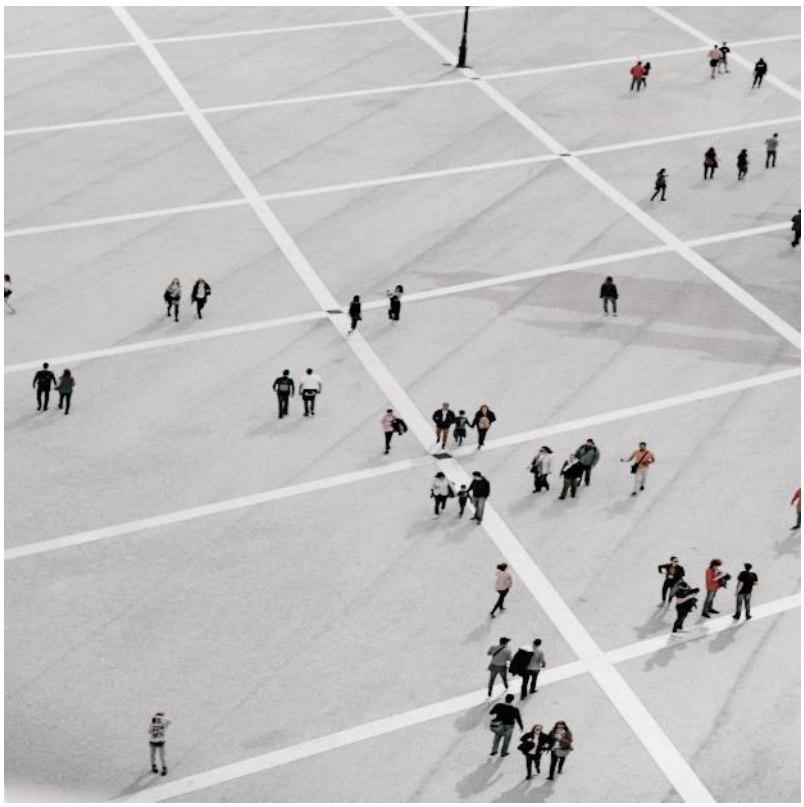
Epinion

Forcensur på folkeskolens prøver

Slutrapport

28. februar 2021



Indholdsfortegnelse

1. Indledning 1
1.1 Resultater af forcensur 2020 og 2021 1
1.2 Fremadrettet plan for forcensur – opdateret drejebog 1
2. Forcensur 2020 2
2.1 Rekruttering af skoler til deltagelse i forcensur 2
2.2 Oparbejdning af opgavesæt og forslag om øveprøver 2
3. Forcensur 2021 4
3.1 Rekruttering af skoler til deltagelse i forcensur 4
3.2 Oparbejdning af opgavesæt 4
3.3 Forcensur i dansk, læsning og retskrivning 5
3.3.1 Datagrundlaget for beregning af omsætningstabeller 5
3.3.2 Ækvivalering og valg af metode 6
3.3.3 Ækvivalering med afsæt i data fra forcensurskoler overfor øveprøver 6
3.3.4 Omsætningstabeller og karakterfordeling 8
3.4 Forcensur i matematik med og uden hjælpemidler 9
3.4.1 Datagrundlaget for beregning af omsætningstabeller 9
3.4.2 Ækvivalering 9
3.4.3 Omsætningstabeller og karakterfordeling 10
3.5 Præsentationer for opgavekommissioner 11
4. Metoder til ækvivalering 13
4.1 Equirank-metoden 13
4.2 Modelmetoden 14
4.3 Konklusion vedr. valg af ækvivaleringsmetode 15
5. Bedømmelse af prøverne i naturfag 17
5.1 Baggrund for omlægning af omsætningstabeller i naturfagene 17
5.2 Analyser af prøvedata med afsæt i maj 2019-prøven 18
5.2.1 Score- og karakterfordelinger for de skriftlige prøver i STEM-fagene 18
5.2.2 Rekalibrering af skalaerne i de tre skriftlige naturfagsprøver 19
5.2.3 Sammenstilling af omsætningstabeller ud fra de 2 principper 20
5.3 Beregnet omsætningstabel med afsæt i 2021-prøven 21
5.3.1 Besvarelser af øveprøver 2021 som datagrundlag 21
5.3.2 Ækvivalering på grundlag af maj 2021-prøverne 22
5.3.3 Ækvivalering på grundlag af tilpassede forcensursæt 24
6. Bedømmelse af prøverne i fremmedsprog 27

6.1 Udvikling i de gennemsnitlige karakterer i sprogfagene 28
6.2 Bedømmelse i sprogprøver 29
6.3 Konsekvenser af tre delprøver i scoringsmodel 30
6.4 En samlet prøve overfor flere delprøver, engelsk, maj 2019 31
6.4.1 Faktoranalyse af 2019-prøven i engelsk 31
6.4.2 Reliabilitet i bedømmelsen af engelsk i 2019 34
6.4.3 Konklusion på analyserne af engelsk 2019-besvarelserne 36
6.5 Grundlag for forcensur i sprogfagene 36
6.5.1 Datagrundlaget for beregning af omsætningstabeller 37
6.5.2 En samlet omsætningstabel for engelsk, maj 2017 38
6.5.1 En fælles sproglig dygtighed 40
6.6 Ækvivalering af tysk og fransk på grundlag af engelsk 42
6.6.1 Ækvivalering af tysk 42
6.6.2 Ækvivalering af engelsk og fransk 43
6.6.3 Modelmetoden og omsætningstabeller for tysk og fransk 44
6.6.4 Karakterfordelinger 47
7. Forcensur for prøverne i 10 klasse 48
7.1 Datagrundlag, dansk 48
7.1.1 Muligheder for at ækvivalere niveau i dansk FP9 og FP10 48
7.1.2 Dansk, maj 2021 som fremtidigt forcensursæt 48
7.2 Datagrundlag, matematik 49
8. Anbefalinger 50
8.1 Anbefalinger mht. de bundne prøver 50
8.2 Anbefalinger vedrørende skriftlige udtrækprøver, FP9 51
8.3 Generelle anbefalinger vedrørende prøver og opgaver 51
9. Bilag 54
9.1 Dokumentation af elevbesvarelser 54
9.2 Tidligere fremsendte dokumenter 54
9.3 Præsentationsmateriale til opgavekommissionsmøderne 54
9.4 Drejebog 55
9.5 Vejledningsmateriale 55
9.6 Responsanalyser til opgavekommissioner 55

1. Indledning

Forcensurens hovedformål er at sikre, at eleverne ligebehandles ved den bedømmelse, de får efter besvarelse af især de afsluttende selvrettende prøver. Variation i opgavesættenes sværhed fra år for år gør det nødvendigt at justere de omsætningstabeller, der består af grænseværdier for de pointintervaller, der ligger til grund for karaktergivningen. På den måde sikres en retfærdig og ensartet bedømmelse af elevernes præstationer uanset årgang og prøvens sværhed.

Projektet er en opfølgning og udbygning af det arbejde med forcensur der blev gennemført i 2017-2019 og som er beskrevet i rapporten Evaluering af forcensur 2017-19 (jf. Bilag 9.2.2)

Formålet med denne rapport er at dokumentere projektaktiviteterne i forbindelse med forcensuren i 2020-2021. Først præsenteres resultaterne af forcensuren 2020-2021 og de benyttede analysemetoder til ækvivalering. Herefter adresseres bedømmelse af prøverne i naturfag, dernæst prøverne i fremmedsprog og slutteligt beskrives en kort analyse af muligheder for forcensur i 10. klasse. Afslutningsvist gives projektets anbefalinger til det videre arbejde med afsæt i erfaringer og resultater fra forcensur i perioden 2017-2021. I bilagene er bl.a. vedlagt datamateriale, drejebog til forcensur med udvalgte vejledningsmaterialer samt præsentationer til opgavekommissionerne.

1.1 Resultater af forcensur 2020 og 2021

I 2020 er der arbejdet med rekruttering af kerne- og udtræksskoler og oparbejdning af prøvesæt i Epinions prøvesystem. De skriftlige prøver FP9 (bundne og udtræk) er i 2020 aflyst, hvorfor gennemførsel af forcensurprøver og ækvivalering ikke er foretaget.

I 2021 er gennemført FP9 i de bundne skriftlige prøver, dansk læsning og retskrivning samt matematik med og uden hjælpemidler. Desuden er der rekrutteret skoler til gennemførsel af øveprøver i udtræksfag (naturfag og fremmedsprog) til brug for øvrige analyser (kapitel 5 og 6). Der er foretaget ækvivalering af de fire prøver, og forslag til omsætningstabeller er leveret til opgavekommissionerne.

1.2 Fremadrettet plan for forcensur – opdateret drejebog

Der er foretaget en opdatering af den eksisterende drejebog for forcensur. Drejebogen er vedlagt som Bilag 9.4 og beskriver detaljeret de aktiviteter, der hidtil er foretaget ifm. forcensur, samt forslag til fremadrettet proces. Drejebogen indeholder:

- **Forcensurens formål og forløb,** herunder**:**
  - Baggrund
  - Hidtidige formål med forcensuren
  - Forcensurens normale årsforløb
- **Rekruttering og vejledning af skoler,** herunder:
  - Incitamenter og forpligtelser
  - Skoletyper, herunder kerneskoler, udtræksskoler og franskskoler
  - Vejledningsmateriale og kommunikation
  - Måltal for rekruttering
  - Rapportering af resultater
  - Særlig proces for forcensur af ikke-selvrettende prøver
  - Forslag til proces for forcensur i 10. klasse
- **Analyse og ækvivalering,** herunder**:**
  - Analysens forløb

  - Datarensning
  - Ækvivalering og forslag til omsætningstabeller
  - Overdragelse til opgavekommissionerne
- **Eksempel: Formidling af resultater til opgavekommissionen i dansk**

2. Forcensur 2020

Da de skriftlige prøver i 2020 blev aflyst som følge af COVID19 og nedlukning af skoler, blev der ikke gennemført forcensurprøver og ækvivalering af 2020-prøverne. Beslutningen om aflysning af prøverne kom så sent i skoleåret, at praktisk taget alle forberedelserne af forcensuren i 2020 var blevet gennemført.

2.1 Rekruttering af skoler til deltagelse i forcensur

Som forberedelse til forcensur 2020 blev gennemført rekruttering af skoler til deltagelse i forcensurprøven forud for FP9. Rekruttering bestod i hovedtræk af:

- **Genrekruttering/fastholdelse af skoler, der tidligere har deltaget i forcensur:** Herunder har der både været kontakt til tidligere *kerneskoler* (der deltager med alle elever i alle fag) og tidligere *udtræksskoler* (deltager i udtræksfag) med henblik deltagelse som kerneskole.
- **Rekruttering af udtræksskoler:** Målrettet rekruttering af skoler med mange elever, der har udtrukket et pågældende udtræksfag, herunder naturfag, engelsk, tysk og fransk.
- **Rekruttering af franskskoler:** Målrettet rekruttering af skoler med mange franskelever med henblik på både 9. og 8. klassernes deltagelse i forcensur.

Oversigten nedenfor viser det rekrutterede antal elever til forcensur 2020.

Tabel 1. Antallet af rekrutterede elever til forcensur 2020.

Prøve (2017+2020)	Antal elever
Biologi	641
Geografi	520
Fysik/kemi	509
Engelsk (læsning, lytning, sprog og sprogbrug)	449
Tysk (læsning, lytning, sprog og sprogbrug)	352
Fransk (læsning, lytning, sprog og sprogbrug), 9. klasse	54
Fransk (læsning, lytning, sprog og sprogbrug), 8. klasse	25
Dansk, læsning	1.101
Dansk, retskrivning	1.101
Matematik uden hjælpemidler	952
Matematik med hjælpemidler	628

2.2 Oparbejdning af opgavesæt og forslag om øveprøver

For at tilbyde størst mulig fleksibilitet til de deltagende skoler har Epinion oparbejdet alle forcensurprøverne (se Tabel 1 ovenfor) i Epinions prøvesystem. De fleste prøver var med få tilpasninger klar fra 2017-19-forcensurprojektet, men enkelte forbedringer og nye prøver kom til – herunder 2017-

prøvesættene i skriftlig tysk og fransk (lytning, læsning, sprog og sprogforståelse), der som nye fag indgår i forcensuren. Prøverne er opsat, så de er visuelt og funktionelt tæt på identiske med opsætningen i Testogprøver.dk.

En enkelt justering blev foretaget af prøverne i fremmedsprog i forlængelse af dialog med opgavekommissionerne: Afholdelsen af prøverne i fremmedsprog implicerer normalt, at eleverne skal have et teksthæfte, og kun selve besvarelsen foregår på computeren – ligesom for Dansk, læsning. Da volumen af tekster imidlertid er begrænsede i fremmedsprog, kunne teksterne til opgaverne (fx i læsning) indarbejdes i selve besvarelsesapplikationen uden at det var til gene for eleverne.

Som bekendt skete der lige før igangsættelse af forcensurprocessen en omfattede nedlukning af samfundet og skolerne, inklusive hjemsendelse af eleverne. Beslutningen om at aflyse prøverne og lade årskarakterer indgå på eksamensbeviset i stedet stiller efter vores opfattelse ny og forstærkede krav til validiteten af karaktergivningen. Da Epinions prøvesystem var færdigudviklet og klar til drift, og da lærerne efterlyste muligheder for at evaluere eleverne med afsæt i prøvelignende opgavebesvarelser og samtidigt give eleverne eksamenstræning/prøveoplevelser, foreslog vi at stille Epinions prøvesystem til rådighed for skolerne, hvorpå der kunne gennemføres øveprøver som uformelle øveprøver på skolen eller som hjemmeprøver med efterfølgende detaljeret feedback til læreren om enkeltelevers præstationer. Oplægget er vedlagt i Bilag 9.2. Ud fra en samlet vurdering valgte ministeriet dog ikke at anvende løsningen.

3. Forcensur 2021

Til FP9 i 2021 er der igen foretaget omfattende aflysninger, men der blev dog gennemført 4 ud af 12 skriftlige prøver. Eleverne har gennemført de bundne prøver, dvs. skriftlig matematik (med og uden hjælpemidler) og skriftlig dansk (læsning og retskrivning). Nedenfor er beskrevet, hvilke aktiviteter der er gennemført i forbindelse med forcensuren i 2021.

3.1 Rekruttering af skoler til deltagelse i forcensur

Som forberedelse til forcensuren 2021 er der foretaget fastholdelse og rekruttering af skoler til deltagelse, som beskrevet i afsnit 2.1 og uddybet i drejebogen i Bilag 9.4.

Til forcensuren 2021 er der rekrutteret elever som vist i nedenstående tabel.

Tabel 2. Antallet af rekrutterede elever til forcensur 2021.

Prøve (i kombination med 2021-prøvesættet)	Antal elever (brutto)
Biologi 2021	742*
Geografi 2021	742*
Fysik/kemi 2021	760*
Engelsk 2017 (læsning, lytning, sprog og sprogbrug)	947*
Tysk 2021 (læsning, lytning, sprog og sprogbrug)	770*
Fransk 2021 (læsning, lytning, sprog og sprogbrug), 9. klasse	212*
Dansk, læsning 2017	1.598
Dansk, retskrivning 2017	1.494
Matematik uden hjælpemidler 2017	1.379
Matematik med hjælpemidler 2017	1.175

* Antallet af elever i udtræksfagene er opgjort inden selve udtrækket.

3.2 Oparbejdning af opgavesæt

De fire 2017-prøver i dansk og matematik er gennemført som forcensurprøver (blandt rekrutterede skoler) eller som øveprøver på STILs Testogprøver blandt øvrige skoler, der ikke har været tilmeldt forcensur. Data fra Testogprøver er efterfølgende blevet udleveret til Epinion af STIL.

Ifm. frigivelsen af 2021-prøvesættene i tysk og fransk til brug for øveprøver på Testogprøver har Epinion oparbejdet disse prøvesæt i Epinions prøvesystem for at kunne supplere dataindsamlingen og styrke datamaterialet.

Forcensurskoler (men ikke de skoler, der brugte muligheden for at gennemføre øveprøver) fik feedback (se eksempel i drejebog, Bilag 9.4) på alle prøvebesvarelser, uanset om prøven var gennemført på Epinions prøvesystem eller på Testogprøver.

Forcensurprøverne blev gennemført efter forudgående aftale med skolen om, at der skulle instrueres omhyggeligt og at eleverne ville få feedback på deres indsats, mens hovedformålet med øveprøverne var at give eleverne en mulighed for at træne op til årets prøve. Der kunne derfor antages at var mere ”på spil” (stakes) ved forcensurskolernes prøveafvikling end ved øveprøverne hvor eleverne blot skulle træne i at logge på testogprøver.dk og eksamenssituationen. Dette gav sig også udslag i, at der blandt deltagerne i øveprøverne var en del elever der besvarede ingen eller få opgaver. Nyere forskning (Attali 2016) peger på at stakes ikke spiller den målbare rolle for resultatet som det ofte antages. Men det kan betyde noget for om eleven overhovedet går ind i besvarelsen med henblik på at gøre sit bedste. Konklusionen heri er, at elever, der først går i gang med at be-

svarer opgaver, glemmer alt om stakes og gør deres bedste – uanset stakes. Men et initialt engagement er nødvendigt og derfor bør der ske en vurdering/frasortering af de elevdata, hvor eleven slet ikke går i gang. Det vurderes således at data fra forcensurskoler og fra øveprøver kan anvendes på lige fod, når blot de helt useriøse besvarelser er sorteret fra. Vi skal yderligere trykprøve denne påstand (nedenfor) ved at gennemføre separate ækvivaleringer på baggrund af henholdsvis forcensurskole- og øveprøvedata.

3.3 Forcensur i dansk, læsning og retskrivning

Mulighederne for at kombinere forcensuren med gennemførelse af øveprøver betød, at det blev muligt at indhente et meget stort datamateriale til brug for beregning af omsætningstabellerne. Der gives her en summarisk gennemgang af resultaterne, mens de mere detaljerede resultater kan findes i Præsentationsmateriale til opgavekommissionsmøderne, der er vist i bilag 9.3.

3.3.1 Datagrundlaget for beregning af omsætningstabeller

Nedenstående Tabel 3 viser datagrundlaget for forcensuren i skriftlig dansk retskrivning og læsning. Forcensur- og øveprøver i dansk er gennemført i perioden 6.-20. april 2021, og FP9 er gennemført d. 5. maj 2021.

Tabel 3: Datagrundlag for forcensur i dansk læsning og retskrivning (før og efter datarensning)

Population	Elever, retskrivning (2017+2021)	Elever, læsning (2017+2021)
Forcensur	1.330	1.409
Øveprøver	10.160	9.676
Brutto Sample	**11.490**	**11.085**
Frasorteret (nulscore i en prøve)	106	-
Frasorteret pga. inhomogen indsats (i parentesen angives datakilde: Øveprøve/Forcensurskole)	198 (195/3)	208 (203/5)
Netto Sample	**11.186**	**10.877**

Integrationen af øveprøver som en del af ækvivaleringen har som nævnt givet et styrket datamateriale, men samtidig har det krævet en udvidet datarensning. Det skyldes som ovenfor nævnt, at der i øveprøvedata kan forventes at optræde en større andel uengagerede besvarelser (fremstår ofte som outliers, men er det ikke nødvendigvis da frasorterings ikke sker på den afhængige variabel) end data fra forcensurskolerne indeholder. I forbindelse med datarensning blev der således frasorteret en række besvarelser, hvor det vurderes, at der med stor sandsynlighed er tale om usammenlignelige præstationer i de to prøver. Elever, der opfylder ét af følgende kriterier, er frasorteret i det endelige datamateriale:

a) Hvis en af de to prøvers score er 0. Her vurderes det, at eleven ikke reelt har gjort en indsats. Frasortering på den afhængige variable (her scoren) skal altid ske med stor varsomhed, og dette kriterium betyder konkret at analysen ikke kan bruges ved fastlæggelse af grænsen mellem -3 og 0 (hvilket traditionelt gøres ud fra praktiske overvejelser end statistiske vurderinger).
b) Hvis elevens indsats er er inhomogen. For at undgå frasortering på grundlag af scoren sker fravurderingen på grundlag af differencen mellem antal gyldige (dvs. ikke-blanke) svar i forcensurprøven (2017-prøven) og FP9 (2021-prøven). Hvis denne difference er større end et konkret vurderet acceptabelt niveau (25 eller mere i læsning eller 30 eller mere i retskrivning) frasorteres opgaven. Her vurderes det, at indsatserne er så forskellige, at de to prøveresultater ikke kan sammenlignes meningsfuldt. Som det fremgår af Tabel 3 (næstsidste

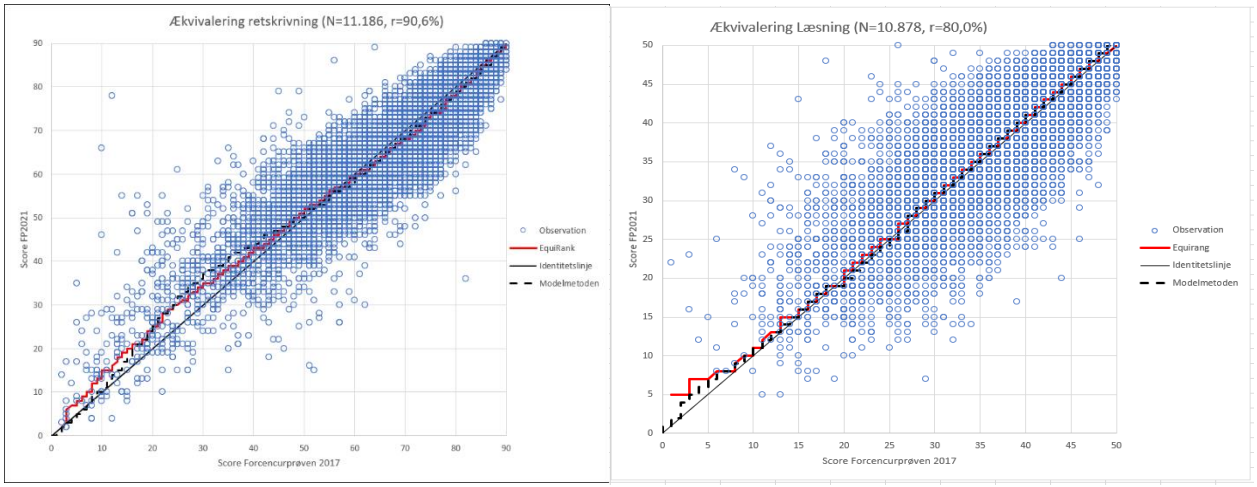
linje, frasorteret pga inhomogen indsats) er der kun skete en yderst begrænset frasortering af data og næsten i alle tilfælde har det været fra øveprøver (i retskrivning er 195 elevbesvarelser ud af de 198 frasorterede fra øveprøver, læsning er det 203 ud af 208).

De efterfølgende beregninger er baseret på det rensede datamateriale (Netto sample i Tabel 3).

3.3.2 Ækvivalering og valg af metode

Nedenfor vises ækvivaleringskurven for dansk retskrivning og læsning. I retskrivning ligger kurven lidt over identitetslinjen (hvor x=y) i nederste halvdel af pointskalaen, dvs. eleverne får typisk flere point i 2021-prøven end i 2017-prøven. I øverste del af pointskalaen får eleverne lidt færre point i 2021-prøven.

Figur 1. Ækvivaleringskurver, dansk retskrivning og læsning 2017/2021 (rensede data).



I læsning er det gennemsnitlige antal korrekte svar meget tæt på identiske i de to prøver, hvorfor omsætningstabellen for 2021 kommer til at være næsten identisk med omsætningstabellen fra maj 2017.

Som det fremgår, er der gennemført to typer af ækvivaleringer: Equirank (den røde kurve) og ækvivalering med modelmetoden (den sorte stiplede kurve). Metoderne har hver især fordele og ulemper og kan give marginalt forskellige resultater uden at det endegyldigt kan afgøres om det ene resultat er mere korrekt end de andet. Den mest markante fordel ved modelmetoden er at den kan give mere robuste resultater hvor der begrænset datagrundlag. De mest markante fordele ved Equirank-metoden er at den er umiddelbart forståelig, gennemskuelige og til at gennemføre/analysere i et regneark. Såfremt der ikke har været særlige begrundelser for andet, har det derfor været praksis at tage afsæt i equirank ækvivaleringen ved fastlæggelse af omsætningstabeller. Resultater fra modelmetoden har været brugt som kontrol. Metoderne er beskrevet nøjere i afsnit 4.

3.3.3 Ækvivalering med afsæt i data fra forcensurskoler overfor øveprøver

Som nævnt i afsnit 3.3.1 bygger dataanvendelsen ved ækvivaleringen på, at data fra forcensurskoler og øveprøver kan betragtes som sammenlignelige når der er renset for inhomogen indsats. Dette er en overvejelse der er baseret på nyere forskning på området (jf afsnit 3.2). Dette skal efterprøves med afsæt i de indhentede retskrivningsdata. Data herfra er særligt egnede til afprøvning af denne antagelse af flere årsager:

1. Der er relativ stærk sammenhæng (r=0,91) mellem resultaterne i forcensurdata og årets prøve så selv små variationer vil kunne identificeres som systematiske,
2. Der er tydelige forskel på strukturen i prøvernes sværhed (afvigelser fra identitetslinjen).

3. Der er forskel på resultaterne ved de to ækvivaleringsmetoder (ikke fuldstændigt overlap mellem de to kurver i Figur 1)
4. Der er relativt mange items i retskrivningsprøven, hvilket alt andet lige betyder en bedre mulighed for at identificere afvigelser.

I Figur 2 fremgår i *venstre side* resultaterne af to t-test, der viser, at eleverne fra forcensurskolerne er marginalt, men lige akkurat signifikant bedre præsterende end elever fra øveprøveskoler, der dog dækker over en anelse større variation i elevdygtighed. Dette mønster et helt parallelt i de to prøver (Øverst: Årets Maj 2021-prøve og Nederst: Forcensurprøven, Maj 2017).

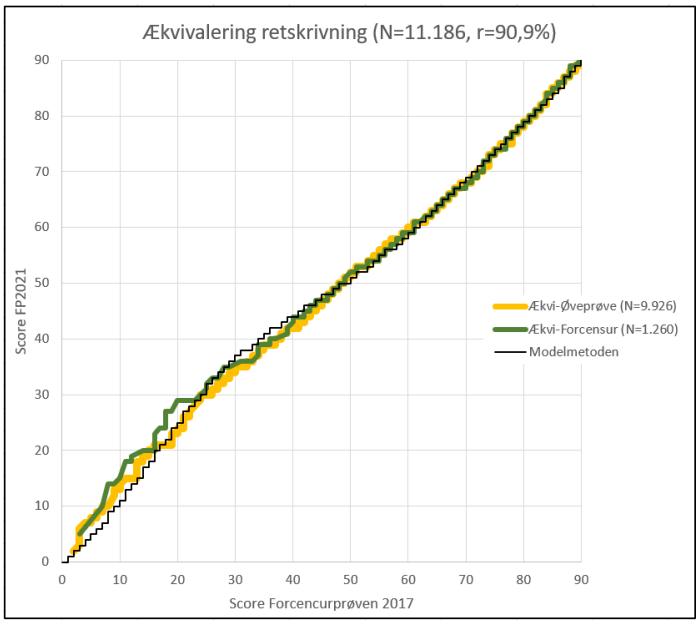
Figur 2: Data for forcensur- og prøveresultater for eleverne på øveprøve- og forcensur-skoler.

