jaar vanuit de wetenschappelijke hoek in de media kritiek uitgeoefend op wat ik deed en hoe ik het deed. En de laatste 15 jaar, sinds ik niet meer bij de Politieke Barometer betrokken ben, voegden diverse onderzoeksbureaus zich hierbij. Hun benadering leek sprekend op die van Wil Foppen uit 1976: er werd op diverse manieren gesuggereerd dat wat ik deed niet aan wetenschappelijke standaarden beantwoordde, daarbij impliciet of expliciet stellend, dat wat men zelf deed, wel aan die standaarden beantwoordde. Maar als je dan hun aanpak bestudeerde bij uitgevoerd onderzoek, dan had men minimaal dezelfde -soorten- problemen, waarvoor men ook oplossingen zocht.

Het palet van de kritiek werd zelfs veelkleuriger. Er werd steeds meer uit de kast gehaald. En dan ook nog vaak met de usual suspects als woordvoerders.

Het interessante daarbij is dat als ik die reacties in media lees of hoor op mijn werk steeds één of meer van de volgende vier basiselementen herken:

A. Weinig of geen echte kennis van hoe ik mijn onderzoek doe (dus ook niet mijn verantwoording gelezen hebbend of zich verdiept in mijn aanpak).
B. De problematiek bij hun eigen onderzoek niet herkennen of via de kritiek op mijn werk de eigen problemen maskeren.
C. Weinig snappen van hoe kansberekening echt toegepast dient te worden, hoewel men het tegendeel suggereert.
D. Jalousie de metier en/of gebrek aan integriteit.

Bij sommige reacties herken ik zelfs alle vier tegelijk.

Ik heb in de afgelopen jaren via mailwisselingen achter de schermen met meerdere betrokkenen discussies gevoerd over hun openbare uitingen of hun eigen onderzoek (en ze gewezen op mogelijke problemen daarbij), en dan zie ik vrijwel steeds één of meer van de bovenstaande vier basiselementen terug. Daarbij blijkt vaak ook een vorm van onwil om logisch na te denken (of erger). Op zo'n basis kan ook geen echte discussie gevoerd worden. Zoals bij voorbeeld ook recentelijk bleek bij de uitzending van DWDD, waar **Siewert van Lienden,** die zoals uit zijn eigen woorden bleek, vooraf met zeker twee marktpartijen had gesproken, en maar een aantal mantra's bleef herhalen – mantra's die ik ook vaker had gehoord. Maar het feit dat een dag later GfK/EenVandaag (waarvan Van Lienden , zoals achteraf bleek, tijdens de uitzending de uitslag al van wist), met een peilingsuitslag zou komen die sterk leek op die van mij in de voorgaande week, maakte in zijn opstelling blijkbaar niets uit. Van Lienden: "Het gaat niet om de cijfers, maar om de aanpak".

Want ook dat is een bijna vast patroon. De belangrijkste trends in het electoraat, zoals die door andere peilingen worden getoond, stel ik doorgaans als eerste vast. (Dat komt zowel door mijn bijzondere aanpak als doordat ik wekelijks meet en de anderen minder vaak peilen.)

Mede door deze onwil om logisch na te denken heb ik aangegeven geen zin meer te hebben om in het openbaar te debatteren met die usual suspects over de methode van peilingen. Ik had het ook bij de bijeenkomst "Peilingoproer" niet gaan doen, ook als ik dan wel in Nederland terug was geweest.
Ik wil echter wel op een aantal inhoudelijke zaken ingaan, en die nu niet -via mailwisselingen- naar bepaalde personen apart sturen, maar op deze manier communiceren. Ik zal op diverse argumenten ingaan die ik de afgelopen jaren in verschillende varianten van verschillende betrokkenen heb gehoord.

.

1. **Een panel met zelfaanmeldingen is slecht**

Het grote manco van ALLE steekproefonderzoeken onder Nederlanders is dat deze tegenwoordig op geen enkele manier meer in de buurt komen van een goede steekproef!

Zelfs toen in de jaren zeventig van de vorige eeuw enquêteurs nog bij de ondervraagden thuis kwamen en een responsecijfer van 75 a 80% werd gerealiseerd, werd dat ideaal niet gehaald. Maar de afwijkingen van dat ideaal waren beduidend minder dan nu. Sindsdien werd het alleen maar -veel- slechter.

Aanvankelijk waren de reponsecijfers bij telefonisch onderzoek nog vrij redelijk, maar inmiddels is het bij bezoek van enquêteurs aan huizen of telefonische benadering huilen met de pet op. Het aantal mensen dat je mag benaderen nam fors af (mede door het "bel me niet register') en als de mensen wel bereikt worden, werkt een steeds kleiner deel mee. Ook via internet is het niet mogelijk om zelfs een kwalitatief redelijke steekproef te krijgen. Of je het dan via zelfaanmelding doet of het kleine deel van de groep die je wel benadert, die meewerkt, het komt vrijwel op het zelfde neer. De kwaliteit van de steekproef is matig als je

geluk hebt, maar vaak slecht. Dat geldt voor al het onderzoek in Nederland dat pretendeert een beeld te geven van de Nederlandse populatie.

Dan kan je twee dingen doen. Geen (steekproef)-onderzoek meer uitvoeren of maximale inspanningen verrichten om toch cijfers te krijgen, die een behoorlijke afspiegeling geven van de werkelijkheid. Ik stel vast dat bij de verschillende marktonderzoekbureaus nog steeds onderzoek wordt verricht en daar veel geld voor ontvangen van bedrijven en instellingen en dat bij de universitaire instellingen dit ook het geval is, ondanks de gesignaleerde grote problemen in het bestand van respondenten. Tevens geldt dat voor het CBS en het SCP. Men probeert er -terecht- nog het beste van te maken.

Zij proberen daarbij, net als ik, de tekortkomingen in hun bestanden zo goed mogelijk onder ogen te zien en daarvoor te corrigeren.

Op zichzelf vind ik dat begrijpelijk. Maar waarvan ik langzamerhand mijn buik vol van heb, is dat dezelfde mensen dan methodische argumenten hanteren om mijn aanpak te diskwalificeren, terwijl er minimaal net zoveel is aan te merken op hun eigen aanpak.

Een goede illustratie hiervan is de kritiek van professor **Andre Krouwel** van de Vrije Universiteit. Eerst gaf hij in de media uitvoerig af over mijn aanpak met een panel van zelfaanmelding. Kort erna las ik in Trouw echter een analyse van zijn hand over wat Nederland vond, op basis van antwoorden gegeven bij het -overigens interessante- instrument van Kieskompas. Alsof die deelnemers aan Kieskompas ook niet een eigen variant zijn van een groep kiezers die zich zelf hebben aangemeld en geenszins een goede afspiegeling vormen van de Nederlandse kiezer.

