Opmerkingen en overdenkingen over ‘validiteit’ en ‘betrouwbaarheid’

In de wetenschap wordt een test pas ‘goed’ gevonden als hij betrouwbaar en valide is. Een test wordt **betrouwbaar** geacht indien de gemeten eigenschappen op verschillende meetmomenten een vergelijkbare/overeenkomstige uitslag geven. Een test wordt **valide** geacht als het gedrag van de hond tijdens de test overeenkomt met zijn gedrag in een niet-test situatie, zoals thuis in zijn eigen omgeving. Daarom wilde ik absoluut voordat de test in het fokreglement zou komen, eerst 40 honden testen en proberen uit te zoeken of de test inderdaad valide en betrouwbaar was. Dit deed ik in 2006/2007.

De validiteit vaststellen is niet bij iedere test even gemakkelijk. Wanneer de test eisen aan een hond stelt die ver buiten zijn normale leven vallen, vertoont de hond vaak gedrag (bijvoorbeeld agressie, hevige stress of grote angst) dat hij in het dagelijks gebruik in de vertrouwde omgeving niet vertoont. Je krijgt dan zogenaamde ‘vals positieven’ (honden die in de test bv bijten, maar in het normale leven niet). Het omgekeerde kan ook overigens (vals negatieven: de hond bijt in het normale leven, maar niet in de test). De eigenaar zegt dan in beide gevallen het gedrag van de hond in de test niet te herkennen.
De validiteit van een test meten via vragenlijsten aan de eigenaar is vaak moeilijk, zo komt in allerlei onderzoek naar voren. Omdat mensen doorgaans van hun honden houden, hebben ze vaak een (onbewust) positief gekleurd beeld van hun hond en komt het oordeel van de eigenaar en die van de gedragsbeoordelaars niet overeen. Eigenaren zijn bovendien vaak zo gewend aan bepaalde gedragingen dat ze die afzwakken of niet kenbaar maken als naar meer ‘lastig’ of ‘storend’ gedrag wordt gevraagd. Ook als er wat belangrijks voor de eigenaar afhangt van de test (bv een positieve beoordeling is nodig om te mogen fokken), geeft hij een positiever gekleurd beeld. Ook in de vragenlijsten van de Retrievertest (die niet voor validatie bestemd waren, maar om de Big Four van de honden vast te stellen) kwamen overwegend naar voren dat mensen erg blij waren met hun hond, ook als de beoordelaars dachten dat bepaald gedrag van de honden wel eens lastig in het gebruik zou kunnen zijn.
Omdat de Retrievertest zo heel dicht stond bij situaties die eigenaars en honden ook uit de dagelijkse praktijk kenden, zeiden vrijwel alle eigenaars na afloop van de test (en vóór het ontvangen van het eindrapport) dat de hond zich gedroeg als altijd en/of dat ze het gedrag uit situaties thuis herkenden. Met de validiteit leek het dus wel goed te zitten.

Er zijn maar weinig gedragstesten die volgens wetenschappelijke normen betrouwbaar zijn. Het probleem is meestal dat in de praktijk om organisatorische redenen testen niet herhaald kunnen worden (wat nodig is om de betrouwbaarheid vast te stellen); meestal omdat dit te duur is en de testende organisatie of deelnemers niet bereid zijn om nogmaals een vrij hoog bedrag neer te leggen voor een herhaling. (De Retrievertest kost bijvoorbeeld per hond €85). Of omdat men geen zin heeft weer een eind te reizen. Men ziet er vaak het nut niet van in omdat men immers al de uitslag heeft. Wetenschap staat niet het hoogst op het prioriteitenlijstje.
Dat de test zo duur is, komt enerzijds doordat er veel mensen nodig zijn bij de test (die meestal minimaal reiskosten kunnen declareren), er een testterrein gehuurd en een gedragsbeoordelaar betaald moet worden, anderzijds doordat er vaak maar relatief weinig honden op een dag getest kunnen worden. De Retrievertest neemt bijvoorbeeld een uur per hond in beslag (een half uur voor de test zelf en een half uur voor de opstelling van het eindrapport en de nabespreking met de eigenaar), zodat er maar 6 of 7 honden op een dag getest kunnen worden. Meer honden is ook voor de beoordelaars/testers niet wenselijk, omdat het zeer intensief en geestelijk belastend werk is.

