Kritische analyse op 'PSI bewijs' kan veel beter.

Dick J. Bierman*

Appeared as: Bierman, D.J. (1998). Discussie - Kritische analyse op 'PSI-bewijs' kan veel beter. *Nederlandsch Tijdschrift voor Psychologie*, **53**, 95-99.

De claim van Bem & Honorton

In een opzienbarend artikel in *Psychological Bulletin* rapporteerden Bem & Honorton in 1994 de uitkomsten van een serie replicatie-onderzoeken naar telepathie of 'anomalous information transfer' zoals zij dat noemden (Bem & Honorton,1994). Een proefpersoon wordt via een inductieprocedure in een licht veranderde bewustzijnstoestand gebracht waarin het veronderstelde paranormale proces beter zou verlopen. Tijdens de zitting zit een andere persoon op een geïsoleere plek naar een willekeurig gekozen doel-plaat of doel-clip te kijken. Aan het einde van de sessie moet de proefpersoon uit een set van vier mogelijke doelen het juiste doel aanwijzen. Het materiaal dat bij deze experimenten werd gebruikt bestond uit 160 mogelijke doelen genummerd van 1 tot en met 160. Deze doelen zijn georganiseerd in sets van vier. Dus de sets zijn {1,2,3,4}, {5,6,7,8},...{157,158,159,160}.

Een selectieprocedure genereert een willekeurig getal in dit interval. Laten we zeggen dat hieruit het doel '7' volgt. De gebruikte beoordelingsset is dan logischerwijze de set {5,6,7,8}. Om een raadkans van 25% te rechtvaardigen moeten de waarschijnlijkheden waarmee de vier mogelijke doelen binnen 1 set voorkomen gelijk zijn.

In 329 sessies werden 106 treffers geboekt. Dat is 24 meer dan je theoretisch mag verwachten en het scoringspercentage van 32% (106 / 329) ligt zeer dicht bij de eerder uit een meta-analyse bepaalde waarde.

In hun artikel laten Bem & Honorton ook zien dat de proefpersonen voorkeuren hebben voor bepaalde platen of clips. Ze bespreken de mogelijkheid dat dit, tezamen met een vertekening in de selectie van doelen, geleid kan hebben tot de extra treffers. Zo'n vertekening kan voorkomen omdat de selectie procedure niet geheel correct is verlopen maar eveneens omdat nu eenmaal eens in de zoveel keren een afwijkende reeks voorkomt en sowieso het 'ideale' geval van precies gelijke frequenties voor alle alternatieven zeer onwaarschijnlijk is. Ze berekenden daarom een `a posteriori raadkans die rekening houdt met de in de experimenten waargenomen selectie- en response-frequenties. Zelfs na toepassing van deze conservatieve procedure claimen Bem & Honorton dat de gevonden 24 extra treffers niet aan toeval zijn toe te schrijven.

De kritiek van Van den Brink

Terecht merkt Van den Brink in zijn kritische analyse in dit tijdschrift op dat de selectie van 1 uit een set van vier mogelijke doelen in de betreffende experimenten een kritisch methodologisch onderdeel vormt (Van den Brink, 1997). Vervolgens toont Van den Brink dat de selectie van doelen niet zo 'random' was als je zou mogen hopen.

Tenslotte stelt hij dat de correctie die Bem & Honorton hiervoor toepassen volstrekt 'zinloos' is. Daarmee zouden, volgens Van den Brink, de resultaten voor een groot deel door een correlatie tussen doelfrequenties en voorkeuren verklaard kunnen worden.

De verdere uitwerking en weerlegging van de conclusie van Van den Brink

In het onderstaande zal ik in de eerste plaats laten zien dat een directe analyse van doelfrequenties van de vier mogelijke doelen binnen 1 set geen randomisatie probleem laat zien welke zou hebben kunnen leiden tot de correlatie van doelfrequenties en voorkeuren. Hierdoor is een correctieprocedure eigenlijk niet nodig.

Niettemin zal ik aantonen dat de door Van den Brink gewraakte correctieprocedure wel degelijk correct is. Toepassing van deze 'conservatieve' procedure kan niet een groot deel maar ongeveer 10% van het effect verklaren.

