FORSVARSAKADEMIET
Institut for Militærpsykologi

50 ÅR MED

INTELLIGENSPRØVEN

BPP

Erik Kousgaard

NOV 2003

INDHOLD

INDLEDNING

Den 1. marts 1952 oprettede Forsvarsministeriet institutionen Militærpsykologisk Arbejdsgruppe (MPA), og selvom den flere gange har skiftet navn - over Militærpsykologisk Tjeneste (MPT), Psykologisk Afdeling (PSA) til det nuværende Institut for Militærpsykologi (IMP) - var den endnu intakt 50 år efter og varetog alle psykologiske opgaver i forsvaret. En af gratulanterne ved 50 års jubilæet i foråret 2002 var den nu pensionerede afdelingsleder og konstituerede direktør for Danmarks Pædagogiske Universitet, cand. psych. Børge Prien, som ved receptionen spurgte mig, om intelligensprøven BPP stadig fungerede og blev anvendt på sessionerne. Når man ved, at BPP står for Børge Priens Prøve, var hans interesse meget forståelig. Jeg kunne fortælle, at prøven stadig blev anvendt og i praksis fungerede upåklageligt, men jeg måtte samtidig erkende, at IMP nok havde forsømt at kontrollere, om den struktur, der fandtes i besvarelsen af delprøvernes enkeltspørgsmål, stadig eksisterede. Den nemme forklaring på denne undladelse er manglende tid, men der er en anden forklaring, som nok er mere rigtig. Da prøven blev udviklet, udførte Børge Prien et så grundigt arbejde såvel ved konstruktion som ved afprøvning, at der gennem årene har eksisteret en grundfæstet tillid i IMP til prøvens stabilitet. Og hvorfor bruge tid på at kontrollere noget, man ikke nærer tvivl om?

Konsekvensen af samtalen ved jubilæet blev, at IMP hen over sommeren besluttede at råde bod på det forsømte. Prøvehæfter fra hen ved 3000 sessionsbehandlede fra efteråret 2001 og foråret 2002 blev udvalgt, så de var nogenlunde repræsentative for en årgang, og de enkelte svar på BPP's 76 spørgsmål blev registreret. Hensigten var bl.a. at sammenligne med de resultater Børge Prien havde opnået på konstruktionstidspunktet, og det var derfor nødvendigt at søge i arkiverne efter originalmateriale. Herved blev det opklaret, at prøven blev taget i rutinemæssig brug tidligere, end vi antog. Misforståelsen var opstået, fordi prøven betegnes BPP-57, men der fandtes en ældre udgave BPP-53 (i folioformat!), som blev anvendt fra november 1953 som rekrutprøve på nyindkaldte værnepligtige bl.a. med henblik på udtagelse af sergentelever. I 1956 blev det besluttet, at prøven skulle flyttes til sessionstidspunktet, så den kunne blive en del af udskrivningsgrundlaget. Det skete ved efterårssessionen 1956, og på grundlag af de indsamlede besvarelser blev der ændret lidt på rækkefølgen af nogle opgaver ud fra et ønske om, at opgaverne skulle have stigen-

de sværhedsgrad. Herved opstod betegnelsen BPP-57, som er identisk med BPP-53 bortset fra, at der er byttet lidt om på opgavernes rækkefølge.

Prøven har altså været i konstant brug i 50 år (se side 15ff om anvendelsen i forsvaret), og IMP fandt det derfor rimeligt at markere jubilæet med dette lille skrift. Næsten alle danske mænd mellem 18-19 år og 70 (og i de senere år også en del kvinder) har udfyldt prøven, så det bliver til et antal mellem 1.5 og 2 mil. BPP er derved uden sammenligning den mest succesfulde danske intelligensprøve, og det siger meget om dens gennemtænkte konstruktion, at den ikke er blevet "slidt" op. Sessionernes omhyggelige administration har været med til at sikre, at opgaverne er forblevet ukendte i offentligheden. En så udbredt brug gennem 50 år kunne ellers let have ført til, at forhåndsviden om prøven havde gjort den værdiløs.

KONSTRUKTION

Psykologisk Laboratorium ved Københavns Universitet havde før oprettelsen af MPA assisteret ved udvælgelsen af forsvarets piloter. Hertil havde man anvendt en intelligensprøve (IGP) efter svensk forbillede. MPA anså det fra starten som en vigtig opgave at få udviklet en dansk intelligensprøve, som kunne danne rygraden i de forskellige udvælgelsesopgaver, som institutionen blev pålagt. Cand.psych. Børge Prien fik overdraget opgaven, som han løste ved at konstruere fire delprøver, der hver består af ret ensartede opgaver. Når denne løsning blev valgt frem for en omnibustest med blandede opgaver, hang det utvivlsomt sammen med, at dr. Georg Rasch - senere professor i statistik ved Københavns Universitet - var statistisk konsulent ved MPA i institutionens første år. Rasch havde i nogle år arbejdet med at formulere et alternativ til den faktoranalytiske behandling af psykologiske tests, og han vidste formentlig allerede på dette tidspunkt, at kun beslægtede opgaver ville kunne leve op til de objektivitetskrav, som han senere formulerede.

Børge Prien har i det efterfølgende notat ud fra sin erindring gjort rede for det udgangspunkt han i foråret 1953 havde for testens konstruktion.

Børge Priens notat fra juli 2003.

Den intelligensprøve, der var brugt af Psykologisk Laboratorium ved udvælgelsen af pilotaspiranter, viste sig uegnet på et bredere udsnit af befolk-

ningen, idet løsningen af opgaverne i høj grad beroede på læsefærdighed. Det var nemlig stort set tekstopgaver.

Nogle kendte udenlandske intelligensprøver var derimod alene baseret på meningsløse figurer, hvilket var en fordel, hvis man ville adskille opgaveløsningsfærdighed fra indlært viden, men disse opgaver blev ensidige på den måde, at personer uden veludviklet visuel forestillingsevne kom til at virke dumme, ligesom de dårlige eller uøvede læsere gjorde det i IGP.

Ønsket blev at fremstille en prøve, bestående af dele, der foruden fælles almentræk kun i begrænset omfang stillede krav til hver sine specielle færdigheder (f.eks. læsefærdighed eller talforståelse), så ingen af disse kom til at dominere totalresultatet. Samtidig skulle alle prøver så vidt muligt være uafhængige af uddannelsesbetinget viden, såsom kendskab til fremmedord eller regning ud over folkeskolens pensum.

Endvidere var det et krav, at prøvetiden ikke måtte være stort mere end en time. Det skyldtes dels praktiske hensyn ved prøvetagningen, dels at mange mennesker begynder at blive trætte, og dermed dårligere ydende, efter 1 times hjernearbejde. Derfor blev antallet af opgaver og dermed antallet af delprøver med et rimeligt opgaveantal begrænset til en tre-fire stykker. Tidsbegrænsningen stiller store krav til den enkelte prøves opbygning, idet lettere opgaver ikke må komme efter sværere opgaver.

Det skulle være en gruppeprøve, der kunne tages på mange (op til hundredvis) personer samtidig. Individualprøver var udelukket af arbejdsmæssige grunde og desuden for tidskrævende, når man af statistiske grunde ønskede store materialer indsamlet på kort tid, så man dels kunne undersøge prøvernes egnethed som måleinstrumenter, dels kunne opstille normtabeller for prøveresultaterne i forskellige referencegrupper til brug ved vurdering af et givet prøveresultat.

Endvidere måtte man finde en balance mellem den nemme og tidsbesparende valgsvarsform, multiple-choice-typen, hvor alternative svarmuligheder er angivet til afkrydsning, og typen med åbne svarmuligheder, hvor den prøvede f.eks. skal skrive et ord, en sætning, eller en begrundelse for sit svar.

Afkrydsningsformen (vælg mellem 3-8 forskellige forslag til svar) er meget populær i udenlandske prøver, da den ikke afhænger af sproglig udtryksfærdighed og er tidsøkonomisk, specielt i rettefasen; men typen åbner mulighed for heldig gætning mellem valgsvarene. Det var godt nok uønsket, men fritsvarsformen var for usikker og for langsom, når prøveresultater skulle foreligge hurtigt efter prøvens afholdelse. Man måtte da finde på en eller flere besvarelsesformer, der ikke lagde op til gætning.

Rammerne var hermed lagt. Så manglede der 'bare' at fastlægge det fælles funktionelle indhold i de ca. 4 delprøver og så disses forskellige iklædning af opgaverne, samt udformning af hensigtsmæssig tidsbegrænsning, at efterprøve den praktiske gennemførlighed og frem for alt give hver delprøve en indre struktur, der kunne gøre den til et måleinstrument. Princippet var her, at man allerede ved afprøvningerne i konstruktionsfasen tog hensyn til, at de anvendte opgaver i hver delprøve skulle kunne ordnes *entydigt efter sværhed (rigtigheds-procent) i vidt forskellige grupper af personer, således at rækkefølgen blev populationsuafhængig.*

Og så skulle prøven for øvrigt være færdig til brug næste år.

Her kan passende indsættes et citat af Georg Rasch. Hans bog fra 1960 (omtales senere) blev genoptrykt i 1980 med et forord, hvor Rasch citeres for følgende om BPP's konstruktion.

> Prien did that in six months. He invented tests, which, when you see them, are rather surprising. He really did invent items of the same sort, from very easy to very difficult, and spaced in a sensible way. We did do some checking in the process and omitted or modified items that did nor seem to be working. It was really a masterpiece. Prien had been told, 'All you have to construct is four different kinds of tests, with very different subject matters and each of them should be just as good as Georg (Rasch) tells us that Raven's tests are. And he did so.

Det er så heldigt, at der er bevaret et meget fyldigt referat af et foredrag, som Prien holdt i november 1953 for MPA's medarbejdere. Her beskrev han, hvilke overvejelser han havde gjort sig om testens indhold, og da det giver et spændende indblik i konstruktionsfasen, vil store dele af dette og næste afsnit være citater fra referatet - alle markeret med kursiv.

Matrixprøven: *Ved at sætte mig ind i en række udenlandske intelligensprøvesæt, stødte jeg gang på gang på matrixopgaverne; dels de engelske Ravenprøver, hvor hver opgave består af nogle figurer, og hvor man skal finde en manglende figur, dels opgaver, der var efterligninger af Raven's matrixopgaver. Det drejer sig i alle disse opgaver om at finde et system, finde relationen mellem de enkelte figurer og vise, at man har fundet den ved selv at tilføje den manglende relat. Jeg tror, at Raven har fat i noget meget væsentligt ved tænkningens psy-*

kologi, når han arbejder med relationskalkyler på den måde. Jeg besluttede mig derfor til selv at arbejde videre med dette system. Der er fra forskellig side rejst kritik mod det Raven'ske matrixprincip, bl.a. fordi testen er en multiple-choice-test: der er til hver opgave 6 eller 8 forskellige valgmuligheder. Det bevirker, at der er mulighed for at gætte sig til det rigtige resultat.

Prien beskriver herefter et forsøg, hvor han lod 100 rekrutter udfylde Raven's test, og han fandt herved, at adskillige af valgmulighederne aldrig blev brugt - de var for lette at vælge fra.

Det betyder, at gætningsmomentet virkelig er blevet farligt. Så længe, der er 8 valgmuligheder, er der en stor chance for at gætte galt, men hvis der f.eks. kun er 2 svar, der er tillokkende, er chancen for at gætte rigtigt betydelig større. Man kan altså i mange tilfælde klare sig igennem ved at udelukke de mest tåbelige forslag og gætte mellem resten.

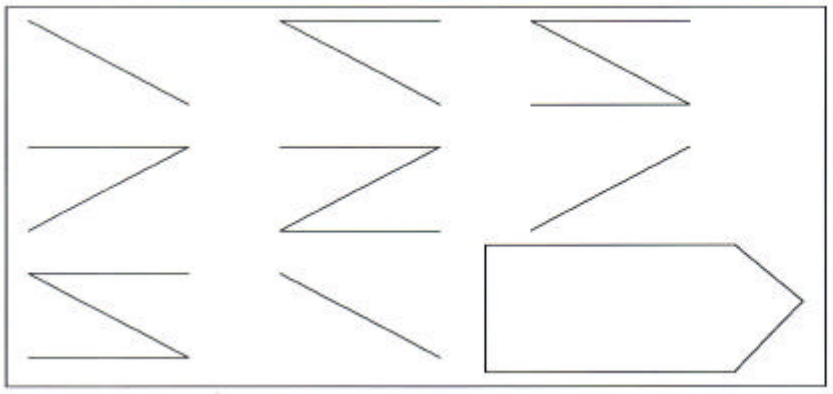
Jeg lagde endvidere mærke til, at der var meget få ubesvarede opgaver, formentlig netop på grund af valgmulighederne. Man ved, at en af dem er rigtig og tager så i hvert fald en af dem.

Jeg havde altså meget stærke motiver til ikke at anvende mutiple-choice-tests og måtte derfor lave prøver, hvor probanden selv skulle frembringe svaret og ikke vælge blandt nogle muligheder.

Det er klart, at jeg så ikke kunne arbejde med de Raven'ske figurer, der til-med kunne være vanskelige at tegne for en uøvet hånd. Jeg måtte finde noget andet og helst noget overindøvet, så selve fremstillingen af svaret ikke blev nogen større vanskelighed. Jeg valgte da bogstaverne, som de allerfleste har overindøvet i betydelig grad. Jeg prøvede at oversætte matrixopgaverne til bog-staver ved at erstatte hver figur eller del af figur med et bestemt bogstav. Der er vitterlig opgaver i sættet her, som er en slags oversættelse af matrixopgaverne.

Som eksempel viser den efterfølgende figur en af de mulige oversættelser af en Raven-lignende matrix til en BPP-lignende opgave. Antallet af liner er oversat til bogstaverne E, T og W, og hældningen af de skrå liner oversættes til X og Y.