Ik wist niet goed of ik moest huilen of lachen toen ik las wat Krouwel op de ochtend van het Oekraïne-referendum voorspelde over de opkomst. Die zou hoog worden, in de richting van 42 procent, want dat hadden de mensen bij het invullen van Kieskompas gezegd. En dat als de opkomst niet hoog zou zijn de mensen toch wel erg tegen hem hadden gelogen. (Wederom: alsof de deelnemers aan Kieskompas een goede weerspiegeling zijn van alle kiezers en alsof er niet al een hele lange ervaring is dat kiezers beduidend meer zeggen op te komen dan ze uiteindelijk opkomen. Een van de uitdagingen bij voorspellingen van de uitslagen van verkiezingen is immers te kunnen bepalen welke groepen kiezers uiteindelijk meer niet opkomen in relatie tot wat ze gezegd hebben dan andere groepen).

Ook professor **Philip van Praag** vertoont dit gedrag. Afgeven op de kwaliteit van mijn steekproef en vervolgen zelf conclusies trekken op basis van een ander onderzoek, waarbij je ook ernstige vraagtekens kan zetten bij de kwaliteit.

In de aanloop van verkiezingen van 2003 was ik voor het eerst actief met een steekproef op internet vanuit mensen die zich zelf hadden aangemeld. Direct na het RTL-debat met het debuut van Wouter Bos, stelde ik in dagelijkse peilingen voor het SBS-programma “Stem van Nederland” vast dat de PvdA snel de grote achterstand op het CDA inhaalde. Toen CDA en PvdA bij mijn dagelijkse peiling vrijwel gelijk stonden gaf de Politieke Barometer nog een verschil van 12 zetels aan. **Reinier Heutink** van de Politieke Barometer had in de publiciteit forse kritiek op mijn aanpak “Dit soort onderzoek kan niet via internet”, “Het is zelfaanmelding”, “Je moest het niet doen op basis van dagelijkse metingen, maar gemiddeldes over 2 weken” (zoals De Politieke Barometer toen deed, omdat hun veldwerk op deze manier toen nog georganiseerd was). Toen ik twee weken later in TweeVandaag zat (toen heette dat nog zo) om over deze kritiek te praten en inmiddels de Politieke Barometer ook een fors kleiner verschil gaf tussen PvdA en CDA, was mijn argument dat ik hetzij blijkbaar toch goed onderzoek had gedaan of telepathisch begaafd was om twee weken van te voren de uitslag van de Politieke Barometer te voorspellen. Ik heb daar tussentijds ook nog dit artikel over geschreven.

Op dinsdag 29 november jl. kwam Gfk met haar uitslag voor EenVandaag uit. En op 6 december kwam TNSNipo met hun uitslag. Beide bureaus pretenderen een soort steekproeftrekking uit te voeren zonder zelfaanmelding. Beide uitslagen lijken voor de meeste partijen sterk op de uitslag van mijn peiling, die ik in de weken voorafgaand hieraan had gepubliceerd. Ook de Peilingwijzer van 21 december bevestigt vrijwel volledig mijn cijfers van een aantal weken eerder. En inmiddels laat I&O ook een sterk toegenomen verschil tussen PVV en VVD zien.

Logischerwijs kan er maar één conclusie worden getrokken:

De door mij gekozen aanpak is minimaal net zo goed als die van de andere bureaus.

2. **De statistische marges zijn zodanig dat minimaal 3 zetels verschuiving tussen peilingen pas wat inhoudelijks kan betekenen.**

Met regelmaat wordt er in reacties van concurrenten of wetenschappers gewezen op de statistische marges die gelden bij steekproefonderzoek. Er worden dan voorbeelden genoemd met enkelvoudige kansberekening, in de vorm van "bij een steekproef met 2000 ondervraagden is er een statistische marge van zeker 3 zetels naar boven en beneden".

En dat wordt vaak op een manier gedaan alsof ik geen enkel verstand van statistiek zou hebben en deze simpele kansberekening formules niet zou kennen.

Allereerst was ik één van de weinige sociaalgeografen, die aan het eind van de zestiger jaren tijdens zijn doctoraal Statistiek als bijvak nam. Ik kreeg een 10 voor mijn mondeling doctoraal van de destijds befaamde statistiekdocent J.C. Spitz. In de jaren doceerde ik, naast vele andere methodologische vakken, ook statistiek aan sociaalgeografen aan de UvA.

Dat ik in 1976 überhaupt met mijn correctiemethodes ben gekomen voor peilingen, was omdat ik niet alleen goed op de hoogte was van statistiek, maar ook de praktijk van het veldwerk bij het onderzoek van nabij had leren kennen bij het marktonderzoekbureau Inter/View, waar ik was komen werken.

Het is inderdaad zo dat als ik een perfecte steekproef trek van bij voorbeeld 2000 ondervraagden bij een willekeurig onderzoek en bijvoorbeeld 30% van de ondervraagden geeft een bepaald antwoord, dat je precies via kansberekening kan vaststellen tussen welke marges dat ligt met 95% zekerheid. (Dat ligt dan in de buurt van plus of min 2%).

Maar dat gaat uit van een ideale steekproef, iets wat bij een onderzoek onder Nederlandse burgers niet gerealiseerd kan worden, zoals hierboven al uiteen is gezet. Het houdt evenmin rekening met het feit dat bij het uitvoeren van het onderzoek, op de wijze waarop ik het doe, de formules van kansberekening veel complexer zijn dan de basisformules uit het handboek statistiek.

Bij een ideale steekproef zijn er niet alleen onzekerheidsmarges op basis van kansberekening (je werkt met een steekproef en niet met de hele populatie, dus als je nog een keer een steekproef trekt kunnen de cijfers wat verschillen, waarvoor formules zijn op basis van kansberekening). Er is ook nog een andere -heel belangrijke- afwijking. In 1976 stelde ik bij NIPO al vast dat via de vraag "wat heeft u gestemd?" gesteld kort na vier Tweede Kamerverkiezingen ervoor er elke keer een oververtegenwoordiging was van gemiddeld 2% van PvdA-kiezers en een ondervertegenwoordiging van gemiddeld 2% van VVD-kiezers. Het bleek te gaan om wat tegenwoordig bekend staat als het "Huiseffect" van, in dit geval, NIPO.

Dat patroon van structurele verschillen is tegenwoordig ook goed te illustreren via de Peilingwijzer. Daar is een groot overzicht van dit "Huiseffect" van de verschillende bureaus.

