Utdanningsdirektoratet

Metodegrunnlag for nasjonale prøver

Prøvene blir nå behandlet med moderne IRT og resultatene er hvert år satt på felles skala som tillater sammenlikning mellom år


Fotograf Jannecke Sanne Normann

Innholdsfortegnelse

Introduksjon

I denne artikkelen beskrives hovedpunktene i det psykometriske (testteoretiske) grunnlaget for nasjonale prøver fra og med 2014. Før 2014 ble prøveresultatene behandlet med klassisk testteori der antall riktige oppgaver for den enkelte elev ble oversatt til mestringsnivåer før resultatene ble aggregert og rapportert.

Ved bruk av klassisk testteori får elever som besvarer like mange oppgaver med ulik vanskegrad samme resultat. Altså vil ikke poengsummen fra prøven fullt ut representere elevens egentlige ferdighetsnivå. Ved å anvende en metode som tar hensyn til at oppgaver har ulik vanskegrad og ulik diskriminering mellom flinke og svake elever, vil en oppnå større presisjon. En slik metode kan dermed brukes til bedre og sikrere rapportering av resultatene.

Fra og med 2014 er resultater fra alle nasjonale prøver basert på bruk av IRT («Item Response Theory») kalibrerings- og skaleringsmetoder der vi anvender en 2-parameter IRT-modell. Med den nye modellen er det også mulig å integrere en ankerprøve som sikrer at samme tall til enhver tid beskriver samme ferdighet. Dette gir oss et måleverktøy som gjør at vi kan si noe om endringer fra et år til det neste.

Nasjonale prøver utvikles av forskjellige fagmiljøer:

- Leseprøve på 5. og 8-9. trinn: Universitetet i Oslo, Institutt for lærerutdanning og skoleforskning
- Engelskprøver på 5. og 8. trinn: Universitetet i Bergen
- Regneprøver på 5. og 8. trinn: Norges teknisk-naturvitenskapelige universitet, Matematikksenteret

Alle fagmiljøene arbeider på grunnlag av et rammeverk for nasjonale prøver, hvor både prøvenes innhold og metodologi er fastsatt. Rammeverket er under revisjon og blir publisert på nytt våren 2016.
Alle prøveoppgaver blir pilotert minst to ganger, i representative elevgrupper, slik at de endelige prøvene bare inneholder oppgaver som vi vet på forhånd at fungerer godt.

Eksempeloppgaver og andre generelle opplysninger om prøvene og oppgavene finnes på http://www.udir.no/Vurdering/Nasjonale-prover. Rapporten er skrevet av Julius K. Bjørnsson med input og støtte fra prøveteamet i avdeling for vurdering 2.

Bakgrunn for endringene i 2014

Når en prestasjon eller ferdighet skal følges og sammenliknes over tid, er det en grunnleggende regel at målingen (prøven) ikke må endres. Dersom en skifter ut gjenstanden for målingen (måleinstrumentet), vet en ikke lenger om endringene i resultatene fra en prøve til en annen skjer fordi måleinstrument endres, fordi ferdighetene i elevgruppen forandrer seg, eller begge deler samtidig. Oppgavene fra nasjonale prøver publiseres alltid etter gjennomføringen, og derfor må det lages nye prøver hvert år.

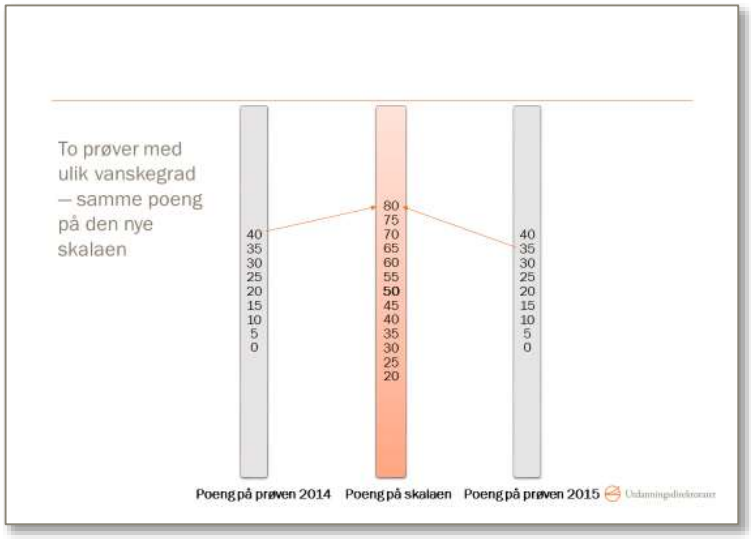
Med bruk av klassisk testteori og vanlige poengsummer fra prøvene, var det derfor umulig å vite om endringer fra år til år skyldtes nye prøver eller endret ferdighet hos elevene eller begge deler. Derfor ble det tekniske grunnlaget for de nasjonale prøvene skiftet ut, slik at i dag brukes IRT-skalering av prøvene og IRT- lenking og ekvivalering fra et år til det neste.

For å måle endring over tid, ble 2014 definert som «år 1» i regning og engelsk på 5. og 8. trinn. Fra og med høsten 2016 blir prøvene i lesing på 5. og 8. trinn også elektroniske, og 2016 er derfor startåret for måling av endring over tid i lesing. Dette betyr at resultater for hele populasjonen kan med hjelp av den nye metoden, sammenliknes over tid.

Elevers poengsummer kan aldri bli helt like på to forskjellige prøver, selv om prøvene er laget på akkurat samme måte, av samme prøvekonstruktører og laget for å måle akkurat samme ferdighet. Det vil alltid være forskjeller i vanskegrad både på de forskjellige oppgavene og på prøvene i sin helhet. For å måle

endring over tid, var det derfor nødvendig å ta i bruk både nye skaleringsmetoder for prøveresultatene og metoder for lenking og ekvivalering av forskjellige prøver. En slik måling av endring over tid krever at **samme tall beskriver samme ferdighet**, selv om målingen er foretatt med forskjellige prøver. En oversikt av metoder brukt til dette formålet kan for eksempel finnes i Tan & Michel, (2011).

Figur 1. er et skjematisk eksempel på denne situasjonen hvor den samme poengsummen på to prøver ikke gjenspeiler samme dyktighetsnivå. Figuren viser også hvordan disse ulike poengsummene kan bindes sammen ved hjelp av en ny felles skala.



Figur 1. Skjematisk bilde av to prøver med ulik vanskegrad satt på samme skala

Det er av overordnet betydning når prøver er viktige og en vil evaluere effekten av f.eks. endringer i utdanningssystemet over tid, at en ekvivalering av forskjellige prøver skjer på en metodologisk forsvarlig og solid måte. Alle retningslinjer for god faglig praksis og rettferdig og riktig bruk av prøver og psykometriske metoder understreker dette (se for eksempel Standards for Educational and Psychological Testing, 1999).

IRT-metoden

IRT (Item Response Theory) er blitt brukt i prøveutvikling i over femti år og har fått status som en standard for behandling av prøver. De aller fleste storskala prøvesystemer, nasjonale prøver og internasjonale komparative undersøkelser bruker i dag IRT-analyse. Som nevnt over, er IRT- metodene en samling verktøy som gir bedre presisjon i målingene og som gir nye muligheter for prøveutviklerne. Vanskegraden til en oppgave i klassisk testteori er oftest uttrykt som en prosentandel personer som får til en oppgave (p-verdi mellom 0 og 1). Dyktigheten til elevene representeres så av total poengsum eller prosentandel riktig av full skåre. Det underliggende problemet med denne metoden er at prøvens vanskegrad er dermed avhengig av hvilke elever som deltar og elevenes dyktigheter er avhengig av hvilke oppgaver som var med i prøven.

Usikkerheten i denne metoden kan kanskje best uttrykkes med et enkelt eksempel: På en dag i 1953 stod to mennesker på toppen av Mount Everest. Og dette var en så vanskelig oppgave at ingen hadde klart den før. På en annen dag i 1996 stod 20 mennesker samtidig på toppen av fjellet. Og ifølge teorien om at vanskegrad uttrykkes med antall personer som klarer oppgaven, ville man kunne konkludere at fjellet var blitt 10 ganger lettere å bestige, siden det var ti ganger flere personer der på den andre dagen. Men dette stemmer naturligvis ikke. IRT metodenes evaluering av vanskegrad fungerer ikke slik, men er en statistisk estimering av fjellets høyde, dvs. en måte å måle fjellets høyde, istedenfor å estimere høyden ut ifra hvor mange klarer å bestige det.

