Statistisk sentralbyrå
Statistics Norway

Reestimering av nasjonale prøveresultater 2014–2021

Kjell Solem Slupphaug og Samuel Jones Gilje

TALL SOM FORTELLER
NOTATER / DOCUMENTS
2025/32

I serien Notater publiseres dokumentasjon, metodebeskrivelser, modellbeskrivelser og standarder.

Publisert: 13. november 2025
Rettet 16. desember 2025: side 3

ISBN 978-82-587-2042-0 (elektronisk)
ISSN 2535-7271 (elektronisk)

Standardtegn i tabeller	Symbol
Ikke mulig å oppgi tall Tall finnes ikke på dette tidspunktet fordi kategorien ikke var i bruk da tallene ble samlet inn.	.
Tallgrunnlag mangler Tall er ikke kommet inn i våre databaser eller er for usikre til å publiseres.	..
Vises ikke av konfidensialitetshensyn Tall publiseres ikke for å unngå å identifisere personer eller virksomheter.	:
Desimaltegn	,

Forord

Statistisk sentralbyrå (SSB) har foretatt en vurdering av muligheten for å oppdatere den offisielle statistikken som bruker resultater fra nasjonale prøver, etter at det ble oppdaget feil i resultatene for årene 2014 – 2021. I dette notatet presenterer vi tre metoder for reestimering av resultater for nasjonale prøver, for bruk på SSBs data, og drøfter styrker og svakheter med de ulike metodene.

Forfatterne vil rette en takk til personer i Udir, Frischsenteret og CEMO, for kommentarer til metodenotatet, med en særskilt takk til Björn Anderson, Henrik Ræder og Ole Røgeberg.

Statistisk sentralbyrå, 10. november 2025

Ann-Kristin Brændvang

Sammendrag

En feil i utregningen av skalapoengene som måler elevers ferdighetsnivå i lesing, regning og engelsk i nasjonale prøver gjennomført i årene 2014 – 2021 ble avdekket i en artikkel publisert av Markussen et al., (2024). Metoden som ble benyttet til å måle utvikling i elevenes ferdighetsnivå over tid ble ikke implementert på riktig måte, og ga dermed et upresist bilde av endringer i ferdighetsnivå over tid. I 2024 estimerte Utdanningsdirektoratet (Udir) skalapoengene på nytt, for alle nasjonale prøver på 5. og 8. trinn i perioden 2014 – 2021, ut ifra en bedre egnet metodikk.

SSB publiserer den offisielle statistikken for nasjonale prøver, og kobler disse resultatene til ulike variabler som innvandrerkategori og foreldrenes utdanningsnivå. Udirs nye resultater kan dessverre ikke kobles direkte til SSB sin gamle data, da Udir ikke har beholdt personidentifiserende informasjon.

Hovedformålet med dette metodenotatet er å vurdere ulike metoder SSB kan bruke til å korrigere den offisielle statistikken, og eventuelt andre statistikker som benytter seg av resultater fra nasjonale prøver – slik at de reflekterer de nye estimatene fra Udir. Vi presenterer tre alternativer til hvordan man kan reestimere skalapoengene: lokal omskalering, global omskalering og en regresjonstilnærming.

Vi vurderer «*goodness of fit*» til de ulike metodene og undersøker hvordan valg av modell påvirker resultater på utvalg av ulike størrelse. Dette er viktig da SSB publiserer statistikk aggregert på ulike gruppenivå, og ikke for enkeltindivider. Et slikt eksempel kan være resultater fra nasjonale prøver i norske fylker. Det er da nøyaktigheten av resultatene på slike gruppenivå som burde anses som aller viktigst, hvor nøyaktigheten på individnivå er av mindre viktighet.

Generelt sett presterer alle metodene nokså likt når man aggregerer på gruppenivå, og forskjellene i *«goodness of fit»* er relativt små. Vi anbefaler derfor at man benytter seg av den globale omskaleringsmetoden, som er lettest å anvende, har færrest antagelser, og som ikke tilføyer ny støy i dataene.

Innhold

1. Introduksjon

I en artikkel som undersøkte metodene benyttet av Utdanningsdirektoratet (Udir) til å estimere skårer på ferdighetsnivå blant elever (*skalapoeng*) på nasjonale prøver for lesing, regning og engelsk (5. og 8. trinn), avdekket Markussen et al., (2024) en potensiell feil i utregningen av disse skalapoengene. Ifølge de var det sannsynligvis en feil i programmet (`XCalibre`), som ble brukt til å estimere skalapoengene. Etter en intern vurdering konkluderte Udir (2024) med at *«Resultatene gir ikke et presist bilde av hvordan elevenes ferdigheter har utviklet seg».*

Når nasjonale prøver ble innført i 2004 ble prøvene brukt for å gi et øyeblikksbilde av skoleelevenes ferdigheter i lesing, regning og engelsk. Resultatene for en gitt elev kunne ses i kontekst av andre elevers skårer innenfor samme prøve, men kunne ikke si noe om ferdighetsutviklingen blant elever over tid. Fra og med 2014 (for prøvene i regning og engelsk) og fra og med 2016 (for prøvene i lesing), endret Udir metodikken for å kunne sammenligne prøveresultatene fra år til år, og dermed fange opp utvikling i elevenes ferdigheter over tid. Det er denne endringen i metodikk som ikke ble riktig implementert. Metoden plukket dermed ikke opp endringer i skalapoeng over tid, og ga ikke riktig bilde av utviklingen i ferdighetsnivå i regning, lesing og engelsk.

Der hvor internasjonale undersøkelser som PISA og TIMSS viste betydelige endringer i ferdighetsnivåene blant grunnskoleelever over tid, viste de offisielle statistikkene basert på tall fra Udir kun minimale endringer i lesing, regning og engelsk (Markussen et al., 2024). Når Udir analyserte resultatene på nytt, med bedre egnet metodikk, fant de at resultatene i realiteten viste større endringer i ferdighetsnivå over tid. SSB benytter tall fra nasjonale prøver i offisiell statistikk, tilgjengeliggjør data fra nasjonale prøver til forskere og i microdata.no, og gjør en rekke analyser og oppdrag der data fra nasjonale prøver inngår. Det har derfor vært viktig for SSB å sikre et grunnlag for å vurdere om SSB skal oppdatere allerede publisert statistikk og tilhørende mikrodata, og hvordan dette kan gjøres.

1.1. Bruk av nasjonale prøver i SSB

SSBs offisielle statistikk basert på nasjonale prøver

I henhold til Nasjonalt program for offisiell statistikk 2024-2027 er SSB som beskrevet under statistikkområdet *Utdanning* ansvarlig produsent av offisiell statistikk over nasjonale prøver.

SSB har i tråd med dette en egen statistikkside over karakterer og nasjonale prøver i grunnskolen[1], og tall for nasjonale prøver publiseres årlig. Resultater for nasjonale prøver 2024 ble publisert 14. november 2024. Statistikken omfatter åtte statistikkbanktabeller[2] med resultater for nasjonale prøver hvorav syv har tidsserier som inkluderer hele perioden 2014-2021 som er berørt av manglene ved Udirs tidligere beregning av skalapoeng.

Nasjonale prøver inngår også i KOSTRA-indikatorer for grunnskolen (elever på mestringsnivå 3-5 på nasjonale prøver 8. trinn i lesing og regning)[3].

[1] https://www.ssb.no/utdanning/grunnskoler/statistikk/karakterer-ved-avsluttet-grunnskole
[2] Karakterer og nasjonale prøver i grunnskolen. Statistikkbanken (ssb.no)
[3] 12255: Utvalgte nøkkeltall for grunnskole (K) 2015 - 2022. Statistikkbanken (ssb.no)

I tillegg brukes elevenes tidligere resultater på nasjonale prøver som bakgrunnsvariabel i en del øvrige statistikker, herunder i en statistikkbanktabell over standpunktkarakterer[4] og to statistikkbanktabeller i statistikk over gjennomføring i videregående opplæring[5].

Frem til korrigerte tall kan foreligge er det i SSBs statistikktabeller informert om manglene ved tidligere estimeringer.

Bruk av nasjonale prøvedata hos SSB i analyser, tabelloppdrag og datadeling med forskere
I tillegg til bruk av data fra nasjonale prøver i faste statistikkbanktabeller benytter SSB data fra nasjonale prøver i en rekke tabelloppdrag og analyser for ulike oppdragsgivere. Nasjonale prøver er også en sentral del av datagrunnlaget for beregning av skolebidragsindikatorer for mellom- og ungdomstrinnet.

Videre er nasjonale prøver etterspurt for utlån av data fra SSB til en rekke ulike forsknings-prosjekter. Data fra nasjonale prøver legges også inn i Nasjonal utdanningsdatabase og deles i mikrodata.no.

1.2. Hvorfor burde SSB oppdatere statistikker?

Som nevnt ovenfor er data for nasjonale prøver 2014-2021 benyttet på en rekke måter i SSB. SSB kobler data om elevprestasjonene på nasjonale prøver til andre egenskaper som foreldrenes utdanningsnivå, innvandringsbakgrunn og sentralitet av bostedskommune. SSBs offisielle statistikk blir brukt for å beskrive trender i disse gruppene, sett opp imot hverandre og den øvrige befolkningen. Når det i etterkant viser seg at disse tallene gir et upresist bilde av utvikling over tid, har det vært viktig med en vurdering av hvordan dette bør håndteres i den offisielle statistikken. Man kunne valgt å la de gamle tallene stå med en forklaring av problematikken, eller reestimere skalapoengene, og deretter oppdatere statistikken samt gjøre det mulig å korrigere underliggende mikrodata som gjøres tilgjengelig for forskere. I dette metodenotatet undersøker vi tre ulike metoder for å reestimere elevenes skalapoeng, og gir en anbefaling for hvilken av metodene SSB bør benytte.

Skalapoeng (2014 – 2021)

Udir har valgt å betegne estimatet på ferdighetsnivået til en elev på nasjonale prøver (2014 – 2021) for skalapoeng. Skalapoeng er et standardisert mål, som er måles i forhold til et referanseår. Ved refereanseåret til et prøveemne, ble skalapoengene standardisert slik at de hadde et snitt på 50, og standardavvik på 10. For regning og engelsk ble snittet og standardavviket satt i 2014, mens for lesing ble det satt i 2016. Skalapoeng er relativt til skalapoengene fra tidligere år. Derav kan skalapoeng endres over tid, for å reflektere endringer i det generelle ferdighetsnivået til norske grunnskole elever. Dermed kan snittet og standardavviket (spredningen) til skalapoengene endres over tid, i takt med endringer i ferdighetsnivå. Endringer i skalapoeng over tid vil også innebære endringen i hvordan elevmassen fordeler seg på de ulike mestringsnivåene på nasjonale prøver, ettersom mestringsnivå tilordnes ut ifra skalapoengene.

I 2024 estimerte Udir skalapoengene på nytt for alle nasjonale prøver i perioden 2014 – 2021, ut ifra en bedre egnet metodikk. Optimalt sett burde det ha vært mulig å koble de nye estimatene direkte på SSBs gamle data, men dette er dessverre ikke mulig. Udir er pålagt å slette personidentifiserende informasjon etter en viss tid, og har derfor ikke individdata på oppgavenivå for nasjonale prøver 2014-2021 med personidentifiserende informasjon. Når Udir har beregnet skalapoengene på nytt har de derfor benyttet avidentifiserte data med elevers svar på enkeltoppgaver, total poengsum og kjønn. De nye skalapoengene kan dermed ikke kobles direkte til personidentifiserende informasjon som f.eks. fødselsnummer. SSB har ikke tidligere hentet inn og lagret data på oppgavenivå for

[4] 08533: Elever fordelt på standpunktkarakterer, etter fag, prøve og mestringsnivå på nasjonale prøver på 8. trinn 2010 - 2023. Statistikkbanken (ssb.no)
[5] Gjennomføring i videregående opplæring – SSB

nasjonale prøver. Data for nasjonale prøver 2014-2021 som er lagret hos SSB med mulighet for kobling via pseudonymisert person-ID omfatter kun data for den enkelte prøven, ikke data om hvordan eleven gjorde det på den enkelte oppgaven. Disse begrensningene gjør det ikke mulig å direkte koble reestimerte skalapoeng fra Udir med SSBs lagrede gamle data for nasjonale prøver. Problemet til SSB har da vært å finne en måte å oppdatere statistikken for 2014-2021 og tilhørende mikrodata i lys av dette.

Det er viktig å notere seg at forskjellene mellom grupper på en gitt prøve innad i et år på nye og gamle skalapoeng vil være relativt like. Altså vil en korrigering av statistikken basert på de nye skalapoengene i mindre grad endre forholdet mellom grupper, innad samme år og prøve. Hvis man derimot sammenligner elever på tvers av år vil bruken av nye skalapoeng kunne gi større utslag. Eksempelvis var det en relativt stor økning i skalapoeng i engelsk 8.trinn fra 2014 til 2021. Forskjellen mellom en gitt gruppe (f.eks Trøndelag fylke) i 2014 og en gitt gruppe i 2021 vil da sannsynligvis være større når man ser på de nye skalapoengene, i forhold til de gamle skalapoengene.

Tabell 1.2.1 Oversikt over tilgjengelige variabler i SSBs data, og nye data fra Udir

Variabler	SSB-data (X_{yt})	Udir-data (Y_{yt})
Personidenter[a]	Ja	Nei
Kjønn	Ja	Ja
Enkeltoppgaver	Nei	Ja
Poengsum	Ja	Ja
Gamle Skalapoeng ($\boldsymbol{\theta_g}$)	Ja	Nei
Nye Skalapoeng ($\boldsymbol{\theta_n}$)	Nei	Ja
Andre variabler	Ja	Nei

Notat: a. Personidenter referer til direkte identifiserende personidentifiserende informasjon som navn, fødselsdato, fødselsnummer. Samt diverse indirekte personidentifiserende informasjon som skole, bostedskommune.

