Rapport fra Følgegruppen for én bedømmer ved folkeskolens prøver

Dolin, Jens; Nielsen, Keld; Rangvid, Beatrice Schindler

Publication date:
2018

Rapport fra

Følgegruppen for én bedømmer ved folkeskolens prøver

Marts 2018

Udarbejdet af

Jens Dolin

Keld Nielsen

Beatrice Schindler Rangvid

Indholdsfortegnelse

Introduktion

Den ændrede bedømmelsesordning ved folkeskolens skriftlige prøver blev indført med lovforslag nr. L181 af 26. maj 2015 (Indførelse af fælles prøve i fysik/kemi, biologi og geografi, ændring af bedømmelsesordning ved folkeskolens skriftlige prøver m.v.) og trådte i kraft den 1. januar 2016. Som daværende undervisningsminister Christine Antorini sagde ved fremlæggelsen: "Med lovforslaget lægges der endvidere op til, at den nuværende bedømmelsesordning i forbindelse med folkeskolens skriftlige prøver ændres, således at prøver med skriftlig besvarelse på 9. og 10. klassetrin fremover alene bedømmes af én statsligt beskikket censor." (Skriftlig fremsættelse af L181, 26. marts 2015).

Som et af de understøttende initiativer til implementering af ordningen med kun én censor blev der nedsat en følgegruppe, der skulle følge den ændrede bedømmelsesordning i to år.

Det fremgår af Kommissoriet af 13. april 2016, at

"Følgegruppen skal for eksempel bidrage til at

- vurdere perspektiverne i udbygning af de vejledende karakterbeskrivelser til at omfatte alle karakterer på skalaen.
- vurdere perspektiverne i censorgrupper med mulighed for faglig sparring om bedømmelserne og den konkrete udformning.
- vurdere behovet for yderligere anonymisering af elevbesvarelser efter det første år med UNI-Login som prøvenummer.
- rådgive ministeren og ministeriet på grundlag af de gennemførte prøver og de opnåede resultater."

Som Medlemmer af følgegruppen blev udpeget

- Professor og institutleder (ved rapportaflevering professor emeritus) Jens Dolin, Institut for Naturfagenes Didaktik, Københavns Universitet.
- Lektor emeritus Keld Nielsen, Institut for Datalogi, Aarhus Universitet.
- Ph.D. og seniorforsker Beatrice Schindler Rangvid, VIVE - Det Nationale Forsknings- og Analysecenter for Velfærd.

Følgegruppen har arbejdet tæt sammen med og er blevet sekretariatsbetjent af Undervisningsministeriets Kontor for Prøver, Test og Eksamen.

Følgegruppen har ladet dele af de statistiske beregninger foretage af professor Helle Sørensen, Laboratorium for Anvendt Statistik, Institut for Matematiske Fag, Københavns Universitet.

Resumé og anbefalinger fra Følgegruppen

Som et af de understøttende initiativer til implementering af ordningen med kun én censor til Folkeskolens skriftlige prøver blev der nedsat en følgegruppe, der skulle følge den ændrede bedømmelsesordning i to år.

Ændringen i bedømmelsesordningen vedrører de skriftlige prøver i dansk, skriftlig fremstilling, matematik med hjælpemidler og fremmedsprog på 9. og 10. klassetrin. Følgegruppen har koncentreret sig om prøverne i matematik og dansk på 9. klassetrin. Den har gennem et spørgeskema undersøgt hvorledes den nye censorordning fungerer set fra censorernes side og ved genretning af et større sample af prøvebesvarelser analyseret problemer vedrørende anonymisering og censorpålidelighed.

I den nye ordning er hver censor medlem af en censorgruppe (”en sparregruppe”) på 4-5 censorer. Hvis en censor føler sig usikker på bedømmelsen af en besvarelse, kan han/hun kontakte en anden censor fra gruppen for at få en ekstra bedømmelse af den problematiske besvarelse.

På baggrund heraf kan vi sige:

Censorernes opfattelse af den nye ordning

Allerede det første år tog censorerne godt mod den nye ordning.

357 censorer – ud af 428 adspurgte – svarede på spørgsmål om, hvordan de synes den nye ordning fungerer alt i alt. 50% af dem svarede, at den nye ordning fungerer ”godt” eller ”særdeles godt”, mens ca. 30% svarede at den fungerer ”nogenlunde”.

349 censorer, der også havde været bedømmere under den gamle ordning, blev spurgt, hvordan de vurderer deres bedømmelsers sikkerhed i den nye ordning i forhold til den gamle. 69% af dem vurderede, at de i den nye ordning havde givet mere sikre bedømmelser end før.

335 censorer rapporterede, at de brugte de vejledende karakterbeskrivelser.

Censorernes tilbagemeldinger viser, at den tilstræbte anonymisering af elevernes besvarelser gennem brug af uni-login ikke funger, fordi en elevs uni-login består af (dele af) elevens fornavn og nogle tal. 359 censorer (det er 84% af de adspurgte) har svaret på, om de på grundlag af uni-login kan gennemskue elevens køn i mere end hver fjerde af de rettede besvarelser. Det svarer 74% af dem ”ja” til.

Censorerne blev også spurgt, om de kan gennemskue elevens etnicitet. 74% af dem svarer ”ja” til, at de ved mere end hver fjerde besvarelse kan se om der er tale om en etnisk dansk elev. 42% af dem svarer ”ja” til, at de ved mere end hver fjerde besvarelse kan gennemskue om der er tale om en elev med anden etnisk baggrund end dansk.

Endelig blev censorerne spurgt, om de brugte sparregruppen til at sparre med. 361 af censorerne svarede på, om de havde kontaktet en anden censor for at sparre. 81 censorer havde kontaktet en anden censor for at sparre. Ud af de 81 censorer, der faktisk kontaktede en anden censor for at sparre, svarede 97% at sparringen hjalp til at gøre den/de relevante bedømmelse(r) mere sikre.

