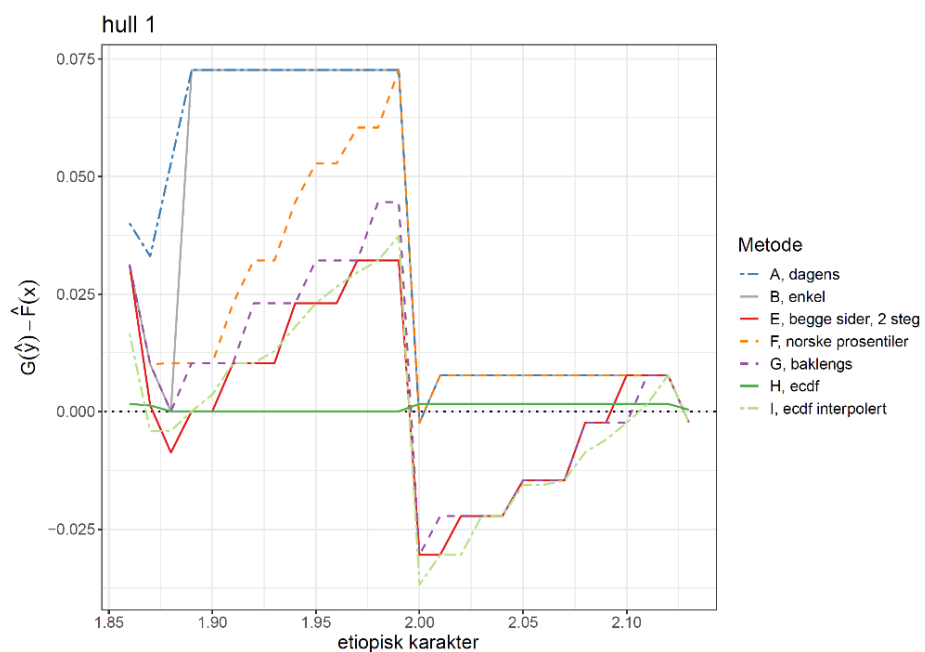


En diskusjon av noen statistiske metoder brukt for å lage omregningstabeller for utenlandske karakterer



Notatnr
Forfattere
Dato

SAMBA/32/22

Ingunn Fride Tvete og Céline Marie Løken Cunen

04.11.2022

Forfatterne

Ingunn Fride Tvette og Céline Marie Løken Cunen er seniorforskere ved Norsk Regnesentral.

Norsk Regnesentral

Norsk Regnesentral (NR) er en uavhengig, ideell og allmennyttig privat stiftelse som utfører oppdragsforskning. NR ble etablert i 1952 og har kontorer i Kristen Nygaards hus ved Universitetet i Oslo. Forskningsområdene er statistisk modellering, geomodellering, jordobservasjon, maskinlæring, bildeanalyse, kunstig intelligens, språkteknologi, digital sikkerhet, digital inkludering og digital transformasjon. NR er et av Europas største miljøer innen anvendt statistisk modellering. Instituttet leverer samfunnsnyttige tjenester og skaper nye verdier for kunder og partnere innen et vidt spekter av bransjer: Teknologi og industri, Finans og forsikring, Klima og miljø, Naturressurser, Hav, Helse og Samfunn. Våre oppdragsgivere er privat næringsliv og offentlige virksomheter både i Norge og internasjonalt, EU og Norges forskningsråd. Vår oppgave er å skape forskningsresultater som brukes og synes.

Direktoratet for høyere utdanning og kompetanse ved Avdeling for regelverk og opptak (Samordna opptak)

Arbeidet presentert i notatet er gjort på oppdrag fra Direktoratet for høyere utdanning og kompetanse ved Avdeling for regelverk og opptak (Samordna opptak), ved avdelingsleder for regelverk og opptak Rigmor Bjørkli.

Tittel

En diskusjon av noen statistiske metoder brukt for å lage omregningstabeller for utenlandske karakterer

Forfattere

Ingunn Fride Tvette og Céline Marie Løken Cunen

Dato

04.11.2022

År

2022

Publikasjonsnummer

SAMBA/32/22

Sammendrag

I 2014 vurderte Norsk Regnesentral (NR) metodene som lå til grunn for å oversette utenlandske karaktersnitt til norske, og dette ble oppsummert i notatet SAMBA/10/14. Notatet førte til noen endringer i metodene. Direktoratet for høyere utdanning og kompetanse ved Avdeling for regelverk og opptak (Samordna opptak) har i årene etter 2014 videreutviklet sine metoder.

Det er nå ønskelig å diskutere noen aspekter knyttet til de statistiske metodene som Samordna opptak i dag benytter i arbeidet med produksjon av omregningstabeller for utenlandske karakterer til norske. Norsk Regnesentral har undersøkt noen problemstillinger Samordna opptak ser på som utfordrende, og resultatene av dette arbeidet er presentert i notatet.

Vi tar for oss følgende spørsmål:

1. Er det endringer i karakterfordelinger over tid?
2. Hvordan håndtere «store hull» i det utenlandske tallmaterialet?
3. Hvordan håndtere tilfeller der noen utenlandske karaktersnitt er større enn største mulige verdi?
4. Hvordan håndtere statistikk fra land der mange får laveste eller høyeste karaktersnitt?
5. Hva gjør man når egeninnsamlet statistikk ikke omfatter alle karakterer i skalaen?

Emneord

Interpolering, karaktersnitt, omregningstabell og prosentiler

Målgruppe

Direktoratet for høyere utdanning og kompetanse ved Avdeling for regelverk og opptak (Samordna opptak)

Tilgjengelighet

Åpen

Prosjektnummer

221092

Satsningsfelt

Teknologi og industri

Antall sider

45

© Copyright

Norsk Regnesentral

Innhold

1	Bakgrunn	7
2	Data	8
3	Analyse 1: Er det endringer i karakterfordelingene over tid?	9
3.1	Deskriptiv analyse av de norske og franske karaktersnittene	9
3.2	Statistiske tester for å sammenligne empiriske fordelinger	12
3.3	Betydning av å legge ulike år med fransk statistikk til grunn	14
3.4	En alternativ måte å oversette karaktersnitt på	15
3.5	Antall år med statistikk og antall karaktersnitt	16
3.6	Oppsummering analyse 1	16
4	Analyse 2: Hvordan håndtere «store hull» i det utenlandske tallmaterialet?	17
4.1	Undersøke kalibrering	19
4.2	Undersøke differensiering	20
4.3	Metoder for å hankses med hull	21
4.4	Resultater for Etiopia	23
4.4.1	Kalibrering	25
4.4.2	Differensiering	29
4.4.3	Konklusjoner	30
4.5	Resultater for International Baccalaureate	30
4.6	Oppsummering av Analyse 2	33
5	Analyse 3: Hvordan håndtere tilfeller der utenlandske karaktersnitt er større enn største mulige verdi?	33
6	Analyse 4: Hvordan håndtere statistikk fra land der mange får laveste eller høyeste karaktersnitt?	35
6.1	Fem alternative metoder	36
6.2	Kalibrering	38
6.3	Oppsummering av analyse 4	39
7	Analyse 5: Hva gjør man når egeninnsamlet statistikk ikke omfatter alle karakterer i skalaen?	39
8	Oppsummering	43
8.1	Anbefalte metoder	44

8.2	Anbefalte rutiner	44
-----	-------------------------	----

1 Bakgrunn

Basert på norsk og utenlandsk karakterstatistikk lager Samordna opptak (SO) omregningstabeller for å oversette utenlandske elevers karaktersnitt til norske karaktersnitt. Hvis mulig lages disse tabellene med innhentet offisiell karakterstatistikk fra andre lands utdanningsmyndigheter. Er ikke dette mulig lages tabeller basert på egeninnsamlet statistikk fra vitnemål som søkere har sendt inn til norske universiteter og høyskoler. Har man heller ikke en slik egeninnsamlet statistikk av tilstrekkelig omfang lages lineære tabeller der en gjør en lineær omregning mellom norsk og utenlandsk karakterskala. Dette prosjektet tar for seg noen problemstillinger knytte til de to første typene omregningstabeller, og er kort oppsummert:

1. Er det endringer i karakterfordelinger over tid?
2. Hvordan håndtere «store hull» i det utenlandske tallmaterialet?
3. Hvordan håndtere tilfeller der noen utenlandske karaktersnitt er større enn største mulige verdi?
4. Hvordan håndtere statistikk fra land der mange får laveste eller høyeste karaktersnitt?
5. Hva gjør man når egeninnsamlet statistikk ikke omfatter alle karakterer i skalaen?

Analysene i punktene over vil ta utgangspunkt i karakterstatistikker for noen utvalgte land. Noen av funnene vil være spesifikke knyttet til det aktuelle landet, mens andre funn vil være av en mer generell art. Dette vil bli kommentert underveis i notatet.

Det er to prinsipper som bør ligge til grunn for god omregningsmetodikk. For å presist kunne beskrive disse to prinsippene må vi innføre litt matematisk notasjon, som også brukes videre i notatet. Vi legger til grunn samme notasjon som i notatet «Samordna opptak og omregningstabeller» (notat SAMBA/10/14).

La Y angi norske karaktersnitt og videre $g(y)$ angi Y s sannsynlighetstetthetsfunksjon og $G(y)$ Y s kumulative fordelingsfunksjon. Den kumulative fordelingsfunksjonen angir, for hvert eneste mulige karaktersnitt y , sannsynligheten for at en norsk elev får et karaktersnitt *lavere eller likt* y . Siden den norske karakterstatistikken er kjent og fullstendig kan $g(y)$ og $G(y)$ anses som *kjente*.

Tilsvarende lar vi X angi utenlandske karaktersnitt, med $f(x)$ og $F(x)$ som henholdsvis sannsynlighetstetthetsfunksjon og kumulativ fordelingsfunksjon. Når man har tilgang til landets offisielle karakterstatistikk kan $f(x)$ og $F(x)$ anses som kjente, men når SO kun har tilgang til egeninnsamlet statistikk så kan *ikke* $f(x)$ og $F(x)$ anses som kjente. Da *estimeres* tettheten med data tilgjengelig, og da brukes notasjonen $\hat{f}(x)$ og $\hat{F}(x)$ der «hatten» angir at dette ikke er den sanne tetthetsfunksjonen, men et estimat basert på et begrenset utvalg.

En omregning er en funksjon, eller prosedyre, $m(x)$ som tar et utenlandsk karaktersnitt x , og gir et omregnet norsk karaktersnitt $\hat{y} = m(x)$. Igjen brukes «hatten» fordi $\hat{y}$ kan

anses som et estimat basert på de dataene vi har tilgjengelig, og den omregningsmetodikken vi har valgt.

En omregning, eller omregningsmetode, regnes som god dersom den i størst mulig grad oppfyller to prinsipper:

1. Kalibrering: gitt antagelsen om at alle skolesystemer er like gode så ønsker vi at omregningen skal være slik at $F(x)$ skal være så lik som mulig $G(\hat{y})$, for alle mulige utenlandske karaktersnitt x :

$$F(x) \approx G(\hat{y})$$

Dette betyr for eksempel at dersom en elev har et snitt x_i som er slik at 70 % av landets elever har lavere eller likt snitt, dvs $F(x_i) = 0.70$, så ønsker man at den omregnede karakteren $\hat{y}_i$ skal være slik at 70 % av norske elever har et lavere eller likt snitt, dvs $G(\hat{y}_i) = 0.70$. Eller sagt på en annen måte, en omregningsmetode er godt kalibrert dersom den omregnede norske karakteren har omtrent lik rangering i Norge som karaktersnittet hadde i landet vi oversetter karakterer fra.

2. Differensiering: her ligger det for det første et minimumskrav om at omregningen må være slik at dersom vi har to utenlandske karaktersnitt x_i og x_j og $x_i < x_j$, så må $\hat{y}_i < \hat{y}_j$. I tillegg så ligger det et viktigere prinsipp om at forskjellige utenlandske karaktersnitt så langt det er mulig skal gi forskjellige omregnede norske karaktersnitt. Dette prinsippet er litt mindre viktig enn kalibrering og vi kan si at gitt et visst nivå av kalibrering, så er økt differensiering fordelaktig.

Dette er to prinsipper vi tar med i betraktning i våre analyser.

For å gjøre beregningene i notatet har vi benyttet statistikkprogrammet R.

2 Data

Datagrunnlaget for analysene omfatter offisiell (OS) og egeninnsamlet statistikk (ES) bestående av enkeltpersoners snittkarakter fra videregående utdanning fra ulike land:

- Norsk statistikk (OS, alle analyser)
- Fransk statistikk (OS, analyse 1 og 3)
- Etiopisk statistikk (ES, analyse 2)
- Statistikk fra International Baccalaureate (OS, analyse 2)
- Algerisk statistikk (ES, analyse 4)
- Russisk statistikk (ES, analyse 4)

- Pakistansk statistikk (ES, analyse 5)

All statistikk vil omfatte karaktersnitt fra elever som har bestått. Videre beskrivelse av statistikken gis i kapitlene der analysene beskrives.

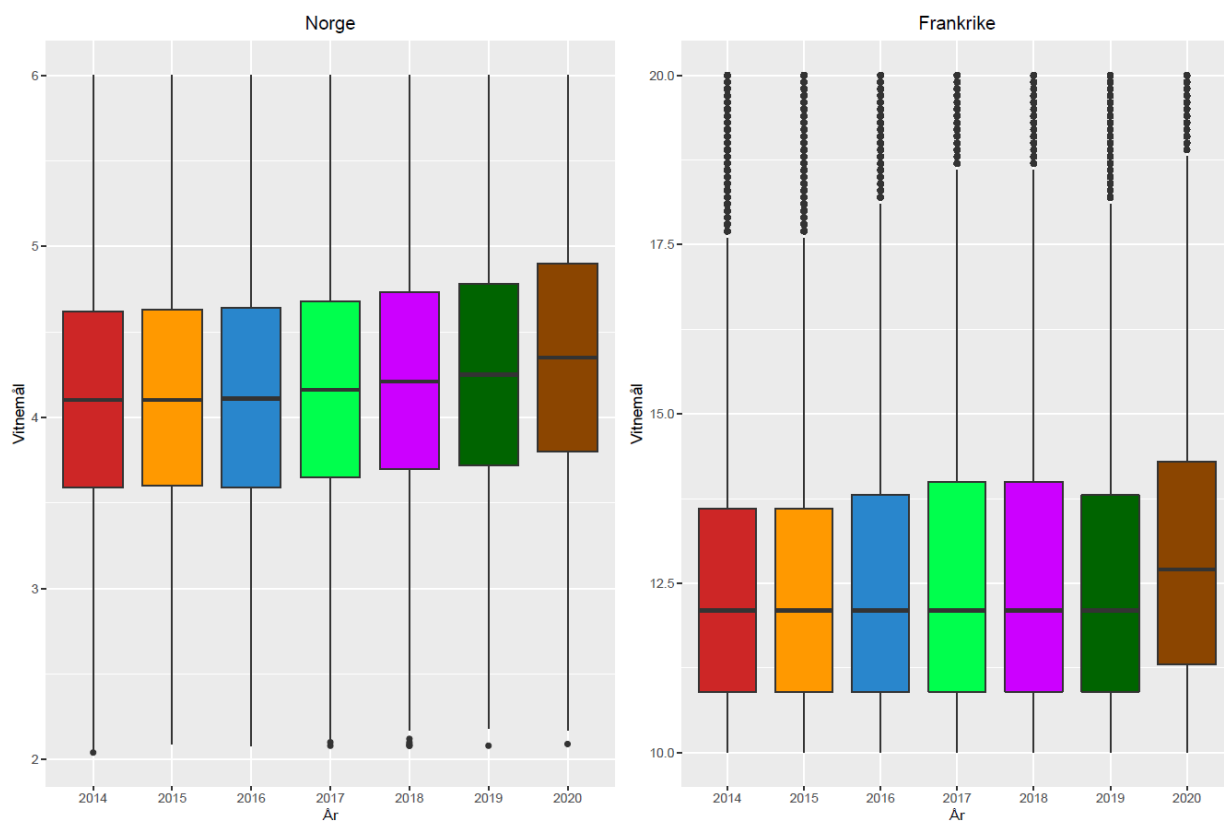
3 Analyse 1: Er det endringer i karakterfordelingene over tid?

Vi skal i dette kapitlet se på karakterfordelingen over tid, og hvorvidt det er tegn til endringer i fordelingen over år. Vi ser på den norske og franske statistikken, som begge er landenes offisielle statistikker. Lavest beståttkarakter i Norge er 2,00 og høyeste er 6,00, med 2 desimalers inndeling, mens tilsvarende tall for Frankrike er laveste beståttkarakter på 10,0 og høyeste på 20,0, men med 1 desimalers inndeling.

Selv om den franske karakterskalaen går til maksimalt 20,0 er det i den offisielle statistikken noen som har fått over 20,0 i karaktersnitt. Høyeste registrerte karaktersnitt er 21,3. I den franske statistikken oppgis gjerne høyeste karakterkategori som «20,0 eller mer». Vi vil i denne analysen behandle denne karakterkategorien som karaktersnitt 20,0. En særskilt analyse der vi ser på utfordringer knyttet til at noen elever har et karaktersnitt større enn maksimal verdi i karakterskalaen diskuteres i analyse 5.