Ved hjelp av IRT kan man estimere oppgavenes vanskegrader uavhengig av elevene som deltok, og man vil kunne beregne elevers dyktigheter uavhengig av nøyaktig hvilke oppgaver de besvarte[1]. Grunnlaget for dette er en såkalt probabilistisk modell hvor sannsynligheten for å få riktig svar på en oppgave uttrykkes som en funksjon av elevenes dyktigheter. Resultatene fra denne metoden gir i prinsippet verdier langs en skala fra pluss til minus uendelig, men for praktiske formål avgrenses gjerne skalaen til å strekke seg fra -3 til +3. Man har valgt å bruke disse tallene fordi de er de samme som beskriver en vanlig normalfordeling (z-skåre) hvor gjennomsnittet er vanligvis satt på 0 og standardavviket på 1. På det grunnlaget blir tallene lett forståelige, men de kunne altså være nesten hvilke tall som helst.

Gode oversikter av disse metodene finnes i Brennan (2006) og Kolen & Brennan(2004). I tillegg finnes det mange beskrivelser tilgjengelige på internett, f.eks.: https://en.wikipedia.org/ wiki/ Item_response_theory. For nærmere opplysninger om det matematiske og statistiske grunnlaget henvises til den grunnleggende boken av Frederick M. Lord (1980).

IRT beskriver den ferdigheten som en oppgave måler med en enkel logistisk funksjon som uttrykkes slik:

$$P(\theta) = c + \frac{1 - c}{1 + e^{-Da(\theta - b)}}$$

Dette er en såkalt tre-parameter modell (3PL) som beskriver sannsynligheten for et riktig svar. Modellen beskriver oppgavens diskriminering[2] (a), oppgavens vanskegrad (b) og hvorvidt gjetting/tilfeldighet har en innflytelse på sannsynligheten for riktig svar (c). I praksis betyr c sannsynligheten for riktig svar, gitt ingen ferdighet. Dette er også ofte kalt «pseudo» gjetting. I analysen av nasjonale prøver brukes ikke denne modellen i de endelige analysene, men en to-parameter modell (2PL) hvor gjettingen ikke er inkludert.

Den modellen vi bruker er:

$$P(\theta) = \frac{1}{1 + e^{-1{,}7a(\theta - b)}}$$

Eller skrevet på en annen måte:

$$P(\theta) = \frac{e^{(\theta - b)}}{1 + e^{Da(\theta - b)}}$$

Her er P(θ) =sannsynligheten for riktig svar, b=oppgavens vanskegrad og a=oppgavens diskriminering, dvs. i hvilken grad oppgaven diskriminerer mellom svake og flinke elever. Θ (theta) representerer selve ferdigheten eller kompetansen oppgaven måler og e er den naturlige logaritmen. Konstanten D eller 1,7 i formelen ovenfor er brukt til å approksimere en kumulativ normal fordeling av ferdigheten («Normal Ogive»).

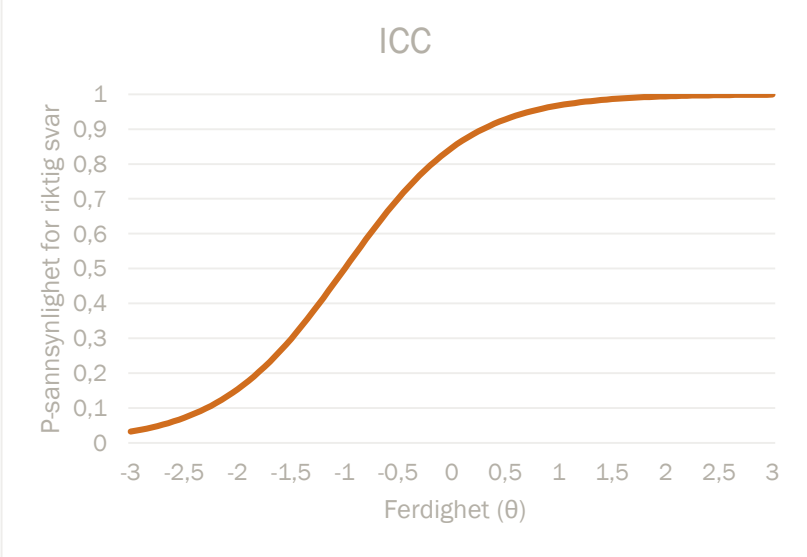
Denne D konstanten har en historisk forklaring og er brukt til å få samsvar med de første IRT modellene som var basert på en kumulativ normalfordeling. Disse gamle modellene var vanskelige å regne ut og på datidens datamaskiner tok utregningene veldig lang tid. Derfor valgte man å gå over til en logistisk modell som var mye lettere å regne ut. Men bruk av konstanten har den konsekvensen at

[1] Skulle vi være strenge her, så er det slik at forholdet mellom sannsynligheten for riktig svar på oppgaver i prøven er uavhengig av hvilke personer som deltok, og vice versa.
[2] Evne til å skille mellom svake og flinke elever.

diskrimineringsparameteren blir lavere, men også på en annen skala og når resultatene tolkes må en vite om den er brukt i analysen eller ikke.

En enklere variant av denne modellen ville være en én-parameter modell (1PL) som kun beskriver vanskegraden til oppgaven. Dette er det samme som den kjente Rasch-modellen (Rasch, 1960). I denne modellen er diskrimineringen til alle oppgaver satt til 1. Ved analyser av nasjonale prøver før 2014, viste det seg imidlertid at oppgavene på prøvene hadde sterkt varierende diskriminering, både analysert med IRT-metoden og også med klassiske metoder (f.eks. med en biserial korrelasjon), og rammeverket for nasjonale prøver setter minimumskrav for diskriminering. Derfor ble en to-parameter modell valgt for nasjonale prøver, en modell som leverer både vanskegrad og diskriminering til hver oppgave.

Figur 2 viser et eksempel på en ICC («Item Characteristic Curve») kurve for en oppgave.



Figur 2. ICC kurve for en enkeltoppgave:

Kurven på figur 2 viser sannsynligheten for riktig svar på denne oppgaven for elever med forskjellige ferdighetsnivåer. Den loddrette aksen viser sannsynligheten for riktig svar og den vannrette aksen viser ferdigheten. Det stedet på ferdighetsskalaen hvor denne sannsynligheten er 50 % (0,5) defineres som oppgavens vanskegrad. Denne kan leses av den vannrette aksen for denne 2PL modellen. Bildet viser en relativt lett oppgave som har en b-verdi (vanskegrad) på -1. Diskrimineringen til oppgaven styrer hvor bratt kurven blir. Kurven for en oppgave som diskriminerer sterkt mellom svake og flinke, vil derfor være brattere i området rundt oppgavens vanskegrad enn en oppgave med svak diskriminering.

Usikkerhet

Klassisk testteori beskriver oppgavers egenskaper utelukkende ut ifra hvordan oppgavene oppfører seg samlet i en prøve og en av de viktigste egenskapene til en prøve ut ifra en slik tilnærming er akseptabel reliabilitet. Den dreier seg om hvorvidt alle oppgavene i prøven måler det samme på en konsistent måte, dvs. hvor pålitelig prøven er. Imidlertid leverer reliabilitetstall for en hel prøve bare ett enkelt estimat på usikkerhet i målingen. Dette er i klassisk testteori uttrykt med den såkalte standard målefeil (SEM «Standard Error of Measurement»). SEM er den samme for alle nivåer av ferdigheten og er kun avhengig av reliabiliteten og standardavviket til prøven i sin helhet:

$$\text{SEM}=SD\sqrt{1-r}$$

Et 95% konfidensintervall for enhver skåre på prøven er da gitt av 1,96*SEM.

I IRT er derimot målefeilen avhengig av elevens skåre, dvs. målefeilen er ulik for ulike områder av skalaen (CSEM «Conditional Standard Error of Measurement»). Den er vanlig lavest rundt gjennomsnittet av målingen og høyest øverst og nederst på skalaen. Dette er en mye bedre representasjon av usikkerheten i målingen bl. a. fordi en vanlig prøve inneholder jo flest oppgaver på midten av ferdigheten og gir derfor mest opplysninger og minst målefeil der.

Konsekvensen av dette er at usikkerheten i målingen blir oftest størst nederst og øverst på skalaen blant de svakeste og de dyktigste elevene. Dette burde en være oppmerksom på når resultatene for disse elevene tolkes. Hver enkelt oppgave har dermed også en målefeil som er gitt av IRT-funksjonen presentert foran.