Prøveresultater Maj 2021-prøven i retskrivning for forcesur-populationen

t-test: To stikprøver med forskellig varians	*Eleverne, der tog Øveprøver*	*Elever fra Forcensurskoler*
Middelværdi	67,0	68,1
Varians	195,0	190,2
Observationer	9926	1260
Hypotese for forskel i middelværdi	0	
fg	1605	
t-stat	-2,590	
P(T<=t) en-halet	0,005	
t-kritisk en-halet	1,646	
P(T<=t) to-halet	0,010	
t-kritisk to-halet	1,961	

Forcensur-data fra Maj 2017 prøven i retskrivning - indhentet april 2021

t-test: To stikprøver med forskellig varians	*Resultater 2017 - Øveprøver*	*Resultater 2017 - Forcensurskoler*
Middelværdi	67,8	68,9
Varians	232,3	228,1
Observationer	9926	1260
Hypotese for forskel i middelværdi	0	
fg	1602	
t-stat	-2,312	
P(T<=t) en-halet	0,010	
t-kritisk en-halet	1,646	
P(T<=t) to-halet	0,021	
t-kritisk to-halet	1,961	



I Figur 2 fremgår i *højre side* tre ækvivaleringskurver: Dels modelækvivaleringen (baseret på hele populationen) som også fremgår af venstre side af Figur 1, dels to særskilte ækvivaleringskurver efter equirank-metoden for henholdsvis med gul øveprøve- og med grøn forcensurskole-populationen. Hvis den gule og grønne kurve havde været fuldstændigt sammenfaldende, havde vi eftervist at der ingen forskel var på, hvordan eleverne på øveprøve-skoler og forcensur-skoler havde besvaret de to prøver i forhold til hinanden. Imideltid ses at der i den lave ende (ved en score under 25 i forcensurprøven) er at der er synlige forskelle mellem de to kurver. Det er dog svært at sige om denne forskel er systematisk, da der er meget begrænset datagrundlag i denne ende af skalaen (25 elever, svarende til 2% af eleverne score lavere end 25 point). Ser man på kurveforløbet for elever der i forcensurprøven (den vandrette akse) har scoret 30 eller derover er der et næsten fuldstændigt sammenfald mellem den gule og grønne kurve. Og da skæringspunktet mellem karaktererne 0 og 2 går mellem 31 og 32 point ved maj 2017-prøven, betyder den forskel, der er mellem de to kurveforløb intet for fastlæggelse af omsætningstabellen.

Konklusionen er altså, at man kommer frem til nøjagtig den samme omsætningstabel ved at anvende de to populationer. Det næste spørgsmål er, om man kan generalisere ud fra denne observation eller om sammenhængen er specifik for retskrivning i 2021. Eller med andre ord, vil man komme til et tilsvarende resultat, hvis man ser på matematik, læsning, fremmedsprog og/eller et af naturfagene? Udgangspunktet for analysen (at stakes ikke spiller nogen afgørende rolle, når først respondenten har "bidt på" og er gået i gang med at besvare opgaverne) var generel og ikke afhængig af fag, og envidere viser andre erfaringer (jf Attali 2016) at de samme mekanismer gør sig gældende for både læseforståelse og matematik.

3.3.4 Omsætningstabeller og karakterfordeling

Nedenfor vises de beregnede omsætningstabeller for 2021-prøverne i læsning og retskrivning som begge er baseret på Equirank-metoden. Ækvivaleringskurven for retskrivning afspejles i omsætningstabellen, idet vi ser en forskydning på op til to point. I bunden af karakterskalaen er 2021-prøven lettere, og der skal fx 36 point til for at bestå – mod 32 point i 2017-prøven. I toppen af skalaen, hvor 2021-prøven er lidt sværere, skal der 76 point til for at opnå karakteren 10.

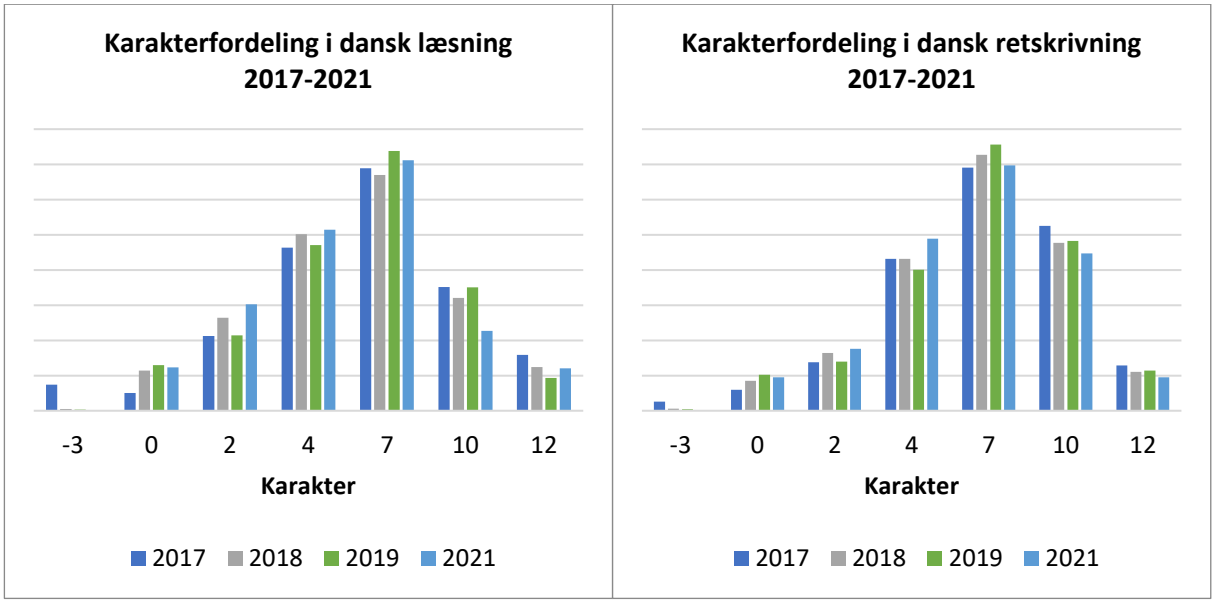
I dansk læsning ligger de to prøver meget tæt op ad hinanden i sværhedsgrad, jf. ovenfor. Der er således kun en enkelt forskydning i omsætningstabellen, idet der i 2021-prøven kræves 45 point for at opnå karakteren 10, hvor der i 2017-prøven kræves 44 point.

Tabel 4. Beregnet omsætningstabel (Equirank) for dansk retskrivning og læsning, 2017-prøven ift. 2021-prøven.

Retskrivning			Læsning		
Karakter	*Omsætning 2017 (point)*	*Beregnet omsætning 2021 (point)*	*Karakter*	*Omsætning 2017 (point)*	*Beregnet omsætning 2021 (point)*
12	86-90	**85**-90	12	48-50	48-50
10	78-85	**76-84**	10	44-47	**45**-47
7	65-77	**64-75**	7	37-43	37-**44**
4	48-64	**50-63**	4	30-36	30-36
2	32-47	**36-49**	2	22-29	22-29
0	2-31	1-**35**	0	1-21	1-21
-3	0-1	0	-3	0	0

De foreslåede omsætningstabeller resulterer i en karakterfordeling for FP9 dansk retskrivning (63.409 elever) og dansk læsning (63.429 elever), der er vist i Figur 3.

Figur 3. Beregnet karakterfordeling i dansk retskrivning og læsning, 2017-2021.



Sammenlignet med 2019, medfører omsætningstabellen i retskrivning en forskydning fra toppen mod bunden af karakterskalaen. En mindre andel elever får karakteren 12, 10 og 7, og en større andel elever får karakteren 4 og 2.

Samme tendens ser vi for læsning, hvor en større andel elever får karakteren 2 og 4, og færre elever får karakteren 7 og 10.

3.4 Forcensur i matematik med og uden hjælpemidler

I matematik betød mulighederne for at kombinere forcensuren med gennemførelse af øveprøver, at der blev indhentet et meget stort datamateriale til brug for beregning af omsætningstabellen til matematik uden hjælpemidler. Derimod blev det ikke muligt at øge datagrundlaget tilsvarende for matematik med hjælpemidler, da prøven ikke er selvrettende og skal rettes af en censor eller lærer. Her gives resultaterne i hovedpunkter, mens de mere detaljerede resultater kan findes i Præsentationsmateriale til opgavekommissionsmøderne, der er vist i bilag 9.3.

3.4.1 Datagrundlaget for beregning af omsætningstabeller

Nedenfor vises datagrundlaget for forcensur i skriftlig matematik med og uden hjælpemidler. 9.941 har gennemført både 2017-prøven og 2021-prøven i matematik uden hjælpemidler. I papirprøven i matematik med hjælpemidler har 889 elever deltaget i begge prøvesæt. Forcensur- og øveprøver er gennemført i perioden 6.-19. april 2021, og FP9 er gennemført d. 4. maj 2021.

Tabel 5: Datagrundlag for forcensur i matematik med og uden hjælpemidler (efter datarensning)

Population	Elever, mat.u.hj. (2017+2021)	Elever, mat.m.hj. (2017+2021)
Forcensur	1.247	889
Øveprøver	8.694	-
Total	**9.941**	**889**

3.4.2 Ækvivalering

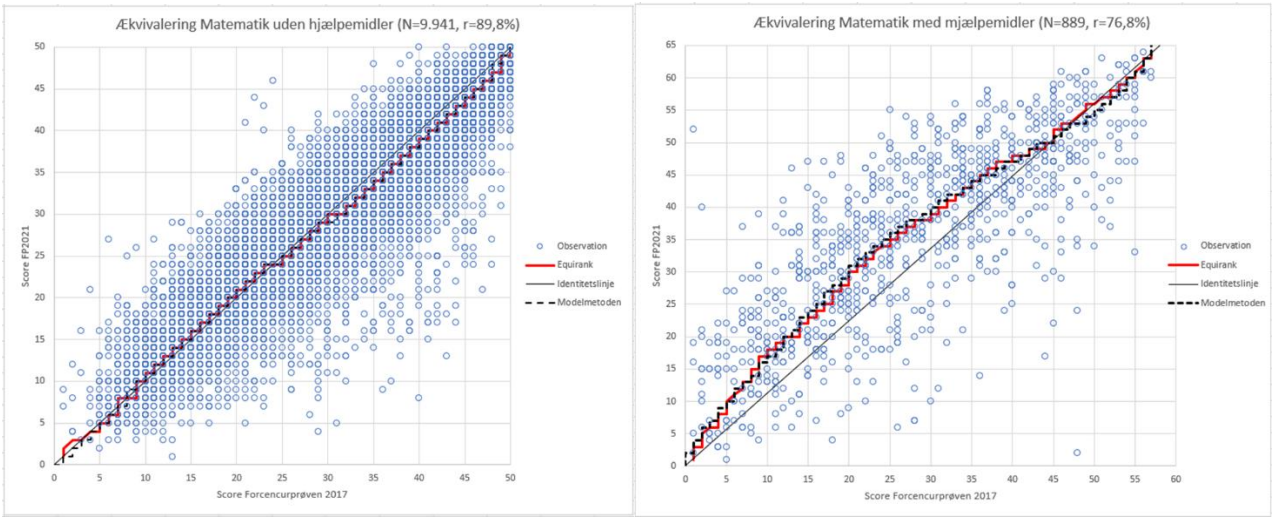
Ligesom i dansk (se afsnit 3.3.2) er der frasorteret en række besvarelser, hvor det vurderes, at der med stor sandsynlighed er tale om usammenlignelige præstationer i de to prøver. Efter frasortering resterer 9.941 elever, hvoraf 12,5 procent kommer fra forcensurskoler. Elever, der opfylder ét af følgende kriterier, er frasorteret i det endelige datamateriale for matematik uden hjælpemidler:

a) Hvis enten 2017- eller 2021-prøvens resultat er 0. Her vurderes der at være en meget stor risiko for, at eleven ikke reelt har gjort en indsats (i alt 37 besvarelser frasorteret, heraf 1 fra forcensurskoler).
b) Hvis differencen mellem antal blanke svar i forcensurprøven (2017-prøven) og FP9 (2021-prøven) er 25 eller større vurderes det, at indsatserne er så forskellige, at de to prøveresultater ikke kan sammenlignes (i alt 94 besvarelser frasorteret, heraf 5 fra forcensurskoler).

Nedenfor vises ækvivaleringskurven for matematik med og uden hjælpemidler. I prøven uden hjælpemidler ligger kurven tæt ved identitetslinjen (det vil sige der hvor andel rigtige i de to prøver er det samme) – dog lidt under linjen i den øverste halvdel af pointskalaen. I denne ende af skalaen får eleverne således lidt færre point i 2021-prøven ift. i 2017-prøven.

I matematik med hjælpemidler ses en betragtelig forskydning af kurven, hvor eleverne får relativt flere point i 2021-prøven ift. i 2017-prøven (med undtagelsen af i toppen af pointskalaen). Årets prøvesæt (2021) er derved væsentligt lettere end 2017-prøvesættet.

Figur 4. Ækvivaleringskurver, matematik med og uden hjælpemidler 2017/2021 (rensede data).



3.4.3 Omsætningstabeller og karakterfordeling

Nedenfor vises de beregnede omsætningstabeller for 2021-prøverne i matematik med og uden hjælpemidler (baseret på equirank-modellen). I matematik uden hjælpemidler, jf. Figur 4, skete der stort set ingen ændringer i bunden af skalaen, mens der ved første gennemgang af besvarelserne skete en vis forskydning i toppen af pointskalaen, hvor det fx så ud til, at eleven i 2021-prøven skal have 43 point for at opnå karakteren 12 – mod 45 point i 2017-prøven. Det så altså ud til, at især de dygtige elever gennemgående havde sværere ved 2021-prøven. Imidlertid viste en gennemgang af fejlsvarerne, at der især for to items vedkommende ved en fejl var rettet for restriktivt (en version af svaret som var rigtigt var ikke kommet med i rettevejledningen, og derfor fik en del elever en eller to fejl flere, end de burde).

Dette blev imidlertid fanget inden omsætningstabellen blev endeligt fastlagt, hvilket er baggrunden for de to versioner af omsætningstabellen i matematik uden hjælpemidler, der fremgår af Tabel 6. Efter denne fejl var korrigeret viste data fra besvarelserne, at forcensurprøven (maj 2017 opgavesættet) og 2021-prøven i matematik uden hjælpemidler var stort set lige svære. Dette resultat kan også være med til at ramme en pæl igennem en mistanke om, at forcensurprøverne (på grund af asymmetri i hvad der er på spil) har systematisk bias, blot der som her er renset for systematiske fejl. Denne sammenhæng har også internationale studier har været med til at belyse[1].

Tabel 6. Beregnet omsætningstabel (equirank) for matematik med og uden hj.m., 2017-prøven ift. 2021-prøven.

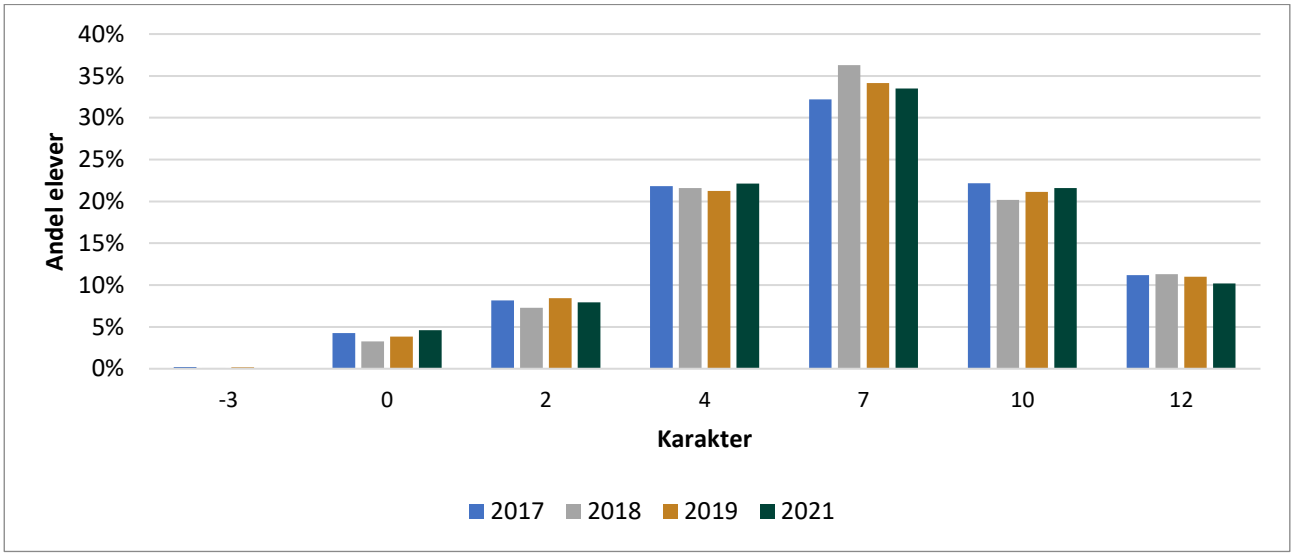
Matematik uden hjælpemidler				Matematik med hjælpemidler		
Karakter	*Omsætning 2017*	*Beregnet omsætning 2021, 1. Version**	*Beregnet omsætning 2021, revideret version**	*Karakter*	*Omsætning 2017*	*Beregnet omsætning 2021*
12	45-50	**43**-50	**44**-50	12	44-58	**50-65**
10	38-44	**36-42**	**37-43**	10	33-43	**42-49**
7	26-37	26-**35**	26-**36**	7	20-32	**29-41**
4	16-25	16-25	**17**-25	4	12-19	**19-28**
2	11-15	11-15	**12-16**	2	7-11	**12-18**
0	1-10	1-10	1-**11**	0	1-6	1-**11**
-3	0-1	0	0	-3	0	0

[1] Attali, Y. (2016). Effort in Low-Stakes Assessments: What Does It Take to Perform as Well as in a High-Stakes Setting? Educational and psychological measurement, 76(6), 1045-1058.

*Omsætningstabellen i matematik uden hjælpemidler blev korrigeret på grund af justeringer i rettevejledningen. Alle omsætningstabeller er baseret på Equirank metoden.

Modsat ser det ud til, at 2021-prøven i matematik med hjælpemidler har været væsentlig lettere end 2017-prøven. Der sker således en forskydning i omsætningstabellen hele vejen op gennem pointskalaen. Alene af den årsag kræver det flere point at opnå en given karakter. Dog skal man være opmærksom på, at der var flere point i 2021-prøven. Man skal derfor være forsigtig med at konkludere, om opgavesættets sværhed ud fra forskelle i omsætningstabellerne (eksempelvis når der i 2017-prøven kun skal 33 point til for at opnå et 10-tal, hvor det i 2021-prøven kræves 42 point, så kan ikke hele forskellen tilskrives forskellig sværhed).

Figur 5. Karakterfordeling i matematik uden hjælpemidler, 2017-2021.



Af Figur 5 ses den karakterfordeling, der opnås blandt alle elever, der har deltaget i FP9 Matematik, når der benyttes de foreslåede omsætningstabeller. Den foreslåede omsætningstabel resulterer i en karakterfordeling for FP9 matematik uden hjælpemidler (63.778 elever). Fordelingen er ganske tæt på 2017-fordelingen, og der er en lille forskydning på max 1 procentpoint i andel.

Karakterfordelingen i matematik med hjælpemidler kan ikke på tilsvarende måde vises, da populationsdata eller større stikprøver endnu ikke er gjort tilgængelig.

3.5 Præsentationer for opgavekommissioner

En central delleverance i projektet er de præsentationer af resultaterne fra forcensuren, der gives til censormøderne i opgavekommissionerne (se Bilag 9.3), hvor omsætningstabellerne fastsættes. Tidligere har det været fire møder (Dansk, Engelsk, Matematik og Naturfag), men i 2021 drejede møderne sig om resultaterne fra de fire prøver fordelt på to opgavekommissioner. Forcensur i dansk læsning og retskrivning er blevet præsenteret på møde med opgavekommissionen i dansk d. 19. maj 2021, og forcensur i matematik med og uden hjælpemidler er præsenteret for opgavekommissionen i matematik d. 20. maj 2021.

Et af de centrale formål med projektet og med præsentationen er at levere et solidt fagligt grundlag til opgavekommissionen og et databaseret og begrundet forslag til årets omsætningstabel for alle skriftlige prøver, der er omfattet af forcensur. Præsentationerne (inkl. besvarelse af spørgsmål og dialog om data) har typisk været af en times varighed og har fulgt følgende struktur:

- **Indledning:** Præsentationsrunde og kort introduktion til forcensuren i praksis.

- **Datagrundlag:** Ækvivaleringsgrundlag, repræsentativitet og datarensning.
- **Opgavesættenes sværhed:** Herunder opgavesværhed, herunder sværhed ift. elevdygtighed.
- **Eksempler på itemanalyser:** Differentiel item funktion samt eksempler på gode og mindre gode opgaver.
- **Ækvivalering:** Ækvivaleringsmodeller og ækvivaleringskurver.
- **Omsætningstabeller og karakterfordeling:** Forslag til årets omsætningstabeller og betydningen heraf for karakterfordelingen og landsgennemsnittet.

Som det fremgår, indgår der i drøftelserne typisk andet og mere end blot omsætningstabellen. Som led i ækvivaleringen (modelmetoden) foretages en raschanalyse, og igennem besvarelserne af disse skabes en betydelig viden og en række informationer om elevernes opfattelse af de enkelte opgavers/items funktionsmåde. Analyserne drejer sig naturligvis om opgaver fra en prøve, der er afholdt, og kan derfor ikke påvirke resultatet heraf. Men da det ofte vil sætte fokus på generelle mekanismer, når der påpeges styrker og svagheder i enkeltopgaver, kan denne viden være af stor betydning for opgavekommissionerne ved deres fremadrettede arbejde med opgaver til næste års prøver.

Analyserne kan kort sagt være med til at sætte fokus på, om prøven bedst muligt kan bruges til at vurdere den enkelte elevs præstation, og om den afspejler elevens dygtighed i faget:

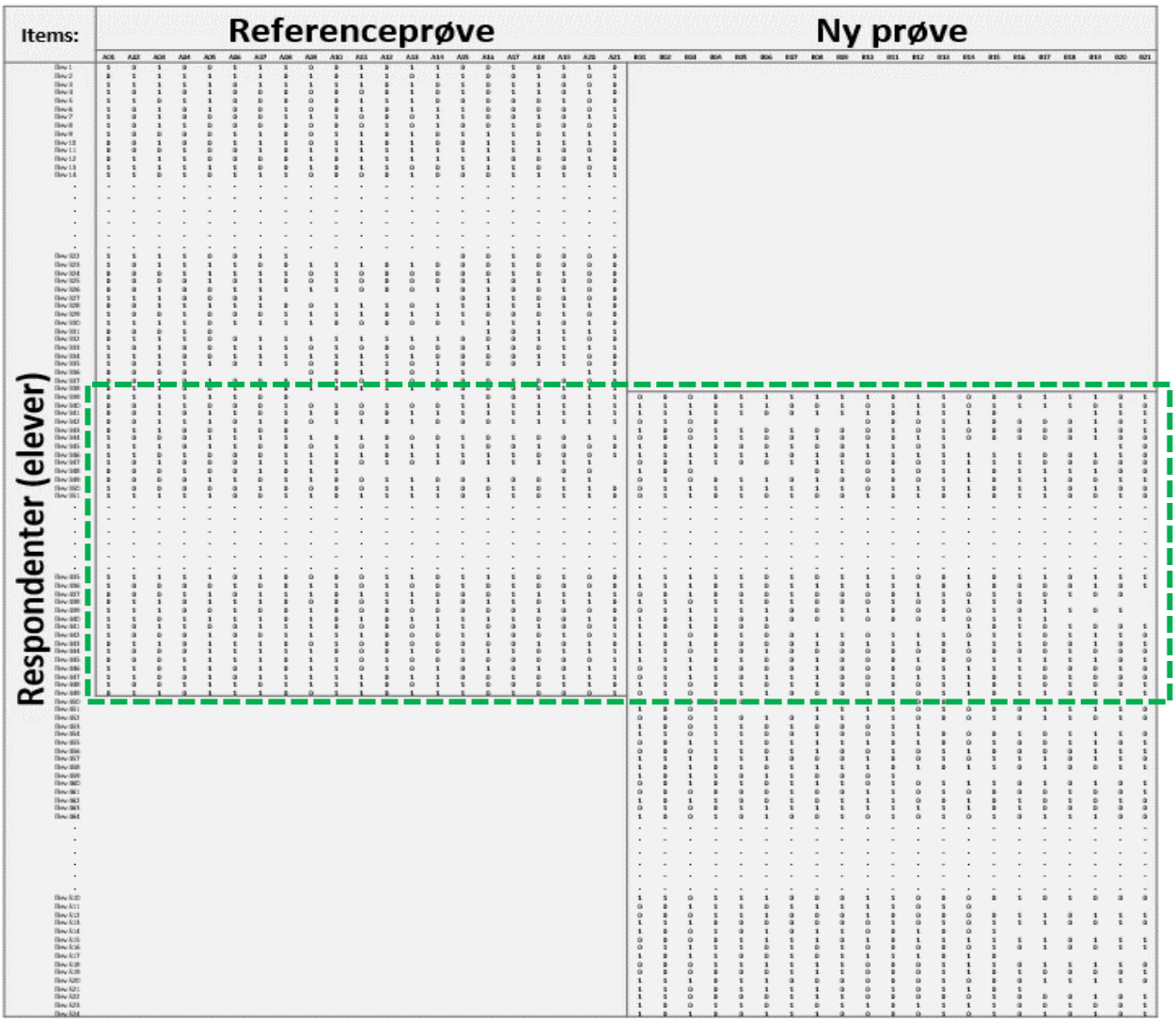
- **Har de anvendte skabeloner for opgaverne været valgt hensigtsmæssigt?** Kunne man få bedre og/eller mere data ud af en opgave, hvis den var blevet stillet på en anden måde?
- **Er det alene elevens dygtighed, der spiller ind på sandsynligheden for korrekt svar?** Er der opgaver, hvor fx elevens køn eller bopæl i landet giver visse elever urimelige vilkår i forhold til andre (differentiel item funktion (DIF)).
- **Er der opgaver, hvor sandsynligheden for korrekt svar ikke i stiger proportionalt med elevens dygtighed i faget?** Er der opgaver, som tilsyneladende ikke måler den samme form for dygtighed som resten af prøven, fx trickopgaver eller meget teksttunge opgaver i naturfag, der giver dårlige læsere svære vilkår (Lav diskrimination/manglende monotonicitet)?
- **Er der opgaver, hvor besvarelsen viser, at opgaven/items er koblet til andre opgaver/items på en uhensigtsmæssig måde?** Fx opgaver, hvor muligheden for at svare rigtigt i højere grad afhænger af svaret på en anden opgave end elevens dygtighed (lokal afhængighed)?
- **Er der opgaver med dårlige distraktorer?** Er der fx gennem uhensigtsmæssigt valg af forkerte svarmuligheder skabt trickopgaver, der snyder de dygtigste elever og favoriserer de elever, der ikke har tilstrækkelige færdigheder?

I præsentationerne vises typiske eksempler på gode og dårlige opgaver fra årets opgavesæt ved at gennemgå item-karakteristik-kurver (ICC'er) fra udvalgte opgaver. Eksempler herpå fremgår af præsentationer 9.3 og af bilag 9.6.

4. Metoder til ækvivalering

Der anvendes to metoder til ækvivalering i forcensurprojektet: Equirank- og model-metoden. Der er fordele og ulemper ved hver af metoderne, som begge baseres på, at der eksisterer en dygtighed hos eleverne, der er fælles for de prøver, som ækvivaleres (typisk referenceprøven og den nye prøve). I Figur 6 er vist besvarelser fra to prøver (opgaverne er søjlerne), hvor en andel af eleverne (rækkerne) har taget begge prøver (indrammet af den stiplede grønne boks). Med udgangspunkt i denne illustration skal de to metoders fordele og ulemper omtales.

Figur 6: Datasæt til ækvivalering af to prøver (eksempel)



4.1 Equirank-metoden

Denne metode baserer sig på den klassiske testteori, og her ser man alene på totalscorer (altså summen af antal rigtige i en prøve). Ved equirank-metoden er det alene eleverne der har taget begge prøver (den grønne stiplede boks), som kan indgå ved beregningen af sammenhængen mellem prøverne, og der er en række forudsætninger, som skal være opfyldt, før metoden kan anvendes med høj grad af fortolkelighed af resultaterne:

- Præstationerne skal afspejle forskelle i dygtighed – det forudsættes altså at eleverne gør deres bedste ved besvarelsen.