Det viste sig imidlertid, at de allerfleste af matrixopgaverne kunne oversæt-tes på forskellige måder, afhængig af hvor langt ned jeg gik i min visuelle ana-lyse af figurerne, og hvad jeg derigennem kom til at betragte som "elementer". Eftersom jeg betragtede f.eks. de enkelte streger eller en kombination af streger



EX	TX	WX
TY	WY	EY
WX	EX	

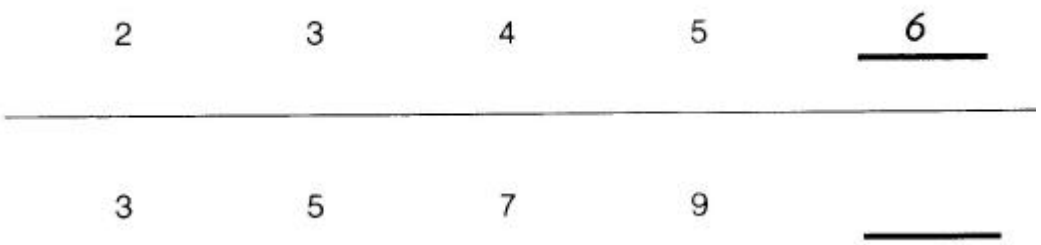
Figur 1. Oversættelse af en Raven-lignende matrix til en BPP-lignende opgave

som elementer, kunne jeg få et helt bånd af oversættelser for hver opgave; tilmed fik de forskellige oversættelser af samme opgave vidt forskellige sværhedsgrader. Meget svære opgaver kunne omformes til ganske lette og omvendt. ...

Samtidig fik jeg indblik i, hvilken rolle perceptions- og gestaltningsprocesser spiller for relationskalkylen i Raven's matrix. Disse processers indflydelse kommer jeg i vid grad til at undgå ved ikke at arbejde med de Raven'ske matrixfigurer. Bogstaverne (lutter versaler) anvender jeg som velkendte grafiske tegn uden deres fonetiske figur eller deres betydning. De må ikke danne ord, og kendskab til deres rækkefølge i alfabetet indgår ikke i opgaveløsningen.

Talrækker: *Jeg skulle nu finde en anden type af opgaver, som kunne være den næste delprøve. Jeg ville meget gerne fortsætte med princippet: Givet en relation; find den og vis, at De har fundet den ved at tilføje den manglende relat.*

Jeg havde brugt princippet med bogstaver i 1. prøve og ville nu gerne gøre det på en anden måde. Jeg fik fat i talserieopgaverne; der findes 2 i den gamle sessionsprøve og et sæt af dem i den svenske prøve. Man ordner en række tal efter et vist system, og probanden skal så finde det næste tal i rækken.

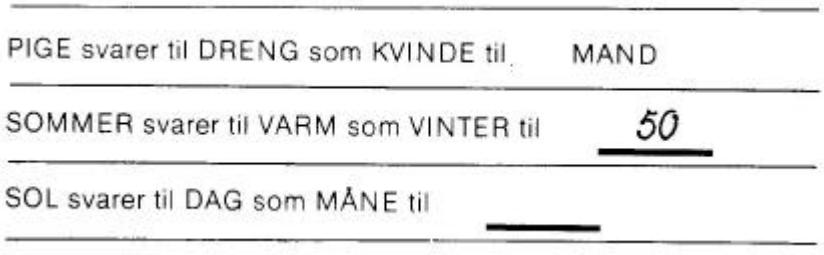


Figur 2. Demonstrationseksempler fra talserieprøven

Ordrelationer: *Det gik altså forholdsvis let at finde et prøvesæt til, som kunne gå efter samme princip som det foregående. Endnu et sæt fandt jeg på ved at lave verbalopgaver: sætninger, hvor det sidste ord manglede; sætninger, der var formede som en konstatering af relationer mellem forskellige ords betydning. Jeg mener herved at få fat i et meget væsentligt område indenfor intelligensen, nemlig en del om begrebsdannelser, en del af de verbale færdigheder, som jo spiller en væsentlig rolle i den hidtil brugte prøve, IGP.*

Af demonstrationseksemplerne ser De, hvordan det hele igen er stillet op som relationskalkyle: det første begreb svarer til det andet ligesom det tredje til det fjerde. Til sætningen: "Sommer" svarer til "varm" som "Vinter" til … skulle man svare kold. Der står 50 i figur 3, det er fordi kold er nr. 50 i ordlisten på siden overfor i prøvehæftet. Ordene i den er indordnet efter deres begyndelsesbogstaver, og hvert ord har et nr., således at opgaven kan besvares med et tal. Årsagen til, at verbalprøven er forsynet med en ordliste, er for det første den, at vi ellers kunne risikere ikke at kunne læse probandens svar. Vi prøvede at sende prøven ud uden ordliste, og det viste sig faktisk, at svarene ofte var ulæselige eller utydelige.

En anden grund til at bruge ordliste er, at man sikrer sig mod nydannelser; man har ganske bestemte ord, som svarene skal være, og disse ord findes i listen. Det er altså et nyt forsøg med opgør med multiple-choice-princippet. Vi præsenterer ganske vist et materiale, hvori svaret findes, men i en sådan mængde, at man ikke uden videre kan gætte; der er 100 ord i hver ordliste. Derimod er det ganske nemt at finde et ord, når man først har tænkt det, idet man blot kan slå det op under dets begyndelsesbogstav. Ordlisten er bygget op over de fejl, vi fik, da vi først sendte prøven ud uden ordliste. Der er altså rig mulighed for at begå de fejl, som man i almindelighed har lyst til at begå. Vi har siden fyldt ordlisten op med ganske almindelige ord, så vi fik lige mange under hvert begyndelsesbogstav.

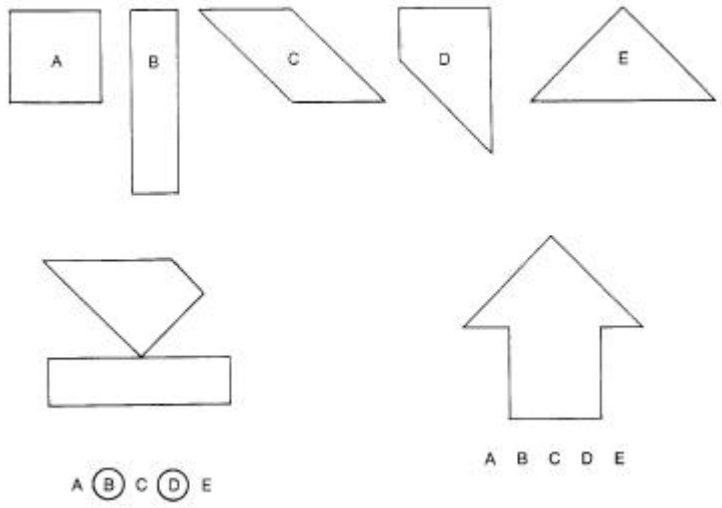


Figur 3. Demonstrationseksempler fra ordrelationsprøven.

Figurprøven: *I de eksisterende intelligensprøver er der ofte opgaver, hvor det gælder om at finde de indbyrdes relationer mellem nogle figurer, ofte geometriske. Sådanne opgaver betegnes i litteraturen som spatialopgaver; de skulle have noget med den "rumlige intelligens" at gøre.*

Jeg tror, at hovedsagen er, at man i disse opgaver skal finde relationerne mellem noget, som er figuralt givet. Man bruger i disse tests geometriske figurer og kalder dem så spatialtests. Det, jeg ville have fat i, var en prøve med figurer, noget visuelt givet, som var meningsløst på samme måde, som visse stavelser kan være det.

Jeg fik en dag fingre i et kinesisk puslespil, Hao-Wan. Det bestod af små puslebrikker, som skulle sammensættes til figurer efter en fortegning. Alle brikkerne skulle bruges til hver figur. Jeg opdagede hurtigt, at det var meget svært. Jeg tænkte, at hvis jeg lavede opgaver, hvor man ikke skulle anvende alle brikkerne, ville opgaven blive nemmere, endda tilstrækkelig nem til at kunne løses på det mentale plan, altså uden virkelig at bruge puslebrikker, men blot ved hjælp af en tegning på et stykke papir. Vi lavede nogle puslebrikker af pap, satte figurer sammen og gav os til at gennemprøve dem. Det er klart, at i en sådan prøve bliver opgaverne sværere, jo mere kompakte de er, dvs. jo flere af brikkernes kanter eller hjørner, der ligger skjult inden i opgavefiguren. Man kan faktisk stille helt vanskelige opgaver af den art. Hver af de brikker man skal bruge til løsning af opgaverne, har et bogstav (A.B.C.D.E). Nedenunder hver opgavefigur står alle 5 bogstaver, og man angiver sit svar ved at sætte ringe rundt om de bogstaver, som svarer til de brikker, man har brugt ved opgave løsningen.



Figur 4. Demonstrationseksempler fra figurprøven.

AFPRØVNING

Der fulgte nu adskellige afprøvningsrunder, som typisk foregik ved, at omkring 100-200 værnepligtige udfyldte prøven. Første trin var at få opgaverne indordnet efter sværhedsgrad, så de letteste opgaver stod først i prøven. Det viste sig

at en opgaves løsningsfrekvens var afhængig af opgavens plads i forhold til andre opgaver. Det var, som om hver opgave befandt sig i et opgavemiljø, og blev den flyttet ud af det, skiftede den karakter. Vi prøvede i et af opgavesættene at sætte samme opgave ind 2 forskellige steder; den fik vidt forskellig løsningsfrekvens. Løsningsfrekvensen var langt mindre, når opgaven blev placeret langt henne i sættet; man havde da åbenbart indstillet sig på, at nu var det svært. Det var ikke fordi, man ikke nåede så langt; vi gav nemlig ubegrænset tid.

Man kunne altså ikke første gang en prøve kom hjem fastlægge sværhedsgraden. Man måtte sende den ud flere gange, men anden gang var der mindre at rette end første gang, og efter nogle gange nåede vi til en forholdsvis stabil rækkefølge.

Desuden lavede vi, hver gang en ny prøve kom hjem, fejlanalyser. For hver opgave blev alle forkerte løsninger skrevet op samt antal personer, der havde begået hver fejl. På den måde fik vi fat i de hyppigste fejl og kunne give os til at tænke over, hvorfor en bestemt fejl var meget hyppig. Nogle gange skyldtes det, at opgaven var virkelig svær og forledte til at gøre en bestemt fejl, hvis man ikke tænkte sig om. I andre tilfælde viste fejlen sig at kunne være en rigtig løsning; opgaven var altså flertydig. Sådanne flertydige opgaver måtte selvfølgelig udrangeres.

Efter at have vist et eksempel på flertydighed fra matrixprøven tilføjes:

Der så altså ud til at være noget flertydigt ved opgaven og desuden en ret stor chance for at få den rigtige løsning ved gætning, altså besvare opgaven rigtigt uden at have fattet systemet. Det var faktisk en virkelig farlig opgave, vi havde fundet frem til, for vi ønskede jo ikke ved intelligensmåling at begunstige en aladdintype, der bare gætter. Hvad værre er, er at den begavede opdager, at systemet gør opgaven flertydig. Han indser, at der må være mere i opgaven; han prøver at finde frem til det, kan måske ikke finde det og lader opgaven stå ubesvaret. Vi kan altså her have det paradoksale forhold, at den anvendte opgave i en vis forstand kan hænde at blive "sværere" for den godt begavede end for den mindre godt begavede.

Prien omtalte herefter, hvordan man undersøgte, hvor lang tid der skulle gives til hver af delprøverne, for at man kunne regne med, at yderligere tid ikke ville give flere rigtigt løste opgaver. Det skete ved, at probanderne skulle skifte til en ny farveblyant, hver gang der var gået 5 minutter. Herved fik man et ganske godt indtryk af hvor længe, der blev løst opgaver, og hvor længe der fremkom rigtige svar. Det var egentlig tanken, at der skulle gives så megen tid, at man kunne regne med, at alle havde præsteret alt, hvad de kunne. Dette kom også til at gælde for matrix- og talrækkeprøven, men for de to andre prøver, ordrelationer og figurer, viste det sig nødvendigt at begrænse tiden noget. Det skyldtes, at hvis der blev givet rigelig tid, var der så mange probander, der kom tæt på at løse alle opgaver, at prøverne ikke kunne skelne klart mellem de dygtigste.

Ved afprøvningerne viste der sig nogle specielle problemer ved figurprøven.

Når vi bygger figurer op af småbrikker, kan det ikke undgås, at jo flere brikker, vi bruger, des større bliver opgavefigurerne. Jeg var lidt bange for, at man sidst i opgavesættet ville kunne se af figurens areal, at alle brikkerne skulle bruges. For at undgå det, lavede jeg en figurprøve, hvor alle opgavefigurernes arealer var lige store. I den prøve skulle man altså ikke bruge selve de viste brikker, men brikker af samme form som dem og med forskellig størrelse i den enkelte opgave. Det viste sig imidlertid ikke at være nødvendigt. Løsningsfrekvenserne for opgaver med og uden transformeret areal fulgtes pænt ad. Det var sværest at løse opgaverne med transformeret areal, men kurverne over løsningsfrekvenserne var blot forskudt i forhold til hinanden, og de var forskudt nogenlunde lige meget for alle opgaver. Man syntes at have forskudt hele prøvens sværhedsgrad og ikke ændret noget ved de enkelte opgavers sværhed i forhold til hinanden, således som jeg havde befrygtet. Det var vi meget glade for, da det var meget tidsrøvende at give instruktion til opgaver med transformeret areal.

Endnu et problem dukkede op. Skal probanderne have lov til at tegne i figurerne, eller skal de ikke? Hvis nogle tegner, og andre ikke, kan det være de i grunden ikke løser samme slags opgaver. Vi ville måske risikere at måle forskellige ting med de samme opgaver, hvis nogle tegnede og andre ikke gjorde det. Vi udsendte derfor prøverne dels med tilladelse, ligefrem opmuntring til at tegne, og dels med strengeste forbud mod at tegne andet end blyantsringene omkring bogstaverne neden under hver opgave.