Informasjon og målefeil

I IRT-modellering har man også et begrep som representerer det motsatte av målefeil, den såkalte informasjonsverdien. Informasjonsverdien er lav når målefeilen er stor og vice versa. Den enkelte oppgaves informasjonsverdi er et produkt av sannsynligheten for riktig svar, sannsynligheten for galt svar og oppgavens diskriminering i annen potens, slik:

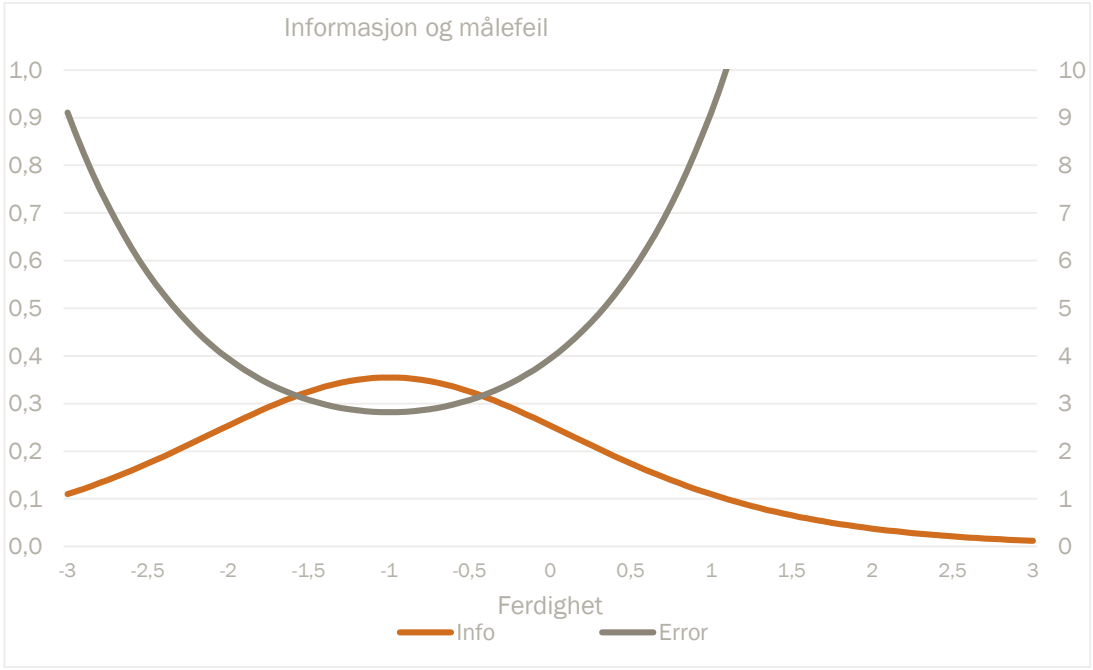
$$\boldsymbol{I = a^2 P(\theta)\, Q(\theta)}$$

Den avhengige[3] målefeilen er direkte relatert til oppgavens informasjonsverdi slik:

$$\text{CSEM} = \frac{\mathbf{1}}{\sqrt{\boldsymbol{I}}}$$

Figur 3. viser et eksempel på både informasjonsverdi og målefeil for den samme oppgaven som ble vist i figur 2. Figuren viser at mest informasjon er levert rundt oppgavens vanskegrad, og at det også er minst målefeil der.

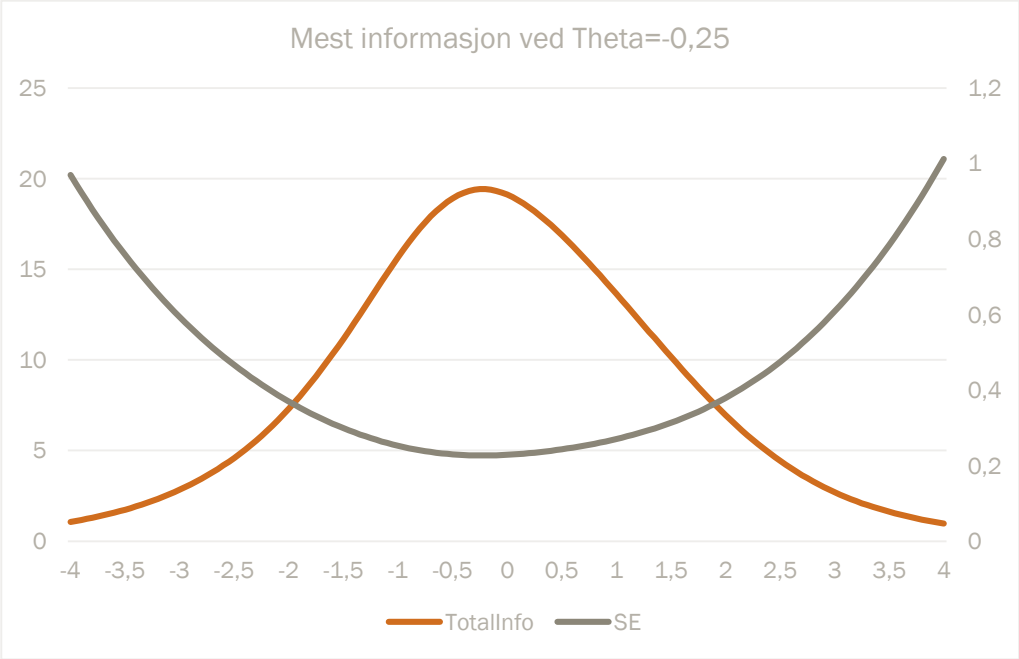
[3] Her betyr avhengig målefeil at den er avhengig av plassering på ferdighetsskalaen, vanlig størst øverst og nederst



Figur 3. Informasjon og målefeil.

En slik analyse blir derfor foretatt for hver eneste oppgave i prøven i tillegg til å bestemme oppgavens sannsynlighetskurve (ICC kurve) og hvor bra denne logistiske modellen passer til de faktiske dataene («Model Fit»).

En av IRT-metodens beste egenskaper er at disse opplysningene fra analyse på hver enkeltoppgave, kan summeres for hele prøven. Figur 4. er et eksempel fra nasjonale prøver 2014 der informasjonsverdi og målefeil vises for en hel prøve.



Figur 4. Informasjon og målefeil for en hel prøve.

Dette er et kraftig verktøy i prøvekonstruksjonen og ved evalueringen av hvordan en prøve fungerer, fordi det gjør det mulig å sette sammen en prøve slik at den måler med høyest presisjon akkurat der en ønsker på ferdighetsskalaen. Hvis prøven skal måle hele ferdigheten, som i de nasjonale prøvene, vil kurvene

se ut som vist i figur 4. Men dette kan se annerledes ut i prøver som har et annet formål enn å måle hele ferdigheten. I de nasjonale kartleggingsprøvene ønsker man for eksempel å identifisere de elevene som lærerne bør bekymre seg ekstra for. For disse prøvene bør toppen på informasjonskurven være høyest rundt prøvens bekymringsgrense, altså ganske langt nede på skalaen til å sikre at sikkerheten i målingen blir best rundt denne grensen.

IRT gir altså mer presise opplysninger på alle ferdighetsnivåer enn klassiske metoder. Disse metodene gir oss derfor et bedre verktøy enn før ble brukt, for å utvikle prøver med de egenskapene vi ønsker at prøvene skal ha.

Dimensjonalitet og lokal uavhengighet

Nasjonale prøver måler stort sett et endimensjonalt konstrukt. Dette vises blant annet gjennom den høye alfa-koeffisienten alle prøvene har. I tillegg viser en faktoranalyse av prøvene i lesing, regning og engelsk, at denne endimensjonaliteten er lik for alle prøvene. Det er viktig å bevare denne egenskapen. Endimensjonalitet gjør IRT-analysen mulig og bedre. Denne egenskapen ved prøvene gjør dessuten tolkning og bruk av prøveresultatene bedre og lettere og mer målrettet og praktisk nyttig for lærere og elever.

Det er overordnet viktig ved bruk av IRT analyse i prøveutviklingen å ivareta analysens grunnkrav om endimensjonalitet og lokal uavhengighet. Lokal uavhengighet betyr ikke at det ikke må være sammenhenger eller korrelasjoner mellom oppgaver i en prøve - slike sammenhenger er naturlige og nettopp det vi vil måle. Lokal uavhengighet betyr derimot at oppgaver som måler på samme nivå, som har samme vanskelighetsgrad, skal være uavhengige av hverandre slik at sannsynligheten for å svare riktig på den ene oppgaven ikke har en effekt på sannsynligheten for å svare riktig på den andre også. Sagt med andre ord, to oppgaver skal ikke være korrelert når ferdigheten er fiksert/kontrollert. Eller som Lord(1980) beskriver det: «*Local independence requires that any two items be uncorrelated when θ is fixed. It definitely does not require that items be uncorrelated in ordinary groups, where θ varies. Note in particular that local independence follows automatically from unidimensionality. It is not an additional assumption».*

I grupper hvor ferdigheten varierer, kan det naturligvis forekomme korrelasjoner mellom oppgaver. Dette betyr at endimensjonalitet og lokal uavhengighet er to sider av samme sak. Hvis man ikke kan påvise endimensjonalitet, er det også stor sannsynlighet for lokal avhengighet. I praksis betyr dette at hvis f.eks. et riktig svar på en oppgave er avhengig av at man har svart riktig på en foregående oppgave, så er ikke lokal uavhengighet til stede, og da vil IRT modellen ikke beskrive ferdigheten på en passende måte.