- Prøvebesvarelser, der inddrages i analysen, skal bestå af to komplette besvarelser, således at ubesvarede opgaver kan fortolkes som forkerte svar.
- Spredningen i opgavernes sværhed skal i et vist omfang passe til spredningen i elevernes dygtighed.
- De to prøver skal måle samme former for dygtighed, således at en elev der klarer godt i den ene prøve også vil klare sig godt i den anden prøve.
- Elevpopulationen skal være rimelig stor (helst ikke under 200) og udvalgt således, at der er god spredning i præstationerne (både dygtige og svage elever).

Den elev, der er dygtigst i den ene prøve (renset for tilfældige udsving), skal således også (i teorien) være dygtigst til den anden prøve – prøverne vil i så fald rangordne eleverne på samme måde (heraf navnet equirank-metoden).

Plotter man alle elevernes resultater i de to prøver mod hinanden vil det teoretisk ideelle resultat altså blive en monotont stigende kurve. Sammenhængen vil kun være lineær i det særlige tilfælde at de to prøvers sværhed (item for item) er den samme. Men man kan sagtens sammenligne resultater af prøver med meget forskellig sværhed og i øvrigt med forskelligt antal opgaver, men så bliver sammenhængen altså ikke-lineær.

Den fuldstændig monotone sammenhæng finder man ikke i praksis, heller ikke selvom prøverne skulle afprøve præcist den samme færdighed, bl.a. fordi tilfældigheder og gode/dårlige dage spiller ind på elevernes besvarelse. Men det er en forudsætning for modellens anvendelse, at der er en solid korrelation mellem de to prøveresultater.

Sammenhængen mellem de to prøver fås ved at rangordne elevernes resultater i hver prøve og derefter plotte resultater med samme rangordensplacering (elevuafhængigt) på hver sin akse (nr. 1 mod nr. 1, nr. 2 mod nr. 2 osv.). Derved får man den monotont stigende kurve, der er omtalt ovenfor, der vil være det bedste bud på sammenhængen mellem de to prøver, hvis men ser bort fra tilfældige udsving.

Fordelen ved denne metode er, at den er umiddelbart transparant, meget enkel at udføre i et regneark og i de fleste tilfælde (hvor flertallet af eleverne har svaret alle opgaver i begge prøver) giver den et lige så solidt resultat som mere komplicerede modeller.

Ulempen er, at det kan være svært at gennemskue, hvad det betyder, hvis ikke alle forudsætningerne er opfyldt, og metoden kræver et særdeles kvalificeret datagrundlag med mange komplette besvarelser fra den samme elevgruppe.

4.2 Modelmetoden

Denne metode baseres på moderne testteori, hvor der kan inkluderes ikke-komplette besvarelser og besvarelser fra elever, der kun har besvaret den ene af prøverne. Yderligere kan inddrages flere end to prøver i ækvivaleringen. Metoden adskiller sig grundlæggende ved at der – med udgangspunkt i alle besvarelser fra de inkluderede prøver – beregnes en fælles dygtighed for hver elev (en raschscore[2]). Udgangspunktet er *at prøverne måler samme konstrukt*, at der kan *estimeres en raschscore pr. elev (person location)* der er fælles for prøverne der ækvivaleres og *at der kan fastlægges en fast relation mellem denne raschscore og totalscoren* (antallet af rigtige svar).

[2] Raschscoren er en relativ score, der nærmere afspejler *andelen* end *antallet* af rigtige svar. Et eksempel: I en prøver på 30 items svares 12 rigtige, 10 forkerte og 8 blanke – spørgsmålet er om andelen af rigtige er 12/30 eller 12/(30-8). I sidstnævnte tilfælde vil altså opnås en større raschscore uanset at antallet af rigtige svar er det samme.

Der beregnes altså en sammenhæng mellem elevens dygtighed (person location), og *hvor mange rigtige svar en elev kan forventes at levere hvis eleven besvarer alle opgaver i prøven*. Output til en sådan analyse fra Rumm2030 er vist i Figur 17, hvor sammenhængen mellem prøveresultatet i matematik uden hjælpemidler og de tre naturfag er illustreret.

I modelmetoden kan der beregnes sammenlignelige estimater af dygtigheden selvom eleverne ikke har besvaret det samme antal items i en prøve. Modellen kan anvendes i to varianter. En styrke ved begge varianter af denne metode er, at man kan inddrage besvarelser fra elever, som ikke har deltaget i begge/alle prøver. Den væsentligste forskel mellem varianterne er hvordan manglende besvarelser i prøven håndteres.

I variant A) behandles ubesvarede spørgsmål som et forkert svar (ligesom ved Equirank-metoden) således at alle elever har den samme maksimale score i hver af prøverne. Denne variant af modelmetoden (også kaldet testlet-modellen eller Leunbachs Powerseriemodel[3]) er den der som regel er anvendt i dette projekts ækvivalering af folkeskolens prøver, da et blankt svar som hovedregel betragtes på linje med et fejlsvar - især i den endelige prøve.

I variant B) kan ubesvarede opgaver behandles i analysen på flere måder: Enten som forkerte svar eller som om de ikke var stillet og derfor udgår af analysen. Det kan være mere retvisende til analyseformål at se bort fra alle eller nogle af de manglende besvarelser i prøver, fx hvor eleven ikke har forholdt sig til alle opgaverne i en low stakes prøve. Det kan fx være relevant hvis eleven (som eksamenstræning) bliver sat til at svare på så mange opgaver som muligt på begrænset tid. Eller hvis eleven kun er blevet præsenteret for et udsnit af opgaverne, der normalt indgår i prøven.

Forskellen er illustreret i eksemplet nedenfor.

Figur 7: Eksempel til illustration af en af forskellene mellem de to varianter

Case: *En elev deltager i en sprogprøve bestående af 3 delprøver med i alt 80 items. Eleven får stillet de to første delprøver med i alt 50 items og besvarer de første 45 items med 25 rigtige – de sidste 5 item besvarer eleven ikke. Eleven bliver slet ikke stillet delprøve 3 (30 items).*

Modelmetoden, variant A: *Eleven vil blive vurderet ud fra 25 rigtige ud af alle 80 items – altså som om alle de ubesvarede opgaver var besvaret forkerte. Andel rigtige bliver 25/80.*

Modelmetoden, variant B: *Eleven vil blive vurderet ud fra 25 rigtige enten ud af de 45 besvarede opgaver eller de 50 opgaver som eleven er blevet præsenteret for fra delprøve 1 og 2. De 30 opgaver fra delprøve tre behandles som ikke eksisterede. Man kan altså afhængig af kontekst vælge om eleven skal vurderes på grundlag af en andel rigtige på 25/45 eller 25/50.*

4.3 Konklusion vedr. valg af ækvivaleringsmetode

Resultater fra Equirankmodellen er lettest at forstå og formidle, og ved et solidt datagrundlag med mange komplette besvarelser for begge prøver i begge ender af skalaen, er den derfor klart at foretrække.

Med modelmetoden kan man udnytte mest muligt af datagrundlaget. Der beregnes en separat score for hver prøve og ved at bruge eleverne, der har besvaret begge prøver (eleverne i den grønne boks i Figur 6) som "anker-respondenter", sikres det, at de to scorer kan placeres på en

[3] Efter Gustav Leunbach, der beskrev denne model i 1976 i Teknisk rapport nr. 19, Dansk Pædagogisk Institut. Metoden er beskrevet nøjere i Adroher, N. D., Kreiner, S., Young, C., Mills, R., & Tennant, A. (2019). Test equating sleep scales: applying the Leunbach's model. BMC medical research methodology, 19(1), 1-13.

fælles Rasch-skala. Så ved ækvivaleringen kan modelmetoden give svar på, hvordan en elev med en given Rasch-score (person location) vil kunne forventes at klare sig i hver af de 2 prøver (målt ved den forventede totalscore). Ved variant A) af modellen behandles alle ubesvarede opgaver som fejlsvar (ligesom ved equirankmodellen). Dette vil som hovedregel være den variant, der er mest naturlig at anvende i dette projekt.

Men i visse tilfælde vil det give et fejlagtigt billede af elevens dygtighed, og i så fald kan man benytte variant B) af modelmetoden, især hvor man skal arbejde med data hvor eleven ikke er blevet præsenteret for eller har forholdt sig til alle opgaver i prøverne. Dette vil give en særlig mulighed for at få præcise estimater på skala-områder (fx de meget lavt præsterende elever), hvor der til low stakes prøver er et utilstrækkeligt datagrundlag, hvis man alene ser på de komplette besvarelser.

5. Bedømmelse af prøverne i naturfag

I forbindelse med den samlede omlægning af naturfagsprøven har der været gennemført forskellige versioner af en online-prøve for hver af de tre naturfag. I den sidst gennemførte ordinære prøve i maj 2019 var prøve-setuppet med 49-50 items og en varighed på 30 minutter.

I 2020 blev gennemført forskellige forsøgsvise afprøvninger af det nye setup, og der blev tilgængeliggjort en eksempelprøve (december 2019-sæt) på Testogprøver.dk med 75-76 items beregnet på en prøve af 60 minutters varighed. Yderligere er der gennemført forsøg med et naturprøve-setup baseret mestendels på åbne besvarelser af en serie opgaver (beregnet på efterfølgende retning af lærer eller censor – et prøvesetup sammenligneligt med prøven i matematik med hjælpemidler), der gennemføres i forlængelse af onlineprøven i faget. En omlægning af forcensurpraksis skal være forberedt, for at forsøgsprøverne med åbne besvarelser kan inkluderes i en samlet karaktergivning.

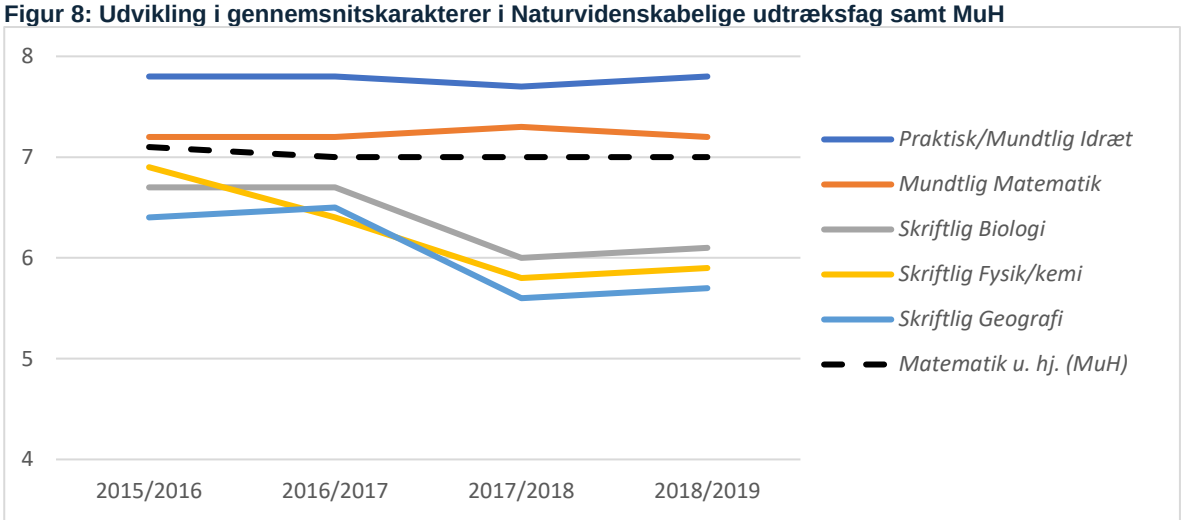
Onlineprøverne i naturfag blev aflyst både i 2020 og i 2021 og erstattet af årskarakterer. Som forberedelse til den afsluttende karaktergivning i 2021 og med henblik på at give eleverne erfaring med at gå til eksamen, blev det opgavesæt, som skulle have været anvendt til maj 2021-prøverne, brugt som øveprøver (i april 2021). Dette gav endvidere et godt datagrundlag for forcensurprojektet. Opgavesættet forventes at danne grundlag for den fremtidige forcensur i de tre naturfag, hvor der kan skabes en intern sammenhæng mellem karakterniveauer i STEM-fagene: Naturfagene og matematik uden hjælpemidler (MuH).

Med henblik på at etablere det bedst mulige grundlag analyseres dels data fra den sidst gennemførte udtræksprøve i naturfagene (2019) og fra årets øveprøver (2021).

5.1 Baggrund for omlægning af omsætningstabeller i naturfagene

De skriftlige udtræksprøver i naturfag bedømmes, som det fremgår af Figur 8, med lavere karakterer end de mundtlige naturvidenskabelige udtræksfag. Det er ikke usædvanligt at mundtlige karakterer ligger højere end skriftlige, men som det fremgår, er spændet vokset betydeligt i perioden 2016-2019.

Figur 8: Udvikling i gennemsnitskarakterer i Naturvidenskabelige udtræksfag samt MuH



Kilde: https://www.uvm.dk/statistik/grundskolen/karakterer-og-test/karakterer-i-grundskolen

Det fremgår således, at forskellen er blevet mere markant siden 2017, hvor den fælles naturfaglige prøve blev introduceret. Et fokusskifte i undervisningen i de tre naturfag som følge heraf kan være en medvirkende årsag til faldet, og måske at der er opstået en ny balance.

5.2 Analyser af prøvedata med afsæt i maj 2019-prøven

Som det endvidere fremgår af Figur 8 ovenfor og Tabel 7 nedenfor, ligger de tre naturfag nu ca. en karakter lavere end matematik uden hjælpemidler, hvor prøveformen er sammenlignelig, og hvor karaktergennemsnittet har været nogenlunde stabilt i perioden. Det bemærkes, at matematik også tilhører STEM-familien, og altså har et stærkt fagligt slægtskab med naturfagene.

Tabel 7: Data fra maj FP2019 – Råscore, karakterer og antal elever (ekskl. -03, dvs. blanke besvarelser)

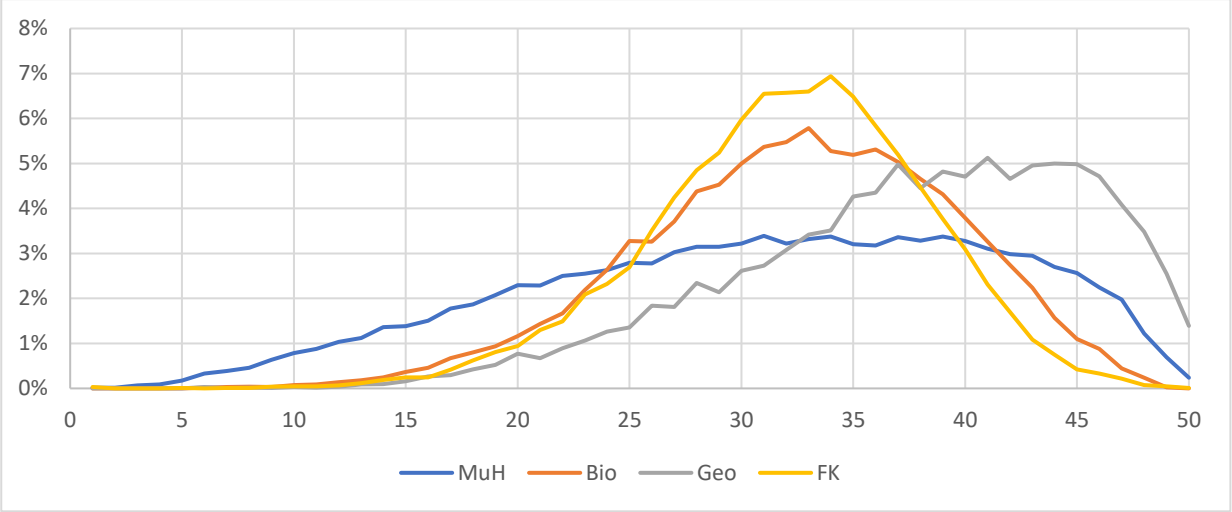
	Gns. råscore (ud af 50)	Gns. karakter	Antal elever
Matematik uden hjælpemidler (MuH)	29,7	6,85	65.636
Biologi	32,5	6,08	12.224
Geografi	37,6	5,69	12.796
Fysik-Kemi	30,5	5,83	12.903

Beregnes de kommende omsætningstabeller på grundlag af de hidtil anvendte metoder i forcensurprojektet (equirank- og/eller model-ækvivalering), vil man sandsynligvis opleve, at disse forskelle fastholdes, og en omlægning af de skriftlige naturfagsprøver til 1-times prøver kan få svært ved at blive betragtet som en positiv nyskabelse, da eleverne vil opleve at blive bedømt gennemsnitligt 1-2 karakterer lavere end ved udtræk af en af de andre prøver i de naturvidenskabelige udtræksfag.

5.2.1 Score- og karakterfordelinger for de skriftlige prøver i STEM-fagene

Det kan derfor overvejes at benytte introduktionen af de nye prøver i naturfagene som anledning til at genskabe sammenhængen mellem bedømmelserne i STEM-fagene, der afvikles som digitale selvrettende prøver. I Figur 9 vises pointfordelingerne for hver af de fire fag.

Figur 9: Fordeling af scores i STEM-fag, maj 2019



Det fremgår, at scorernes fordelinger er ret forskellige: Biologi og Fysik-Kemi minder om hinanden, hvor Geografi-opgaverne er lettere end de to andre fags opgaver (fordelingen er meget højreskæv). Fordelingen af scorerne i matematik uden hjælpemidler (MuH) er grundlæggende anderledes (meget fladere med større spredning, jf. også Figur 10, nedenfor), hvilket beror på, at opgaverne i MuH fortrinsvis besvares med åbne tal-svar (”skriv svaret”), hvor der overvejende måles produktive færdigheder, og hvor man ikke kommer så langt med et kvalificeret skøn, sat over for Multiple Choice-opgaverne i forskellige formater der systematisk anvendes i naturfagsprøverne (denne opgaveform måler i højere grad receptive færdigheder).

Forskellene udjævnes dog i betydeligt omfang af omsætningstabellerne, jf. Figur 10, der viser de resulterende karakterfordelinger i de fire fag.

Figur 10: Karakterfordelinger i STEM-fag, maj 2019

5.2.2 Rekalibrering af skalaerne i de tre skriftlige naturfagsprøver

Ønskes der etableret en bedre sammenhæng mellem bedømmelserne i de fire prøver, er der principielt to forskellige måder at gribe det an på:

1) Enten kan man så at sige nulstille systemet ved fx at tilpasse omsætningstabellerne til, at karakterskalaerne giver hver deres normalfordeling med et ønsket gennemsnit (fx 7),
2) eller også kan man vælge Matematik uden hjælpemidler (MuH) som gennemgående skala, ved at definere et baseline-niveau for de øvrige skalaer.

1. Den første metode er meget enkel, men har den ulempe, at den ikke nødvendigvis skaber en naturlig sammenhæng mellem STEM-skalaerne, og at en total nulstilling vil bero på en række mere eller mindre arbitrære valg (fx hvorfor en normalfordeling? Hvorfor ikke lægge gennemsnittet på 6,5 eller på 8?).

2. Den anden metode har flere fordele. Det vil umiddelbart virke mere naturligt at lægge sig op ad MuH, da det er en prøve/et fag i samme familie. Forudsætningen for denne argumentation er dog, at det kan sandsynliggøres, at der er en klar sammenhæng mellem præstationen i MuH og hver af de tre naturfag. Om der er en sådan sammenhæng, kan belyses ved at se på sammenhængen imellem 2019-prøven i MuH og 2019-prøverne i de tre naturfag. Dette er illustreret i Tabel 8.

Tabel 8: Korrelation mellem scorer for elever, der både har besvaret MuH (N=37.897) og et naturfag i 2019

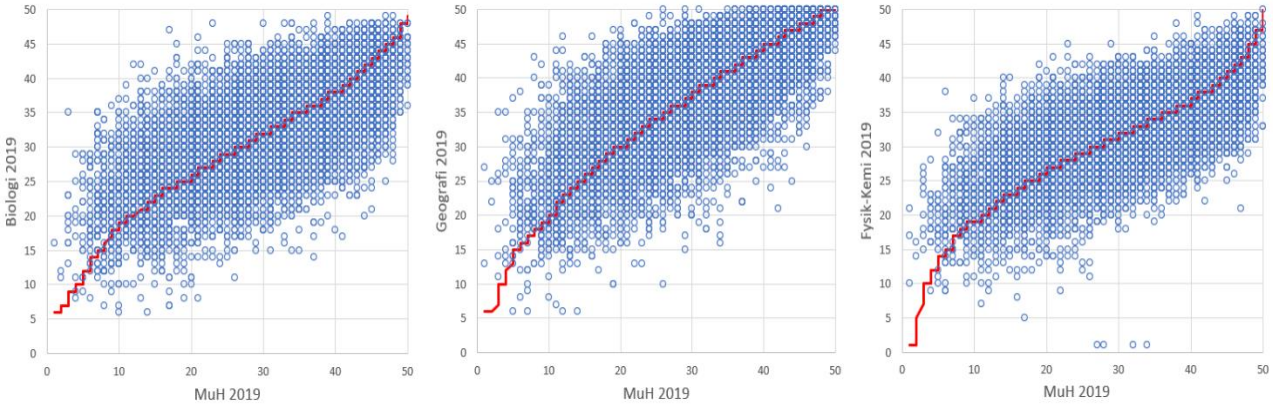
Udtræksfag, maj 2019		Score i naturfag 2019		Score i MuH, 2019		Korrelation	Antal elever (N=37.897)
		Gns. score	Std.afv.	Gns. score	Std.afv.		
Elever udtrukket i:	Biologi	32,5	6,96	30,8	10,39	0,66	12.220
	Geografi	37,6	7,74	30,8	10,30	0,74	12.780
	Fysik-Kemi	32,0	6,07	30,9	10,29	0,68	12.898

Note: Elever med nul-score i en af prøverne er frasorteret: Der er frasorteret 53 elever, heraf havde 30 i 0 i matematik, 4 havde 0 i Biologi, 6 havde 0 i Geografi og 18 havde 0 i Fysik/kemi. Derudover er der registreret 33 elever med 3 scorer – nemlig 0 i et af fagene og derudover 2 positive scorer. Disse 33 elever indgår i opgørelsen med de positive scorer.

Som det fremgår af Tabel 8, er der en pæn korrelation mellem MuH og hver af de tre naturfag (mellem 0,66 og 0,74). Dette bestyrker den faglige holdbarhed i at benytte matematikscoren som benchmark for fastlæggelse af udgangsniveauet i omsætningstabellen for de tre nye naturfagsprø-

ver. Som det endvidere fremgår af Figur 11, er der (som forventet pga. fordelingernes forskellighed, jf. Figur 9) tydelige, men ikke helt lineære sammenhænge mellem MuH og hvert af de tre naturfag.

Figur 11: Sammenhængen mellem Maj 2019 prøveresultatet i Matematik u. hj. og hvert af de tre naturfag



Den ikke-lineære sammenhæng sætter en naturlig grænse for størrelsen af korrelationen (estimat af den lineære sammenhæng). Set i lyset af de ovenstående beskrevne sammenhænge vurderes det som umiddelbart mest hensigtsmæssigt at arbejde videre med metode 2. Konsekvensen vil være illustreret i Figur 11, hvor den røde linje (equirank ækvivaleringskurven) definerer den reviderede (genberegnede) omsætningstabel.

5.2.3 Sammenstilling af omsætningstabeller ud fra de 2 principper

Nedenfor anskueliggøres, hvordan det ville have kunnet påvirke omsætningstabellerne fremadrettet, hvis man på denne måde benytter matematik som en reference. Med afsæt i 2019-prøven og resultaterne herfra (jf. Figur 11), er det beregnet, hvilke ændringer det ville have betydet for omsætningstabellerne i 2019. Dette er gengivet i Tabel 9 (genberegnet, Equirank) sammen med den gældende omsætningstabel (gældende dvs. beregnet ved ækvivalering ift. 2017).

Tabel 9: Omsætningstabeller skriftlige naturfag 2019, gældende og genberegnede med MuH som reference

Karakter	Matematik uden hjælpemidler (MuH)		Biologi				Geografi				Fysik-Kemi			
			Gældende		Genberegnet		Gældende		Genberegnet		Gældende		Genberegnet	
	Fra	Til	Fra	Til	Fra	Til	Fra	Til	Fra	Til	Fra	Til	Fra	Til
-3	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	10	1	21	1	18	1	23	1	20	1	21	1	19
2	11	16	22	26	19	23	24	31	21	26	22	26	20	23
4	17	25	27	31	24	29	32	38	27	34	27	32	24	29
7	26	36	32	37	30	35	39	44	35	41	33	36	30	34
10	37	43	38	42	36	40	45	48	42	46	37	41	35	39
12	44	50	43	50	41	50	49	50	47	50	42	50	40	50

Note: Alle oplysninger og omsætningtabeller er beregnet på grundlag af Equirankmetoden.

Som det fremgår, vil det for alle karakterer i alle tre prøver betyde en forskydning i retning af, at der skal færre point til at udløse en givet karakter (2-3 point i biologi og fysik-kemi, 2-5 point i geografi). Det vil altså betyde en forskydning af omsætningstabellen i retning af højere karakterer for samme præstation på alle niveauer i alle tre prøver. Den genberegnede omsætningstabel er designet til, at karaktergennemsnittene i alle tre naturfagsprøver kommer så tæt på gennemsnittet for de samme elevers karakter i matematik uden hjælpemidler.

Med de nye prøveformater i naturfag (forlænget tid og flere opgaver) er der flere point at give karakter ud fra. Det betyder, at arbejdet med at ramme den ønskede fordeling i alle tilfælde bliver

enklere og mere præcist. En del af problemet med at fastsætte omsætningstabellerne i de tre naturfag med afsæt i det hidtidige prøveformat (ca. 50 point pr. prøve) er, at når besvarelserne fordeler sig med en relativt lav spredning (især for fysik-kemi og biologi, se fordelingerne i Figur 9 og standardafvigelser i Tabel 8), så bliver det relativt få point, der adskiller de enkelte karakterer midt på skalaen. Med de nye prøver, der er forlænget til en time, og hvor der er væsentligt flere items, vil dette problem sandsynligvis være afhjulpet i betydelig grad.

5.3 Beregnet omsætningstabel med afsæt i 2021-prøven

I forbindelse med årets forcensur blev der gennemført øveprøver i de tre naturfag med afsæt i det opgavesæt, som var skrevet til maj 2020-prøven, men altså først afviklet i foråret 2021. Samtidig blev gennemført øveprøver i matematik uden hjælpemidler med afsæt i maj 2017-sættet, som har dannet grundlag for forcensuren siden 2017. Dette datamateriale giver grundlag for, at man fremadrettet kan anvende det nye prøveformat for naturfagene og data fra 2021 øveprøven som grundlag for fremtidens forcensur i forbindelse med onlineprøver.

Dette forbereder samtidig, at onlineprøven kan blive et fagligt solidt grundlag for karaktergivningen, også såfremt forsøgene med de nye naturfagsprøver med åbne besvarelsesformater resulterer i, at de implementeres i fuld skala.

5.3.1 Besvarelser af øveprøver 2021 som datagrundlag

Af Tabel 10 fremgår resultaterne fra øveprøverne i marts-april, 2021. I sidste kolonne fremgår antallet af frasorterede besvarelser, og i næstsidste kolonne antallet, der har kunnet anvendes i forbindelse med disse analyser. Opgavebesvarelser, der ikke vurderes som valide mål for elevernes dygtighed, er således udeladt i analyserne. I praksis kræves, at scoren er over 0 og et vist minimum af besvarede opgaver for at scoren kan betragtes som et gyldigt mål for elevens dygtighed. Det konkrete antal fastlægges ud fra en konkret vurdering fra datamaterialet og prøven indhold. For matematik er tærsklen sat ved, at der skal være besvaret mindst 10 items (matematikprøven består hovedsageligt af åbne spørgsmål som det det kræver en indsats at besvare). I naturfagene kræves mindst halvdelen af opgaverne besvaret. Af tabellen fremgår den gennemsnitlige råscore (det samlede antal rigtige svar i de gyldige besvarelser) for den samlede prøve. Gennemsnittet af råscore er beregnet på baggrund af besvarelserne i den afholdte prøve (FP), og antallet af items i denne fremgår under tabellens kolonne antal items/FP2021. Endvidere fremgår det totale antal items i et foreslået fremtidigt forcensursæt (FC21), hvor de items, der ikke fungerer rimeligt i de tre naturfagsprøver, er frasorteret. Frasorteringen er sket på grundlag af Raschanalyser af de enkelte prøvers opgavebesvarelser.