Zo laat de Politieke Barometer scores zien voor de VVD en PVV die fors verschillen van die van Peil.nl, terwijl het patroon van het verloop in de tijd bij die partijen redelijk op elkaar lijkt. I&O laat voortdurend cijfers zien voor de PvdA die hoger zijn dan die van de andere bureaus en de PVV, die lager zijn dan van de andere bureaus.

Daaruit is al op te maken dat per losse peiling de omvang van de structurele verschillen met de werkelijkheid groter kunnen zijn dan die berekend worden op basis van de formule van de kansberekening. De marges kunnen wel meer dan twee keer zo groot zijn dan alleen op basis van kansberekening wordt bepaald.

Mede daarom besloot ik 40 jaar terug dat je zoveel mogelijk vanuit het beschikbare materiaal moet proberen die marges aanzienlijk te verkleinen. Zowel de structurele marges als de marges op basis van kansberekening.

Ja, inderdaad, ook die op basis van kansberekening!

Aan de hand van het volgende voorbeeld laat ik zien dat bij een goede aanpak ook die laatste marges kleiner kunnen worden. Stel, je trekt een steekproef van 2500 ondervraagden 1 dag *na* de verkiezingen. Je vraagt de respondenten welke partij men gestemd heeft (recall) en op welke partij men zou stemmen als er vandaag verkiezingen zouden zijn (stemintentie). Laten we even aannemen dat 3% van de ondervraagden zeggen dat ze vandaag een andere partij stemmen dan gisteren op de verkiezingsdag. Per partij verschilt de uitslag van gisteren dan heel weinig met die van vandaag, want een deel van de bewegingen is immers tegengesteld aan elkaar.

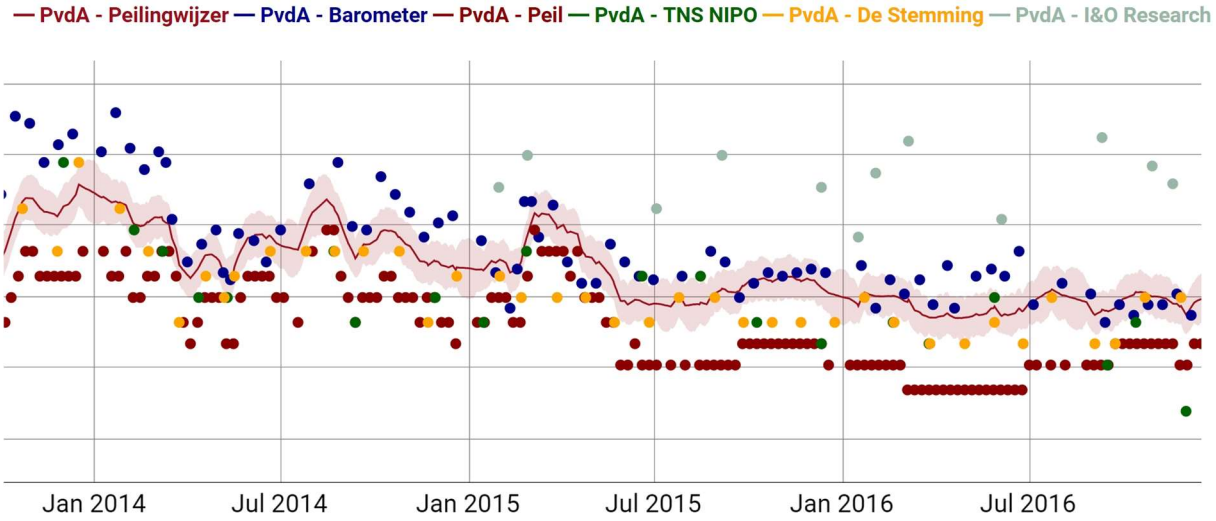
Laten we even de andere mogelijke wegingscomponenten buiten beschouwing houden, maar deze steekproef wegen op het stemgedrag bij de verkiezingen de dag ervoor (recall). Bij het *on*gewogen materiaal zegt bij voorbeeld 22,1% gisteren partij A gestemd te hebben en 22,3% zegt het nu te stemmen. Maar Partij A scoorde gisteren in het echt 20,0%. Door naar die feitelijke uitslag van 20,0% te wegen komen we bij een uitslag van “gisteren gestemd op Partij A” (recall) op 20,0%. Dezelfde weging op “Stemt vandaag Partij A” (stemintentie) komt dan uit op 20,2%.

Dan is het dus niet dat je te maken hebt met een marge van de kansberekening van bij voorbeeld 2% rond 20,2% omdat je een steekproef hebt van 2000. Nee, de marge die dan berekend wordt moet dan alleen gericht zijn *op de 3% mensen die van partij zijn veranderd*. Dat percentage kan in de hele populatie iets hoger of lager zijn. Maar de onzekerheidsmarge van de kansberekening is dan over de eindscore per partij nog maar een fractie van wat je normaal zou berekenen bij een normale ideale steekproef van een onderzoek over een willekeurig onderwerp. In dit voorbeeld zou die marge maar ongeveer 0,2% zijn, in plaats van 2% (dus fors minder dan 1 zetel)!

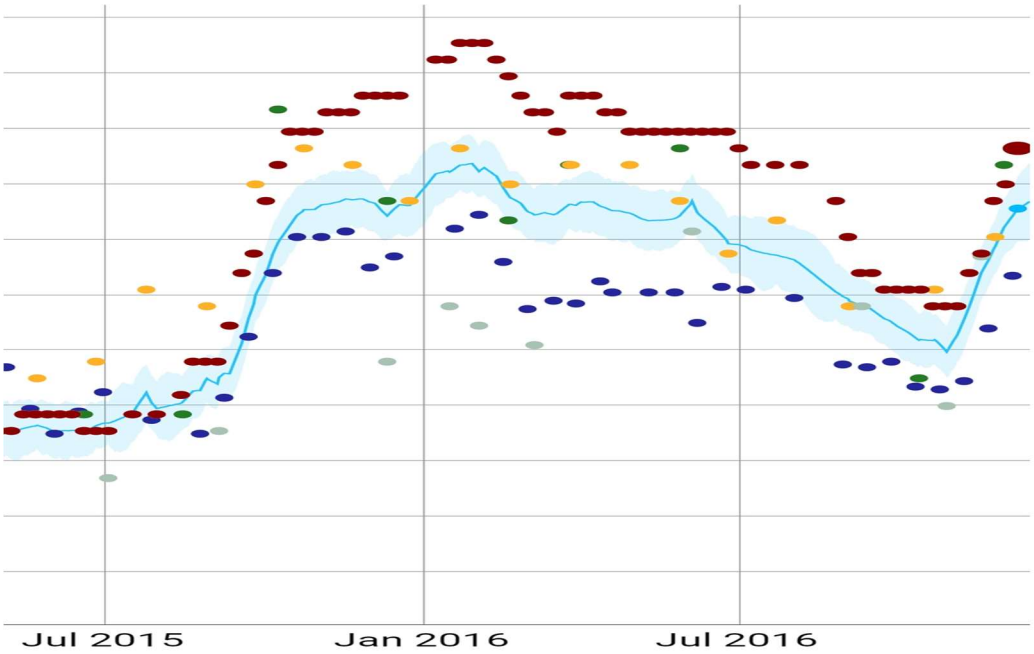
Naarmate de verkiezingen verder in het verleden liggen nemen de verschuivingen toe t.o.v. het vorig stemgedrag (recall) en zal dus de statistische marge op basis van kansberekening voor het totaalpercentage groter worden. Maar ook dan is deze marge beduidend kleiner dan in het geval van de enkelvoudige kansberekening, die door critici steevast als voorbeeld wordt gegeven wanneer men beweert dat het minstens 3 zetels zou moeten zijn.