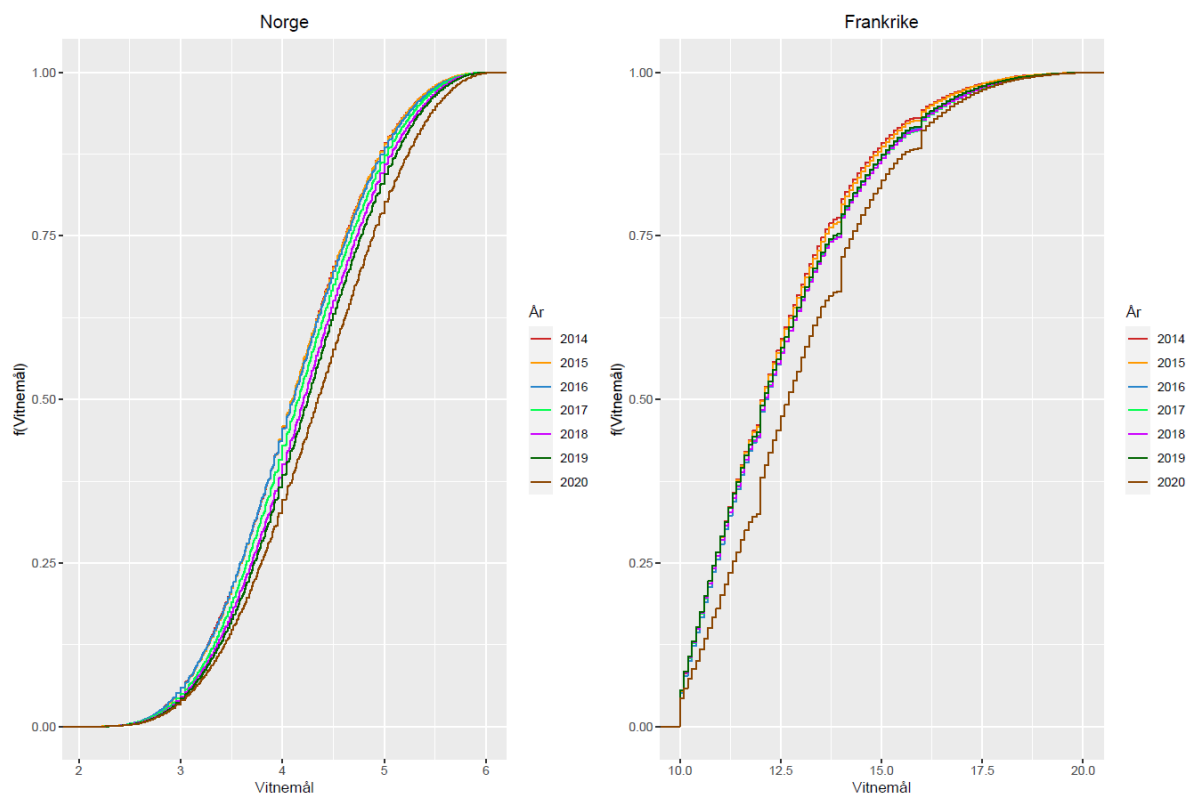
3.1 Deskriptiv analyse av de norske og franske karaktersnittene

Vi er interessert i hvordan karaktersnittene i de to landene fordeler seg over årene 2014 til 2020, og spesielt om dette endrer seg over tid. Figur 1 viser boksplott for gjennomsnittskarakterene for hvert år. Selve boksen i boksplottet starter ved første kvartil (Q1) og ender ved tredje kvartil (Q3) slik at boksen representerer 50 % av «midt-dataene», og streken i midten av boksen angir median-verdien. Strekene ut fra boksen angir avstand til minste og største verdi, når vi ser bort ifra karaktersnitt som skiller seg særskilt ut i verdi, også kalt «uteliggere» («outliers»). Hva som definerer en «uteligger» kan variere, men typisk (som her) er dette verdier som er større enn $Q3 + 1.5 \cdot (Q3 - Q1)$ eller mindre enn $Q1 - 1.5 \cdot (Q3 - Q1)$.



Figur 1. Boksplott basert på norsk statistikk (til venstre) og fransk statistikk (til høyre).

Den empiriske kumulative fordelingsfunksjonen (ecdf) viser andelen karaktersnitt som er mindre eller lik et gitt karaktersnitt (for største karaktersnitt blir denne andelen da 1,00), og disse fordelingene for karakterstatistikken for de ulike årene er vist i figur 2.



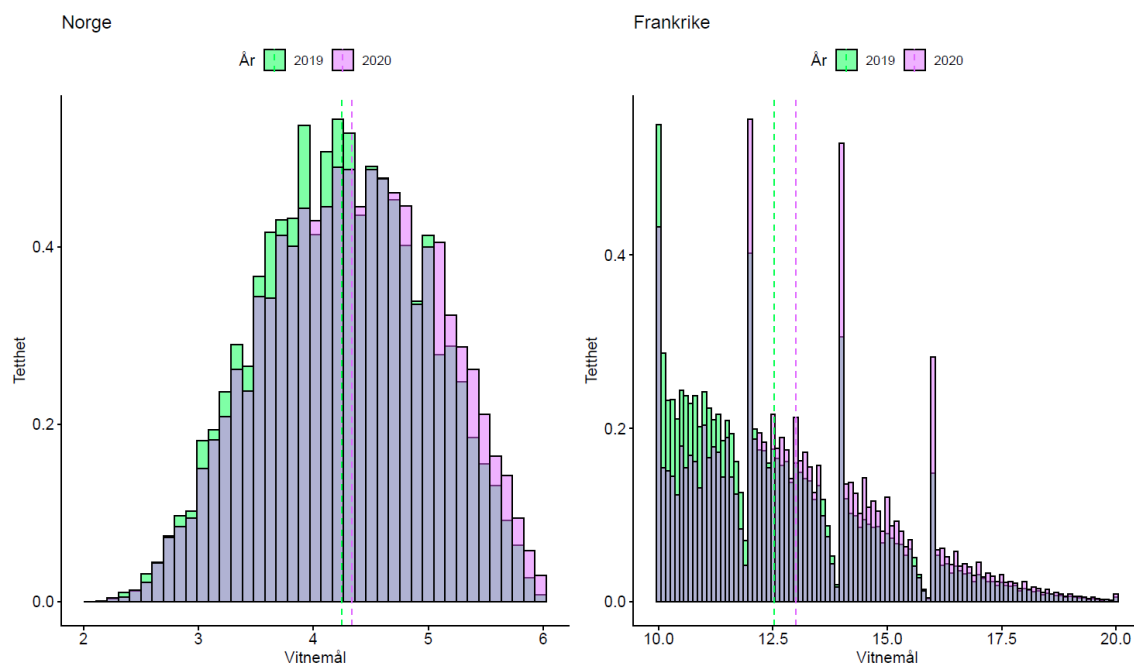
Figur 2. Empiriske kumulative fordelingsfunksjoner basert på norsk statistikk (til venstre) og fransk statistikk (til høyre).

Vi ser av figur 1 og 2 at den norske statistikken synes å ha nokså lik fordeling i 2014, 2015 og 2016, med en forskyving i karakterskalaen fra 2016 til 2017 og videre over de siste årene. Spesielt ser vi en endring fra 2019 til 2020. Den franske statistikken, derimot, synes å ha en mer lik fordeling i årene 2014 til 2019, men med en tydelig endring fra 2019 til 2020. Tabell 1 viser oppsummerende tall for karakterstatistikken for de ulike årene.

Tabell 1. Median, standardavvik (std), Q1 (25. prosentil) og Q3 (75. prosentil) for norsk og fransk statistikk over årene 2014–2020.

År	Norsk statistikk		Fransk statistikk	
	Median (std)	(Q1, Q3)	Median (std)	(Q1, Q3)
2014	4,10 (0,70)	(3,59 , 4,62)	12,1 (1,94)	(10,9 , 13,6)
2015	4,10 (0,70)	(3,60 , 4,63)	12,1 (1,96)	(10,9 , 13,6)
2016	4,11 (0,71)	(3,59 , 4,64)	12,1 (2,03)	(10,9 , 13,8)
2017	4,16 (0,70)	(3,65 , 4,68)	12,1 (2,06)	(10,9 , 14,0)
2018	4,21 (0,71)	(3,70 , 4,73)	12,1 (2,05)	(10,9 , 14,0)
2019	4,25 (0,71)	(3,72 , 4,78)	12,1 (2,03)	(10,9 , 13,8)
2020	4,35 (0,74)	(3,80 , 4,90)	12,7 (2,07)	(11,3 , 14,3)

Endring i karakterfordelingene fra 2019 til 2020 for begge land ser vi av histogrammene for de to karakterstatistikken i figur 3.



Figur 3. Histogrammer over norsk og fransk karakterstatistikk for 2019 og 2020. Den stiplede linja angir gjennomsnittet for de to karakterstatistikkene (2019 og 2020).

3.2 Statistiske tester for å sammenligne empiriske fordelinger

Det finnes ulike statistiske tester for å vurdere om det er signifikante forskjeller mellom to populasjoner (her: karakterstatistikker). Vi har valgt å se på en ofte brukt test for å sammenligne to populasjoner: Mann-Whitney-Wilcoxon testen (også kalt "Wilcoxon rangsumtest" eller "Wilcoxon tautvalgstest»), og undersøker om:

- H_0 (nullhypotesen): statistikk for to år har samme fordeling
- H_A (den alternative hypotesen): statistikk for et år er høyere (eller lavere) enn et annet

Dette er en såkalt ikke-parametrisk test (vi antar ikke at datene følger en kjent fordeling). Alle karaktersnittene rangeres fra lav til høy, og den såkalte p-verdien som beregnes avhenger av de gjennomsnittlige rangeringene i de to gruppene.

Vi sammenligner to og to karakterstatistikker, og dette gir da totalt 21 ulike tester for hvert av landene, og resultatene er oppsummert i tabell 2 og 3.

Tabell 2. Wilcoxon to utvalgs test, norske data, sammenligner karakterstatistikker for to ulike år (2014–2020). Røde ruter indikerer p-verdi over 0,01 og grønne ruter indikerer p-verdier under 0,01*.

	2015	2016	2017	2018	2019	2020
2014	0,7896	0,1738	<0,001	<0,001	<0,001	<0,001
2015		0,2704	<0,001	<0,001	<0,001	<0,001
2016			<0,001	<0,001	<0,001	<0,001
2017				<0,001	<0,001	<0,001
2018					<0,001	<0,001
2019						<0,001

*Svært lave p-verdier oppsummeres gjerne som <0,001

Tabell 3. Wilcoxon to utvalgs test, franske data, sammenligner karakterstatistikker for to ulike år (2014–2020). Røde ruter indikerer p-verdi over 0,01 og grønne ruter indikerer p-verdier under 0,01*.

	2015	2016	2017	2018	2019	2020
2014	<0,001	<0,001	<0,001	<0,001	<0,001	<0,001
2015		<0,001	<0,001	<0,001	<0,001	<0,001
2016			0,009	0,014	<0,001	<0,001
2017				0,882	<0,001	<0,001
2018					<0,001	<0,001
2019						<0,001

*Svært lave p-verdier oppsummeres gjerne som <0,001

Med lave p-verdier¹ ser vi at for de fleste år konkluderer testene med at karakterstatistikken er ulike over år, med noen unntak. Når vi snakker om signifikans og p-verdier er det viktig å merke seg at de påvirkes av hvor mange observasjoner (her: karaktersnitt) vi har i karakterstatistikken. Med karakterstatistikker med mange karaktersnitt kan selv veldig små forskjeller få en så lav p-verdi at de betegnes som statistisk signifikante. Det bør også påpekes at her gjøres flere hypotesetester, og vi har ikke tatt særskilt hensyn til at vi er i en såkalt «multippel hypotesetestingssituasjon». Enkelt forklart kan det sies at hvis vi gjør veldig mange hypotesetester vil det være en viss sannsynlighet for at en av testene slår ut feilaktig (såkalt «falsk positiv test»). Denne risikoen vil alltid være til stede når man gjør en hypotesetest, men med veldig mange tester kan denne risikoen bli uakseptabelt stor. Dette kan man adressere², men vil ha lite å si når p-verdiene er svært små. I den ene sammenligningen mellom fransk statistikk for 2016 og 2018 vil dette kunne ha noe å si.

Ønsker vi å teste om karakterstatistikken for **alle** årene for hvert av landene er like eller ikke kan vi gjøre en såkalt ANOVA-analyse der vi tester:

- H₀ (nullhypotesen): det er ikke forskjeller mellom gjennomsnittene i karakterstatistikken
- H_A (den alternative hypotesen): det er forskjeller mellom gjennomsnittene i karakterstatistikken

¹ p-verdien forteller oss om sannsynligheten for å observere det vi gjør dersom karakterstatistikken er like.

² Ved for eksempel å gjøre en såkalt «Bonferroni korreksjon».

Denne testen konkluderer med at det er forskjeller over år for begge landene da vi i begge testene får en p-verdi mindre enn 0,0001.

3.3 Betydning av å legge ulike år med fransk statistikk til grunn

Basert på funnene i disse analysene kan vi si at det har vært noen endringer i karakterfordelingene over tid, og spesielt gjelder dette fra 2019 til 2020. Dette innebærer at en fransk søker med eldre vitnemål med et nokså høyt karaktersnitt vil kunne få litt dårligere uttelling og likedan at en søker med eldre vitnemål med et nokså lavt karaktersnitt vil kunne få litt bedre uttelling hvis også fransk statistikk for 2020 legges til grunn for omregningstabellen. Dette kan vi studere i ulike scenarioer.

Vi har forstått det slik at norsk statistikk for de siste 10 årene som hovedregel legges til grunn når omregningstabeller lages. Vi vil ta utgangspunkt i dette, og se på tre ulike alternativer når vi lager omregningstabeller fra franske til norske snittkarakterer:

1. Vi legger fransk statistikk for 2014 til 2019 til grunn.
2. Vi legger fransk statistikk for 2014 til 2020 til grunn.
3. Vi legger kun fransk statistikk for 2020 til grunn.

Vi bruker den norske prosentiltabellen laget av SO som baserer seg på årene 2012-2021. Omregningstabeller for de tre scenarioene for de ti laveste og de ti høyeste prosentilene i tabellene er vist i tabell 4.

Vi ser av tabellen at legges fransk statistikk for årene 2014–2020 til grunn (Alternativ 2) vil søkere med fransk karaktersnitt på den laveste delen av skalaen kunne få oversatt et noe dårligere norsk karaktersnitt enn hvis statistikk for 2020 ikke er med (Alternativ 1), mens hvis vi legger kun statistikk for 2020 til grunn (Alternativ 3) vil søkeren få et enda dårligere norsk karaktersnitt. For søkere med karaktersnitt i den øvre delen av den franske karakterskalaen ser vi at et gitt norsk karaktersnitt svarer til et noe høyere fransk karaktersnitt når kun 2020-statistikken fra Frankrike legges til grunn sammenlignet med statistikk for 2014–2020, og videre at samme norske karaktersnitt vil kunne svare til enda et litt lavere fransk karaktersnitt hvis vi legger statistikk for 2014–2019 til grunn.

Tabell 4. Omregningstabeller; franske karakterintervall med tilhørende norsk oversatte karaktersnitt for de 10 laveste og 10 høyeste prosentiler der den norske prosentiltabellen er laget av SO, og baserer seg på årene 2012-2021.						
	Alternativ 1 (2014–2019)		Alternativ 2 (2014–2020)		Alternativ 3 (2020)	
prosentil	fransk	norsk	fransk	norsk	fransk	norsk
6	(10,0 – 10,0)	2,88	(10,0 – 10,0)	2,88	(10,0 – 10,0)	2,79
9	(10,1 - 10,1)	3,17	(10,1 - 10,1)	3,13	(10,1 - 10,1)	3,01
11	(10,2 - 10,2)	3,25	(10,2 - 10,2)	3,21	(10,2 - 10,2)	3,13
13	(10,3 - 10,3)	3,32	(10,3 - 10,3)	3,32	(10,3 - 10,3)	3,17
15	(10,4 - 10,4)	3,38	(10,4 - 10,4)	3,38	(10,4 - 10,4)	3,25
18	(10,5 - 10,5)	3,48	(10,5 - 10,5)	3,45	(10,5 - 10,5)	3,29
20	(10,6 - 10,6)	3,53	(10,6 - 10,6)	3,50	(10,6 - 10,6)	3,35
22	(10,7 - 10,7)	3,58	(10,7 - 10,7)	3,56	(10,7 - 10,7)	3,42
25	(10,8 - 10,8)	3,65	(10,8 - 10,8)	3,63	(10,8 - 10,8)	3,45
27	(10,9 - 10,9)	3,70	(10,9 - 10,9)	3,65	(10,9 - 10,9)	3,50

...						
92	(15,6-15,9)	5,21	(15,7-15,9)	5,21	(16,0-16,1)	5,21
93	-	-	(16,0-16,0)	5,25	(16,2-16,3)	5,25
94	(16,0-16,1)	5,30	(16,1-16,2)	5,30	(16,4-16,5)	5,30
95	(16,2-16,4)	5,35	(16,3-16,4)	5,35	(16,6-16,7)	5,35
96	(16,5-16,7)	5,40	(16,5-16,7)	5,40	(16,8-17,0)	5,40
97	(16,8-17,0)	5,46	(16,8-17,1)	5,46	(17,1-17,3)	5,46
98	(17,1-17,5)	5,54	(17,2-17,5)	5,54	(17,4-17,7)	5,54
99	(17,6-18,1)	5,62	(17,6-18,2)	5,62	(17,8-18,4)	5,62
100	(18,2-19,9)	5,74	(18,3-19,9)	5,74	(18,5-19,9)	5,74
101	(20,0-20,0)	6,00	(20,0-20,0)	6,00	(20,0-20,0)	6,00

3.4 En alternativ måte å oversette karaktersnitt på

Et alternativ til denne måten å oversette utenlandske karaktersnitt på kan være å sammenligne den empiriske kumulative fordelingsfunksjonen (ecdf) for de to landenes statistikk og så koble norsk og utenlandsk karaktersnitt ved å finne ecdf-ene som er mest like (eventuelt velge norsk karaktersnitt som har ecdf mest mulig lik, men eventuelt større enn utenlandsk ecdf). Som ellers kan vi gjøre unntak i «bunnen av skalaen» hvis vi har et «hull» større enn 5 prosentiler. Med ecdf-metoden vil man da få koblet hvert franske karaktersnitt med et norsk karaktersnitt istedenfor slik som nå når et intervall av franske karaktersnitt svarer til et norsk karaktersnitt. Man vil altså få en større grad av differensiering i omregningen. Dette er vist som et eksempel i tabell 5.