1.3. Hvordan burde SSB oppdatere statistikker?

Siden det ikke er noen enkel koblingsnøkkel mellom de nye skalapoengene til Udir, og de gamle skalapoengene til SSB, må man finne alternative metoder for å få gode estimater på nye skalapoeng i SSBs data. I dette notatet presenterer vi tre forskjellige metoder, som kan benyttes til å reestimere skalapoengene ut ifra SSB sine data. Metodene refereres til som:

1. Global Omskalering
2. Regresjonstilnærming
3. Lokal Omskalering

Metodene presenteres nærmere i seksjon 2.2

Vi vurderer de ulike tilnærmingene opp mot hverandre etter to hovedprinsipp: «*goodness of fit*» (GOF) og «*parsimony*». Kort oppsummert vil man ha en modell som presterer så godt som mulig, men som samtidig er enkel å kommunisere utad, lett å anvende, har minimale antakelser, og tilføyer minst mulig ny støy til dataene.

For å vurdere modellene opp mot hverandre, bruker vi Udir sitt anonyme datasett med alle besvarelsene til alle prøvene for 5. og 8. trinn for årene 2014 – 2021. Dette er fordi det er i Udir sine nye data vi kan få de beste estimatene på gamle og nye skalapoeng, i ett og samme datasett. Altså har vi i de nye dataene både tilgang til de faktiske nye estimatene til Udir, men også tilgang til svar på enkeltoppgaver, som lar oss estimere gamle skalapoeng, ut ifra Markussen et al., 2024.

Vi vil vurdere 1. Hvordan modellene presterer på individnivå, og 2. Hvordan de presterer på gruppenivå. Det er den sistnevnte som vi anser til å være av størst relevans, gitt at SSB sine tall publiseres på gruppenivå, og ikke individnivå.

1.4. En introduksjon til IRT-modeller

Ferdighetsnivå på nasjonale prøver er målt gjennom det Udir kaller for skalapoeng, som er estimert ved bruk av *Item Response Theory* (IRT). Kort oppsummert forsøker IRT-modeller å identifisere karakteristikker ved oppgaver, som beskriver sannsynligheten for å besvare (response) en oppgave (item) riktig, gitt distribusjonen av et underliggende (latent) ferdighetsnivå hos respondentene (Lord, 1952; Embretson & Reise, 2000). I denne seksjonen vil vi gi en kort introduksjon til estimeringen av IRT-modeller, og estimeringen av faktorskårer – som blir benyttet til å kalkulere skalapoeng. I Udir sine estimeringer har man benyttet seg av en *'Generalized Partial Credit'* (GPC) model. GPC-modellen er en mer komplisert IRT-modell, som tillater ordinale oppgaver, med graderte skårer (e.g., en oppgave hvor man kan få en poengsum fra 0-4). Når dette er sagt er de fleste oppgavene på nasjonale prøver dikotome. Altså man får enten riktig (1) eller feil (0) skår. For dikotome oppgaver tilsvarer GPC-modellen det samme som en *'2 parameter logistic'* (2PL) modell (Embretson & Reise, 2000). For enkelhets skyld, vil vi i denne seksjonen fokusere på 2PL-modellen.

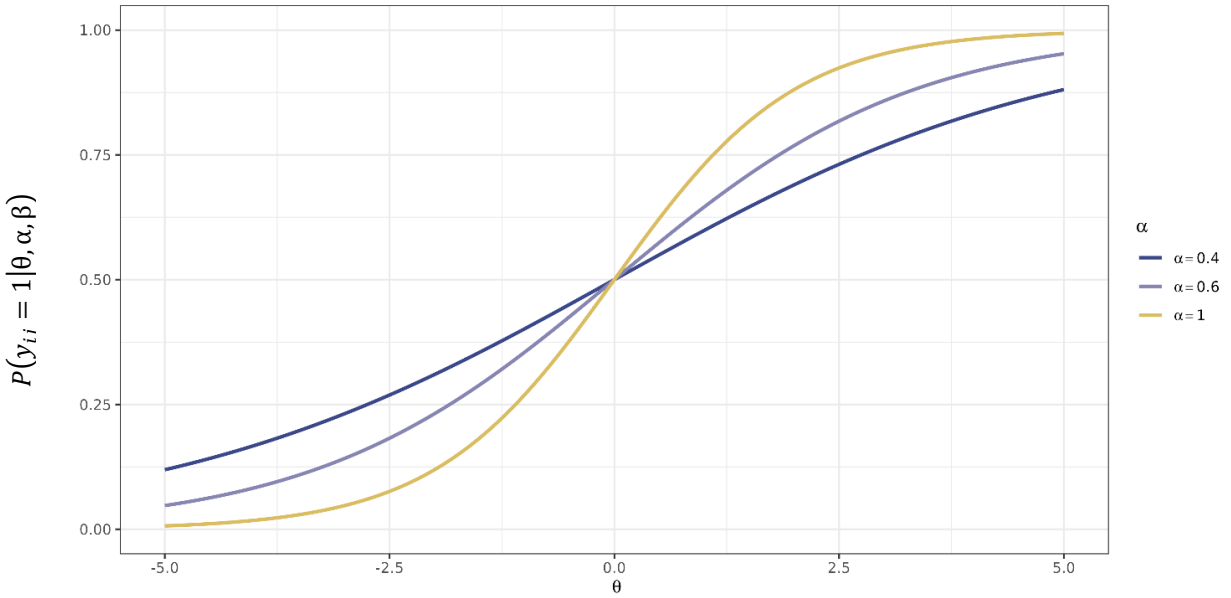
2PL IRT-Modeller

I 2PL-modellen er item karakterisert av to parameter α og β, som gir en sannsynlighetsfordeling for at en person med en gitt verdi θ_i (en latent variabel), har fått skåren 1 på item j.

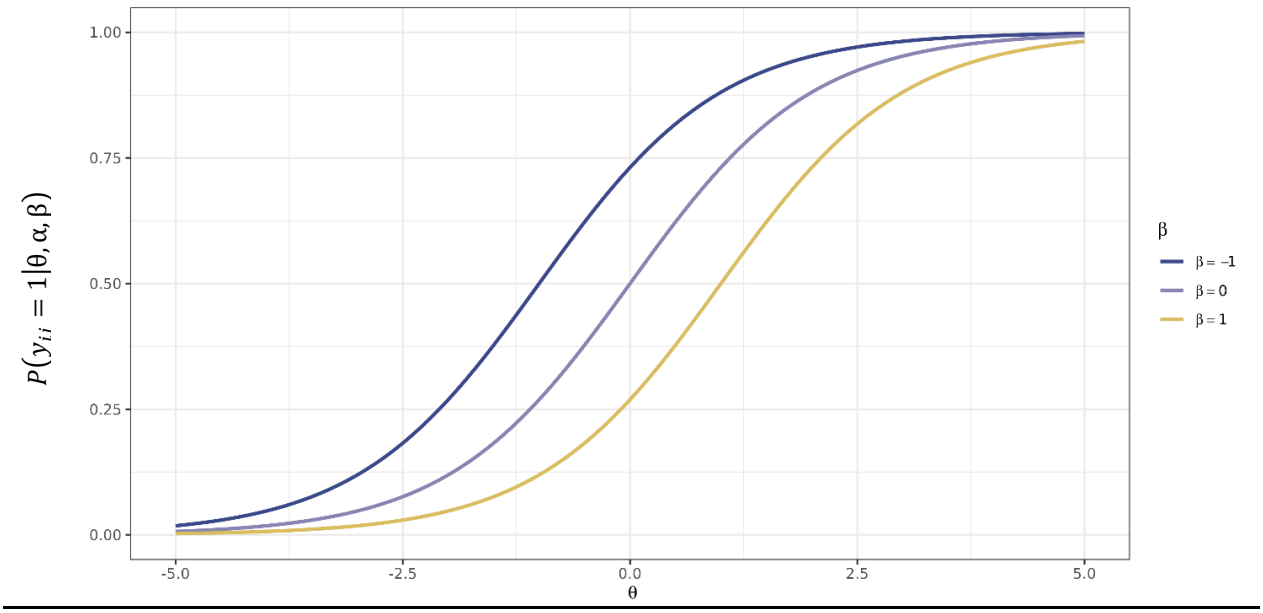
$$P(y_{ij} = 1|\theta_i \alpha_j, \beta_j) = \frac{e^{\alpha_j(\theta_i - \beta_j)}}{1 + e^{\alpha_j(\theta_i - \beta_j)}}$$

α måler den diskriminerende effekten til itemet (hvor raskt sannsynligheten $P(y_{ij} = 1)$ endres ved endring i θ_i). Dette gjøres ved å endre stigningen på funksjonen. Ved en bratt stigning er det en rask og stor endring i sannsynligheten, noe som da gir et større skille mellom respondenter som vist i **Figur 1.4.1**. β representerer i praksis vanskelighetsgraden til itemet, og representerer verdien av $\boldsymbol{\theta}$ hvor 50 prosent av respondentene vil få en skår på 0 og 1 se **Figur 1.4.2**. Når vanskelighetsgraden β øker må respondenten ha en høyere $\boldsymbol{\theta}$ verdi for å ha samme sannsynlighet for en skår av 1.

Figur 1.4.1 2PL α - diskriminering



Figur 1.4.2 2PL β – vanskelighetsgrad



I teorien har hver respondent (i) sitt eget parameter θ_i som representerer respondentens underliggende verdi av $\boldsymbol{\theta}$. I praksis blir ikke denne verdien eksplitt identifisert for hvert individ. I stedet modellerer man θ_i implisitt, via $\boldsymbol{\theta}$ sin tilhørende sannsynlighetsfordeling. Dette er ofte gjort ved å bruke et numerisk approksimert integral av sannsynligheten for å observere en gitt verdi av $\boldsymbol{\theta}$ i dataene, ganger sannsynligheten for å observere dataen gitt verdien av $\boldsymbol{\theta}$. For eksempel kan man beregne sannsynligheten for å observere gitt data, gitt itemenes (oppgavenes) parameter ved å benytte en «*fixed Gauss-Hermite Quadrature*» ved

$$\ell = \sum_{i=1}^{N} ln\left(\sum_{q=1}^{Q} w_q \prod_{j=1}^{J} P(y_{ij}|\theta_q, \alpha_j, \beta_j)\right)$$

Hvor ℓ er den naturlige logaritmen av sannsynligheten for å observere de observerte dataene ($\boldsymbol{y}$), gitt $\theta_q, \alpha_j, \beta_j$. Videre er y_{ij} svaret på item j til repsondent i, α_j er diskrimineringsevnen- og β_j er vanskelighetsgraden til et item j, N er antall respondenter, J er antall item, Q er antall integreringspunkt (antall «*quadratures*»), og w_q er den assosierte vekten til et integreringspunkt q langs $\boldsymbol{\theta}$. Vektene w_q er gitt ved

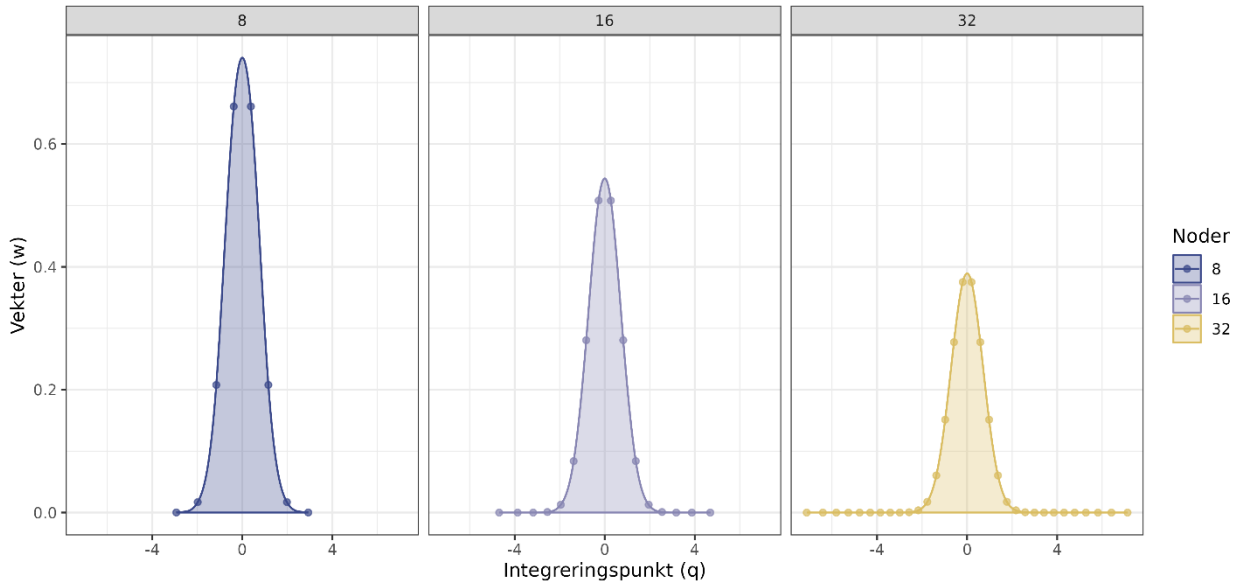
$$w_q = \frac{2^{Q-1} Q! \sqrt{\pi}}{Q^2 \left[H_{Q-1}(x_q)\right]^2}$$

Hvor H_Q er den Q. hermit polynomfunksjonen. Hvor den vektede summen av $f(x_q)$ approksimerer integralet av normalfordelingen ganger $f(x)$, slik at

$$\sum_{q=1}^{Q} w_q f(x_q) \approx \int_{-\infty}^{\infty} e^{-x^2} f(x)\, dx$$

Hvor $f(x)$ byttes ut med ℓ. w_q er et estimat på sannsynligheten for at en observasjon befinner seg innenfor et gitt intervall langs $\boldsymbol{\theta}$. Ved flere integreringspunkt blir intervallene representert av w_q mindre. Ved $\lim_{Q \to \infty} w_q$ vil vektene approksimere normalfordelingen.

Figur 1.4.3 ***Fixed Gauss-Hermite Quadrature***



Det er verdt å merke seg at funksjonene brukt til å kalkulere logen av sannsynligheten (ℓ) benyttet i både `mirt` og `XCalibre` er mer kompliserte enn «*fixed Gauss-Hermite Quadrature*» funksjonen, men vil dele mange av de samme hovedtrekkene, i form av et numerisk approksimert integral (Chalmers, 2012).