In de conclusie zal ik vooral ingaan op de rol van de kritische methodoloog in experimenteel (para)psychologisch onderzoek.

Bierman, D.J. (1998). Discussie - Kritische analyse op 'PSI-bewijs' kan veel beter. *Nederlandsch Tijdschrift voor Psychologie*, **53**, 95-99.

Vertekeningen in de selectie van doelen

Van den Brink volgt de mening van Bem & Honorton waar zij stellen dat een chi^2 toets ter beoordeling van het aselecte karakter van het doelselectie proces in de range [1,160] niet adequaat is vanwege te kleine aantallen selecties per doel. Dus kiest Van den Brink noodgedwongen een andere weg. Hij laat zien dat, als je op een indirecte manier naar de uitkomsten van de doelselectie kijkt, er afwijkingen te zien zijn van een nette willekeurige uitkomst. Hij toetst namelijk de verdeling van het aantal doelen met een bepaalde selektiefrequentie. Hij vindt dan een onder-representatie van doelen die drie keer gekozen zijn. Dit maakt de selectie verdacht.
Maar, de geconstateerde afwijking van randomness **over** alle doelsets heen hoeft allerminst te wijzen op een bias voor de doelen **binnen** de sets. En een vertekening in de doelselectie **binnen** de sets, tezamen met voorkeuren voor bepaalde platen binnen de sets, is nodig om tot een mogelijke 'normale' verklaring van Bem & Honortons resultaten te komen.
Vanwege het **indirecte** karakter is Van den Brinks manier van toetsen niet de meest logische. Een meer voor de hand liggende directe toetsing van de waarschijnlijkheden waarmee de doelen **binnen** de set worden gekozen is om direct te kijken naar hoe vaak het eerste doel, het tweede doel, het derde doel en het vierde doel binnen de set gekozen werden. Ik vind dan:

nr. binnen set	frequentie	theoretische frequentie
eerste	75	82.25
tweede	84	82.25
derde	88	82.25
vierde	82	82.25

chi2 = 1.08 ; df= 3; p= 0.80.

De conclusie uit deze analyse is dat er niets, maar dan ook absoluut niets, verdachts met de selectie van doelen binnen één set aan de hand is. Alleen op grond van deze eenvoudige analyse al kunnen we verwachten dat Van den Brinks voorgestelde model voor een normale verklaring van de extra treffers nooit meer kan opleveren dan een zeer kleine correctie.

Het is overigens wel duidelijk dat de selectieprocedure heeft geleid tot een afwijkende frequentieverdeling van gebruikte sets. Van den Brink besteedt hier aandacht aan bij de bespreking van de oververtegenwoordiging van set 20 (doelen 77.78.79.80). Maar naast de zeer extreme oververtegenwoordiging van set 20 zijn eigenlijk alle sets van 1-20 oververtegenwoordigd. Met andere woorden: de doelen van 1-80 zijn tamelijk systematisch over- en die van 81 tot 160 ondervertegenwoordigd. Elders is geprobeerd een verklaring voor deze achteraf geconstateerde afwijking te vinden. Maar zelfs analyse van de gebruikte software kon geen verheldering brengen (Bierman et al, 1998).

Deze afwijkende verdeling van gebruikte sets is op zich consistent met de door Van den Brink aangetoonde verschillen in doel frequenties. Maar de essentie is of de platen/clips die vaker doel waren ook de aantrekkelijkste doelen waren. Wie zal zeggen of het omgekeerde niet het geval was? Alleen zo'n correlatie kan het aantal treffers systematisch beïnvloeden. Een correlatie kan men nimmer afleiden uit een univariaat gegeven zoals de doelfrequenties; die moet worden uitgerekend. Ik heb daarom per set de relatieve keuze-frequenties bepaald voor elk van de vier platen. Omdat dit gebaseerd is op weinig gevallen zal deze maat voor de aantrekkelijkheid van de platen niet zeer betrouwbaar zijn. Daarna heb ik deze aantrekkelijkheidsscores gecorreleerd met de doelfrequentie. De productmoment correlatie blijkt 0.078 (N= 160; p = 0.328). Dit laat alweer zien dat er vooralsnog weinig grond is voor het idee dat Bem & Honortons resultaten het gevolg zijn van zo'n correlatie. Maar zoals we hieronder zullen zien is een preciezere schatting mogelijk.