Dette kan være en alternativ fremgangsmåte hvis man har offentlig statistikk fra andre land tilgjengelig og ønsker at ulike utenlandske karaktersnitt i størst mulig grad skal kobles mot et norsk karaktersnitt³, og ikke at et intervall av utenlandske karaktersnitt skal svare til et enkelt norsk karaktersnitt. Her har vi for enkelhets skyld brukt en prosentilskala som går opp til 100 og ikke 101 (slik SO gjorde tidligere).

Tabell 5. Omregningstabeller; franske og norske empiriske kumulative fordelingsfunksjoner (ecdf), avrundet til 4 desimaler, med tilhørende karaktersnitt*.

fransk snitt	ecdf, Frankrike	norsk snitt	ecdf, Norge
17,6	98,0586	5,62	98,0056
17,7	98,2450	5,64	98,2517
17,8	98,4243	5,66	98,3785
17,9	98,5537	5,67	98,5235
18,0	98,7184	5,69	98,6920
18,1	98,8565	5,71	98,8815
18,2	98,9828	5,73	98,9831
18,3	99,1046	5,75	99,1487
18,4	99,1929	5,76	99,2276
18,5	99,2974	5,78	99,2906
18,6	99,3847	5,79	99,3663
18,7	99,4652	5,81	99,4857

³ Er statistikken liten kan man i prinsippet også her ende opp med intervaller, men med offisiell statistikk er dette lite sannsynlig.

18,8	99,5404	5,83	99,5570
18,9	99,5906	5,84	99,6124
19,0	99,6553	5,85	99,6596
19,1	99,709	5,87	99,7018
19,2	99,7552	5,88	99,7592
19,3	99,7965	5,89	99,7858
19,4	99,8249	5,90	99,8256
19,5	99,8586	5,92	99,8718
19,6	99,8854	5,94	99,883
19,7	99,9068	5,95	99,9289
19,8	99,9275	5,95	99,9289
19,9	99,9403	5,95	99,9289
20,0	100,000	6,00	100,000
*Norsk statistikk med 409 493 karaktersnitt, fransk statistikk fra 2014 - 2020			

3.5 Antall år med statistikk og antall karaktersnitt

Vi har i analyse 1 sett at det å inkludere karakterstatistikk fra ulike år kan ha betydning for hvilket norsk karaktersnitt et utenlandsk karaktersnitt tilordnes. Dette gjelder spesielt når karakterfordelingen endrer seg fra et år til et annet. Dette er relatert til problemstillinger som ble diskutert i notatet fra 2014⁴, der vi så på hvordan standardfeilen som funksjon av karaktersnittet varierte med antallet karaktersnitt vi la til grunn, og der vi kunne vise hvordan standardfeilen gikk ned når antallet karaktersnitt økte.

Det er flere hensyn å ta når karaktersnitt skal oversettes. Først må vi vurdere hvor store endringer i karakterfordelingen over år som kan være akseptabelt. Det finnes ulike mål på hvor forskjellige to (empiriske) fordelinger kan være, for eksempel Kullback–Leibler divergens, men slike mål er ikke lett anvendbare i praksis. Det vil være fornuftig å følge med på karakterfordelingene fra år til år og beregne oversettelsestabeller med og uten siste år for å se på betydningen av eventuelle endringer i karakterfordelingene.

I utgangspunktet er det ønskelig å ha et stort datagrunnlag når oversettelsestabeller skal lages. Den årlige offisielle statistikken for Frankrike omfatter mange vitnemål, mens for andre land vil det å kun se på siste år innebære at man får et betydelig mindre datagrunnlag, og spesielt vil dette få betydning hvis antallet blir så lite at man må ty til lineær omregning av karaktersnittene.

3.6 Oppsummering analyse 1

Vi har sett på fransk og norsk karakterstatistikk for årene 2014–2020 og beregnet oversettelsestabeller for franske til norske karaktersnitt. Vi fant at:

- Karakterfordelingene i Norge var nokså like i 2014, 2015 og 2016, med en forskyving i skalaen fra 2016 til 2017 og videre i 2018 og 2019. Spesielt ser vi en endring fra 2019 til 2020.

⁴ Tvete og Løland, Samordna opptak og Omregningstabeller, SAMBA/10/14

- Karakterfordelingene i Frankrike endret seg ikke mye fra 2014 til 2019, men med det er en tydelig endring fra 2019 til 2020.
- Å inkludere den franske 2020-statistikken hadde noe betydning for hvordan franske karaktersnitt ble oversatt til norske.
- Vi kan også oversette utenlandske karaktersnitt ved å sammenligne den empiriske kumulative fordelingsfunksjonen (ecdf) til to lands statistikker (mer) direkte.
- Det kan være nyttig å rutinemessig følge med på lands statistikker over tid og gjøre noen scenarioanalyser for hva det betyr å inkludere eller ikke inkludere et års statistikk når man beregner oversettelsestabellene.
- Er det endringer i karakterfordelingene over tid bør man vurdere hvorvidt søkere med eldre vitnemål bør plasseres på den norske karakterskalaen etter en særskilt (eldre) omregningstabell.

4 Analyse 2: Hvordan håndtere «store hull» i det utenlandske tallmaterialet?

Nåværende omregningsmetodikk krever at man håndterer «store hull» i hundreprosentiltabellene basert på utenlandske karakterstatistikker. Hullene er hundreprosentiler som ikke har fått tilegnet en karakter, som vist i ett utsnitt av hundreprosentiltabellen for Etiopia i tabell 6. Der har vi et stort hull for hundreprosentilene 50 til 54, og et lite hull for 56 og 57. Merk forresten at Etiopia har et karaktersystem som går fra 1 til 4 (best), og der karaktersnittene rapporteres med 2 desimaler.

Etiopia		
Hundreprosentil	Minimums karakter	Maksimums karakter
48	2,61	2,63
49	2,64	2,70
50	-	-
51	-	-
52	-	-
53	-	-
54	-	-
55	2,71	2,71
56	-	-
57	-	-
58	2,72	2,80

Hva er egentlig «hull» og hvor kommer de fra? Blant de 200 etiopiske karaktersnittene tabellen er basert på er det 97 elever som har karaktersnitt 2,70 eller lavere, 48,5 %. Dermed blir den øvre grensen for 49-hundreprosentilen 2,70. Om vi går til neste mulige

karaktersnitt 2,71 viser det seg at det er hele 12 elever blant de 200 som har dette snittet. Det er dermed 109 personer som har dette snittet eller lavere, 54,5 %. Dermed er vi kommet til 55-hundreprosentilen. Hull kan altså skyldes *opphopninger* eller *ujevnheter* i datamateriale tabellen er basert på. I eksempelet over var det en tydelig opphopning på det etiopiske karaktersnittet 2,71. Jo større datamateriale hundreprosentil bygges på, jo lavere sannsynlighet er det for hull (særlig store hull)⁵. **Store hull bør derfor hovedsakelig opptre i hundreprosentiltabeller basert på egeninnsamlet statistikk.** Unntaket er hull i bunnen eller toppen av hundreprosentiltabellen som nok kan opptre selv i hundreprosentiltabeller basert på offisiell statistikk. I denne analysen tar vi ikke for oss hull i bunnen og toppen av hundreprosentiltabellen, men begrenser oss til hull som opptrer midt i tabellen.

Et annet unntak gjelder karaktersystemer med *veldig grove skalaer*. Der vil hull i prosentiltabellen forekomme når det er færre mulige karakterer enn 100 (altså færre mulige karakterer enn antall hundreprosentiler). Dette gjelder for eksempel karaktersystemet til International Baccalaureate (IB) der det kun er 22 mulige poengsummer/karaktersnitt (eventuelt 26, hvis man tar med poengsummer som er teoretisk mulige å få, men som ikke figurerer i den offisielle karakterstatistikken). Huller som skyldes grove skalaer er egentlig et annet fenomen enn huller som skyldes opphopninger i statistikken, og må egentlig ikke håndteres på noen bestemt måte. Vi kommer tilbake til dette på slutten av dette kapitlet.

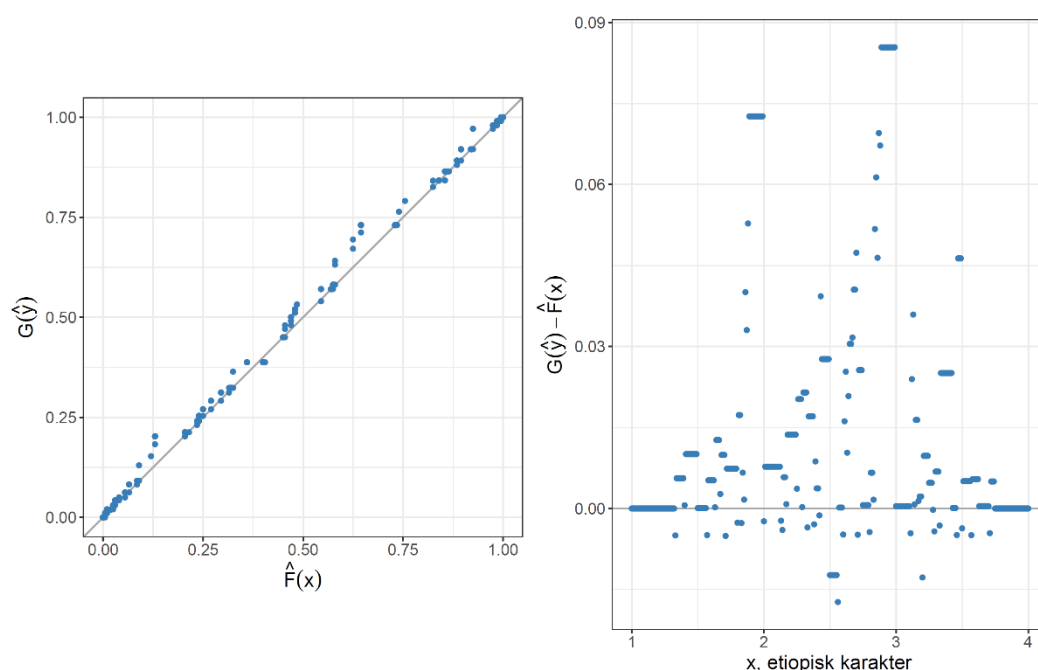
I dette kapitlet skal vi bruke omregningen av etiopiske karaktersnitt som illustrasjon, for å vise forskjeller og likheter mellom ulike metoder for å håndtere hull. I tillegg kommer vi tilbake til IB omregningen på slutten. Vi starter med to delkapitler der vi diskuterer hvordan vi skal sammenligne og vurderer ulike omregningsmetoder. Deretter beskriver vi 9 ulike metoder for å hankses med hull, før vi anvender metodene på Etiopia-omregningen og vurderer resultatene.

Som nevnt innledningsvis er det to prinsipper som bør ligge til grunn for en god omregningsmetodikk, og vi vil videre i kapittel 4 bruke notasjonen introdusert i kapittel 1. Vi skal diskutere hvordan figurer og scorer kan brukes til å vurdere hvor godt en bestemt omregning oppfyller et gitt prinsipp, og videre sammenligne ulike metoder med tanke på kalibrering og differensiering. Dette kan man se på ved å studere ulike figurer, eller ved å definere noen scorer som måler graden av for eksempel differensiering.

⁵ Fordi jo større datamateriale (= karakterstatistikk) jo større må opphopningene være for at de skal gi utslag i store hull.

4.1 Undersøke kalibrering

For å vurdere graden av kalibreringen kommer vi til å se på figurer som figur 4 under.



Figur 4. Eksempel på kalibreringsplott.

Den horisontale aksen i plottet til venstre viser $\hat{F}(x)$ for alle mulige etiopiske karaktersnitt x . Hvert punkt i plottet tilsvarer et etiopisk karaktersnitt der vi ser rangeringen i Etiopia langs den horisontale aksen mot rangeringen i Norge for den tilsvarende omregnede karakteren langs den vertikale aksen. Dersom en omregningsmetode er perfekt kalibrert, vil alle punktene ligge på den diagonale grå linjen. Avvik fra den diagonale linjen forteller oss noe om for hvilke karaktersnitt omregningsmetoden fungerer bedre eller dårligere, og i hvilken retning avvikene går. Dersom punktene ligger over den grå linjen betyr dette at den omregnede karakteren får en bedre rangering i Norge enn det opprinnelige karaktersnittet hadde i Etiopia, altså at omregningsmetoden er litt for fordelaktig for disse karaktersnittene. Dersom punktene ligger under den grå linjen, er det omvendt.

Vi kan se nærmere på avvikene i plottet til høyre. Der ser vi på differansen mellom $G(\hat{y})$ og $\hat{F}(x)$ for hvert eneste mulige etiopiske karaktersnitt: differanser som er forskjellige fra 0 viser brudd på kalibrering. For den omregningsmetoden vi ser på her er avvikene oftere positive enn negative, det vil si at omregningsmetoden ofte gir en litt for høy omregnet norsk karakter, særlig for etiopiske karakterer rundt 2 og 3.

I begge plottene bruker vi $\hat{F}(x)$ med en «hatt» for å understreke at den etiopiske kumulative fordelingsfunksjonen $F(\cdot)$ er estimert basert på egeninnsamlet statistikk.

Vi kan få et generelt inntrykk av graden av kalibrering ved å definere scorer. En vanlig score, her kalt K , oppnår vi ved å summere opp de kvadrerte avvikene mellom $G(\hat{y})$ og $\hat{F}(x)$ for alle mulige verdier av x_i (som vi indekserer fra $i = 1$ til n under):

$$K = \sqrt{\sum_{i=1}^n [G(\hat{y}_i) - F(x_i)]^2}.$$

K vil alltid være et positivt tall. Perfekt kalibrering vil gi $K = 0$, jo større K jo dårligere kalibrering.

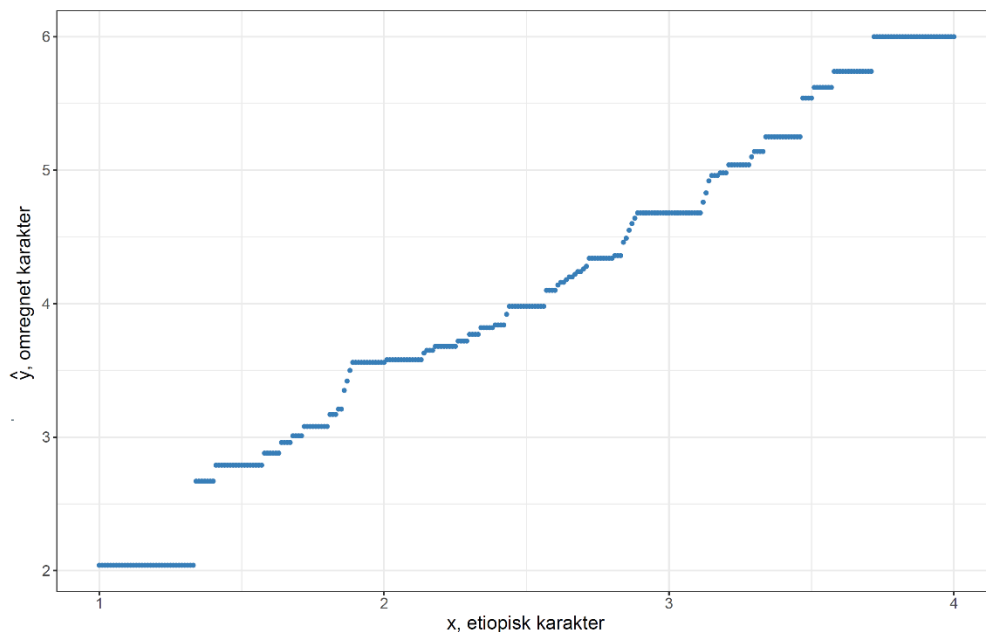
I tillegg kan vi se på (den lineære) korrelasjonen mellom $G(\hat{y})$ og $\hat{F}(x)$:

$$C = \text{corr}(G(\hat{y}), \hat{F}(x)).$$

Perfekt kalibrering gir en korrelasjon på 1, mens et tall nærmere 0 tyder på dårligere kalibrering.

4.2 Undersøke differensiering

Vi kan undersøke differensiering ved å se på en figur med x , det etiopiske karaktersnittet, mot $\hat{y}$, den omregnede norske karakteren:



Figur 5. Eksempel på plott for å se på grad av differensiering.

Områder der prikkene ligger flatt etter hverandre viser deler av den etiopiske karakterskalaen som blir regnet om til samme norske karakter, for eksempel ser vi at med denne omregningen får alle etiopiske karakterer mellom 1 og ca 1.4 den norske karakteren 2. Vi ser at det er flere slike intervaller i denne figuren.

