Peter Allerup
Professor,
DPU - Danmarks institut for
Pædagogik og Uddannelse,
Aarhus Universitet

Nationale test

Ønskede egenskaber

og indfriede krav?

De danske nationale test fulde navn er egentlig "De Nationale It-baserede Adaptive Obligatoriske Test". Det fulde navn fremhæver, til forskel fra det forkortede, flere vigtige sider af disse test i sammenligning med mange andre typer af test. Altså nogle signaler om, hvad der gør netop disse test specielle. Testene er resultatet af nogle krav om udvikling af nye evalueringsinstrumenter, som fandt sted for ca. 15 år siden, og i dag indgår testene rutinemæssigt i skolens hverdag. Derfor er det også et passende tidspunkt at spørge, om nogle af de krav og intentioner, der oprindeligt lå bag ved udviklingen, faktisk er blevet indfriet og kan demonstreres opfyldt gennem egenskaber, som man kan undersøge via indsamlede data i dag. Denne artikel har til formål at beskrive et par af de oprindelige krav: Sammenlignelighed af testresultater, adaptivitet i forhold til elevers færdighedsniveau og evne til at formidle brugbare evalueringsresultater – set i lys af deres aktualitet i den pædagogiske verden dengang og nu. Vi kan med et kik over skulderen i dag spørge, om det blev, som vi ønskede og krævede det – eller om der mangler noget?

Lidt historisk

Indtil udgangen af 1990'erne benyttede danske lærere sig, i den daglige evaluering, hovedsageligt af test, som de enten selv fremstillede, eller som de hentede på det dengang eksisterende Dansk Psykologisk Forlag (i dag Hogrefe). Der var tale om både formative (diagnostiske) og normative test og på det øverste plan – Undervisningsministeriet – administrerede man de årlige afgangsprøver der, som normative prøver, alene havde til formål at udstyre eleverne med en passende fornemmelse af, hvor deres præstation lå sammenlignet med andre elevers samt at give Undervisningsministeriet en overordnet fornemmelse af elevernes præstationsniveau. Trods små tilløb til at ændre på de grundlæggende procedurer, fx i dansk retstavning, hvor der en overgang anvendtes målorienteret karaktergivning, var det klart for alle, at de anvendte normative prøver og test manglede nogle fundamentale egenskaber. Det drejede sig først og fremmest om egenskaber vedrørende sammenlignelighed mellem resultater opnået fra ét tidspunkt, fx et givent år, til et andet tidspunkt og måske med en anden elevgruppe, sådan som tilfældet er, når man vil studere udviklingen af præstationer i afgangsprøven i 9. klasse. Når man ser bort fra nogle bestemte af Dansk Psykologisk Forlags test, var det indlysende umuligt at give mening til sammenligninger mellem forskellige år og elevpopulationer. Det er lidt ironisk, at det var en dansk statistiker, Georg Rasch, som i 1960'erne formulerede nogle modeller (Rasch 1960), som passende anvendt over for konkret testkonstruktion faktisk gjorde det muligt at foretage 'objektive' sammenligninger af den nævnte type over forskellige tidspunker og over forskellige elevpopulationer. Ironien består i, at mens man i de nationale NAEP test i USA allerede i 1980'erne indførte konstruktionsprincipper ud fra disse såkaldte Rasch Modeller, skete det kun i Danmark i forbindelse med nogle af Dansk Psykologisk Forlags test – og her hovedsageligt begrundet i nogle familiære relationer til forlaget. Det officielle Danmark kørte videre som hidtil, og det var først i slutningen af 1990'erne, at Undervisningsministeriets satsning på D*anmarks strategi for uddannelse, læring og IT* med undertitlen *Vi skal videre* fremlagde planer for, hvordan de uddannelsespolitiske målsætninger fremover kunne styrkes ved hjælp af IT, herunder brug af den moderne teknologi, i forbindelse med test af elever.

Det var ministeriets uddannelsesdirektør Kim Mørch Jacobsen, som efter en rejse til Norge i starten af 00'erne viderebragte de første idéer om konstruktion af It-baserede adaptive test, udarbejdet på de moderne psykometriske vilkår, som Rasch modellen specificerer. Den validitet, som testen sikres netop gennem Rasch Modellen, løser de nævnte problemer med objektive sammenligninger over tid og over elevpopulationer. Principperne for Rasch Modellen gennemgås kort nedenfor (se også JAM 2004). Forskellige arbejdsgrupper fremkom med tanker om de praktiske og teoretiske problemer, som skulle løses i forbindelse hermed, men det afgørende skub til igangsættelse af arbejdet med nationale test kom fra en helt anden vinkel.

