Appendiks 1: Om baggrund og teori bag valg af skala

De nationale test gav i 2010 for første gang danske lærere mulighed for at foretage en egentlig måling på en skala af deres elevers præstationer på grundlag af videnskabelige principper. Af de nationale test er testene i læsning og matematik uden tvivl de vigtigste. Dels fordi sproglig og matematisk forståelse er grundlaget for tilegnelsen af stoffet i alle de andre fag. Dels fordi matematiske færdigheder og læsningens resultat – forståelsen af det skrevne budskab – er blandt de sværeste færdigheder at følge udviklingen i: Der er brug for redskaber til at observere det uobserverbare – testen er for læreren hvad termometer og blodtryksmåler er for lægen.

De nationale test opgøres på Raschskalaer, som er den ultimative videnskabelige standard indenfor test. Rasch-skalaer måles i enheden Logits – Skalaen går principielt fra minus til plus uendelig (i praksis fra ca. -7 til +7) og nulpunktet fastlægges som regel som den midterste opgave i opgavebanken. På denne skala måles både elevers dygtighed og opgaves sværhed. Denne skalatype har en række belejlige egenskaber for dem, der anvender den til videnskabeligt og professionelt brug – først og fremmest at det er en skala hvor længden af forskellige intervaller på skalaen er direkte sammenlignelige; Raschskalaen er en såkaldt intervalskala[1]. Det vil sige at en forbedring i læsefærdighed fra -2 til 0 logitter er lige så stor som en forbedring fra 1 til 3 og dobbelt så stor som forbedringen fra 1½ til 2½. Betydningen af forskelle målt på skalaen varierer ikke afhængigt af hvor på skalaen man befinder sig. Man taler om at målinger på denne skala er invariante.

Men den er samtidigt meget abstrakt og både lærere, elever og forældre har behov for en reference, som de kan forholde sig til. For at gøre testresultaterne forståelige blev det valgt at rapportere resultater fra de nationale test af systemet på såkaldte percentil-skalaer, som går fra 1 til 100 point og opdeler landets elever i 100 grupper (procentdele) efter stigende præstation – de midterste elever får 50 point og den dygtigste procentdel får 100 point osv.

Skalaer til brug for formidling af resultater

Til brug for dialog med hjemmet og elever blev percentilskalaerne yderligere forenklet til to skalaer: Dels norm- og dels kriterieskalaen.

Normskala: En fem-trinskalaer, der minder om karakterer med fordeling efter ECTS/7-trinsskalaen (1-10=Klart under middel, 11-35=Under middel, 36-65=Middel, 66-90=Over middel og 91-100=Klart over middel).

Kriterieskala: Er en seks-trinskala med bedømmelse ift. en faglige forventning til elevens niveau. Ligesom normskalaen er den dannet som en gruppering af skalascorer, men grænseværdierne er valgt ud fra ministeriets fagfolks vurdering af, hvilke sværheder af opgaver, det er nødvendigt at kunne besvare korrekt, for at leve op til at være henholdsvis: Ikke-tilstrækkelig, Mangelfuld, Jævn, God, Rigtig god, Fremragende.

Mens normskalaen arbejder med samme intervaller for alle test- og profilområder, er kriterieskalaen kun defineret for dansk læsning og matematik. For afkodning i dansk læsning er der den særlige ting at sige om kriterieskalaen, at ”God” er det øverste niveau, hvilket vil sige at skalaen kun har fire trin. Baggrunden er, at mht. afkodning opnås en automatiseret (dvs. perfekt) afkodning forholdsvis hurtigt, så kategorierne ”Rigtig god” og ”Fremragende” efter fagfolkenes/Ministeriets vurdering ikke giver mening: ”God” er det bedst opnåelige – men det kunne lige såvel være blevet betegnet som ”Fremragende”, fordi det bliver ikke bedre.

[1] Intervalskala: En skala hvor alle afstande er lige store og der derfor kan regnes gennemsnit, summer og differencer. Temperatur-, længde- og vægtskalaer er gode eksempler på denne type. En oversigt over forskellige skalatyper kan findes her: S. S. Stevens (1946), On the Theory of Scales of Measurement, Science, Vol. 103.