DIF («Differential Item Functioning»)

En av fordelene ved å bruke IRT analyse for behandling av prøveoppgaver, er at den gjør mulig å utforske om oppgavene og dermed prøven i sin helhet favoriserer en bestemt gruppe over en annen, f.eks. gutter over jenter, eller innfødte over innvandrere, for å nevne et par eksempler. Dette er en situasjon hvor oppgaven er f.eks. vanskeligere for en person av en viss type enn en annen som ikke er av den typen, gitt at begge har samme underliggende ferdighet. Det er derfor veldig viktig å skille mellom reelle forskjeller i ferdighet og skjevheter av denne typen. Reelle forskjeller vil vi gjerne at prøven måler, men samtidig vil vi utelukke måleskjevheter som kan gi feil resultater og føre til feilaktige konklusjoner.

Det er klart at alle oppgaver i en prøve har en viss relasjon (f.eks. korrelasjon) til den underliggende ferdigheten. Hvis denne relasjonen er den samme i to grupper (f.eks. jenter og gutter) så favoriserer oppgaven ikke det ene kjønnet over det andre og alle forskjeller mellom gruppene som oppgaven viser, er da reelle forskjeller. På den andre siden, hvis relasjonen ikke er den samme i begge grupper, så foreligger det en DIF som gjør det vanskelig å si noe om reelle forskjeller mellom gruppene. Når IRT er brukt til å utforske om DIF er til stede, blir oppgaveparametrene estimert separat for begge målgrupper (f.eks. gutter og jenter) og hvis de ikke varierer mellom gruppene så konkluderer man at oppgaven ikke måler forskjellig i de to gruppene. Bak dette ligger ganske komplisert matematikk som vi ikke skal gå inn

i her, men bare konstatere at utforsking av DIF i prøveoppgaver er en kjerneoppgave for prøveutviklingen, siden gruppeforskjeller er ofte det som er de viktigste resultatene fra en prøve. I nasjonale prøver blir derfor alle oppgaver testet for kjønns DIF og oppgaveutviklerne prøver å utelukke slike oppgaver eller hvis ikke det er mulig, å balansere favoriseringen av det ene kjønnet over det andre i prøven med oppgaver som går i motsatte retninger, helst på samme ferdighetsnivå, slik at prøven som en helhet ikke favoriserer det ene kjønnet over det andre.

Endring over tid

Som beskrevet tidligere er to prøver som er satt sammen for å skulle måle det samme og som inneholder samme typer oppgaver, likevel aldri helt like i vanskegrad. Men i mange situasjoner har vi behov for å kunne sammenlikne resultater fra to eller flere slike «like» prøver. Hvis vi vil se endringer i ferdigheten over tid, eller hvis vi vil sammenlikne to varianter av samme prøve, er det derfor nødvendig å foreta en prosedyre som kalles lenking. Dersom man lykkes med å foreta en slik lenking mellom to prøver, kan skårene på en prøve holdes sammen med (eller ekvivaleres med) skårene fra en annen prøve.

I IRT-tilnærmingen blir den ene prøven ekvivalert med den andre ved at en del av oppgavene i de to prøvene er identiske. Denne delen av prøven kalles ofte for ankerprøve og blir i nasjonale prøver gjennomført som en integrert del av de ordinære prøvene over flere år. Nasjonale prøver inneholder derfor en prøveversjon som inneholder disse ankeroppgavene og denne versjonen besvares av et tilfeldig utvalg på rundt 6 % av hele populasjonen elever. Ankeroppgavene fra 2014 ble gjentatt i 2015, og da brukt til å skalere 2015-prøvene slik at de havner på samme målestokk som prøvene i 2014. Et skjematisk bilde av modellen vises i figur 5.

Figur 5. Skjematisk oppsett av ankerprøve og kohortprøve.

Dette ankerdesignet er ofte kalt NEAT som står for «Non-Equivalent Groups with Anchor Test». Forskjellige ankerdesign og metoder for lenking kan en lese om f.eks. i Kolen & Brennan (2014). Ankerprøven i nasjonale prøver er altså en integrert del av prøven (om lag 20 oppgaver for regning og engelsk) som et representativt tilfeldig utvalg av elevene besvarer de ulike årene. Ankeroppgavene ble kalibrert med prøvene i 2014 og så inkludert i prøvene i 2015, med oppgaveparametrene fra 2014. Det fører til at prøvene fra 2014 og 2015 blir på samme skala og det blir så sikkert som mulig at samme tall beskriver samme ferdighet begge årene.

Forskjellige testsett

For regning og engelsk er alle prøveoppgavene delt på fire såkalte testsett som inneholder alle oppgavene. Tre av disse testsettene har samme oppgaver, men i ulik rekkefølge. Dette er gjort for å motvirke den tendensen hos mange elever å ikke svare på oppgaver på slutten av prøven. Det fjerde testsettet blir besvart av rundt 6 % av elevene (ofte rundt 3500), som er tilfeldig utvalgt, og inneholder f.eks. i regneprøven 30 oppgaver fra den vanlige prøven og 20 ankeroppgaver. Alle testsettene blir så kalibrert sammen, 0g er dermed satt på samme skala.

I elektronisk lesing som lanseres høsten 2016, blir dette designet litt annerledes, siden leseprøvene består av tekster som hver har mange tilhørende oppgaver. Der lager man en hel ekstra prøve og et tilsvarende stort utvalg elever f.eks. på 8. trinn, får en prøve hvor der er 7 tekster, 5 vanlige tekster og 2 ankertekster. I lesing blir det derfor fem forskjellige testsett som vist på figur 6. Tilsvarende design er brukt for 5. trinn, men med kun 5 tekster.

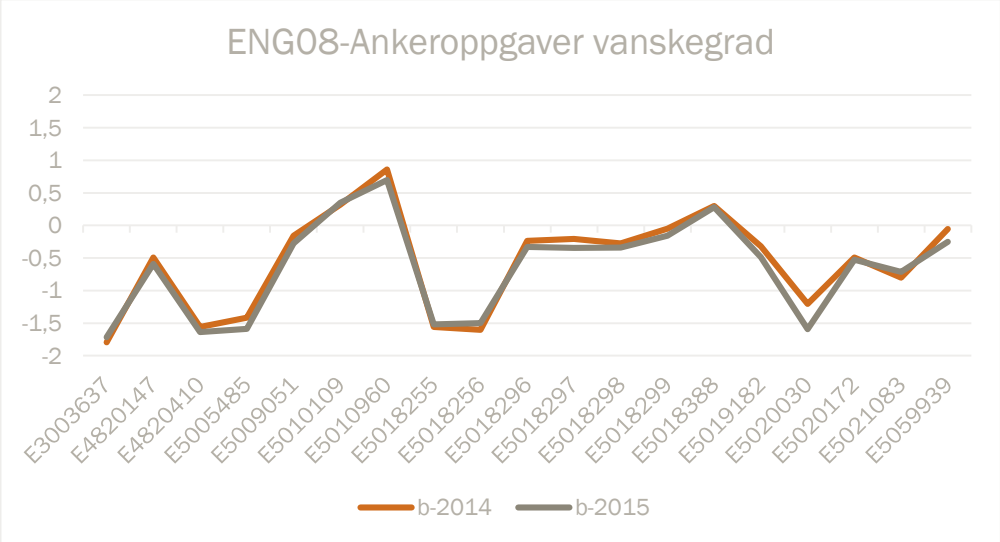
Prøve	Antall elever	Tekst 1	Tekst 2	Tekst 3	Tekst 4	Tekst 5	Tekst 6	Tekst7
Kohort	56 000	Kohort 1	Kohort 2	Kohort 3	Kohort 4	Kohort 5	Kohort 6	Kohort 7
Anker 1	1 000	Anker 1	Anker 2	Kohort 3	Kohort 4	Kohort 5	Kohort 6	Kohort 7
Anker 2	1 000	Kohort 1	Kohort 2	Anker 3	Anker 4	Kohort 5	Kohort 6	Kohort 7
Anker 3	1 000	Kohort 1	Kohort 2	Kohort 3	Kohort 4	Anker 5	Anker 6	Kohort 7
Anker 4	1 000	Ny tekst	Kohort 2	Kohort 3	Kohort 4	Kohort 5	Kohort 6	Anker 7

Figur 6. Ankerdesign for elektronisk leseprøve

Som beskrevet tidligere, er alle testsettene kalibrert sammen hvert år, inkludert ankerprøven og i regning og engelsk var 2014 det første året av måling av endringer over tid. Da ble gjennomsnittet på hver prøve satt på 50 skalapoeng med standardavviket 10, som beskrevet her litt senere. Dette betyr at oppgaveparametrene fra 2014 på ankeroppgavene er brukt igjen 2015. Dette setter prøvene fra begge disse årene på samme skala og gjør resultatene fra dem sammenliknbare.