En enkel score for den generelle graden av differensiering er

$$D = \frac{\#\{\text{omregnede karakterer}\}}{\#\{\text{etiopiske karakterer}\}}.$$

Firkanttegnet betyr «antall». D er altså antall unike karakterer etter omregning delt på antall etiopiske karakterer. Det er 301 etiopiske karakterer mellom 1 og 4 (med 2 desimaler), og for den omregningen vi ser her blir disse regnet om til kun 53 forskjellige norske karakterer, det vi si at $D = 0,18$ i dette eksempelet. Perfekt differensiering gir $D = 1$.

4.3 Metoder for å hankses med hull

Vi har sett på ni ulike metoder for å hankses med hull, se tabell 7. Metode A er metoden som brukes nå. Metodene B til G er varianter av dagens metode, mens metodene H og I er varianter av ecdf-metodikken som er beskrevet i kapittel 3.4.

Tabell 7. Metoder for å hankses med hull som vi har brukt i Etiopia-eksempelet.			
	Navn på metode	Kort beskrivelse	Fylling av hull?
A	Dagens	Metoden som er i bruk nå	Ja
B	Enkel	Bruker bare hundreprosentilene direkte, fyller ingen hull	Nei
C	Gammel, «trappetrinnsmetoden»	Metoden som brukes for å fylle hullene i Excel-arket «Bakgrunnsmateriale til omregningstabell for Etiopia, A-F»	Ja
D	Begge sider, ett steg	Justerer begge sider av prosentilhullet, ett steg i hver ende	Ja
E	Begge sider, to steg	Justerer begge sider av prosentilhullet, to steg i hver ende	Ja
F	Norske prosentiler	Fyller hullet ved å bruke den norske karakterstatistikken i prosentilhullet	Ja
G	Baklengs	Som dagens metode bare fra bunnen og oppover	Ja
H	ecdf	Bruker den empirisk kumulative fordelingsfunksjonen	Nei
I	ecdf interpolert	Bruker en lineær interpolering av den empirisk kumulative fordelingsfunksjonen	Ja, og endrer potensielt alle karaktergrenser

For en mer detaljert beskrivelse av dagens metode, Metode A, henviser vi til SO sitt notat «Metoder og dataprogram for omregningstabeller, juli 2022». Vi starter i bunnen av tabellen (fra lave til høyere prosentiler) og fyller store hull etter hvert som vi støter på disse. Store hull fylles ved å interpolere mellom minimumskarakteren to steg før hullet til minimumskarakteren ett steg etter hullet (se blåfargede tall i tabell 9). Deretter justeres maksimumskarakterene på passende vis. Til slutt finner vi omregnede norske karakterer ved å matche hundreprosentilene i den etiopiske tabellen med de norske hundreprosentilene.

Metode C er en eldre variant av Metode A der det interpoleres direkte mellom de norske karakterene som er omfattet av hullet. Den brukes for å fylle hullene i Excel-arket «Bakgrunnsmateriale til omregningstabell for Etiopia, A-F», og den beskrives mer inngående der.

Metode B bruker bare hundreprosentilene direkte uten å fylle noen hull. Vi illustrerer forskjellene på Metode B og Metodene A, D, E, F og G ved å se et eksempel fra Etiopia-tabellen (se tabell 9). De fargede tallene i hvert kolonnepar viser hvilke karaktergrenser som er endret av de ulike metodene. Med Metode B endres det naturligvis ingen karaktergrenser. Metode A endrer grensene mellom minimumsgrensen to steg før hullet (2,61) og minimumsgrensen ett steg etter hullet (2,71). Metode D og F er begge to slik at de interpolerer mellom minimumsgrensen ett steg før hullet (2,64) og maksimumsgrensen ett steg etter hullet (2,71). Disse metodene endrer altså karaktergrensene rundt prosentilhullet i noe mindre grad enn Metode A. De er forskjellige ved at Metode D gjør lineær interpolasjon, mens Metode F interpolerer ved å bruke den norske karakterstatistikken for de tilsvarende hundreprosentilene. Tabell 8 viser den norske karakterstatistikken rundt prosentilhullet.

Tabell 8. Den norske karakterstatistikken rundt prosentilhullet.							
Hundreprosentil	49	50	51	52	53	54	55
Antall	5225	2786	5349	2827	5344	3071	4835
Prosent	17,7	9,5	18,2	9,6	18,2	10,4	16,4

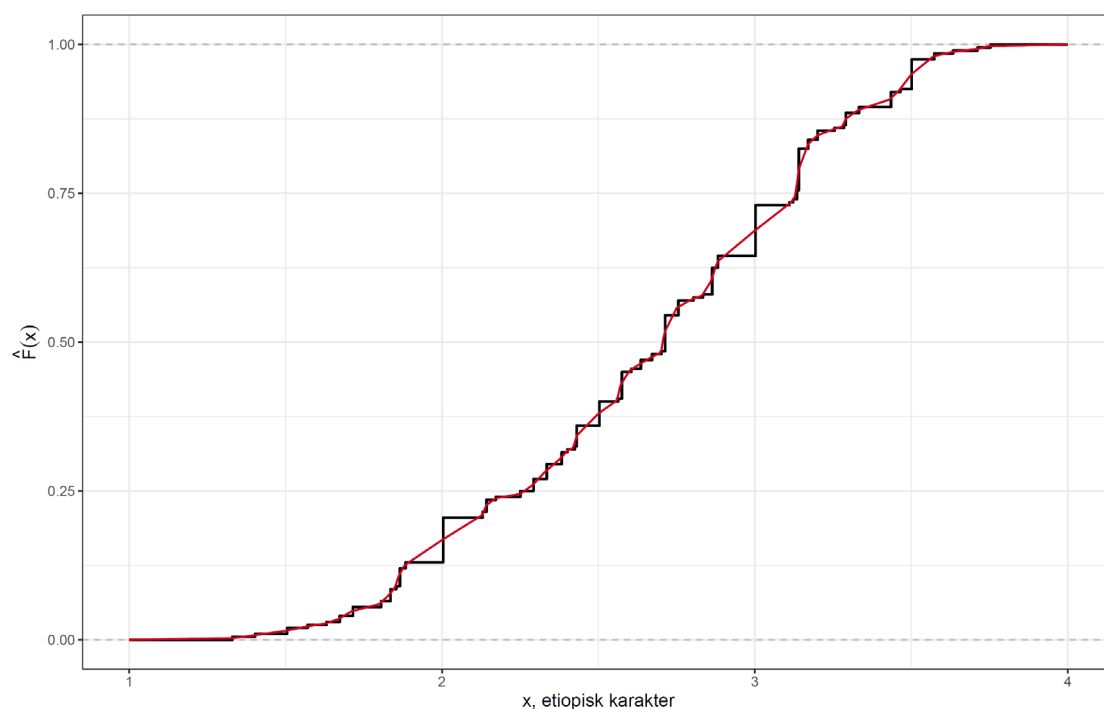
Her ser vi at det for eksempel er en noe større andel karakterer i 51-prosentilen sammenlignet med 50-prosentilen. Vi interpolerer ved å bruke disse prosentandelene, som gjør at 51-prosentilen vil få en noe større del av tallmassen mellom 2,64 og 2,71 enn 50-prosentilen. I praksis gir ikke dette forskjellige karaktergrenser i eksempelet under (Metode D og F har de samme grensene), og som vi skal se lenger nede er Metode D og Metode F generelt veldig like for Etiopia-dataene. Metode E interpolerer mellom minimumsgrensen to steg før hullet (2,61) og maksimumsgrensen to steg etter hullet (2,80, siden det er et lite hull rett etter det store hullet må vi gå helt ned til 58-prosentilen). Metode E fører dermed til større endringer rundt prosentilhullet enn Metode A. Metode G er ment å være en «baklengs» versjon av Metode A og starter dermed med å fylle hullene fra toppen av tabellen. Den interpolerer mellom maksimumsgrensen ett steg før hullet (2,70) og maksimumsgrensen to steg etter hullet (2,80).

Tabell 9. Eksempel på fylling av hull i hundreprosentiltabell med ulike metoder.												
	Metode B Enkel		Metode A Dagens		Metode D		Metode E		Metode F		Metode G	
%	Min.	Maks.	Min.	Maks.	Min.	Maks.	Min.	Maks.	Min.	Maks.	Min.	Maks.
48	2,61	2,63	2,61	2,61	2,61	2,63	2,61	2,62	2,61	2,63	2,61	2,63
49	2,64	2,70	2,62	2,63	2,64	2,64	2,63	2,63	2,64	2,64	2,64	2,70
50	-	-	2,64	2,64	2,65	2,65	2,64	2,65	2,65	2,65	2,71	2,71
51	-	-	2,65	2,66	2,66	2,66	2,66	2,67	2,66	2,66	2,72	2,72
52	-	-	2,67	2,67	2,67	2,67	2,68	2,69	2,67	2,67	2,73	2,73
53	-	-	2,68	2,69	2,68	2,68	2,70	2,70	2,68	2,68	2,74	2,74
54	-	-	2,70	2,70	2,69	2,69	2,71	2,72	2,69	2,69	2,75	2,76
55	2,71	2,71	2,71	2,71	2,70	2,71	2,73	2,74	2,70	2,71	2,77	2,77
56	-	-	-	-	-	-	2,75	2,76	-	-	2,78	2,78
57	-	-	-	-	-	-	2,77	2,77	-	-	2,79	2,79
58	2,72	2,80	2,72	2,80	2,72	2,80	2,78	2,80	2,72	2,80	2,80	2,80

Metodene H og I forlater hundreprosentilene og bruker isteden den fulle empiriske kumulative fordelingen (ecdf), som er definert i kapittel 3. I kapittel 1 skisserte vi at en omregningsmetode burde være slik at $F(x) = G(\hat{y})$. Av dette følger det følgende Metode H: $\hat{y} = G^{-1}[\hat{F}(x)]$, der $G^{-1}[\cdot]$ er den inverse kumulative fordelingsfunksjonen, også kalt kvantilfunksjonen, basert på norsk karakterstatistikk, og $\hat{F}(\cdot)$ nok en gang er den estimerte/empiriske kumulative fordelingsfunksjonen (ecdf) basert på egeninnsamlet etiopisk karakterstatistikk, se den svarte linjen i figur 6.

Metode H er en naturlig utvidelse av hundreprosentil-metodikken. Metode I er en variant av Metode H som kan skrives som $\hat{y} = G^{-1}[\tilde{F}(x)]$, der $\tilde{F}(x)$ et annet estimat på den kumulative fordelingsfunksjonen enn den vanlige empirisk kumulative fordelingsfunksjonen som vi brukte for Metode H. Her bruker vi en stykkevis lineær interpolering av $\hat{F}(x)$, der man trekker en rett linje mellom midten av hvert steg i den stegvise ecdf-en, se den røde linjen i figur 6. Vi tar altså utgangspunkt i stegene i $\hat{F}(x)$, den svarte linjen, og trekker rette linjer innenfor hvert steg - $\tilde{F}(x)$ blir altså en sammensetning av flere rette linjer med ulike stigningstall.

Metode I kan sees på som en generell versjon av «hullfyllingsmetodene» A og C–G, der vi istedenfor å kun endre karaktergrensene rundt store hull justerer alle karaktergrensene. I nederste og øverste steg bruker Metode I grensene for karakterskalaen (for eksempel 1 og 4 for Etiopia) for å trekke første og siste rette linje.

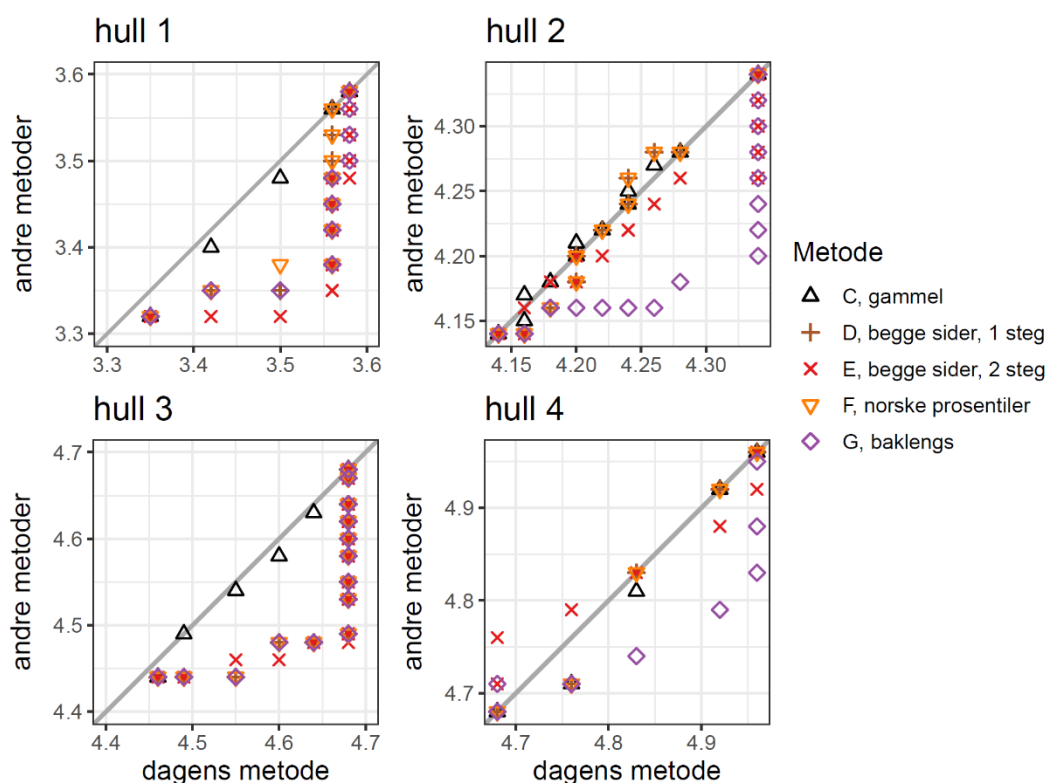


Figur 6. Stykkevis lineær interpolasjon (rød) av den empirisk kumulative fordelingsfunksjonen (svart) basert på 200 etiopiske karaktersnitt.

4.4 Resultater for Etiopia

Vi starter med å sammenligne de omregnede norske karaktersnittene vi oppnår med de ulike metodene beskrevet over, før vi etterpå skal sammenligne dem med tanke på kalibrering og differensiering. Først ser vi på de fem metodene som er varianter av dagens metode («hullfyllingsmetoder»): C, D, E, F og G. I figur 7 ser vi på de omregnede norske karaktersnittene inni hvert av de store prosentilhullene. Det er fire

store prosentilhull i hundreprosentiltabellen for Etiopia. Utenom disse fire hullene gir de fem metodene naturligvis like omregninger.



Figur 7. Omregnet norsk karaktersnitt med dagens metode mot omregnet norsk karaktersnitt med noen varianter av dagens metode.

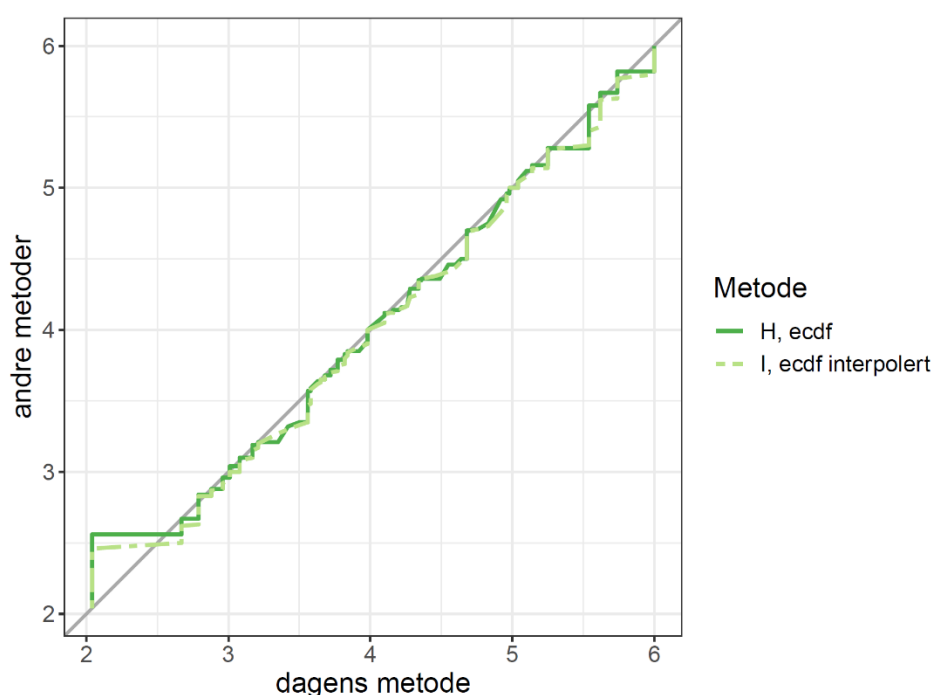
Hvert panel i figuren over representerer et hull. På den horisontale akse ser vi det omregnede norske karaktersnittet med dagens metode, mens på den vertikale akse ser vi det omregnede norske karaktersnittet med noen varianter av dagens metode. Dersom punktene i figurene ligger på den diagonale grå linjen betyr det at man får samme omregning med dagens metode som med en annen metode. Punkter som ligger under diagonallinjen viser at den andre omregningsmetoden gir lavere karaktersnitt enn dagens metode, mens punkter over diagonallinen viser det omvendte.