Typer af skalaer

Percentil-, norm- og kriterieskalaen er alle ***ordinalskalaer***[2], som ved rangordning af elever er meget velegnede til at give læreren et forståeligt mål for elevernes faglige niveau i forhold til ***det gennemsnitlige niveau*** (normen) eller ift. hvad fagfolk betragter som god læsning i Danmark. De to grupperede skalaer (norm og kriterie) inddeler eleverne i meget store grupper. Skal man have en mere præcis angivelse af elevens score, vil man være henvist til at anvende percentilskalaen. Men denne skalatype har 2 alvorlige mangler i forhold til brug i det pædagogiske arbejde:

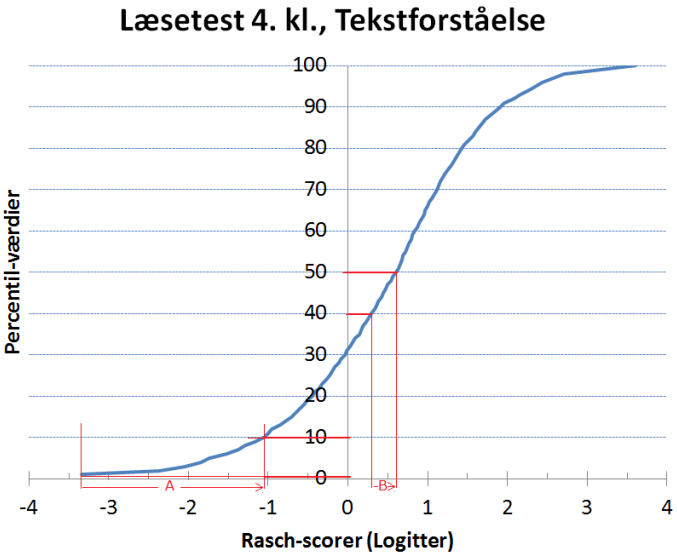
1) Dels kan en progression på fx 10 percentiler repræsentere en meget forskellig fremgang forskellige steder på dygtigheds-skalaen. *Eksempel: En elev scorer "1" i 2010 og "10" i 2012. En anden elev scorer "40" i 2010 og "50" i 2012. Spørgsmålet man som lærer kan stille sig er "Betyder en progression på ca. 10 point nogenlunde det samme faglige fremskridt?*"
2) Dels kan resultaterne fra de fire læsetest (til hhv. 2., 4., 6. og 8. klasse), der er opgjort på hver deres skalaer, ikke sammenlignes indbyrdes. *Eksempel: En elev i 4. klasse scorer "60" i 2011 og 2 år senere i 2013 scorer eleven "45" 6. klasse testen. Spørgsmålet man som lærer kan stille sig er, om "eleven læsefagligt er gået frem, tilbage eller er han stået stille?"*

For at besvare disse to spørgsmål skal vi dels kunne placere resultaterne fra de fire forskellige test på den ***samme sammenlignelige skala***, og dels skal denne skala skal være en ***intervalskala***, hvilket den bagved liggende Rasch-skala er, som det tidligere er nævnt.

Ad 1: Placering på skalaen.
Det vil være normal logisk tænkning at en fremgang på 10 point er nogenlunde det samme uanset hvor man befinder sig på skalaen. Men sådan er det ikke nødvendigvis.

Af figur 1 fremgår forholdet mellem Percentil- og Rasch-skalaen (hvor man måler dygtighed) for tekstforståelse i 4. klasse i de nationale test. En fremgang fra 1 til 10-percentilen svarer til næsten 2½ enheder (logits) på Raschskalaen (markeret med "A" i figuren), mens en fremgang fra 40 til 50-percentilen svarer til knap 1/3 logit (markeret med "B"). En fremgang fra 1 til 10 point repræsenterer en ***forbedring af læsefærdighederne, der er ca. 7 gange så stor***, som fremskidt fra 40 til 50-percentilen.

Figur 1



Uanset at de fleste lærere har en god fornemmelse af deres elever, så er det altså helt afgørende for vurderingen af progression, at man måler på en egnet skala, hvor man ved hvad en enhed betyder. Især hvis man vil se på en elevgruppes (fx en klasses) samlede fremskridt, vil valg af den rette skala betyde forskellen mellem en solid måling og et ufortolkeligt, måske stærkt misvisende tal: Alene på grund af valg af skala, kan der vise sig målefejl på gennemsnitsresultater svarende til næsten to års normale progression for en klasse.

[2] Ordinalskala: En skala, hvor observationer (fx elevernes præstationer) kan rangordnes, men hvor værdierne på skalaen kun er knyttet til en rangorden og ikke kan bruges til at regne videre på en teoretisk holdbar måde, selvom man ofte gør det – fx ved beregning af karaktergennemsnit.

Ad 2: Udviklingsarbejdet og valg af fælles skalaer

For hvert profilområde er der fire percentilskalaer med bagved liggende Raschskalaer, der måler samme egenskab (fx Sprogforståelse 2., 4., 6. og 8. klasse).