Ankeroppgaver

Vi vet av erfaring at ankeroppgavene endrer seg over tid. De kan bli ødelagt på grunn av at innholdet blir utdatert, eller at de for eksempel av mange forskjellige grunner blir lettere eller vanskeligere over tid. Det er derfor nødvendig å analysere om ankeroppgavene har endret seg, etter hver gjennomføring. Et eksempel på samme ankeroppgaver to år på rad er vist på figur 7, hvor vanskegrad for ankeroppgavene i engelsk 8. trinn er vist for 2014 og 2015.



Figur 7. Vanskegrad av ankeroppgaver i engelsk 2014 og 2015.

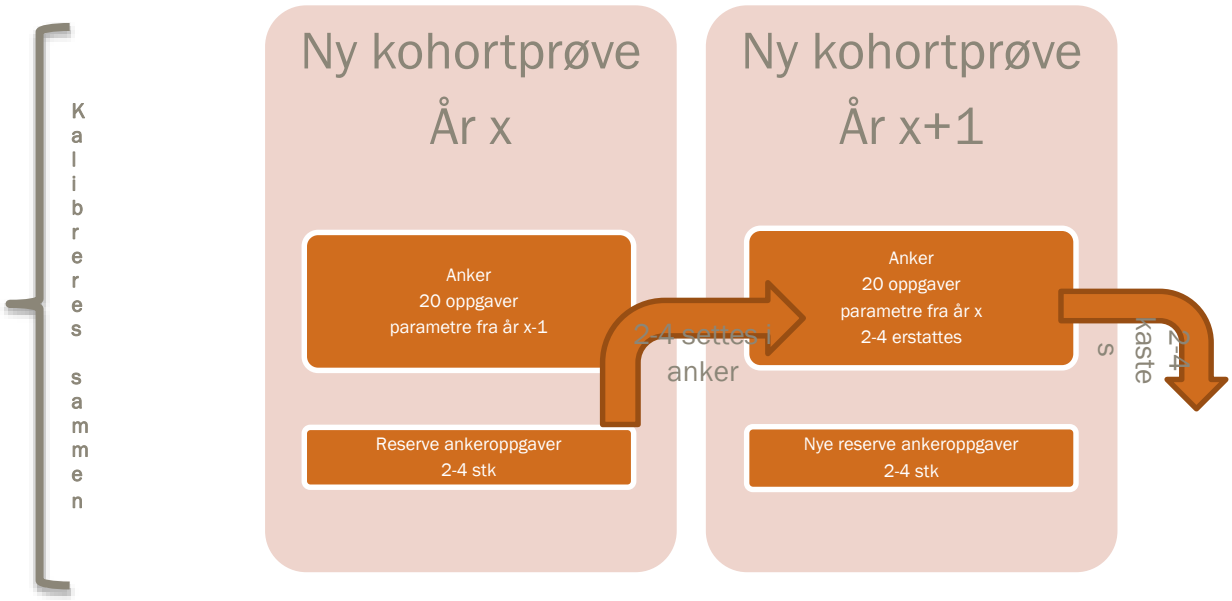
Her viser bildet at ankeroppgavene i engelsk har ikke endret seg signifikant, men disse endringene blir testet statistisk på mange måter, ut ifra vanskegradens standardmålefeil og med DIF analyse av ankeroppgavene fra begge årene, kalibrert sammen. I dette tilfellet var fire ankeroppgaver signifikant endret, to ble litt lettere og to ble litt vanskeligere så i det store og hele var ankeret uendret.

Dersom det har skjedd endringer, og de er over en viss grense, så må de oppgavene som endrer seg erstattes. Dette krever at det da må foreligge nye erstatnings ankeroppgaver som allerede er utprøvd og kalibrert sammen med en kohortprøve og en ankerprøve, og som dermed allerede er på samme skala som forrige ankerprøve. På denne måten blir det mulig å vedlikeholde ankeret med nye oppgaver og også fornye det over tid, og samtidig sikre at samme skala blir brukt hvert år. Og dette medfører at nye oppgaver i ankeret vil ikke endre det eller skalaen, siden de ble opprinnelig kalibrert sammen med ankeret.

For engelsk og regning sin del er det naturlig å sette inn to til fire nye ankeroppgaver hvert år, og ta ut to til fire gamle når vi har et sett med 20 ankeroppgaver. I lesing vil endringen være proporsjonalt større siden en hel tekst med tilhørende oppgavesett må byttes ut. Figur 8. viser et skjematisk bilde av denne ankerfornyelsen.

Det er viktig å understreke at kalibrering av nye ankeroppgaver skjer sammen med **kohort og anker året før de blir brukt i ankeret**, fordi det er eneste måten å sikre at ankeret vedlikeholdes riktig og at skalaen på oppgavene og prøvene blir den samme fra det første året og videre.

I engelsk og regning må man derfor inkludere minimum to til fire nye ankeroppgaver hvert år som kalibreres med hele settet og i lesing må en til to nye tekster inkluderes for samme formål hver gang. Ankerdesignet for lesing har denne muligheten.



Figur 8. System for ankerfornyelse

Skalapoeng

IRT-analysen leverer resultater i form av såkalte theta-verdier (θ) som er omtrent fra -3 til +3 med et gjennomsnitt på 0 for estimering av elevprestasjon. Alle oppgavenes vanskegrader er også uttrykt på samme skala. Av flere grunner er det imidlertid ikke spesielt hensiktsmessig å bruke disse verdiene direkte i rapporteringen, blant annet fordi negative verdier fort kan misforstås og bli tolket som fravær av den dyktigheten som prøven måler. I stedet er det brukt en skala som uttrykker elevenes dyktigheter som et positivt tall og uten bruk av desimaler. Dette er såkalte skalerte skårer (Tan & Michel, 2011) som er vanlig praksis i alle store testsystemer. I nasjonale prøver er disse kalt skalapoeng.

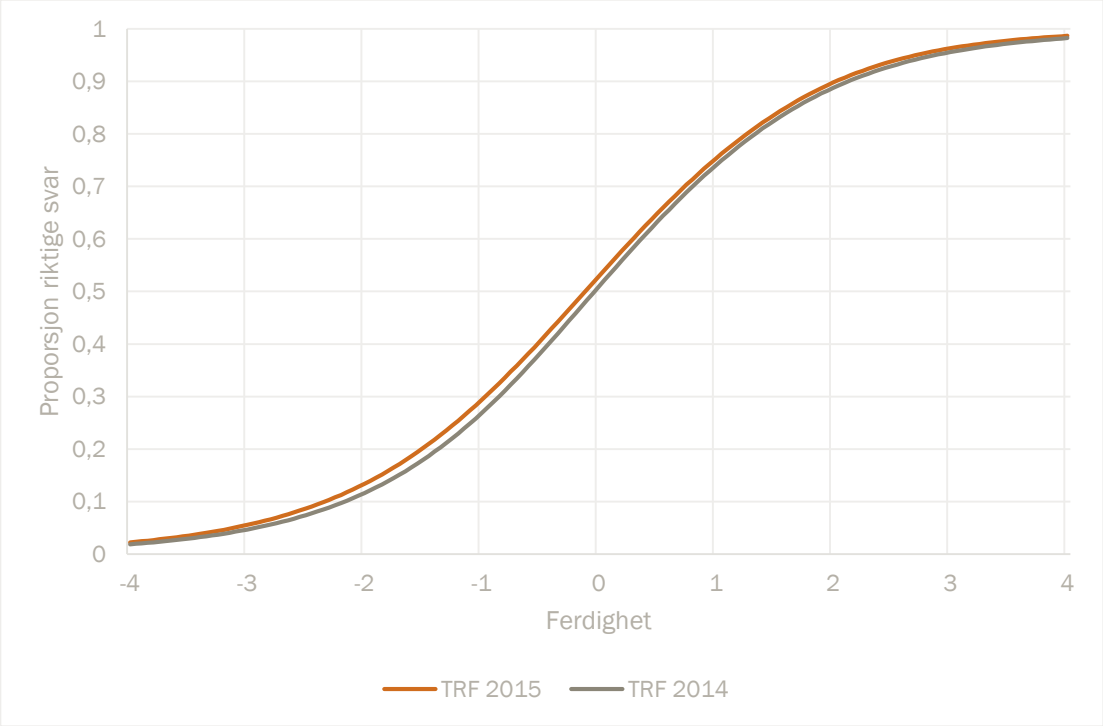
Mer presist, uttrykkes elevenes skåre i de nasjonale prøvene på en skala med et nasjonalt gjennomsnitt på 50 poeng og med standardavviket 10. Dette er en enkel lineær transformasjon av tallene fra IRT-analysen, der Skalapoeng= θ*10+50. Her kunne vi ha brukt hvilken som helst skala, men denne ble

valgt fordi den har en tallrekke som likner på poengsummen fra prøvene og derfor har en sammenliknbar presisjon. Hvis vi hadde valgt lavere tall, ville det ha vært nødvendig å bruke desimaler. Dette anså vi ikke som ønskelig. Hvis vi hadde valgt et høyere tall, som for eksempel et gjennomsnitt på 500, så ville det ha gitt et inntrykk av høyere presisjon enn prøvene egentlig har.