Tabel 10: Elever, der deltog i øveprøver i marts-april 2021 (efter datarensning)

	Gns. råscore	Antal items		Antal elever	Antal frasorterede
		FP2021	FC21		
Matematik uden hj.midler maj 2017 (MuH)	31,2	50	-	10.298	78
Biologi, maj 2021	51,4	76	62	2.895	25
Geografi, maj 2021	48	75	59	2.916	21
Fysik/Kemi, maj 2021	44,6	76	66	2.864	31

Der beregnes to sæt af omsætningstabeller: Udgangspunktet er et sæt af omsætningstabeller for de tre naturfag, der baseres på de oprindelige naturfaglige prøver indeholdende alle item (jf kolonnen FP2021 i Tabel 10). Disse fremgår af afsnit 5.3.2

Men yderligere er beregnet et sæt omsætningstabeller (jf. afsnit 5.3.3), hvor de items, der statistisk set forringer grundlaget for omsætningstabellerne er fjernet fra de prøver der indgår (jf kolonnen FC21 i Tabel 10).

5.3.2 Ækvivalering på grundlag af maj 2021-prøverne

I Tabel 11 vises korrelationen mellem de forskellige prøvesæt i naturfag, maj 2021 og prøven i matematik uden hjælpemidler maj 2017.

Tabel 11:Tabel over korrelationer mellem prøvesættene indbyrdes

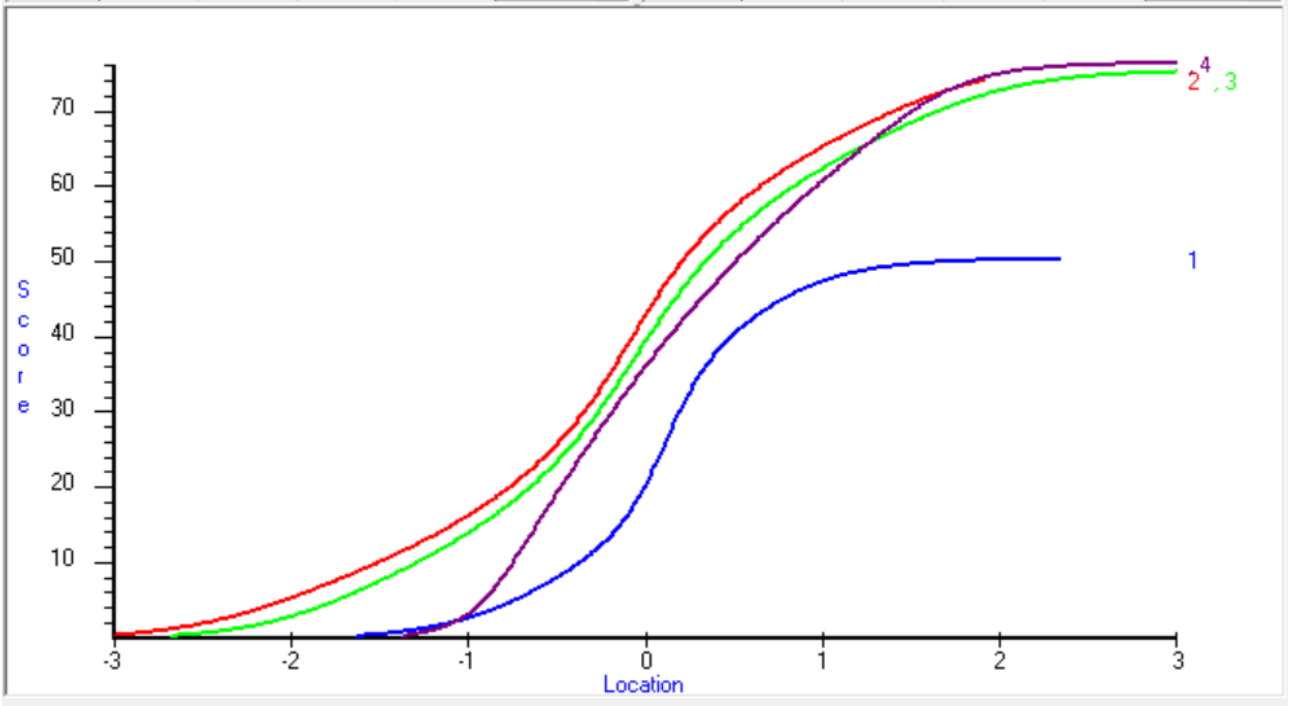
	MuH17	Bio FP2021	Geo FP2021	FK FP2021
Matematik uden hjælpemidler, maj 2017	1	0,650	0,700	0,688
Biologi, maj 2021 (FP2021)	0,650	1	0,816	0,802
Geografi, maj 2021 (FP2021)	0,700	0,816	1	0,804
Fysik/Kemi, maj 2021 (FP2021)	0,688	0,802	0,804	1
Antal elever	10.298	2.895	2.916	2.864

Som det fremgår, er korrelationen mellem scorerne i de tre naturfag indbyrdes større end med MuH, hvilket er forventeligt givet et relativt stærkt slægtskab mellem de tre fag, både hvad angår fagligt indhold og opgaveformater (udelukkende MC-formater). Det fremgår dog også, at korrelationen med matematik (på trods af det anderledes opgaveformat) er ganske høj og svarer til det, der blev fundet i 2019 jf. Tabel 8. I projektet har der (jf. afsnit 4 om Metoder til ækvivalering) været anvendt to forskellige teknikker og modeller til ækvivalering. Af Figur 12 fremgår en ækvivalering af matematik og de 3 naturfag med modelmetoden (jf afsnit 4).

Figur 12: Ækvivalering med model-metoden af 4 skalaer (MuH17, Biologi, Geografi og Fysik/kemi FP2021)

Item	M17	B21	G21	FK21
M17	X			
B21		X		
G21			X	
FK21				X

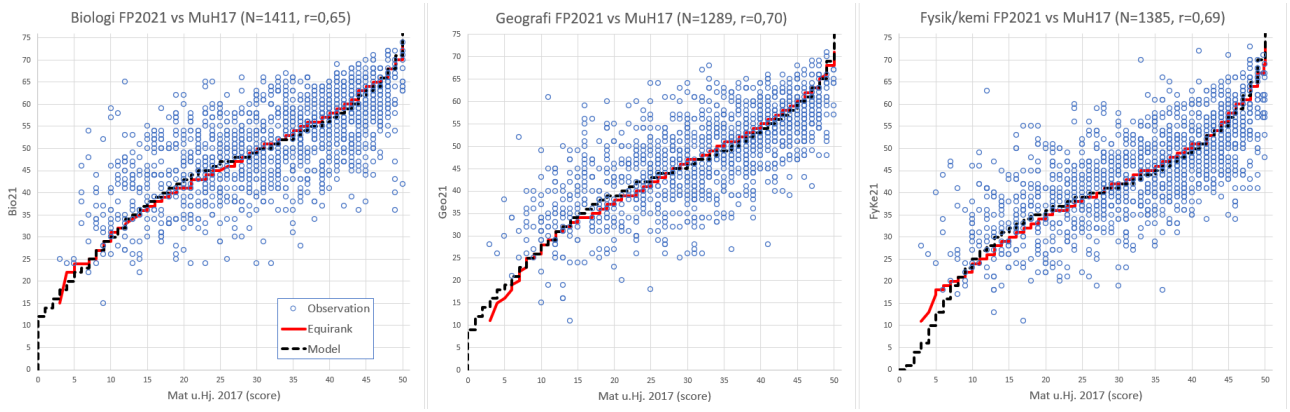
Score	M17	B21	G21	FK21
0**	-1,636	-3,073	-2,683	-1,374
1	-1,317	-2,724	-2,358	-1,202
2	-1,102	-2,479	-2,134	-1,090
3	-0,956	-2,306	-1,979	-1,017



Princippet er, at man på den vandrette akse kan aflæse elevens dygtighed (Person Location) på en fælles logit skala, og på den lodrette hvor mange rigtige svar eleven med den dygtighed kan forventes at score i hver af de 4 prøver. Der er forklaret lidt mere uddybende til denne type af visning i næste afsnit.

I Figur 13 er vist resultater af to forskellige ækvivaleringsmetoder: Dels ækvivaleringen baseret på klassiske metoder (equirank, den røde kurve der viser en ækvivalering baseret på rangorden af råscorer) og dels ækvivaleringen baseret på Rasch-analyse (den sorte stiplede kurve, modelmetoden, hvor hver prøve behandles som ét polytomt item hvor blanke svar i prøven behandles som forkerte svar – det kaldes også for et superitem eller en testlet)[4].

Figur 13: Sammenhængen mellem resultatet i MuH og hver af prøvesættene i de tre naturfag (forår 2021)



Nedenfor er vist den omsætningstabel, som elevbesvarelserne i foråret 2021's forcensur og øveprøver har ført frem til for hver af de tre dedikerede forcensursæt i Biologi, Geografi og Fysik/kemi.

Tabel 12: Beregnede omsætningstabeller (equirank) for naturfagenes forcensursæt med MuH17 som reference

Karakter	Matematik u. hjælpemidler, maj 2017 (MuH17)		Biologi, FP2021		Geografi, F P2021		Fysik/kemi, F P2021	
	Fra	Til	Fra	Til	Fra	Til	Fra	Til
-3	0	0	0	0	0	0	0	0
0	1	10	1	30	1	27	1	23
2	11	16	31	38	28	34	24	31
4	17	25	39	45	35	42	32	38
7	26	36	46	54	43	51	39	47
10	37	43	55	61	52	57	48	54
12	44	50	62	76	58	75	55	76

Modellerne har hver især fordele og ulemper, men i stort set alle de hidtidige tilfælde i nærværende projekt har det vist sig, at resultaterne fører frem til stort set de samme omsætningstabeller. Modsat i ovennævnte eksempel, hvor resultaterne er stort set identiske findes der er tilfælde, hvor de to metoder giver væsentlig forskellige resultater, og hvor modelmetoden er at foretrække frem for den equirank som anvendes her. Dette er omtalt nærmere i afsnit 4.

[4] Se mere uddybende beskrivelse i afsnit 4 eller i Bilag 9.2, Rapporten Evaluering af forcensur 2017-19 (UvM, 2019). Link: https://nordicmetrics.com/wp-content/uploads/2020/02/Bilag-4A.1-Rapport-fra-tidligere-forcensur.pdf

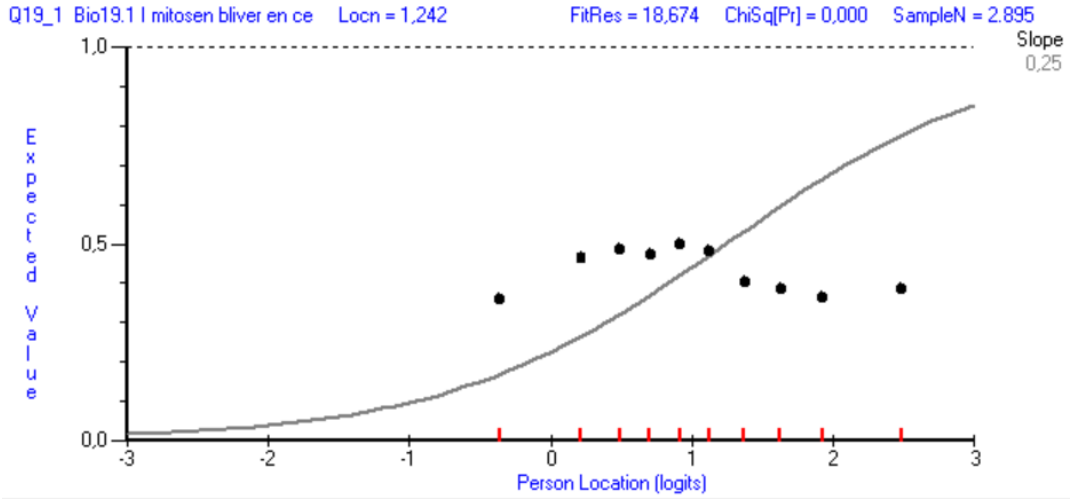
5.3.3 Ækvivalering på grundlag af tilpassede forcensursæt

Som nævnt er der yderligere beregnet et sæt omsætningstabeller, hvor de items, der statistisk set forringer grundlaget for omsætningstabellerne er fjernet fra de prøver der indgår (jf kolonnen FC21 i Tabel 10).

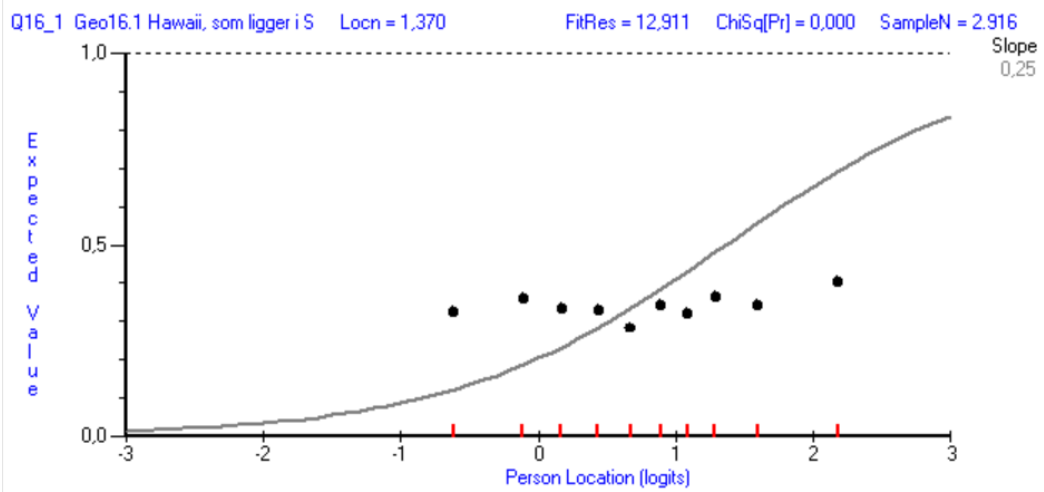
Disse omsætningstabeller beregnes efter Raschanalyser af de tre naturfagsprøver. Det skal understreges, at der ikke kan forventes en tilpasning til alle Raschmodellens krav pga. den faglige spredning, der er i naturfagene. Men det vil være rimeligt at forvente en form for fællesmængde af faglighed indenfor det enkelt naturfag, når der skal gives en samlet karakter.

Ved en indledende gennemgang blev identificeret opgaver, hvis items alle diskriminerede under det sædvanlige minimumskrav (opgaver hvor lavt præsterende elever har større sandsynlighed for at svare rigtigt end dem der scorer højt i de øvrige opgaver). Tre eksempler fra hvert sit naturfag på sådanne opgaver er vist i Figur 14, Figur 15 og Figur 16.

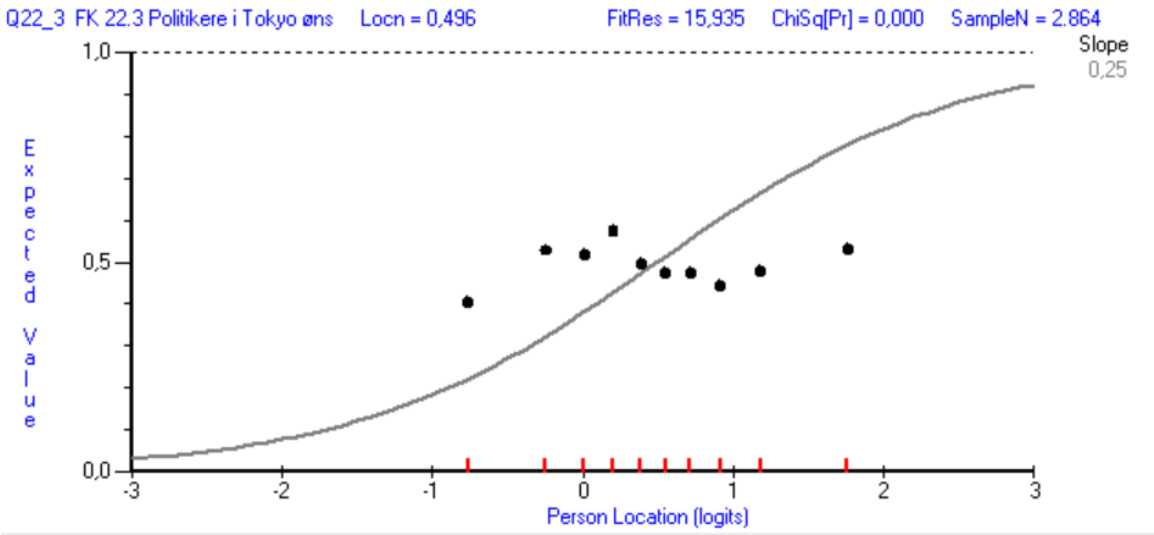
Figur 14: Eksempel på problematisk item fra Biologi FP2021-sættet



Figur 15: Eksempel på problematisk item fra Geografi FP2021-sættet



Figur 16: Eksempel på problematisk item fra Fysik/kemi FP2021-sættet



Disse opgaver bidrager til, at elever der scorer lavt i den samlede prøve (de lavt præsterende elever) har mindst lige så stor sandsynlighed for at svare rigtigt som de dygtigste elever. Opgaver som disse forringer prøvens måleegenskaber og det vil derfor give en mere retvisende vurdering af elevernes dygtighed, hvis de udgik af opgavesættet, især når det skal anvendes til forcensur.

I analysen blev disse og tilsvarende items udeladt af opgavesættet og Raschanalysen blev gentaget for de resterende opgaver. Herefter blev vurderet individuelt, hvilke items der fungerede på en sådan måde, at de medvirkede til at reducere den samlede analyses statistiske pålidelighed (reliabilitet). Ud fra de statistiske oplysninger (herunder ICC-kurver, jf. Bilag 9.6) blev de dårligst fungerende items identificeret. Ikke alle tilbageværende items lever op til Raschmodellens krav (fx et Fit-residual < 2,5), men ud fra en pragmatisk tilgang anbefales det, at de tages med, da de bidrager positivt til, at det samlede opgavesæt giver en mere reliabel måling (PSI) af elevernes dygtighed i faget, end hvis de frasorteres.

Med denne fremgangsmåde blev der frasorteret hhv. 14, 16 og 10 items i opgavesættene for biologi, geografi og fysik/kemi maj 2021 (for en oversigt over frasorterede opgaver/items, se bilag 9.6, 1-3, Responsanalyser til opgavekommissioner), hvorefter det antal items, som er vist i Tabel 10 (kolonne FC21), resterer.

De tilpassede opgavesæt er dem, det umiddelbart anbefales at lade være grundlag for den fremtidige forcensur med mindre at der er andre grunde end de rent statistiske der taler imod[5], og de kaldes herefter forcensursættene i afsnit 5.3.3I Tabel 13 illustreres korrelationen mellem de forskellige forcensursæt i naturfagene og prøven i matematik uden hjælpemidler (MuH).

Tabel 13:Tabel over korrelationer mellem de tilpassede forcensursæt i naturfagene indbyrdes og MuH17

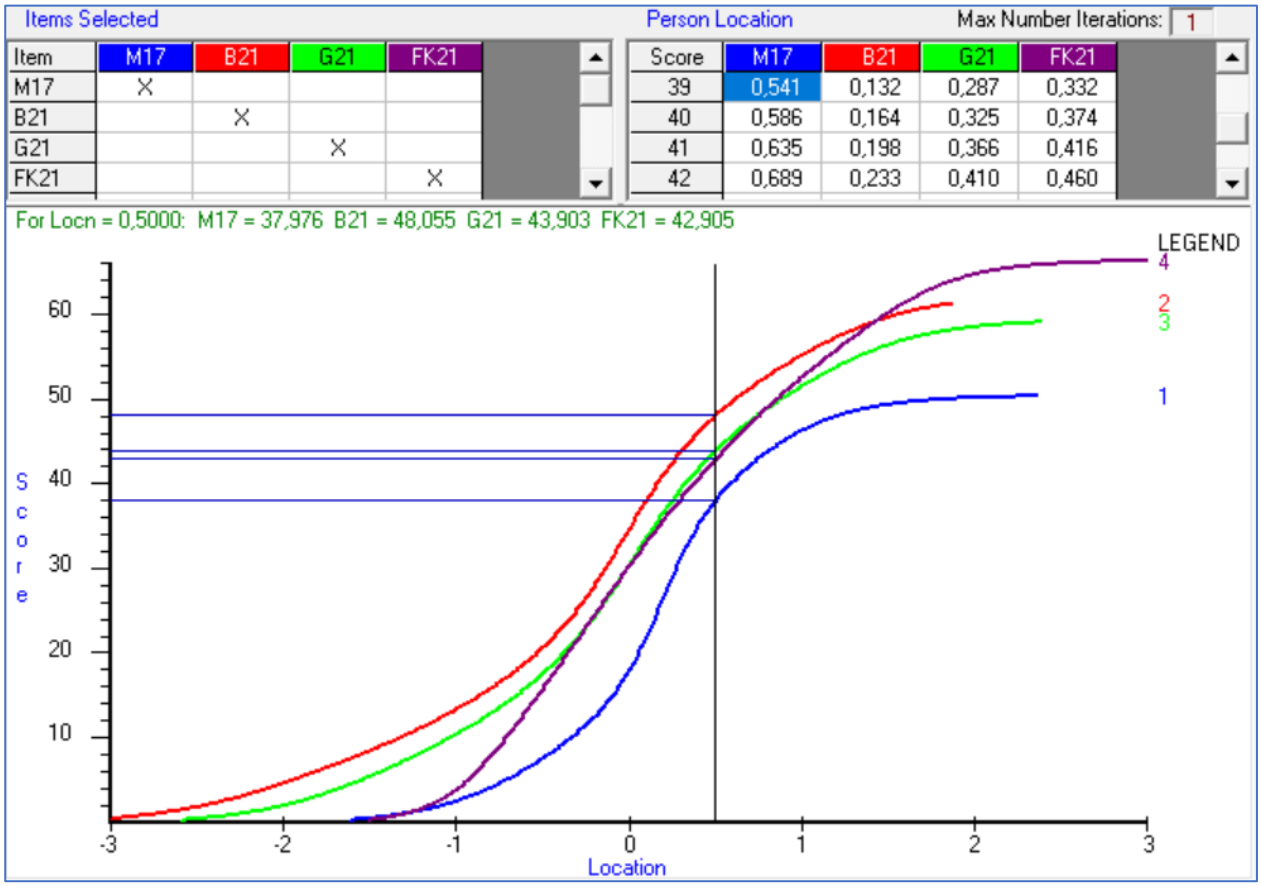
	MuH17	B21FC	G21FC	FK21FC
Matematik uden hjælpemidler, maj 2017 (MuH17)	1,000	0,656	0,703	0,692
Biologi, maj 2021, forcensursæt (B21FC)	0,656	1,000	0,821	0,806
Geografi, maj 2021, forcensursæt (G21FC)	0,703	0,821	1,000	0,810
Fysik/Kemi, maj 2021, forcensursæt (FK21FC)	0,692	0,806	0,810	1,000
Antal elever	10298	2895	2916	2864

[5] En endelig beslutning om forcensursæt bør inkludere en gennemgang af opgaverne og en faglig indholdsmæssig vurdering, som det ikke har været mulig at foretage, da opgavesættene ikke har været tilgængelige ved denne analyse, der således alene beror på responsdata og statistiske analyser heraf.

Som det fremgår, er korrelationen mellem de tre forcensursæt (FC) i naturfag og MuH den samme eller marginalt større end i den tilsvarende opgørelse for hele opgavesættet (jf.Tabel 11). Stærke ligheder er forventeligt givet den store fællesmængde, men det kan dog bemærkes at korrelationen alt andet lige vokser når der frasorteres de dårligst fungerende items.

Af Figur 17 fremgår hvordan scorerne i de 4 skalaer (MuH17, B21FC, G21FC, FK21FC) forholder sig til hinanden, når de er belyst ved modelmetoden (Raschanalyse).

Figur 17: Ækvivalering med model-metoden af 4 skalaer (MuH17, B21FC, G21FC, FK21FC)



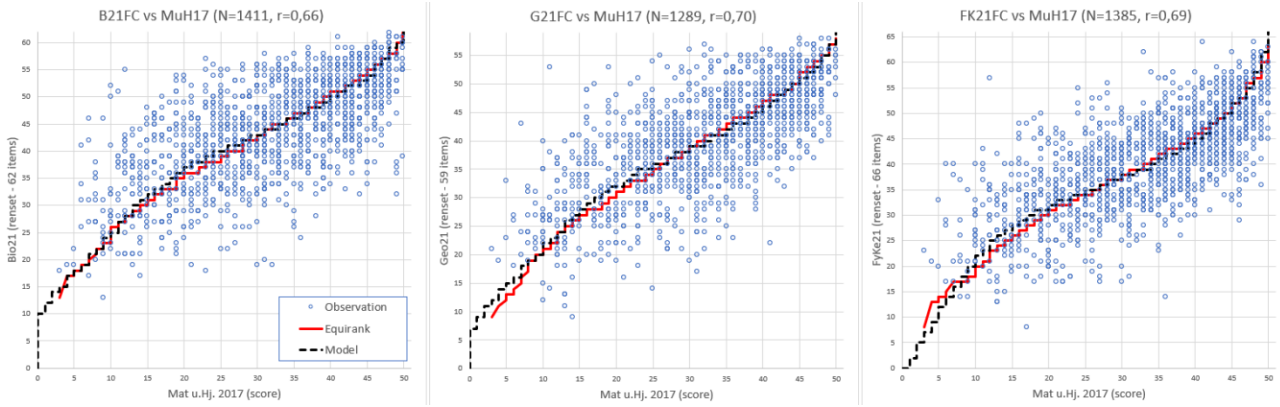
Denne figur fremstår tæt på identisk med Figur 12 – ved nærmere eftersyn findes at der her er noget færre items.

Princippet er, at man på den vandrette akse kan aflæse elevens dygtighed (Person Location) på en fælles logit skala – i figuren er markeret dygtigheden svarende til 0,5 logits. Skæringspunkterne med de fire kurver (vandrette linjer) markerer hvor mange rigtige svar der skal til i hver prøve for at opnå raschscoren 0,5 logits.

Af Figur 18 fremgår sammenhængen mellem resultatet fra de tilpassede resultater fra øveprøver/forcensur i foråret 2021 i opgavesættet i matematik (MuH 2017) og forcensursættene (forår 2021) i hver af de tre naturfag. Her er resultater af to forskellige ækvivaleringsmetoder: Dels ækvi-

valeringen baseret på klassiske metoder (equirank, den røde kurve der viser en ækvivalering baseret på rangorden af råscorer) og ækvivaleringen baseret på Rasch-analyse (den sorte stiplede kurve, modelmetoden, hvor hver prøve behandles som ét polytomt item – en såkaldt testlet)[6].

Figur 18: Sammenhængen mellem resultatet i MuH og Forcensursættene i hver af de tre naturfag (forår 2021)



I Tabel 14 (nedenfor) er vist den omsætningstabel, som elevbesvarelserne i foråret 2021's forcensur og øveprøver har ført frem til for hver af de tre dedikerede forcensursæt i Biologi, Geografi og Fysik/kemi.

Tabel 14: Beregnede omsætningstabeller (equirank) for naturfagenes forcensursæt med MuH17 som reference

Karakter	Matematik u. hjælpemidler, maj 2017 (MuH17)		Biologi, FC21		Geografi, FC21		Fysik/kemi, FC21	
	Fra	Til	Fra	Til	Fra	Til	Fra	Til
-3	0	0	0	0	0	0	0	0
0	1	10	1	25	1	21	1	21
2	11	16	26	33	22	28	22	28
4	17	25	34	40	29	35	29	34
7	26	36	41	47	36	42	35	41
10	37	43	48	52	43	48	42	49
12	44	50	53	62	49	59	50	66

6. Bedømmelse af prøverne i fremmedsprog

De tre udtrækprøver skriftlig engelsk, tysk og fransk har en del parallelle problemstillinger i forhold til naturfagene. De har fx til fælles, at de er alternative muligheder for udtræk i samme fagblok, og at de har et fagligt slægtskab indbyrdes. En væsentlig del af overvejelserne om naturfagenes indbyrdes relationer kan derfor forventes at være relevante også ved sammenligning af sprogprøverne, især i forhold til de tre fags indbyrdes forskelle i karakterniveau. Men der er også en række forskelle på opgavefelterne på især to områder.