Faktorskårer

I etterkant vil man kunne estimere verdien av $\boldsymbol{\theta}$ til en respondent, gitt de estimerte parameterne (α, β) til itemene (oppgavene) som respondenten har svart på. Det finnes forskjellige tilnærminger for å estimere $\boldsymbol{\theta}$. For eksempel kan man maksimere logen av sannsynligheten ved:

$$L(\boldsymbol{\theta}_i|\boldsymbol{y_i}, \theta_i, \alpha_j\beta_j) = \sum_{j=1}^{J} ln\, P(y_{ij}|\theta_i, \alpha_j\beta_j)$$

Slik at de estimerte skårene $\boldsymbol{f}$ av $\boldsymbol{\theta}$ gis ved

$$\boldsymbol{f} = arg\, max_\theta\{L(\boldsymbol{\theta}|\boldsymbol{y_i}, \alpha_j, \beta_j)\}$$

Hvor P er funksjonen som evaluerer den sannsynligheten for å observere y_{ij} gitt $\theta_i, \alpha_j\beta_j$. Siden man allerede vet verdiene til y_{ij}, α_j og β_j trenger man kun å finne verdien θ_i som maksimerer sannsynligheten for å observere svarene ($y_{i1}, y_{i2}, \ldots y_{ij}$) til en gitt person (Embretson & Reise., 2000). Slike estimater på latente variabler omtales ofte som faktorskårer (Mehmetoglu & Venturini, 2021). Udir sine skalapoeng er de estimerte faktorskårene for $\boldsymbol{\theta}$ ($\boldsymbol{f}$), hvor skalapoengene ($\boldsymbol{f}^*$) gis ved

$$\boldsymbol{f}^* = 10\boldsymbol{f} + 50$$

Merk at dette er det samme som å forhåndsdefinere μ og σ til 50 og 10 før man estimerer modellen.

Fixed common item parameters

En av fordelene med IRT-modeller, er at de åpner for muligheten til å sammenligne resultater fra flere utvalg, selv om de ikke har samme oppgavesett. Det finnes flere måter å gjøre dette på, men Udir valgte å gjøre dette ved å benytte det de kalte for ankeroppgaver. Ankeroppgaver er i essens

oppgaver som er repetert i flere årganger, som da gjør det mulig å sammenligne prestasjoner på tvers av år.

I utgangspunktet er snittet (μ) og standardavviket (σ) til fordelingen av $\boldsymbol{\theta}$ arbitrært. Altså det er likeså riktig å si at μ er 50, og σ er 10, som det er å si at μ er 0, og σ er 1. Hvis μ og σ er uvisst på forhånd, er modellen som regel ikke identifiserbar (dvs., det er et uendelig antall gyldige kombinasjoner av parameterne ($\mu, \sigma, \alpha, \beta$)). Hvis man derimot allerede vet parameterne for en eller flere av itemene ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$), kan man bruke disse til å begrense antall gyldige kombinasjoner av parameterne. Altså man kan ikke identifisere parameterne til itemene ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$), hvis parameterne (μ, σ) til $\boldsymbol{\theta}$ ikke er identifisert på forhånd. På samme måte kan man heller ikke identifisere parameterne (μ, σ) til $\boldsymbol{\theta}$, hvis ingen av oppgaveparameterne ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) er identifisert på forhånd. Hvis man i 2014 først estimerer en modell hvor μ er 0 og σ er 1, kan man neste år repetere noen oppgaver (ankeroppgaver), som man allerede har estimert parameterne ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) på. Hvis man da forhåndsdefinerer parameterne ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) til ankeroppgavene, kan man frigjøre parameterne (μ, σ) til $\boldsymbol{\theta}$. Hvis man da for eksempel finner at μ er høyere i 2015, impliserer det et høyere ferdighetsnivå enn i 2014. Denne tilnærmingen refereres ofte til som «*fixed common item parameters*» (FCIP; Li et al., 1997).

1.5. Hva gikk feil i tidligere beregning av skalapoengene for nasjonale prøver 2014-2021?

Markussen et al., (2024) gir en forklaring på det de mener gikk galt i `XCalibre`, programvaren som Udir benyttet i tidligere beregning av skalapoeng for nasjonale prøver. De peker på at μ og σ aldri var frigjort i årene etter 2014. Altså `XCalibre` forhåndsdefinerte både parameterne på ankeroppgavene ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) til fjorårets verdier, og μ og σ til 0 og 1. Det var da ikke mulig for `XCalibre` å finne nye estimat på snittene og standardavvikene i ferdighetsnivå, siden de allerede var forhåndsdefinert.

Den eneste grunnen til at vi allikevel ser noe variasjon i snitt på skalapoeng fra år til år, er på grunn av at man ikke ser på parameterne μ og σ estimert (altså forhåndsdefinert) i modellen, men på snittet og standardavviket som kommer ut ifra faktorskårene (estimatet av $\boldsymbol{\theta}$), estimert i etterkant. Vanligvis vil snittet og standardavviket på faktorskårene ligge veldig tett opp imot μ og σ i modellen, men det er ikke alltid tilfellet hvis modellen ikke har en konsistent løsning. Altså i `XCalibre`-modellen har man forhåndsdefinert både parameterne for ankeroppgavene ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) samt μ og σ. Hvis det da blir en endring i ferdighetsnivå ($\boldsymbol{\theta}$) over tid, vil dette være en selvmotsigelse. Altså, modellen sier først at elevene gjør det bedre enn hva man gjorde før, men samtidig sier modellen også at elevene gjør det akkurat like bra som før. Når man da estimerer faktorskårer vil man få variasjon i både snittet og standardavviket, som man ikke ville ha observert i modellparameterne μ og σ. Dette forklarer da hvorfor man observerer nok variasjon, til at engelsk 8. trinn hadde en økning til 51 skalapoeng i 2021.

Denne feilen hadde ikke en uniform effekt på elevene på nasjonale prøver. Kortfattet førte det til at forholdet mellom ankerelevene og de øvrige elevene ble feilestimert. Relativt til hverandre/ gjennomsnittet fikk de øvrige elevene en overestimering i takt med positive endringer, mens ankerelevene fikk en underestimering. I tilfeller hvor det var en reduksjon i skalapoeng over tid så man den motsatte effekten. Siden det er langt færre ankerelever enn øvrige elever, var effekten mer markant for ankerelevene enn de øvrige elevene. I hovedsak balanserte disse effektene hverandre ut, slik at snittet av overestimeringen til de øvrige elevene og underestimeringen til ankerelevene utjevnet hverandre. Dermed var det liten endring i snittet for hele populasjonen, tross endringer blant ankerelevene og de øvrige elevene.

2. Metoder

2.1. Estimering av gamle skårer $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$

For å kunne følge utvikling i elevenes ferdigheter over tid må resultatene fra et referanseår kobles til resultatene fra påfølgende år. Som nevnt i del 1.3 valgte Udir lenkemetoden *«Fixed common item parameters»* (FCIP). FCIP metoden krever at ankerelevene er bredt fordelt over ferdighetsnivået i populasjonen, noe som Markussen et al., (2024) undersøkte. De sammenlignet ankerelevenes svar på «kohort oppgaver» (oppgaver som alle elevene svarte på), da bare en viss andel av oppgavene som en ankerelev svarer på er ankeroppgaver. De fant ingen signifikante forskjeller mellom ankerelever og andre elever innad i kullene på 95% konfidens nivå, med unntak av lesing i 5. trinn. Der var det allikevel små nok forskjeller til at de konkludert med at *«utviklingen over tid ikke er påvirket av skjeve utvalg av ankerelever».*

Modellen er videre utviklet til å ta hensyn til at endringer i elevpopulasjonen kan påvirke elevenes sannsynlighet for å korrekt besvare en oppgave, uavhengig av deres ferdighetsnivå i et fag. For eksempel kan en anta at andelen elever som svarer riktig på en oppgave angående analog klokke i regneprøven påvirkes av deres erfaring med analoge- fremfor digitale klokker – uten at det reflekterer det underliggende matematiske ferdighetsnivået. Slike ankeroppgaver burde i så fall identifiseres og tas ut av analysen. Slike oppgaver kan identifiseres med såkalt «*differential item response function*» (DIF).

Udir har brukt `DIF` funksjonen i `mirt` pakken (Chalmers, 2012) for å analysere ankeroppgaver for tegn av DIF. De har deretter tatt en kvalitativ vurdering av resultatene, før de har valgt ut hvilke oppgaver de ekskluderer fra analysen (Udir 2024 – referat). I vår estimering av gamle skalapoeng har vi brukt en liste over ankeroppgaver med DIF egenskaper levert av Udir, til å ekskludere de gjeldende oppgavene. Listen er gjengitt i vedlegg A.

Videre er vår estimering av gamle skalapoeng basert på funnene i Markussen et al., (2024), hvor både parameterne til ankeroppgavene ($\boldsymbol{\alpha}$, $\boldsymbol{\beta}$) og parameterne til $\boldsymbol{\theta}$ (μ, σ) er forhåndsdefinert. Der andre modellspesifikasjoner brukt i `mirt` avviker mellom det som er presentert i Markussen et al., (2024) og Udir sitt referat (Udir 2024 – referat) har vi valgt å følge Udir sine spesifikasjoner. Koden er vedlagt i vedlegg B.

De estimerte skalapoengene som skal reflektere de gamle skalapoengene til Udir ($\boldsymbol{\theta}_{\boldsymbol{g}}$) refereres videre til som $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$. Hvor $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ simpelthen er de estimerte faktorskårene ($\boldsymbol{f}$) fra en IRT-modell med modellspesifikasjonene spesifisert ovenfor.

$$\boldsymbol{f} = arg\ max_{\theta}\{L(\boldsymbol{\theta}|\boldsymbol{y_i}, \alpha_j, \beta_j)\}$$

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}} = 10\boldsymbol{f} + 50$$

Videre refereres vårt estimat av de nye skalapoengene til Udir ($\boldsymbol{\theta}_{\boldsymbol{n}}$) som $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$. Vi vil nå beskrive tre forskjellige alternativer for å estimere $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$.

2.2. Tre alternativer for å estimere nye skårer $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$

Global omskalering

Den enkleste metoden er en omskalering av gamle skalapoeng slik at både gjennomsnittet (μ) og standardavviket (σ) for hele populasjonen blir likt de nye resultatene fra Udir, for hver prøve og hvert år:

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} = \frac{\sigma_{\theta n}}{\sigma_{\theta g}}\left(\boldsymbol{\theta}_{\boldsymbol{g}} - \mu_{\theta g}\right) + \mu_{\theta n}$$

Hvor $\mu_{\theta n}$ og $\sigma_{\theta n}$ er gjennomsnittet og standardavviket til $\boldsymbol{\theta}_{\boldsymbol{n}}$, og $\mu_{\theta g}$ og $\sigma_{\theta g}$ er gjennomsnittet og standardavviket til $\boldsymbol{\theta}_{\boldsymbol{g}}$. Dette antar at feilen i `XCalibre` i hovedsak påvirker snittet og standardavviket til utvalget, og kun i mindre grad rangeringer av skalapoeng i dataene. Det forventes systematiske forskjeller mellom ankerelever og øvrige elever, men ikke særlig endring i rangering innad disse (Markussen et al., 2024). Forskjellene mellom ankerelevene og de øvrige elevene er derimot relativt små, og det burde ikke ha en stor effekt på resultatene. Altså selv om det kan være en relativt stor forskjell mellom nye og gamle skalapoeng for et gitt individ, vil forskjellen mellom individer være relativt like ved både nye og gamle skalapoeng.

Regresjonstilnærming

En mer kompleks, men potensielt mer nøyaktig, metode er å bruke en regresjonstilnærming. Dette innebærer å finne sammenhengen mellom *nye skalapoeng* og variablene *gamle skalapoeng*, *kjønn*, og *poengsum* i Udir sin nye data, ved bruk av en regresjonsmodell. Deretter kan man estimere et sett med koeffisienter, som lar oss danne et estimat på nye skalapoeng i SSBs gamle data. Dette gjøres da ved en vektet sum av *gamle skalapoeng*, *kjønn* (som dummy variabel) og *poengsum*.

Regresjonstilnærmingen har følgende fremgangsmåte. Siden vi har tilgang til respondentenes svar på enkeltoppgavene i det nye datasettet fra Udir, $\boldsymbol{Y}_{\boldsymbol{yt}}$ (hvor y = år, og t = prøve), kan vi danne et estimat ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$) på de gamle `XCalibre`-skårene ($\boldsymbol{\theta}_{\boldsymbol{g}}$), slik som forklart i del 2.1. Vi har da et anonymt datasett med både de nye skårene til Udir ($\boldsymbol{\theta}_{\boldsymbol{n}}$), samt et estimat på de gamle skårene ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$). Vi gjør deretter en regresjonsanalyse for å finne sammenhengen mellom $\boldsymbol{\theta}_{\boldsymbol{n}}$ og $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$, gitt kjønn ($\boldsymbol{k}$) og poengsum ($\boldsymbol{p}$). Videre antar vi at relasjonen mellom $\boldsymbol{\theta}_{\boldsymbol{n}} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}} \mid \boldsymbol{k}, \boldsymbol{p}$ i datasett $\boldsymbol{Y}_{\boldsymbol{yt}}$ er proporsjonal til sammenhengen mellom $\boldsymbol{\theta}_{\boldsymbol{n}} \sim \boldsymbol{\theta}_{\boldsymbol{g}} \mid \boldsymbol{k}, \boldsymbol{p}$ i vårt gamle datasett $\boldsymbol{X}_{\boldsymbol{yt}}$.

Det burde være god grunn til å anta dette, gitt at både $\boldsymbol{\theta}_{\boldsymbol{g}}$ og $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ er kalkulert på samme data, med forholdsvis lik metodikk. Selv om konklusjonen til Markussen et al., (2024) skulle være feil (brukt i estimeringen av $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$) ville man fortsatt forventet at relasjonene er relativt proporsjonale til hverandre. Dette er fordi man ser at korrelasjonene mellom $\boldsymbol{p}$ og $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}/\boldsymbol{\theta}_{\boldsymbol{g}}$ er relativt høye i begge datasett, hvor som regel ligger et sted mellom $.90 - .95$ (noe som er forventet Emertson, 2001). Dette viser da at begge deler mye av sin varians, bare gjennom delt varians med $\boldsymbol{p}$. I verst tenkelig scenario burde korrelasjonskoeffisienten (r) mellom $\boldsymbol{\theta}_{\boldsymbol{g}} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ være lik $r(\boldsymbol{\theta}_g, \boldsymbol{p})r(\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}, \boldsymbol{p})$. Derimot, dersom konklusjonene til Markussen et al., (2024) skulle være feil ville det sannsynligvis vært bedre å benytte en av de andre metodene.