Vi ser at alle de fem variantene av dagens metode gir omregnede karaktersnitt som nesten alltid er noe lavere enn det man får av dagens. Den «gamle» metoden (Metode C, svarte trekkanter) er ofte svært lik nåværende metode, og vi kommer derfor ikke til å studere denne metoden videre (da den uansett er mer kronglete å regne ut enn dagens metode). Metode D «begge sider, 1 steg» gir nesten alltid identisk resultat som Metode F «norske prosentiler» (fordi begge metoder justerer ett steg på begge sider av prosentilhullet, som diskutert over), og derfor i studerer vi bare Metode F «norske prosentiler» videre og lar Metode D ligge.

Punkter av samme farge som ligger vertikalt oppå hverandre angir et karaktersnittintervall av etiopiske karaktersnitt der dagens metode gir nøyaktige samme omregnede karaktersnitt, mens den andre metoden gir ulike omregnede karaktersnitt. For eksempel ser Metode E «begge sider, 2 steg» og Metode G «baklengs» ut til å ofte gi mer differensierte karaktersnitt enn dagens metode (A).

Videre kan vi sammenligne dagens metode med de to ecdf-metodene, Metode H og I (figur 8). Her ser vi på hele karakterskalaen, siden ecdf-metodene kan endre alle omregningene sammenlignet med dagens metode. Begge metodene gir oftere noe lavere karaktersnitt enn dagens metode. Unntaket er rundt 2.5 der ecdf-metodene gir et høyere karaktersnitt enn dagens metode. Dette klare avviket fra diagonallinjen skyldes hovedsakelig ett eneste etiopiske karaktersnitt, nemlig 1.33 (se tabell 10).

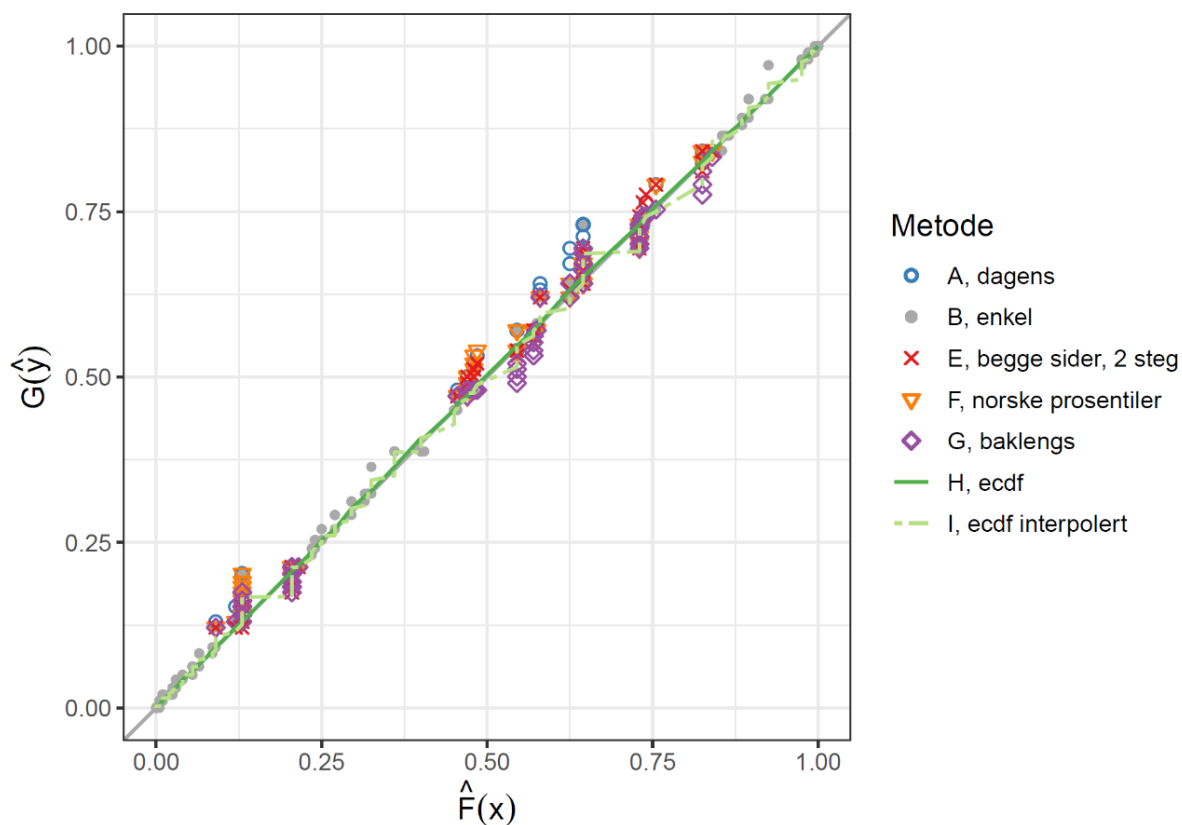
Tabell 10. Et utsnitt av omregningstabellene basert på dagens metode og de to ecdf-metodene, der disse gir ganske ulikt resultat.			
Etiopisk snitt	Norsk snitt (A, dagens)	Norsk snitt (H, ecdf)	Norsk snitt (I, ecdf interpolert)
1,32	2,04	2,04	2,46
1,33	2,04	2,56	2,46
1,34	2,67	2,56	2,50



Figur 8. Omregnet norsk karaktersnitt med dagens metoder mot omregnet norsk karaktersnitt med ecdf-metodene (fargede linjer).

4.4.1 Kalibrering

Figur 9 viser vi kalibreringsplot for de syv gjenværende metodene. Vi skal se nærmere på hvert enkelt hull under. Generelt ser vi at alle metodene er relativt godt kalibrert (punktene ligger ganske nærme diagonallinjen). Metode H «ecdf» er tilnærmet perfekt, per konstruksjon. Merk at metodene som baserer seg på prosentiler og fylling av hull kun varierer i hullene (de fargede punktene), mens de to ecdf-metodene er forskjellige fra resten overalt og er dermed vist som linjer istedenfor punkter. For de fleste av metodene ligger punktene oftere over diagonallinjen enn under, dette betyr at $G(\hat{y})$ er større enn $\hat{F}(x)$, altså at det omregnede karaktersnittet er ofte litt for høy sammenlignet med det opprinnelige etiopiske karaktersnittet (de omregnede karaktersnittene får ofte en litt bedre rangering i Norge enn de hadde i den etiopiske statistikken). Unntaket er metoden «baklengs» som ofte gir litt for lav omregnet karaktersnitt.

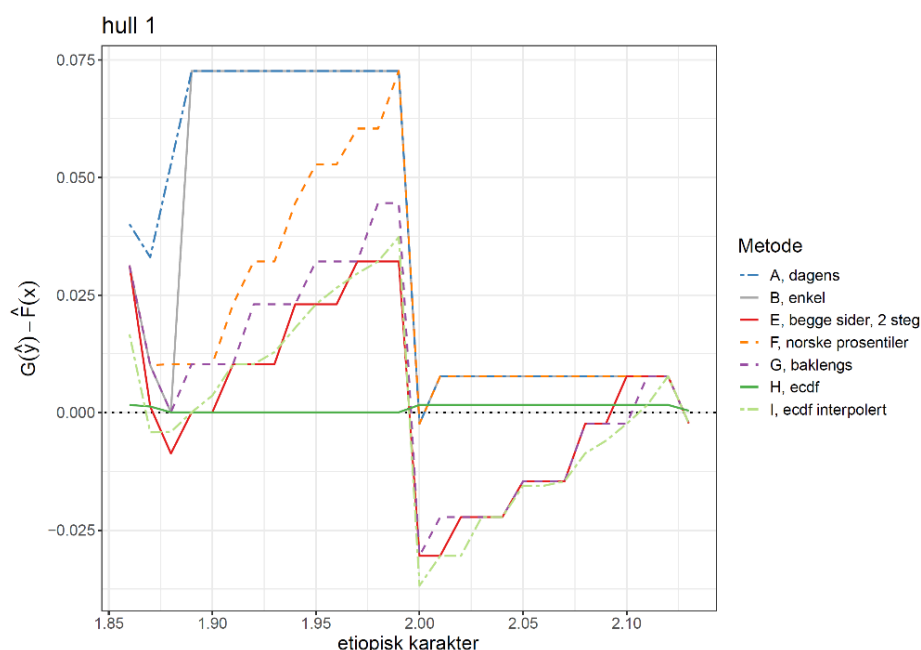


Figur 9. Kalibreringsplott for Etiopia-eksempelen. Ulike omregningsmetoder er vist med ulike fargede symboler/linjer.

Vi bruker scorene vi definerte i kapittel 4.1 for å finne ut hvilken metode som er generelt best kalibrert (tabell 11). For K-scoren er lavere score bedre. For C-scoren er høyere score best. Metode H «ecdf» har K-score nær null per definisjon, og Metode I «ecdf interpolert» er nest best ifølge denne scoren. Alle de nye metodene er bedre kalibrert enn dagens metode ifølge begge scorene.

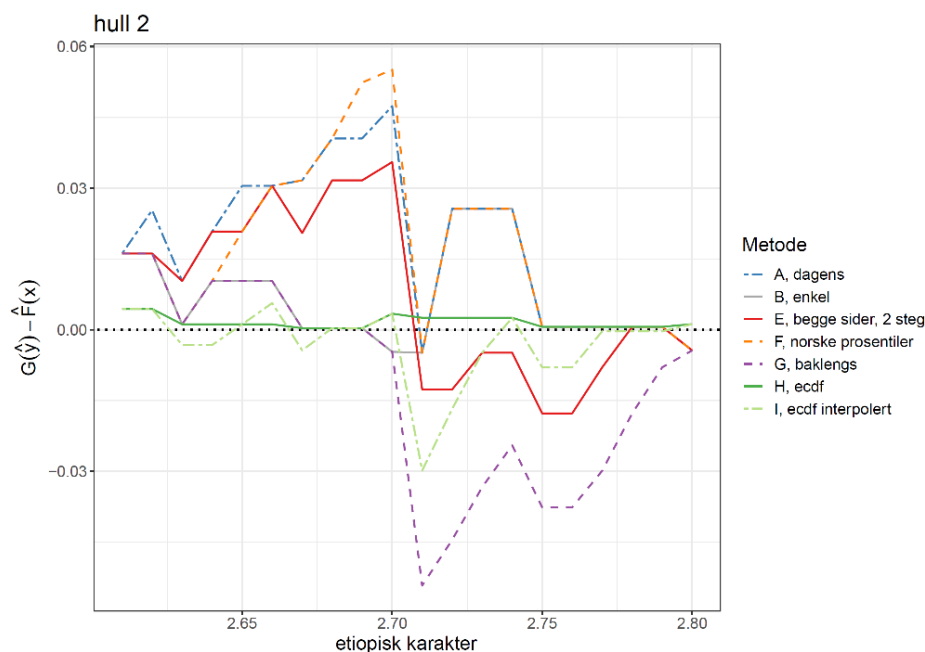
Tabell 11. Kalibreringsscorer for Etiopia-eksempelen*.			
	Metode	K	C
A	Dagens	0,454	0,9982
B	Enkel	0,420	0,9984
C	Gammel	0,449	0,9982
D	Begge sider, 1 steg	0,295	0,9992
E	Begge sider, 2 steg	0,257	0,9993
F	Norske prosentiler	0,289	0,9992
G	Baklengs	0,278	0,9992
H	ecdf	0,04	≈ 1
I	ecdf interpolert	0,185	0,9996
*Scorene K og C er definert i kapittel 4.1			

For å se litt nærmere på forskjellene mellom metodene ser vi på figurene 10–13 med differansen mellom $G(\hat{y})$ og $\hat{F}(x)$ innenfor hvert hull (plottet mot den opprinnelige etiopiske karakteren i hullet). Her har vi vist metodene som linjer for at det skal være lettere å se forskjellene mellom dem. Jo nærmere de fargede linjene er 0, jo bedre er metodene kalibrert.



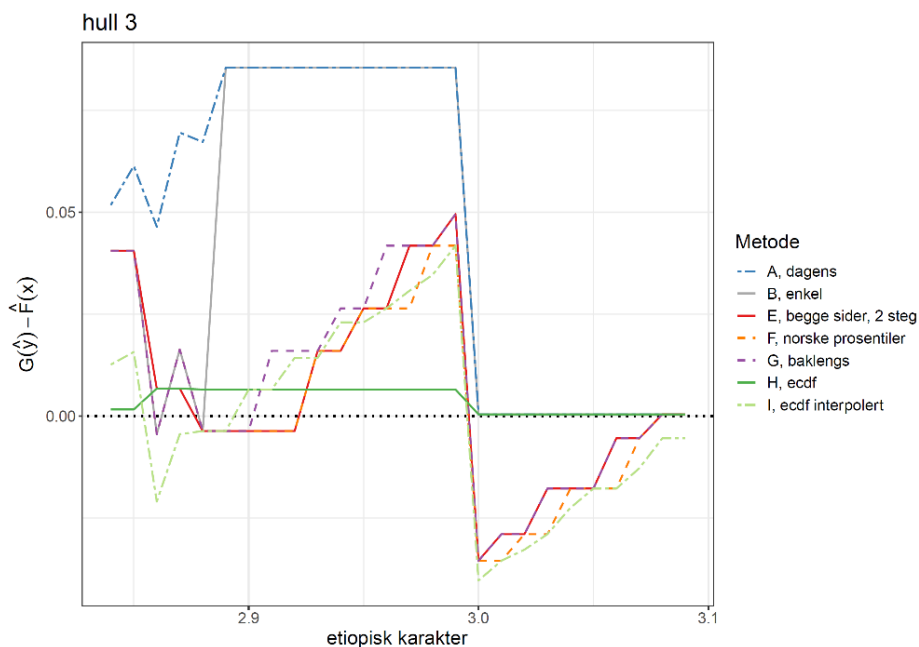
Figur 10. Kalibreringsplott, differansen i rangering, for hull 1 i Etiopia-eksempelet. Ulike omregningsmetoder er vist med ulike fargede linjer.

I hull 1 har metodene A «dagens», B «enkel» og F «norske prosentiler» relativt lik kalibrering, og gir omregninger som nesten alltid er for høye. Vi ser at Metode A og F endrer tallene etter prosentilhullet i liten grad: ingen endring etter Etiopisk karaktersnitt 1,89 for «dagens» og fra 1,99 for «norske prosentiler». Metodene E «begge sider, 2 steg», G «baklengs» og I «ecdf interpolert» endrer tallene etter prosentilhullet i stor grad (frem til ca 2,10). Disse siste metodene (E,G,I) gir et omregnet karaktersnitt som er litt for høyt først (før 2) og deretter litt for lavt.



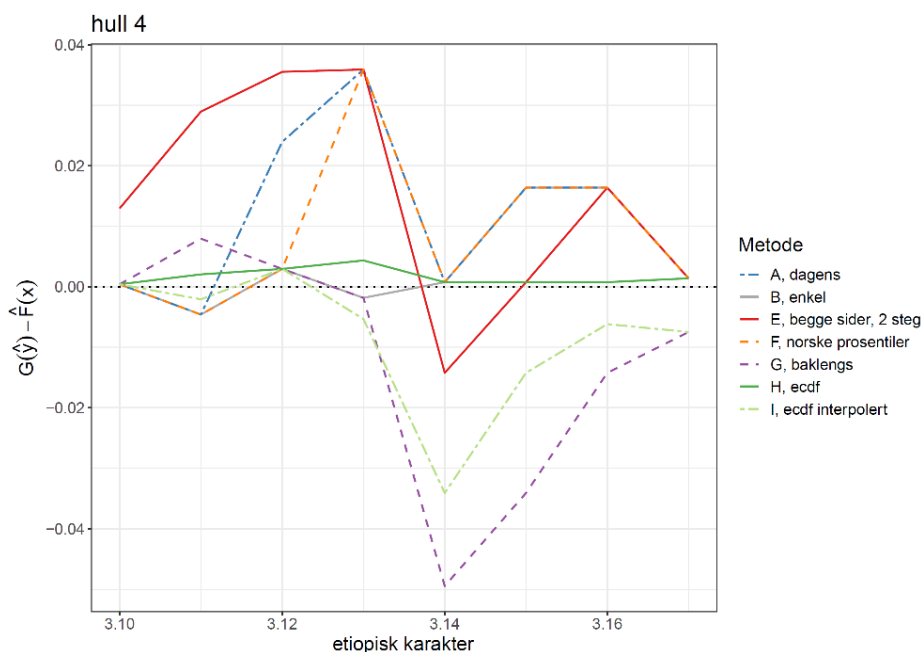
Figur 11. Kalibreringsplott, differansen i rangering, for hull 2 i Etiopia-eksempelet. Ulike omregningsmetoder er vist med ulike fargede linjer.

I hull 2 har Metode I «ecdf interpolert» og Metode B «enkel» god kalibrering for mesteparten av intervallet, unntatt for 2,71 for Metode I og mellom 2,72 og 2,74 for Metode B. Metodene A «dagens» og F «norske prosentiler» gir omregninger som nesten alltid er litt for høye. Mens Metode E «begge sider, 2 steg» gir omregninger som først er litt for høye og deretter litt for lave. Metode G «baklengs» gir klart for lave omregnede karaktersnitt fra 2,70 og videre.



Figur 12. Kalibreringsplott, differansen i rangering, for hull 3 i Etiopia-eksempelet. Ulike omregningsmetoder er vist med ulike fargede linjer.