Ved at undersøge løsningsfrekvenserne i de to sæt opgaver i forhold til hinanden fandt man:

Det er åbenbart ikke så vidt forskellige opgaver, der løses i de to tilfælde. Der er i hvert fald ingen grund til at forbyde probanderne at tegne. Vi ville også let være kommet i den situation, at nogle havde tegnet trods forbudet, og vi ville så ikke kunne bedømme deres prøve.

Endelig forsøgte vi os med demonstrationsbrikker, der alle var formindskede eller forstørrede i konstant forhold til størrelsen af brikkerne i opgavefigurerne. Der viste sig ligesom før blot en forskydning i løsningsfrekvenserne. Det er sværere at arbejde med forstørrede eller formindskede end med brikker af samme størrelse som opgavefigurernes. Men det var sværere på en ensartet måde for alle opgaverne; de enkelte opgavers sværhedsrækkefølge blev ikke forrykket.

ANVENDELSE

Internt. BPP prøven blev i august 1956 introduceret på sessionerne ved en skrivelse fra Indenrigsministeriet til alle udskrivningskredse. Her blev formålet med prøven defineret således:

> Ved prøven tilsigter man fortrinsvis
> a) at blive opmærksom på de værnepligtige, der er uegnede eller mindre egnede til tjeneste i forsvaret på grund af ringe intelligens,
> b) at udfinde de særligt intelligente, samt
> c) gennem prøveresultatet at skaffe forsvarets enkelte afdelinger nogen forhåndsviden om de mødende.

Man havde som allerede nævnt anvendt intelligensprøver før 1956, men da skete det først efter indkaldelsen til forsvaret. Den væsentligste ændring ved overførslen til sessionerne var, at det nu var muligt at kassere en værnepligtig før han blev indkaldt, hvis hans intelligens var så lav, at der var stor risiko for, at han ikke kunne gennemføre værnepligten. Det havde man nok kunnet før, men det var da mere tilfældigt, om man blev opmærksom på den ringe intelligens. Der blev indført et mindstekrav til resultatet på BPP, og det er stadig sådan, at man kasseres, hvis resultatet er for ringe.

Det er flere gange blevet påvist, at værnepligtige med lav intelligens i højere grad end andre bliver efterkasseret, dvs. kasseret i løbet af indkaldelsesperioden, og kravet er blevet hævet et par gange i årenes løb. I dag har 7-8 pct. af de

værnepligtige et BPP resultat, der er lavere end det krævede, og de bliver derfor kasseret. I en del tilfælde ville de alligevel være blevet kasseret af fysiske grunde, så det er vanskeligt at give et præcist tal for, hvor mange der bliver afvist alene på grund af et lavt BPP resultat.

Ved konstruktionen af BPP prøven lagde Prien stor vægt på at undgå multiple-choice opgaver, og det er derfor yderst vanskeligt at slumpe sig til et for godt resultat, men man kan naturligvis altid bevidst gøre sig ringere, end man er. På sessionerne vil man derfor ikke kassere en værnepligtig alene på grundlag af et lavt BPP tal. Det skal bekræftes f.eks. gennem iagttagelse, samtale eller skole-resultater. Det sker kun sjældent, at sessionen er nødt til at underkende BPP resultatet. Det er kun få, der bryder sig om at gøre sig mindre kløgtige, end de er, og de fleste bliver nok også så grebet af opgavernes udfordring, at de yder alt, hvad de kan, selvom de måske har et ønske om at blive kasseret.

Som det fremgår af Indenrigsministeriets skrivelse skal BBP også tjene det formål at finde de bedst begavede. Her tænkes specielt på udtagelse af elever til sergentskolerne, hvor der er et absolut mindstekrav til BPP resultatet. Også ved besættelsen af specielle funktioner kan tjenestestedet tage hensyn til intelligen-sen, hvis ikke den civile uddannelse i sig selv lover tilstrækkeligt.

IMP deltager i udvælgelsen af personel til mange uddannelser og specielle jobs i forsvaret. Udvælgelsesproceduren består typisk af en række tests samt i større eller mindre omfang kontakt med en psykolog, der bedømmer personligheds-faktorer som f.eks. motivation. BPP indgår altid i testbatteriet, men der stilles ikke i disse tilfælde absolutte krav til intelligensen. I stedet for gives der en samlet bedømmelse af ansøgeren ud fra alle testresultater og personvurderinger - men BPP indgår med ret stor vægt.

BPP spiller også en vigtig rolle, hvis en personelgruppe i en undersøgelses-sammenhæng skal beskrives. F.eks. i en undersøgelse af, om de værnepligtige, der melder sig til frivillig tjeneste i den Danske Internationale Brigade, udgør en særlig gruppe. Det vil være naturligt at sammenligne dem med andre værnepligtige med hensyn til personlighedsmæssige forhold, uddannelse, opvækstforhold mm. samt BPP.

Det er formentlig en helt enestående situation, at næsten hele den mandlige befolkning i et land er blevet intelligensmålt med den samme prøve i 50 år.

Resultaterne har derfor også stor interesse uden for forsvaret. Af anonymitetshensyn er det desværre ikke muligt at gøre det store materiale frit tilgængeligt for forskningen, men en hel del undersøgelser har dog i årenes løb kunnet udnytte målingerne. I nogle tilfælde har summariske opgørelser været tilstrækkeligt, og i andre tilfælde har IMP udført nogle mellemregninger, så anonymiteten ikke er blevet krænket.

Tom Teasdale. Som eksempler på den forskning, der er baseret på BPP, kan nævnes nogle af de resultater, som lektor ved Psykologisk Institut, Københavns Universitet Tom Teasdale i samarbejde med andre forskere har opnået.

- I en dansk-amerikansk langtidsundersøgelse deltog 232 danske mænd, som i den forbindelse blev testet med en anden intelligensprøve, WAIS. Denne test administreres individuelt i modsætning til BPP, som udfyldes i grupper med ca. 30 personer. Det har åbenbart været diskuteret, om denne forskel havde indflydelse på resultatet, uden at det dog var blevet undersøgt. Man benyttede derfor lejligheden til at sammenligne WAIS resultatet med det BPP resultat, mændene tidligere havde opnået, da de var på session. I gennemsnit var der gået godt fire år mellem de to prøvetidspunkter, men til trods herfor var der en høj korrelation på 0.82 mellem resultaterne.

- I mange industrialiserede lande har man konstateret tydeligt stigende resultater ved intelligensmålinger for fødselsårgangene efter 1940. I Danmark er BPP resultaterne blevet undersøgt for årgangene 1940-80, og her viser den samme stigning sig - dog i mindre omfang fra 1970 til 1980. Stigningen er størst for dem, der har en kort skolegang, mens studenternes resultater ikke er steget. Generelt gælder det, at man har gået flere år i skole, og det betyder, at de bedste elever forlader de korte uddannelser til fordel for længere, og de lange uddannelser får tilført en del lidt tungere elever. Herved er udgangsniveauet alt andet lige faldet noget ved både de korte og de lange skoleuddannelser. Når der alligevel sker en stigning ved de korte uddannelser kan det forklares ved, at der i perioden er ydet en stor indsats i skolerne for at støtte de svageste elever, mens der ikke har været tilsvarende tiltag for de dygtigste elever.

- I flere lande er det blevet påvist, at der er regionale forskelle i scores på intelligenstests. Generelt scores der højere i regioner med store byområder end i landdistrikter. To forklaringsmuligheder er blevet foreslået. For det første kunne forskellen skyldes, at de bedre uddannede og de socialt højest placerede

gennem generationer er flyttet fra land til by. En anden mulighed er, at vilkårene - specielt uddannelsesvilkårene - er bedst i byområderne. For at belyse de regionale forskelle er de sessionsbehandledes BPP resultater og uddannelsesresultater blev opdelt på de syv udskrivningskredse, som Danmark er inddelt i. Der var også i Danmark en klar tendens til højere BPP resultater i de udskrivningskredse, hvor befolkningstætheden er størst, og her havde de sessionssøgende også haft længere skolegang. Det interessante var, at variationen mellem udskrivningskredsene i BPP næsten forsvinder, hvis skoleuddannelsen holdes fast. 92% af variationen kan forklares ved forskellen i uddannelse, og den resterende variation er uafhængig af befolkningstætheden. Det taler for den anden forklaringsmulighed, at forskellen i BPP skyldes forskel i adgang til længerevarende uddannelse, som til gengæld er en funktion af befolkningstætheden.

- Ved hjælp af BPP resultater er der også blevet ydet et bidrag til diskussionen om, det er den genetiske arv eller opvækstmiljøet, der har størst betydning for den målte intelligens. Det er sket ved at undersøge et antal hel- og halvsøskende par, som ved fødslen er blevet bortadopteret til forskellige hjem. Man har registeret BPP, højde og skoleuddannelse, og opdelt parrene i følgende fire grupper:

A: 28 par helsøskende, der er opvokset adskilt
B: 64 par halvsøskende, der er opvokset adskilt
C: 24 par ubeslægtede, der er opvokset sammen (adopterede af de samme)
D: 73 par helsøskende, der er opvokset sammen (hos forældrene)

Korrelationerne mellem parrene vises i de tre sidste søjler i nedenstående tabel. De to søjler - arv og miljø - viser, hvilken størrelsesorden man kan forvente korrelationen har i de fire grupper, hvis arv henholdsvis miljø er ene om at bestemme størrelsen af en måling. Hvis arven er eneansvarlig, må korrelationen være 0 i gruppe C, hvor der ikke er noget slægtskab. Størst korrelation, R, fås for helsøskende (gruppe A og D), og en mindre korrelation, r, fås for halvsøskende (gruppe B). Hvis miljøet er ene om indflydelse på en måling, vil korrelationen være 0 for parrene i grupperne A og B, mens den vil være af samme størrelse for parrene i grupperne C og D, hvor opvækstmiljøet har været det samme.

Gruppe	arv	miljø	Højde	BPP	Uddannelse
A helsøsk. /adskilt	R	0	0.27	0.47	0.38
B halvsøsk./adskilt	r	0	0.20	0.22	0.00
C ubeslægt./sammen	0	R	0.03	0.02	0.43
D helsøsk. /sammen	R	R	Ikke målt	0.52	0.67

Som ventet passer korrelationerne for Højde godt med en hypotese om, at arven er eneafgørende, og det samme gælder for BPP resultatet. Derimod forklares Uddannelsen bedst som et samspil mellem arv og miljø. Effekten er nogenlunde den samme i gruppe A og C, hvor hver sin faktor er i spil, mens der er en forstærket effekt i gruppe D, hvor begge faktorer spiller ind. Undersøgelsen er altså et stærkt argument for, at intelligensen bestemmes af den genetiske arv og ikke af miljøet.

Børge Prien. I 1947 undersøgte Georg Rasch rekrutternes intelligensprøveresultater målt med IGP. Han opdelte dem efter geografisk opvækstmiljø, fædrenes sociale rang (defineret ved erhverv) og egen uddannelse. Der var tydelige geografiske forskelle i intelligens, og den gennemsnitlige intelligens steg op gennem de sociale rangklasser og med stigende uddannelse. Rasch viste, at af disse tre variabler var det dog kun uddannelsen, som havde relevans for beskrivelsen af rekrutternes intelligens, idet intelligensen i de enkelte uddannelsesgrupper var uafhængig af socialklasse og geografi - det sidste svarer til et af Teasdales resultater. Det tilsyneladende sammenfald mellem intelligens og faderens sociale rang kunne forklares ved, at sønner af socialt højt placerede fædre havde fået en længere uddannelse. Af Raschs undersøgelse kan man dog ikke slutte, at intelligens alene er et resultat af uddannelse; intelligensforskelle kan have eksisteret før uddannelsesvalget og have været medbestemmende for valget. Der var da også andre undersøgelser, der viste, at børns intelligens før skolealderen og i de tidligste skoleår - hvor der endnu ikke er forskel i uddannelsen - stiger med faderens uddannelsesniveau.

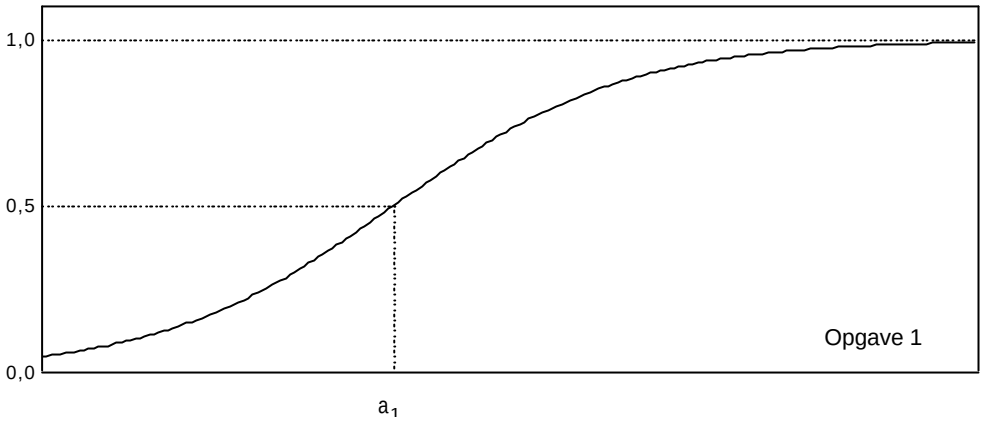
Børge Prien har belyst, hvilken betydning faderens uddannelse har for barnets uddannelsesvalg, og har herved bidraget til en øget forståelse af samspillet mellem barnets intelligens forud for differentieret undervisning, uddannelsesvalg og faderens uddannelse. Det skete ved i 1962 at intelligensmåle ca. 1000 børn i 6. klasser med BPP, efter at det var påvist, at den kunne anvendes på børn fra 5. klasse og opefter. I 1960'erne havde man delt folkeskole, hvor børnene blev opdelt i en almen og en boglig linie fra 6. klasse, og målingen skete ved starten af 6. klasse, hvor valget var foretaget, men før børnene havde modtaget forskellig undervisning. Ved nogle skoler havde man udelte klasser, og de blev også repræsenteret i undersøgelsen. Fædrene blev opdelt i fire grupper efter de uddannelseskrav, der blev stillet til deres erhverv.