Gitt denne antagelsen kan vi da bruke koeffisientene $\boldsymbol{\beta}$ (β_0, β_1, β_2, β_3) fra regresjonsanalysen til å estimere $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ i $\boldsymbol{X}_{\boldsymbol{yt}}$.

$$\boldsymbol{\theta}_{\boldsymbol{n}} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}} \mid \boldsymbol{k}, \boldsymbol{p} = \beta_0 + \beta_1 \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}} + \beta_2 \boldsymbol{k} + \beta_3 \boldsymbol{p} + \boldsymbol{\varepsilon}$$

Hvor $\boldsymbol{\varepsilon}$ er feilleddet. Når vi har fått koeffisientene $\boldsymbol{\beta}$ fra regresjonsanalysen på det nye datasettet $\boldsymbol{Y_{yt}}$ kan vi anvende disse på det gamle datasettet $\boldsymbol{X_{yt}}$, for så å bytte ut $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ med $\boldsymbol{\theta_g}$, slik at vi får et estimat ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$) på $\boldsymbol{\theta_n}$.

$$\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} = \beta_0 + \beta_1 \boldsymbol{\theta_g} + \beta_2 \boldsymbol{k} + \beta_3 \boldsymbol{p}$$

Siden feilleddet er ikke tatt med når vi kalkulerer $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, må vi må vi omskalere $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ i etterkant, hvis vi ønsker at variansen til $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ skal være lik variansen til $\boldsymbol{\theta_n}$.

$$\widehat{\boldsymbol{\theta}}^{*}_{\boldsymbol{n}} = \frac{\sigma_{\theta n}}{\sigma_{\hat{\theta}_n}}\left(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}} - \mu_{\hat{\theta}_n}\right) + \mu_{\theta n}$$

I tillegg til at $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ er omskalert i etterkant, er $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ omskalert på forhånd, slik at variansen og snittet er likt variansen og snittet til $\boldsymbol{\theta_g}$. Hvis man ikke gjorde dette, ville det proporsjonalt sett bli feil blanding av $\boldsymbol{\theta_g}$ $\boldsymbol{k}$ og $\boldsymbol{p}$ i $\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$.

Lokal omskalering

Den siste tilnærmingen er nok den mest kompliserte, men har også mulighet for å være den mest nøyaktige. Med lokal omskalering bruker vi samme omskaleringsfunksjon som den globale, men istedenfor å ta hele populasjon for en prøve i ett, deler vi den opp i undergrupper basert på poengsum ($\boldsymbol{p}$) og kjønn ($\boldsymbol{k}$). Hver undergruppe omskaleres slik at gjennomsnitt og standardavvik for undergruppene blir like i de nye og gamle datasettene. En vil da kunne finne forskjellen i gamle og nye skalapoeng for en gitt krysning av prøve, år, kjønn og poengsum.

For eksempel vil man gjøre en separat omskalering for undergruppen jenter ($\boldsymbol{k}$ = 0) med poengsum 37 ($\boldsymbol{p}$ = *37*) i engelsk 8. trinn i 2017 ($\boldsymbol{X_{y=2017\ t=ENG08}}$ og $\boldsymbol{Y_{y=2017\ t=ENG08}}$) slik at snittet og standard-avviket blir likt som for samme undergruppe i begge datasett.

$$\mu(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \mu(\boldsymbol{\theta_n})|_{k=0,\ p=37}$$

$$\sigma(\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}) = \sigma(\boldsymbol{\theta_n})|_{k=0,\ p=37}$$

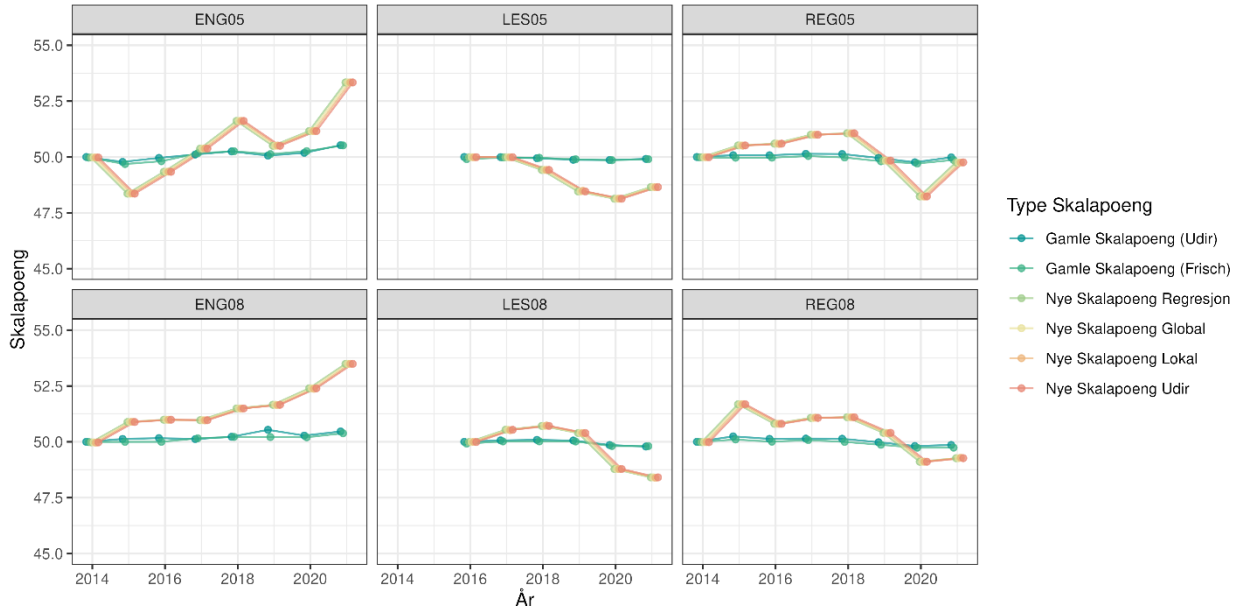
Tilnærmingen kan også ses som en versjon av regresjonstilnærmingen, hvor man behandler kjønn og poengsum som nominelle (kategoriske) variabler, hvor man har saturert alle interaksjonseffekter i modellen, slik at det er en skalering av gamle skalapoeng opp imot nye, for hver undergruppe av kjønn × poengsum. I motsetning til regresjonstilnærmingen antar metoden ikke at forholdet mellom $\boldsymbol{\theta_n} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}} \mid \boldsymbol{k}, \boldsymbol{p}$ i datasett $\boldsymbol{Y_{yt}}$ er proporsjonal til sammenhengen mellom $\boldsymbol{\theta_n} \sim \boldsymbol{\theta_g} \mid \boldsymbol{k}, \boldsymbol{p}$ i vårt gamle datasett $\boldsymbol{X_{yt}}$. Metoden antar derimot at det er en høy korrelasjon mellom $\boldsymbol{\theta_n}$ og $\boldsymbol{\theta_g}$. Som nevnt er det god grunn til å anta dette, gitt korrelasjonen fra poengsum til både $\boldsymbol{\theta_n}$ og $\boldsymbol{\theta_g}$. Gitt at poengsum er den samme variabelen i begge datasett, vil korrelasjonen på sitt laveste være lik $r(\boldsymbol{\theta}_n, \boldsymbol{p})r(\boldsymbol{\theta}_g, \boldsymbol{p})$. Den viktigste antakelsen for metoden er at $\boldsymbol{p}$ og $\boldsymbol{k}$ er de samme variablene i datasett $\boldsymbol{X_{yt}}$ og $\boldsymbol{Y_{yt}}$. Det vil si at man for hver deltaker i vil finne at $p_{ix} = p_{iy}$ og $k_{ix} = k_{iy}$. Hvor ix er den i. deltakeren i datasett $\boldsymbol{X_{yt}}$ og iy er den y. deltakeren i datasett $\boldsymbol{Y_{yt}}$.

I teorien burde dette være sant, gitt at man kan identifisere den i. deltakeren i begge datasett. I realiteten er dette mer problematisk, da det finnes små forskjeller i datasettene mellom SSB og Udir. SSB har for eksempel imputert kjønn basert på fødselsnummer, mens Udir har brukt innrapportert kjønn. Dette betyr at vi ikke får en perfekt match i undergrupperinger i dataene. I noen tilfeller finner man til og med grupper som kun eksisterer i ett av datasettene. I tilfeller hvor en under-gruppe i $\boldsymbol{X_{yt}}$ ikke finnes i $\boldsymbol{Y_{yt}}$ må man i stedet benytte seg av den globale omskaleringsmetoden. I praksis er forskjellene i undergrupperingene små, og man vil sjeldent måtte benytte seg av den globale omskaleringsmetoden for en gitt undergruppe i $\boldsymbol{X_{yt}}$.

3. Analyse og resultater

På det overordnede nivået (skalapoeng per prøve per år) er det per definisjon ingen observerbar forskjell mellom de forskjellige reestimeringsmetodene – siden alle metodene sørger for å ende opp med samme snitt og standardavvik som θ_n. Dette kan ses i **Figur 3.1**

Figur 3.1 Skalapoeng på gamle nasjonaleprøver reestimert med tre ulike metoder. 2014–2021



I figuren ovenfor ser man snittet til de forskjellige versjonene av skalapoengene, på de forskjellige årene, tilhørende de forskjellige prøvene. Hvis man ser på de reestimerte skalapoengene ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$) ser man at alle har samme snitt som de nye skalapoengene fra Udir ($\boldsymbol{\theta}_{\boldsymbol{n}}$). Grunnen til at man kan se alle de forskjellige linjene i grafen, er fordi vi har lagt til et horisontalt skift på x-aksen, slik at de forskjellige $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, og $\boldsymbol{\theta}_{\boldsymbol{n}}$ er synlige. Hvis man derimot ser på y-aksen vil man se at alle har samme verdier. Samtidig kan man også se Udirs gamle skalapoeng ($\boldsymbol{\theta}_{\boldsymbol{g}}$), samt Frisch-estimatet ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$). Der kan man se at de stort sett følger hverandre, men at det i noen tilfeller er noen små forskjeller (e.g., Engelsk 8. trinn, 2019).

Det er ikke noe poeng i å tolke for mye ut ifra resultatene i anvendelsen på vår gamle data, siden vi ikke har noen god måte å vurdere '*goodness of fit*' (GOF) mellom faktiske ($\boldsymbol{\theta}_{\boldsymbol{n}}$) og estimerte ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$) verdier på nye skalapoengene, siden vi ikke har tilgang til de nye skalapoengene til Udir ($\boldsymbol{\theta}_{\boldsymbol{n}}$)**.** I stedet burde vi se på effekten av reestimeringsmetodene våre, i en setting hvor vi faktisk kan gjøre en direkte sammenligning mellom $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ og $\boldsymbol{\theta}_{\boldsymbol{n}}$.

3.1. Hvordan Vurdere '*Goodness of Fit*' mellom $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ og $\boldsymbol{\theta}_{\boldsymbol{n}}$

For å vurdere GOF mellom $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ og $\boldsymbol{\theta}_{\boldsymbol{n}}$ trenger vi en setting hvor vi har tilgang til begge. Dette er såklart ikke mulig i våre gamle datasett (med reestimerte skalapoeng $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$), siden vi ikke har tilgang til $\boldsymbol{\theta}_{\boldsymbol{n}}$. Et alternativ er å bruke de nye dataene fra Udir. I de nye dataene har vi ikke bare tilgang til $\boldsymbol{\theta}_{\boldsymbol{n}}$, men også tilgang til svar på enkeltoppgaver. Dette betyr at vi kan kalkulere estimerte gamle skalapoeng ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$). med Frisch-tilnærmingen, og deretter bruke reestimeringsmetodene til å forsøke å gjenskape de nye skalapoengene til Udir. Vi vil da kunne sammenligne de faktisk nye skalapoengene, med hva våre reestimeringsmetoder vil predikere. Det er verdt å merke at dette ikke en perfekt metode siden det ikke perfekt representerer differansen mellom $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ og $\boldsymbol{\theta}_{\boldsymbol{n}}$, siden vi ikke har tilgang til $\boldsymbol{\theta}_{\boldsymbol{g}}$. Metoden

er dermed avhengig av at forholdet mellom $\boldsymbol{\theta_n} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ er proporsjonalt til forholdet mellom $\boldsymbol{\theta_n} \sim \boldsymbol{\theta_g}$. Gitt at antakelsene til Markussen et al., (2024) om hva som gikk galt i `XCalibre` er korrekte, burde dette være en nokså uproblematisk antagelse. I tilfellet hvor dette ikke stemmer, vil nok GOF-målene fortsatt være nokså gode, siden det uansett er god grunn til å anta en høy korrelasjon mellom $\boldsymbol{\theta_g} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ (se **Kapittel 2.2**).

Kort oppsummert har alle metodene systematiske feil. Når dette er sagt, er ikke målet nødvendigvis å ha en perfekt estimering på individnivå, men heller en *'god nok'* metode, som er nøyaktig på gruppenivå. Vi vil dermed ikke legge stor vekt på konsekvensene på individnivå. Vi vil heller fokusere på hvor godt metodene fungerer når man aggregerer på forskjellige gruppenivå, avhengig av gruppenes størrelser. Altså hva vil for eksempel konsekvensene være for et fylke, en kommune, en skole, eller en skoleklasse?

Per definisjon blir snittene og standardavvikene til de reestimerte skalapoengene på globalt nivå (med den totale populasjonen for hver prøve og hvert år) like de faktiske nye skalapoengene til Udir, som vist i **Figur 3.1**. Vi ser nå på effekten på mindre grupper av ulike størrelser og sammenligner ankerelevene med de andre elevene.