Men hverken enheden (hvilken læsefremgang betyder en tilvækst på en logit) eller 0-punktet (den gennemsnitlige opgaves sværhed i hver af testene) kan på forhånd antages at være den/det samme for de fire Raschskalaer.

Disse relationer mellem testene blev i projektet undersøgt ved at en række skoler gennemførte test i nov.-dec. 2012 for elever i 3., 5. og 7. klasse, der kunne tage en frivillig national test på både klassetrinnet over og under (fx tog elever i 5. klasse læsetesten udarbejdet såvel til 4. som til 6. klasse).

Når den samme elev scorede forskelligt i de to test, kunne forskellen kun (se note 5) skyldes to forhold:

- forskelle i testenes sværhed og
- hvad man kan kalde "målefejl" (som følge af fx elevernes gode/dårlige dage, koncentration, kendskab til indhold af konkrete opgaver, teknik-bøvl, instruktion og andre omstændigheder ved testafviklingen).

På grundlag af ca. 1.800 besvarelser fra 900 elever (hver besvarede to test indenfor samme uge) kunne de parvise relationer mellem testene til 2. & 4. klasse, 4. & 6. klasse og 6. & 8. klasse undersøges på grundlag af elevernes scorer og skalaerne ækvivaleres profilområde for profilområde.

På det grundlag kan man placere præstationerne på fælles skalaer med de rigtige egenskaber, givet visse forudsætninger er opfyldt:

- At opgaverne i hver af de 12 opgavebanker hver især følger en raschmodel. Dette krav er afprøvet som et led i etableringen af opgavebankerne, hvor alle skalaerne ved idriftsættelsen levede op til Raschmodellen krav.
- At hver af fire test måler præcis de samme underliggende egenskaber for eleverne - samme slags sprogforståelse, afkodning og tekstforståelse[3].

Om valg af skalatype: Man kunne have valgt at arbejde videre med en af Raschskalaerne til testene (en logit-skala – gående fra minus uendelig til plus uendelig med den midterste opgave i opgavebanken som nulpunkt), eller en hvilken som helst anden intervalskala (Fx en PIRLS- eller en PISA-skala). Men samtidigt er det vigtigt at valget af skala skal forbindes med en pædagogisk problemstilling og ikke rangordning, kontrol, politik og administration. På grundlag af amerikanske forskningsbaserede erfaringer og tidligere arbejde med danske elever/lærere/skoler blev valgt en skala, der har været gode erfaringer med (Lexile-skalaen[4]) som model for de fælles skalaer.

[3] Ved udviklingen af testene var det aftalt, at testene/testopgaverne skulle fremstilles/afprøves på den måde, men det er ikke så vidt vides undersøgt til bunds om dette er sket og/eller holder efter testene er sat i drift - og i givet fald for alle klassetrin og for alle profilområder. Dette vil ikke kunne undersøges tilbundsgående inden for dette projekts rammer. Det ville kræve at man analyserede de enkelte opgaver og opgavers besvarelser (altså Rasch- og/eller Item Respons analyser), hvilket er uden for dette projekts scope.

[4] Lexile-skalaen er udviklet i USA og bruges til både at beregne en teksts sværhed og elevers læsefærdigheder (Se http://Lexile.com). Når man kender elevens dygtighed og tekstens sværhed og har dem opgjort på samme skala, kan man let udvælge tekster, der lige præcis passer til eleven. I nærværende sammenhæng anvendes skalaen alene til beregning af elevernes dygtighed.

Konklusion

Den oprindelige begrundelse for at afrapportere på enten percentilskalaer (1-100-skalaer) eller de karakterlignende norm- og kriterieskalaer (*ordinalskalaer*) holder stadigvæk. Denne type af skaler er egnet til formidling, fordi den afspejler måden, man sædvanligvis tænker uddannelsesresultater på *udenfor* uddannelsessystemet. Men hvis skolen skal kunne arbejde professionelt med testene som pædagogisk redskab til monitorering af udvikling, så kræver det, at denne tilgang suppleres med udviklings-/progressionsskalaer. Elevernes faglige udvikling skal kunne følges over tid således, at man skal kunne beregne meningsfuld progression og gruppegennemsnit (klasse og skole), så man kan vurdere klassenes udvikling.

For hvert profilområde er der i testsystemet 4 skalaer: Percentilskalaer, normskalaer, kriterieskala og Raschskala. Disse skalaer er specifikke eksempelvis for testning i læsning i 2-, 4-, 6- og 8-klassetesten. Beregnerens fælles progressionsskala er fælles for alle fire test, så et resultat i 2. klassestesten er direkte sammenligneligt med et resultat i 6. klassestesten. Deres indbyrdes relation er vist for S-skalaen (sprogforståelse) på linealen til højre.