I hull 3 har metodene A «dagens» og B «enkel» relativt lik kalibrering, og gir omregninger som er for høye mellom 2,89 og 2,99. Metodene E «begge sider, 2 steg», F «norske prosentiler, G «baklengs» og I «ecdf interpolert» er ganske likt kalibrert: før 3,00 gir de ofte litt for høye omregnede karaktersnitt og deretter litt for lave.



Figur 13. Kalibreringsplott, differansen i rangering, for hull 4 i Etiopia-eksempelet. Ulike omregningsmetoder er vist med ulike fargede linjer.

I hull 4 har Metode B «enkel» god kalibrering frem til 3,15. Metodene A «dagens», E «begge sider, 2 steg» og F «norske prosentiler» gir omregninger som oftest er for høye. Mens Metodene G «baklengs» og I «ecdf interpolert» oftest gir for lave omregninger.

4.4.2 Differensiering

Vi ser på D scoren for å vurdere i hvor stor grad metodene differensierer mellom ulike etiopisk karaktersnitt, se tabell 12. Jo nærmere 1 jo bedre. Den enkle Metode B uten fylling av hull differensierer minst, som forventet, mens Metode I «ecdf interpolert» differensierer mest. De fleste metodene differensierer mer enn dagens metode.

Tabell 12. Differensieringsscore for Etiopia-eksempelet*.		
	Metode	D
A	Dagens	0,176
B	Enkel	0,146
C	Gammel	0,186
D	Begge sider, 1 steg	0,209
E	Begge sider, 2 steg	0,229
F	Norske prosentiler	0,209

G	Baklengs	0,223
H	ecdf	0,179
I	ecdf interpolert	0,708
*Scoren D er definert i kapittel 4.2		

4.4.3 Konklusjoner

Basert på analysen av omregningsmetoder av etiopiske karaktersnitt ser vi følgende:

- Metode H (ecdf) er best på kalibrering, men gir en relativt lav grad av differensiering (ca lik som dagens metode).
- Metode I (ecdf interpolert) har både relativt god kalibrering og differensiering. Det samme gjelder til en viss grad metode E (Begge sider, 2 steg) og G (Baklengs), men disse har dårligere scorer enn I (ecdf interpolert).
- Metoden H (ecdf) er enkel og enhetlig å bruke, men må kanskje studeres litt nærmere (særlig med tanke på hull i endene). Da vil man «slippe» å måtte lage hundreprosenttabeller.
- En kan tenke seg andre interpoleringsmetoder for den empirisk kumulative kurven. For eksempel kunne en vurdere å bruke parametriske modeller. Dette må i så fall undersøkes nærmere i et eget arbeid.

Dette er kommentarer basert på vår analyse av etiopisk karakterstatistikk, og det er ikke opplagt at det vi her har funnet ut er direkte overførbart til alle andre omregningssituasjoner, selv om hovedprinsippene nok vil være de samme.

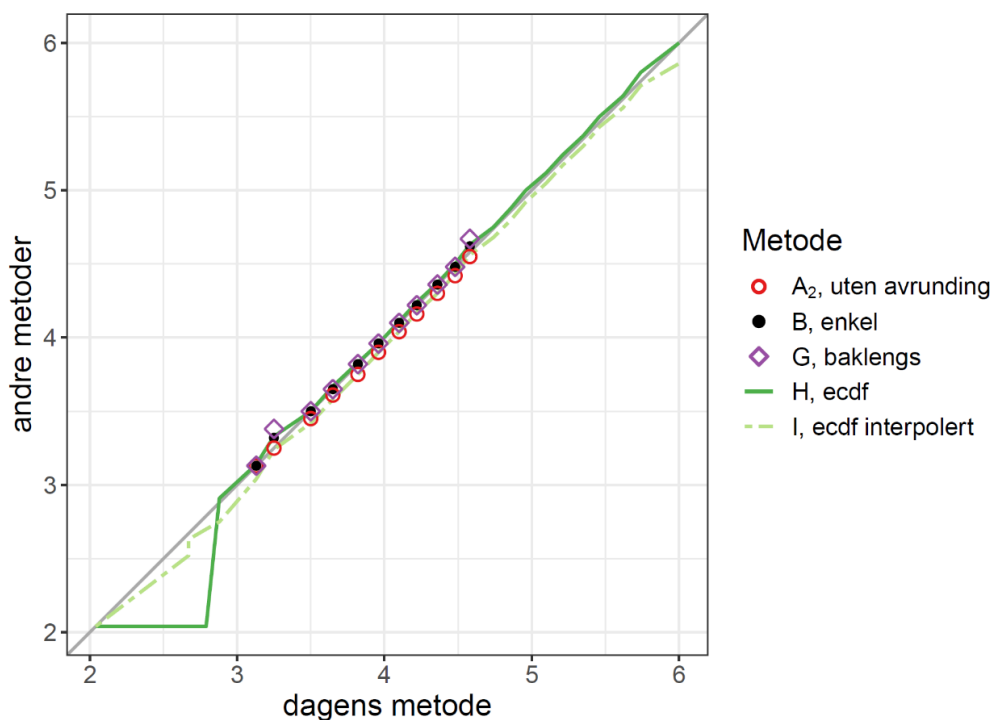
4.5 Resultater for International Baccalaureate

Som vi skriver i starten av kapittel 4 så har International Baccalaureate (IB) en grov karakterskala som kun består av maksimum 26 mulige poengsummer/karaktersnitt. Det er skalaen som gir hull i hundreprosenttabellen. For IB har vi offisiell karakterstatistikk som omfatter mange hundretusen elever. Dette er altså en ganske annerledes type hull enn for Etiopia-eksempelet over.

På grunn av denne uvanlige karakterskalaen er dagens metode blitt endret noe for bruk på IB-omregningen sammenlignet med den generelle metoden beskrevet i kapittel 4.3. Blant annet gjør dagens metode for IB data utstrakt bruk av avrunding i flere steg. En annen effekt av den grove skalaen er at mange av metodene beskrevet i tabell 7 vil gi nøyaktig samme omregning for IB: alle metoder som kun justerer ett steg i hver ende av prosentilhullet vil gi samme svar som Metode B «Enkel». I tillegg er det grunn til å anta at å justere to steg i hver ende vil være en dårlig idé i dette tilfellet og derfor studerer vi ikke den metoden her.

Tabell 13. Metoder for å hankses med hull som vi har brukt i IB-eksempelet.			
	Navn på	Kort beskrivelse	Fylling av hull?
A ₁	Dagens	Metoden som er i bruk nå	Ja
A ₂	Uten avrunding	Som dagens bare uten avrunding i hvert steg	Ja
B	Enkel	Bruker bare prosentilene direkte, fyller ingen hull	Nei
G	Baklengs	Som dagens metode bare fra bunnen og oppover	Ja
H	ecdf	Bruker den empirisk kumulative fordelingsfunksjonen	Nei
I	ecdf interpolert	Bruker en lineær interpolering av den empirisk kumulative fordelingsfunksjonen	Ja, men endrer potensielt alle karaktersnitt

Alle metodene ser ut til å gi nokså like omregnede karaktersnitt som dagens metode i IB-tilfellet. Unntaket er Metode H helt nederst i karakterskalaen der denne skiller seg en del fra de andre, se figur 14. Dette skyldes at de tre laveste karaktersnittene i skalaen er «teoretiske» og aldri observert i karakterstatistikken.

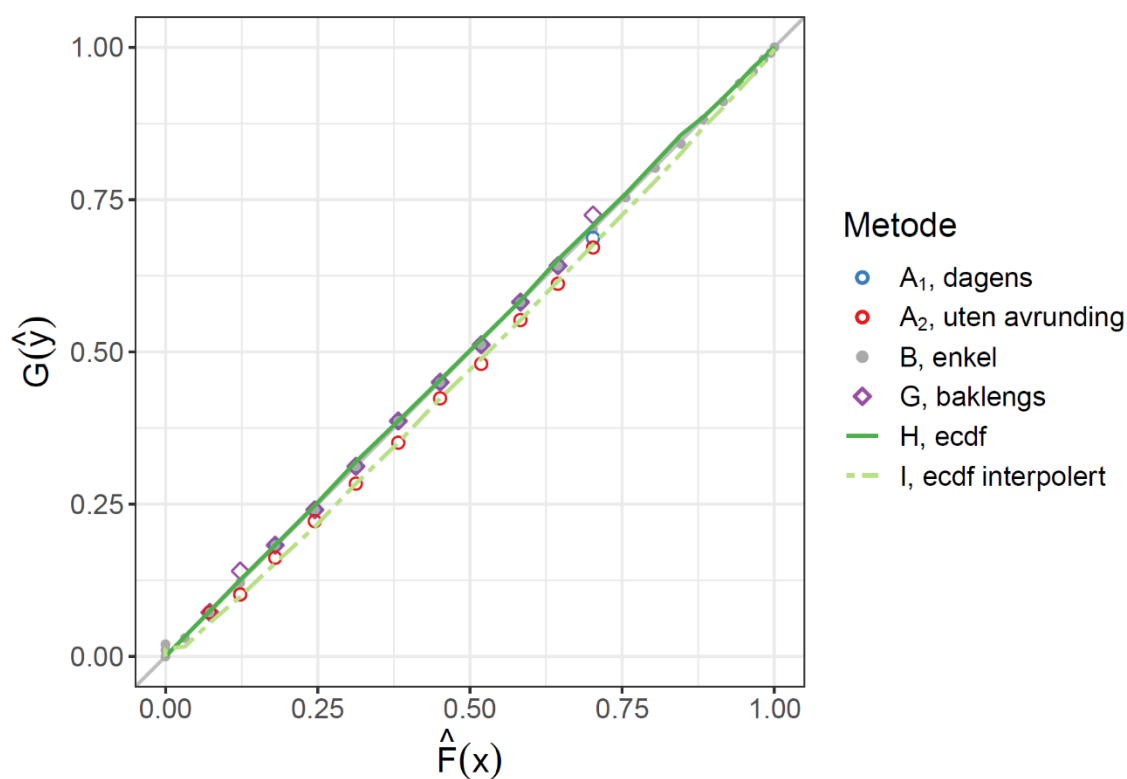


Figur 14. Omregnet norsk karaktersnitt med dagens metode mot omregnet norsk karaktersnitt med noen varianter av dagens metode, for IB-eksempelet.

Alle metodene er relativt godt kalibrert her, se figur 15 og tabell 14. Dagens metode er middels godt kalibrert sammenlignet med de andre. Metode I «ecdf interpolert» gir nesten alltid en litt for lav omregnet karaktersnitt og scorer dårligst med tanke på

kalibrering. Metode A₂ «uten avrunding» har også dårlige kalibreringsscorer. Metode H «ecdf» scorer best (igjen, per definisjon), Metode B «enkel» scorer nest best.

Tabell 14. Kalibreringsscorer for IB-eksempelet.			
	Metode	K	C
A ₁	Dagens	0,039	0,9998
A ₂	Uten avrunding	0,029	0,9999
B	Enkel	0,095	0,9991
G	Baklengs	0,041	0,9998
H	ecdf	0,019	≈ 1
I	ecdf interpolert	0,107	0,9994
*Scorene K og C er definert i kapittel 4.1			



Figur 15. Kalibreringsplott for IB-eksempelet. Ulike omregningsmetoder er vist med ulike fargede symboler/linjer.

Alle metodene gir godt differensierte omregninger, se tabell 15. Metode H «ecdf» har en noe lavere grad av differensiering sammenlignet med de andre metodene, og dette skyldes bunnen av karakterskalaen der de tre laveste mulige karaktersnittene får samme omregnede norske karaktersnitt. Dette kan utbedres ved å lage en variant av Metode H som behandler bunnen (og eventuelt toppen) av skalaen litt annerledes, jmfør diskusjonen i analyse 5.

Tabell 15. Differensieringsscore for IB-eksempelet.		
	Metode	D
A ₁	Dagens	0,962
A ₂	Uten avrunding	0,962
B	Enkel	0,962
G	Baklengs	0,962
H	ecdf	0,885
I	ecdf interpolert	1
*Scoren D er definert i kapittel 4.2		

4.6 Oppsummering av Analyse 2

Vi har sett på etiopisk karakterstatistikk og karakterstatistikk fra IB, og diskutert problemstillinger knyttet til «store hull» i deres statistikk. Vi har følgende anbefalinger basert på de analysene vi har gjort:

- ecdf-metoden vil per konstruksjon alltid være best med tanke på kalibrering, men kan gi en ganske lav grad av differensiering (når ecdf-en er basert på egeninnsamlet statistikk).
- Dersom man er i en situasjon som i Etiopia-eksempelet (egeninnsamlet statistikk, fininndelt karakterskala) så kan man vurdere å bruke varianter av ecdf-metoden, som «ecdf interpolert» eller lignende.
- Dersom man er i en situasjon som Etiopia-eksempelet (fin karakterskala og egeninnsamlet statistikk) og man ønsker å fortsette å bruke hundreprosentiltabellene er det sannsynligvis bedre å bruke metoder som endrer grensene rundt prosentilhullene i større grad enn dagens metode (som for eksempel «begge sider, 2 steg»).
- Det er ikke nødvendig å fylle hull i situasjoner som IB-eksempelet (grov skala og offisiell karakterstatistikk). For IB er det best å kun bruke prosentilene direkte (Metode B «Enkel») eller å bruke ecdf-metoden (gjerne med noe spesielt for bunnen/toppen av skalaen).

5 Analyse 3: Hvordan håndtere tilfeller der utenlandske karaktersnitt er større enn største mulige verdi?

Det finnes tilfeller der tilsendt statistikk fra andre land inneholder enkelte karaktersnitt som har høyere verdi enn største verdi på landets skala. Et eksempel er statistikk fra Frankrike der den offisielle skalaen går fra 10,0 til 20,0, mens tilsendt statistikk viser at noen har fått en snittkarakter på over 20,0. Tabell 16 viser antallet elever med snittkarakter større enn 20,0 i årene 2014 til 2018.

Tabell 16. Oversikt over antallet elever med ulike karaktersnitt over største offisielle karaktersnitt 20,0 i 2014–2018.					
Karaktersnitt	2014	2015	2016	2017	2018
20,1	47	40	64	62	96
20,2	38	26	44	53	66
20,3	29	25	30	47	55
20,4	21	9	15	26	22
20,5	16	13	22	33	28
20,6	14	9	17	20	16
20,7	9	2	14	16	16
20,8	7	2	15	11	10
20,9	0	1	0	6	8
21,0	2	2	2	3	3
21,1	0	0	2	3	3
21,2	0	0	1	2	1
21,3	0	0	0	0	2
Sum	183	129	226	282	326

Vi ser to mulige måter å håndtere dette på i arbeidet med å lage omregningstabeller:

1. Karaktersnitt over 20,0 behandles som om de var 20,0, og prosentilskalaen beregnes deretter.
2. Utvide karakterskalaen til største observerte verdi (her 21,3), og prosentilskalaen beregnes deretter. Dette vil kunne innebære at SO årlig må beregne oversettelsestabeller for de landene det gjelder.

Hvilken maksimal verdi på karakterskalaen vi legger til grunn vil kunne få betydning i toppen av karakterskalaen, og tabell 17 viser dette.

Tabell 17. Omregningstabell for franske til norske karaktersnitt, øverste prosentiler når vi lar største franske karaktersnitt være 20,00 eller 21,30.						
	Øvre grense er 20,00			Øvre grense er 21,30		
Prosentil	Nedre fransk	Øvre fransk	Norsk karakter	Nedre fransk	Øvre fransk	Norsk karakter
95	16,2	16,4	5,35	16,2	16,4	5,35
96	16,5	16,7	5,40	16,5	16,7	5,40
97	16,8	17,0	5,46	16,8	17,0	5,46
98	17,1	17,5	5,54	17,1	17,5	5,54
99	17,6	18,1	5,62	17,6	18,1	5,62
100	18,2	19,9	5,74	18,2	21,2	5,74
101	20,0	20,0	6,00	21,3	21,3	6,00

Resultatene i tabell 17 er kanskje ikke så overraskende i lys av hvordan omregningen gjøres, og all den tid det er svært få som har fått en snittkarakter over 20,0. Lager vi en

omregningstabell med ecdf-metoden, som beskrevet i kapittel 3.4, gir dette større forskjeller i omregningstabellen, som vist i tabell 18.

Tabell 18. Omregningstabell; franske og norske empiriske kumulative fordelingsfunksjoner (ecdf), avrundet til 4 desimaler, med tilhørende karaktersnitt*							
Øvre grense er 20,00				Øvre grense er 21,30			
Fransk kar,	Fransk prosentil	Norsk kar,	Norsk prosentil	Fransk kar,	Fransk prosentil	Norsk kar,	Norsk prosentil
19,8	99,9341	5,95	99,9289	19,8	99,9341	5,95	99,9289
19,9	99,9454	5,96	99,9536	19,9	99,9454	5,96	99,9536
20,0	100,000	6,00	100,000	20,0	99,9625	5,97	99,9546
				20,1	99,9726	5,97	99,9546
				20,2	99,9800	6,00	100,000
				20,3	99,9861	6,00	100,000
				20,4	99,9892	6,00	100,000
				20,5	99,9928	6,00	100,000
				20,6	99,9953	6,00	100,000
				20,7	99,9972	6,00	100,000
				20,8	99,9987	6,00	100,000
				20,9	99,9991	6,00	100,000
				21,0	99,9995	6,00	100,000
				21,1	99,9998	6,00	100,000
				21,2	99,9999	6,00	100,000
				21,3	100,000	6,00	100,000
*Norsk statistikk med 409 493 karaktersnitt, fransk statistikk fra 2014–2020 og prosentilskalaen går som i kapittel 3.4 for enkelhetsskyld opp til 100 og ikke 101.							