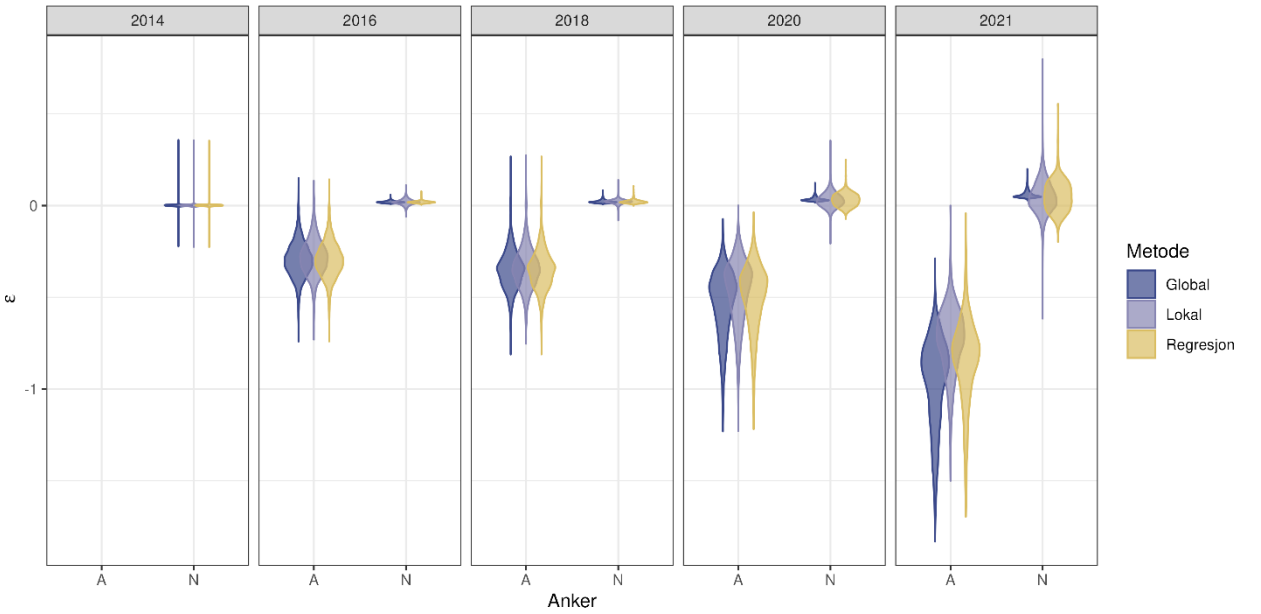
3.2. Resultater

Som nevnt, er ingen av metodene en perfekt tilnærming, hvor alle har systematiske feil. **Figur 3.2.1** viser differansen ($\boldsymbol{\varepsilon}$) mellom $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ og $\boldsymbol{\theta_n}$, hvor

$$\boldsymbol{\varepsilon} = \boldsymbol{\theta_n} - \widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$$

Ankerelevene er markert som *A* på x-aksen, mens de andre elevene er markert med *N*. Vi viser samme diagram for engelsk 8. trinn (ENG08) i fem ulike år, som viser hvordan elevgruppene kommer ut i $\boldsymbol{\varepsilon}$, etter hvert som ferdighetsnivået på engelsk 8. trinn øker over tid. Dette er en konsekvens av at de gamle (`XCalibre`) skalapoengene til elevene ble tvunget av modellen til å passe til referanseåret, og ikke fritt til å bevege seg med ankerelevene som de skulle.

Figur 3.2.1 Fordeling av feilledd for ankerelever og øvrige elever



Notat: A = ankerelevene, N = andre elever. I 2014 var det ingen ankerelever siden det var det første året.

I **Figur 3.2.1** kan man se at forskjellen mellom ankerelever og de andre elevene øker sammen med ferdighetsnivået i populasjonen. For en metode uten bias, ville alle fordelingene vært sentret på 0. Her ser vi derimot at ankerelevene er nærmere -1 i 2021. Dette antyder at $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ i snitt underestimerer skalapoengene til ankerelevene, med i underkant av ett poeng (hvor standardavviket er cirka 10). Selv om effekten har høyest konsekvens for ankerelevene, er den også til stede for de andre elevene. Per definisjon vil $\boldsymbol{\varepsilon}$ for alle reestimeringsmetodene være sentrert på 0, altså

$$E[\boldsymbol{\varepsilon}] = 0$$

Dette betyr at en systematisk underestimering av $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ for ankerelevene vil føre til en tilsvarende overestimering av $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ for de øvrige elvene. Denne effekten er derimot sterkt påvirket av gruppe-størrelsene. Siden det er såpass få ankerelever (cirka 5 prosent av utvalget) vil effekten være langt tydeligere for ankerelevene, enn for de andre elevene. På en side kan det ses på som problematisk at de forskjellige reestimeringsmetodene gir systematiske feil, på både ankerelver og de andre elvene i utvalgene. Når dette er sagt er det viktig å huske at disse står i motsetning til hverandre, slik at $E[\boldsymbol{\varepsilon}] = 0$. Selv om man da vil forvente å ha systematiske feil på $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ for hver eneste elev i utvalget, vil disse systematiske feilene kanselleres når man aggregerer på gruppenivå. For enhver gruppe (mer enn ett individ) vil man ha et forventet feilledd på 0. Man vil likevel forvente større variasjon i feilleddet hos en mindre gruppe, sammenlignet med en større gruppe. Spørsmålet blir da hvordan variasjonen i $\boldsymbol{\varepsilon}$ påvirkes av størrelsen på en gitt gruppe.

Gruppestørrelse

For å teste effekten av gruppestørrelser på variasjonen i feilleddet ($\boldsymbol{\varepsilon}$), utførte vi en simulering hvor vi valgte ut tilfeldige grupper av elever, og kalkulerte snittet til $\boldsymbol{\varepsilon}$. Ved stor variasjon i snittet til $\boldsymbol{\varepsilon}$ for en gitt gruppestørrelse, vil det i snitt være en større usikkerhet vedvarende hvor godt $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ reflekterer $\boldsymbol{\theta}_{\boldsymbol{n}}$. Ved lav variasjon i snittet til $\boldsymbol{\varepsilon}$ for en gitt gruppestørrelse, vil det være derimot være høyere sikkerhet, og man vil kunne si at snittet til $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ er et godt estimat på snittet til $\boldsymbol{\theta}_{\boldsymbol{n}}$. La n representere gruppestørrelsen og R antall simuleringer. Snittet av feilleddet for en gruppe r kan da regnes ut som

$$\mu_{\varepsilon r} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_{ri}$$

ε_{ri} representerer da feilleddet for person i (av n) i simulering r (av R). Videre kan man regne ut standardavviket av $\mu_{\varepsilon r}$ på tvers av simuleringene. Siden man allerede vet den forventede verdien til $\mu_{\varepsilon r}$, kan man benytte seg av formelen for standardavviket for populasjonen (hvor man deler på n, ikke $n\ -1$)

$$\sigma_{\mu\varepsilon} = \sqrt{\sum_{r=1}^{R} \frac{(\mu_{\varepsilon r} - 0)^2}{n}}$$

$$\sigma_{\mu\varepsilon} = \sqrt{\sum_{r=1}^{R} \frac{{\mu_{\varepsilon r}}^2}{n}}$$

I **Tabell 3.2.2** og **3.2.3** kan vi se forskjellene i standardfeil i engelsk 8.trinn for 2015 og 2021, for de forskjellige reestimeringsmetodene. I **Tabell 3.2.4** viser vi standardavvikene for feilleddet til engelsk 8. trinn 2014-2021. Generelt sett ser man at den lokale omskaleringsmetoden har de laveste standardfeilene, etterfulgt av regresjonstilnærmingen, og til slutt den globale omskaleringsmetoden. Forskjellene mellom metodene blir større etter hvert som forskjellen mellom $\boldsymbol{\theta}_{\boldsymbol{n}}$ og $\boldsymbol{\theta}_{\boldsymbol{g}}$ øker. Vi ser

derfor veldig liten forskjell mellom metodene i 2015, og større forskjell i 2021. Videre blir forskjellen mellom metodene mindre tydelige, etter hvert som n (gruppestørrelsen) øker.

Tabell 3.2.2 Standardfeil på differansen mellom $\widehat{\theta}_n$ og θ_n for ulike gruppestørrelser n. Engelsk 8. trinn. 2015

ENG08 2015	regresjon		lokal omskalering		global omskalering	
n	μ	σ	μ	σ	μ	σ
5	0,000	0,030	0,000	0,029	0,000	0,030
50	0,000	0,010	0,000	0,009	0,000	0,010
500	0,000	0,003	0,000	0,003	0,000	0,003
5000	0,000	0,001	0,000	0,001	0,000	0,001

Talla er gjennomsnitt og standardfeil for 10 000 iterasjoner.

Tabell 3.2.3 Standardfeil på differansen mellom $\widehat{\theta}_n$ og θ_n for ulike gruppestørrelser n. Engelsk 8. trinn. 2021

ENG08 2021	regresjon		lokal omskalering		global omskalering	
n	μ	σ	μ	σ	μ	σ
5	0,000	0,103	0,001	0,092	0,001	0,110
50	0,000	0,033	0,000	0,029	0,000	0,035
500	0,000	0,010	0,000	0,009	0,000	0,011
5000	0,000	0,003	0,000	0,003	0,000	0,004

1. Tallene er gjennomsnitt og standardfeil for 10 000 iterasjoner.

Som følge av *'central limit theorem'* (CLT) burde fordelingen av $\mu_{\varepsilon r}$ approksimere en normalfordeling med kjente karakteristikker, etter hvert som n øker i størrelse. I teorien burde da det forventede standardavviket av $\mu_{\varepsilon r}$ være proporsjonalt til standardavviket til $\boldsymbol{\varepsilon}$ (σ_ε). Mer nøyaktig burde fordelingen av $\mu_{\varepsilon r}$ ha standard avvik ($\sigma_{\mu\varepsilon}$) som følger:

$$\sigma_{\mu\varepsilon} = \frac{\sigma_\varepsilon}{\sqrt{n}}$$

Gitt at dette stemmer, burde det være mulig å danne et estimat på den forventede variasjonen i feilleddet til en gruppe med gruppestørrelse n. Man vil da kun trenge en oversikt over standard-avviket til $\boldsymbol{\varepsilon}$ på den gitte prøven, det gitte året. Her ser vi eksempelvis en slik oversikt for engelsk 8. trinn.

Tabell 3.2.4 Standardavvik differanse mellom predikerte og reelle nye skalapoeng. ENG08. 2014-2021

		$\sigma(\epsilon)$		
Prøve	År	Regresjon	Lokal Omskalering	Global Omskalering
ENG08	2014	0,017	0,017	0,017
	2015	0,068	0,066	0,068
	2016	0,080	0,078	0,080
	2017	0,056	0,055	0,057
	2018	0,092	0,091	0,093
	2019	0,077	0,070	0,081
	2020	0,139	0,133	0,143
	2021	0,234	0,209	0,249

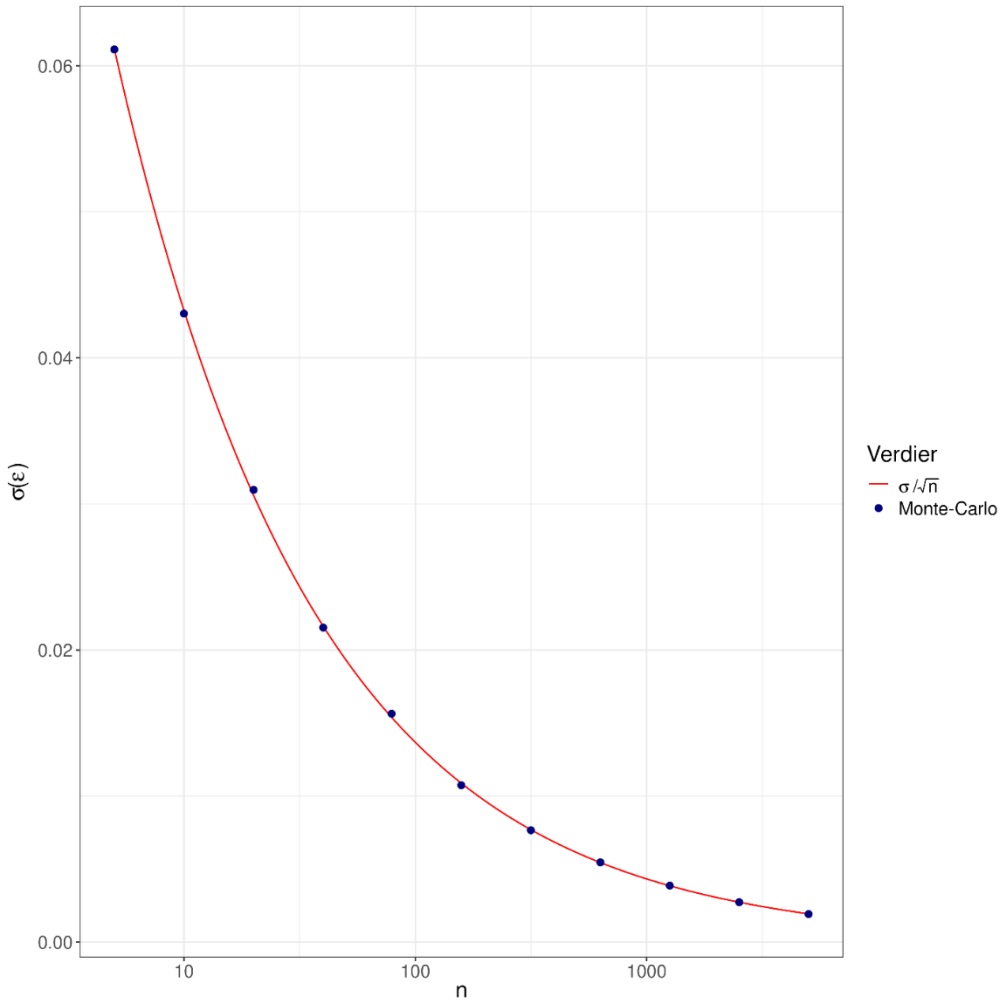
Notat: $\sigma(\epsilon)$ representerer standardavviket til differansen mellom estimerte og reelle nye skalapoeng, for de forskjellige metodene. Tall for alle prøver presentert i **Vedlegg C**.

For eksempel vil man forvente å i snitt bomme med $\pm .249/\sqrt{9} = \pm .083$ for en gruppe med størrelse 9, tatt i fra engelsk 8. trinn, 2021, hvor man har benyttet seg av den globale omskaleringsmetoden for å kalkulere $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$. Dette burde da ses i kontekst til at standardavviket til skalapoengene cirka ligger på 10. Dette tilsvarer da en feilmargin på $\pm .0083$.