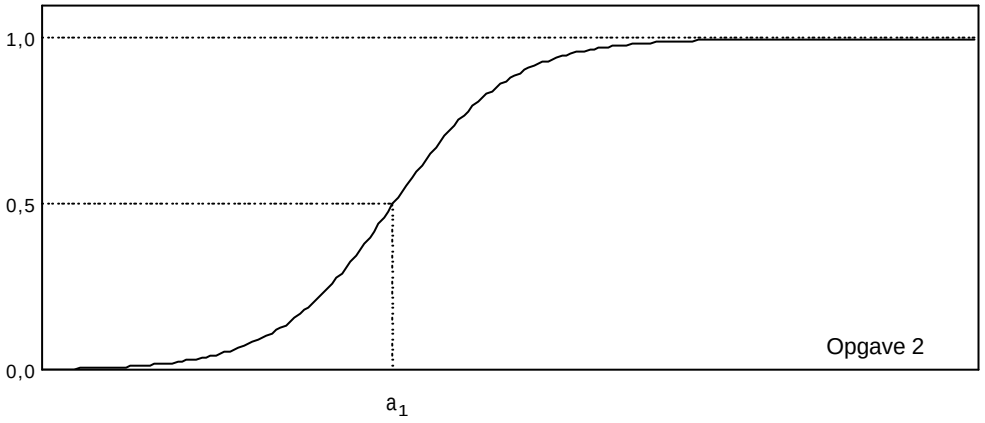
Undersøgelsen bekræftede, at børnenes intelligens var stigende med fædrenes uddannelsesniveau, hvilket formentlig kan henføres til forskel i genetisk arv og/eller i opvækstmiljøets stimulering. For alle fædregrupper var børnenes gennemsnitlige intelligens højere i de boglige klasser end i de almene klasser, mens de udelte klasser placerede sig imellem: børnenes intelligens havde altså betydning for uddannelsesvalget. Men valget var også styret af fædrenes uddannelse. Det fremgår af, at mens børnene i de delte klasser med de laveste intelligenser alle kom i den almene linie, og alle med de højeste intelligenser kom i den boglige linie, så var valget for den store mellemgruppe i høj grad afhængig af faderens uddannelse. F.eks. var der i en intelligensgruppe lidt under middel kun 15 pct. af børnene med fædre i de laveste uddannelsesgrupper, som kom i den boglige linie, mens 52 pct. af børn med samme intelligens men med fædre i de højeste grupper kom i den boglige linie. I mindre udtalt grad fandt man samme forskel i de andre mellem-intelligensgrupper. I de delte klasser kom i alt 36 pct. af børnene med fædre i laveste uddannelsesgruppe i den boglige linie, mens det gjaldt 80 pct. af børnene fra den højeste uddannelsesgruppe, og denne forskel var altså et resultat af forskel i såvel intelligens som i fædrenes uddannelse.

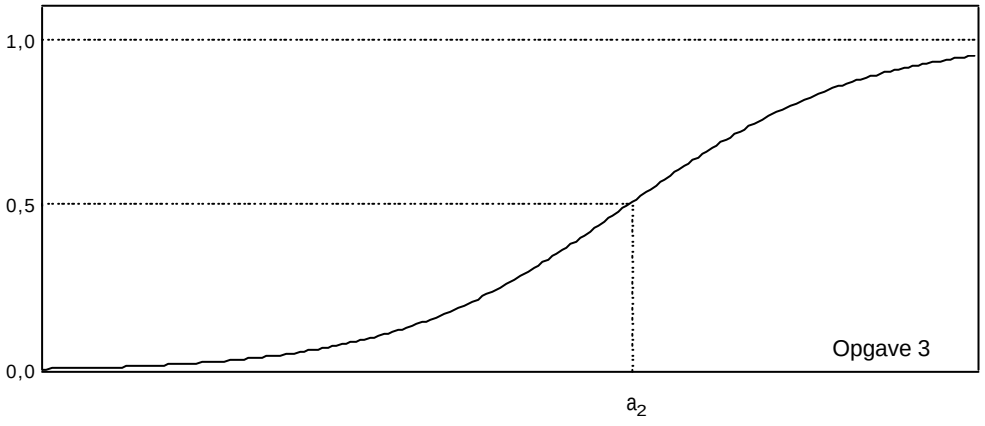
Et sideresultat af undersøgelsen var, at Prien kun fandt minimale forskelle mellem piger og drenge. For hele gruppen var det gennemsnitlige antal løste opgaver ca. 30, og pigernes gennemsnit var en lavere end drengenes i de almene og i de udelte klasser og to lavere i de boglige klasser.

Georg Rasch. Da BPP blev konstrueret og afprøvet, var den samtidig fødselshjælper for en vigtig nyskabelse i psykometrien, og denne lidt specielle anvendelse er den sidste, der skal omtales. Det er allerede nævnt, at Georg Rasch deltog som statistisk konsulent i afprøvningsfasen, men forud for det, havde han i nogle år beskæftiget sig med lignende opgaver og var meget optaget af, hvordan man skal definere en opgaves sværhed. Hvis det skal have mening at sige, at opgave 1 er lettere end opgave 2, må det skulle gælde for alle personer. I en gruppe personer, som er gode til at løse opgaver, skal der være flere som kan løse opgave 1 end opgave 2 - og det skal også gælde for en gruppe af dårlige opgaveløsere. Begge grupper skal have større sandsynlighed for at klare opgave 1 end 2. Definitionen af opgavesværhed skal altså være uafhængig af, hvilken population opgaverne bliver stillet til.

I figur 5 er det søgt illustreret ved for tre opgaver at vise forbindelsen mellem testpersoners dygtighed (de vandrette akser) og deres sandsynlighed for at løse







Figur 5. Tre eksempler på sandsynligheden for et rigtigt svar som funktion af dygtighed.
Vandrette akser: dygtighed og Lodrette akser: sandsynlighed

opgaven (de lodrette akser). Hvis man er meget dygtig (til højre i figuren) er sandsynligheden for et rigtigt svar tæt på 1 for alle opgaverne. Er man derimod en dårlig opgaveløser (til venstre i figuren), er sandsynligheden for et rigtigt svar næsten 0. En person med dygtigheden a_1 har f.eks. 50 pct. sandsynlighed for at løse både opgave 1 og 2, men han har en mindre sandsynlighed (ca. 20 pct.) for at løse opgave 3.

Det er den samme kurve, der er indtegnet for opgaverne 1 og 3 bortset fra, at kurven i opgave 3 er forskudt mod højre. For enhver dygtighed er der altså størst sandsynlighed for at svare rigtigt på opgave 1, og det er derfor berettiget at sige, at opgave 1 er lettere end opgave 3.

Opgave 2's kurve afviger fra de to andre. Der er 50 pct. sandsynlighed for et rigtigt svar ved samme dygtighed a_1 som i opgave 1, men kurven er stejlere i omegnen af a_1. Det betyder, at området, hvor det er meget usikkert, om der vil blive svaret rigtigt eller forkert, er mindst for opgave 2: Opgave 2 har størst diskriminationsevne. Men hvilken af opgaverne 1 eller 2 er sværest at løse? Alle med en dygtighed mindre end a_1 har tydeligvis større sandsynlighed for at klare opgave 1 end opgave 2, så opgave 1 er lettest for de mindst dygtige. For de dygtigste er det omvendt. Hvis dygtigheden er større end a_1, er sandsynligheden størst for at løse opgave 2. Opgave 2 er altså lettest for de dygtigste testpersoner. Man kan derfor ikke sige, at den ene opgave er sværere end den anden. Hvis de to opgaver var egnede til børn og blev stillet i 3. klasse, ville man få flest rigtige svar på opgave 1, mens man i 9. klasse ville få flest rigtige svar på opgave 2 - resultatet afhænger af populationen.

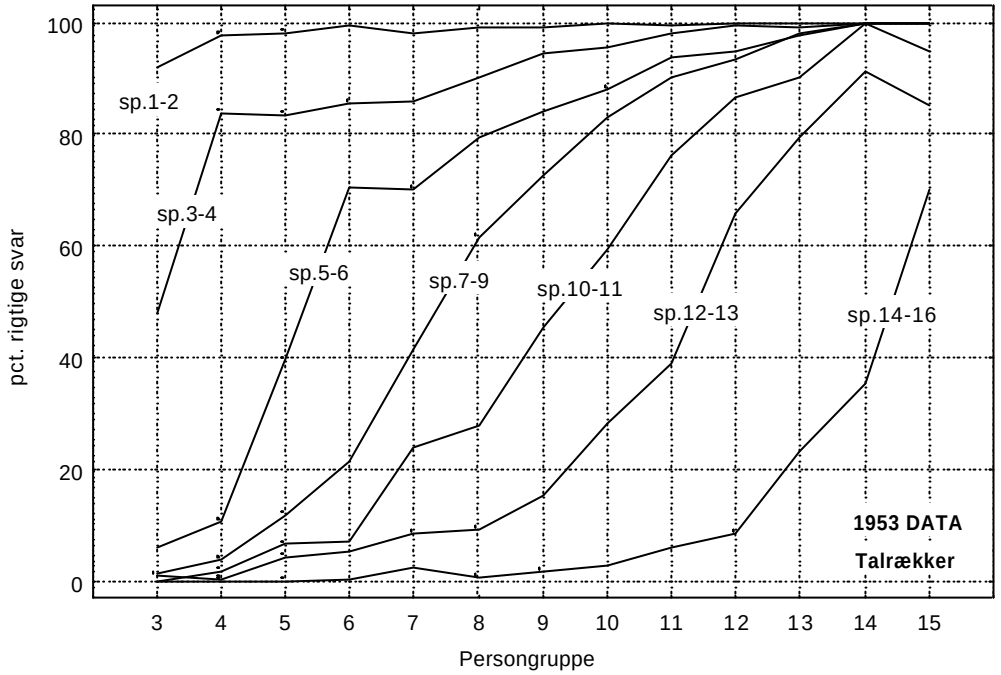
En af Raschs opgaver havde netop været at bygge bro mellem læseprøver, der blev anvendt på forskellige klassetrin, og der er det uheldigt, hvis rækkefølgen af opgaverne efter sværhed ikke er den samme for forskellige aldre. Han formulerede derfor hvilke krav, de enkelte opgaver (items) i en test skal leve op til, hvis man skal kunne tale om en opgaves sværhed uafhængigt af testpersonerne - og om personers dygtighed uafhængigt af opgaverne. Betingelsen for, at opgaverne i en test har en sværhed uafhængigt af populationen, er således, at sandsynligheden for rigtige besvarelser skal være voksende funktioner af dygtigheden, og at de tilhørende grafer ikke krydser hinanden - dvs. at opgaverne har samme diskriminationsevne.

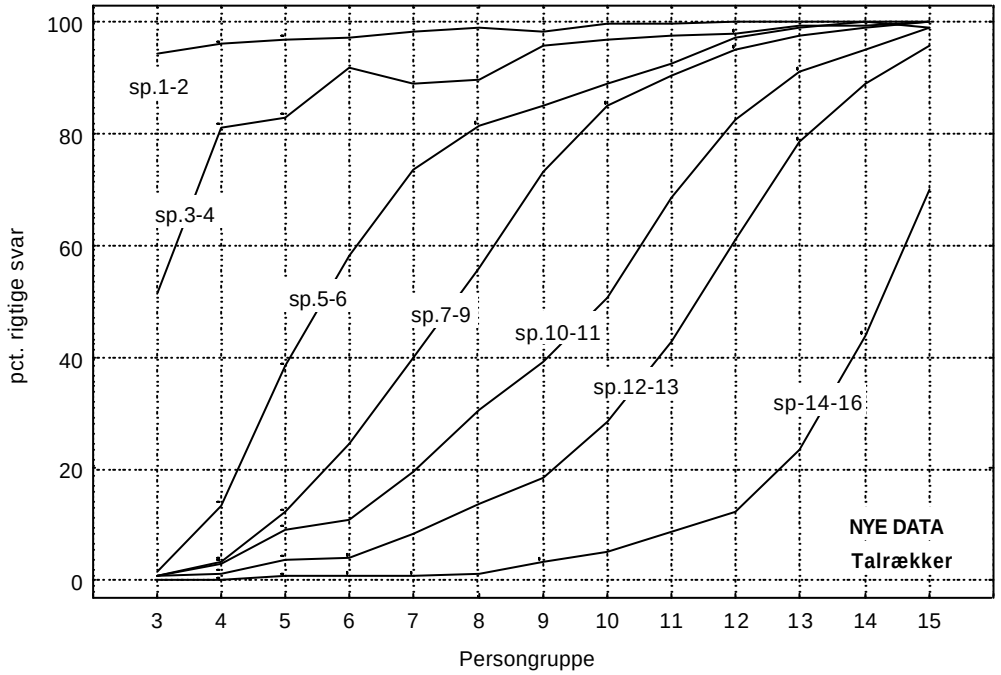
I 1960 havde Rasch afklaret sine ideer så meget, at han kunne skrive bogen ”Probabilistic Models for Some Intelligence and Attainment Tests”, som skulle blive lidt af en bestseller. Hans ideer fandt så megen interesse, at mange arbejdede videre ud fra dem, og Raschs betydning kan f.eks. illustreres ved antallet af gange han og hans model bliver omtalt i bogen ”Handbook of Modern Item Response Theory” fra 1989. I bogen har 29 forskere hver bidraget med et afsnit om forskellige sider af itemanalyse, og i 19 af bidragene henvises der til Raschs arbejde.