Iets anders is dat naarmate de verkiezingen verder weggelegen zijn mensen meer fouten maken ten aanzien van het weergeven van hun laatste stemgedrag. Via Peil.nl heb ik dat kunnen vaststellen, omdat ik van een deel van het panel tijdens de vorige verkiezingen zelf heb kunnen vaststellen of en hoe men gestemd heeft. Inmiddels geeft 17% van de ondervraagden het verkeerde antwoord. Dus als een bureau een nieuwe steekproef neemt en ijkt op het vorig stemgedrag op basis van wat mensen nu daarover zeggen, dan zit die 17% er dan ook al fout in. Daardoor kan bijvoorbeeld de winst van de PVV op dit moment onderschat worden en het verlies van de PvdA onderschat. En dat komt omdat meer mensen zeggen zich te herinneren PVV gestemd te hebben, terwijl ze dat toen niet gedaan hebben en bij de PvdA

herinnert een deel van de respondenten zich niet meer dat ze PvdA hebben gestemd. Het zou me niets verbazen als de peilingen van **I&O**, die pas sinds 2014 met dit soort peilingen bezig zijn, dus na de verkiezingen van 2012, onder dit patroon lijden. De PVV is bij dat bureau fors lager dan de rest en de PvdA is bij dat bureau fors hoger dan de rest, zoals uit onderstaande grafieken van Peilingwijzer blijkt. Kijk naar de lichtgroene bolletjes in vergelijking met de rest.



En dit is de grafiek van de PVV.



Op basis van het cijfermateriaal zelf van mijn peiling is eenvoudig te bewijzen dat de statistische marges fors lager zijn dan als je simpelweg alleen de marge berekent op basis van de omvang van de steekproef van die week en het percentage dat een partij behaalt.

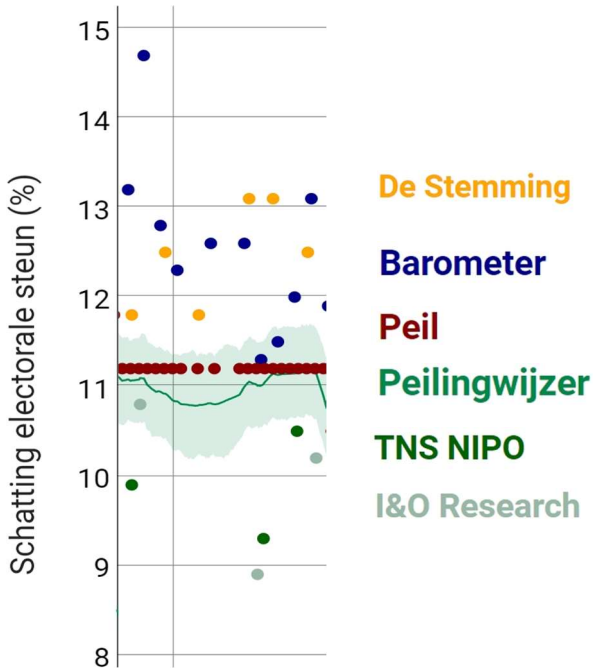
Stel dat op één dag achter elkaar 50 keer een ideale steekproef wordt getrokken van 2500 Nederlanders en gevraagd wordt welke partij ze gestemd hebben en je nummert ze van 1 tot 50. En stel verder dat er geen enkele bewerkingsslag wordt uitgevoerd. Dan heb je inderdaad per afzonderlijke steekproef bij de grootste partijen een marge van plus of min 2% (3 zetels). Dat houdt in dat aangenomen mag worden dat in 95% van de metingen de werkelijke score in de populatie binnen die marge van plus of min 2% zit rond het gevonden percentage in die steekproef. Hoe vaak, kun je je nu afvragen, zal het gebeuren dat een

steekproef qua zeteltal voor *alle* partijen *volledig identiek* is aan de steekproef met het voorafgaande volgnummer? Het berekenen van die kans vergt een complexe berekening, maar zal, gezien het feit dat bij alle 10 partijen sprake is van marges bij de kleinere partijen van 1 en bij de grotere partijen van 3 zetels, vrij klein zijn.

Laten we nu eens de 50 peilingen nemen die ik in 2016 heb gepubliceerd. Die zijn niet op een en dezelfde dag genomen, maar wekelijks. In die periode zou het ook nog zo kunnen zijn dat kiezers van voorkeur zijn veranderd. Desondanks is het in 2016 bij maar liefst 27 (!) van de 50 peilingen (54%) zo geweest dat ik in de daaropvolgende week geen enkele verandering in zetels heb gerapporteerd. In de zomer van 2016 trad zelfs 5 weken achter elkaar geen enkele verschuiving in zetelaantallen op. De kans dat dit gebeurt is uitermate klein als het werkelijk zo zijn dat de statistische marge per grote partij 3 zetels is. Dan is het vrijwel onmogelijk dat 5 keer achter elkaar voor alle 10 partijen dezelfde score eruit kwam.

Dat laat al zien dat de statistische marges van Peil.nl beduidend kleiner zijn dan die door de criticasters aan Peil.nl wordt toegeschreven.

Aan de hand van de onderstaande grafiek kan dat nog verder uitgewerkt worden. Dit komt van de site van Peilingwijzer en betreft de scores van het CDA bij Peilingwijzer en de 5 Nederlandse peilingen tussen eind mei 2016 en begin november 2016.