Figur 3.2.5 viser den estimerte standardfeilen som en funksjon av gruppestørrelse (n), for regning 8. trinn 2021. De mørkeblå punktene er resultater fra våre simuleringer på de ulike gruppestørrelsene

(likt som rapportert i tabellene for engelsk 8. trinn 2015 og 2021). Den røde linjen viser den teoretiske forventingen til $\sigma_{\mu\varepsilon}$, som en funksjon av n ($\sigma_{\mu\varepsilon}/\sqrt{n}$). I **Figur 3.2.5** kan vi se at den observerte standardfeilen for μ_ε tett følger den teoretiske fordelingen, med kun minimale avvik (som teoretisk sett vil bli mindre, desto flere iterasjoner (R) vi har i simuleringen vår). Hvis man da senere må gjøre en vurdering av feilmarginen til et snitt på $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, skal det være tilstrekkelig å vite gruppe-størrelsen (n) og standardavviket til $\boldsymbol{\varepsilon}$. I **Vedlegg C** er det rapportert standardavvik for alle metodene for å estimere $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$, per prøve, per år. Man kan da bruke denne tabellen til å gjøre en vurdering av feilmargin, gitt n.

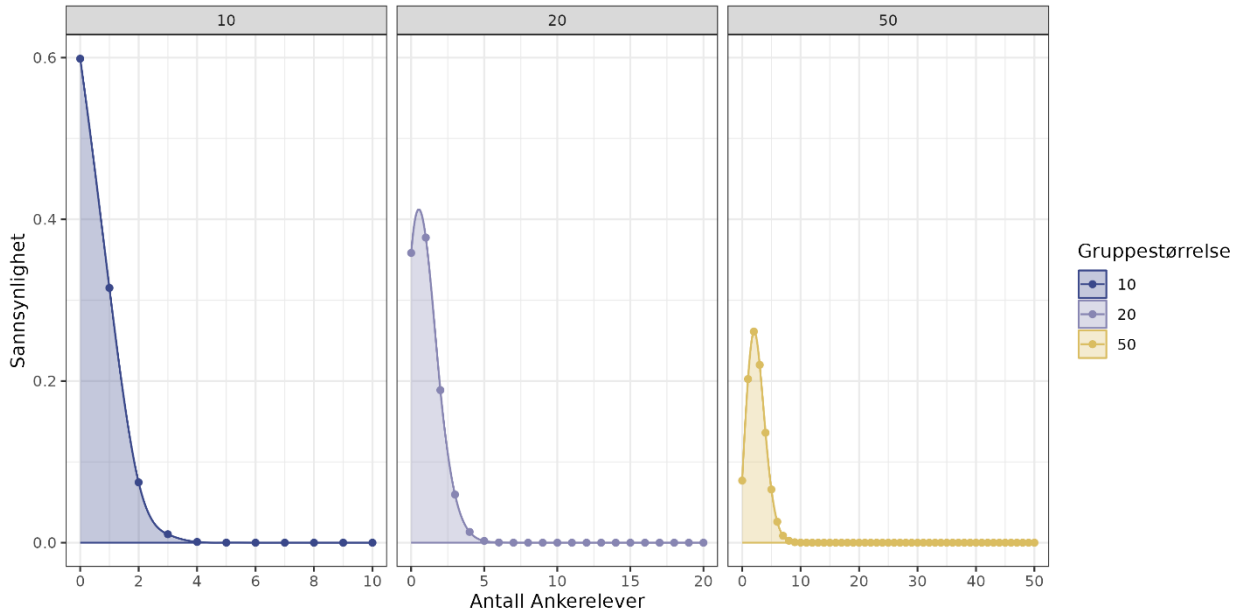
Figur 3.2.5 Estimert standardfeil for gruppestørrelse n. Regning 8. trinn. 2021



Forventet bias ved kjent gruppestørrelse

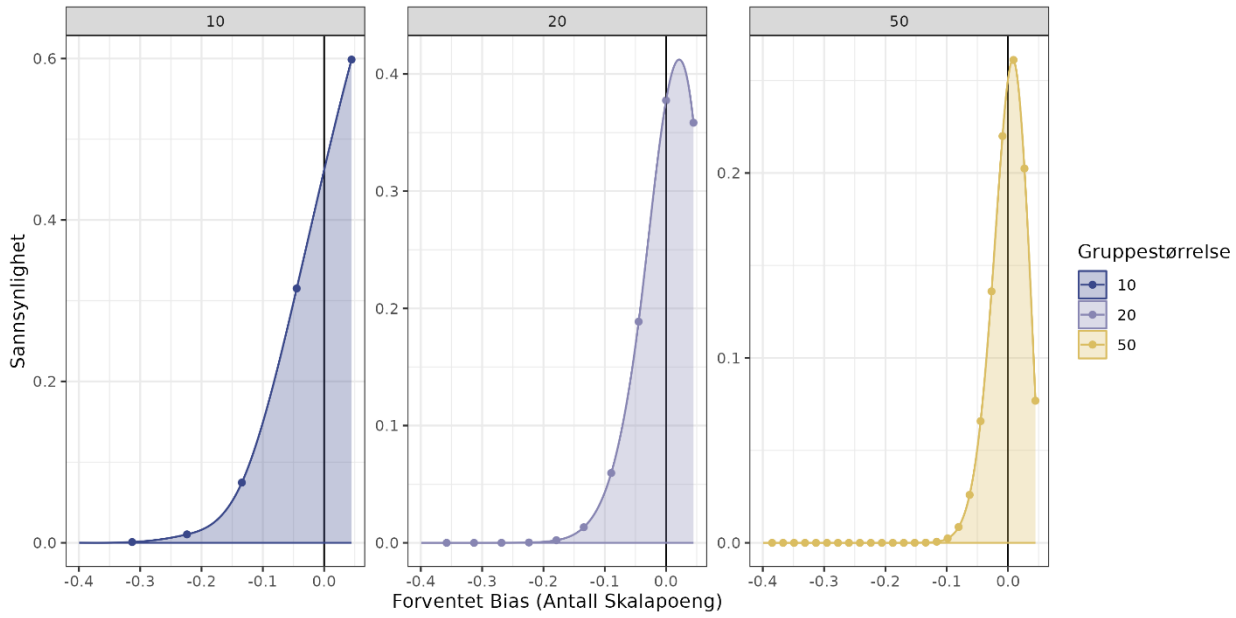
Selv om $\widehat{\boldsymbol{\theta}}_{\boldsymbol{n}}$ forventes å være et «unbiased» estimat på $\boldsymbol{\theta}_{\boldsymbol{n}}$, gjelder dette kun i tilfeller hvor ankerelevstatus er ukjent. Som nevnt har begge gruppene (ankerelever/øvrige elever) sin egen bias, som i snitt balanserer ut hverandre. Hvis man derimot har en gruppe hvor man vet det er en overrepresentasjon av en av gruppene, vil man ha et forventet bias. Det forventede biaset avledes da ut ifra fordelingen av ankerelever/øvrige elever, for en gitt gruppestørrelse. Generelt sett har andelen ankerelever ligget på cirka 5 prosent (cirka 3000 / 60 000). I **Figur 3.2.6** kan vi se sannsynlighetsfordelingen for å observere et gitt antall ankerelver, ved en gitt gruppestørrelse, under antakelsen om at andelen ankerelever ligger på 5 prosent. Dette kan enkelt regnes ut for hvilken som helst gruppestørrelse, ved bruk av den hypergeometriske fordelingen (Dodge, 2008).

Figur 3.2.6 Sannsynlighet antall ankerelever, etter gruppestørrelse. Engelsk 8. trinn. 2021



Siden man vet sannsynligheten for å observere en gitt fordeling av ankerelver/øvrige elever gitt gruppestørrelsen og det assosierte forventede biaset, kan det konstrueres en sannsynlighets-fordeling av forventet bias for en gitt gruppestørrelse. Det største biaset for ankerelvene lå på cirka -.85 for engelsk 8. trinn i 2021 (Se **Figur 3.2.1**). **Figur 3.2.7** viser sannsynlighetsfordelingen for det forventede biaset, gitt en bias for ankerelever på -.85, og en 5 prosent andel ankerelever.

Figur 3.2.7 Sannsynlighetsfordeling forventet bias, etter gruppestørrelse. Engelsk 8. trinn. 2021



Når man ser på fordelingen av det forventede biaset kan man se at sannsynligheten for store bias er lav. For eksempel ved en gruppestørrelse på 10, er det tilnærmet 0 prosent sannsynlighet for å få en bias under -0.3. Dette må da tas i betrakting til at standardavviket på skalapoeng ligger på cirka 10 poeng, noe som da tilsvarer en forventet bias på -0.03 standardavvik. Forenklet kan man bruke standardavvikene fra **Tabell C1** til å estimere standardfeil, og danne en sannsynlighetsfordeling

rundt det forventede biaset. Hvis man eksempelvis har en gruppe på 20 elever med 4 ankerelever fra engelsk 8. trinn vil man ha et forventet bias på litt under -0.1.

Sannsynlighetsfordeling av feilledd etter gruppestørrelse

Man kan deretter estimere forventet standardfeil, for en gitt gruppestørrelse, og fordeling av ankerelever. Den forventede standardfeilen (σ_v) kan estimeres ut ifra standardavvikene for $\boldsymbol{\varepsilon}$, til ankerelevene (σ_k) og de øvrige elevene (σ_n).

$$\sigma_v = \frac{\sqrt{k\sigma_k^2 + n\sigma_n^2}}{n+k}$$

Hvor k er antall ankerelever, n er antall øvrige elever i gruppen. For engelsk 8. trinn 2021 ved bruk av den globale omskaleringsmetoden var

$$\sigma_k = 0.240$$

$$\sigma_n = 0.019$$

Den forventede standardfeilen, for en gruppe med 4 ankerelever, og 16 øvrige elver gis da ved

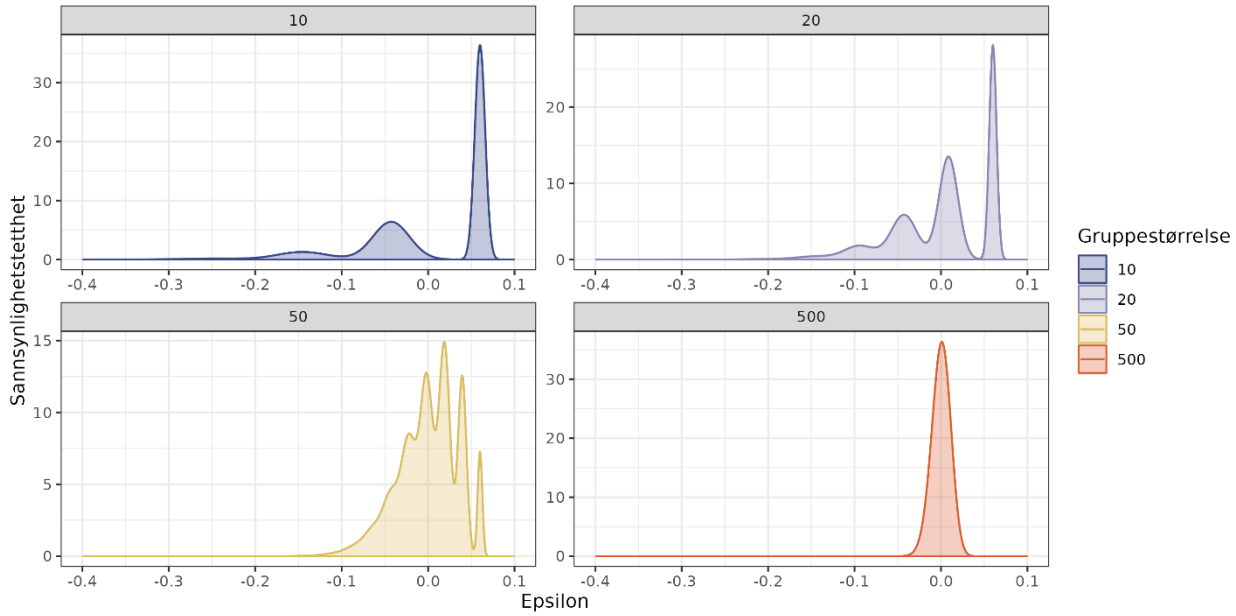
$$\sigma_v = \frac{\sqrt{4 \cdot 0.240^2 + 16 \cdot 0.019^2}}{20} = 0.024$$

Som nevnt burde fordelingen av bias approksimere normalfordelingen etter hvert som n øker, men den vil ikke være normalfordelt for små verdier av n. Fordelingen gis den vektede summen av sannsynligheten for å observere en gitt bias ved et gitt antall ankerelever, for alle ankerelever

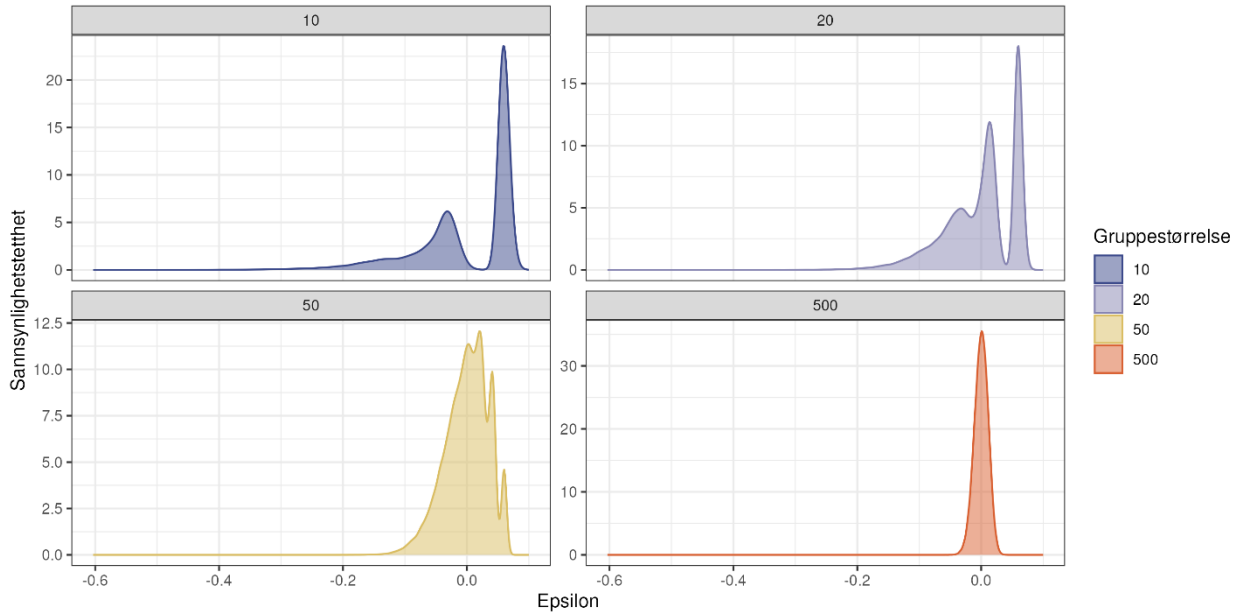
$$P(\varepsilon|N,K,n,k,b_k,b_n,\sigma_k,\sigma_k) = \sum_{k=0}^{s} \frac{\binom{K}{k}\binom{N-k}{n-k}}{\binom{N}{n}} N\left(\frac{kb_k + nb_n}{k+n}, \frac{\sqrt{k\sigma_k^2 + n\sigma_n^2}}{n+k}\right)$$