For det første i forhold til bedømmelsesmetoder og itemformater, hvor naturfagsprøverne hver især består af en simpel række ligeværdige items, hvor sprogfagene består af en mere kompleks model, hvor der indgår tre multiple-choice prøver (med hver deres omsætningstabel) og en essayopgave, der alle fire skal sammenvejes til brug for beregning af én karakter.

[6] Se mere uddybende beskrivelse i afsnit 4 i denne rapport eller i Bilag 9.2, Rapporten Evaluering af forcensur 2017-19 (UvM, 2019). Link: https://nordicmetrics.com/wp-content/uploads/2020/02/Bilag-4A.1-Rapport-fra-tidligere-forcensur.pdf

For det andet i forhold til referenceprøven, hvor man på naturfagsområdet var nødt til at hente en reference udefra (MuH) for at skabe et fælles grundlag for en ny baseline, men hvor man med engelsk har en reference blandt de tre sprogfag, der kan danne afsæt for en baseline for tysk og fransk.

6.1 Udvikling i de gennemsnitlige karakterer i sprogfagene

De tre sprogfag hører til den humanistiske fagblok i udtræksfag.

Gennemsnitlige karakterer i udtræksfag ses af Tabel 15, og heraf fremgår, at karaktererne i tysk og fransk gennemgående ligger relativt lavt i fagblokken.

Tabel 15: Karakterer i udtræksfag fra folkeskolens prøver i 9. klasse

Prøve	2015/2016	2016/2017	2017/2018	2018/2019
Naturvidenskabelige udtræksfag				
Skriftlig **Biologi**	6,7	6,7	6,0	**6,1**
Skriftlig **Fysik/kemi**	6,9	6,4	5,8	**5,9**
Skriftlig **Geografi**	6,4	6,5	5,6	**5,7**
Praktisk/Mundtlig **Idræt**	7,8	7,8	7,7	**7,8**
Mundtlig **Matematik**	7,2	7,2	7,3	**7,2**
Humanistiske Udtræksfag				
Skriftlig **Engelsk**	7,2	7,3	7,3	**7,3**
Mundtlig **Fransk**	6,9	7,1	6,3	**6,3**
Skriftlig **Fransk**	6,5	6,8	6,1	**6,3**
Mundtlig **Historie**	7,8	7,8	7,7	**7,8**
Mundtlig **Kristendomskundskab**	7,8	7,9	7,8	**7,9**
Mundtlig **Samfundsfag**	7,9	7,8	7,8	**7,8**
Mundtlig **Tysk**	6,1	6,0	6,0	**6,3**
Skriftlig **Tysk**	7,8	7,1	6,5	**7,3**

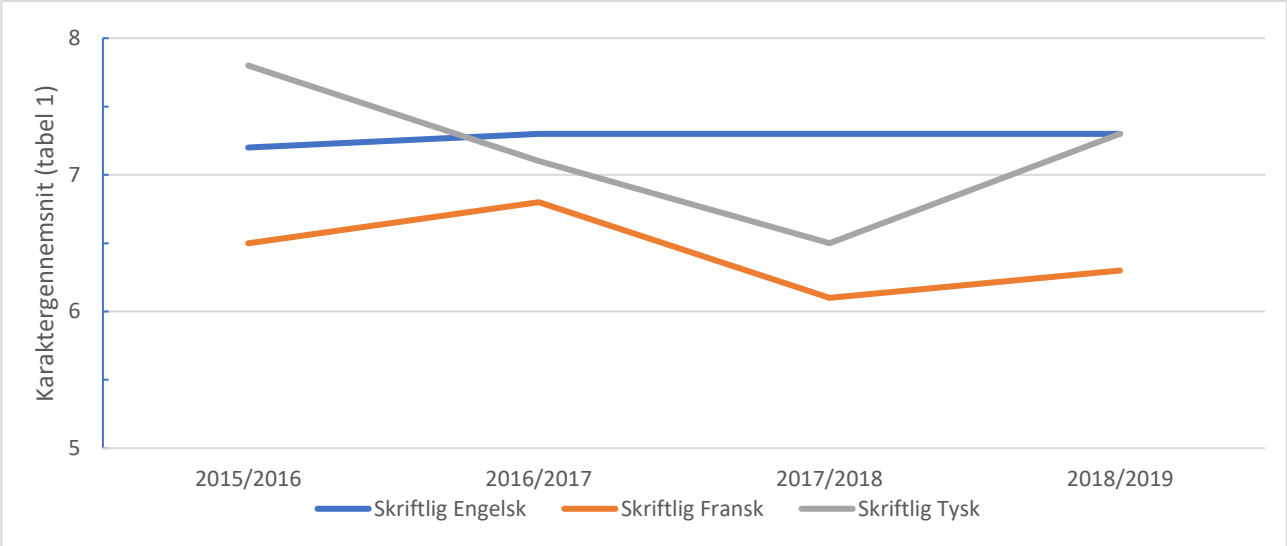
Note: Gennemsnittet udregnes blandt elever i 9. klasse med en karakter i den respektive prøve eller delprøver. Specialklasseelever og privatister indgår ikke i tabellen. Tallene for 2018/2019 indeholder ikke resultater fra sygeprøver.
Kilde: Indberettede karakterer til Styrelsen for It og Læring
Note: Se https://www.uvm.dk/-/media/filer/uvm/stat/pdf19/191114-elever-i-grundskolen-2018-19.pdf?la=da

I Figur 19 er vist karakterudviklingen i de skriftlige udtræksfag for hver af de tre fremmedsprog. Som det ses, er der meget små udsving i gennemsnitskarakteren i engelsk fra år til år, mens udsvingene er betydelige fra år til år især i tysk men også i fransk. Det skyldes givetvis en kombination af mange forskellige forhold (tilfældigheder i udtræk, forskellige antal elever, der tager prøverne), men prøveresultaternes sammenlignelighed fra år til år er en sandsynlig hovedforklaring.

Ideen med ækvivaleringsbaseret forcensur er netop, at der gennem tilpasning af omsætningstabellen korrigeres for prøvernes varierende sværhed fra år til år, således at udsving alene kan henføres til variationer i elevernes præstation på en sammenlignelig skala fra år til år.

Engelsk var det første fag, hvor der blev gennemført en ækvivaleringsbaseret forcensur (som en forsøgsordning i undervisningsministeriet i 2010), og der har i engelsk været gennemført ækvivaleringsbaseret forcensur lige siden.

Figur 19: Skriftlige udtræksfag, sprog



I 2017 blev prøverne i engelsk besvaret på papir, og oplysninger om besvarelserne blev i forcensurprojektet indsamlet på skolerne i Excel-baserede retteark. Dette sæt fungerede godt og har siden dannet grundlag for forcensur. Fra 2018 blev engelskprøverne altovervejende gennemført digitalt.

I tysk og fransk har forcensuren i samme periode været gennemført af opgavekommissionerne på grundlag af årets data (i 2017 på grundlag af eget indsamlet sample, og i 2018 og 2019 gennem digital afvikling) men uden en forcensurprøve som reference.

For at rette op på dette var det intentionen at gennemføre forcensur 2020 og 2021 i både fransk og tysk (hvilket ikke skete, da disse prøveafholdelser blev aflyst begge år) gennem anvendelse af 2017-prøverne som reference. Men i modsætning til engelsk og de øvrige prøver, hvor der blev foretaget forcensur i 2017, blev der ikke indsamlet et stærkt datagrundlag til vurdering af de to prøvesæts (skriftligt fransk og tysk, maj 2017) egnethed som ”baseline”, da de to prøver i 2017 blev besvaret på papir, og kun karakterer blev registreret (besvarelser og scorer blev ikke registreret og indberettet af censorer).

6.2 Bedømmelse i sprogprøver

Der er ingen tvivl om, at danske elever normalt klarer en samtale bedre på engelsk end på tysk eller fransk (og naturligvis endnu bedre på dansk). Men det vil ikke være rimeligt/fair at bedømme fagene ud fra en sådan kriterieskala, hvor det er ”den indfødtes” sprogfærdighed, der sætter benchmarket, og i så fald vil det være uklogt at vælge fransk i skolen, hvis man som elev interesserer sig for at opnå det bedst mulige gennemsnit i folkeskolens eksamen. For at ligebehandle de tre sprogfag må præstationer med andre ord vurderes i forhold til kravene i fælles mål, og hvad der rimeligvis kan forventes af danske elever med hensyn til niveau.

De tre skriftlige udtræksprøver i fremmedsprog engelsk, tysk og fransk er strukturmæssigt identiske: For hver prøve gives fire karakterer. Der er tre sæt opgaver, der kan rettes automatiseret ud fra objektive kriterier (dvs. uden efterfølgende censorvurdering af besvarelsen), og der fastsættes en omsætningstabel for hver af disse. Yderligere er der en essayopgave, der skal bedømmes af lærer og/eller censor ud fra overordnede retningslinjer. De tre sæt opgaver er en læse- og en lytteprøve (hver 20-25 items), der hver vægter 25% af karakteren, og en prøve i sprog og sprogbrug (21-30 items) der tæller for 15% af karakteren. Samlet afgør disse tre delprøver (med i alt 62-80 items) 65% af karakteren. De resterende 35% af karakteren gives på grundlag af essayopgaven,

hvor bedømmelsen sker af en censor uafhængigt af de objektive prøver. De fire delprøver bedømmes hver for sig med en karakter, og den samlede karakter beregnes som et vægtet gennemsnit (afrundet til nærmeste karakter) af de fire karakterer.

Vægtningen mellem de fire delprøver i faget er en faglig/politisk beslutning. Med hensyn til forholdet mellem essayopgaven (35%) og de tre automatisk rettede prøver under ét (i alt 65%) vurderes det ikke umiddelbart muligt at anvise entydigt bedre sammenvejningsmetoder.

Men hvordan det virker, når de tre automatisk rettede delprøver scores og samles til én karakter, analyseres nedenfor, og det skal samtidigt belyses, om man med enkle midler kan gøre det på en mere retfærdig og enkel måde, så bedømmelsen bliver mere retvisende for elevens præstation.

6.3 Konsekvenser af tre delprøver i scoringsmodel

Der sker altså en omsætning af delprøver til karakterer, hvorefter der tages et gennemsnit af disse.

I praksis kan dette medføre et betydeligt tab af opløsning undervejs i beregningsprocessen, da det rent matematisk svarer til, at der rundes af i mellemregninger. Dette kan i sidst ende betyde, at det snarere er beregningsprincipper og tilfældigheder end elevens faglige præstation, der afgør om eleven får fx 4 eller 7.

Tabel 16: Omsætningstabeller i sprog: Eksempel fra engelsk, FP2019, de tre delprøver i skriftlig engelsk

			Lytning (Lyt)	Læsning (Læs)	Sprog & Sprogbrug (S&S)
Karakteren	12	gives for fgl. point	**24-25**	**24-25**	**26-30**
Karakteren	10	gives for fgl. point	**21-23**	**20-23**	**22-25**
Karakteren	7	gives for fgl. point	**16-20**	**15-19**	**17-21**
Karakteren	4	gives for fgl. point	**11-15**	**10-14**	**14-16**
Karakteren	2	gives for fgl. point	**8-10**	**6-9**	**11-13**
Karakteren	0	gives for fgl. point	**1-7**	**1-5**	**1-10**
Karakteren	-3	gives for fgl. point	**0**	**0**	**0**

Dette kan illustreres med et eksempel: Af Tabel 16 fremgår omsætningstabellerne for hver af de tre delprøver. En elev svarer rigtigt på hhv. 15, 14 og 17 opgaver (dvs. svarende til karaktererne 4, 4 og 7) i lytning (Lyt), læsning (Læs) og sprog & sprogbrug (S&S). Eleven får 4,7 afrundet til karakteren 4 i vægtet gennemsnit af de tre prøver.

Hvis eleven svarer blot én opgave mere rigtigt i enten Læs eller Lyt, får eleven karakteren 7 i delprøven, og gennemsnittet bliver således 5,8, som afrundet giver karakteren 7. Derimod kan eleven svare op til 7 opgaver mere rigtigt i S&S (og få 10 i denne delprøve, hvilket giver samlet gennemsnit 5,4 – afrundes til 4) og stadig kun få den samlede karakter 4. Disse forskelle er affødt af vejningen (at S&S tæller mindre i gennemsnittet end Lyt og Læs), at der regnes på gennemsnit af karakterer (med betydelige fejl som følge af afrundede mellemregninger) og princippet i omsætningstabellen.

Men tilbage står, at det gældende princip betyder, at ét yderligere rigtigt svar på én delskala, kan betyde mere end yderligere 7 rigtige svar på en anden delskala. Hvis det skal kunne forklares, må det rimeligvis betyde, at opgaverne på de tre delskalaer måler noget forskelligt, som ikke er direkte sammenligneligt. Vi skal nedenfor undersøge, om vi kan finde tegn på, at delprøverne måler noget forskelligt. Engelsk bliver brugt som case, fordi det er her, vi finder det største og mest egnede da-

tagrundlag. Da de tre delskalaer måler det samme, og bedømmelsesprincipperne i øvrigt er sammenlignelige fremmedsprogene imellem, vil det være en rimelig antagelse, at det, vi finder som gældende for skriftlig engelsk, også gælder for tysk og fransk.

6.4 En samlet prøve overfor flere delprøver, engelsk, maj 2019

I Tabel 17 nedenfor vises korrelationen mellem delprøverne og den samlede prøve.

Tabel 17: Korrelation mellem råscorer (delprøver engelsk, 2019), N=11.709

	Eng. Læsning	Eng. Lytning	Eng. Sprog	Eng. Samlet
Eng. Læsning	1,00	0,79	0,80	0,94
Eng. Lytning	0,79	1,00	0,72	0,90
Eng. Sprog	0,80	0,72	1,00	0,92
Eng. Samlet	0,94	0,90	0,92	1,00

Der ses en rimelig korrelation mellem delprøverne indbyrdes. Den prøve, der udskiller sig mest, er lytteprøven, hvilket giver god mening, da den i afviklingsform (lytteopgaver) afviger fra de to andre. Så spørgsmålet er, om de tre skalaer måler noget substantielt forskelligt, eller om de med lige så god mening kunne betragtes som en fælles skala i forståelse af engelsk. En metode til statistisk at undersøge, om de tre prøver måler hvert deres konstrukt, er at anvende en eksplorativ faktoranalyse.

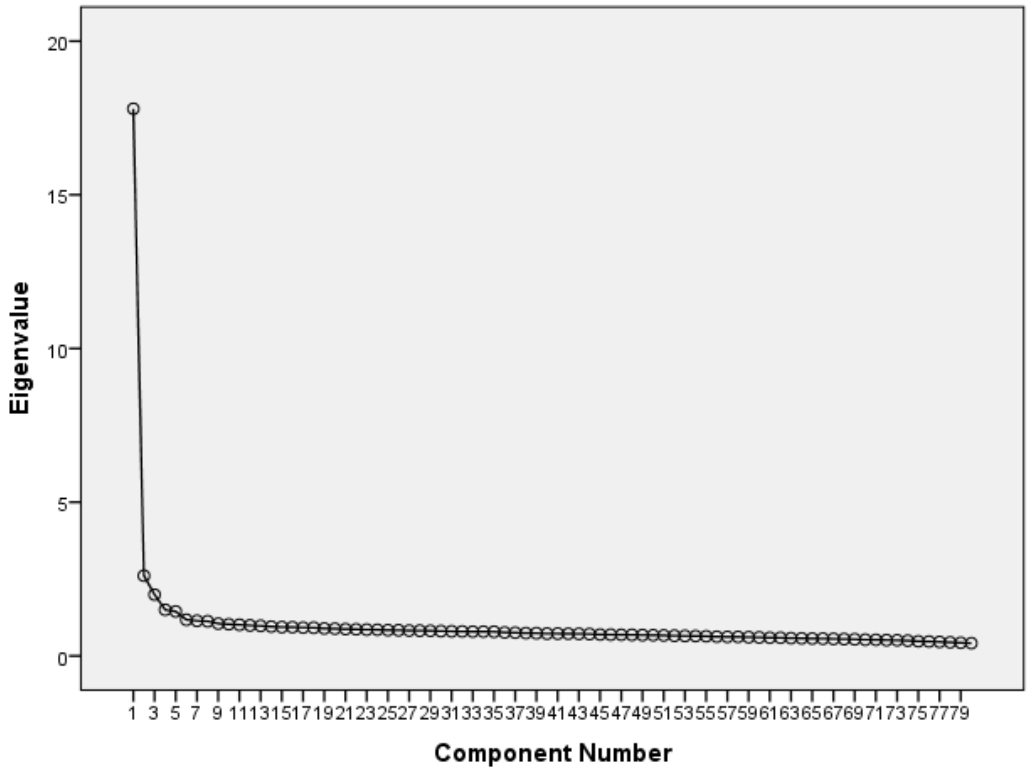
6.4.1 Faktoranalyse af 2019-prøven i engelsk

Der er derfor gennemført en *Principal Component Analyse* (PCA - en version af faktoranalyse) hvor antallet af komponenter (dvs. dimensioner eller faktorer) er fastsat til tre for at undersøge, om de tre delskalaer viser sig som hver deres konstrukt (component).

Af Figur 20 fremgår et Scree-plot af eigenværdier ved PCA-analyse af alle 80 items i engelsk. Dette illustrerer, hvor stor en andel af variationen, der kan forklares med hvor mange komponenter.

I et scree-plot opgøres for hver komponent en ”eigenvalue”, der er en parameter, der er udtryk for, hvor stor en del af den samlede variation den enkelte komponent kan forklare. Der skal maksimalt anvendes lige så mange komponenter, som antallet af items (her 80) for at forklare 100 procent af variationen. Som det fremgår af Figur 20, viser det sig, at der kun er én fremtrædende komponent (med en eigenværdi på 17,8).

Figur 20: Scree-plot af eigenværdier ved PCA-analyse af alle 80 items i engelsk



Af Tabel 18 fremgår det, at den første komponent forklarer 22% af den totale variation, men at de efterfølgende komponenter kun har meget begrænset forklaringskraft af den samlede variation.

Tabel 18: Andel af den forklarede variation med de tre komponenter

	Initial Eigenvalues (Extraction Sums of Squared Loadings)			Rotation Sums of Squared Loadings
Component	Total	% of Variance	Cumulative %	Total
1	17,799	22,249	22,249	16,105
2	2,608	3,259	25,508	8,228
3	1,993	2,491	28,000	6,608

Det ses endvidere, at de tre komponenter kun forklarer 28% af den samlede variation. Et normalt krav til en velfungerende faktoranalyse er, at mindst halvdelen af variationen skal kunne forklares – hvilket må siges at være uden for rækkevidde her, uanset om flere komponenter inddrages.

Af Tabel 19 fremgår, hvor meget de enkelte items loader på hver af de tre komponenter.

Der er langt flest items, der loader mest på den første komponent. Alle tre delskalaer i engelskprøven er repræsenteret, men items i Sprog & Sprogbrug (S&S) er noget underrepræsenteret her (dog forekommer 10 items, dvs. 1/3 af alle S&S-items i første komponent). Bortset fra det er der så godt som ingen sammenhæng med, hvilken delprøve item'et kommer fra, og hvilken komponent det loader på. De to efterfølgende komponenter er domineret af S&S men forklarer hver især under 3 pct. af den samlede variation. Så, som det også ovenfor var konklusionen, er der næppe belæg for at arbejde med mere end én betydende komponent (faktor).

Denne komponent kunne måske betegnes som 'engelsk forståelse'.

Tabel 19: Pattern matrix (Extraction Method: PCA, Rotation Method: Oblimin with Kaiser Normalization).

1. component dominerer			
Item	Component		
	1	2	3
L22	0,680	-0,029	0,063
L02	0,662	-0,203	-0,061
L18	0,642	-0,189	0,002
L24	0,633	0,037	0,039
R24	0,602	0,074	0,017
L14	0,592	-0,012	0,059
S16	0,591	-0,136	-0,232
S20	0,590	-0,213	-0,239
S19	0,567	-0,069	-0,196
L03	0,560	-0,111	-0,042
S11	0,557	-0,160	-0,309
L19	0,545	0,040	0,105
R03	0,534	-0,038	-0,077
L17	0,533	-0,011	0,038
S18	0,517	-0,105	-0,223
S17	0,515	-0,110	-0,244
L15	0,499	0,144	0,032
L09	0,498	0,024	-0,093
S15	0,496	-0,045	-0,159
R09	0,492	-0,013	-0,253
R08	0,478	0,280	-0,103
R19	0,477	0,119	-0,026
L25	0,465	0,087	0,133
L01	0,458	-0,042	0,005
L07	0,457	-0,050	-0,018
R13	0,449	0,229	-0,167
R04	0,447	-0,142	-0,166
L12	0,445	0,105	0,129
L13	0,441	0,248	0,133
S12	0,429	-0,057	-0,167
L16	0,418	0,046	0,074
R07	0,417	0,209	-0,127
R22	0,415	0,231	0,078
R16	0,399	0,325	-0,149
L23	0,387	0,118	0,027
R05	0,373	0,243	-0,108
R21	0,368	0,120	-0,023
R06	0,351	0,152	-0,092
S30	0,349	0,212	-0,194
S14	0,348	-0,087	-0,253
R12	0,340	0,302	-0,181
R02	0,339	0,208	-0,017
L08	0,332	0,118	0,159
L04	0,320	0,131	0,058
R23	0,310	0,193	0,030
R17	0,291	0,273	-0,201
L11	0,289	0,129	-0,083
R01	0,245	0,086	-0,012
L06	0,245	0,223	0,072
L20	0,238	0,210	0,143
R25	0,222	0,211	-0,015
R20	0,086	0,047	-0,028

2. component dominerer			
Item	Component		
	1	2	3
S29	0,108	0,529	-0,188
S27	0,145	0,496	-0,031
S26	0,242	0,450	-0,236
S13	0,044	0,430	-0,058
S28	0,319	0,414	-0,056
R11	0,315	0,414	-0,144
R18	0,022	0,405	0,075
S25	-0,065	0,391	-0,126
R14	0,308	0,390	-0,084
L05	0,227	0,368	0,089
S04	0,062	0,367	-0,184
S21	0,274	0,361	-0,091
S23	0,033	0,342	-0,294
R15	0,317	0,326	-0,129
S05	-0,081	0,314	-0,240
R10	0,267	0,292	-0,067
S24	0,159	0,284	-0,087
S22	0,015	0,269	-0,003
L21	-0,082	0,231	0,019
L10	0,041	0,186	0,012

3. component dominerer			
Item	Component		
	1	2	3
S08	0,200	0,010	-0,609
S09	0,325	-0,103	-0,537
S03	0,041	0,172	-0,527
S06	-0,084	0,183	-0,525
S07	0,197	-0,034	-0,480
S02	-0,047	0,185	-0,451
S01	0,134	0,002	-0,374
S10	-0,004	-0,136	0,139

Farveforklaringer:

- Grøn markering: Factor-load over 0,30 (normal tommelfingerregel om sikkert niveau for loading)
- Gul – 25 Lytteitems (L1-L25), Rød – 25 Læseitems (R1-R25), Blå – 30 items i Sprog & sprogbrug (S1-S30).

Vi har således forsøgt statistisk at afprøve en teori om at Lytte, Læsning og Sprog & Sprogbrug i engelsk udgør hver sit konstrukt ved at undersøge, om deres items loader på hver deres komponent i faktoranalysen. Det viser sig, at der ikke er grundlag for en sådan teori. Dels er det stærkt

tvivlsomt, om der er mere end én komponent (faktor), og dels viser det sig, at items fra alle delprøver loader på denne dominerende komponent. Opgaverne fra hver sin delprøve ser altså ud til (ud fra et psykometrisk perspektiv) i højere grad at måle samme form for dygtighed (der måske kunne betegnes som 'generel forståelse af engelsk') end at udgøre hvert sit konstrukt.

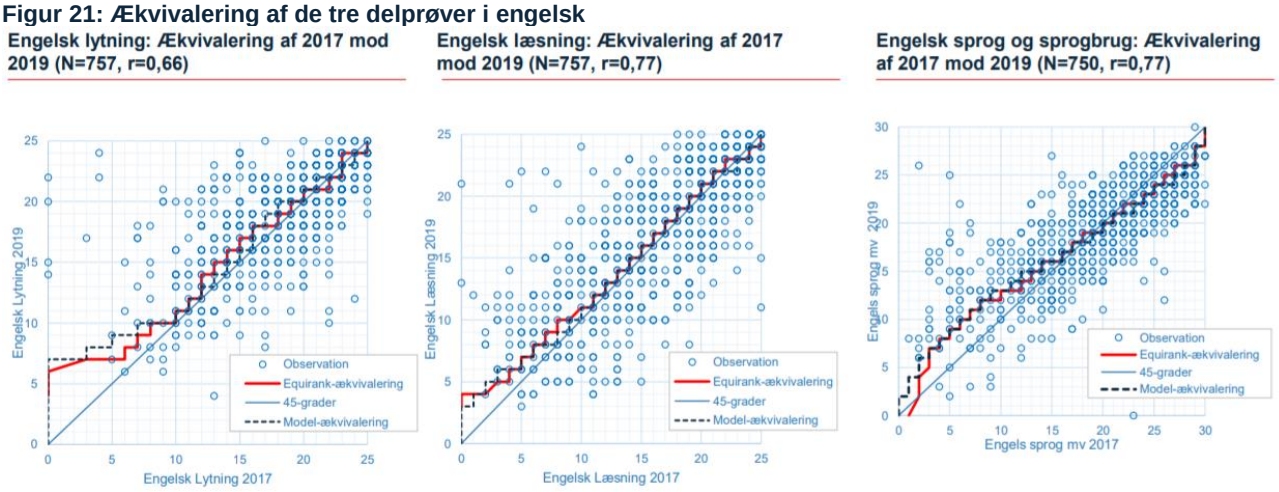
Det betyder også, at der ikke findes noget klart måleteknisk argument for at lade opgaverne fra de tre skalaer veje noget forskelligt i det samlede resultat, og at det derfor vil være logisk at arbejde med en enklere løsning, hvor alle items indgår på lige vilkår i analysen af resultatet.

Det næste spørgsmål, der skal belyses, er, hvordan det påvirker den samlede prøves reliabilitet, hvis man sammenlægger engelskprøven til én skala i stedet for tre separate.

6.4.2 Reliabilitet i bedømmelsen af engelsk i 2019

Ved forcensuren i 2019 blev gennemført en beregning af hvilke gevinster, der kunne være i at betragte engelskprøverne i Lytning, Læsning og Sprog & Sprogbrug som en samlet prøve i stedet for at beregne tre separate karakterer. Ækvivalering af de tre delprøver hver for sig fremgår af Figur 21.

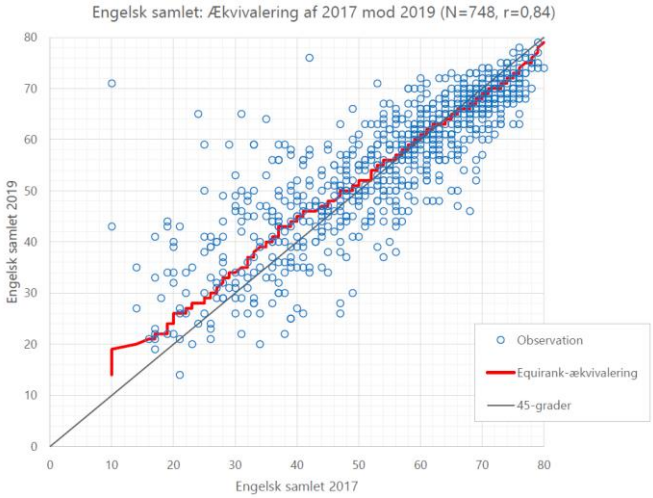
Figur 21: Ækvivalering af de tre delprøver i engelsk



Kilde: Bilag 9.2- Rapporten Evaluering af forcensur 2017-2019

I forbindelse med præsentationen til opgavekommissionen blev det også vist, hvordan ækvivaleringen ville se ud, hvis de tre prøver på hhv. 25, 25 og 30 items blev behandlet som en samlet prøve 80 items.