I Raschs bog beskrives først to Poisson modeller for læseprøver og dernæst den logistiske model med to parametre (testpersonens dygtighed og opgavens sværhed), som Rasch betegnede ”a structural model for items in a test”. Det er denne model, som gjorde Rasch berømt, og nu omtales den blot som Raschs model. Hele denne del af bogen illustreres med data fra de fire delprøver i BPP’en indsamlet for 1094 rekrutter ved den sidste afprøvning i september 1953. I bogen er der så detaljerede dataoplysninger om delprøven Talrækker, at det er muligt at foretage en grafisk sammenligning mellem de nutidige og de 50 år gamle testresultater for denne delprøve.

Øverst i figur 6 ses en gentegning af den figur, som Rasch brugte til at vise, at sandsynligheden for et rigtigt svar for alle spørgsmål er en voksende funktion af dygtigheden. Altså, at sandsynligheden opfører sig som vist i figur 5. Dygtigheden for de enkelte rekrutter er ukendt, så Rasch opdelte i stedet for personerne i grupper efter deres totale antal rigtige svar på prøvens 17 spørgsmål. Her udnyttede Rasch det forhold, at hvis sandsynligheden er voksende, så vil en ordning af personerne efter antal rigtige svar også være en ordning efter dygtighed. Det følger af, at hvis en gruppe personer er lidt dygtigere end en anden gruppe, vil de have lidt større sandsynlighed for at svare rigtigt på alle spørgsmål og derfor også i gennemsnit have et lidt større antal rigtige svar. Følgelig vil en persongruppe med f.eks. 8 rigtige svar også i gennemsnit være dygtigere end den gruppe, der kun har 7 rigtige svar.

Rasch tog ikke de grupper med, der havde ganske få rigtige svar eller svaret rigtigt på næsten alle opgaver. Han lagde ligeledes spørgsmål sammen, som havde næsten samme totale rigtighedsprocent og udelod det sværeste spørgsmål nr. 17. Begge dele skete for at udjævne nogle af de tilfældige udsving, der kunne være. På figuren kan man se, at alle kurverne herefter er monotont voksende med få undtagelser. Den nederste del af figur 6 viser de samme kurver for

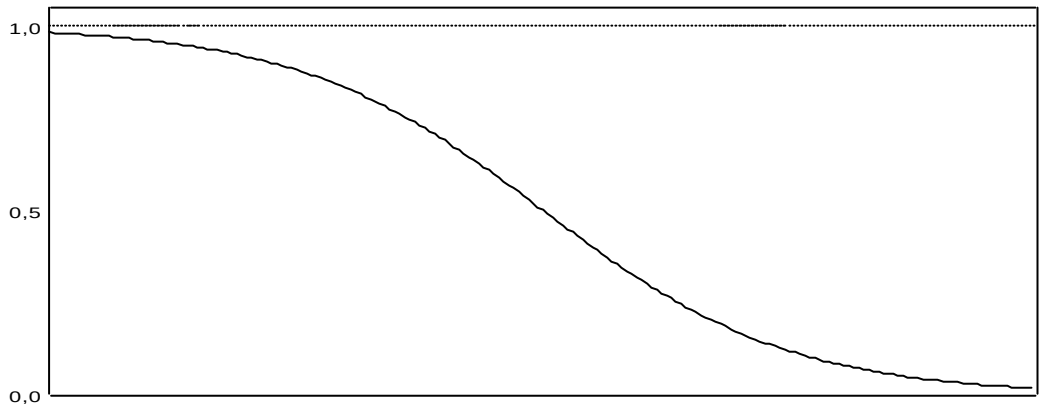




Figur 6. Hyppigheden af rigtige svar som funktion af det samlede antal rigtige svar

de nye data, som er indsamlet for 2734 sessionssøgende fra årene 2001-02. Hovedindtrykket af de to figurer er det samme, men for de nye data er der dog færre undtagelser fra monotonien, og det kan til dels skyldes, at der her er mere end dobbelt så mange testpersoner.

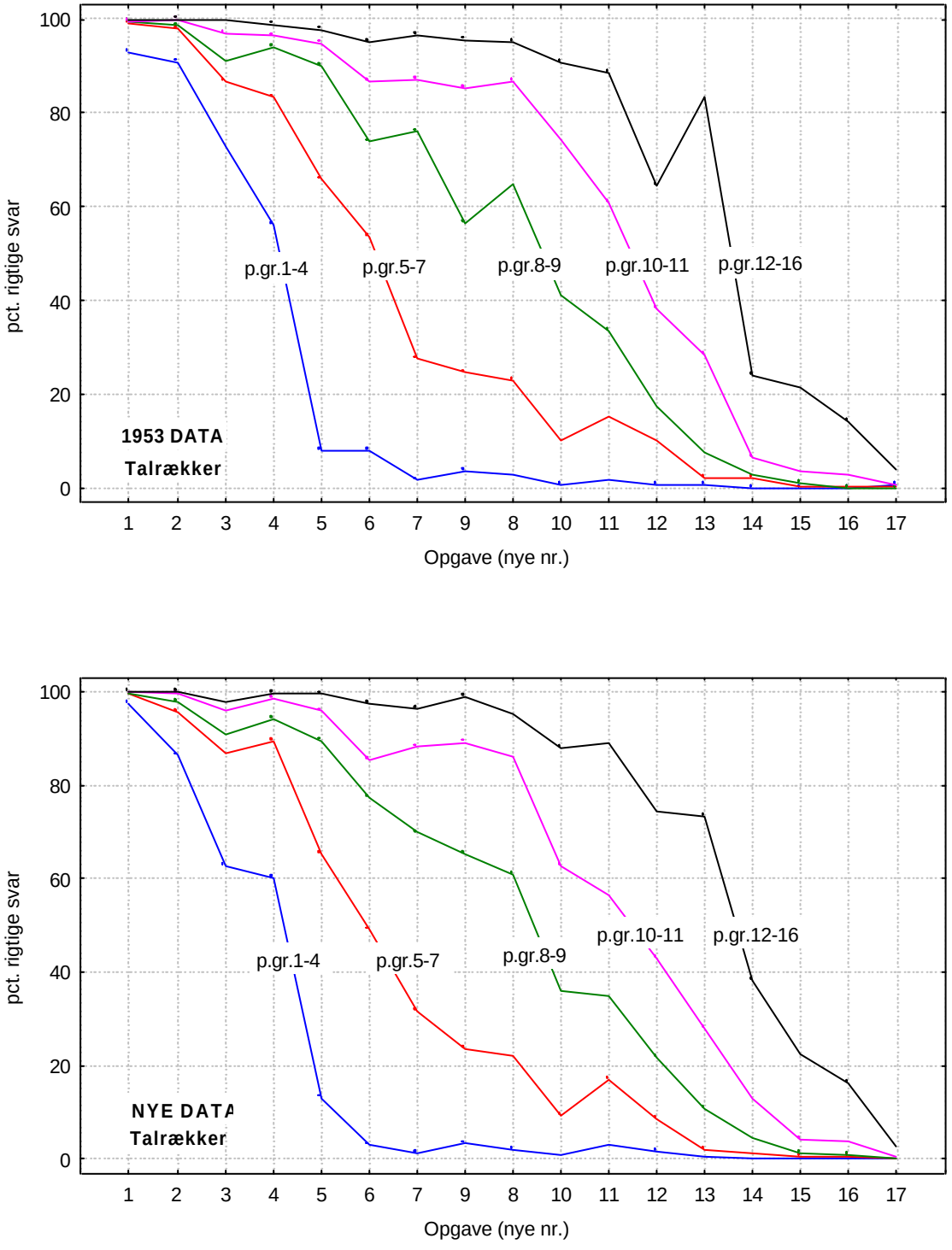
I stedet for at undersøge om graferne krydser hinanden, valget Rasch at undersøge en persons sandsynlighed for et rigtigt svar som funktion af opgavens sværhed. Hvis begrebet sværhed skal have mening, må denne sandsynlighed være en aftagende funktion af sværheden som vist i figur 7. Jo sværere en opgave er, jo mindre er sandsynligheden for, at en bestemt person kan løse opgaven - og det skal gælde for alle personer. Altså skal alle de sandsynlighedskurver, der kan tegnes for forskellige personer være aftagende. Man kan vise, at hvis disse kurver er aftagende, så vil sandsynlighedskurverne som funktion af dygtighed (figur 5) ikke skære hinanden - de to betingelser er ækvivalente.



Figur 7. Sandsynligheden for et rigtigt svar som funktion af sværhed

Figur 8 viser for oven gentegningen af Raschs figur til undersøgelse af sandsynligheden for et rigtigt svar som funktion af sværhed - her konkretiseret ved antal rigtige svar på opgaverne. For at stabilisere graferne har Rasch måttet foretage en gruppering af de persongrupper, som bliver bestemt af det totale antal rigtige svar. Stort set har graferne det ønskede aftagende forløb, men der er dog nogle afvigelser, som kun til dels er forsvundet i den nederste del af figur 8, som er tegnet på grundlag af de nye data.

Hvis sandsynligheden for et rigtigt svar (næsten) er en voksende funktion af personernes samlede antal rigtige svar (figur 6), vil dette antal ordne personerne efter dygtighed . Og hvis sandsynligheden for et rigtigt svar for alle personer



Figur 8. Hyppigheden af rigtige svar som funktion af opgaverne ordnet efter stigende sværhed

(næsten) er en aftagende funktion af antallet af forkerte svar på opgaverne (figur 8), vil dette antal ordne opgaverne efter sværhed . Men Rasch ville et skridt videre. Han ville måle dygtighed og sværhed på en kvantitativ skala og ikke blot på en ordinal. Det forudsætter, at sandsynlighedsfunktionen specificeres, og Rasch viste, hvordan funktionen skal se ud for at gøre en kvantitativ måling mulig. Han gav også en anvisning på, hvordan dette kunne kontrolleres grafisk, uden at der dog her skal gås nærmere ind på disse grafer, idet der senere er udviklet mere præcise tests, som de nye data vil blive afprøvet med i næste afsnit.

Raschs konklusion om BPP's fire delprøver var, at Matrixopgaverne og Talrækkeopgaverne tilsyneladende kunne måles på en kvantitativ skala, mens Ordrelationerne og Figuropgaverne helt sikkert ikke kunne. Man behøver ikke at undre sig over, at de vel gennemarbejdede BPP dele ikke til fulde lever op til kravene i Raschs model. Modellen er så restriktiv, at en stor samling opgaver/spørgsmål sjældent som helhed vil kunne tilfredsstille den. Den repræsenterer idealet, som det kan være fornuftigt at have for øje, men det kan også være fornuftigt at stille sig tilfreds med mindre, hvis brugen af testen eller spørgeskemaet stiller mindre krav end en kvantitativ måling. Ellers kan man ende i den situation, at statistiske idealer gør det psykologiske indhold af et spørgeskema tandløst.

NY AFPRØVNING

De nye data stammer fra sessionerne i efteråret 2001 og foråret 2002 og de 2734 prøvehæfter er udvalgt, så landets syv udskrivningskredse er repræsenteret proportionalt med antallet af sessionsbehandlede, men det har ikke været praktisk muligt at foretage en egentlig tilfældig udvælgelse. Disse data testes i det følgende mod Raschs model, og det viser sig, at de ikke til fulde tilfredsstiller denne model. Det er dog ikke nogen stor overraskelse på baggrund af Raschs resultater i 1950'erne. Derefter vil det blive undersøgt, om de nye data kan tilfredsstille en mindre restriktiv model, Mokkens ikke-parametriske model.

Raschs model. En almindelig anvendt testmetode er baseret på, at testpersonerne opdeles i to grupper. For hver gruppe estimeres opgavernes sværhed (a-værdierne i figur 5), og det testes, om disse to sæt estimater adskiller sig

signifikant fra hinanden. Det må ikke være tilfældet, hvis Raschs model beskriver besvarelserne. Ved testen anvendes EDB programmet LPCM-WIN.