Hvor N er populasjonsstørrelsen, K er antall ankerelever i populasjonen, s er gruppestørrelse, k er antall ankerelever i gruppen, n er antall øvrige elever, b_k er biaset for ankerelever, b_n er biaset for de øvrige elvene, σ_k og σ_n er standardavviket av $\boldsymbol{\varepsilon}$ for ankerelevene og de øvrige elvene, samt $N(\mu,\sigma)$ er normalfordelingen. De teoretiske fordelingene ved forskjellige gruppestørrelser vises i **Figur 3.2.8.** I **Figur 3.2.9** vises de observerte fordelingene fra simuleringene benyttet i **Figur 3.2.5**

Figur 3.2.8 Teoretisk sannsynlighetsfordeling feilledd, etter gruppestørrelse. Engelsk 8. trinn. 2021



Figur 3.2.9 Observert sannsynlighetsfordeling feilledd, etter gruppestørrelse. Engelsk 8. trinn. 2021



4. Oppsummering og konklusjon

Funnene til Markussen et al., (2024) viser at metoden som ble tatt i bruk for å kunne sammenligne resultater fra nasjonale prøver over tid, ikke fungerte etter hensikten. Den originale estimeringen av skalapoengene førte til et uriktig bilde av elevenes ferdighetsutvikling i lesing, regning og engelsk. Etter en ny analyse med en mer hensiktsmessig metode fremkom det at endringene i ferdighetsnivå faktisk var større enn tidligere rapportert. Dette innebærer at den offisielle statistikken over nasjonale prøver har måttet vurderes på nytt, da særlig med tanke på trend. I forlengelsen av dette reises spørsmålet om hvordan tidligere publisert offisiell statistikk bør oppdateres, for å gi et mer nøyaktig og rettferdig bilde av utviklingen i norske grunnskoleelevers ferdigheter over tid.

Vi har presentert tre alternativer for å reestimere skalapoengene: den lokale- og den globale omskaleringsmetoden, samt regresjonstilnærmingen. Vi drøfter nå styrker og svakheter med de tre metodene før vi konkluderer med en anbefaling.

Den lokale omskaleringsmetoden prestert best på «*goodness of fit*» (GOF), men det er likevel bare små forskjeller mellom de tre metodene. Forskjellene er dermed ikke store nok, til at man kan velge metode basert på GOF alene. Videre blir forskjellene mellom metodene mindre, etter hvert som gruppestørrelsene man aggregerer på øker. SSB sin offisielle statistikk er publisert på gruppenivå hvor den minste gruppen er kommune. Dette betyr at man stort sett aggregerer på relativt store gruppenivå, og det vil da være tilnærmet ingen forskjell mellom metodene.

Selv om alle metodene introduserer systematiske feil på individnivå, noe som er spesielt tydelige for ankerelevene, er disse feilene ikke systematiske på aggregert nivå. Videre følger variasjonen i feilleddene en godt beskrevet teoretisk fordeling, avhengig av standardavviket til feilleddet som helhet, og gruppestørrelsen. Dette betyr at man lett kan gjøre en vurdering av hvor godt en reestimering av skalapoengene reflekterer de nye skalapoengene til Udir. Man trenger da kun å vite gruppestørrelsen, og standardavviket til feilleddet, for en gitt prøve og år. En slik oversikt er inkludert i **Vedlegg C**.

I regresjonstilnærmingen foreligger det en sterk antakelse om at forholdet mellom $\boldsymbol{\theta_n} \sim \widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$ er proporsjonalt til forholdet mellom $\boldsymbol{\theta_n} \sim \boldsymbol{\theta_g}$. Selv om det er god grunn til å tro at dette stemmer, vil brudd på antakelsen være problematisk for nøyaktigheten ved metoden. Dette vil i hovedsak avhenge av hvorvidt antakelsene til Markussen et al., (2024) om akkurat hva som gikk galt i `XCalibre` stemmer.

Ett problem med den lokale omskaleringsmetoden er at det er avvik i undergrupperingene i SSBs gamle data ($\boldsymbol{X_{yt}}$) og Udirs nye data ($\boldsymbol{Y_{yt}}$) (se **Kapittel 2.2**). I praksis er nok ikke dette et stort problem. Det er verdt å merke seg at dette er ikke et problem for den globale omskaleringsmetoden og regresjonstilnærmingen.

Gitt at det er relativt små forskjeller mellom de forskjellige metodene, mener vi at det er god grunn til å favorisere den globale omskaleringsmetoden, grunnet dens enkelhet, færre antakelser, samt at den ikke tilføyer ny støy i dataene. Metoden avhenger kun av 4 parametere ($\mu_{\theta g}, \mu_{\theta n}, \sigma_{\theta g}, \sigma_{\theta n}$) for et gitt år og en gitt prøve, og kan anvendes i mange ledd av databehandlingen. Dette betyr at man kan reestimere nye skalapoeng, selv om man ikke har tilgang til hele populasjonen for en gitt prøve et gitt år. For både regresjonstilnærmingen og den lokale omskaleringsmetoden trenger man ikke bare tilgang til kjønn og poengsum, men også tilgang til Udirs nye data. For regresjonstilnærmingen må man videre ha tilgang til et estimat på de gamle skalapoengene ($\widehat{\boldsymbol{\theta}}_{\boldsymbol{g}}$).

Regresjonstilnærmingen og den lokale omskaleringsmetoden vil i praksis bety at SSB må distribuere data på individnivå, med de reestimerte skalapoengene, for alle som har lyst på tilgang til de

reestimerte resultatene. Ved å benytte den globale omskaleringsmetoden, kan man derimot som f.eks. en ekstern forsker reestimere skalapoengene i sitt eget datasett, kun med tilgang til 4 verdier ($\mu_{\theta g}, \mu_{\theta n}, \sigma_{\theta g}, \sigma_{\theta n}$). Videre må ikke metoden anvendes på individnivå, men kan også anvendes på aggregert nivå. Gitt at skalapoengene er lineært (f.eks., gjennomsnitt og summering) aggregert innad samme prøve og år kan man anvende metoden likt på de aggregerte resultatene, og det tilføyes derav ikke noen nye feilkilder (støy) i dataene. Parameterne kan publiseres offentlig i en tabell, for hver prøve i hvert år.

Til slutt er den globale omskaleringsmetoden også metoden som er enklest å kommunisere ut til et bredt publikum, uten teknisk forståelse av metodikken bak estimeringen av skalapoengene i nasjonale prøver.

Referanser

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1 – 29.

Dodge, Y. (2008). Hypergeometric Distribution. In: The Concise Encyclopedia of Statistics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_182

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

Li, Y. H., Griffith, W. D. & Tam, H. P. (1997). *Equating Multiple Tests via an IRT Linking Design: Utilizing a Single Set of Anchor Items with Fixed Common Item Parameters during the Calibration Process.* Paper presented at the annual meeting of the Psychometric Society, Knoxville, TN. https://files.eric.ed.gov/fulltext/ED418130.pdf

Markussen, S., Ræder, H. G., Røgeberg, O. & Raaum, O. (2024). Skoleferdigheter i endring: Utviklingen over tid målt ved nasjonale prøver. *Acta Didactica Norden*, 18(1). Art. 1. https://doi.org./10.5617/adno.10310

R Core Team (2022). *R: A language and environment for statistical computing.* Vienna, Austria: R Foundation for Statistical Computing.

Udir (2024). *Feil i resultater fra nasjonale prøver 2014 – 2021*. Artikkel. Utdanningsdirektoratet. Versjon 18.03.2024 https://www.udir.no/tall-og-forskning/feil-i-resultater-fra-nasjonale-prover-20142021/

Udir (2024 - referat). *Oppsummering fagseminar om analyse av nasjonale prøver 2014 – 2021*. Referat. Utdanningsdirektoratet.

Kuo, T. C., & Sheng, Y. (2016). A comparison of estimation methods for a multi-unidimensional graded response IRT model. *Frontiers in Psychology*, *7*, 880.

Mehmetoglu, Mehmet & Venturini, Sergio. (2021). Structural Equation Modelling with Partial Least Squares Using Stata and R. 10.1201/9780429170362.