I december 2003 fremlagdes resultaterne af 2. runde af den internationale PISA undersøgelse, den 1. runde var i år 2000, og undersøgelsens resultater var i Tyskland årsag til en ophidset stemning. Stemningen i Danmark var helt modsat, og op til det planlagte pressemøde så det ud til, at der var en gennemgående accept af resultaterne. Men resultaterne blev pludselig, af daværende undervisningsminister Ulla Tørnæs, på selve pressemødet, drejet til at være en 'katastrofe' for det danske undervisningssystem. Der var ikke mange af de, som havde kendskab til de faktiske resultater, der var enige i den udlægning, men den politiske handlekraft satte omgående ind, og næsten samme dag blev forligspartierne indkaldt med henblik på at gøre noget aktivt for at løse problemerne omkring dette 'dårlige' PISA resultat: Indførelsen af obligatoriske nationale test.

I næste afsnit beskrives nogle centrale krav, som blev formuleret i forbindelse med projektudbuddet. Det blev bragt i udbud i to omgange, første gang med en økonomisk ramme på 32 mio. kr, hvilket var helt utilstrækkeligt, og derefter en ny udbudsrunde med en økonomisk ramme på 62 mio. kr. I denne anden udbudsrunde modtog undervisningsminister Bertel Haarder to seriøst gennemskrevne tilbud inden for den økonomiske ramme. Danmarks Pædagogiske Universitet (DPU) var den ene byder og det rådgivende ingeniørfirma Cowi Consult stod for det andet. En underleverandør til DPU's tilbud sprang fra i sidste øjeblik, og dette medførte at tilbuddet var ugyldigt, og arbejdet blev dermed lagt i hænderne på Cowi Consult, som iflg. egne beskrivelser fokuserer deres spidskompetencer således, at "Med eksperter i verdensklasse inden for ingeniørkunst, miljø og samfundsøkonomi angriber vi udfordringerne fra mange forskellige vinkler, så vi skaber mere sammenhængende løsninger for vores kunder" (cowi.dk). Med flere forsinkelser og løbende udvidelse af den økonomiske ramme til mere end 110 mio. kr., inklusive bøder til Cowi, overtog Undervisningsministeriet ansvaret for testene ca. i 2010. Det var allerede fra starten aftalt mellem forligspartierne, at der efter nogen tid med praktisk anvendelse af testene i skolen skulle laves en opfølgende evaluering af forskellige aspekter af implementeringen af testene. Denne blev gennemført af Rambøll (2013).

Nogle ønsker om egenskaber ved de nationale test

Allerede fra starten ønskede man, at der blev udviklet test for mange klassetrin og i så mange fag som muligt. Dermed lå også i luften, som et stærkt ønske, at man skulle kunne sammenligne resultaterne fra år til år, således at effekten af centrale initiativer, fx skolereformer kan måles og evalueres dynamisk over en række år. Med de eksisterende afgangsprøver er dette, som nævnt, umuligt. Der findes i øjeblikket 10 obligatoriske test i skolens fag samt to frivillige test i faget dansk som andetsprog, men det er hensigten at udvide listen både mht. fag og klassetrin.

De centrale dele af den nye skolereform belyser alle nogle egenskaber vedrørende sammenlignelighed mellem elever fra forskellige klasetrin og mellem de samme elever set over flere år:

- **Alle elever bliver udfordret, så de bliver så dygtige, som de kan**
- **Andelen af de allerdygtigste elever i dansk og matematik skal stige år for år**
- **Andelen af elever med dårlige resultater i de nationale test for læsning og matematik skal falde år for år**
- **Eleverne skal på sigt kunne det samme i 8. klasse, som de i dag kan i 9. klasse**

Det står allerede ret klart fra denne forkortede oversigt, at det, man ønsker, er, at de nationale test kan leve op til egenskaber om at gøre sammenligninger over år og mellem elevgrupper mulige, således at man fra et resultat fra de nationale test fra en konkret elev skal kunne sammenligne dels med sig selv på et senere tidspunkt, dels med andre elevers resultater. Der kræves faktisk, at de nævnte sammenligninger skal kunne udføres på 'forskningsmæssigt valide vilkår', hvilket er noget mere end blot en mulighed.