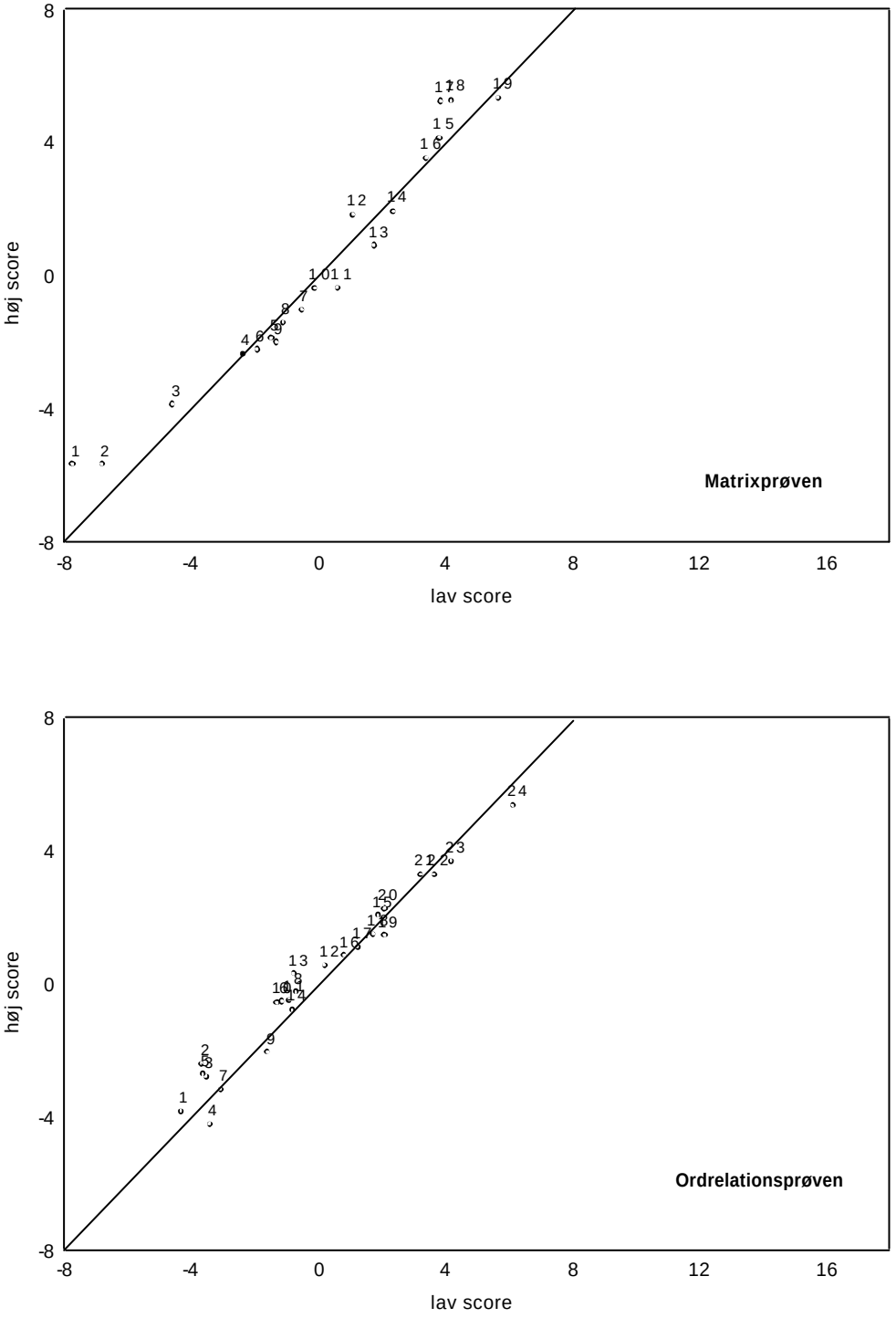
Opdelingen af personerne kan ske på mange måder f.eks. efter deres skoleuddannelse. Denne oplysning mangler dog for et stort antal, så i stedet for opdeles efter antallet af rigtige svar: den ene gruppe består af den halvdel (ca.), der har det laveste antal rigtige svar, og den anden gruppe er dem, der har det højeste antal rigtige svar. Den χ^2-test, som bruges til at sammenligne de to sæt sværhedsgrader, er udledt af professor Erling B. Andersen, der i sin værnepligtstid fungerede som statistiker ved Militærpsykologisk Tjeneste i 1963-64.

I figurerne 9 og 10 er de to sæt estimerede opgavesværheder tegnet op mod hinanden for hver delprøve. På figurerne er den identitetslinie indtegnet, som punkterne ligger på, hvis sværhederne estimeres til samme værdier i de to grupper. En hel del punkter afviger fra den tilhørende linie, f.eks. har opgave 13 i Matrixprøven en sværhed på 1.82 for dem med lav score og en lavere sværhed på 0.84 for dem med høj score. Andersens test viser da også, at forskellen mellem de to sæt sværheder for alle fire delprøver er stærkt signifikante ($p<0.1\%$). Tabellen nedenfor viser χ^2-testværdien med det tilhørende frihedsgradsantal samt 0.1% signifikansgrænsen, som i alle tilfælde er væsentlig mindre end testværdien.

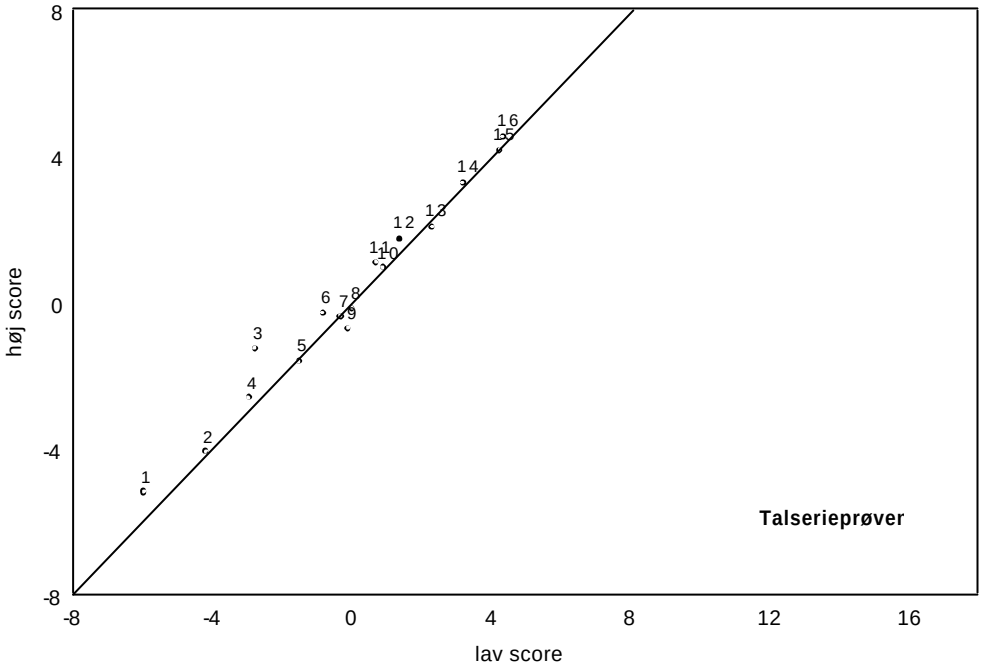
Deltest	χ^2**-testværdien**	**Frihedsgrader**	**0.1% grænsen**
Matricer	267.5	18	42.3
Ordrelationer	309.3	23	49.7
Talserier	135.8	15	37.7
Figurer	182.0	17	40.8

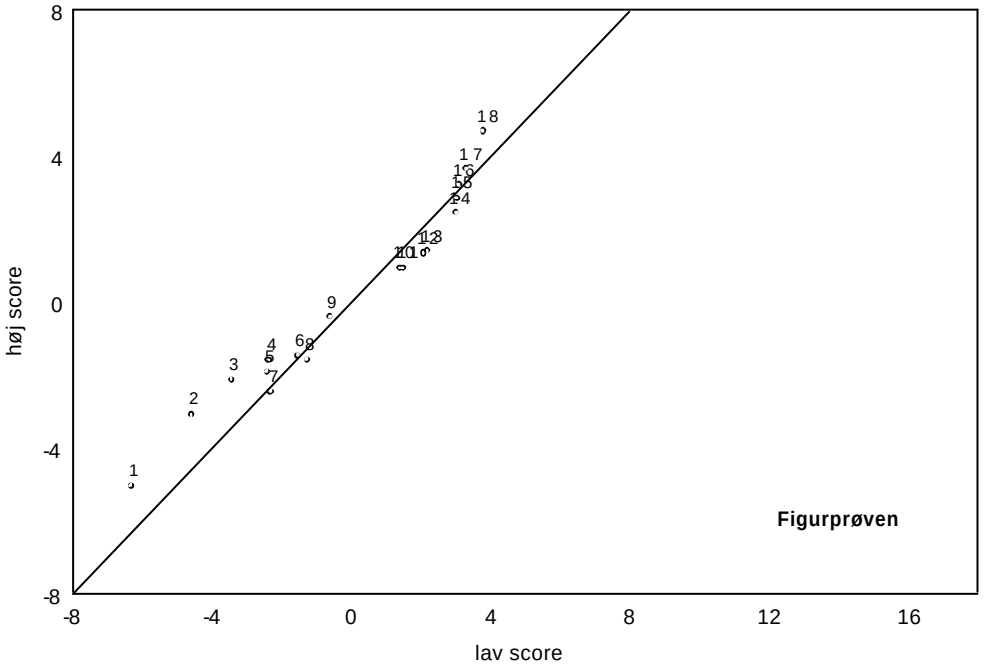
Mokkens model. Denne model testes i to trin

- Først undersøges om sandsynligheden for et rigtigt svar for alle opgaver er voksende funktioner af personernes dygtighed, jvf. figur 5. Det er tidligere omtalt, at hvis det er tilfældet, vil antallet af rigtige svar give en ordning af testpersonerne efter dygtighed. Dette kalder Mokken for monoton homogenitet.
- Hvis der er monoton homogenitet, undersøges om de samme sandsynlighedsfunktioner skærer hinanden. Som omtalt er det ækvivalent med, at sandsynligheden for et rigtigt svar som funktion af sværheden for



Figur 9. Raschs model. Opgavesværhed estimeret for personer med lav henholdsvis høj score





Figur 10. Raschs model. Opgavesværhed estimeret for personer med lav henholdsvis høj score

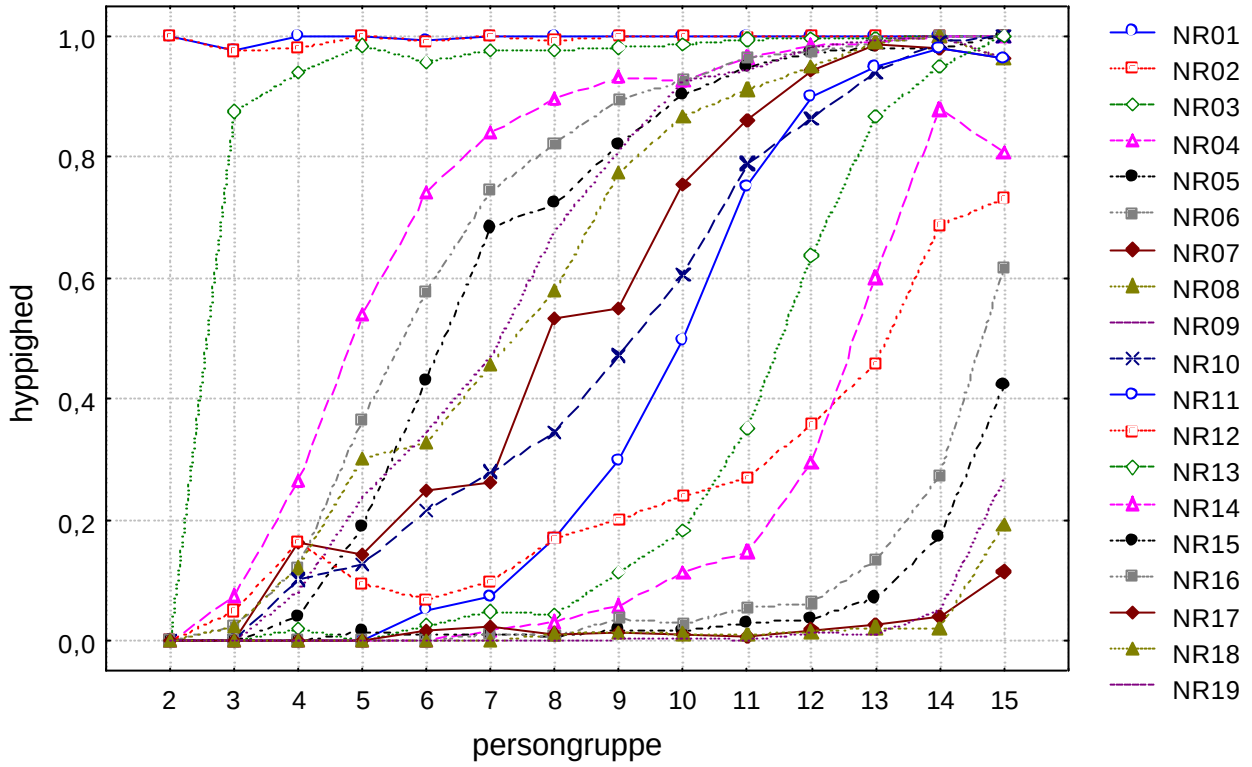
alle dygtigheder er aftagende funktioner, jf. figur 7. Mokken taler da også om dobbelt monotoni, mens Rasch talte om holomorfi. Hvis dette kan vises, vil antallet af forkerte svar på opgaverne ordne opgaverne efter sværhed.

Undersøgelsen af, om de nye data opfylder kravene i Mokkens model, sker ved hjælp af programmet MSP5. Før man tester vil det dog være naturligt at visualisere forholdet ved hjælp af figurer svarende til figur 6, hvis kurver er voksende, hvis der er monoton homogenitet, og ikke skærer hinanden, hvis der er dobbelt monotoni. Rasch lagde nogle opgaver og nogle persongrupper sammen for at udjævne mindre afvigelser. Det er ikke sket i figur 11, hvor kurverne vises for Matrixprøven, men persongrupperne 0 og 1 samt 16 til 19 er ikke indtegnet, da de tilsammen kun omfatter 5 personer. Stort set er kurverne for alle opgaver voksende, men der er dog nogle undtagelser, hvis signifikans undersøges ved hjælp af programmet.

Det er tydeligt, at opgave 12 afviger fra de andre opgaver ved at have en fladere kurve (mindre diskriminationsevne), således at kurven skærer flere af de andre kurver. F.eks. vil de dygtigste, persongrupperne 11-15, have lettere ved at klare opgave 13 end 12, mens det omvendte er tilfældet for persongrupperne 2-10. Matrixopgaverne som helhed kan derfor ikke være dobbelt monotone, men det er muligt, at hvis opgave 12 udelades, vil de andre 18 opgaver være dobbelt monotone. Dette kan også testes med programmet.

MSP5 tester monoton homogenitet ved for hver opgave at sammenligne hyppigheden af rigtige svar for hvert par af persongrupper, dvs. hvert punktpar på opgavens kurve i figur 11. Hvis hyppigheden er lavest for persongruppen med det højeste nummer testes, om faldet er signifikant. I programmet defineres persongrupperne lidt anderledes end i figur 11, idet
- svaret på den aktuelle opgave ikke indgår i den samlede sum af rigtige svar, som definerer persongrupperne.
- persongrupper med få personer lægges sammen, så der er mindst 50 personer i hver gruppe.

