Status for arbeidet med re-estimering av skalapoeng fra nasjonale prøver

**Om dette dokumentet:**

Arbeidet med å re-estimere skalapoengene fra nasjonale prøver har fått stor offentlig interesse. Vi ønsker derfor å ha en stor grad av åpenhet om arbeidet. I dette dokumentet vil vi forsøke å gi jevnlige statusoppdateringer. Vær klar over at Frischsenteret er en liten organisasjon uten dedikerte ressurser til informasjonsarbeid. Vi har heller ingen ekstern finansiering av dette arbeidet. Ressursbruken til informasjonsarbeid m.m. er derfor begrenset. For ytterligere informasjon kan prosjektgruppen kontaktes direkte:

Simen Markussen, Frischsenteret

Ole Jørgen Røgeberg, Frischsenteret

Oddbjørn Raaum, Frischsenteret

Henrik Galligani Ræder, CEMO UiO

**11.09.2023:**

*Status for forskningsartikkelen:* Artikkelen ble innsendt til vitenskapelig tidsskrift på våren 2023. Vi har mottatt kommentarer fra fagfeller og jobber nå med en revisjon. Vi tror den reviderte versjonen vil bli innsendt i løpet av September. Vi har vært nødt til å fjerne artikkelutkastet fra våre nettsider pga. opphavsrettslige regler med forlaget.

*Pågående arbeid med å lage korrigerte mikrodata for skalapoeng 2014-2021:* Våre resultater avviker til dels betydelig fra den offisielle statistikken. Data for skalapoeng benyttes både i statistikk og forskning. Vi jobber nå med å utrede mulighetene for å lage en korrigering av den offisielle statistikken. Vi trodde lenge dette kunne gjøres som en lineær transformasjon av de offisielle tallene. Per nå tror vi ikke det er tilfelle og en korrigering må gjøres på en mer omfattende måte.

Vi har vært i kontakt med leverandøren av analyseprogrammet Xcalibre (ASC) og har i dag mottatt to frie lisenser av programmet. Når vi får dette til å virke vil vi forsøke å replisere de offisielle tallene og dersom vi er i stand til å finne akkurat de innstillingene som gjenskaper de offisiell tallene eksakt vet vi også hvordan de opprinnelige estimeringene er gjort.

Vi har i dag derfor bedt Udir om å oversende alle Xcalibre-rapportene fra 2014-2021. Det vi vil ha er både kontrollfiler (tekstfiler som brukes under estimeringen - liste over alle items i et år og hvilke verdier som ble lagt inn for ankeroppgavenes parametre) og selve de XCalibre rapportene.

Dersom dette er vellykket tror vi at vi skal kunne lage en slags 1:1-nøkkel på individnivå (egentlig eksakt «skalapoengnivå») fra de offisielle tallene til de nye. Mulighetene for dette avhenger av at vi ikke får for mange dubletter i skalapoeng. Helt presist er vi avhengig av at det ikke finnes tilfeller hvor ankerelever og andre elever får samme poengskår. For å unngå dette trenger vi så mye informasjon som mulig om elevene. I det anonyme datasettet vi har mottatt per nå er det kun informasjon om elevenes kjønn. Vi har derfor i dag bedt Udir om så sjekke hvorvidt det finnes mer informasjon (så som skolekode). Helt presis har vi spurt Udir om følgende:

*«a) Dere er helt sikre på at det ikke lenger finnes oppgavedata med personnummer? Dersom dette finnes, og vi kan få tilgang en avidentifisert versjon av dette (evt. via SSB) så vil det være relativt enkelt å rette opp.*

*b) I filene vi har fått fra dere finnes kun kjønn. I filene fra SSB finnes også skolekommune m.m. Har dere oppgavedatasett med mer enn kjønn? Alt av detaljer er viktig her fordi vi får færre dubletter og dermed en mer presis "crosswalk".*

*c) Har dere datasett for beregnede skalapoeng med mer enn tre desimaler? Vi har fått bekreftet fra SSB at de kun har 3 desimaler i data fra dere, men kanskje dere har flere desimaler? Det vil hjelpe veldig.  Vi ser i de gamle Xcalibre-rapportene at parameteranslag for alle items i rapportene kun er oppgitt med tre desimaler, mens de bruker fire desimaler i kontrollfilene. Det kan bety at dere har fått ut estimatene fra XCalibre med fire desimaler via en annen output-rute enn rapportene.*

d) Worst case: Vet dere noe om hvorvidt sorteringen i filene vi har fått fra dere på oppgavenivå er den samme som i datasettene overlevert SSB? Dette siste er et halmstrå så vi håper vi kan krysspeile med den andre informasjonen.*»*

**12.09.2023:**

Udir har gitt tilbakemelding på at vi får oversendt kontrollfiler og rapporter fra Xcalibre.

Udir har også avklart at oppgavedata kun inneholder kjønn og ikke personnummer, skolekommune eller andre kjennetegn.

Udir har trolig ikke skalapoeng med mer enn tre desimaler, men skal dobbeltsjekke. De vil også undersøke om sorteringen av data kan være den samme i anonyme oppgavedata og individdata oversendt SSB.