In deze periode van ruim 5 maanden haalde het CDA bij *alle* 25 achtereenvolgende wekelijkse peilingen van Peil.nl 17 zetels (in deze grafiek vertaald naar 11.2%). Goed te zien is dat bij de andere 4 peilingbureaus in deze periode het CDA schommelde. Bij de maandelijkse meting van De Stemming (EenVandaag) betrof dat 2 zetels variatie, bij de tweewekelijkse meting van de Politieke Barometer (Ipsos) betrof dat 5 zetels, bij TNS Nipo (met 3 peilingen) betrof dat 2 zetels en bij I&O (met 3 peilingen) ging het om 3 zetels.

Nu is het zo dat het principe van de kansberekening als volgt werkt. Stel je neemt een oneindig aantal steekproeven met dezelfde omvang. En je kijkt wat de gemiddelde waarde van het CDA is. Vervolgens wordt er gekeken naar de standaarddeviatie, dat is de mate waarin de gevonden waardes van het CDA bij al die steekproeven gespreid zijn rond die gemiddelde waarde. Als je de gevonden waardes afzet in een histogram dan krijg je een normaalverdeling (dus een soort klokvorm) met als hoogste punt het gemiddelde en dan loopt het aan de beide kanten af. De mate waarin het rechts en links van dat hoogste punt afloopt

hangt samen met die standaarddeviatie, de spreiding rondom het gemiddelde. Hoe kleiner die standaarddeviatie is hoe “smaller” de klokvorm. 95% van de waardes zullen liggen tussen het gemiddelde min 2 maal de standaarddeviatie en het gemiddeld plus 2 maal de standaarddeviatie. Dat betekent dat 1 van de 20 steekproeven buiten die marge valt.

Als je maar één steekproef trekt dan is het als het ware één willekeurige uit dit oneindig aantal getrokken steekproeven. Op basis van het gemiddelde en standaarddeviatie kan je dan berekenen waar de echte scores in de populatie liggen met 95% zekerheid. (Want 1 van de 20 kan er buiten liggen).Maar als je slechts één steekproef trekt dan heb je wel een waarde voor het CDA. Maar je weet niet hoe die waarde zich verhoudt tot de echte waarde in de populatie. Je gaat er dan van uit dat met 95% betrouwbaarheid het ligt binnen 2 maal de standaarddeviatie rechts en links van die waarde. Maar met slechts één steekproef kan je ten aanzien van de waarde van het CDA geen standaarddeviatie berekenen. (Die krijg je pas bij meerdere steekproeven). Met behulp van de formules uit de kansberekening kan echter berekend worden wat die standaarddeviatie zal zijn op basis van de grootte van de steekproef. Bij een kleine steekproef is de standaarddeviatie groter dan bij een grote steekproef (Een 4 keer grotere steekproef geeft een 2 keer kleinere marge). Bij een uitkomst van 11,2% voor het CDA en een omvang van de steekproef van 3000 levert de formule een standaarddeviatie op van 0,6%. En zal dus 95% van alle uitkomsten liggen tussen 11,2% min 1,2% en 11,2% plus 1,2%. Dat is tussen 10,0 en 12,4%. Een marge van 2 x 1,2% = 2,4%.

Maar bij 25 scores achter elkaar, waarbij de score dusdanig is dat er geen verschuiving van 1 zetel optreed (0,67%), zoals bij deze uitkomsten van Peil.nl, is er een standaarddeviatie van ongeveer 0,15%. En ligt de marge dus slechts tussen 11,2% – 2 x 0,15% en 11,2% + 2 x 0,15%. De marge is dus 0,6 groot. Een 4 keer kleinere marge dan via de enkelvoudige kansberekening wordt bepaald. Dat zou hetzelfde effect zijn als een 16 keer grotere steekproef (dus geen 3000, maar 48000).

In dit geval kan dus gesteld worden dat als 25 keer achter elkaar het CDA 17 zetels scoort een stijging van 1 of daling bij Peil.nl van 1 wel significant is!

De schommelingen van de andere vier peilingen vertonen trouwens wel het patroon dat je kan verwachten op basis van de enkelvoudige berekeningen van de marges in relatie tot de omvang van de steekproeven van die bureaus. (In die periode zien we ook dat Peilingwijzer amper verschuift. Dat komt deels door Peil.nl, maar de scores van de andere peilingen leverden blijkbaar geen duidelijke andere beweging op).

> **De conclusie luidt dat de betrouwbaarheidsmarges op basis van de kansberekening van peiling tot peiling bij Peil.nl beduidend kleiner zijn dan bij de andere bureaus.**

Dat brengt met zich mee dat een verschuiving van 1 à 2 zetels bij Peil.nl van de ene peiling tot de andere peiling duidelijk meer betekenis heeft dan eenzelfde verandering bij de andere bureaus. Botweg debiteren dat pas bij een verschuiving van 3 zetels bij Peil.nl sprake is van een significante verschuiving is ronduit onzin en laat zien dat men zich niet serieus heeft verdiept in de aanpak van Peil.nl en/of weinig kaas heeft gegeten van hoe kansberekening werkt.

Maar er is, afgaande op de scores van de andere bureaus door de jaren heen, meer aan de hand. Met regelmaat stel ik bij de gerapporteerde verschuivingen vast dat het niet alleen de marges van de kansberekening betreft, maar dat er ook iets anders het geval moet zijn geweest bij de uitvoering van het onderzoek. Zo laat de peiling van EenVandaag van eind december 2016 zien dat GroenLinks in een maand van 9 naar 15 zetels is gestegen, terwijl de voorafgaande drie Politieke Barometers juist aangaven dat GroenLinks van 15 naar 10 is gedaald. In diezelfde periode haalde GroenLinks bij mijn 9 wekelijkse metingen steeds 15 zetels, inmiddels is dat wekelijks al 4 keer 14 zetels. De omvang van de verschuiving bij EenVandaag en die van de Politieke Barometer zijn dermate groot dat als je er alleen kansberekening op zou toepassen de kans heel erg klein is dat dit het gevolg is van toeval. Peil.nl laat een stabiel patroon zien,

EenVandaag een zeer forse stijging en Politieke Barometer een forse daling. Juist door die tegengestelde bewegingen van de peilingen van die twee bureaus en de stabiele score van Peil.nl kan je haast met zekerheid zeggen dat er dus een ander probleem was bij Politiek Barometer en EenVandaag dan alleen het kansberekening-stuk!