Det hører med til billedet, at skolereformen jo er iværksat *efter* indførelsen af de nationale test og derfor ikke kan tage æren for at generere de krav til egenskaber, som lige er nævnt.

Set fra en målteoretisk synsvinkel, også kaldet psykometrisk vinkel, må man efter løsning af en række tekniske problemer omkring implementeringen af de nationale test i dag sige, at de, som ét af de få af slagsen i verden, faktisk indfrier disse krav, der kort refereres til som egenskaben at kunne formidle 'objektive sammenligninger'.

Mens 'objektivitets'-egenskaben ikke har sat sig spor i det fulde navn for testen, har en anden egenskab til gengæld gjort det: 'Adaptiv'. Når testene kaldes adaptive, hentydes der til, at selve testafviklingen foregår ved, at eleven løbende udsættes for opgaver, som er passende i forhold til elevens dygtighed. Det betyder konkret, at man tilstræber, at svage såvel som stærke elever får præsenteret opgaver i testforløbet, som de har ca. 50% chance for at svare rigtigt på. Ved starten af et testforløb ved man ikke, om man har med en svag eller stærk elev at gøre, og man starter derfor i 'midten' med en middelsvær opgave. Afhængigt af om eleven kan svare rigtigt eller forkert på opgaven (faktisk et par opgaver og ikke kun én opgave) vælges derefter nye opgaver, som er enten sværere eller lettere end den (de) første opgaver, og man bruger alle de opnåede svar til at beregne et skøn over, om man har en stærk eller svag elev foran sig. Valget af opgaver foregår ved, at der er opbygget en opgave-bank bestående af mange opgaver inden for hvert fag og med forskellige sværhedsgrader. Hver gang eleven skal præsenteres for en ny opgave, vælges den næste opgave tilfældigt blandt de opgaver i opgavebanken, som har den valgte 'tilpassede' sværhedsgrad, og som ligger i det faglige domæne, som eleven i øjeblikket bliver testet i. Inden for hvert af folkeskolens hovedfag har praksis udviklet sig sådan, at der er skabt tre faglige sub-domæner. I matematik er der fx tale om algebra, geometri og anvendelse. Tilsammen udgør resultaterne fra de tre domæner en profil af eleven, hvilket er årsagen til, at nogle betegner de nationale test som 'profiltest'. Eleven præsenteres løbende for nye opgaver inden for hvert af de tre sub-domæner eller profilområder, indtil der er en vis statistisk sikkerhed for at kunne sige, at eleven kan besvare endnu en opgave af samme sværhedsgrad som den sidste med chancen 50%. Denne sikkerhed afhænger af antallet af opgaver eleven har været igennem og det faktiske forløb med skiftende sværere/lettere opgaver, som eleven har oplevet. Som regel vil eleven opleve, at den statistiske sikkerhed er på plads ved besvarelsen af ca. 20 opgaver i hvert profilområde.

Den adaptive egenskab blev i sin tid tænkt ind i konstruktionen af de nationale test af flere grunde. Dels eksisterede der allerede ganske mange adaptive test på markedet, man kan prøve en del af dem ved at aktivere de såkaldte CAT test (**C**omputerized **A**daptive **T**esting) på web. Dels lå der en drivkraft i at kunne tilbyde især svage elever en mulighed for at opleve at svare rigtigt på flere opgaver, end de var vant til under sædvanlige test. Endelig kan man sige, at når man er i gang med at opgradere betydningen af IT i undervisningen og ved testsituationer, jf bemærkningerne oven for, så ligger det lige for at udnytte situationen med eleven foran en computer til at gå væk fra det 'stive' såkaldte lineære testsystem på papir med et fast antal opgaver, som er ens for alle elever.