Det må understrekes at valget av tall for skalaen ikke bare er et psykometrisk hensyn, men styres av mange andre ting som nevnt over. Den valgte skalaen er en såkalt t-skåre. Fordelen er at den er vanlig og velkjent i tillegg til å være enkel og forståelig. Den enkle transformasjonen av theta-verdiene fra IRT-analysen bevarer i tillegg fordelingen av skårer i hele elevgruppen på en god måte, noe som er nødvendig for alle skalerte skårer.

Test respons funksjonen

Som nevnt tidligere, så leverer IRT-analysen en sum av informasjonsverdier og målefeil for hele ferdighetsfordelingen i hver prøve (Figur 4.). Men den leverer også en såkalt TRF, en Test Respons Funksjon, som er en god beskrivelse av hvordan sammenhengen er mellom de skalerte skårene i form av skalapoeng og antall riktig besvarte oppgaver på prøven. Figur 9. viser slike kurver for en av de nasjonale prøvene for årene 2014 og 2015 der 2015-prøven allerede er lenket til 2014-prøven.



Figur 9: Regning 8. trinn 2014 og 2015 etter lenking

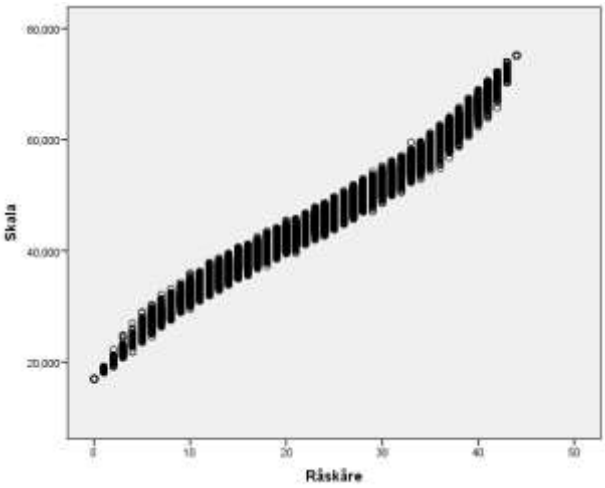
Figur 9. viser at det er forskjell mellom prøvene disse to årene. Prøven fra 2015 er litt endret slik at elevene må svare riktig på noen flere oppgaver i 2015 enn i 2014 for å få samme resultat. Oppgavene i 2015-prøven er gjennomsnittlig litt lettere enn i 2014. Og det er viktig å understreke her at denne forskjellen er ofte som vist på bildet, ikke den samme alle steder på ferdighetsskalaen. På figuren ser vi at over +1 på ferdigheten er prøvene disse to årene nesten helt identiske, men under gjennomsnittet er det litt forskjell på dem. Uten en operasjon der vi lenker prøvene sammen ved hjelp av ankeroppgavene, ville det ha vært umulig å si noe om endring over tid, for hele populasjonen og/eller mindre grupper.

En analyse av denne typen blir gjort for alle prøvene, hvert år, slik at prøveutviklerne får gode opplysninger om hvordan hele prøven virker og hvordan den har endret seg fra forrige år. Disse opplysningene bør også brukes i prøveutviklingen, slik at prøvene blir så like som mulig fra år til år.

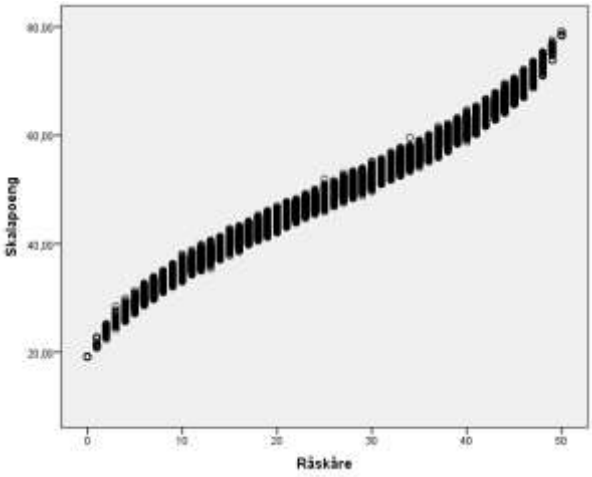
Skåring av individresultater

Når en bruker en to-parameter modell i IRT, så er det endelige resultatet for hver elev, en vurdering av besvarelsene på en slik måte at hvis eleven svarer riktig på vanskelige oppgaver, blir resultatet bedre enn hvis eleven svarer riktig på like mange lette oppgaver og hvis eleven svarer riktig på mange høyt diskriminerende oppgaver blir resultatet også bedre. Derfor er det ikke en én til én korrespondanse mellom råskåre-poengsum (antall riktige svar) og skalapoengene, men en maksimum sannsynlighets estimering («Maximum likelihood estimation») av svarmønsteret til eleven. Skåringsprosedyren tar hensyn til antall riktig besvarte oppgaver og oppgavenes vanskegrad og diskriminering, bruker altså hele svarmønsteret til eleven for å finne fram til det endelige resultatet.

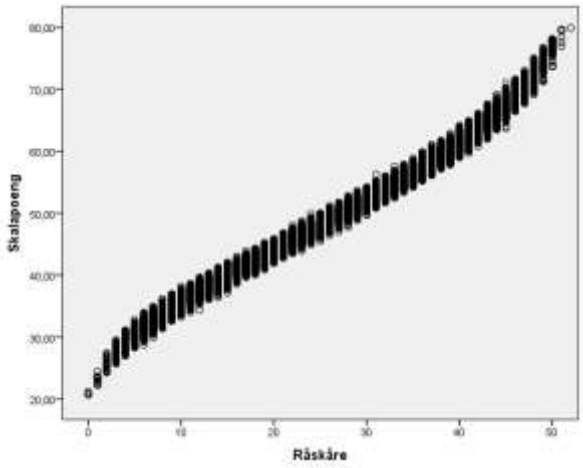
I grunnen er denne maksimum sannsynlighets estimeringen av hver elevs ferdighet, en oppsummering av alle sannsynligheter for svar på alle oppgaver i prøven for det svarmønstret eleven hadde, som resulterer i den mest sannsynlige plassering av eleven på ferdighetsskalaen. Og denne estimeringen er en mange ganger mer presis estimering av ferdigheten enn en enkel oppsummering av poeng ville kunne være. Det finnes noen forskjellige metoder for å gjøre denne estimeringen, men nasjonale prøver bruker en såkalt EAP («Expected a Posteriori») metode som ansees å være den mest presise. Sammenhengen mellom råskåre og skalapoeng vises på figurene 10, 11 og 12. for prøvene i lesing, regning og engelsk på 8. trinn 2015. Her er det tydelig at samme råskåre kan føre til ulikt antall skalapoeng, avhengig av hvilke oppgaver eleven har besvart riktig.



Figur 10. Råskåre og skalapoeng, nasjonale prøver i lesing på 8. trinn 2015



Figur 11. Råskåre og skalapoeng, nasjonale prøver i regning på 8. trinn 2015



Figur 12. Råskåre og skalapoeng, nasjonale prøver i engelsk på 8. trinn 2015

Mestringsnivåer

I tillegg til å rapportere resultatene som skalapoeng for hver elev og som gjennomsnitt for ulike grupper, blir både oppgaver fra prøven og elever plassert på mestringsnivåer. Dette bygger på den egenskapen til IRT analysen å kunne plassere både oppgaver og elever på samme skala. For å gi mer mening til resultatene blir alle oppgavene plassert på mestringsnivåer, 5 nivåer for 8. trinn og 3 nivåer for 5. trinn. Dette ble gjort opprinnelig for engelsk og regning etter prøvegjennomføringen i 2014, på en slik måte for 8. trinn at 10 % av elevene havner på nederste nivå, 20 % på nivå 2, 40 % på nivå 3, 20 % på nivå 4 og 10 % på nivå 5. Denne statistiske inndelingen bygger på praksis fra nasjonale prøver før 2014. For 5. trinn havner 25 % på nivå 1, 50 % på nivå 2 og 25 % på nivå 3 det første året.

Det må legges vekt på her at denne statistiske inndelingen skjer bare det første året av måling av utvikling over tid. Grensene i skalapoeng som ble etablert for regning og engelsk i 2014 og de grensene som blir etablert for lesing høsten 2016, vil holdes fast i årene fremover. Dette betyr at det blir mulig å følge med på antallet elever på hvert mestringsnivå. Figur 13 viser grensene for mestringsnivåene i skalapoeng.