Vedlegg A: Ankeroppgaver med og uten DIF

```
# ENG08
anker_eng08_non_dif <- list(
`2014-2015` = c("3003637", "4820147", "4820410",  "5010960", "5018255",  "5018296", "5018297",
                "5018298", "5018299", "5018388", "5019182", "5020172", "5059939", "5064622"),
                # DIF: "5005485", "5020030", "5018256", "5009051", "5021083", "5010109",
`2015-2016` = c("4820147", "4820410", "5009051", "5010960", "5018255", "5018256", "5018296",
                "5018297", "5018298", "5018299", "5020172", "5059939", "5064622", "5074233",
                "5099276", "5099278", "5099287", "5099393"),
                # DIF: "3003637", "5099289", "5099283", "5018388",
`2016-2017` = c("3003637", "4820410", "5020172", "5043498",  "5064622", "5074233", "5099276",
                "5099278", "5099283", "5099287", "5099289", "5099393", "5108235", "5115043",
                "5120349") # DIF: "5059939", "5112895", "5010960",
`2017-2018` = c("3003637", "4820410",  "5020172", "5043498", "5059939", "5074233", "5099276",
                "5099278", "5099287", "5099289", "5099393", "5108235", "5115043", "5116045",
                "5129358", "5147253", "5148138"),
                # DIF: "5099283", "5112895", "5010960",
`2018-2019` = c("3003637", "5020172", "5059939", "5074233", "5099276", "5099278", "5099283",
                "5099287", "5099289", "5099393", "5116045", "5129358", "5147253", "5159083"),
                # DIF: "5043498", "4820410", "5115043", "5112895", "5108235", "5010960",
`2019-2020` = c("3003637",  "5020172", "5043498", "5074233", "5099393", "5112895", "5116045",
                "5129358", "5147253", "5159083", "5179332", "5179659"),
                # DIF: "4820410", "5115043", "5108235",
`2020-2021` = c("3003637", "5020172", "5043498", "5074233", "5099393", "5108041", "5108235",
                "5112895", "5116045", "5129358", "5147253", "5159083", "5179332","5179659",
                "5192499", "5192501", "5192505", "5192511", "5192520")
                # DIF:"5115043",
)

# ENG05
anker_eng05_non_dif <- list(
`2014-2015` = c("5018277", "5018280", "5018820", "5018857", "5018893", "5019452", "5019464",
                "5019982", "5019984", "5020486", "5020563", "5020599", "5020721"),
                # DIF: "5020723", "5064369", "5009011", "5019713", "5019980", "5019986",
                #      "5019988", "5019466", "5019460", "5019454", "5019474",
`2015-2016` = c("5009011", "5018277", "5018280", "5018893", "5019452", "5019454", "5019460",
                "5019464", "5019466", "5019474", "5019982", "5019984", "5019986", "5019988",
                "5020486", "5020563", "5020723", "5064369", "5074793"),
                # DIF: "5018820", "5098038", "5019980", "5097167",
`2016-2017` = c("5009011", "5018893", "5019980", "5019982", "5019984", "5019986",
                "5019988", "5020486", "5020563", "5064369", "5074793", "5109466"),
                # DIF: "5112170", "5020723", "5018820", "5098038", "5097167",
`2017-2018` = c("5018274", "5018893", "5019980", "5019984", "5019986", "5019988", "5020486",
                "5020563", "5097167", "5098038", "5109466", "5112170", "5145691", "5145715",
                "5147530", "5148574"),
                # DIF: "5064369", "5019982", "5145713", "5145707", "5145694",
`2018-2019` = c("5018274", "5019980", "5019982", "5019984", "5019986", "5064369", "5097167",
                "5109466", "5120043", "5145691", "5145694", "5145698", "5145707", "5145713",
                "5167995"),
                # DIF: "5167142", "5112170", "5147530", "5018893", "5098038", "5019988",
                #      "5145715", "5148574"
`2019-2020` = c("5019980", "5019982", "5019984", "5019986", "5019988", "5112170", "5120043",
                "5145691", "5145694", "5145698", "5145707", "5145713", "5145715", "5148574",
                "5167142", "5177109", "5180775", "5181049", "5181415"),
                # DIF: "5109466", "5167995",
`2020-2021` = c("5019980", "5019982", "5019984", "5019986", "5019988", "5109466", "5112170",
                "5120043", "5145691", "5145694", "5145698", "5145707", "5145713", "5145715",
                "5167142", "5177109", "5180775", "5181415", "5181467", "5186798", "5187867")
                # DIF: "5181049",
)

# LES05
anker_les05_non_dif <- list(
`2016-2017` = c("5113085", "5113089", "5113096", "5113098", "5117772", "5117775", "5117873",
                "5117877", "5117880", "5117884", "5117895", "5118182", "5118813", "5118815",
                "5118844", "5118848", "5119234", "5146328", "5146330", "5146697", "5146701",
                "5146703") # DIF: "5118139", "5118162", "5118175", "5117779", "5113087",
`2017-2018` = c("5117772", "5117775", "5117779", "5117873", "5117877", "5117880", "5117884",
                "5117895", "5118139", "5118162", "5118175", "5118182", "5118813", "5118815",
                "5118844", "5118848", "5146328", "5146330", "5146697", "5146701", "5146703",
                "5175908", "5175910", "5175912", "5175914", "5175916", "5175918"),
                # DIF: None
`2018-2019` = c("5117772", "5117775", "5117779", "5117873", "5117877", "5117880", "5117884",
```

```
                "5117895", "5118139", "5118162", "5118175", "5118182", "5146328", "5146697",
                "5146701", "5146703", "5175908", "5175910", "5175912", "5175914", "5175916",
                "5175918", "5185121", "5185123", "5185125", "5185127", "5185129"),
                # DIF: None
'2019-2020' = c("5117772", "5117775", "5117873", "5117877", "5117880", "5117884",
                "5117895", "5146697", "5146701", "5146703", "5175908", "5175910", "5175912",
                "5175914", "5175916", "5175918", "5185121", "5185123", "5185125", "5185127",
                "5185129", "5195241", "5195243", "5195245", "5195247", "5195248", "5195251"),
                # DIF: "5117779",
'2020-2021' = c("5117772", "5117775", "5117779", "5146701", "5146703", "5175908", "5175910",
                "5175912", "5175914", "5175916", "5175918", "5185121", "5185123", "5185125",
                "5185127", "5185129", "5195241", "5195243", "5195245", "5195247", "5195248",
                "5195251", "5199825", "5199827", "5199829", "5199831", "5199833", "5199835")
                # DIF: None
)

anker_les08_non_dif <- list(
'2016-2017' = c("5150154", "5150156", "5150158", "5150160", "5150162", "5150164",
                "5150166", "5150168", "5150170", "5150172", "5150174", "5150176", "5150178",
                "5150180", "5150182", "5150184", "5150186", "5150188", "5150190", "5150193",
                "5150195", "5150197", "5150199", "5150201", "5150203", "5150205", "5150207",
                "5150209", "5150211", "5150213", "5150215", "5150218", "5150222", "5150224",
                "5150226", "5150228", "5150230", "5150232", "5150235", "5150237", "5150239",
                "5150241", "5150243", "5150245"), # DIF: "5150220",
'2017-2018' = c("5150154", "5150156", "5150158", "5150160", "5150162", "5150164",
                "5150166", "5150168", "5150170", "5150172", "5150174", "5150176", "5150178",
                "5150180", "5150182", "5150184", "5150186", "5150188", "5150190", "5150193",
                "5150195", "5150197", "5150199", "5150201", "5150203", "5150205", "5150209",
                "5150211", "5150213", "5150215", "5150218", "5150220", "5150222", "5150224",
                "5150226", "5150228", "5150230", "5150232", "5150235", "5150237", "5150239",
                "5150241", "5150243", "5150245"), # DIF: "5150207",
'2018-2019' = c("5150154", "5150156", "5150158", "5150160", "5150162", "5150164",
                "5150166", "5150168", "5150170", "5150172", "5150174", "5150176", "5150178",
                "5150180", "5150182", "5150184", "5150186", "5150188", "5150190", "5150193",
                "5150195", "5150197", "5150199", "5150201", "5150203", "5150205", "5150207",
                "5150209", "5150211", "5150213", "5150215", "5150237", "5150239", "5150241",
                "5150243", "5150245", "5185299", "5185301", "5185303", "5185305", "5185307",
                "5185309", "5185311", "5185313"), # DIF: "5150235",
'2019-2020' = c("5150154", "5150156", "5150158", "5150160", "5150162", "5150164",
                "5150166", "5150168", "5150170", "5150172", "5150174",  "5150193", "5150195",
                "5150197", "5150199", "5150201", "5150203", "5150205", "5150207", "5150209",
                "5150211", "5150213", "5150215", "5150235", "5150237", "5150241", "5150243",
                "5150245", "5185299", "5185301", "5185303", "5185305", "5185307", "5185309",
                "5185311", "5185313", "5195182", "5195184", "5195186", "5195188", "5195190",
                "5195194"), # DIF: "5195192", "5150239", "5150176",
'2020-2021' = c("5150154", "5150156", "5150158", "5150160", "5150162", "5150164", "5150166",
                "5150168", "5150170", "5150172", "5150174", "5150176", "5150193", "5150195",
                "5150197", "5150199", "5150201", "5150203", "5150205", "5150207", "5150209",
                "5150211", "5150213", "5150215", "5185299", "5185301", "5185303", "5185305",
                "5185307", "5185309", "5185311", "5185313", "5195182", "5195184", "5195186",
                "5195188", "5195190", "5195192", "5195194", "5199840", "5199842", "5199846",
                "5199848", "5199850") # DIF: "5199844",
)

anker_reg05_non_dif = list(
'2014-2015' = c("5061574", "5061575", "5061576", "5061577", "5061578", "5061579",
                "5061580", "5061581", "5061582", "5061584", "5061585", "5061586", "5061587",
                "5061589", "5061590", "5061593"), # DIF: "5061591", "5061592",
'2015-2016' = c("5061574", "5061575", "5061576", "5061577", "5061578", "5061579",
                "5061581", "5061582", "5061584", "5061585", "5061587", "5061589", "5061590",
                "5061591", "5061592", "5061593", "5093468", "5115175"),
                # DIF: "5061586", "5061580",
'2016-2017' = c("5061574", "5061575", "5061579", "5061580", "5061582", "5061584",
                "5061585", "5061586", "5061587","5061590", "5061591", "5061592", "5061593",
                "5104916", "5115175"),
                # DIF: "5061577", "5093468", "5061578", "5061581", "5061589",
'2017-2018' = c("5061574", "5061575", "5061578", "5061579", "5061581", "5061582",
                "5061584", "5061585", "5061586", "5061589", "5061590", "5061592", "5061593",
                "5083327", "5093468", "5115175", "5126045", "5130957", "5134677"),
                # DIF: "5104916",
'2018-2019' = c("5061574", "5061575", "5061578", "5061582", "5061584", "5061592",
                "5083327", "5110248", "5115175", "5126045", "5130957", "5134677", "5148717",
                "5162017", "5163087"),
                # DIF: "5104916", "5093468", "5061581", "5061589", "5061585",
'2019-2020' = c("5061574", "5061578", "5061582", "5061584", "5083327", "5088906",
                "5104916", "5115175", "5126045", "5130957", "5133810", "5134677", "5148717",
                "5162017", "5163087", "5177969", "5180155"),
                # DIF: "5110248", "5061581", "5061592",
```

```
‘2020-2021‘ = c(“5061578”, “5061581”, “5061582”, “5061584”, “5083327”, “5088906”,
                “5115175”, “5126045”, “5130957”, “5133810”, “5134677”, “5148717”, “5162017”,
                “5163087”, “5177969”, “5180155”, “5186403”, “5186614”, “5190922”)
                # DIF: “5104916”,
)

anker_reg08_non_dif <- list(
‘2014-2016‘ = c(“5041670”, “5041687”, “5041693”, “5041701”, “5041704”, “5041705”,
                “5041736”, “5041775”, “5041798”, “5041809”, “5041826”, “5041829”, “5041845”),
                # DIF: “5041745”, “5041792”, “5041835”, “5041818”, “5041729”, “5041664”,
                #      “5041797”,
‘2015-2016‘ = c(“5041664”, “5041670”, “5041687”, “5041701”, “5041704”, “5041705”,
                “5041729”, “5041736”, “5041745”, “5041775”, “5041792”, “5041797”, “5041798”,
                “5041809”, “5041826”, “5041829”, “5041845”), # DIF: “5041693”, “5041835”,
‘2016-2017‘ = c(“5041664”, “5041687”, “5041693”, “5041701”, “5041704”, “5041705”,
                “5041729”, “5041792”, “5041797”, “5041798”, “5041809”, “5041826”, “5041829”,
                “5041835”, “5103456”, “5111635”),
                # DIF: “5041845”, “5041736”, “5111684”, “5041670”,
‘2017-2018‘ = c(“5041664”, “5041693”, “5041701”, “5041704”, “5041705”, “5041797”,
                “5041798”, “5041809”, “5041829”, “5041845”, “5103456”, “5111635”, “5111684”,
                “5121555”, “5129176”, “5132153”, “5132700”),
                # DIF: “5041792”, “5041826”, “5041687”,
‘2018-2019‘ = c(“5041664”, “5041693”, “5041705”, “5041792”, “5041826”, “5041829”,
                “5041845”, “5111635”, “5111684”, “5121555”, “5129176”, “5132153”, “5132700”,
                “5161723”, “5162535”, “5167723”),
                # DIF: “5041701”, “5103456”, “5041809”, “5161744”,
‘2019-2020‘ = c(“5041792”, “5041809”, “5041826”, “5041845”, “5103456”, “5111635”,
                “5121555”, “5129176”, “5132153”, “5161744”, “5162535”, “5167723”, “5176244”),
                # DIF: “5041705”, “5041701”, “5132700”, “5041693”, “5111684”, “5161723”,
                #      “5041829”,
‘2021-2021‘ = c(“5041693”, “5041701”, “5041829”, “5041845”, “5103456”, “5121555”,
                “5129176”, “5132153”, “5132700”, “5160397”, “5161723”, “5161744”, “5162535”,
                “5162567”, “5167723”, “5176244”, “5186213”, “5187369”)
                # DIF: «5041705», «5111684»,
)
```

Vedlegg B: Kode for FCIP lenke metoden

```
# mirtFCIP takes the following arguments:
#       y = a list containing datasets with results from a test for each year
#       anchors = a list of anchors to use: we used Udir's list of non-DIF tasks
#          reproduced in Vedlegg 1
#       fixed = fixes parameter estimates for mu and sigma to those from the first year when true:
#        this estimates the scores obtained using XCalibre
mirtFCIP <- function(y, anchors = anchors, fixed=TRUE) {
  ModelSout <- vector("list", length(y))

   # 1. Performs IRT analysis on reference year (first year in list
   mirtEstT1 <- mirt::mirt(y[[1]], 1, TOL = 0.00001, itemtype = 'gpcm',
                           technical = list(NCYCLES = 2000))

   # 2. Sets model parameters to first element in results list
   ModelSout[[1]] <- mirtEstT1

   # 3. For each year in list:
   for(k in 1:(length(y) - 1)) {
     # 4. Sets parameters from previous year.
     mirtPars <- mirt::mod2values(ModelSout[[k]])

     # 5. Parameter structure for IRT model fitted to current year.
     mirtParsTk <- mirt::mirt(y[[k + 1]], 1, TOL = 0.00001, itemtype = 'gpcm',
                              technical = list(NCYCLES = 2000), pars = 'values')
    # 6. Extract list of relevant anchor items Udir's list of all anchor items (see Appendix A)
    anchorsTk <- anchors[[k]]

    # 7. Fix values for anchors
    for(i in 1:length(anchorsTk)){
      mirtParsTk[mirtParsTk$item == anchorsTk[i] & mirtParsTk$name == 'a1', 'value'] <-
        mirtPars[mirtPars$item == anchorsTk[i] & mirtPars$name == 'a1', 'value']
      mirtParsTk[mirtParsTk$item == anchorsTk[i] & mirtParsTk$name == 'a1', 'est'] <- FALSE

      for(j in 1:(sum(mirtParsTk$item == anchorsTk[i] & mirtParsTk$name |>
                  stringr::str_detect('d')) - 1)){
        mirtParsTk[mirtParsTk$item == anchorsTk[i] &
                   mirtParsTk$name == stringr::str_c('d', j),  'value'] <-
          mirtPars[mirtPars$item == anchorsTk[i] &
                   mirtPars$name == stringr::str_c('d', j), 'value']

        mirtParsTk[mirtParsTk$item == anchorsTk[i] &
                   mirtParsTk$name == stringr::str_c('d', j), 'est'] <- FALSE
      }
    }

    mirtParsTk[mirtParsTk$name == "MEAN_1" | mirtParsTk$name == "COV_11", 'est'] <- !fixed
    # if fixed then FALSE

    # 8. Estimating single group model with non-DIF anchor item parameters fixed
    mirtEstTk <- mirt::mirt(y[[k + 1]], 1, TOL = 0.00001, itemtype = 'gpcm',
                            technical = list(NCYCLES = 2000), pars = mirtParsTk)

    # 9. Add results for year k to dataframe with results from all years
    ModelSout[[k + 1]] <- mirtEstTk
  }

  ModelSout
}
```

Vedlegg C: Differanse mellom predikerte og reelle nye skalapoeng

Tabell C1 Standardavvik differanse mellom predikerte og reelle nye skalapoeng. Alle prøver, 2014–2021

		$\sigma(\epsilon)$		
Prøve	År	Regresjon	Lokal Omskalering	Global Omskalering
ENG05	2014	0,036	0,035	0,036
	2015	0,136	0,133	0,137
	2016	0,061	0,059	0,062
	2017	0,113	0,111	0,113
	2018	0,222	0,222	0,222
	2019	0,130	0,128	0,132
	2020	0,193	0,191	0,193
	2021	0,358	0,335	0,360
ENG08	2014	0,017	0,017	0,017
	2015	0,068	0,066	0,068
	2016	0,080	0,078	0,080
	2017	0,056	0,055	0,057
	2018	0,092	0,091	0,093
	2019	0,077	0,070	0,081
	2020	0,139	0,133	0,143
	2021	0,234	0,209	0,249
LES05	2016	0,058	0,058	0,058
	2017	0,053	0,051	0,053
	2018	0,066	0,065	0,066
	2019	0,125	0,119	0,131
	2020	0,160	0,159	0,160
	2021	0,124	0,121	0,128
LES08	2016	0,044	0,044	0,044
	2017	0,043	0,043	0,043
	2018	0,056	0,055	0,056
	2019	0,056	0,055	0,056
	2020	0,086	0,085	0,087
	2021	0,090	0,089	0,091
REG05	2014	0,033	0,033	0,033
	2015	0,057	0,056	0,057
	2016	0,067	0,066	0,067
	2017	0,105	0,104	0,105
	2018	0,122	0,112	0,126
	2019	0,048	0,047	0,049
	2020	0,212	0,212	0,212
	2021	0,039	0,039	0,039
REG08	2014	0,020	0,020	0,020
	2015	0,113	0,107	0,115
	2016	0,077	0,075	0,078
	2017	0,081	0,081	0,082
	2018	0,079	0,079	0,079
	2019	0,036	0,036	0,037
	2020	0,114	0,114	0,114
	2021	0,092	0,092	0,092

Notat: $\sigma(\epsilon)$ representerer standardavviket til differansen mellom estimerte og reelle nye skalapoeng, for de forskjellige metodene.