Lager vi omregningstabeller på denne måten ser vi at med en karakterskala som går opp til 21,3 vil franske elever med snittkarakter 20,0 og 20,1 få oversatt disse til norsk snittkarakter 5,97 og at alle med høyere snittkarakter vil få norsk karaktersnitt 6,00.

Hvordan dette skal håndteres vil avhenge av flere ting. Er det relativt få elever med et karaktersnitt større enn største mulige snittkarakter og det er naturlig å tro at en slik karaktermåling bygger på en misforståelse av bruken av karakterskalaen kan det synes fornuftig å trukkene disse til den største verdien i karakterskalaen. Hva som oppleves som rettferdig for elevene som faktisk har oppnådd høyeste karaktersnitt er også noe å ta hensyn til.

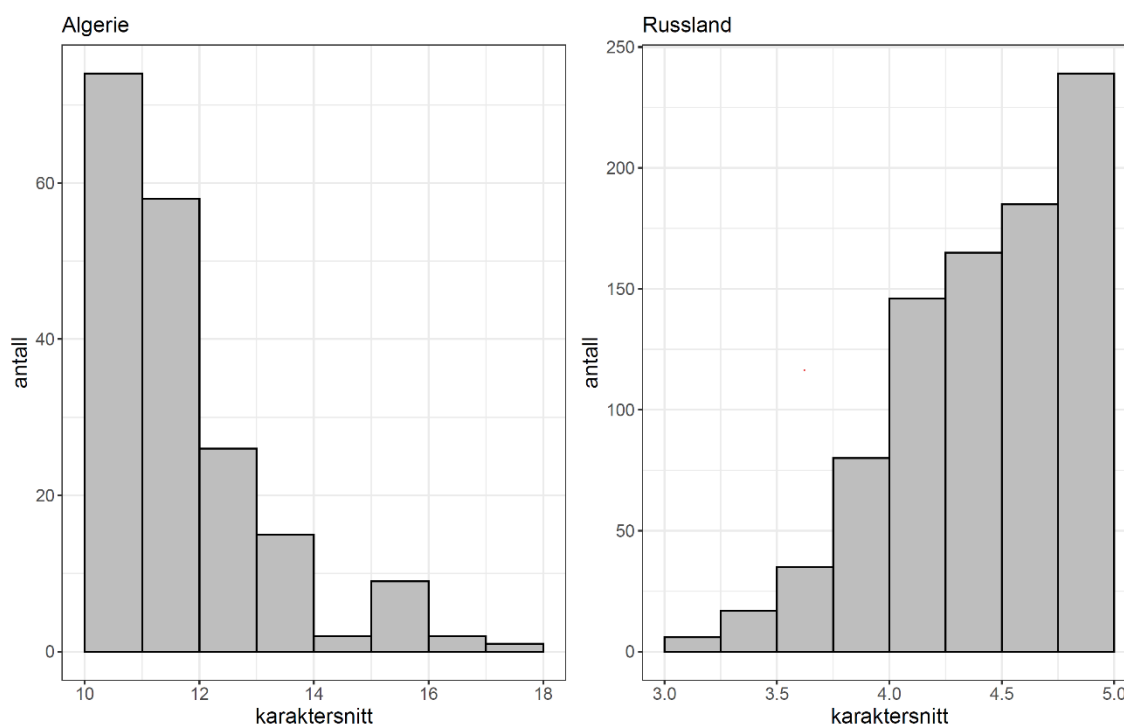
6 Analyse 4: Hvordan håndtere statistikk fra land der mange får laveste eller høyeste karaktersnitt?

I noen land får en stor andel elever laveste eller høyeste karaktersnitt. Dette var tema i notatet SAMBA/10/14 og her ser vi på dette spørsmålet igjen. Vi skal se på karakterstatistikk for Algerie, som har en bunntung karakterfordeling, og Russland, som har en topptung karakterfordeling.

Algerie har en karakterskala der de beståtte karakterene går fra 10 til 20 (best). Snittkarakter rapporteres med to desimaler. Karakterstatistikken er basert på

egeninnsamlede data, der norske og svenske tall kombineres. Totalt har vi fått tilgang til 187 elevers karaktersnitt, og blant disse har hele 21 elever, 11 %, et karaktersnitt som er nøyaktig likt 10. Karakterfordelingen er klart bunntung, med en spesielt stor opphopning av karaktersnitt nøyaktig lik 10, se venstre panel i figur 16.

Russland har en karakterskala der de beståtte karakterene går fra 3 til 5 (best). Snittkarakter rapporteres med to desimaler. Karakterstatistikken er basert på egeninnsamlede data. Totalt har vi fått tilgang til 873 elevers karaktersnitt, og blant disse har hele 114 elever, 13 %, et karaktersnitt som er nøyaktig lik 5. Karakterfordelingen er klart topptung, med en spesielt stor opphopning av karaktersnitt nøyaktig lik 5, se høyre panel i figur 16.



Figur 16. Histogrammer over algerisk og russisk karakterstatistikk.

6.1 Fem alternative metoder

Vi skal se på noen ulike omregningsmetoder, og konsentrere oss om hvordan disse regner om topp- og bunnkarakterene. Vi skal se på fem metoder listet opp i tabell 19.

Tabell 19. Metoder for å håndtere bunn- og topptung skala.

	Navn på	Kort beskrivelse av hvordan metoden ordner bunn/topp
A	Dagens	Metoden som er i bruk nå. Bruker mediankaraktersnitt i topp og/eller bunn.
B	Enkel	Bruker bare prosentilene direkte, justerer ikke endene av skalaen.
C	Gammel	Metoden som ble brukt før, bruker gjennomsnittskaraktersnitt i topp og/eller bunn.
H	ecdf	Bruker den empiriske kumulative fordelingsfunksjonen.
I	ecdf interpolert	Bruker en lineær interpolering av den empiriske kumulative fordelingsfunksjonen.

Som ellers i dette notatet, består Metode A i å finne de utenlandske hundreprosentilene og avstemme disse med norske hundreprosentiler for å komme frem til en norsk omregnet karaktersnitt for hvert utenlandske karaktersnitt. Når den utenlandske karakterfordelingen regnes som bunn- eller topptung justeres omregningen for henholdsvis bunn- og toppkarakteren. La oss ta Algerie som eksempel på det førstnevnte. Som nevnt får 11,22 % av elevene snittkarakteren 10, og den snittkarakteren havner dermed på 12-prosentilen med nåværende metodikk. Den norske 12-prosentilen gir karakteren 3,29, og dermed burde egentlig elever fra Algerie med snitt på 10 få norsk omregnet karaktersnitt på 3,29. Isteden får de en omregnet karaktersnitt som tilsvarer *medianen i bunnen av den norske karakterfordelingen*, som i dette tilfellet blir 3,08. Bunnen av fordelingen defineres som alle norske karaktersnitt som er lavere enn 12-prosentilen, 3,29.

Russiske snittkarakterer i intervallet mellom 4,96 og 5,00 tilsvarer 101-prosentilen. Uten justering ville disse snittkarakterene blitt omregnet til norsk karaktersnitt 6,00. Med Metode A, får de isteden karakteren 5,35, som tilsvarer medianen i toppen av den norske karakterfordelingen. Her er toppen av fordelingen definert som alle norske karaktersnitt som er høyere enn 5,04, som tilsvarer den russiske 87-prosentilen som er den neste prosentilen i den russiske prosentiltabellen (under 101).

Metode B bruker prosentilene direkte og endrer ikke omregningen for hverken topp- eller bunnkarakteren. Metode C er en eldre variant av Metode A (som var den daværende metoden i SAMBA/10/14) der man istedenfor medianen bruker gjennomsnittet mellom 2,00 og 3,08 for det algeriske eksempelet, og mellom 5,04 og 6 for det russiske. Metodene H og I er beskrevet tidligere i dette notatet.

Resultatet av omregningen fra utenlandsk til norsk karaktersnitt vises i tabellene 20 og 21. Fargede tall angir snittkarakterer der de ulike metodene avviker fra Metode C. Metode A, B og C er kun forskjellige for den aller laveste karakteren i det algeriske eksempelet, og for karakterer over 4,96 i det russiske eksempelet. Metode H og I skiller seg i større grad fra de andre metodene.

For Algerie gir Metode H «ecdf» den høyeste norske omregnede karakteren for den algeriske snittkarakteren 10, Metode B «enkel» er ganske nærme, Metode I og A gir noe lavere karakterer og Metode C gir klart lavest norsk karakter.

Tabell 20. Omregning fra opprinnelig algerisk karaktersnitt til norsk karaktersnitt (i bunnen).					
Algerie	Metode A	Metode B	Metode C	Metode H	Metode I
10,00	3,08	3,25	2,67	3,29	3,05
10,01	3,29	3,29	3,29	3,29	3,30
10,02	3,29	3,29	3,29	3,29	3,30
10,03	3,29	3,29	3,29	3,32	3,30
10,04	3,32	3,32	3,32	3,32	3,32

I det tilsendte bakgrunnsmateriale fra SO ser vi at metoden for å omregne russiske karaktersnitt er litt annerledes enn det som gjøres når de algeriske karaktersnittene omregnes. Metode A i tabell 21 er derfor ikke helt identisk med Metode A i tabell 20. Som man kan se av tabellen under, endres ikke kun omregningen for toppkarakteren

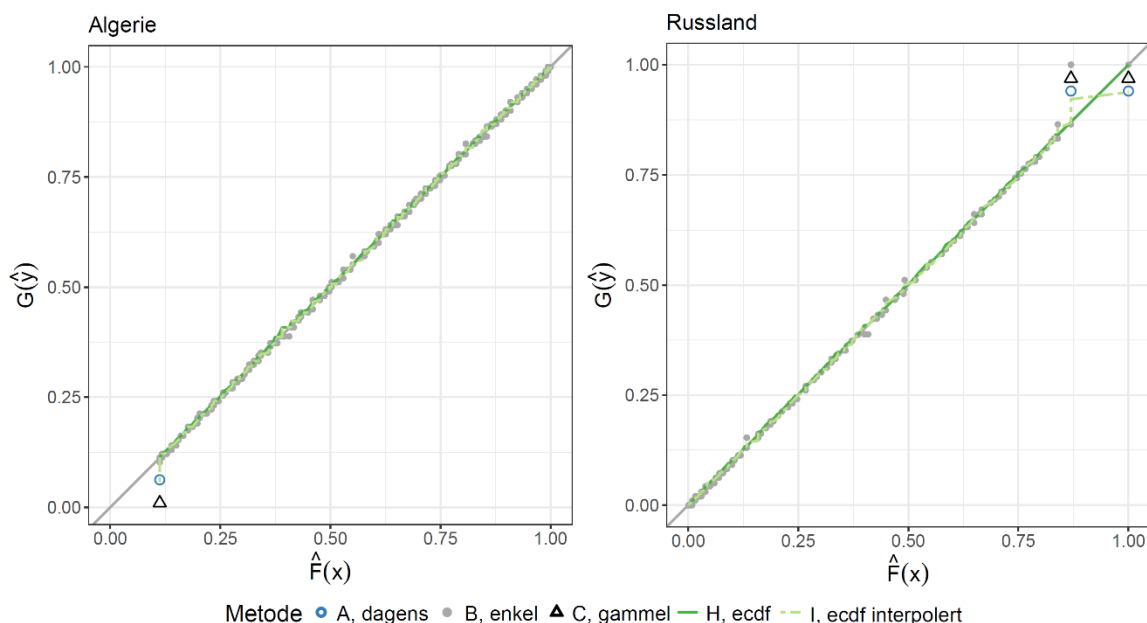
5,00, men hele intervallet som h rer til 101-prosentilen (4,96 til 5,00). Metode B tilegner den norske toppkarakteren 6 til alle russiske karakterer mellom 4,96 og 5,00, mens Metode H kun tilegner 6,00 til den russiske toppkarakteren 5,00. De andre metodene gir et maksimal norsk omregnet karaktersnitt som er lavere enn 6,00: Metode A og I gir lavest omregning, mens Metode C gir en middels h y norsk omregnet karakter. Metode I differensierer mer mellom de russiske toppkarakterene enn de andre metodene.

Tabell 21. Omregning fra opprinnelig russisk karaktersnitt til norsk karaktersnitt (i toppen).

Russland	Metode A	Metode B	Metode C	Metode H	Metode I
4,94	5,04	5,04	5,04	4,96	5,00
4,95	5,04	5,04	5,04	5,07	5,05
4,96	5,35	6,00	5,52	5,07	5,10
4,97	5,35	6,00	5,52	5,07	5,15
4,98	5,35	6,00	5,52	5,07	5,20
4,99	5,35	6,00	5,52	5,07	5,26
5,00	5,35	6,00	5,52	6,00	5,33

6.2 Kalibrering

Vi kan sammenligne de fem metodene med tanke p  kalibrering for disse eksemplene ogs . Figur 17 tyder p  at alle metodene er rimelig godt kalibrert her. Metode A og C skiller seg fra B kun i henholdsvis bunnen (for Algerie) og toppen (for Russland).



Figur 17. Kalibreringsplot for omregningsmetoder fra algeriske til norske karaktersnitt (venstre), og fra russiske til norske karaktersnitt (h yre).

Man  nsker seg fortsatt omregningsmetoder med god kalibrering og differensiering, men i denne problemstillingen er vi motivert ut ifra andre hensyn: ettersom et karaktersnitt lik 6 er s v rt uvanlig i den norske statistikken, s  har det syntes SO

urimelig at for eksempel over 13 % av russiske søkere til Norge skulle få 6 som norsk omregnet karakter.

Et annet, men noe mindre viktig, hensyn man kan tenkes å bry seg om her, er at man helst skulle unngå store hopp i de norske omregnede karakterene. Det kan for eksempel virke negativt at Metode A gir norsk karaktersnitt på 3,08 for den algeriske snittkarakteren 10, mens en elev med et marginalt høyere snitt på 10,01 vil få en langt høyere norsk omregnet karaktersnitt på 3,29.

Vi vil presisere at det ikke nødvendigvis er slik at man må behandle en bunntung og topptung utenlandsk karakterfordeling på samme måte.

Rimelighetshensynet som ble løst formulert over kan gi støtte for at utenlandske karaktersystemer med en svært topptung fordeling bør få en noe nedjustert omregning i toppen, men medfører det at man skal nedjustere omregningen i bunnen dersom det utenlandske karaktersystemet er bunntungt? Kanskje ikke: 11 % av algeriske elever får snittkarakteren 10, 11 % av norske elever får snittkarakteren 3,29 eller lavere, mens kun 6 % av norske elever får snittkarakteren 3,08 eller lavere – bør en algerisk elev med 10 i snitt få 3,29 eller 3,08? Men på den andre siden: 10 er laveste ståkarakter, mens 3,29 er en ganske høy norsk karakter.

6.3 Oppsummering av analyse 4

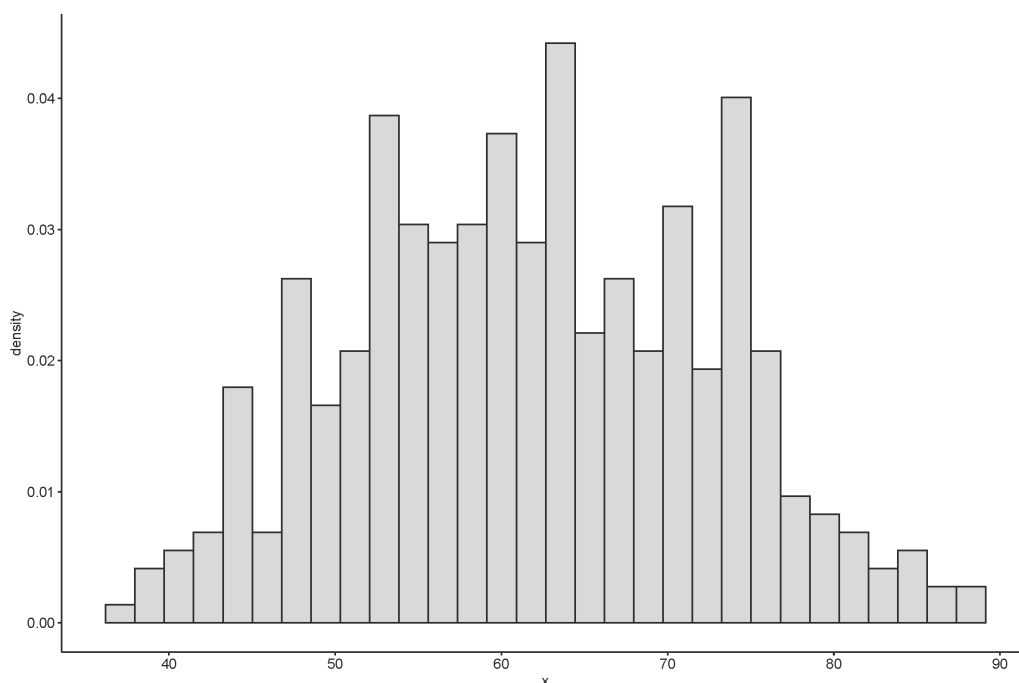
Vi har studert karakterstatistikk for Algerie, med sin bunntunge karakterfordeling, og Russland, med sin topptunge karakterfordeling, og sammenlignet fem ulike omregningsmetoder. Basert på dette mener vi at:

- Ved topptung karakterfordeling bør øverste del av omregningstabellen justeres. Dagens metode ser ut til å fungere greit, men et alternativ er metode I, i og med at vi da får en mere jevn omregning i øverste del av karakterskalaen.
- I likhet med å justere en topptung karakterfordeling kan det være naturlig å også justere en bunntung karakterfordeling. Det er ikke automatisk gitt at topp og bunn justeres på samme måte. Hva som gjøres vil typisk være en avveining av ulike hensyn som skal tas. Hvis SO ikke ønsker å justere bunnen kan de bruke Metode H «ecdf», mens hvis SO derimot ønsker å justere bunnen kan de bruke Metode I (eller dagens metode).