8. trinn	Mestringsnivå 1	Mestringsnivå 2	Mestringsnivå 3	Mestringsnivå 4	Mestringsnivå 5
Engelsk	til og med 36	37 til 43	44 til 55	56 til 62	63 og høyere
Regning	til og med 36	37 til 44	45 til 54	55 til 62	63 og høyere

5. trinn	Mestringsnivå 1	Mestringsnivå 2	Mestringsnivå 3
Engelsk	til og med 42	43 til 56	57 og høyere
Regning	til og med 42	43 til 56	57 og høyere

Figur 13. Mestringsnivåer for engelsk og regning.

Fagmiljøene som utvikler alle oppgaver lager i tillegg beskrivelser av ferdigheten for alle nivåer, slik at elevens plassering på nivå blir meningsfull for både elever og lærere og kan benyttes i videre læring for eleven.

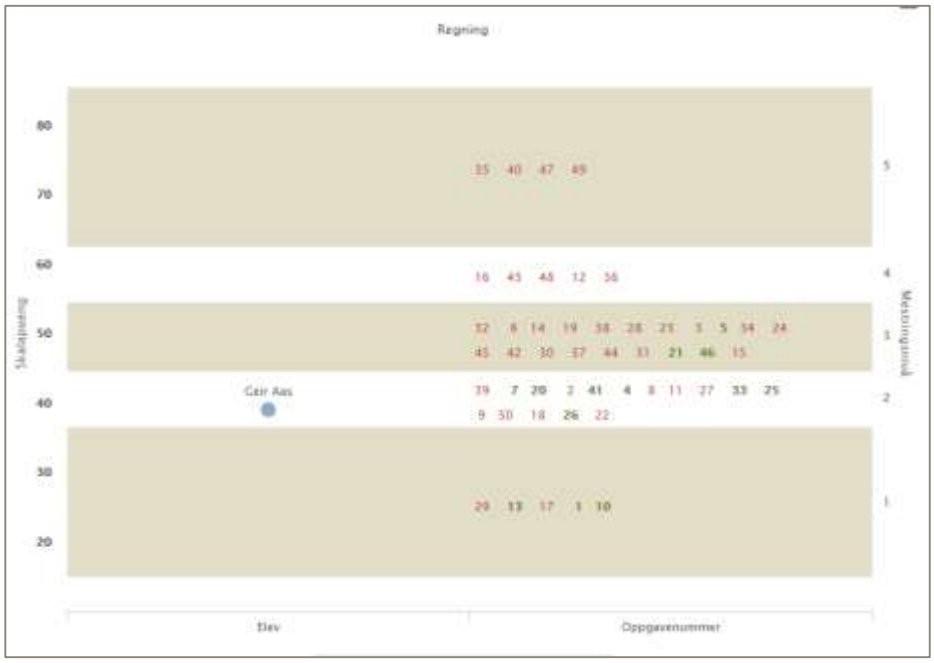
I følgende tabell vises et eksempel på hvordan oppgaver fra en prøve faller på de ulike mestringsnivåene. Eksemplet er fra prøven i regning for 8. trinn i 2014 og tabellen viser navnet på oppgaven og tilhørende vanskegrad i skalapoeng. Her vises det veldig klart at det er få oppgaver på øverste og nederste nivå men flest på nivå 3. Hvis en ser på denne fordelingen av oppgaver på ferdighetskalaen, så er lett å forstå hvorfor usikkerheten (avhengig målefeil) i målingen er størst nederst og øverst på skalaen.

Nivå 1		Nivå 2		Nivå 3		Nivå 4		Nivå 5	
R5070475	32,14	R5064667	37,67	R5041693	46,12	R5041829	55,40	R5061830	65,19
R5058460	32,35	R5065984	37,92	R5059281	46,16	R5058589	55,89	R5063786	65,27
R5058733	35,27	R5041664	38,90	R5061848	46,26	R5041775	56,24	R5061782	69,39
R5062646	36,16	R5060221	40,30	R5065964	46,34	R5059020	56,42		
		R5058725	40,43	R5041729	46,41	R5041797	56,78		
		R5041826	41,34	R5065992	46,80	R5041818	56,86		
		R5058731	41,83	R5041705	46,88	R5065947	56,95		
		R5061863	41,94	R5065949	47,02	R5058726	57,62		
		R5041809	42,19	R5065981	47,54	R5059132	58,60		
		R5059804	42,66	R5061881	47,96	R5059296	59,14		
		R5063754	43,33	R5086018	48,45	R5059319	59,21		
		R5041745	43,55	R5058072	48,56	R5061549	59,91		
		R5061115	43,65	R5059343	48,76	R5041670	60,26		
		R5062376	43,90	R5059789	48,79	R5041704	60,63		
		R5085926	44,13	R5041736	49,54	R5065960	60,84		
		R5065976	44,48	R5065971	49,63	R5064630	61,21		
		R5065870	44,54	R5060216	49,75	R5058566	61,47		
		R5059788	45,61	R5061867	50,14				
		R5062467	45,68	R5065937	50,26				
		R5041701	45,82	R5061757	50,33				
				R5041798	50,82				
				R5065988	50,85				
				R5061720	51,30				
				R5041845	51,50				
				R5059783	51,64				
				R5065994	51,76				
				R5059145	51,93				
				R5041687	53,81				
				R5061113	53,88				
				R5041792	53,98				
				R5064665	54,02				
				R5041835	54,49				
				R5065975	54,57				
				R5064536	54,89				

Her er alle oppgaver vist, både kohortprøven (58 oppgaver) og ankerprøven (20 oppgaver) og disse er på samme skala. Ved nærmere analyse av individresultater er det derfor mulig å se hvilke oppgaver ligger på samme nivå som elevens ferdighet og hvilke ligger rett over og under den og i tillegg se hvilken usikkerhet er bundet til plasseringen av eleven på nivå. Sammenliknet med mestringsnivåene før 2014, er presisjonen i dette økt vesentlig for alle elevresultater, i tillegg til det at nå er oppgavene også plassert på samme skala og dermed på samme mestringsnivåer som tabellen viser. Dette gir lærere helt nye muligheter for videre arbeide med elevene basert på prøveresultatene

Rapportering av resultatene

En av de gode egenskapene til IRT-analysen, som beskrevet tidligere, er at vanskegraden til oppgavene i prøven og elevenes skåre kommer på samme skala. Dette gjør det mulig å rapportere resultatene på en mer meningsfull måte enn tidligere. Figur 14. viser et eksempel på rapportering av resultater for en enkeltelev i analyserapporten som blir publisert i PAS og lærere har tilgang til for sine elever.