Figur 22: Engelsk samlet

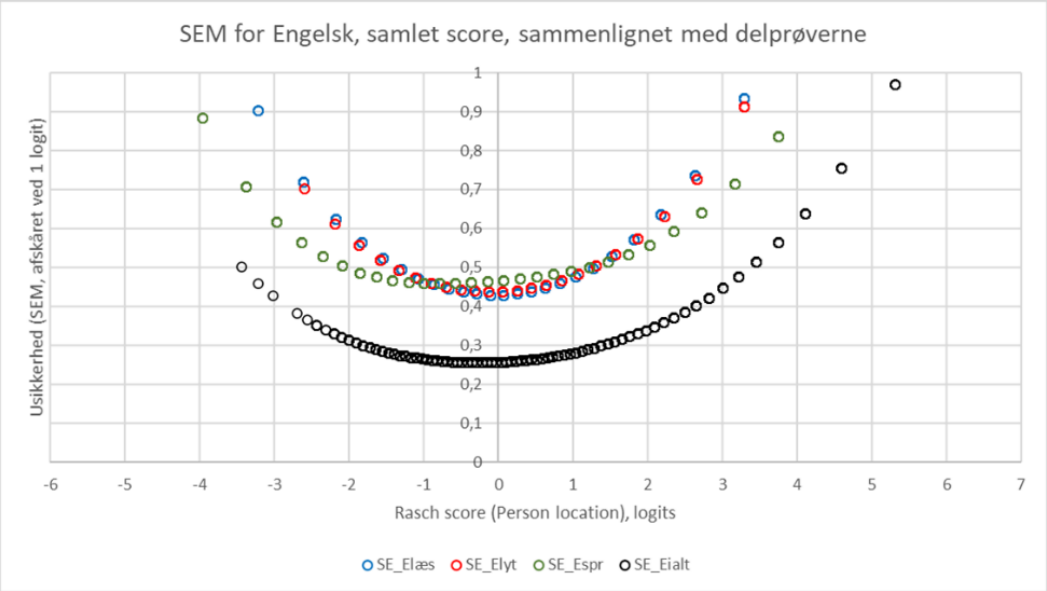


Resultatet, som fremgår af Figur 22, viser, at sammenhængen mellem 2017 og 2019-prøverne bliver mere tydelig ved at benytte alle items, idet korrelationen mellem de to prøver stiger mærkbart.

Reliabiliteten på en test eller prøve er omvendt korreleret med usikkerheden. Det vil sige, at når standardfejlen (SEM eller den statistiske usikkerhed) reduceres, øges prøveresultatets pålidelighed. I Figur 23 er vist sammenhængen mellem elevresultater (målt i Raschscorer) samt usikkerheden for de tre delprøver og for den beregnede samlede skala.

Alle fire skalaer har lavest usikkerhed (optimal target) lige omkring eller lidt under 0-punktet på raschskalaen. Dette svarer for alle fire skalaer til, at omkring halvdelen af opgaverne er korrekt besvaret, og ser man på omsætningstabellerne, giver det for alle fire skaler karakteren 4. Ude i enderne af skalaen er usikkerheden væsentligt større.

Figur 23: Statistisk usikkerhed (N=11.709) for de tre delprøver og samlet



Høj international standard for usikkerhed på en prøves resultat er, at SEM maksimalt bør ligge på 0,3. Kun den samlede skala kan honorere et sådant krav og kun for en del af skalaen. SEM er under 0,3, når Raschscoren er mellem -1,8 og 1,4. Omsat til antal rigtige svar bliver det til mellem 16 og 61 korrekt besvarede items. I Tabel 20 er gengivet den beregnede omsætningstabel for den samlede prøve, og det fremgår, at den samlede prøve fungerer statistisk på meget højt niveau for præstationer mellem 16 og 61 point, der giver karakterer mellem 0 og 10.

Tabel 20: Beregnet omsætningstabel for Samlet engelsk

Samlet omsætning 2019	Karakterer	Antal	Procent
72-80 point	12	873	7%
61-71 point	10	3.696	32%
47-60 point	7	3.507	30%
33-46 point	4	2.178	19%
22-32 point	2	1.087	9%
1-21 point	0	368	3%
0 point	-3	0	0%
		11.709	100%

I forbindelse med afrapporteringen af resultaterne af forcensuren til engelsk-opgavekommissionen blev beregnet en fælles score baseret på metoden med vægtede gennemsnit (de gældende principper). Der blev endvidere foretaget en konsekvensberegning af, hvad det vil betyde at arbejde med en samlet og mere reliabel score/omsætningsmodel, hvor eleverne ikke kan blive straffet/belønnet på grund af afrunding før beregning af samlet karakter. Det viser sig, at det samlede gennemsnit kun ændrer sig helt marginalt (fra 7,03 til 7,08) men at næsten hver 7'ende elev ville have fået en anden karakter, der ud fra statistiske principper ville være mere retvisende for deres præstation: Den alternative omsætningstabel, der arbejder med en samlet score, ville føre med sig, at 1.646 elever (eller 14 % af samtlige elever, der gennemførte i engelskprøven), ville have fået en anden karakter, end den de fik: 937 elever ville have fået en højere karakter og 709 en lavere karakter.

6.4.3 Konklusion på analyserne af engelsk 2019-besvarelserne

Det munder ud i en række delkonklusioner angående engelskprøven:

- Det medfører utilsigtede forskelle (uretfærdigheder) i karaktergivningen, når der gives separate karakterer for Lytning, Læsning og Sprog & Sprogforståelse, som herefter sammenvejes til én karakter.
- Der er er gode *indholdsfaglige argumenter* for tre separate delprøver, men der kan ikke findes *psykometriske argumenter* for at opdele scoringen af de tre prøver, tværtimod.
- Der vil uden problemer – med det samme datamateriale – kunne udarbejdes vejledende omsætningstabeller til hver delprøve (fx til pædagogisk anvendelse ved terminsprøver osv.).
- Der vil ikke være noget problem i at implementere en sådan omlægning uden at ændre på prøvestruktur/format.

Besluttes det at bedømme på grundlag af én større skala (i stedet for tre mindre, som hidtil) opnås en substantielt højere reliabilitet af karaktergivningen, at beregningen bliver mere enkel og gennemskuelig, samt at der bliver bedre overensstemmelse mellem karakter og samlet præstation (retfærdighed). Ændringen er en beregningsteknisk forbedring af den individuelle elevkarakter, den vil ikke medføre generelle niveauspring og kan ske helt uden ændringer i prøveformatet. Det vurderes endvidere stærkt sandsynligt, at tilsvarende resultater kan findes ved at gennemføre tilsvarende analyser på data fra de to andre sprogfag.

6.5 Grundlag for forcensur i sprogfagene

Ved iværksættelse af forcensur/øveprøverne i foråret 2021 blev det besluttet at benytte anledningen til at få indsamlet et datagrundlag, så det kan danne grundlag for en fremtidig forcensur også i tysk og fransk på et så kvalificeret og retfærdigt grundlag som muligt.

Som nævnt er der etableret en praksis for faglig stærk forcensur i engelsk med afsæt i opgavesættet fra Maj-2017. Dette har medvirket til at lægge grunden for et stabilt karakterniveau uden tilfældige udsving fra år til år. Der er erfaringsmæssigt en solid sammenhæng mellem elevpræstationer i sprogfag, idet de elever, der klarer sig godt i ét sprogfag, rent statistisk også præsterer godt i andre sprog. Dette giver et umiddelbart godt grundlag for at tale om et fælles udgangsniveau for de tre skriftlige udtrækprøver i engelsk, tysk og fransk.

Konkret har det været overvejelsen at give mulighed for at fortsætte med maj-prøven 2017 som forcensursæt i engelsk og samtidigt lade maj-2021-prøverne i tysk og fransk være forcensursæt fremover med en beregnet omsætningstabel, der i udgangsåret (maj-2021-prøven) tilstræber, at alle tre fag retter ind efter karakterniveauet i engelsk.

For at dette skal kunne lade sig gøre optimalt, kræves dels at opgavesættene i tysk og fransk har et psykometrisk og fagligt højt niveau, dels at der i forbindelse med forcensur/øveprøverne forekommer et større antal elever, der både tager prøven i engelsk (maj 2017-prøven) og enten tysk eller fransk, og dels at der er en solid korrelation mellem scorerne i de tre sprogfag.

6.5.1 Datagrundlaget for beregning af omsætningstabeller

I alt 3.122 elever deltog i forårets forcensur- og øveprøver i de tre sprogfag med et resultat, der kunne anvendes ved fastlæggelse omsætningstabeller. Antallet af prøvebesvarelser fremgår af Tabel 21.

Tabel 21: Deltagende elever i forcensur- og øveprøver, marts-april 2021

Fag og opgavesæt:	Engelsk, maj 2017	Tysk, maj 2021	Fransk, maj 2021
1. Samlet antal besvarelser, brutto	2.105	1.634	400
2. Anvendelige besvarelser (Modelmetoden)	2.096	1.633	400
3. Anvendelige besvarelser (Equirank)	1.885	1.532	392
4. Elever med engelsk og andet sprog	-	812	112
5. Karaktergennemsnit i engelsk, maj 2017-prøven	7,33	7,11	9,14

Som det fremgår af *2. række* i Tabel 21 vil langt de fleste besvarelser kunne anvendes ved beregning efter modelmetoden, hvor også delvise besvarelser kan indgå (fx manglende besvarelser af hele delskalaer eller delvise besvarelser af delskalaer, således at der kan gives et reliabelt billede af elevens samlede dygtighed). Til brug for ækvivalering er der særlig interesse for de elevers besvarelse, hvor der foreligger svar fra begge de to skalaer, der skal sammenholdes. Fordelingen på fagkombinationer fremgår af Tabel 22 (diagonalen repræsenterer kun ét fag, idet ingen elever tog prøve i alle tre fag).

Tabel 22: Kombinationer af afholdte prøve med anvendelige besvarelser (modelmetoden) i forcensuren 2021

	Engelsk	Tysk	Fransk	*Antal prøver, I alt*
Engelsk	1.132	850	114	*2.096*
Tysk	850	740	43	*1.633*
Fransk	114	43	243	*400*
Antal prøver, I alt	*2.096*	*1.633*	*400*	*4.129*

En del af disse elever manglede som sagt at gennemføre større dele af prøverne (fx hele delprøver), eller også var besvarelsen så mangelfuldt/sporadisk, at data ikke blev vurderet egnede til anvendelse til forcensur med Equirank-ækvivalering, der stiller større krav til et komplet datamateriale (jf. afsnit 4 om Metoder til ækvivalering). Antallet af egnede prøvebesvarelser til equirank-ækvivalering (frasorteret de delvise besvarelser) fremgår af *3. række i Tabel 21*.

Tabel 23: Korrelation mellem samlede scorer for elever, der gennemførte forcensur-/øve-prøver i to fag

Korrelationsmatrice, råscore	Engelsk	Tysk	Fransk
Engelsk, maj 2017 (80 items)	1,00	0,64	0,69
Tysk, maj 2021 (62 items)	0,64	1,00	0,55
Fransk, maj 2021 (68 items)	0,69	0,55	1,00

Som tidligere nævnt bør der endvidere være en solid korrelation mellem scorerne i de enkelte prøver. Som det fremgår af Tabel 23, ligger korrelationen mellem de samlede scorer i de tre fag i intervallet 0,65-0,72, hvilket vurderes som tilfredsstillende og tilstrækkeligt til formålet.

Af *4. række i Tabel 21* ses antallet af elever, der både har besvaret engelskprøven og en af de to andre prøver. Som det fremgår, er der tilsyneladende et godt grundlag for at foretage en ækvivalering af engelsk og tysk, men et spinkelt grundlag for at ækvivalere engelsk og fransk.

En forudsætning for en stærk Equirank-ækvivalering er, at der, ud over at være et større antal komplette besvarelser, også er repræsenteret en solid faglig spredning med en rimelig datamængde i begge ender af skalaen (både stærke og svage præstationer). Af *5. række i Tabel 21* fremgår karaktergennemsnittet for alle valide prøvebesvarelser i engelskprøven for eleverne i række 4. Dette skal udgøre et mål for elevgrundlaget for ækvivalering. Det samlede engelsk gennemsnit ligger på 7,3, hvilket er på niveau med landsresultatet for maj 2017-prøven (jf. Tabel 15). Eleverne, der har besvaret både engelsk og tysk, ligger lidt under dette niveau, mens eleverne, der har besvaret både fransk og engelsk, ligger på et gennemsnitligt meget højt niveau i engelskfærdigheder.

Der vurderes på det grundlag at kunne gennemføres en ækvivalering på delprøve-niveau af tysk og engelsk ud fra disse besvarelser, men det er et spørgsmål, om der er tilstrækkeligt datamateriale til at foretage equirank-ækvivaleringer af delskalaer for så vidt angår fransk. Ikke blot er datagrundlaget begrænset, men eleverne, der har taget prøverne i både engelsk og fransk, er yderligere sprogligt så højt præsterende, at der kommer til at mangle data i bunden af skalaerne.

Problemet bliver yderligere forstærket, hvis der regnes på delprøveniveau, hvor der til hver delprøve kun er et mindre antal opgaver. Det vil dog hjælpe, hvis der i stedet beregnes en samlet omsætningstabel, da det enkelte prøveresultat bliver betydeligt mere reliabelt med et større antal opgaver, end hvis der regnes på delprøveniveau. Imidlertid findes der ikke en formel samlet omsætningstabel for engelsk, så en sådan må først konstrueres.

6.5.2 En samlet omsætningstabel for engelsk, maj 2017

Når en samlet omsætningstabel for engelsk skal beregnes, er det et væsentligt hensyn, at den skaber størst mulig samlet sammenlignelighed bagud.

Tabel 24: Omsætningstabellerne for delprøver i Skriftlig engelsk, maj 2017-prøven (fra prøvebanken.dk)

Engelsk, Lytning Omsætningstabel 2017		
Karakter	Pointinterval	
-3	0	0
0	1	5
2	6	8
4	9	14
7	15	20
10	21	23
12	24	25

Engelsk, Læsning Omsætningstabel 2017		
Karakter	Pointinterval	
-3	0	0
0	1	4
2	5	8
4	9	13
7	14	19
10	20	23
12	24	25

Engelsk, Sprog & Sprogbrug Omsætningstabel 2017		
Karakter	Pointinterval	
-3	0	0
0	1	7
2	8	12
4	13	16
7	17	22
10	23	27
12	28	30

Med udgangspunkt i de 1.885 tilnærmelsesvist komplette besvarelser på engelskprøverne beregnes for hver elev en karakter efter følgende principper: Først slås scorerne op i omsætningstabellerne vist i Tabel 24, hvorved der findes en karakter for hver delprøve (K-Lyt, K-Læs og K-S&S).

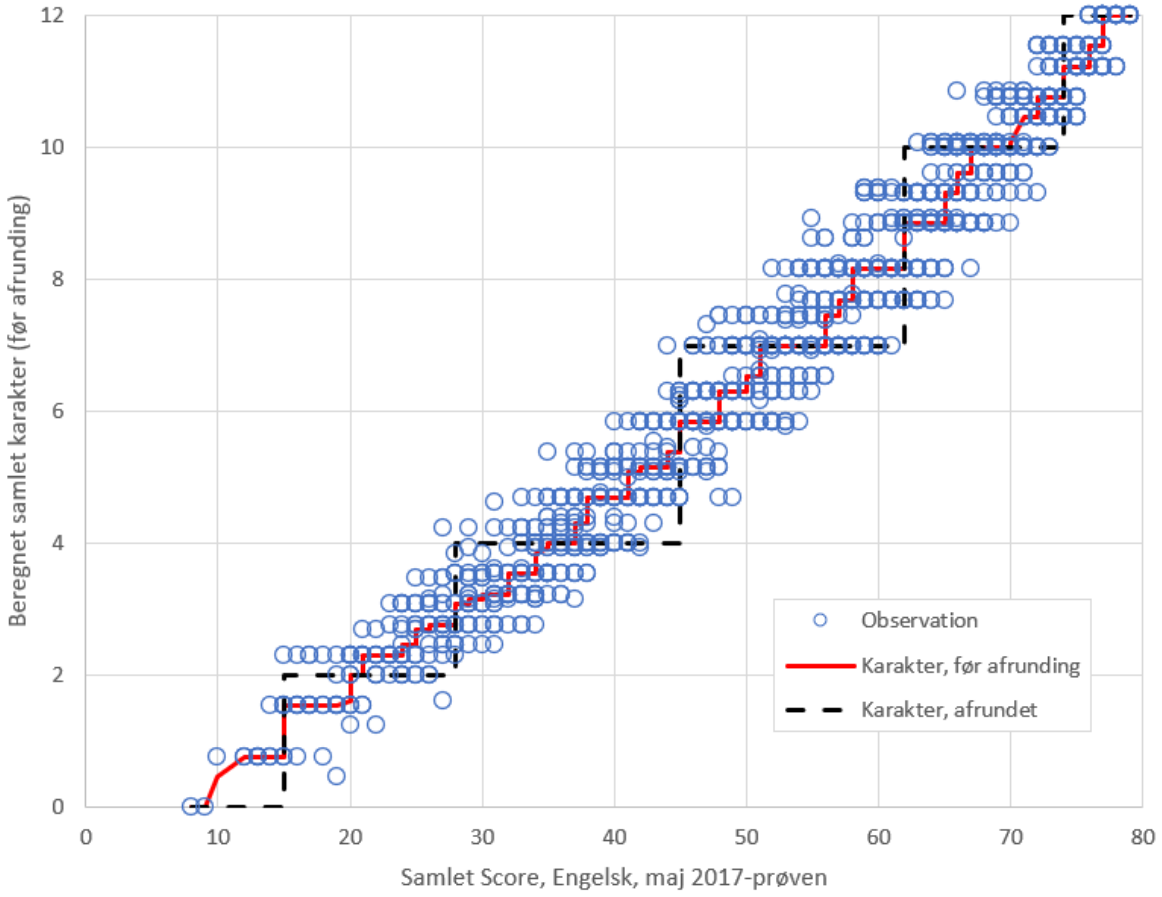
Dernæst sammenvejes de med de vægte, der gælder for karaktergivningen (25%, 25% og 15%), og det samlede resultat tæller 65% af den samlede karakter. Således findes den samlede karakter (ikke afrundet) for de tre delprøver (K-engelsk) på følgende måde:

K-engelsk = (K-Lyt*0,25 + K-Læs*0,25 + K-S&S*0,15) / 0,65

I Figur 24 er som observationer (blå prikker) vist sammenhængen mellem den samlede score og den beregnede samlede karakter før afrunding.

Det fremgår, at den samlede score kan resultere i et spænd på mere end 2 karakterer, og der må derfor benyttes ækvivalering (Equirank) for at identificere den mest sandsynlige sammenhæng mellem score og karakter (den røde kurve).

Figur 24: Sammenhæng mellem samlet score og samlet karakter i engelskprøven, Maj 2017 (N=1885)



Når der herefter afrundes til nærmeste karakter på 12-skalaen (den sorte stiplede linje), har man grundlaget for en omsætningstabel.

Spørgsmålet er, om denne afrunding tilfører nogen systematisk bias. Det kan undersøges ved at se på middelværdien og spredningen i de 2 fordelinger, hvilket fremgår af Tabel 25.

Tabel 25: Middelværdi og spredning i de to karakterberegninger, engelsk maj 2017-prøven

	Karakter, før afrunding (rød) = Observation (Blå)	Karakter, afrundet (sort)
Middelværdi	7,326	7,328
Standardafvigelse	2,830	2,951

Som det fremgår, er spredningen i den afrundede karakter en anelse større, men middelværdien er den samme. Vi kan altså tillade os at konkludere, at den afrundede karakter i Figur 24 er den bedst mulige beskrivelse af sammenhængen mellem den samlede score og samlede karakter for engelsk maj 2017-prøven.

Den resulterende omsætningstabel er gengivet i Tabel 26, og den vil danne grundlag for det videre arbejde med forslag til at tilvejebringe en bedre sammenhæng mellem de skriftlige sprogprøver i udtræksfagene.

Tabel 26: Samlet omsætningstabel for den selvrettende del af prøven (ekskl. Essayprøven) i engelsk maj 2017

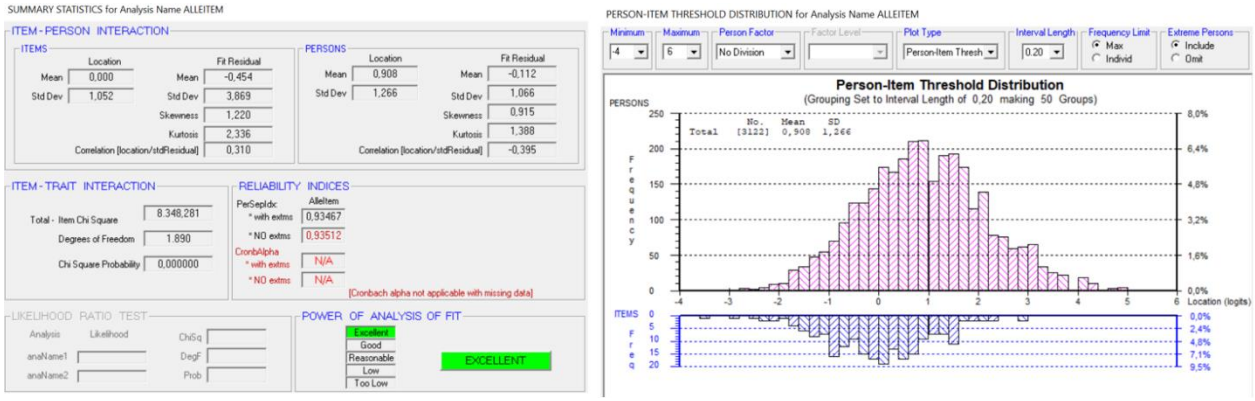
Karakter	Pointinterval	
-3	0	0
0	1	15
2	16	27
4	28	44
7	45	61
10	62	73
12	74	80

6.5.1 En fælles sproglig dygtighed

Som nævnt ovenfor, er der intuitivt et godt grundlag for at tale om et fælles niveau for de tre skriftlige udtræksprøver i engelsk, tysk og fransk. Antagelsen understøttes af en betydelig korrelation mellem scorer i sprogfagene, jf. Tabel 23. Vi skal afprøve, om en sådan dygtighed også genfindes i datamaterialet, og hvad vi i givet fald kan sige om de tre prøvers relative sværhed.

Indledningsvis foretager vi en Raschanalyse med alle 210 items fra de tre prøver for at undersøge, om der er tegn på, at opgaverne fra de tre prøver tilsammen ikke kan beskrives som tilhørende samme skala.

Figur 25: Analyse 1 - Raschanalyse af alle 210 items fra de tre prøver (N=3122)

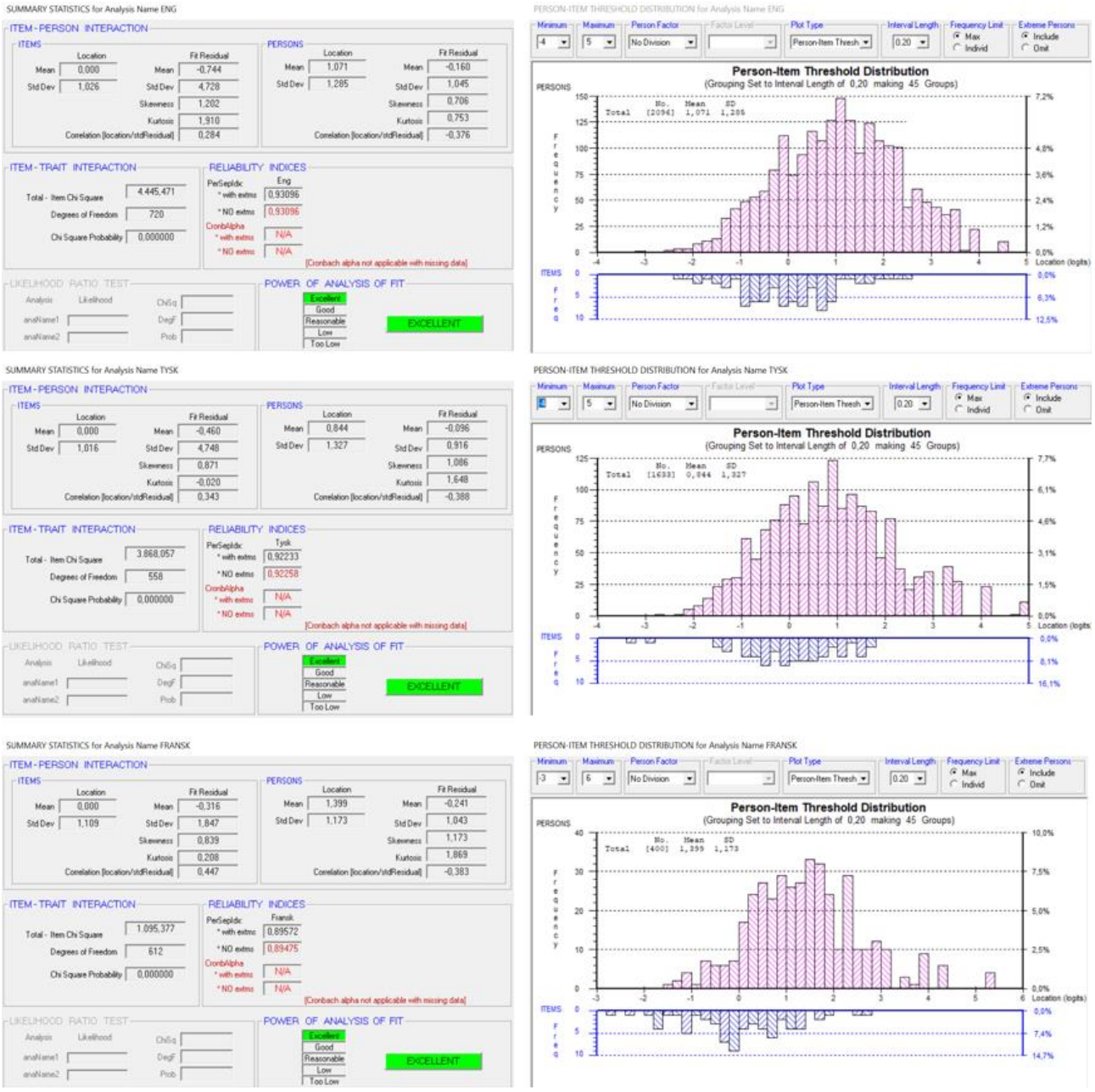


Det ses af Figur 25, at der konstateres en PSI (Person Separations Index[7]), på 0,935, som er et ganske højt niveau, og endvidere at der er en rimelig tilpasning mellem opgavernes sværhedsfordeling og elevernes dygtighed (de hhv. blå- og rødskraverede fordelinger til højre i figuren).

Til sammenligning er der i Figur 26 vist tilsvarende resultater fra Raschanalyser af engelsk, tysk og franskprøverne hver for sig. Som det ses, er PSI i alle tilfælde på enkeltprøve-niveau lavere end på de tre prøver tilsammen i Figur 25, hvilket er et tegn på, at prøverne kan fungere sammen.

[7] PSI er et mål for reliabilitet, der kan antage værdier mellem 0 og 1 – på linje med Cronbach Alpha, der ikke her kan beregnes pga. manglede værdier i data.

Figur 26: Analyse 2-4 - Raschanalyse af engelsk, tysk og franskprøverne hver for sig



Af Tabel 27 fremgår oversigt over resultater af analyserne af de tre prøver hver især og samlet.