7 Analyse 5: Hva gjør man når egeninnsamlet statistikk ikke omfatter alle karakterer i skalaen?

Noen land produserer ikke offisiell karakterstatistikk og SO må da legge til grunn egeninnsamlet statistikk når omregningstabellen skal lages. Antallet utenlandske karaktersnitt vil variere, men uansett vil det være ett utvalg, og typisk vil dette utvalget ikke inneholde alle mulig karaktersnitt som kan oppnås. Hvilke karaktersnitt på karakterskalaen som det er sannsynlig at ikke er representert i den egeninnsamlede statistikken vil avhenge av den underliggende karakterfordelingen. Som vi har sett i SAMBA/10/14 og i dette notatet kan ulike lands karakterfordelinger variere; noen er tilnærmet klokkeformede, som den norske, mens andre er skjeve med lang hale, som den franske og andre igjen kan være mer flate i fordelingen, som den tyske.

I noen egeninnsamlede statistikker mangler karaktersnitt i bunn og/eller topp-sjiktet, noe som innebærer at karakterspennet i laveste og/eller høyeste prosentil kan bli stort slik det gjøres i dag. Dette gjelder for eksempel egeninnsamlet statistikk fra Pakistan. Karakterskalaen i Pakistan går fra laveste beståtte karaktersnitt på 30 til høyeste på 100, og karaktersnittene gis med 2 desimaler. I den egeninnsamlede statistikken, som består av 410 karaktersnitt, er laveste karaktersnitt 36,45 og høyeste 87,64. Histogrammet i figur 18 viser fordelingen av egensamlede karakterer.



Figur 18. Pakistansk karakterstatistikk.

Vi ser på tre mulige måter å håndtere dette på når vi lager omregningstabellen (alternativ 1–3):

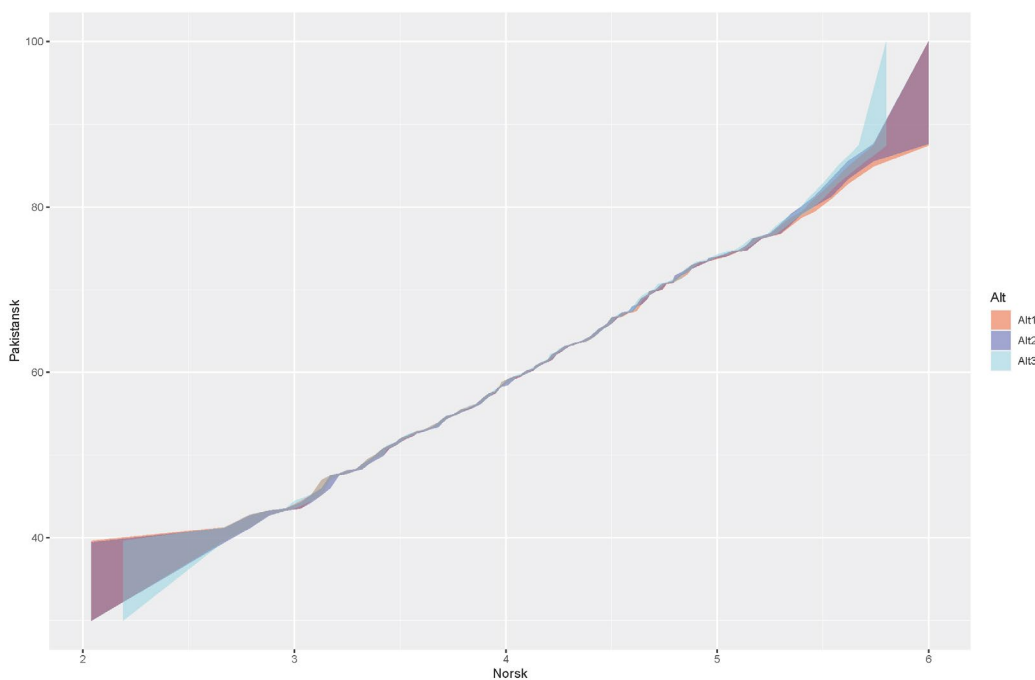
1. La innsamlet minimums- og maksimums-verdi representere faktisk minimums- og maksimums-verdi.
2. Legge inn fiktiv minimums- og maksimums-verdi
3. Se på tilsvarende andel av den norske statistikken og så lage omregningstabell
 - a. Hva er avstand fra laveste innsamlede karaktersnitt til minimumskarakter?
 - b. %-vis del av skalaen?
 - c. Hva er avstand fra største innsamlede karaktersnitt til maksimumskarakter?
 - d. %-vis del av skalaen?

- e. Kutte tilsvarende (%-vis) av norsk skala og så lage omregningstabellen slik SO har gjort tidligere.

Som i analyse 3 legger vi til grunn den norske statistikken med 409 493 karaktersnitt. Vi legger inn i etterkant at laveste intervall for pakistanske karaktersnitt skal være 30,00 og høyeste 100,00 slik at alle søkere med pakistanske vitnemål blir plassert på skalaen.

Tabell 22 viser omregningstabeller for de tre alternativene for de minste og største prosentilene, og figur 19 viser norske karaktersnitt versus øvre og nedre grense i de tilsvarende pakistanske karaktersnittintervallene for de tre omregningsalternativene.

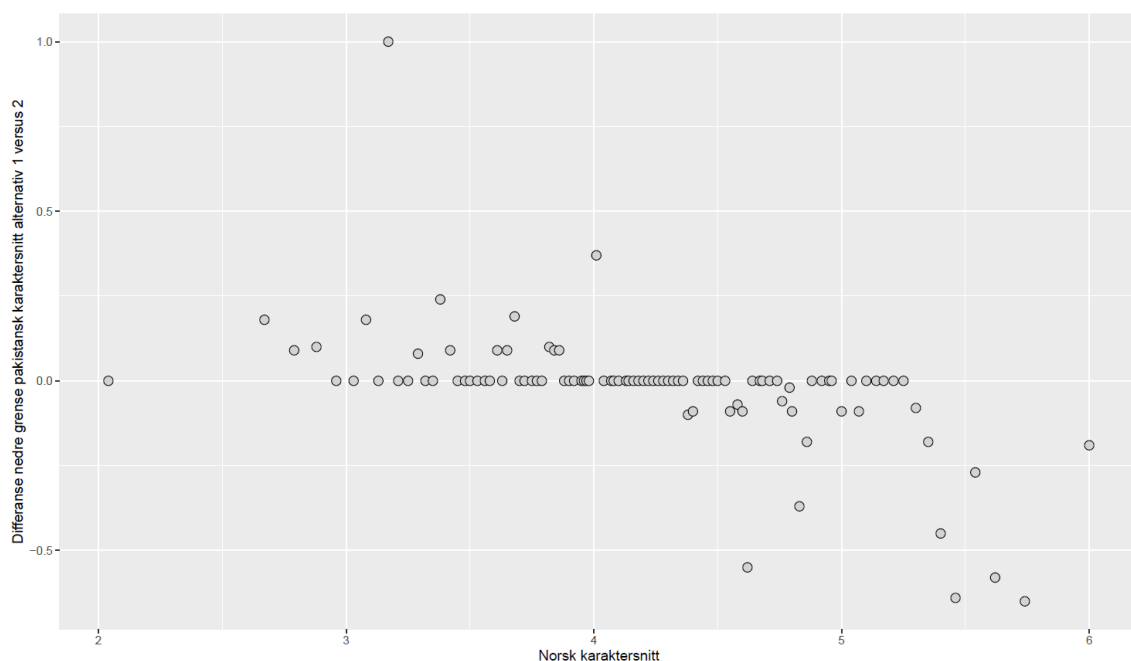
Tabell 22: Resultater, tre alternativer; prosentiler, pakistansk karaktersnitts nedre og øvre grense (P _N og P _Ø) og tilhørende norske karaktersnitt (N)									
	Alternativ 1			Alternativ 2			Alternativ 3		
Prosentil	P_N	P_Ø	N	P_N	P_Ø	N	P_N	P_Ø	N
1	30,00	39,63	2,04	30,00	39,45	2,04	30,00	39,63	2,19
2	39,64	41,27	2,67	39,46	41,18	2,67	39,64	41,27	2,67
3	41,28	42,80	2,79	41,19	42,70	2,79	41,28	42,80	2,79
4	42,81	43,30	2,88	42,71	43,30	2,88	42,81	43,30	2,88
5	43,31	43,55	2,96	43,31	43,55	2,96	43,31	43,55	2,96
6	43,56	44,45	3,03	43,56	44,27	3,03	43,56	44,45	3,00
7	44,46	45,20	3,08	44,28	45,20	3,08	44,46	45,20	3,08
8	45,21	47,00	3,13	45,21	46,00	3,13	45,21	47,00	3,13
9	47,01	47,54	3,17	46,01	47,54	3,17	47,01	47,54	3,17
10	47,55	47,73	3,21	47,55	47,73	3,21	47,55	47,73	3,21
...									
92	76,19	76,45	5,21	76,19	76,45	5,21	76,19	76,45	5,20
93	76,46	76,73	5,25	76,46	76,81	5,25	76,46	76,73	5,24
94	76,74	77,73	5,30	76,82	77,91	5,30	76,74	77,73	5,28
95	77,74	78,73	5,35	77,92	79,18	5,35	77,74	78,73	5,33
96	78,74	79,45	5,40	79,19	80,09	5,40	78,74	79,45	5,38
97	79,46	81,00	5,46	80,10	81,27	5,46	79,46	81,00	5,43
98	81,01	82,82	5,54	81,28	83,40	5,54	81,01	82,82	5,50
99	82,83	84,90	5,62	83,41	85,55	5,62	82,83	84,9	5,57
100	84,91	87,45	5,74	85,56	87,64	5,74	84,91	87,45	5,67
101	87,46	100,00	6,00	87,65	100,00	6,00	87,46	100,00	5,80



Figur 19. Norsk karaktersnitt versus øvre og nedre grense i det tilsvarende pakistanske karaktersnittintervallet for de tre omregningsalternativene.

Alternativ 1 og 2 gir ganske like resultater. Forskjellene er litt større for høye karaktersnitt, og dette er naturlig siden avstanden fra minste observerte verdi (36,45) til minste mulige verdi (30,00) relativt liten, mens tilsvarende avstand i toppen av skalaen er litt større (87,64 versus 100,00). Dette gjelder spesielt for alternativ 3 versus de to andre alternativene. Med alternativ 1 får elever med de høyeste karaktersnitt litt bedre uttelling enn med alternativ 2 som forventet. Med alternativ 3 kreves det igjen litt bedre pakistansk snittkarakter for å få tilsvarende norsk karaktersnitt sammenlignet med de to første alternativene. Samtidig ser vi det motsatte i bunnen av skalaen, selv om forskjellene der er mindre.

For en norsk snittkarakter kan vi plote differansen mellom tilhørende minste pakistanske snittkarakter når vi legger metodene i alternativ 1 versus 2 til grunn, og figur 20 viser dette.



Figur 20. Norsk karaktersnitt versus differensen av nedre grense i det tilsvarende pakistanske karaktersnittintervallet for metodealternativ 1 versus 2.

I den komplette tabellen som tabell 22 viser et utdrag fra finner vi at for et norsk karaktersnitt på 4,00 eller lavere må pakistanske søkere ha litt bedre karaktersnitt med alternativ 1 sammenlignet med alternativ 2 og motsatt for høyere norske karaktersnitt. For eksempel, med alternativ 1 får søkere et norsk karaktersnitt 5,00 med et (minimum) karaktersnitt fra Pakistan på 73,74, mens med alternativ 2 er tilsvarende tall 73,88, og videre med alternativ 3 er tilsvarende tall 74,01.

I en karakterstatistikk som inneholder karaktersnitt enda lengre unna lave og/eller høye verdier på skalaen vil vi kunne se større forskjeller. Det bør nevnes at i prinsippet har vi samme problemstilling for den norske statistikken, der vi ikke har observert karaktersnitt under 2,04, mens skalaen altså går helt ned til 2,00.

Hvordan utfordringen med manglete karaktersnitt i endene av karakterskalaen i produksjonen av omregningstabeller skal håndteres er ikke opplagt. Vi har her vist hva ulike omregningsmetoder kan bety for omregningstabellen for pakistanske karaktersnitt. I dette tilfellet gav det ikke store utslaget å ignorere at skalaen faktisk slutter på 100,00 og ikke 87,64. Men hva hvis statistikken kun innehold karaktersnitt mellom 45 og 65? Da hadde det kanskje ikke vært så enkelt å ikke ta hensyn til at skalaen faktisk begynner på 30,00 og slutter på 100,00. Dette er nok en vurdering som bør gjøres i hvert enkelt tilfelle, og hvordan man håndterer dette vil nok også avhenge av andre hensyn å ta, som hvor mange søkere som vil bli plassert relativt lavt eller høyt på den norske skalaen sammenlignet med andre lands søkere.

8 Oppsummering

Vi har i dette notatet tatt for oss for noen problemstillinger som oppstår i produksjonen av omregningstabeller. Vi har sett på endringer i karakterfordelinger over årene 2014–2020 for franske og norske karaktersnitt. Vi har diskutert hvordan såkalte «store hull» i

utenlandsk statistikk kan håndteres, basert på statistikk fra Etiopia og International Baccalaureate (IB). Vi har studert fransk statistikk med tanke på utfordringer knyttet til det å ha karaktersnitt større enn største mulige verdi på karakterskalaen. Vi har diskutert metoder for å oversette karakterstatistikker som er topptunge eller bunntunge, vist med statistikk fra Algerie og Russland. Til slutt har vi sett på utfordringer knyttet til å skulle omregne utenlandsk statistikk der deler av toppen og/eller bunnen av karakterskalaen mangler, som er tilfellet for den pakistanske statistikken.

Noen av disse problemstillingene har vært matematiske, og resultatene av analysene har kunnet peke i en (eller flere) tydelige retninger. Andre utfordringer har vi kunnet belyse analytisk uten at det nødvendigvis peker seg ut en tydelig løsning. Arbeidet gir grunn for metodiske anbefalinger og videre noen praktiske råd.

8.1 Anbefalte metoder

«Store hull»

- Med lignende statistikk som for Etiopia (egeninnsamlet statistikk, fininndelt karakterskala) kan varianter av ecdf-metoden, som «ecdf interpolert» eller lignende, være gode valg.
- ecdf-metoden vil per konstruksjon alltid være best med tanke på kalibrering, men kan gi en ganske lav grad av differensiering (når ecdf-en er basert på egeninnsamlet statistikk).
- Hvis det er ønskelig å fortsette å bruke hundreprosentiltabellene er det trolig bedre å bruke metoder som endrer grensene rundt prosentilhullene i større grad enn dagens metode (som for eksempel «begge sider, 2 steg»).
- Med lignende statistikk som for IB (offisiell karakterstatistikk og grov skala) er det ikke nødvendig å «fylle hullene». For IB er det best å kun bruke prosentilene direkte (Metode B «Enkel») eller å bruke ecdf-metoden (gjerne med noe spesielt for bunnen/toppen av skalaen).

Topptung/bunntung fordeling

- Ved topptung karakterfordeling bør øverste del av omregningstabellen justeres. Dagens metode kan fungere greit, men et alternativ er Metode I (ecdf interpolert), i og med at vi da får en mer jevn omregning i øverste del av karakterskalaen.
- Hvis SO ønsker å justere bunnen i en bunntung fordeling kan Metode I (ecdf interpolert) eller dagens metode brukes, mens hvis SO ikke ønsker å justere bunnen kan Metode H (ecdf) brukes.

8.2 Anbefalte rutiner

Endringer i karakterfordelinger over tid

- Det kan være nyttig å rutinemessig følge med på lands statistikker over tid og undersøke hva det betyr å inkludere eller ikke inkludere et ekstra års statistikk når en beregner oversettelsestabellene.

- Er det endringer i karakterfordelingene over tid bør det vurderes om hvorvidt søkere med eldre vitnemål skal plasseres på den norske karakterskalaen etter en særskilt (eldre) omregningstabell.

Manglende lave og høye karakterer og karaktersnitt over størst mulige verdi

- Hvordan manglende lave og høye karaktersnitt i statistikken skal håndteres er en vurdering som bør gjøres i hvert enkelt tilfelle, og løsningen vil nok også avhenge av andre hensyn, som hvor mange søkere som vil bli plassert relativt lavt eller høyt på den norske skalaen sammenlignet med andre lands søkere.
- Er det relativt få elever med et karaktersnitt større enn største mulige snittkarakter og det er naturlig å tro at en slik karaktergivning bygger på en misforståelse av bruken av karakterskalaen kan det synes fornuftig å trunkere disse til den største verdien i karakterskalaen.

Avslutningsvis gjentar vi at en alternativ måte å oversette utenlandske karaktersnitt på kan være å ta utgangspunkt i den empiriske kumulative fordelingsfunksjonen til den norske og den utenlandske statistikken. Dette kan undersøkes nærmere for flere lands statistikker. Som vi har sett i dette notatet så vil et slikt metodevalg påvirke alle problemstillingene vi har diskutert.