Set i bakspejlet må man erkende, at der findes både positive og negative sider af den måde, den adaptive egenskab er implementeret på ved de nationale test. Det positive findes hurtigt frem i den omstændighed, at det for første gang nu er blevet muligt at teste elever med helt forskellige (adaptivt tilpassede) opgaver og alligevel være i stand til at sammenligne testresultaterne. De nationale test udnytter her objektivitetsegenskaben i samklang med den adaptive egenskab under Rasch modellen til at kunne gennemføre valide sammenligninger og målinger, som ikke har kunnet gennemføres før. PISA og IEA's TIMSS og PIRLS-undersøgelser opererer alle med mange adskilte opgavehæfter opbygget af forskellige opgaver og trækker i princippet, som de nationale test på samme objektivitetsegenskab: At kunne sammenligne elever, der ikke har besvaret de samme opgaver. De negative sider, af den måde det adaptive princip virker, er detaljeret beskrevet i Allerup (2013), og her skal alene nogle få hovedforhold nævnes: Trods passende elevinstruktion ser det ikke ud til, at eleverne forstår, at det er ' maskinen' og ikke dem selv, der bestemmer det antal opgaver, de får stillet – der går prestige i at få så få opgaver stillet som muligt. Det er tilsvarende lidt 'flovt' at være den sidst testede i computerlokalet. Alle andre

har forladt lokalet/er stoppet, og man sidder alene tilbage. Også her er der tale om manglende forståelse af de tekniske præmisser bag ved testen. Fra kvalitative analyser af testforløbet blev det klart, at en del elever gik meget op i selve antallet af opgaver, de var mest interesserede i at få så mange opgaver som muligt – hurtigst muligt! (Kousholt 2012). Det er ikke nemt at forstå, hverken for dygtige eller svage elever, at når de mødes efter testsessionen og udveksler erfaringer om, at begge parter har løst ca. 50% af opgaverne korrekt, så er det faktisk fint i overensstemmelse med deres gensidige, ret præcise fornemmelse af hinandens daglige, faglige niveau. Fordi tildelingen af 'næste opgave' foregår tilfældigt ud fra maskinens valg af opgaver med passende sværhedsgrad, opleves det tit, at den samlede række af opgaver ikke giver fagligt mening for eleven. Den manglende sammenhæng skal ses i relation til, at fx PISA, TIMSS og PIRLS alle benytter sig af opgaver, hvor man fx gennemlæser større tekststykker først, før man mødes af de opgaver/spørgsmål, som testen er bygget op af.

Endelig gælder det ved alle former for evaluering, at selve formidlingen af resultaterne er en vigtig del af processen. Dette gælder ikke mindst formidling af resultater fra en test, fordi der her ofte kræves brug af forskellige tekniske størrelser for at kunne forstå betydningen af testresultatet. I forbindelse med mange af folkeskolens præstationstest har det været gængs praksis at regne antallet af rigtige besvarede opgaver ud og bruge dette tal, når eleven skal have en vurdering af, om resultatet er 'godt' eller 'mindre godt'. Et bestemt antal rigtige besvarelser fører til en bestemt karakter på 7-trinsskalaen. Omsætningen mellem antal rigtige og en karakter gælder alene for den konkrete prøve og over for de konkrete elever, som har taget prøven. Det er på forhånd fastlagt hvor mange procent, der skal have de forskellige karakterer [1]. Man kan derfor ikke sammenligne præstationerne fra år til år og mellem elever, som ikke har deltaget i samme prøve. Mange tror, at man er bedre stillet med formidlingen af resultatet ved at udregne antallet af rigtige i *procent* i stedet for blot at gå ud fra selve antallet af rigtige besvarelser. Men skal procenten så udregnes i forhold til alle opgaver eller, som Dansk Psykologisk Forlag har gjort det i mange år, som en procent i forhold til det antal opgaver, som eleven har nået at se på? Uanset valget får man imidlertid ikke et 'rigtigere mål' for elevpræstationen.

Ved de nationale test bliver eleverne løbende præsenteret for nye opgaver, indtil der er ca. 50% chance for at svare rigtigt på den næste opgave. Det er derfor umuligt at formidle antallet af korrekt løste opgaver eller en form for procent rigtige som det egentlige mål for elevdygtigheden. Som omtalt hører det med til selve konstruktionsgrundlaget for opgaverne i opgavebanken at en bestemt statistisk model, Rasch Modellen, skal kunne beskrive sandsynligheden for, at en elev svarer korrekt på opgaven. Faktisk blev opgavebanken i første omgang konstrueret ud fra rene faglige principper og hensyn, men efterfølgende måtte ca. halvdelen af opgaverne udgå på grund af manglende tilpasning til Rasch Modellen [2].