Waardoor komt het dat bij mijn peilingen van week tot week de scores van de partijen zo stabiel zijn en verschuivingen consistent? Omdat ik bij mijn peilingen sinds 2002 een nieuwe manier van onderzoek toepas, die voordien niet mogelijk was. Vanaf het begin heb ik dat ook in mijn onderzoekverantwoording gepubliceerd, maar waarvan het lijkt dat de criticasters dit of hetzij nog nooit gelezen hebben of niet kunnen of willen begrijpen.

Als je een steekproefonderzoek doet dan is het probleem dat een nieuwe steekproef uit andere mensen bestaat en dat verschuivingen in politieke voorkeur die je vaststelt mede veroorzaakt kunnen worden doordat je je andere respondenten hebt. Als je op één en dezelfde dag twee verschillende steekproeven neemt met de omvang van de Politieke Barometer (circa 1000 respondenten) kunnen er per partij verschillen ontstaan van 1 tot 3 zetels! Die zeggen dus niets inhoudelijks, terwijl normaliter de electorale verschuivingen per partij van week tot week de 1 à 2 zetels niet overschrijdt (zie daarvoor de grafieken van Peilingwijzer).

Dat is de reden dat ik via wegingen en ijkingen die ik sinds 1976 heb toegepast, geprobeerd heb verschuivingen zo nauwkeurig mogelijk te meten. Toch hield ik verschuivingen over, die niet echt zijn, maar het gevolg van de onderzoeksopzet. Juist omdat de daadwerkelijke electorale verschuivingen van week tot week doorgaans klein zijn, heeft het mij altijd geërgerd als bij peilingen verschuivingen kwamen van 3 à 4 zetels, die vervolgens bij de volgende peiling weer werd geneutraliseerd. Dit kwam in het overgrote deel van de gevallen niet door bewegingen in het electoraat, maar door de onnauwkeurigheid van het meetinstrument. Maar over die verschuivingen werden dan toch inhoudelijke conclusies getrokken, gebaseerd op drijfzand.

Daarom ben ik vanaf 2002, toen ik met een internetpanel ben gaan werken, anders te werk gegaan. Bij de vraag op welke partij men nu gaat stemmen als het verkiezingen zijn, ziet de respondent bij Peil.nl al vooraf het antwoord ingevuld dat de respondent de vorige keer zelf heeft gegeven. De respondent moet die keuze actief veranderen om aan te geven dat de eigen keuze inmiddels veranderd is. (Per week wordt nooit het hele panel ondervraagd, maar doorgaans tussen de 3000 en 5000 personen, zodat men om de zoveel weken deze vraag krijgt voorgelegd).

Min peilingsuitslagen van week tot week zijn daarmee gebaseerd op de veranderingen in keuze op het niveau van individuele respondenten! (panelgegevens).

Allereerst viel me daarbij op dat de electorale verschuivingen van keer op keer beduidend lager waren dan het beeld dat andere peilingen geven (en, om eerlijk te zijn, ik zelf ook lange tijd dacht). Doorgaans maakt meer dan 90% van de respondenten dezelfde keuze als kort ervoor. Een groot deel van de verschuivingen compenseert elkaar ook nog eens (een deel van een partij weg en een ander deel naar die partij toe, zogenaamde “Browse bewegingen”). Het grootste deel van de bij andere bureaus gemelde verschuivingen zijn afgeleiden van het onderzoeksinstrument, maar geen weergave van de werkelijke bewegingen onder het electoraat.

In de tweede plaats viel me op dat als er duidelijke verschuivingen onder de kiezers waren, die toe te wijzen waren aan ingrijpende -politieke- gebeurtenissen waar veel aandacht voor was in de media.

Natuurlijk zijn er ook methodische kanttekeningen te maken bij deze (panel)aanpak. **De vraag is echter hoe relevant deze nog zijn als aangetoond kan worden dat het overgrote deel van de verschuivingen van enige omvang die ik heb getoond later ook door de andere peilingen werden bevestigd? (maar dan wel vaak op een veel grilliger manier, en soms pas weken later).**

Deze aanpak onderstreept verder dat ten aanzien van de verschuiving van peiling op peiling het niet zo is dat pas bij een verschuiving van 3 zetels er sprake is van statistische significantie. In sommige gevallen kan dat zelfs al bij 1 zetel!

Een goed voorbeeld daarvan is wat er gebeurde in de weken voorafgaande aan de Kamerverkiezingen van 2012. Bij het eerste debat (van RTL) waren maar 4 partijen uitgenodigd, D66 niet. Ik zou de laatste 7 uitzendingen voor de verkiezingen dagelijks in DWDD komen met de stand van zaken. Dat begon pas 8 dagen na het RTL-debat. Tussen dat RTL-debat en de uitzending heb ik wel 6 dagen metingen gedaan en ik stelde via mijn manier van meten van verandering op respondent-niveau dat D66 het slechter deed, terwijl dat niet zo was bij de deelnemers aan dat debat. Ik stelde daar in mijn peilingen ook nog een aantal expliciete vragen over.

En ja hoor het bekende patroon van de critici diende zich weer aan: Enkele dagen na de verkiezingen van 2012 verscheen er in NRC-Handelsblad een artikel van **Jean Tillie**, **Tom van der Meer** en **Armen Hakhverdian** over de peilingen met als tussenkop “DWDD was het ergste”. In dat artikel stond onder meer:

“Elk van die dagelijkse peilingen liet verschuivingen zien die zo klein waren – één zetel, twee zetels – dat ze hoogstwaarschijnlijk zijn ontstaan door toevalligheden. Zo werd in de eerste uitzending de daling van D66 in de peiling van De Hond met slechts een enkel zeteltje opgeblazen tot iets betekenisvols. Konden Pechtold en De Hond die daling verklaren? De Hond haastte zich om de aanleiding te zoeken in Pechtolds afwezigheid bij het Carrédebat van de avond ervoor, hoewel de verandering statistisch boterzacht was. DWDD draaide door en creëerde nieuws, ook als er geen nieuws was.”

In de eerste plaats betrof het niet het Carrédebat van de avond ervoor, maar het RTL-debat van 8 dagen ervoor. (Al 3 keer in de laatste 4 keer is het RTL-debat de gamechanger van de verkiezingen geweest, na het debat in 2012 schreef ik dit erover). Aan dit zogenaamde Premiers-debat deden 4 partijen mee (VVD, PVV, PvdA en SP). Het viel me meteen op bij de eerste (niet gepubliceerde) meting van de dag erna dat er een grote beweging was tussen SP en PvdA (die in de twee weken erna fors doorzette). Maar ook dat D66, die niet aanwezig was bij het debat, wat weg was gezakt. Ik heb vervolgens dagelijks metingen gedaan, met daarin ook specifieke vragen hierover. In totaal dus 6 verschillende metingen. Pas bij eerste uitzending van DWDD, 8 dagen na het debat, kwam ik naar buiten met de cijfers van dat moment. D66 was op elk van de 6 dagen (met steeds verschillende steekproeven) elke dag met circa 1% gedaald t.o.v. de score vlak voor het debat. De daling was consistent en hing samen met het niet aanwezig zijn bij het RTL-debat, hetgeen ik uit de specifieke vragen en de overgangenanalyse op individueel respondentniveau kon afleiden.