I tabellen under figur 11 vises resultatet af testen. Efter opgavenummeret vises, i hvor mange pct. af de parvise sammenligninger hyppigheden af rigtige svar er lavest for persongruppen med det højeste nummer - dvs. negative ændringer, der strider mod monoton homogenitet. Antallet af sammenligninger varierer



Figur 11. Matrixprøven. Hyppigheden af rigtige svar som funktion af det samlede antal rigtige svar.

opgave i	% med fald	antal signif. 5%	H_i
1	75	0	0.42
2	47	0	0.38
3	16	0	0.32
4	2	0	0.39
5	2	0	0.39
6	2	0	0.40
7	2	0	0.37
8	0	0	0.36
9	0	0	0.42
10	0	0	0.34

opgave i	% med fald	antal signif. 5%	H_i
11	0	0	0.44
12	4	0	0.13
13	2	0	0.42
14	0	0	0.38
15	5	0	0.20
16	8	0	0.26
17	62	0	0.03
18	64	0	0.11
19	14	0	0.31
total	**11**	**0**	**0.35**

meget for opgaverne på grund af sammenlægning af persongrupper; for matrixprøven er antallet mellem 18 og 55. Den næste søjle viser, at ingen af de negative ændringer er signifikante på 5%-niveauet. Alligevel må man nok være lidt skeptisk over for et par opgaver - nr. 1, 17 og 18 - på grund af det store antal ikke-signifikante negative ændringer.

Mokken bruger et mål H for opgavesættets skalerbarhed, som defineres på følgende måde: For hvert opgavepar (i,j) defineres H_{ij} som korrelationen mellem svarene (rigtigt, forkert) på de to spørgsmål divideret med den maksimale korrelation, der kunne opnås med de givne andele af rigtige svar. Hvis $H_{ij}=1$, betyder det, at alle der har svaret rigtigt på den sværeste opgave også har svaret rigtigt på den letteste opgave (en Guttman skala). Hvis besvarelserne på de to spørgsmål er uafhængige af hinanden, er $H_{ij}=0$, og hvis de, der svarer rigtigt på den sværeste opgave, har vanskeligere ved at klare den letteste opgave end dem, der har svaret forkert på den sværeste, er $H_{ij}<0$. For hver opgave (i) defineres H_i som en art gennemsnit af opgavens H_{ij}-værdier med de andre opgaver. Endelig defineres H også som en art gennemsnit af alle H_i-værdierne.

Mokken stillede to krav til H-værdierne for at betegne et opgavesæt skalerbart
- alle $H_{ij} > 0$
- alle $H_i = 0.30$, herved er H også større end 0.30.
Hvis alle $0.30 = H_i < 0.40$ tales om en svag skala, hvis $0.40 = H_i < 0.50$ betegnes skalaen moderat, og hvis alle $0.50 = H_i$ kaldes skalaen stærk.

Med 19 opgaver i Matrixprøven er der $19 \cdot 18/2 = 171$ H_{ij}-værdier, hvoraf 14 er negative. Af tabellen under figur 11 kan man yderligere se, at H_i er temmelig lav for opgaverne 12, 17 og 18. Matrixprøven er altså som helhed ikke skalerbar i Mokkens forstand, selvom H=0.35 for hele skalaen. Hvis man udelader de tre nævnte opgaver, vil de resterende 16 opgaver opfylde alle Mokkens betingelser for at være moderat skalerbare med et totalt H på 0.41. Det betyder ikke så meget om opgave 17 og 18 er med, idet kun ca. 1.5 pct. svarer rigtigt på disse opgaver, men opgave 12 med ca. 25 pct. rigtige svar er problematisk; den burde ikke være med. Det blev nævnt ovenfor, at opgave 12 også er problematisk med hensyn til dobbelt monotoni. Opgaven adskiller sig fra de fleste andre ved, at matricen kun indeholde to bogstaver (naboopgaverne har mindst seks bogstaver), og den ser derved umiddelbart ud til at være lettere, end den egentlig er. Samtidig står den øverst på den anden opgaveside, og det er derfor sandsynligt, at selv de, der har store problemer med de første 11 opgaver på første

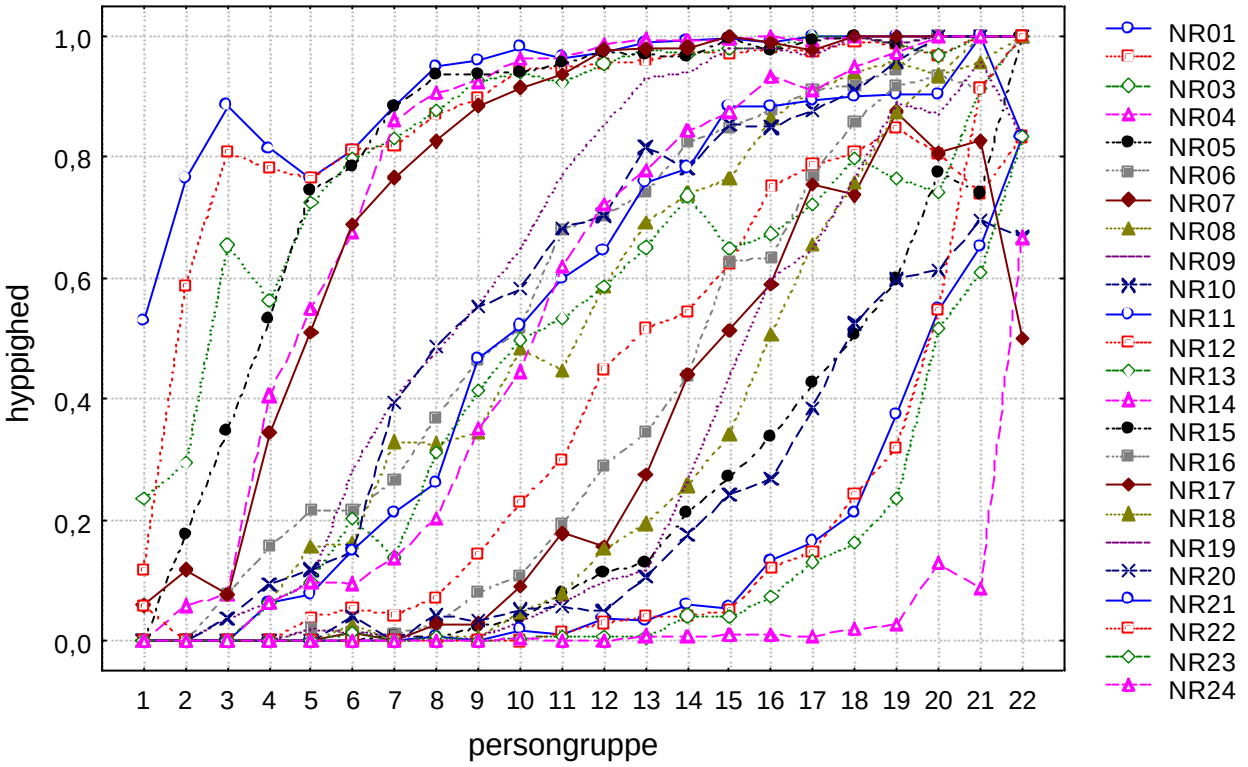
side, forsøger sig med opgave 12. Da der i realiteten - med kun to bogstaver - er et ret begrænset antal fornuftige løsningsmuligheder, har man en ikke ubetydelig chance for at gætte rigtigt, selvom man ingen ide har om systemet i opgaven. Dette er sikkert forklaringen både på det lave H_i og den lave diskriminationsevne.

I MSP5 programmet testes dobbelt monotoni på tre forskellige måder, som ikke vil blive beskrevet nærmere her. Som man kunne forvente, er der problemer med opgaverne 11 til 14. Alle problemer stammer dog fra opgave 12, så hvis den udelades, vil de øvrige opgaver være dobbelt monotone dog med det forbehold, at skalerbarheden for opgaverne 17 og 18 er problematisk.

Figur 12 over hyppigheden af rigtige svar for Ordrelationsprøvens 24 opgaver er lidt uoverskuelig på grund af de mange opgaver. Testene viser, at opgave 13 ikke helt følges med de andre opgaver, og med den viden kan man også se på figuren, at de dygtigste probander (persongrupperne større end 14) har uforholdsmæssigt svært ved at løse opgaven. Grunden kan være, at denne opgave - ligesom opgave 12 i Matrixopgaverne - er placeret øverst på den anden opgaveside og derfor tiltrækker sig større opmærksomhed fra de dårlige opgaveløsere end andre opgaver. En anden mulig forklaring er, at opgaven relaterer to genstande, hvoraf den ene næsten er gået af brug i de forløbne 50 år. Næsten alle kender sikkert genstanden af navn, men en del kender måske ikke dens egenskaber, og det er vigtigt for at løse opgaven.

Ingen af opgavesættets H_{ij}-værdier er negative, og af tabellen under figur 12 ses, at fem af faldene på figurens kurver er signifikante; to af dem findes på kurven for opgave 13. Tabellen viser også, at opgave 13 er den eneste, som har et H_i mindre end 0.30. Hele opgavesættet er altså tæt ved at være monotont homogent og skalerbart med H=0.37. Hvis opgave 13 udelades, forsvinder alle de fem signifikante fald, alle H_i er = 0.30, og det samlede H vokser til 0.39. De resterende 23 opgaver er altså monotont homogene og skalerbare.

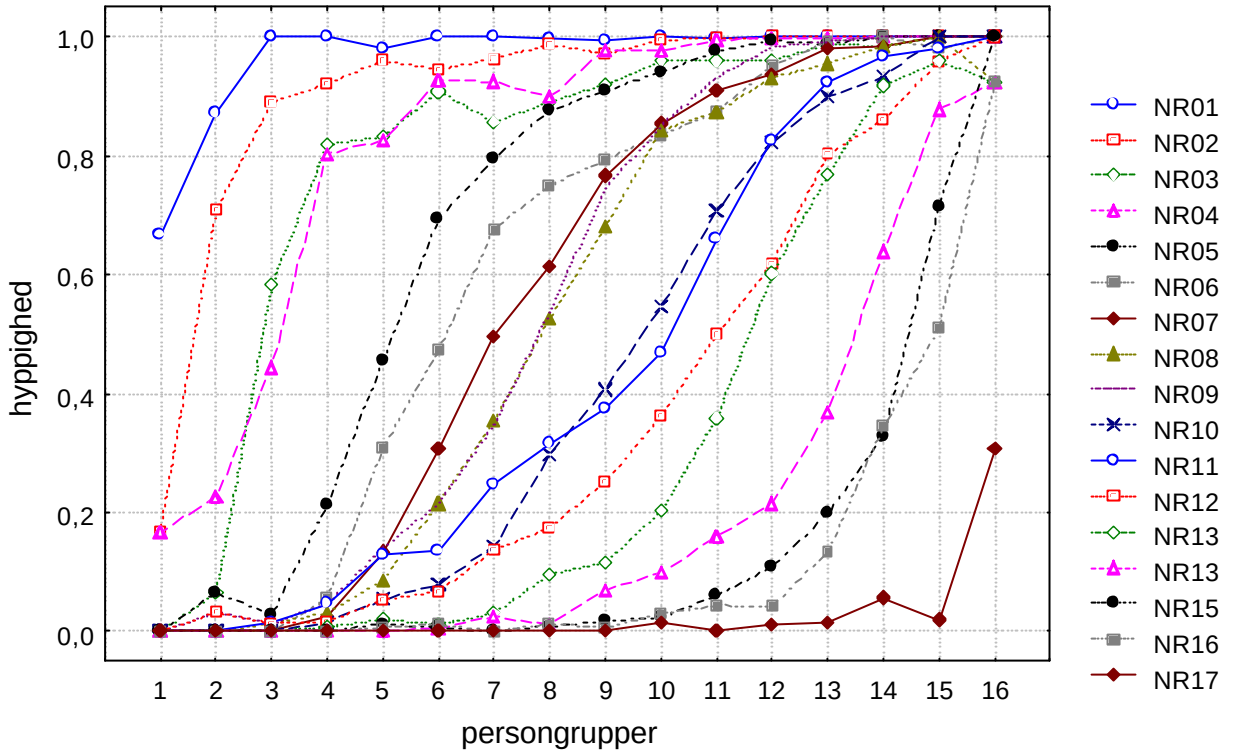
Når det drejer sig om dobbelt monotoni, er opgave 13 igen problematisk, og selv når den opgave udelades, er der stadig mindre afvigelser for en del opgaver, så man må konkludere, at sættet skal reduceres ret meget, før en delmængde af opgaverne er dobbelt monotone.