1 *10% af eleverne på karakteren 12, 25% på 10, 30% på 7, 25% på 4 og 10% på karakteren*

2 *Hvorved verden i forhold til normal praksis jo sættes på hovedet: Normalt forkaster man den statistiske model, hvis modellen ikke passer til 'virkeligheden' (data), men her prioriterer man modellens 'evne' til at danne objektive sammenligninger og laver om på virkeligheden (dvs. finder en ny opgave).*

Den simpleste Rasch model opererer med to sæt parametre: θ1,..., θk, som måler (k) opgavesværheder og σ1,..., σn, som måler (n) elevernes færdigheder (JAM,2004 og Allerup,1994, 2007). Med disse to sæt teoretiske mål (sværhed = θi og færdighed = σv) er Rasch sandsynligheden for et korrekt svar til item nr. i fra elev nr. v (dvs. avi=1) følgende:

$$p(a_{vi} = 1) = \frac{e^{\theta_i + \sigma_v}}{1 + e^{\theta_i + \sigma_v}}$$

Denne statistiske model er speciel ved på enntydig facon at knytte nogle basale praktiske krav mht. brugen af resultaterne fra en test (scoreværdierne) til opfyldelsen af Rasch Modellen. Rasch viste, at de tre følgende udsagn er ækvivalente: (i): elevscorerne (og item scorerne) udtømmer al viden om 'sværhed' og 'dygtighed' (sufficiens), (ii): Det er muligt at sammenligne elevfærdigheder med en hvilken som helst subgruppe af items (skal føre til samme resultat) og (iii) Rasch modellen er en gyldig statistisk beskrivelse af data (rigtigt/forkert) fra(n) elever til (k) opgaver.

Den første egenskab kan kaldes en 'validering' af den praktiske anvendelse af elevscorer. Den anden egenskab kaldes 'specifik objektivitet' og er ekstrem anvendelig ved testsituationer, hvor ikke alle elever får de samme opgaver, herunder De Nationale Test, PISA og IEA's TIMSS og PIRLS undersøgelser.

Når en elev sidder foran computeren ved Nationale Test kendes sværhedsgraderne for samtlige opgaver på forhånd fra en tidligere omfattende test. Elevens færdighedsniveau kendes i sagens natur ikke på forhånd, men undervejs i det adaptive system, trin for trin, opgave for opgave, ny-beregnes et mål for elevdygtigheden (ud fra samtlige foregående svar) og computeren vælger 'næste' opgave, så det forventes, at der er 50% chance for at svare rigtigt.

Den skala, som elevdygtigheden tilhører, ligger fra ca. -3.00 til 3.00 med den 'neutrale' elev målt i midten til værdien 0. Det er samme statistiske model, man anvender ved de internationale PISA test og IEA's TIMSS og PIRLS undersøgelser. Men her har man parallelforskudt de oprindelige skala til at være en skala omkring 500 ± ca. 200 i stedet for 0.00 ± 3.00, en ren matematisk manøvre, som ikke har betydning for fortolkningen, men måske respekterer den kendsgerning, at negative værdier sender forkerte signaler til modtageren. Lidt som den diskussion der udspillede sig ved fastlæggelsen af numeriske værdier på 7-trinkarakterskalaen med karakteren -3. Ved de internationale evalueringer har man vænnet sig til at aflæse elevpræstationerne på skalaen centreret omkring 500, og når danske elever ligger på 492, ved alle, at det er under gennemsnittet. Men hvor meget? Betyder de 8 point virkelig noget, som når man fx går fra karakteren 10 til fx 4 på 7-trinskalaen? Der hersker nogen uklarhed omkring dette spørgsmål, fordi der dels er en statistisk vinkel på det (signifikant forskel eller ej) og dels en politisk. Når det politiske system imidlertid har 'vedtaget', at grænsen 407 er en præstationsgrænse, hvorunder man ikke kan forvente, at elever vil være i stand til at klare en senere ungdomsuddannelse, formidles der direkte en usandhed (se fx Allerup et al 2014). Der er gode, tekniske grunde til at arbejde med mål for elevfærdigheder på 500-skalaen, fordi statistiske analyser af elevpræstationer, sammenligninger mellem forskellige elever og mellem samme elevs præstationer på flere tidspunkter *kun* kan ske meningsfuldt på denne skala. Man kan derfor udveksle data fra de internationale undersøgelser og 'regne videre' på dem i forhold til forskellige baggrundsvariabler, hvis elevpræstationerne er udregnet netop på denne 500-skala.