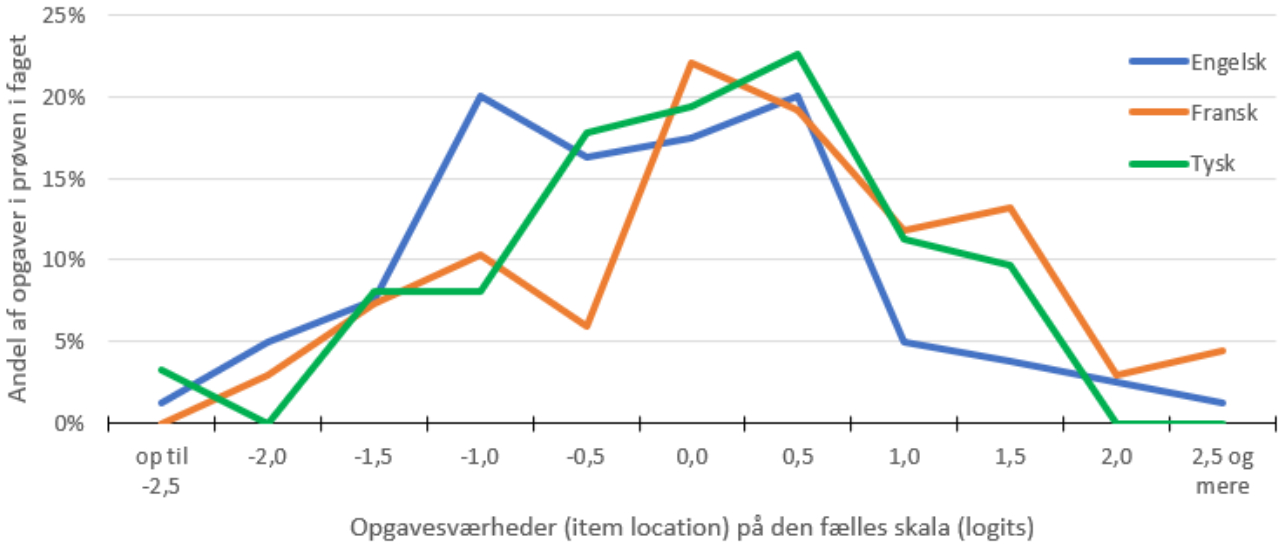
Tabel 27: Oversigt over antal elever, relative opgavesværheder og elevdygtighed i de 4 analyser

Analyse nr.:	1-4	1 (Samlet)	2-4 (engelsk, tysk og fransk)	
	Antal elever	Gennemsnitlig relativ sværhed (item location)	Person Location (Elevdygtighed)	
			Middelværdi (Mean)	Spredning (SD)
Alle tre prøver (210 item)	3.122	0 (Reference)	0,916	1,262
Engelsk (80 item)	2.096	-0,233	1,081	1,278
Tysk (62 item)	1.633	-0,006	0,851	1,326
Fransk (68 items)	400	0,280	1,396	1,172

Når opgaverne fra de forskellige prøver indgår i samme analyse (som i Figur 25), giver det mulighed for at sammenligne opgavesværheder. Det ses af Tabel 27, at engelskopgaverne gennemsnitligt er de letteste (0,23 logit under gennemsnittet), at tyskopgaverne udgør gennemsnittet, og at franskopgaveren gennemsnitligt er de sværeste (0,28 logit over gennemsnittet).

Den relative sværhedsfordeling fremgår af Figur 27, og selvom der gives et mere nuanceret billede af sværhedsfordeling, fremgår det, at opgaverne i engelsk er lettest, og at fransk er sværest.

Figur 27: Opgavernes relative sværhedsfordeling i fagene (målt på samme skala)



Endvidere fremgår det af Tabel 27 (i højre side), at elevernes dygtighed til at løse de opgaver, de er sat overfor, varierer mellem fagene: Det der skiller sig ud er, at eleverne, der har besvaret franskprøven – altså den med de sværeste opgaver – er dem med det klart bedste resultat.

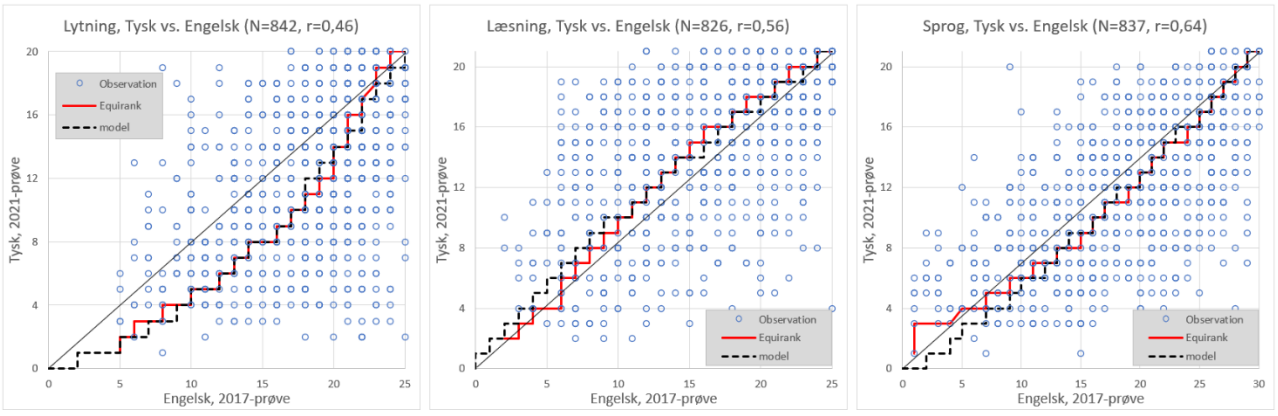
6.6 Ækvivalering af tysk og fransk på grundlag af engelsk

For begge fag gennemgås ækvivalering med både equirank- og modelmetoden (variant B, jf afsnit 4.2) såvel på delskala som samlet for prøven.

6.6.1 Ækvivalering af tysk

Af Figur 28 fremgår ækvivaleringer med både equirank- og modelmetoden (variant B, jf afsnit 4.2) af engelsk og tysk på delskalaniveau.

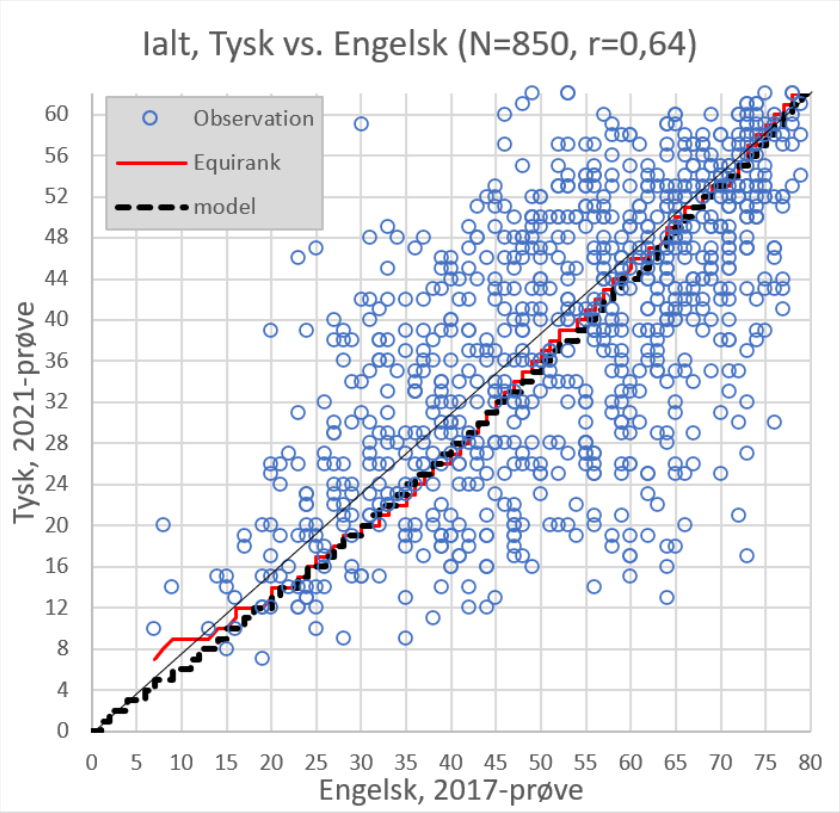
Figur 28: Ækvivalering af delskalaer i tysk og engelsk med equirank-metoden



Der fremgår, at der er et rimeligt datagrundlag, men også nogle forskellige strukturer og styrker i sammenhængene mellem delprøverne i de to prøver (især hvad angår lytning er korrelationen begrænset). Det er tydeligt, hvad angår delprøverne i lytte og sprog & sprogbrug, at prøven i tysk har forekommet sværest (ækvivaleringskurven ligger under identitetslinjen), mens prøven i læsning synes at have været lidt lettere på tysk end på engelsk.

Ser man på den samlede ækvivalering, som fremgår af Figur 29, ser man (i venstre side), at der er en rimelig fordeling af besvarelserne med data i begge ender af fordelingen - dog mangler der data i den helt svage ende af prøverne).

Figur 29: Ækvivalering af samlet prøve med begge ækvivaleringsmetoder (højre side alle tre fag).

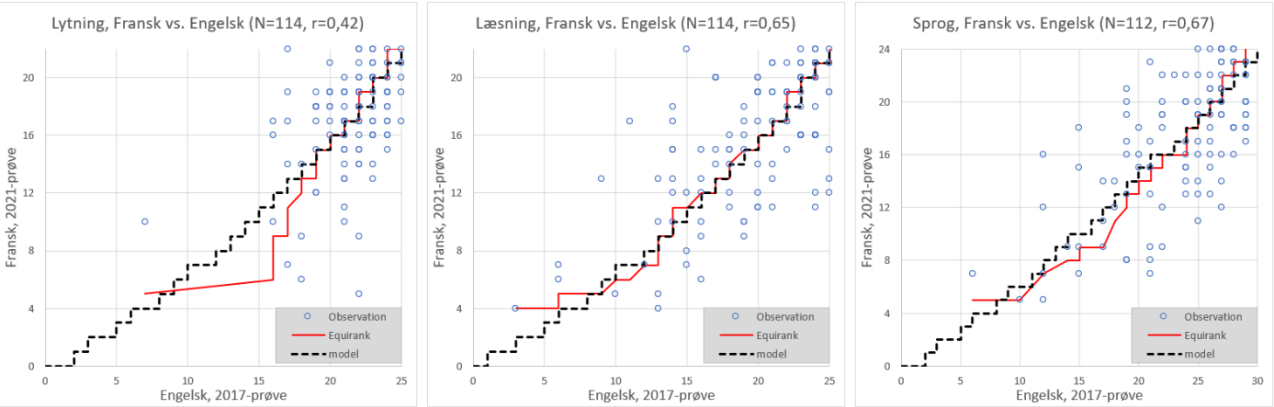


Der forekommer elever, der klarer sig væsentligt bedre i den ene prøve end i den anden, men der er en rimelig sammenhæng (korrelation på 0,64) i datamaterialet. Det fremgår at der er en tæt ved fuldstændig overensstemmelse mellem resultaterne af ækvivaleringer efter de to metoder.

6.6.2 Ækvivalering af engelsk og fransk

Af Figur 30 fremgår ækvivaleringer med afsæt i equirank-metoden på delskalaniveau af engelsk og fransk.

Figur 30: Ækvivalering af delskalaer i fransk og engelsk med både model- og equirank-metoden

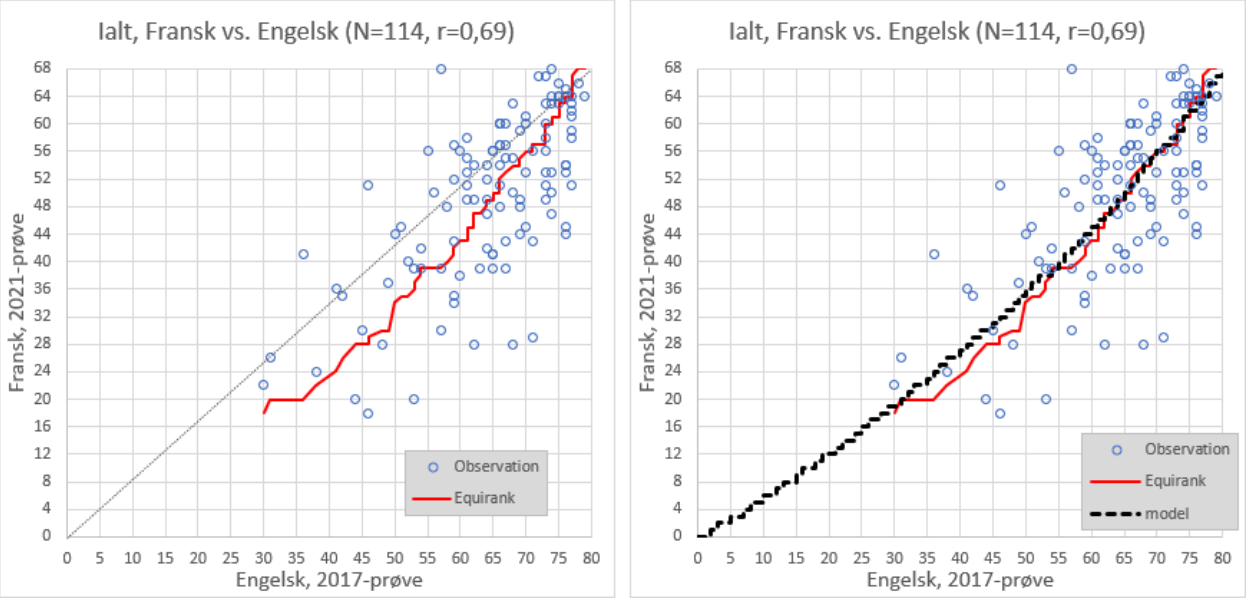


Det fremgår tydeligt, at det meget begrænsede datagrundlag giver problemer for en equirank-ækvivalering. Især i forhold til lytning, hvor der helt mangler elever i den lave ende af skalaen. Grundlaget forekommer bedre for læsning og sprog & sprogbrug og der er i det begrænsede materiale en

rimelig pæn sammenhæng (korrelation på 0,65 og 0,67). Det er endvidere tydeligt, at alle dele af prøven i fransk (især i lytning) har været sværere for eleverne end engelsk (ækvivaleringskurven ligge under identitetslinjen).

Ser man på den samlede ækvivalering, som fremgår af Figur 31, ser man (i venstre side), at der, på grund af det begrænsede elevgrundlag og elevgrundlagets gennemgående gode præstationer i engelsk, helt mangler data i den svage ende. Der er ingen besvarelser med færre end 30 rigtige i engelsk, og da grænsen mellem karaktererne 02 og 4 går ved 28 rigtige (jf. Tabel 26), kan der i denne ækvivalering ikke findes data der kan hjælpe med at finde cutscores under karakteren 4.

Figur 31: Ækvivalering af samlet prøve tysk og engelsk med begge ækvivaleringsmetoder



Det er imidlertid positivt, at det materiale, der trods alt er, viser en pæn korrelation imellem fransk og engelsk-præstationerne uden outliers i bunden. Det betyder også, at equirank-ækvivaleringen i den lave ende nok kan fortolkes som rimeligt gode data. Ser man på højre side, er der også en fin sammenhæng mellem ækvivaleringskurverne fra de to metoder. Og her skal vi huske på, at mens equirank-ækvivaleringen baseres på kun 114 elevbesvarelser, trækker modelmetoden på besvarelse fra 400 franskprøver og over 2.000 elevers engelskprøver.

Det skal understreges, at modelmetodens anvendelighed ved begrænset og/eller skævt fordelt datamateriale er ubestridelig, men også at svaghederne i datamaterialet ikke fuldt ud kan opvejes af en teknisk løsning. Specielt i den lave ende bliver omsætningstabellens grænseværdier meget afhængige af enkeltelevers præstationer, når der blandt de 400 elever, der besvarede franskprøven, kun var tre elever, der fik under 20 rigtige (én fik 16 og to fik 18).

Men det er et produkt af, at der meget få elever, der præsterer i den ende af skalaen – for langt flertallets af elevernes vegne vil karaktergivningen bygge på meget solide datamaterialer.

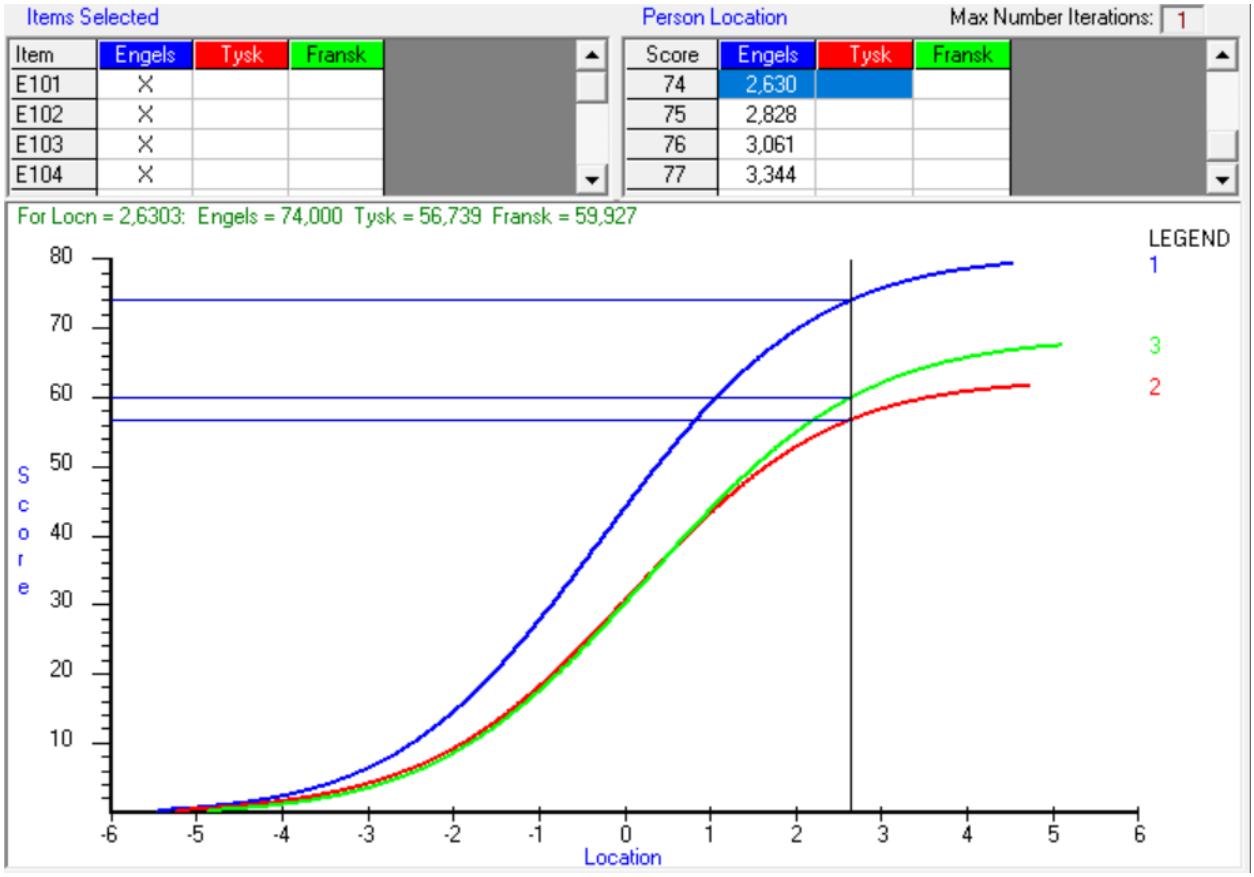
6.6.3 Modelmetoden og omsætningstabeller for tysk og fransk

Af de foregående afsnit kan vi konkludere, at de to ækvivaleringsmetoder viser samme resultat i tysk, og at modelmetoden formentlig er den eneste, der kan levere et resultat i fransk – i hvert fald i den lave ende. Vi skal derfor beregne omsætningstabellerne ud fra modelmetoden.

I Figur 32 er vist den beregnede sammenhæng mellem den fælles dygtighedsskala (location – den vandrette akse) og de dertil svarende scorer i hver af de tre prøver (aflæses på den lodrette akse).

Fremgangsmåden er herefter at slå op i omsætningstabellen for engelsk samlet (se Tabel 20) og finde grænseværdierne (scoren, antal rigtige) for de enkelte karakterer, finde logit-værdien (location) for denne score og herefter undersøge hvilken score denne location svarer til i de to andre fag.

Figur 32: Sammenhængen mellem engelsk, tysk og franskprøverne, som beregnet ved modemetoden



I Figur 32 er vist, hvordan scoren 74 i engelsk (den laveste karakter, der udløser 12 i engelsk) svarer til en location på 2,6303, der igen svarer til scoren 57 i tysk og scoren 60 i fransk, som altså også skal udløse karakteren 12. På den måde beregnes omsætningstabellerne for tysk og fransk samlet. Resultatet er gengivet i Tabel 28.

Tabel 28: Beregnede omsætningstabeller (modelmetoden) for engelsk, tysk og fransk (samlet).

Karakter	Engelsk, maj 2017		Tysk, maj 2021		Fransk, maj 2021	
-3	0	0	0	0	0	0
0	1	15	1	9	1	9
2	16	27	10	17	10	17
4	28	44	18	30	18	30
7	45	61	31	45	31	45
10	62	73	46	56	46	59
12	74	80	57	62	60	68

Som det fremgår, er omsætningstabellerne for tysk og fransk frem til karakteren 10 identiske. Dette er udslag af et rent tilfælde i dette års prøver, som dækker over, at prøven i fransk har flere items som til gengæld gennemgående er sværere. Derfor kan den samme score i fransk og tysk udløse den samme karakter uanset, at der er seks items flere i fransk.

Aktuelt scores prøverne ved omsætningstabeller på hver af de tre delskalaer. Disse skalaer er der også beregnet omsætningstabeller for og af Tabel 29 fremgår de for Tysk samlet og for hver af de tre delskalaer. Tilsvarende fremgår omsætningstabellerne for fransk af Tabel 30, nedenfor. Begge sæt omsætningstabeller er baseret på modelmetoden (variant B, j afsnit 4.2).

Tabel 29: Beregnede omsætningstabeller (modelmetoden) for tysk, maj 2021. Samlet og de tre delskalaer.

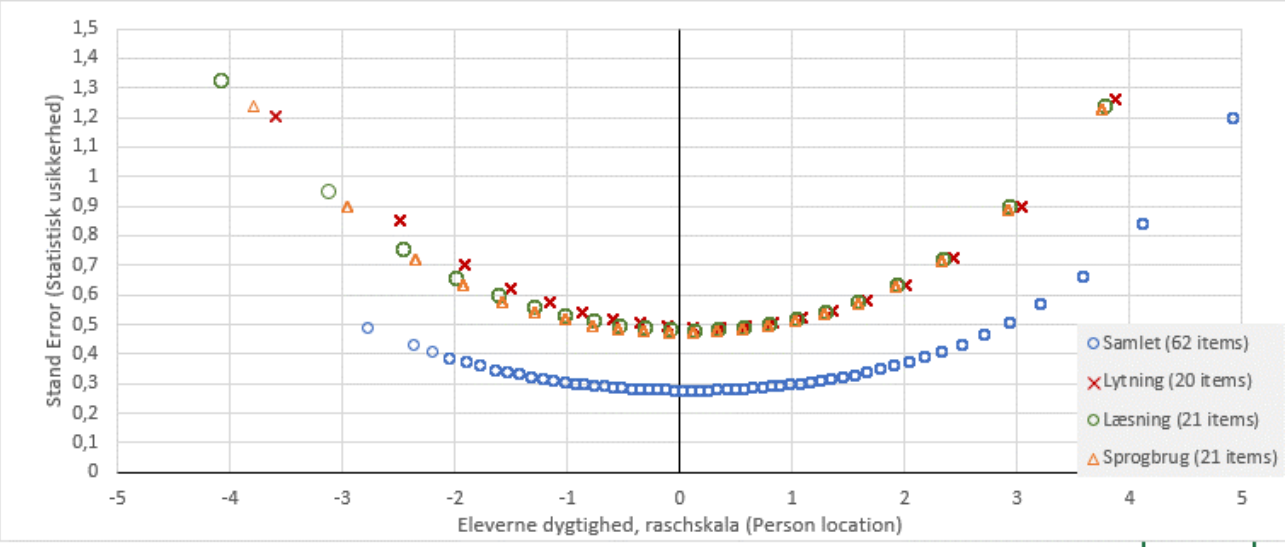
Karakter	Tysk samlet		Tysk, Lytning		Tysk, Læsning		Tysk, sprogbrug	
-3	0	0	0	0	0	0	0	0
0	1	9	1	1	1	4	1	3
2	10	17	2	3	5	8	4	7
4	18	30	4	7	9	12	8	9
7	31	45	8	14	13	14	10	14
10	46	56	15	18	15	19	15	18
12	57	62	19	20	20	21	19	21

Tabel 30: Beregnede omsætningstabeller (modelmetoden) for fransk, maj 2021. Samlet og de tre delskalaer.

Karakter	Fransk samlet		Fransk, Lytning		Fransk, Læsning		Fransk, sprogbrug	
-3	0	0	0	0	0	0	0	0
0	1	9	1	2	1	2	1	4
2	10	17	3	4	3	5	5	7
4	18	30	5	10	6	9	8	11
7	31	45	11	15	10	14	12	16
10	46	59	16	19	15	20	17	21
12	60	68	20	22	21	22	22	24

Udfordringerne ved at operere med 3 omsætningstabeller med få items (i steder for én større) er, som beskrevet tidligere i dette afsnit, at det tilfører en betydelig statistisk usikkerhed og samlet lavere reliabilitet i målingen. Dette er illustreret for tyskprøvens vedkommende (franskprøven ville vise stort set det samme) i Figur 33.

Figur 33: Statistisk usikkerhed (SE) som funktion af elevdygtighed (Tysk, maj 2021. Samlet og de tre delprøver.)



Godt 80 pct. af eleverne ligger indenfor plus/minus 2 logits og her vil Standard Error (SE) på den samlede tysk-score ligge under 0,4 Standard, hvilket er rimeligt godt, mens SE på de tre delskalaer vil ligge fra 0,5 og 0,7, hvilket er i udkanten af et acceptabelt niveau efter international almindelige standard.

6.6.4 Karakterfordelinger

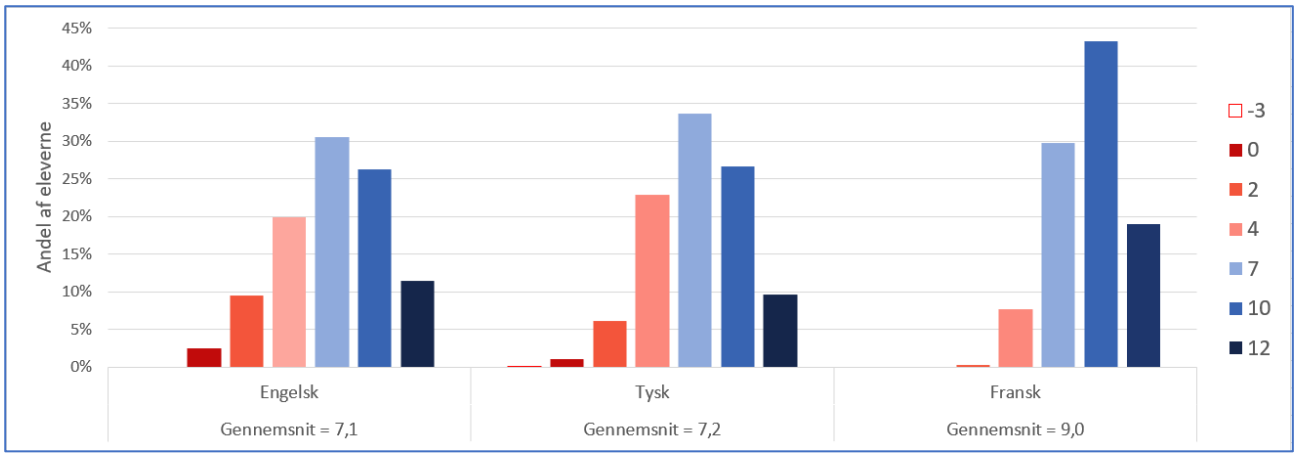
Gennem anvendelse af omsætningstabellerne vist i Tabel 28 kan beregnes karakterfordelingen ved de afholdte forcensur- og øveprøver i april 2021. Disse fordelinger fremgår af Tabel 31.