 September 2017

De betrouwbaarheid van de Retrievertest vast proberen te stellen was dan ook veel ingewikkelder. Er was geen enkel animo om honden te laten hertesten, zelfs als dit gratis kon. Ik heb wel een paar keer een hond kunnen hertesten op de Retrievertest, maar dat waren er maar een paar, waaronder die van mezelf. Deze honden gedroegen zich overigens vrijwel hetzelfde (zie onder 'hergeteste honden' voor de bevindingen). Daarom zocht ik mijn toevlucht in vergelijking met de andere testen van de FRC: de puppytest en de fokdagtest. Deze had ik in 2005 herzien zodat ze dezelfde opbouw, dezelfde beoordelingswijze en dezelfde (maar minder) testonderdelen hadden.
De puppytest had als zelfde testonderdelen:

1. Betreden en onderzoeken vreemde plaats (buiten) met onbekende mensen.
2. Lokken door tester
3. Vastpakken stilliggende voorwerpen
4. Neusgebruik (pens)
5. Apporteren van 4 verschillende voorwerpen. (Sok, pluche bal, konijnenvel, eendenveer)
6. Doorzettingsvermogen/probleemoplossend vermogen en apporteren
7. Akoestische prikkel
8. Visuele prikkel
9. Correctie
10. Apporteren van de 4 voorwerpen
11. Gedrag t.o.v. 'nephond'

De Fokdagtest had:

1. Betreden vreemde ruimte met onbekende mensen en (on)bekende honden
2. Lokken door tester
3. Apporteren van vel
4. Neusgebruik (eend)
5. Schot
6. Apporteren van vel

Bij de 16 honden die ook de puppytest en fokdagtest hadden gedaan bleek dat bepaald gedrag (apporteren en sociale omgang met mensen) constant was, maar zeker niet álle gedrag. (Voor een uitgebreide bespreking zie onder). Inmiddels is uit voortdurende, handmatige vergelijkingen de afgelopen 10 jaar tussen verschillende testen wel duidelijk dat de score van bepaalde eigenschappen constant zijn. Dat is dus een bewijs voor betrouwbaarheid van bepaalde testonderdelen. Maar niet van de test als geheel.

In de wetenschappelijke literatuur over gedragstesten wordt eigenlijk altijd gesproken over de betrouwbaarheid en validiteit van een hele test. Wat onvoldoende uit de verf komt is echter dat een test meestal uit diverse, van elkaar verschillende testonderdelen bestaat met een veelheid aan prikkels, situaties en interacties die verschillende emoties oproepen bij zowel mens als dier.

Bij het analyseren aan de hand van scoreformulieren en videobeelden welke onderdelen van de test betrouwbaar waren, in de zin dat de honden in verschillende testen (puppytest, fokdagtest, retrievertest) hetzelfde gedrag bij een overeenkomstig testonderdeel lieten zien,

viel op dat dit alleen gebeurde als de hond daar geen van de negatieve emoties bij had die te herleiden zijn op Panksepps drie negatieve emotionele hersensystemen: FEAR, GRIEF en RAGE. Als de hond neutraal of positief gestemd was bij aanvang van het testonderdeel dan steeg de betrouwbaarheidsscore naar grote hoogte. Wanneer een hond echter onzeker of angstig was (FEAR), last had van het verlies van steun van een vertrouwde omgeving, eigenaar of andere honden uit het gezin (GRIEF) of geïrriteerd of gefrustreerd was (RAGE), dan zag je in de verschillende testen veel meer wisselend gedrag.

Negatieve emoties overschaduwen naar mijn mening dus veel gedragingen die men met een gedragstest in kaart wil brengen dusdanig, dat ze zorgen voor een hoge onbetrouwbaarheidsscore bij het valideren en de betrouwbaarheid meten van de testen zelf, maar vaak ook voor vertekening in de individuele beoordeling van een hond.

Dit is iets dat in de wetenschappelijke literatuur over de betrouwbaarheid en validiteit van gedragstesten regelmatig wordt gesignaleerd, bijvoorbeeld bij angstige asielhonden die na herplaatsing pas na enkele weken doorgebracht te hebben in het nieuwe gezin hun ‘ware gezicht’ in negatieve of positieve zin (durven te) laten zien. Of het hoge percentage vals positieven en vals negatieven bij agressietesten.