De nationale test og de nævnte internationale undersøgelser har alle Rasch modellen som statistisk baggrund for beregningerne af opgave-sværhedsgrad og elevdygtighed, og derfor er 500-skalaen lige så naturlig for de nationale test som for de internationale. Imidlertid vil man bemærke, at formidlingen af resultater fra de nationale test foregår på en helt anden skala: En såkaldt percentilskala. Denne skala forsøger at kombinere den velkendte 7-trin skalas markering af om præstationen er 'god' eller 'svag' med en gruppering af værdierne fra 500-skalaen. Den såkaldte elevprofil bestående af tre værdier på de fag-

lige sub-domæner er bygget på disse 'oversættelser' fra 500-Rasch skalaen til den anvendte percentil-skala. De er som grupperede værdier mere unøjagtige at 'regne videre' på, end hvis man havde adgang til de bagomliggende ugrupperede 500-skalaværdier (Allerup 2005).

Set fra en statistisk synsvinkel er det lidt pudsigt, at de officielle udmeldinger foregår ved hælp af de omtalte percentilværdier i tre profilben. Man 'tvinges' dermed til at foretage analyser mv i et sprog, som forhindrer præcise statistiske sammenligninger, sammenlignet med de analyser, som kunne være gennemført, hvis man i stedet for anvender værdier fra 500-skalaen. Fx kan en grundig evaluering af skolereformen ikke gennemføres ud fra de officielle elevprofiler, man bliver nødt til at inddrage tallene fra 500-skalaen.

Afslutning

Tidligere var prøver og test en række ens opgaver, som blev skrevet ind på papir og præsenteret for alle eleverne i en udvalgt lektion i løbet af en skoledag. Indførelsen af de nationale test har ændret meget på dette billede, hvor elever samles i et edb-rum med hver sin computer og løbende besvarer opgaver, som de præsenteres for. Opgaver som er forskellige fra elev til elev, og som opholder eleverne forskelligt, rent tidsmæssigt, mens computeren foretager statistiske vurderinger af, hvor præcist den kan beregne elevdygtigheden, før den lader skærmen skifte farve som tegn på, at nu kan eleven forlade computeren – og straks levere den profil, som bygges op af besvarelserne på hver sit profilområde. Det må betragtes som en i princippet erhvervet gevinst, at resultater fra de nationale test, til forskel fra mange gammeldags, lineære test, nu kan levere sammenligninger over tid, så man kan følge eleverne og lave sammenligninger mellem forskellige elevgrupper. Det ligger formodentligt endnu som en mulig gevinst at høste alle fordele af det adaptive testprincip, som på det teoretiske plan virker, men som i praksis udviser mangler. Endelig ser det ud til, at de rammer, som benyttes til formidling af testresultaterne, ikke er afstemt i forhold til den teoretiske ramme for Rasch modellen, der født i, hvilket øjensynligt frembyder problemer for modtagerne også, hvad enten de er elever, lærere eller forældre.

Litteratur

Allerup, P. 2013: Evaluering af de nationale tests - Ekspertvurdering 2. Undervisningsministeriet (Kvalitets- og Tilsynsstyrelsen). Bilag til Evaluering af de nationale test i folkeskolen.

Kousholt, K. (under udgivelse): Børn som deltagere i social test-praksis. Pædagogisk Psykologisk Tidsskrift, særnummer.

Kousholt, K. 2012: De nationale test og deres betydninger – Set fra børnenes perspektiver. Pædagogisk Psykologisk Tidsskrift, 49(4), 273-290.

Allerup, P., Klewe, L. og Torre, A., 2013: Unges valg og fravalg i ungdomsuddannelserne; kvantitativt perspektiveret. Aarhus Universitet, Institut fort Uddannelse og Pædagogik (DPU).

Allerup, P.: "Identification of group differences Using PISA scales". I PISA According to PISA – Does PISA Keep What It Promises? Stefan T. Hopmann/Gertrude Brinek Austria: University of Vienna (bog), Vienna, 2007.

Allerup, P.: Rasch Measurement, theory of. The International Encyclopedia of Education, second edition, Pergamon Press (1994).

Allerup, P.: "Statistik og Test – nogle forudsætninger og muligheder". Kroghs Forlag 2005, p. 140.

Rasch, G. 1960: Probabilistic Models for Some Intelligence and Attainment Tests. Munksgaard og 1980: Univ. of Chicago Press ISBN 0-941938-05- . LC# 80-16546.

JAM Press, 2004: Introduction to Rasch Measurement: Theory, Models, and Application, ISBN: 0-9755351-1-0. (www.jampress.org).