Figur 12. Ordrelationer. Hyppigheden af rigtige svar som funktion af det samlede antal rigtige svar.

opgave i	% med fald	antal signif. 5%	H_i
1	11	0	0.42
2	7	0	0.30
3	9	0	0.37
4	8	0	0.54
5	9	1	0.42
6	2	0	0.32
7	11	1	0.49
8	3	0	0.32
9	4	0	0.46
10	5	1	0.31
11	7	0	0.33
12	5	0	0.34

opgave i	% med fald	antal signif. 5%	H_i
13	10	2	0.23
14	1	0	0.39
15	2	0	0.35
16	3	0	0.40
17	8	0	0.38
18	7	0	0.41
19	6	0	0.44
20	8	0	0.34
21	6	0	0.39
22	7	0	0.43
23	19	0	0.44
24	21	0	0.47
total	**7**	**5**	**0.37**



Figur 13. Talrækker. Hyppigheden af rigtige svar som funktion af det samlede antal rigtige svar.

opgave i	% med fald	antal signif. 5%	H_i
1	64	0	0.59
2	7	0	0.48
3	6	0	0.34
4	8	0	0.52
5	2	0	0.58
6	2	0	0.47
7	0	0	0.51
8	0	0	0.51
9	0	0	0.54

opgave i	% med fald	antal signif. 5%	H_i
10	0	0	0.47
11	0	0	0.43
12	0	0	0.43
13	2	0	0.51
14	2	0	0.47
15	9	0	0.48
16	11	0	0.47
17	14	1	0.45
total	**5**	**1**	**0.48**

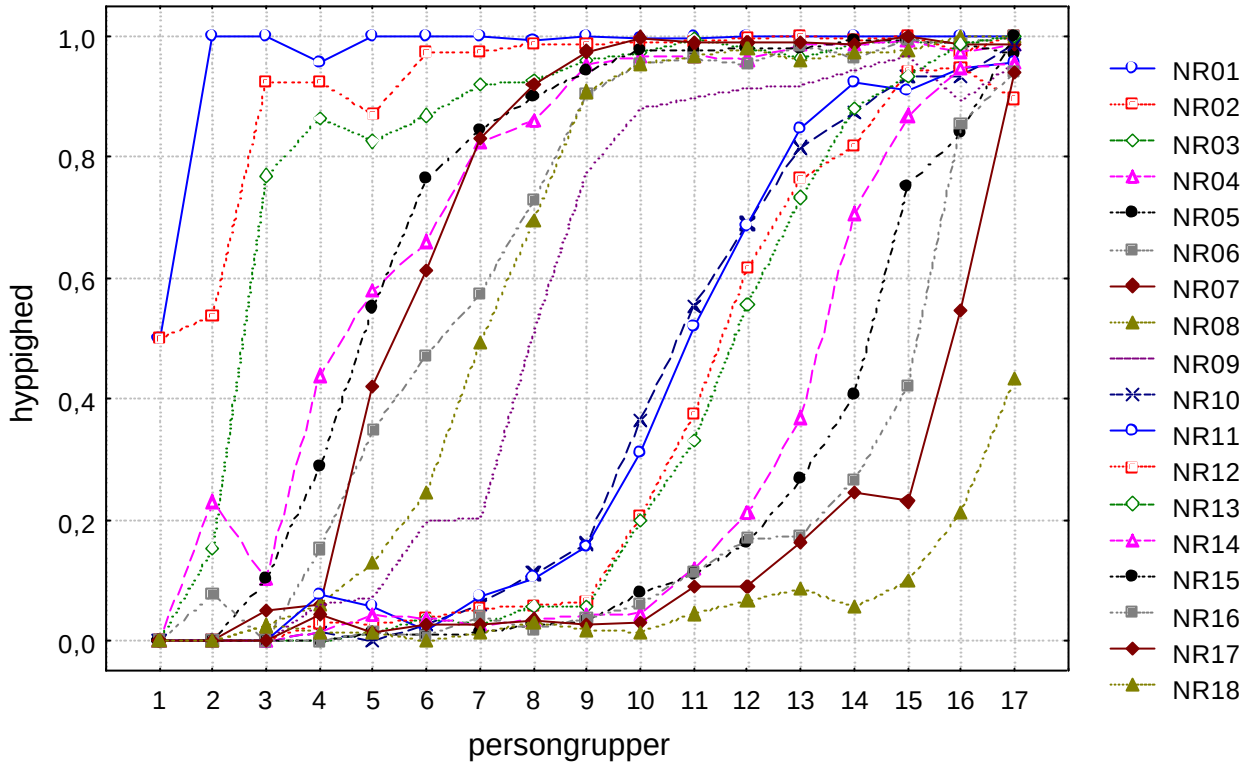
Delprøven med Talrækker er mere homogen end de to foregående delprøver - prøvens opgaver er illustreret i figur 13. Alle H_{ij} er positive, og alle H_i er = 0.43 bortset fra H_3, som dog også er større end 0.30. Der er et enkelt signifikant fald på kurven for opgave 17, men det er et fald på blot 0.02 mellem persongrupperne 10 og 11, som man næppe behøver at tillægge større betydning. Alle testene af dobbelt monotoni falder positivt ud, så hele opgavesættet er dobbelt monotont.

I delprøven med Figurer er der to negative H_{ij} værdier - begge optræder for opgave 1. Endvidere ses af tabellen under figur 13, at der er et stort antal negative ændringer på kurven for denne opgave. Ingen af disse er dog signifikante, og H_1 er høj =0.53. Opgave 1 er så let, at kun 14 af de 2734 probander har fejl i den, og det drejer sig sikkert i hovedsagen om sjuskefejl, idet nogle med helt op til 10 og 11 rigtige i de øvrige 17 opgaver har fejl i opgave 1. Under alle omstændigheder har det med 99.5 pct. rigtige svar minimal betydning, om opgave 1 medtages eller ej. De øvrige 18 opgaver er monotont homogene med ret høje H_i-værdier og H=0.50.

Med hensyn til dobbelt monotoni er der lidt uklarhed med opgaverne 4 og 7, som i nogen grad krydser ind over nr. 5 og 6, se figur 13. De øvrige opgaver er dobbelt monotone.

Konklusion. Da Rasch i 1950'erne afprøvede sin model på BPP's delprøver, stod det klart, at Ordrelationerne og Figurprøven ikke levede op til modellens krav. Derimod mente Rasch, at de to andre prøver, Matrix- og Talserieprøven, kunne beskrives med hans model. Det ser dog ikke ud til at være rigtigt, når de nye data holdes op mod modellen. Årsagen til det ændrede resultat er næppe, at besvarelsesmønstrene har ændret sig men snarere, at man har stærkere tests til rådighed, end den - i hovedsagen - grafiske metode, som Rasch måtte benytte.

Hvis BPP's delprøver havde opfyldt Raschs model, ville der have været tale om en kvantitativ måling af evnen til at løse opgaver af de pågældende typer. Det er imidlertid ikke afgørende for forsvarets brug af BPP, at testen giver en kvantitativ måling. Man skal blot have probanderne rangordnet efter deres dygtighed, og det stiller væsentligt mindre krav til besvarelserne af opgaverne. Ved hjælp af Mokkens model er det undersøgt, om delprøverne er monotont homogene og skalerbare og herved kan give den ønskede rangordning af probanderne.



Figur 13. Figurer. Hyppigheden af rigtige svar som funktion af det samlede antal rigtige svar.

opgave i	% med fald	antal signif. 5%	H_i
1	64	0	0.53
2	22	0	0.37
3	10	1	0.31
4	7	0	0.42
5	9	0	0.47
6	3	0	0.50
7	16	0	0.57
8	6	0	0.56
9	2	0	0.51

opgave i	% med fald	antal signif. 5%	H_i
10	1	0	0.50
11	4	0	0.50
12	2	0	0.51
13	3	0	0.54
14	7	0	0.54
15	1	0	0.52
16	4	0	0.48
17	12	0	0.46
18	16	1	0.42
total	**8**	**2**	**0.50**

Resultatet fremgår af skemaet

Delprøve	Antal opgaver	Skalerbar, hvis følgende opgaver udgår	Dobbelt monoton, hvis disse opgaver også udgår
Matrixprøven	19	nr. 12, 17, 18	ok
Ordrelationer	24	nr. 13	mindst halvdelen
Talserieprøven	17	ok	ok
Figurprøven	18	måske nr. 1	måske nr. 4, 7

Skemaet viser, at Talserieprøven er skalerbar i sin helhed, mens der er enkelte mangler ved de tre andre prøver. Fem opgaver fordelt på disse prøver er ikke skalerbare, men kun de to burde helt sikkert udgå. Det er opgave 12 i Matrix-prøven og opgave 13 i Ordrelationerne. De tre andre problematiske opgaver er enten så lette eller så vanskelige, at de kun har betydning for meget få proban-der. Det er påfaldende, at begge de to mest problematiske opgaver står øverst på den pågældende delprøves anden side, og det er tænkeligt, at en anden placering ville have løst problemet - men til gengæld ville en anden opgave på denne plads måske så have givet problemer.

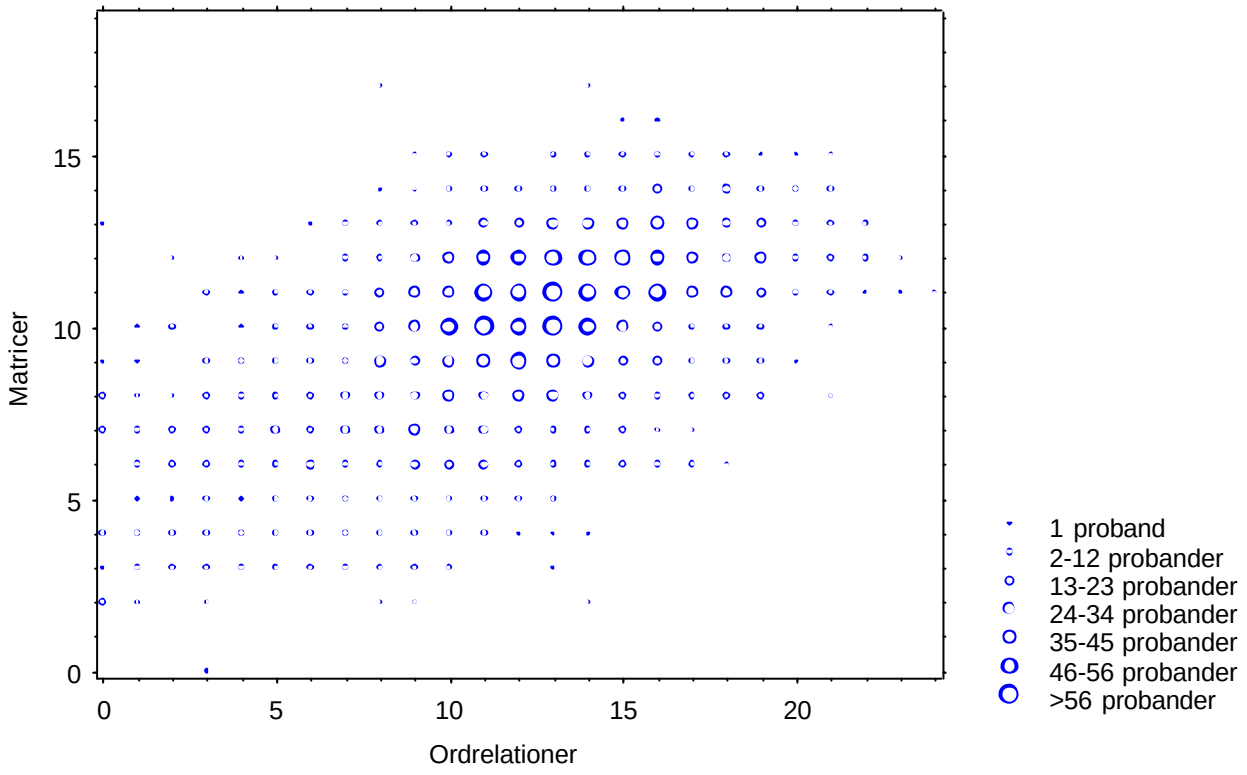
De skalerbare opgaver er også dobbelt monotone i Matrix- og i Talserieprøven, mens der muligvis skal udelades yderligere to opgaver i Figurprøven, før den er dobbelt monoton. Ved Ordrelationerne skal der udelades så mange ekstra opga-ver for at opnå dobbelt monotoni, at det ikke giver mening at forsøge at rang-ordne denne delprøves opgaver.

BPP SCORE

Delprøverne giver fire raw scores (antal rigtige svar) for probanderne. Bortset fra de få afvigelser fra monoton homogenitet som blev påvist i foregående afsnit, giver disse raw scores hver sin rangordning af probanderne efter deres evne til at løse opgaver af de fire typer. Hvis disse rangordninger er (næsten) ens, betyder det, at det er den samme evne, der estimeres af alle delprøverne, men man kan let overbevise sig om, at det ikke er tilfældet.

I figur 14 er antallet af rigtige svar på de to første delprøver plottet ind for de 2734 probander. Det er tydelig, at nok er der en positiv sammenhæng mellem de to antal rigtige svar, men der er også plads til store afvigelser, f.eks. varierer antallet af rigtige svar i Matrixprøven fra 3 til 15 for de 213 probander, der har

10 rigtige svar i Ordrelationerne. Evnerne til at løse de to slags opgaver er altså beslægtede, men de er på ingen måde identiske. Det samme gælder for de andre delprøver, hvilket fremgår af tabellen under figur 14, som viser Spearman's rangkorrelationer mellem delprøverne (BPP omtales senere). Alle prøverne har noget til fælles - Figurprøven mindst - men de er samtidig meget forskellige.



Figur 14. Plot af antal rigtige svar på to delprøver for 2734 probander.

Spearman's rangkorrelationer for BPP's delprøver.

Prøve	Matricer	Ordrelationer	Talserier	Figurer	BPP
Matricer	1.00	0.55	0.57	0.46	0.77
Ordrelationer	0.55	1.00	0.57	0.46	0.84
Talserier	0.57	0.57	1.00	0.41	0.80
Figurer	0.46	0.46	0.41	1.00	0.72
BPP	0.77	0.84	0.80	0.72	1.00

Ved den praktiske anvendelse af BPP har det altid været det samlede antal rigtige svar, der har været brugt som raw score - delprøvernes resultater har ikke været anvendt. BPP scoren er altså et udtryk for en "opsummering" af de fire evner, som Prien anså for væsentlige bidrag til at beskrive en persons intelligens. Man kan derfor ikke forvente, at de 78 opgaver som helhed skal være monotont homogene. En analyse ved hjælp af Mokkens model viser da også, at der højst kan udvælges 45-50 opgaver, som tilsammen er skalerbare.

Den rangordning af probanderne, som BPP scoren giver, er snævert forbundet med delprøvernes rangordninger - jvf. tabellen under figur 14, men den kan ikke tolkes som en rangordning efter en bestemt evne hos probanderne. Der er tale om et blandingsprodukt, som gennem årene har bevist sin store anvendelighed.