Udir ba i dag om å få oversendt oppgaveparametere fra våre estimeringer i Mirt og har fått dette oversendt.

Vi er i gang med å estimere med Xcalibre, men foreløpig kun for 2021. Vi får da nøyaktig samme resultater som Udir.

**13.09.2023:**

Vi har i dag mottatt kontrollfiler for alle år fra Udir for engelsk og lesing.

Vi har også kommet et stykke videre i å forstå hva avvikene skyldes. Vi har laget et syntetisk datasett med ankeroppsett ala det som benyttes i nasjonale prøver. Gruppe 1 har sanne theta-verdier som er N(0,1), mens gruppe 2 har N(0.5, 1.2). Fordelen med denne tilnærmingen er at vi faktisk har en fasit.

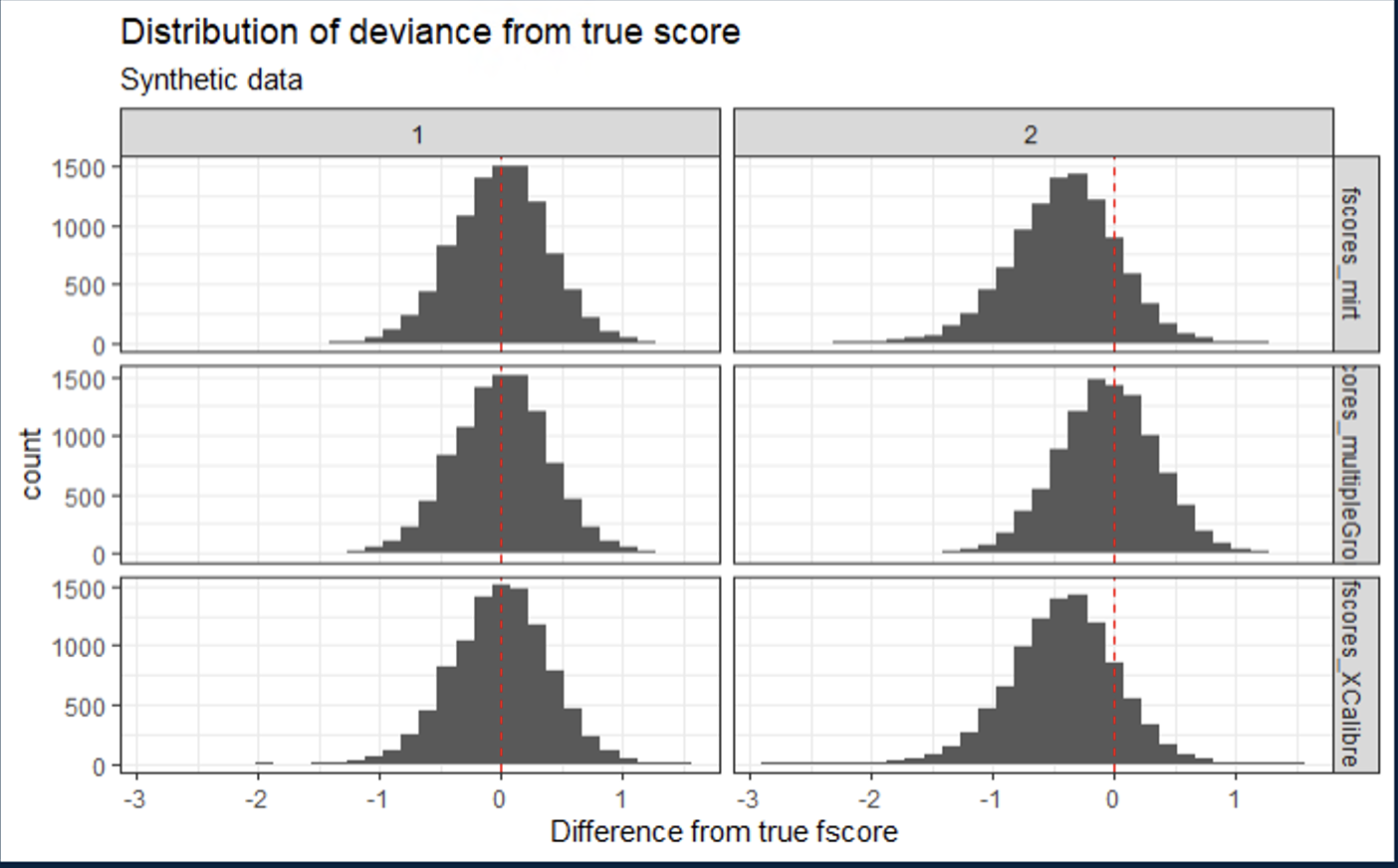
Vi estimerer så på tre måter:

1. mirt-funksjonen med FCIP

2. multipleGroup fra mirt-pakken

3. XCalibre (settinger av typen udir har brukt)

Resultatet illustreres ganske godt av figuren under, som viser fordelingen av avvik på elevnivå mellom anslått og sann theta (fscore) for gruppe 1 (i kolonne 1) og 2, for de tre ulike metodene. Både metode 1 og 3 har systematisk feil for gruppe 2.