Tabel 31: Karakterfordeling ved forcensur- og øveprøver i engelsk, tysk og fransk

Karakter	Engelsk		Karakter	Tysk		Karakter	Fransk	
	Antal elever	Fordeling		Antal elever	Fordeling		Antal elever	Fordeling
-3	0	0,0%	-3	1	0,1%	-3	0	0,0%
0	52	2,5%	0	17	1,0%	0	0	0,0%
2	199	9,5%	2	100	6,1%	2	1	0,3%
4	416	19,8%	4	374	22,9%	4	31	7,8%
7	639	30,5%	7	549	33,6%	7	119	29,8%
10	550	26,2%	10	435	26,6%	10	173	43,3%
12	240	11,5%	12	157	9,6%	12	76	19,0%
I alt	**2096**	**100%**	**I alt**	**1633**	**100%**	**I alt**	**400**	**100%**

Af Figur 34 fremgår data vist grafisk, og her er også angivet gennemsnit. Det skal bemærkes, at der i denne opgørelse er medtaget besvarelser, som i andre analyser i denne rapport er frasorteret[8], hvorfor der i engelsk forekommer en del 00 og 02-karakterer, som typisk ikke optræder, når prøven afholdes for alvor og ikke som øveprøve. Det betyder også, at engelsk-gennemsnittet er lidt lavere end fx i Tabel 25, hvor det opgøres til 7,3 baseret på 1.885 besvarelser.

Figur 34: Karakterfordeling og gennemsnit ved forcensur- og øveprøver i engelsk, tysk og fransk



Som det fremgår, bliver tyskkarakteren beregnet til nogenlunde det hidtidige niveau (jf. 2018/19 i Tabel 15) mens karakteren for skriftlig fransk (9,0) bliver løftet med mere end 2½ karakter i forhold til tidligere. Det kan lyde overraskende, men dette niveau svarer ret præcist til, hvad disse elever havde fået i prøven i engelsk (jf. Tabel 21). Det dog ikke sandsynligt, at dette spring hovedsageligt skal tilskrives en tidligere for lav vurdering af præstationerne i franskprøven. Det skal måske i højere grad henføres til, at de elever, der deltog i forcensur- og øveprøver i år i fransk, udgjorde et særligt fagligt stærkt udsnit af eleverne. Sandsynligvis er forklaringen en kombination af disse faktorer. Hvor meget af forskellen, der kan henføres til hvad, kan de kommende års forcensur og karaktergivning måske give et kvalificeret bud på.

[8] Fx er frasorteret ukomplette besvarelser, herunder elevbesvarelser, hvor hele delskalaer mangler (som i afsnit 6.5.2), hvor der er godt 200 besvarelser, der ikke blev vurderet som repræsentative for elevernes dygtighed, når blank tælles som fejl, og som derfor udgik af analysen ved beregning af omsætningstabellen.

7. Forcensur for prøverne i 10 klasse

Forcensur vil på sigt kunne gennemføres for 10. klasse-prøverne i dansk, matematik og sprogfag. Ved prøveafholdelsen i maj 2021 blev gennemført prøver i matematik og i dansk, hvor prøven og datamaterialet vil kunne danne grundlag for fremtidig forcensur, såfremt prøverne lever op til de forventede psykometriske kvalitetsmæssige krav til en prøve. Prøver i de øvrige fag var afløst af årskarakterer i maj 2021.

7.1 Datagrundlag, dansk

Prøven i dansk, læsning og sprogbrug udgjorde en selvrettende prøve, hvor der blev indsamlet et komplet besvarelsesmateriale via ministeriets prøveafviklingssystem. En andel af eleverne, der gik i 10. klasse frem til sommerferien 2021, har endvidere besvaret de tilbudte øveprøver i læsning og retskrivning, der var FP9 opgavesæt fra maj 2017. På det grundlag var det hensigten at beregne et grundlag for at kunne ækvivalere niveauet i prøverne målrettet 9. kl. og 10 kl., der jo giver de samme rettigheder mht. optag på andre uddannelser mv.

7.1.1 Muligheder for at ækvivalere niveau i dansk FP9 og FP10

Det viste sig, at relativt få af eleverne i 10. klasse havde gjort brug af muligheden af at anvende øveprøver, jf. Tabel 32. Disse knap 90 elever ville kunne være afsæt for en nogenlunde sikker sammenlignelighed af niveau i de to prøveformer. Imidlertid viste det sig, at stort set ingen af disse, havde taget 10. klasse prøven (samlet fire elever).

Tabel 32: Forcensur- og øveprøver i læsning og retskrivning, maj 2017-prøver

	Elevens klassetrin		
	9. klasse	10. klasse	I alt
Læsning, FP9, Maj 2017-prøven	11.299	87	11.386
Retskrivning, FP9, Maj 2017-prøven	11.657	89	11.746

Der var altså ikke et datagrundlag for den sammenlignende analyse af de to prøver, som var planlagt.

7.1.2 Dansk, maj 2021 som fremtidigt forcensursæt

For at belyse sammenhængen mellem opgaverne i prøven, og hermed om den måler de samme form for færdigheder, er der foretaget en Raschanalyse. Da intentionen er, at prøven skal afprøve forskelligartede færdigheder, forventes ikke, at den skal leve fuldstændig op til Raschmodellens krav, men dog skal der helst ikke forekomme opgaver i prøven, der medfører, at det samlede prøveresultat får en klart lavere reliabilitet, end hvis de bliver udeladt af analysen (som det var tilfældet ved naturfagsprøverne, jf. afsnit 5.3). Som det fremgår af Tabel 33 blev prøven i dansk FP10 maj 2021 besvaret af 28.604 elever.

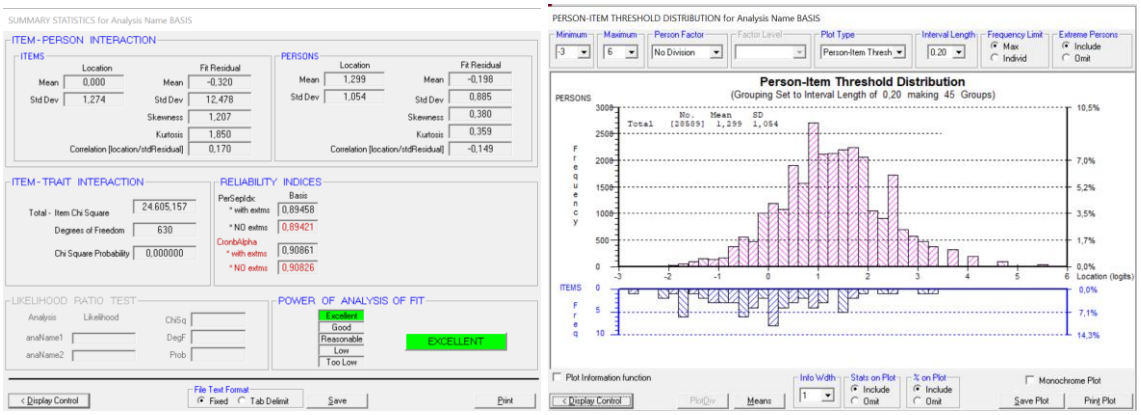
Tabel 33: Antal elevbesvarelser af prøven i dansk FP10 maj 2021

	Antal besvarelser
Antal besvarelser	28.604
Heraf mindst 50% ubesvarede spørgsmål (blanke svar)	15
De resterende (retvisende) svar	28.589

Når man benytter andel rigtige svar som grundlag for analyse, kommer man implicit til at antage, at de ubesvarede spørgsmål (blanke svar) er udtryk for, at eleven ikke kunne svare rigtigt på opgaven. Hvis eleven ikke har gjort et forsøg på at svare rigtigt på opgaverne, bliver antal rigtige svar ikke noget godt mål for elevens dygtighed. For at eliminere de mest misvisende besvarelser fra analysen udelades derfor de besvarelser, hvor færre end halvdelen af de 70 spørgsmål i opgavesættet er besvaret. Dette drejede sig kun om 15 elever på landsplan.

Raschanalysens overordnede resultat af de besvarelser der er vurderet egnede til formålet fremgår af Figur 35.

Figur 35: Raschanalyse af besvarelser af prøven i dansk FP10 maj 2021 (N=28.589)



Som det fremgår af ovenstående, repræsenterer besvarelserne samlet set en solid reliabilitet (Cronback alpha=0,91), og af figurens højre side ses en betydelig spredning og i øvrigt en rimelig tilpasning mellem items sværhed og elevernes dygtighed (prøvens targetting).

En nøjere analyse viser, at der er en række opgaver, hvor besvarelserne passer mindre godt til Raschmodellen. Dem kunne man overveje at revidere eller tage ud af opgavesættet ved fremtidig anvendelse. Men samtidig viser det sig, at såfremt de dårligst fungerende items tages ud af analysen, reduceres den samlede reliabilitet (om end meget beskedent). På den baggrund vurderes det, at opgavesættet i Dansk FP10 maj 2021 vil kunne fungere hensigtsmæssigt i den foreliggende form som grundlag for fremtidig forcensur i skriftlig dansk.

7.2 Datagrundlag, matematik

Prøven i maj 2021, matematik til 10. kl. er en censorrettet prøve med struktur, der svarer til FP9 i matematik med hjælpemidler. I lighed hermed indsamler opgavekommissionen et meget stort materiale fra landets censorer til brug for forcensur. Opgaver af den type, hvor formuleringen af opgaven er mere overordnet, og hvor eleven selv skal foretage en detaljeret formulering, har det med at passe bedre til Raschmodellen, end der hvor opgavekonstruktøren skal formulere svarmuligheder (i MC-format eller lignende).

Det bør dog undersøges, om data passer nogenlunde til de overordnede krav i Raschmodellen. Dette kræver et datagrundlag, og et sådant eksisterer allerede (så vidt vides) og kan tilvejebringes i regnearksformat på opgave- og personniveau (elevoplysninger i anonymiseret form) og således danne grundlag for en analyse af prøvens psykometriske egenskaber. Medmindre en analyse af besvarelserne viser uventede negative træk i opgavesættet, vil det med rimelighed kunne antages, at også opgavesættet i Matematik FP10 maj 2021 vil kunne fungere rimeligt som grundlag for fremtidig forcensur i Matematik til FP10, eksempelvis fra førstkommende prøver. Det bør overvejes at indhøste de første erfaringer med forcensur i dansk og matematik på 10. klasse niveau før yderligere introduceres (fx sprogfag).

8. Anbefalinger

På baggrund af det omfattende arbejde, der er udført med forcensur fra 2017 til 2021, herunder dialog med ministeriet, skoler, opgavekommissioner og andre faglige fora, er der blevet identificeret en række muligheder for forbedringer af organisering af forcensur og prøvepraksis. I forbindelse med den første delrapportering ultimo 2017 blev givet en række anbefalinger (Bilag 9.2.1, afsnit 5 om anbefalinger). Nogle anbefalinger vil kræve ændring af større omlægninger af praksis, rammer, regler og/eller politiske beslutninger vedr. prøverne.

I dette afsnit er fokus på forslag, som drejer sig alene om de prøver og delprøver, der bedømmes ved brug af en omsætningstabel og som umiddelbart vil kunne gennemføres.

Der fokuseres således på de områder/initiativer, hvor der kan opnås forbedringer af prøvekonstruktion, scoringsmodeller og/eller organisering af forcensur/prøveafvikling, uden at løsningerne medfører nævneværdigt omlægninger/ekstraarbejde/ulemper, og hvor initiativerne kan implementeres administrativt indenfor gældende regler og rammer. Der er blandt andet gennemgået og konsekvensvurderet nogle af disse muligheder i denne rapport.

8.1 Anbefalinger mht. de bundne prøver

Der er generelt etableret en solid og gennemprøvet praksis og velfungerende scoringsmodeller, både hvad angår 9. klasseprøverne i dansk (læsning og retskrivning) og matematik (matematik med og uden hjælpemidler). Hvad angår FP9 er det kun ifm. matematik med hjælpemidler, hvor der vurderes at være et potentiale for forbedringer. Forcensur for FP10 efter metoderne anvendt ved FP9 anbefales etableret (jf. afsnit 7).

Anbefaling 1: Systematisk indsamling af rette-data vedr. FP9, matematik med hjælpemidler på elevniveau i BUVM.

I matematik med hjælpemidler er datagrundlaget for forcensuren begrænset (ca. 1.000 elever med den hidtidige organisering), og den dataindsamling (ca. halvdelen af landets elever i regneark), der hidtil er foretaget af opgavekommissionen af censorbedømmelser på opgaveniveau til brug for opgavekommissionens eget ikke-ækvivaleringsbaserede prøve-/forcensurarbejde, vil med fordel kunne gennemføres af ministeriet i samarbejde med opgavekommissionen (fx med bevarelse af at der rettes i regnearksskabeloner men indrapporteres i en central SharePoint-løsning). Herved kan datasikkerhed forbedres, og den anonymisering af eleverne, som den nuværende løsning fordrer (og som forhindrer at disse data kan bruges til ækvivalering på personniveau i forcensuren), kan undgås. Herved kan den omfattende dataindsamling, der har været organiseret af opgavekommissionen, spille bedre sammen med det ækvivaleringsbaserede arbejde med omsætningstabellen. En vellykket løsning vurderes at forudsætte opgavekommissionens fortsatte inddragelse og engagement.

Anbefaling 2: Etablering af forcensur i dansk og matematik ved FP10 med afsæt i maj 2021 sættet

Se begrundelse og dokumentation i afsnit 7, Forcensur for prøverne i 10 klasse.

Anbefaling 3: Systematisk indsamling af rette-data vedr. FP10, matematik på elevniveau i BUVM.

Med henblik på at tilvejebringe et godt grundlag for den fremtidige forcensur i FP10, matematik, foreslås at løsningen fra anbefaling 1 (ovenfor) også implementeres for FP10.

8.2 Anbefalinger vedrørende skriftlige udtræksprøver, FP9

Her vil anbefalingerne især bygge på analyserne i afsnit 5, Bedømmelse af prøverne i naturfag og afsnit 6, Bedømmelse af prøverne i fremmedsprog.

Anbefaling 4: Implementering af nye forcensursæt i naturfagene, jf. afsnit 5.

Ved anvendelse af de forcensur-opgavesæt i naturfag (med tilhørende omsætningstabeller), som er foreslået i afsnit 5, vil gevinsten være en bedømmelsesmæssig ligebehandling af eleverne, uanset hvilket af de tre naturfag der udtrækkes, jf. begrundelser og dokumentation i afsnit 5.

Anbefaling 5: Opdatering af scoringsregler i naturfagsprøver

I naturfag har der hidtil været en særlig regel om pointfradrag indenfor opgaven ved fejlsvar. Denne regel er ikke i overensstemmelse med almindelig god psykometrisk praksis og betyder mindre retfærdige bedømmelser. Den foreslås derfor ændret, så den modsvarer de almindelige regler, der gælder for alle de øvrige prøver i FP9. Kan implementeres uden ændring af prøverne.

Anbefaling 6: Systematisk indsamling af rette-data hvis censorrettede naturfagsprøver indføres

Med henblik på at tilvejebringe et godt grundlag for forcensur i del B af de nye naturfagsprøver foreslås, at løsningen fra anbefaling 1 (med retteskabelon og elektronisk indberetning af rettedata til BUVM for censorer) også implementeres for disse prøvers vedkommende – helst forud for at prøverne sættes i drift.

Anbefaling 7: Implementering af forenklede scoringsregler i skriftlige sprogprøver

Som beskrevet i afsnit 6 medfører de nuværende scoringsregler med delscorer, høj usikkerhed og afrundinger en række uhensigtsmæssigheder. Herunder kan karakterberegningen resultere i tilfældige og ikke fagligt baserede forskelle mellem elever, der har præsteret lige godt i henhold til en mere præcis vurdering af den samme besvarelse. Med dette udgangspunkt foreslås karakteren i både i engelsk, tysk og fransk beregnet på et mere retvisende grundlag, med én samlet score og én omsætningstabel, som beskrevet i afsnit 6. Eventuelt kan der parallelt udarbejdes omsætningstabeller for delscorer til pædagogisk anvendelse. Kan implementeres uden ændring af prøverne.

Anbefaling 8: Prøverne fra maj 2021 som grundlag for forcensur i tysk og fransk

Ved anvendelse af opgavesæt fra maj 2021 (med tilhørende omsætningstabeller, jf. afsnit 6) til forcensur i de to fag, som er foreslået i afsnit 6, vil gevinsten være en bedømmelsesmæssig ligebehandling af eleverne, uanset hvilket af de tre skriftlige sprogfag de har udtrukket, jf. begrundelser og dokumentation i afsnit 6.

8.3 Generelle anbefalinger vedrørende prøver og opgaver

Nogle anbefalinger er et produkt af generelle observationer og ikke knyttet til særlige prøver.

Anbefaling 9: Kombineret øveprøve og forcensurprøve

Ved forcensuren i april 2021 blev forcensurprøven også udbudt som øveprøve. Dette gav et meget stærkt datamateriale med en begrænset ekstraindsats. Det foreslås at anvende den model fremover, uanset at det næppe kan forventes helt så stor deltagelse i et år uden epidemi og fuldt prøveprogram.

Anbefaling 10: Forcensurprøve med feedback

Et incitament til at forcensurskolerne satte de fornødne ressourcer af til arbejdet med en kvalificeret deltagelse i forcensur var, at der blev leveret et detaljeret feedback på individ- og opgaveniveau til den enkelte klasse. Det foreslås at anvende den model fremover, da det er med til at fastholde skolernes interesse og giver et godt grundlag for at have tiltro til, at datagrundlaget er retvisende.

Anbefaling 11: Sammenlignelighed mellem karakterer fra FP9 og FP10

Hvis karaktergennemsnit fra FP9 og FP10 skal give de samme rettigheder med hensyn til fx efterfølgende optag på uddannelser, bør de også kunne sammenlignes uden videre. Det var hensigten at foretage en sådan ækvivalering i år, men der viste sig ikke at være tilstrækkeligt datamateriale til at gøre det muligt (jf. afsnit 7.1). Det foreslås derfor at inkludere i baseline for ækvivaleringen af prøverne i 10. klasse, at de niveaumæssigt er sammenlignelige med prøverne i samme fag i 9. klasse. Da det kun gælder fastlæggelse af niveau, stiller det ikke samme krav til spredning og kan derfor gøres pålideligt ved et begrænset antal besvarelser fra elever, der skal have 10. klasse prøven, og som tager forcensur/øveprøver fra den tilsvarende FP9-prøve.

Anbefaling 12: Kvalitetssikring af svarregler

Der anvendes svarregler, der var forståelige og rimelige, dengang pen- og papirudgaven af prøven var udgangspunktet, og hvor eleverne rent fysisk kunne sætte kryds, hvor de ville. Men der er stadigvæk mulighed for at svare ugyldigt – fx sætte flere krydser hvor der står, at man kun må sætte ét kryds.

Dette kan udløse bias i besvarelsen, hvis fx to svarmuligheder principielt kan være rigtige, og den dygtige elev forsøger at markere dette med to krydser. Når der ikke længere tilbydes mulighed for pen og papirbesvarelse til eleverne, anbefales det at opdatere svarreglerne. Dette kan implementeres uden ændring af prøverne i øvrigt.

Anbefaling 13: Kvalitetssikring af anvendte spørgsmålsskabeloner

I henhold til erfaringer og almindelig god skik i forhold til item- og testkonstruktion er der en række retningslinjer der bør følges[9]. Eksempelvis skal der kun anvendes tekst, information og grafik der er nødvendige for opgaveløsningen, der skal hentes mest mulig information om elevens præstation i forhold til elevens indsats og at det i videst muligt omfang undgås, at der forekommer lokal afhængighed[10] mellem opgaverne indbyrdes. Det betyder også at der er nogle opgaveformater der generelt bør undgås og andre, der kun bør benyttes undtagelsesvist og kun i specifikke sammenhænge. I visse opgaver er det en betingelse, at en bestemt skabelon anvendes. Men som regel kan en opgave stilles på mange forskellige måder med forskellige skabeloner, og her anbefales at anvende de bedst mulige skabeloner/opgaveformater i de enkelte opgaver/prøver for at forbedre prøvernes reliabilitet.

[9] Se fx Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. Applied measurement in education, 15(3), 309-333.

[10] Lokal afhængighed betyder, at besvarelsen af en opgave ikke blot afhænger af elevens dygtighed men også (måske i højere grad) af besvarelsen af andre opgaver. I Raschmodellen kræves lokalt uafhængige opgaver, og selvom der ikke er krav til, at prøverne skal opfylde kriterierne i for Raschmodellen, ligger der implicit en forventning til lokal uafhængighed, når scoringen sker på grundlag af råscoren. Det hænger fx ikke godt sammen, hvis den ene opgave giver to point og en anden tilsvarende opgave kun giver et point, hvilket i realiteten er det, der sker.

Der tilbydes i applikationen (”opgavebygger”) der anvendes ved konstruktion af folkeskolens testlignende prøver, en længere række skabeloner, der ikke alle fungerer optimalt fra et målingsmæssigt perspektiv. Eksempelvis er der opgaver hvor a) selve skabelonen skaber lokal afhængighed, b) ineffektive opgavetyper hvor kun en mindre del af informationerne udnyttes, og c) opgaver med flere spørgsmål, der alligevel kun kan scores rigtigt eller forkert. Her er et eksempel på hver slags:

a) Opgaver der skaber lokal afhængighed. *Eksempel:* Matchopgaver – hvis 4 spørgsmål skal matches med 4 svar, så vil én fejl pr. definition udløse endnu en fejl (der er i realiteten kun 3 delopgaver som giver 1, 2 eller 4 point – hvis de første tre delopgaver er parret korrekt er der ingen mulighed for at svare forkert på den sidste).
   Anbefaling: Hvis formatet anvendes, bør hele opgaven kun scores som 1 point (rigtigt eller forkert). En bedre mulighed er at øge antallet af svarmuligheder fx ved at hver svarmulighed kan genbruges.
b) Ineffektive opgavetyper. *Eksempel:* Opgaver hvor der skal sættes et specifikt antal X’er. Hvis der fx er syv udsagn og opgaven går ud på at finde to udsagn, der er sande. Instruktionen lyder ”Sæt to X’er” – dette giver to items med hver 7 svarmuligheder til det første X og 6 til det andet X (besvarelsen af første delopgave begrænser besvarelsen af næste). Dette giver altså 2 items med 7 x 6 = 42 mulige udfald.
   Anbefaling: Hvis man i stedet beder eleven om at forholde sig de udsagn, der er sande, bliver der i realiteten 7 items ud af opgaven. Den eneste ændring i opgavens formulering er at instruktionen ændres til ”Sæt gerne flere X’er”. På den måde ændres opgaven til 7 items med hver to svarmuligheder (X eller blank). Det giver i alt 7^2 mulige kombinationer svarende til 128 mulige udfald. Det betyder altså godt 3 gange så meget data, uanset om der regnes udfaldsrum (128 i forhold til 42 muligheder) eller antal point (7 i forhold til 2).
c) Polytome opgaver, der kun kan scores dikotomt. *Eksempel:* Rækkefølgeopgave. Fire tekstlinjer skal placeres i den rigtige rækkefølge (1-2-3-4), så det giver mening som en sammenhængende tekst. Her vil den rigtige rækkefølge af 2 eller 3 linjer (fx 2-3-4-1) ikke kunne give point. Enten er hele opgaven korrekt besvaret eller også er den forkert.
   Anbefaling: Hvis formatet anvendes, bør hele opgaven kun scores som 1 point (rigtigt eller forkert).

Dette kan implementeres med ubetydelige ændringer af prøverne i øvrigt.

9. Bilag

9.1 Dokumentation af elevbesvarelser

Samlede og systematiserede elevbesvarelser fra diverse prøver (Excel.filer). **NB Fortroligt – omfatter elevers UNI-Login (personhenførbar oplysning).**

Bilag 9.1.1 Materialet består af et ark med 5 faneblade med data fra årets prøver: Det ene indeholder data fra FP9 Matematik med hjælpemidler (FP9_2021_MMH) indsamlet af Epinion på regneark med censorindberetninger – dette datasæt er begrænset (1166 personers prøveresultater). De fire øvrige datasæt er komplette data indsamlet i Testog-prøver (FP9 Dansk – Læs og retskr. og FP9 Mat uden hj. Samt FP10 Dansk) (***Filnavn: FP9+FP10_2021_DK_MAT_Final.xlsx***).

Bilag 9.1.2. Materialet består af et regneark med 10 faneblade med data (scorer på item og prøveniveau) fra årets forcensurprøver. Der er 5 prøvesæt fra maj 2017 (Dansk læs og ret, Matematik m. og u. hjælpemidler samt engelsk - de tre delprøver) og 5 prøvesæt fra maj 2021 (Bio, Geo og Fysik/Kemi samt Tysk og Fransk). Matematik m. hjælpemidler er indsamlet af Epinion, de tre naturfag (Bio, Geo og F/K) er indsamlet i TestogPrøver, og de resterende 6 prøver er indsamlet med begge systemer. Alle resultater er på både item og totalscore-niveau. (***Filnavn: Forcensur_2021_data_final_Revideret01122021.xlsx***)

Bilag 9.1.3. Materialet består af en stor matrice i ét faneblad: Hovedet på matricen er prøver fra 2012-2021 fra de fag der har været i spil. I alt er der data fra 90 prøver. Forspalten er elever, der har besvaret prøverne. Frem til 2021 er ID et løbenummer – for Forcensurdata fra i 2021 er ID UNI-login. Cellerne i tabellen er scorer fra prøverne. Der er kun medtaget de rækker, som indeholder data, der har kunnet anvendes ifm. forcensur. Tabellen indeholder oplysninger fra 26040 personer og i alt 163.387 prøver. (***Filnavn: Forcensur samlet matrice final.xlsx***)

9.2 Tidligere fremsendte dokumenter

Bilag 9.2.1 "Afholdelse af forcensur og evaluering heraf", Delrapport, 2017

Bilag 9.2.2 "Evaluering af forcensur 2017-19" Slutrapport december 2019, Link: https://nordicmetrics.com/wp-content/uploads/2020/02/Bilag-4A.1-Rapport-fra-tidligere-forcensur.pdf

Bilag 9.2.3 Forslag om tilpasning af Epinions prøveformat til online-hjemmeprøver

9.3 Præsentationsmateriale til opgavekommissionsmøderne

Bilag 9.3.1. Præsentation til opgavekommissionen i dansk, 19. maj 2021

Bilag 9.3.2. Regneark med visning af karakterfordelinger pba. omsætningstabel i dansk

Bilag 9.3.3. Præsentation til opgavekommissionen i matematik, 20. maj 2021

Bilag 9.3.4. Regneark med visning af karakterfordelinger pba. omsætningstabel i matematik

9.4 Drejebog

Bilag 9.4. Revideret drejebog til afholdelse af forcensur

9.5 Vejledningsmateriale

Bilag 9.5.1. Vejledning til kerneskoler

Bilag 9.5.2. Vejledning til afholdelse af forcensurprøve i dansk læsning og retskrivning

Bilag 9.5.3. Prøveplan og levering af resultater

Bilag 9.5.4. Tilmeldingsblanket

Bilag 9.5.5. Eksempel på feedback på prøveresultater

9.6 Responsanalyser til opgavekommissioner

Bilag 9.6.1 Itemkarakteristikkurver (ICC) mv. for alle items – Biologi FP 2021.

Bilag 9.6.2 Itemkarakteristikkurver (ICC) mv. for alle items – Geografi FP 2021.

Bilag 9.6.3 Itemkarakteristikkurver (ICC) mv. for alle items – Fysik/Kemi FP 2021.

Epinion

Epinion Aarhus

Fellow Workspace
Mariane Thomsens Gade 4b
8000 Aarhus C
Denmark
T: +45 87 30 95 00
E: aarhus@epinionglobal.com

Epinion Copenhagen

Ryesgade 3F
2200 Copenhagen N
Denmark
T: +45 87 30 95 00
E: copenhagen@epinionglobal.com