Dat negatieve emoties ander gedag overschaduwen en vertekenen (en positieve emoties niet of veel minder) verklaart ook waarom apporteren wel een hoge voorspellende waarde heeft: apporteren is sociaal spel en een angstige hond speelt niet. Een hond die wel speelt met een vreemde, heeft vertrouwen en laat dus eerder zijn ‘gewone’ doorsnee sociale gedrag zien. Het verklaart ook waarom pups en jonge honden die bij een test op onbekend terrein als ‘angstig’ worden gescoord dat in hun eigen vertrouwde omgeving helemaal niet zijn en vaak ook helemaal niet uitgroeien tot angstige volwassen honden: het verlies aan veiligheid in de onbekende omgeving (GRIEF) vertekent hun ‘normale’ gedrag.

Wat naar mijn mening bij statistische onderzoeken naar betrouwbaarheid en validiteit van een test in de weg kan zitten is dat:
- De test als een totaal wordt onderzocht en niet op afzonderlijke onderdelen.
- Negatieve emoties (zoals angst) in bepaalde testonderdelen niet worden gescoord, terwijl ze in de eerste test niet (geen vertekend beeld) en in de hertest wel aanwezig waren (vertekend beeld). Zoiets kan gemakkelijk gebeuren bij een hertest met een andere tester, die bv door zijn uiterlijk de hond onzeker maakt. Datzelfde kan gelden voor de validiteit. In de test is de hond angstig (bv omdat hij met een kunsthand wordt geaaid) maar in de thuissituatie is hij dat niet omdat hij gewoon met de hand wordt geaaid.

Iets dergelijks kan gebeuren als men in testonderdelen die bijvoorbeeld meten of de hond wil spelen, niet aantekent dat negatieve emoties (bv GRIEF of FEAR) de hond belemmeren om te spelen zodat men niet kan beoordelen of de hond wil spelen. Vaak wordt alleen aangetekend dat de hond niet speelt: hij scoort dan negatief op speelsheid.
Dit laatste gebeurt relatief vaak, omdat de scores zijn afgestemd op wat het testonderdeel moet meten (bv vriendelijkheid, speelsheid).

Wat hier naar mijn mening achter schuilgaat is vaak de premisse dat het gedrag van een dier tot op grote hoogte constant en voorspelbaar is, mits men maar de ‘juiste’ correct uitgevoerde gestandaardiseerde prikkel aanbiedt: als je een hond op correcte wijze uitnodigt tot spel of vriendelijk benadert, kun je zijn ‘speelsheid’ meten.

Als een dier dus bij de hertest heel wat anders doet (omdat hij zich niet prettig voelt), wordt al snel geconcludeerd dat de betrouwbaarheid van de test niet goed is. Dat lijkt me dus niet altijd per definitie terechte conclusie. Er zal naar mijn mening veel beter gekeken en geregistreerd moeten worden welke emoties in het spel zijn. Dat is geen eenvoudige zaak omdat het nog relatief nieuw is en men de expressie van wat verschillende emotionele hersensystemen zijn, niet voldoende (her)kent en in de beoordelingen vaak op een grote hoop gooit(bv angst).

Een andere premisse is, dat een test voorspellende waarde moet hebben. Het idee dat gedrag voorspelbaar kán zijn, is ontstaan in een tijd dat gangbare wetenschapsopvatting over dieren was, dat zij geen bewuste emoties hadden. Daaruit volgde dat men niet onderkende dat een groot deel van het gedrag gestuurd/gemotiveerd werd door emoties (zoals nu door hersenonderzoek van o.a. Panksepp is vastgesteld). Gedrag werd beschouwd als een (tamelijk eenvormige) reactie op een externe prikkel. Men dacht daarom aanvankelijk dat als alles gestandaardiseerd is en de prikkels die worden toegediend ‘dus’ hetzelfde zijn, de reactie ook van het dier hierop ook hetzelfde zal zijn. Dit verklaart ook dat de overschaduwende negatieve emoties niet meegewogen worden. Een misvatting dus naar mijn mening en ervaring (Zie over ‘’standaardisatie’ meer hieronder).