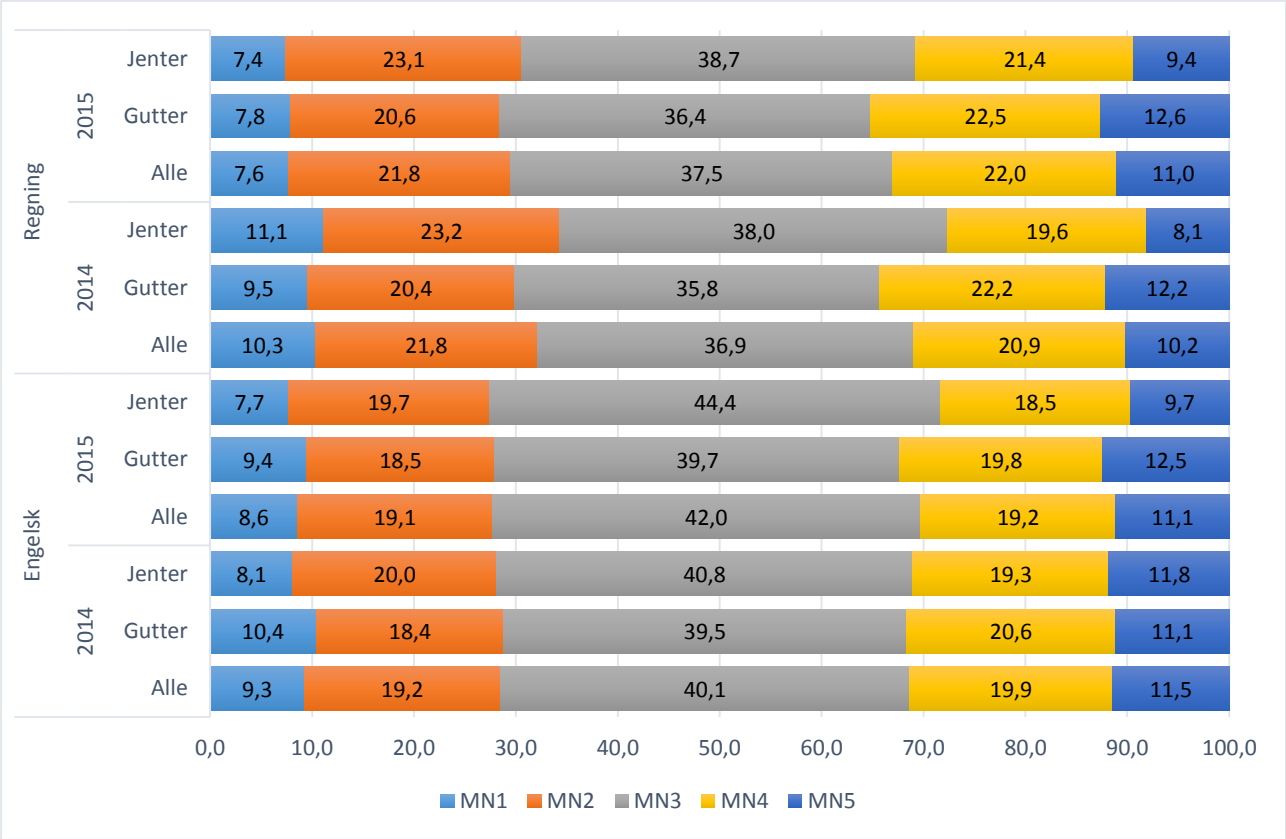


Figur 14. Eksempel fra analyserapporten i PAS (fiktive data)

Eksemplet viser at eleven (blå prikk) er plassert på mestringsnivå 2, og tallene til høyre viser oppgavene i prøven og hvilke av dem eleven klarte (grønn skrift) og ikke klarte (rød skrift). I tillegg kommer det tydelig frem hvor den enkelte oppgave er plassert på skalaen. Dette er nyttig informasjon for læreren. Det gir blant annet mulighet til å konkretisere ved hjelp av oppgaver, hva som kjennetegner elevers dyktigheter på ulike nivåer. Det er viktig for enhver lærer å se på ikke bare hvilke oppgaver eleven klarte men også på hvilke oppgaver eleven nesten klarte og ved å gå inn i oppgavene kan læreren avdekke forhold og kunnskapshull hos eleven. I eksemplet her er det to oppgaver på laveste nivå (1) som eleven ikke klarte og de må læreren se på og prøve å forstå hvorfor ikke eleven klarte dem siden de måler en vesentlig lavere ferdighet enn eleven sannsynligvis har. Det samme gjelder for oppgaver fra et nivå langt over elevens ferdighet som han likevel klarer, der må læreren finne ut ved å se på oppgaven og innholdet til den, hva kan forklare det.

På denne måten kan man også over tid, når man har kjent vanskegrad for mange oppgaver over flere år, beskrive typiske kjennetegn på progresjon langs skalaen. På direktoratets hjemmeside kan en se en film der denne analyserapporten presenteres: http://demorapport.udir.no/ .

Resultatene brukes også til å se hvordan ulike grupper presterer, alt fra nasjonalt nivå ned til skoler og klasser. Siden mestringsnivåene er satt fast første året av måling av endring over tid, er antallet elever på hvert mestringsnivå veldig viktig informasjon for videreutvikling av hele systemet. På figur 15 vises et eksempel på hvordan disse dataene på systemnivå kan brukes i tillegg til å rapportere gjennomsnitt for de ulike gruppene.



Figur 15. Proporsjon elever på hvert mestringsnivå 2014 og 2015. Regning og engelsk for 8. trinn.

Bildet viser at i det store og hele så er ikke store endringer fra år til år, men noen av dem er veldig interessante, f.eks. det at antallet elever på laveste mestringsnivå i regning ser ut til å synke fra 2014 til 2015. På skoleporten.no er det mulig å se mange forskjellige visninger av disse resultatene, for forskjellige perioder og grupperinger.

Detaljene og databearbeiding

Det er naturligvis mange detaljer i disse metodene og den databehandlingen som er nevnt, som ikke er blitt omtalt her, men hensikten med dette avsnittet er å nevne de viktigste, slik at de som er interessert selv kan lese mer om dette og finne den relevante litteraturen.

I analysearbeidet med nasjonale prøver blir det brukt mange typer analyseprogrammer som SAS, SPSS og EXCEL, men IRT-analysen er hovedsakelig utført med programmet Xcalibre (Assessement Systems Corporation, 2014). Analysen er også delvis gjentatt og verifisert med programmet IRTPRO (SSI, 2011).

Som nevnt tidligere baserer analysen seg på bruk av en 2PL modell, og det brukes også en blandet modell for de prøvene som også har polytome oppgaver, dvs. oppgaver med graderte svar («Samejima Graded Response Model» - SGRM).

Kalibreringen av oppgaver er sentrert på den estimerte ferdigheten (theta) og EM-algoritmen i kalibreringen benytter en MML-Marginal Maximum Likelihood estimering. Det siste bidrar til å sikre så langt som mulig at oppgaveparametrene blir invariante (uavhengige av gruppe elever) og at estimeringen av de svakeste og flinkeste elevene blir mer korrekt enn ved bruk av metoder som bruker andre maximum likelihood estimeringsmetoder. Det første året (2014) av bruk av disse metodene ble derfor gjennomsnittet på alle prøvene satt til 50 med standardavviket 10, men på grunn av ankringen mellom år vil disse tallene kunne endre seg over tid.

IRT-analysen er alltid kjørt med test av DIF (Differential Item Functioning) for kjønn. Dette gjøres for å finne ut hvilke av oppgavene favoriserer enten jenter eller gutter. Her bør det legges vekt på at dette er ikke en utprøving av forskjeller mellom kjønnene, men utprøving av om noen oppgaver favoriserer det ene kjønnet fremfor det andre, gitt samme nivå av ferdighet i det prøven måler. Dette er altså en utprøving av om prøven inneholder systematiske skjevheter («bias»).

Ved estimeringen av individers ferdighet (skåring av resultater) anvendes den såkalte EAP – Expected a Posteriori metoden, som er en av de nyere metodene for å estimere (skåre) enkeltindividbesvarelser på en hel prøve.

Alle analyser er gjort flere ganger av minst to personer, uavhengig av hverandre. Deretter er resultatene sammenliknet og kvalitetssikret i alle trinn av analysearbeidet.

For nærmere forklaringer av disse tekniske begrepene anbefales Embretson & Reise (2000).

Oppsummering

Hensikten med denne artikkelen har vært å redegjøre for at analysen av nasjonale prøver foretas på et solid psykometrisk grunnlag. Ved å føre sammen den høye faglige kvaliteten på oppgaveutviklingen i fagmiljøene som leverer oppgaver til nasjonale prøver, de nye elektroniske prøvene, IT-systemene og disse psykometriske metodene for skalering, kalibrering og lenking mellom år, så blir resultatet et robust prøvesystem som kan gi pålitelig og nyttig informasjon til både lærere, skoleledere, skoleeiere og det nasjonale nivået.

Artikkelen gir en oversikt over de metodene som brukes i analysen av nasjonale prøver fra og med 2014. Metodene som blir brukt er godt forankret i forskningslitteraturen og har et stort potensiale for å forbedre nasjonale prøver. Imidlertid er psykometriske skalerings - og kalibreringsmetoder i rask utvikling og vi vil derfor endre vårt metodegrunnlag i tråd med utviklingen på dette feltet.

Mange små detaljer i analyseprosessen er imidlertid utelatt i denne fremstillingen. Dere som er interessert kan se på referanselisten; her er litteratur på mange nivåer som detaljert beskriver mange av de aspektene som bare er så vidt nevnt her.

Referanser

Brennan, R.L. (Ed) (2006). Educational Measurement. American Council on Education, Praeger.

Embretson, S.E. & Reise, S.P. (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates, Inc.

Guyer, R. & Thompson, N.A. (2014). Users Manual for Xcalibre item response theory calibration software, version 4.2.2 and later. Woodbury MN: Assessment Systems Corporation.

Kolen, M. J., & Brennan, R. L. (2014). Test equating, scaling, and linking (3rd ed.). New York: Springer.

Lord, F.M. (1980). Applications of Item Response Theory to Practical Testing Problems. Routledge Taylor and Francis Group, New York and London.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Chicago: University of Chicago Press.

Standards for Educational and Psychological Testing (1999). American Educational Research Association, American Psychological Association, & National Council of Measurement in Education.

SSI:Scientific Software International. (2011) IRTPRO: Users Guide. Lincolnwood IL.

Tan, X & Michel, R. (2011). Why do Standardized Testing programs Report Scaled Scores? Why not just report the raw or percent-correct scores? Educational Testing Service: R&D Connections no. 16, september 2011.

Xcalibre version 4.2. (2014). Assessment Systems Corporation.

Telefon 23 30 12 00
www.utdanningsdirektoratet.no

Utdanningsdirektoratet