**20.09.2023:**

Vi har på sett og vis jobbet langs fire spor:

A: Forskningsartikkelen. Denne er nå ganske snart klar for å sendes til tilbake til tidsskriftet.

B: Replisering av Udirs resultater: Så langt har vi kun mottatt XCalibre-rapporter for Regning, men for disse kan replisere Udirs resultater perfekt.

C: Simuleringsøvelse som tester de ulike metodene i en stilisert verden med en klar «fasit». Øvelsen omtalt over er nå mer eller mindre avsluttet. Vi er derimot svært mottakelige for tilbakemeldinger og innspill. Notatet finner du ved denne lenken: [pdf](https://www.dropbox.com/scl/fi/pxby9vx5g8fugvk4g3gxh/2023-09-19-Simulation-test.pdf?rlkey=59icb3i4n1nfcpjhakuhn9tll&dl=1)

For å kjøre koden bør man ha Rmd-dokumentet, som ligger [her](https://www.dropbox.com/scl/fi/ce2ze0dbxlw7b3jgmaql8/2023-09-19-Simulation-test.Rmd?rlkey=1mjgr1mf0gcmrfcongc8lriqd&dl=1). Notatet dokumenter ganske klart at Xcalibre, kjørt etter manualen ikke er i stand til å gjenfinne de reelle endringene over tid. For ankerelevene gir modellen en viss endring mens det for kohortelevene er veldig små endringer.

Vi tester Multigroup (gir riktig svar) mot to FPC-varianter, Xcalibre og R-mirt. Begge disse to siste gir samme, og feil, svar. Vi har i dag sendt notatet til Udir og til softwareselskapet bak Xcalibre.

D: Korrigerte skalapoeng for 2014-2021: Vi tror det skal være mulig å lage korrigerte skalapoeng men avventer noe hva vi gjør her.

**28.9.2023:**

Vi har nå oppdatert simuleringsnotatet etter å ha hatt en liten a-ha-opplevelse ang hva som er default-innstillinger i mirt-pakken i R. Her er en kort forklaring (les notatet under for en mer komplett beskrivelse samt kode):

I det forrige notatet var det på sett og vis uklart om forskjellen i estimater skyldes «samkalibreringsmetoden», altså den *simultane* (som i multi-group) eller *sekvensielle* (FPC, brukt både i mirt og Xcalibre). Fra notatet kunne man få inntrykk av det var slik sekvensiell samkalibering som ga feil svar. Vi har nå forstått dette bedre.

Om vi tar et lite skritt tilbake kan vi si at en vesentlige forskjell mellom metodene benyttet av oss og Udir er følgende: Vi gjør en simultan estimering, det vil si at vi estimerer alt på en gang, alle år. Dette er jo ikke i praksis mulig å gjøre for Udir siden resultatene skal ut årlig. De gjør derfor i stedet en sekvensiell estimering, såkalt FCP, hvor man først estimerer for det første året. Neste år låser man da parameterne for ankeroppgavene og estimerer resten.

For å bedre forstå hvorvidt denne sekvensielle samkalibreringen er årsaken til forskjellene i resultater gjennomfører vi også en slik sekvensiell estimering ved hjelp mirt-pakken i R. Her kan man da velge hva som skal estimeres og hva som skal være forutbestemt.

I "default" er det sånn at også populasjonsparametere (dvs. hvilken fordeling ferdighetene skal være trukket fra) er låst. De er da låst til N(0,1) som er standard. Om vi ikke endrer dette får vi ved mirt-pakken mer eller mindre nøyaktig det samme som man får med Xcalibre. Men, dette modelloppsettet er på sett og vis "overdeterminert" ved at både oppgaveparametere for ankeroppgaver og populasjonsparametere for ferdighetene er forhåndsbestemt. Dette blir i våre øyne helt feil. Det betyr jo i praksis at man først bestemmer seg for hva ferdighetene i år 2 skal være, for så å estimere dem. Det har også den konsekvens at ankerelevene blir riktigere estimert enn kohortelevene (se i notatet).

Når vi så, i denne mirt-pakken, løser opp denne forutsetningen og ber den om å estimere populasjonsparametere for ferdigheter fritt får vi samme resultat som med "multigroup", bare med større varians (som vi skal forvente da simultanestimering alltid er å foretrekke om det er mulig).

Så, for å oppsummere: Det viser seg å være to vesentlige forskjeller på metodene brukt av oss og Udir:

- måten lenkingen eller samkalibreringen foregår på: Dette er ikke årsaken til forskjellene i resultat.

- fri estimering eller forhåndsbestemte parametere for ferdighetsfordelingen: dette er årsaken.

Her er notatet.

<https://www.dropbox.com/scl/fi/uodgmb20miczpybvcpw1i/2023-09-25-Simulation-test.pdf?rlkey=hjplhs4mg8292e0a0ame441wg&dl=0>

Notatet er 27.9.23 sendt til ASC (selskapet bak Xcalibre), Udir og SSB.