Dus de door mij gemelde verandering was statistisch allesbehalve boterzacht en bovendien waren de ‘feiten’ in het artikel onjuist voor het debat zelf en de datering ervan.

Onder de dekking van hun academische posities en/of titels hadden de auteurs zich -wederom- noch verdiept in mijn onderzoeksaanpak in het algemeen, noch in de specifieke aanpak van dat moment. De meest basale uitgangspunten die je zou moeten hanteren als je als een wetenschapper dit soort zaken publiceert werden zo met voeten getreden. Ook uit het emailverkeer dat ik daarna had, bleken ze niet te willen erkennen dat ze fors de fout in waren gegaan. Het was een bevestiging van elk van de bovenvermelde vier basiselementen van de criticasters waar ik mee te maken heb.

Elke keer weer als ik een wetenschapper of collega-bureau weer de 3 zetels statistische marge hoor zeggen t.a.v. mijn werk, weet ik dat hij of zij zich niet heeft verdiept in mijn aanpak en verantwoording. (En ik betwijfel of hij of zij echt meer van kansberekening weet dan ik).

3. "natte-vinger werk" oftewel "het is verboden goed na te denken"

Wat zich rond peilingen van DENK heeft voltrokken in de afgelopen maanden is een vorm van sublimatie van de processen die ik eerder heb gezien.

Toen eind 2014 de twee PvdA-kamerleden van Turkse herkomst uit de partij stapten werd ik gebeld door een journalist die me vroeg of zij enige kans zouden maken als ze met een eigen partij zouden starten. Mijn antwoord was dat ik de indruk had dat Nederlanders van Turkse herkomst, sterk aan Turkije gebonden waren, want het was mij vaak opgevallen dat bij het Eurovisie Songfestival in Nederland (en in Duitsland) bij het televoten vrijwel steeds 12 punten aan Turkije werd gegeven.

Daar werd toen nog een soort grapje over in De Volkskrant gemaakt.

In mei jl, nadat Sylvana Simons ook tot DENK was toegetreden, heb ik me verder verdiept in de keuze van met name Nederlanders met een Turkse herkomst. O.a. stelde ik vast dat bij de Tweede Kamerverkiezingen in 2012 meer dan 80.000 voorkeurstemmen waren uitgebracht op kandidaten met een Turkse naam. Omdat het een bekend gegeven is bij peilingen dat in Nederland autochtone kiezers (sterk) ondervertegenwoordigd zijn, heb ik datgene gedaan, wat ik vanaf 1976 steeds heb gedaan, in relatie tot een duidelijke ondervertegenwoordiging van groepen in de peiling (zoals dit het geval was bij de SGP, waarvan de kiezers ook beduidend minder aan dit soort onderzoeken meededen): een manier gezocht om die ondervertegenwoordiging te compenseren. Half mei 2016 heb ik van mijn hele panel degenen een uitgebreide vragenlijst voorgelegd, die van niet Westerse herkomst waren, islamiet of van origine uit Suriname of de Nederlandse Antillen kwamen (in eerste of tweede generatie). Van dat onderzoek heb ik in mei 2016 verslag gedaan.

Dankzij deze extra aandacht voor Nederlanders van Turkse afkomst kwam DENK in mijn peiling uit op 2 zetels. Na de couppoging in Turkije en andere ontwikkelingen in het najaar nam dit aantal toe tot 3, omdat er inmiddels ook wat aanhang bij kwam uit andere groepen uit de samenleving. Inmiddels is dit nog steeds het aantal, (maar er is door mij nog geen peiling gedaan na het opstappen van Sylvana Simons).

In die eerste maanden dat DENK bij mij op 2 zetels uitkwam, meldde geen van de andere peilingen dat DENK überhaupt zetels had.

Toen ik begin oktober werd geïnterviewd in De Telegraaf en DWDD ter gelegenheid van mijn 40 jarige jubileum als opiniepeiler gaf ik desgevraagd aan wat mij onderscheidde van de andere peilers en dat betrof de manier waarop ik met DENK was omgegaan. Terwijl zij steeds nul zetels aangaven voor deze partij, door de sterke ondervertegenwoordiging van kiezers met een allochtone herkomst, zocht ik naar een manier om hier een oplossing voor te vinden. Als je immers DENK op 0 hebt staan, terwijl het wel 1 of meer zetels zouden zijn, dan heb je ten slotte ook teveel zetels bij een of meer andere partijen.

Een maand geleden publiceerde NRC een artikel over peilingen.

Daarin ook mijn uitleg over hoe ik met DENK aan de slag was gegaan. (Met daarin ook nog een cryptische beschrijving dat ik 2 jaar geleden desgevraagd over de mogelijke potentie van de twee uitgetreden Kamerleden had aangegeven dat ik aannam dat Nederlanders van Turkse herkomst een grote band met Turkije hadden, aan de hand van het televoten bij het songfestival. Iets wat trouwens onlangs bij een onderzoek van het SCP werd bevestigd, waar van de onderzochte groepen allochtonen bij de groep met een Turkse herkomst de band met het land van herkomst het grootst was).

En ja hoor de rapen waren gaar. Professor **Joop van Holsteijn** (van de categorie dat hij geen gelegenheid zal laten voorbijgaan om geringschattend over mijn werk te doen) vond dit “natte vingerwerk”. En ook andere onderzoekers hadden commentaar. En het ging zelfs zo ver dat ik in een debat belandde met **Sywert van Lienden** in DWDD – die blijkbaar uitvoerig met andere onderzoekers had gesproken -met kritiek over mijn aanpak. Een kritiek, die aangaf dat hij weinig snapte van wat een goede peiling was.

Het geheel kreeg een nog onaangenamere geur toen niet alleen bleek dat de trends die ik vanaf de verkiezing van Trump had aangegeven over de stijging van de PVV (en vervolgens een versnelling door de rechtszaak tegen Wilders) over de electorale groei van de PVV na enige tijd door de andere bureaus werd bevestigd. Maar vooral toen ook nog andere bureaus in november of december 1 of 2 zetels aan DENK gaven. Die bureaus gaven daar expliciet bij aan dat ze hun aanpak hadden veranderd om de allochtone kiezers beter te kunnen meten. Dat was nu net precies datgene wat ik bij het item in DWDD al had voorspeld, dat zowel voor wat betreft de PVV en DENK de andere bureaus hetzelfde beeld zouden gaan geven dat ik al gaf.