Bierman, D.J. (1998). Discussie - Kritische analyse op 'PSI-bewijs' kan veel beter. *Nederlandsch Tijdschrift voor Psychologie*, **53**, 95-99.

Aantrekkelijkheidsverschillen

Hoe zit het dan met de verschillen in aantrekkelijkheid van de verschillende platen? Die zijn nu eenmaal niet te vermijden en de Bem & Honorton bespreken deze dan ook uitgebreid met name in het kader van een serie waarbij maar één set doelen werd gebruikt (serie 302: doelen 87, 88, 89 en 90). Het bleek hierbij overduidelijk dat proefpersonen een sterke voorkeur hadden voor bepaalde doelen. Gevoegd bij de verdeling van doelen in deze serie 302 berekenen ze een actuele raadkans van 34% i.p.v. de theoretische 25%. De actuele of a posteriori raadkans is een raadkans die wordt berekend aan de hand van de (schatting van) de voorkeuren en de echte frequenties waarmee de doelen in het experiment voorkomen. Ter illustratie een voorbeeld:

We vragen een paragnost te voorspellen welke kleur er in 10 opeenvolgende roulettepogingen zal opkomen. Hij voorspelt 9 maal rood en 1 maal zwart. De roulette produceert ook 9 maal rood en 1 maal zwart en de paragnost heeft alle 10 goed. De actuele raadkans is dan: relatieve frequentie van 'doel rood' * 'voorkeur rood' + relatieve frequentie 'zwart' * 'voorkeur zwart' oftewel 0.9 * 0.9 + 0.1 * 0.1 = 0.811. De reden voor deze hoge raadkans is dat de nogal opvallende voorkeur van de paragnost overeenkomt met een wat afwijkende trekking van rood en zwart. Zou de paragnost bij de theoretisch kansverwachting van een 50%, vijf extra treffers hebben gescoord bij toepassing van de gecorrigeerde actuele raadkans zijn dat nog geen twee extra treffers. Het uitgaan van actuele raadkansen i.p.v theoretische is dan ook conservatief t.o.v. de psi-hypothese.

De bevinding dat binnen set 20 een actuele raadkans is van 34%, gebruikt vd Brink overigens op een incorrecte manier. Ik citeer:

"....Alleen op grond hiervan [bedoeld wordt dat set 20 veel vaker dan volgens kans in alle series bij elkaar gebruikt is; djb] zou al verwacht kunnen worden dat het aantal onder toeval verwachte successessen in de tien autoganzfeld experimenten hoger ligt dan 25% waartegen het gevonde succespercentage van 32 is afgezet. *Bem en Honorton vonden immers op grond van de resultaten in experiment 302 dat het onder toeval verwachte percentage successen voor set 20 niet gelijk is aan 25 maar aan 34%.* (mijn cursivering; djb)....."

Ofwel Van den Brink begrijpt het onderliggende model voor die praktische raadkans van 34% niet, en dat kan ik me niet voorstellen, ofwel dit een wel zeer **discussiabele** wending. Immers die 34% is niet een eigenschap van de set. Die 34% komt uit de combinatie van actuele doelselectie en de voorkeuren van de proefpersonen. Die voorkeuren zijn, zo mogen we aannemen, een eigenschap van de combinatie van de set en de proefpersoonpopulatie. Maar de actuele keuzes van elk der mogelijke doelen is voor elke serie anders en daarmee de actuele raadkans.