Ik vond dit een gênante vertoning, met Van Holsteijn en **Peter Kanne** in een dubieuze hoofdrol. Evident speelden weer de volgende drie basiselementen:

- Weinig of geen echte kennis van hoe ik mijn onderzoek doe (dus ook niet mijn verantwoording gelezen hebbend of zich verdiept in mijn aanpak).
- De problematiek bij hun eigen onderzoek niet herkennen of via de kritiek op mijn werk de eigen problemen maskeren.
- Jalousie de metier en/of gebrek aan integriteit.

4. Peilingwijzer als beste indicator van de ontwikkelingen van politieke voorkeur

Hierover heb ik deze week een apart stuk geschreven, waarbij ik uiteengezet heb dat als je over belangrijke keerpunten in de ontwikkeling van de aanhang van de verschillende politieke partijen wilt beschikken Peil.nl daar een accuraat en actueler beeld van geeft dan Peilingwijzer. U treft dat stuk hier aan.

Daarin schrijf ik het volgende over wat het doel is van Peil.nl:

Via Peil.nl wil ik vlak voor de verkiezingen een goede indruk krijgen van wat de uitslag gaat worden. Dat gebeurt dan met een prognose (met marges) voor de uiteindelijke uitslag. (Al die prognoses sinds 2003 van alle soorten verkiezingen die sindsdien zijn gehouden zijn te vinden op de site Peil.nl). In de weken ervoor kan je via Peil.nl volgen hoe de strijd tussen de partijen zich ontwikkelt. Dan is het van groot belang dat de scores van de verschillende partijen dicht bij de werkelijkheid liggen (daarmee bedoel ik de verhoudingen van dat moment onder alle Nederlanders) en dus ook de uitslag van de verkiezingen weinig verschilt van de laatste peiling.

Als er geen verkiezingen in aantocht zijn dan gebruik ik de peiling naar politieke voorkeur vooral om te laten zien wat de effecten zijn onder de kiezers van structurele ontwikkelingen in politiek en maatschappij (zoals bij voorbeeld de samenwerking van VVD en PvdA in een kabinet of de groeiende tweedeling onder hoog- en laagopgeleiden bij hun stemgedrag). En wat de electorale reactie van kiezers op bepaalde -ingrijpende- politieke gebeurtenissen is. Daarbij is het belangrijker om de verschuiving in de tijd te zien, dan dat het absolute niveau per partij, echt heel dicht ligt bij de nationale werkelijkheid.

Dat is mede zo, omdat de ervaring leert dat in de laatste weken voor verkiezingen zich forse verschuivingen voordoen. (Zo was dat bij de laatste 4 verkiezingen in de laatste 2 maanden minimaal 10 zetels bij één van de partijen en soms wel bijna 20).

Dus het doel van Peil.nl is om over de gehele periode tussen de verkiezingen de verschuivingen in voorkeur naar aanleiding van gebeurtenissen zo goed en snel mogelijk weer te geven en in de aanloop naar de verkiezingen de electorale steun per partij zo nauwkeurig mogelijk weer te geven.

Ten slotte

Uit de verdere geschiedenis kan ik nog een waslijst van voorbeelden geven over deze discussies, die vaak onder de vlag gingen van “methodisch”. Denk aan de grove onderschatting van de Politieke Barometer van de opkomst van Fortuyn in 2001-2002 en de onwil daarna om dat te erkennen. Tot en met het onjuist aangeven dat na de moord op Fortuyn de LPF steeg, terwijl die juist daalde.

Het niet onderkennen van de kracht van de PVV in 2004 en later en mij beschuldigen dat ik in het onderzoek “zelf verstopte paaseieren vond”.

Op de site van Stukroodvlees.nl is in 2013 een veelzeggend overzicht gemaakt van deze kritiek en mijn reacties.

Het was steeds hetzelfde patroon: Onder dekking van hun wetenschappelijke titel kreeg ik kritiek die één of meerdere van de door mij onderscheiden basiselementen in zich hadden. Elk van die vier zijn voor wetenschappers een doodzonde. En de criticasters van onderzoekbureaus (niet van alle trouwens, maar Reinier Heutink destijds en Peter Kanne nu spelen daar wel een hoofdrol bij en Joop van Holsteijn die blijkbaar adviseur is bij de peiling van EenVandaag) hadden daarbij duidelijk een dubbele agenda.

Ik werd uitgenodigd om bij de bijeenkomt op 12 januari “Peilingoproer” een rolletje te spelen. Niet alleen is die bijeenkomst, net zoals de vorige vergelijkbare bijeenkomst, op een moment dat ik op vakantie in Cuba ben, maar ook is de opzet niet zodanig dat de discussie die ik zou willen voeren en die ik hierboven heb beschreven, tot zijn recht komt. Daarvoor staan te veel andere -best relevante- zaken op de agenda, en heeft een aantal van de mogelijke discussianten zich in het verleden al zo misdragen, dat het geen enkele zin meer heeft een vorm van debat met ze hierover aan te gaan. Hoe kan ik dat doen als men keer op keer heeft laten zien niet te (willen) begrijpen wat ik doe of -nog erger- zich in de media gedraagt zoals Joop van Holsteijn?

Ik ben met dat soort discussies klaar. De kwaliteit, argumenten en het gebrek aan werkelijke interesse in hoe en waarom ik het heb gedaan van mensen als Foppen, Heutink, Van Praag, Krouwel, Tillie, Van der Meer, Kanne, Van Holsteijn stuit me tegen de borst.

Daarom ben ik in dit stuk uitgebreid en onderbouwd ingegaan op die kritiek, zodat iedereen vervolgens zelf zijn oordeel kan vormen.

Ik stuur dit stuk naar alle betrokkenen en zal het ook op mijn website publiceren. Ik hoop dat het gedeeld wordt met de aanwezigen tijdens het congres op 12 januari.

Ik zal proberen of ik op tijd in Nederland ben om toch aanwezig te zijn, maar ik weet niet of dat lukt.

Ik zal, net zoals ik het de afgelopen 40 jaar heb gedaan, op een gewetensvolle en onderzoeksmatig vaak innovatieve, maar zeker verantwoorde wijze, mijn onderzoeken blijven uitvoeren en via Peil.nl publiceren.

Ik wens de aanwezigen op de 12^{e} een interessant congres toe.

Drs. Maurice de Hond

Peil.nl