Van den Brink kan heel makkelijk nagaan wat voor die tien series de actuele raadkans op set 20 is geweest. Op de procedure om deze raadkans uit te rekenen heeft hij geen kritiek en hij beschikte over de noodzakelijke gegevens. Ik neem ook aan dat hij dat voor set 20 heeft gedaan, hij refereert er zelfs naar in zijn eigen verhaal, maar dat de uitkomst hem niet beviel. Immers voor die tien experimenten tezamen blijkt de actuele raadkans op de 23 sessies met set 20: 28% (tewijl het scoringspercentage in die sessies 43.5 % is!). M.a.w. de veronderstelde combinatie van voorkeuren en doelselectie bias waarvan Van den Brink claimt dat ze een groot deel van de totale resultaten zou kunnen verklaren, verklaart hier welgeteld 0.9 hit! Van den Brink schrijft in zijn conclusie: "..... het valt niet uit te sluiten dat minstens een deel van aangetoonde PSI-effect toegeschreven kan worden aan interactie tussen responsevoorkeuren en ongelijke doelstimulus frequenties. *Met name de ongewoon hoge frequentie waarmee set 20 in de tien autoganzfeld experimenten voorkomt steunt deze zienswijze.* (mijn cursivering)..... " Inderdaad 0.9 hit van de 24 extra hits zijn hiermee te verklaren. Nog geen 4% dus.

Bem's raadkans berekening over alle series is wel degelijk zinnig

De procedure om de actuele raadkans uit te rekenen die ik hierboven toepas op die ene set 20 kan mutatis mutandis worden uitgevoerd voor het hele experiment hoewel dat volgens Van den Brink 'zinloos' zou zijn vanwege de te lage frequenties waarmee de verschillende doelen voorkomen. Mijns inziens betekenen die lage frequenties alleen dat de berekende raadkans onbetrouwbaar geschat is maar de uitkomst is wel degelijk de meest waarschijnlijke

schatting. Het kan zijn dat Van den Brink zo 'op het oog' ziet dat het betrouwbaarheidsinterval bij dergelijke frequenties zo groot zal zijn dat je met recht het predikaat 'zinloos' mag gebruiken. Dat reviewers van een toptijdschrift 'zinloze' analyses laten passeren is inderdaad iets waar je je als methodoloog over kunt opwinden. Maar ik wind me ook wel een beetje op over het gebruik van 'op het oog' evaluaties. Veel beter is natuurlijk om dat betrouwbaarheidsinterval uit te rekenen. Dat is analytisch misschien moeilijk maar ontzettend simpel met Monte Carlo methoden. Ik heb dus maar eens 5000 maal het hele experiment gesimuleerd en naar de verdeling van die actuele raadkansen gekeken. En wat blijkt? De reviewers van 'Psychological Bulletin' waren kennelijk toch niet zo achterlijk. De gemiddelde berekende actuele raadkans uit de Monte Carlo simulatie is 0.250. Dat is geen openbaring maar de standaard deviatie is slechts 0.008! En wat misschien nog wel tekenender is: de minimale en maximale raadkans in 5000 simulaties zijn 0.221 en 0.284 en de verdeling is in alle opzichten volstrekt normaal. De conclusie moet zijn dat Bem de 'correctieprocedure' waarbij de theoretische raadkans door de actuele raadkans wordt vervangen terecht heeft uitgevoerd.
De uitkomst van de correctieprocedure voor response- en targetbias voor alle 10 series tezamen is dat de meest waarschijnlijke actuele raadkans iets kleiner dan 26% is. Indien rekening wordt gehouden met response biases die niet worden gedreven door aantrekkelijkheid van de targets maar door de positie die de targets innamen in de beoordelingsvolgorde dan wordt die actuele raadkans zelfs nog iets kleiner en kan hooguit 10% van de extra treffers verklaren.

Conclusie

Laat ik met nadruk stellen dat ik de conclusie van Van den Brink deel dat onafhankelijke replicaties nodig zijn om meer zekerheid over de psi-hypothese te krijgen. Ook een kritische analyse van de reeds geproduceerde data is absoluut nodig, reden waarom ik hem in mei 1995 de ruwe data op sessie-nivo schriftelijk aanbood. Het betreft hier uiteindelijk geclaimde verschijnselen die een grote verandering in het algemeen aanvaarde wetenschappelijke wereldbeeld kunnen brengen hoewel de claim van Van den Brink dat deze verschijnselen in tegenspraak zouden zijn met natuurkundige processen moet berusten op een klassiek Newtoniaanse misvatting.
Ook de constatering van de merkwaardige afwijkingen van frequentieverdeling van doelsets die wijst op problemen in de randomisatie is volledig terecht. Maar zijn stelling dat dit een belangrijk deel van de resultaten zou verklaren is onterecht.
Ik constateer samenvattend de volgende divergenties met Van den Brink, namelijk:

a) Dat een directe analyse van de frequentieverdeling waarmee de doelen binnen sets worden gekozen laat zien dat binnen de set de procedure volledig aselect was.
b) Dat actuele raadkans van set 20 binnen de 329 sessies geen 34% maar 28% is en dat dit op zich minder dan 4% van het totale gevonden effect kan verklaren.
c) Dat de 'correctie-procedure' die Bem toepast niet 'zinloos' is zoals Van den Brink beweert maar zinnig. De reviewers van Psychological Bulletin hebben hier niet gefaald maar zouden dat wel gedaan hebben als ze de opmerking dat deze procedure 'zinloos' is hadden laten passeren.
d) Toepassing van deze procedure laat zien dat de meest waarschijnlijke over-all correctie rekening houdend met response bias voor de targets zelf en voor de plaats van de targets in de beoordelings sekwentie ongeveer 2.3 hits bedraagt oftewel 10% van het totale effect.

Het diepere probleem dat ik heb met de wijze waarop Van den Brink de kritische analyse doet is dat hij een ‘normaal’ model postuleert en dan ‘circumstantial evidence’ aanvoert om dat model kwalitatief te onderbouwen terwijl directe kwantitatieve analyses mogelijk zijn. Een typisch voorbeeld van kwalitatief redeneren waar het ook kwantitatief kan valt te zien in de secundaire argumenten van Van den Brink. Hierbij wordt ‘optional stopping’ als mogelijk reden voor extra ‘normale’ hits genoemd. Zoals genoegzaam bekend ben ik geen statisticus maar er zijn dacht ik prachtige kwantitatieve mogelijkheden om dit effect gewoon te schatten.

Is het niet de plicht van de methodoloog om ‘alternatieve modellen’ zo ver mogelijk door te rekenen en het niet te laten bij opmerkingen als dat het splitsen van 56 sessies het

met voeten treden van methodologische regels behelst? Methodologische regels zijn zinvol in zoverre ze ons behoeden voor het trekken van verkeerde conclusies. Deze subjectieve splitsing doet echter niets toe of af aan de toetsing of het totaal resultaat aan toeval te wijten is en daar gaat het hier met name om.
Ik denk dat het niet de taak van een methodoloog is om experimenten verdacht te maken. Het is zijn taak zo nauwkeurig als maar kwantitatief mogelijk is de kans te schatten dat een onterechte conclusie wordt getrokken.

In de eeuw serieus onderzoek naar claims als telepathie is evenveel mankracht gaan zitten als in drie maanden regulier psychologisch onderzoek hedentendage. Mede daarom is het juist om zuinig en kritisch met de data om te gaan. Ik waardeer het werk dat Van den Brink in deze kritische analyse heeft gestopt. Er zijn te veel mensen die experimentele data, die niet overeenkomen met hun wereldbeeld, zonder enige analyse als onbetrouwbaar terzijde schuiven. Net zoals er veel mensen zij die een resultaat dat wel overeenkomt met hun verwachtingen klakkeloos aanvaarden. Een kritische analyse is dus altijd welkom maar in dit geval had de uitvoering beter gekund.

Referenties

Bem, D.J. & Honorton, C. (1994). Does PSI exist? Replicable evidence for an anomalous process of information transfer. *Psychological Bulletin, 115,* 4-18.

Bierman, D.J. , Berger, R.E. & Broughton, R. Notes on Random Target selection: The PRL Autoganzfeld targt - and targetset distributions revisited. *Journal of Parapsychology (in press).*

Brink, W.P. van den (1997). Repliceerbaar bewijs voor PSI: een kritische analyse. *Nederlands Tijdschrift voor Psychologie, 52 (6),* 249-254.

* Universiteit van Amsterdam, Faculteit Psychologie, vakgroep Psychonomie, Roetersstraat 15, 1018 WB Amsterdam. e-mail: bierman@psy.uva.nl.