Censorerne fremhævede følgende som positivt ved den nye ordning:

- Bedømmelsesproceduren er mere effektiv, blandt andet fordi censor ikke skal vente på sin medcensor, og fordi der ikke skal bruges tid på kommunikation med denne.
- Bedømmelsen er mere sikker, fordi censor har mere ro og tid til bedømmelsen og kan være mere systematisk, fordi rettearbejdet ikke skal koordineres med en eller flere medcensorer.
- Bedømmelsen er mere objektiv, fordi censor ikke er udsat for pres fra klassens lærer.

Censorernes bekymringer i forbindelse med den nye ordning omhandlede:

- Faglige konsekvenser af at censor ikke mere giver fagligt feedback til klassens lærer.
- Elevernes manglende retssikkerhed, herunder faren for at (uundgåelige) fejl i forbindelse med bedømmelsesarbejdet ikke bliver opdaget, når der kun er én bedømmer.
- Uklarhed på skolerne omkring klageprocedurer.
- Uklarhed omkring praktiske forhold som deadlines, kopiering og forsendelse af opgaver.

(Undersøgelsen i 2016 omtales nærmere i Bilag 3)

I forbindelse med prøverne i maj/juni 2017 har følgegruppen spurgt et antal udvalgte censorgrupper, hvilken betydning eksistensen af en sådan gruppe havde haft, eller kunne forventes at få, for deres arbejde, særligt med henblik på bedømmelsen af vanskeligt vurderbare besvarelser.

De adspurgte grupper fandt ret enstemmigt, at det er en god ide med censorgrupper, da medlemmerne kan have behov for sparring, men at der stadig er en del problemer med at medlemmerne er tilbageholdende med ”ulejlige” hinanden i en travl hverdag.

Grupperne fandt også, at med den nye ordning er problemer som ombytning af besvarelser, manglende overholdelse af deadlines, fejl i forbindelse med afholdelse af prøven, snyd (herunder kopiering fra internettet) etc. vanskeligere at opdage end før, og de peger på, at denne type af problemer og fejl ikke kan afhjælpes i forbindelse med oprettelsen af censorgrupper, således at der må findes andre løsninger.

Tilbage står, at censorernes største bekymring i forbindelse med den nye ordning - både i 2016 og 2017 - ikke er spørgsmålet om sikkerheden af deres bedømmelser, men derimod at den unikke viden, som censor efter afsluttet bedømmelse har om en klasses præstationer, og dermed indirekte om lærerens undervisning, nu ikke mere kommunikeres tilbage til læreren i form af en telefonsamtale om klassens karakterer. Dermed tabes vigtig information, der kan bidrage til formativ evaluering af lærerens undervisning.

Følgegruppen deler den opfattelse, at den manglende tilbagemelding fra censor til lærer er den største udfordring, som skabes af den nye ordning.

(Undersøgelsen i 2017 omtales nærmere i Bilag 4)

Problemer vedrørende anonymisering og censorpålidelighed

Vi finder kun små og statistisk usikre effekter af anonymiseringen med UNI-login på elevernes karakterer mht. køn og etnicitet. Det vurderes derfor at ulemperne ved en anonymisering med UNI-login sammenlignet med en fuld anonym bedømmelse er forholdsvis små.

Censorernes rettepålidelighed ved 9. klasses prøverne i dansk og matematik er hhv. 0,40 og 0,72, målt som sandsynligheden for at to karakterer givet af to forskellige censorer er ens. Dette er lavere end, men sammenligneligt med, udenlandske erfaringer.

Den lave rettepålidelighed i dansk skyldes dels vanskelighederne ved i dansk at opstille stringente rettekriterier (uden at påvirke undervisningen i negativ retning) og dels at bedømmelse af essay-lignende svarformater nødvendigvis må afspejle bedømmernes helt legitime forskellige fagopfattelser.

Vi kan ikke sammenligne pålideligheden af bedømmelse af prøver foretaget af lærer og én censor med pålideligheden af bedømmelsen foretaget af kun én censor. Men vi kan sammenligne pålideligheden af én censor med pålideligheden af to censorer. Både i dansk og matematik giver to censorer (et censorpar giver en fælles karakter) en højere pålidelighed end én. Det er især påfaldende i dansk, hvor sandsynligheden for at to forskellige censorpar giver samme karakter er 10%-point højere end sandsynligheden for at to enkeltcensorer giver samme karakter (det tilsvarende tal for matematik er 7%-point).

Af denne grund finder følgegruppen, at det er vigtigt, at sparregrupperne kommer til at fungere optimalt, da de netop er indført for at sikre, at der - i hvert fald lokalt – bliver formuleret en række fælles kriterier for bedømmelse af årets opgaver, samt at der i forbindelse med en bedømmelse, som en censor finder problematisk, kan inddrages en anden censor.

Anbefalinger

Med udgangspunkt i vores undersøgelser anbefales det at:

1. Der etableres et system, som sikrer tilbagemelding fra censor til lærer om klassens besvarelser. Det kan fx foregå ved at censor skal sende en oversigt over – og en karakteristik af – sin bedømmelse til klassens lærer, når bedømmelsen er overstået.

   For at gøre tilbagemeldingen præcis og overkommelig bør der indføres en rutine, hvor STUK hvert år (og for hvert fag) udarbejder et skema, som censor skal udfylde på grundlag af klassens besvarelser og derefter sende til læreren. Skemaet kan med fordel gøres elektronisk. Det bør eventuelt gøres muligt for læreren at ringe tilbage til censor, hvis læreren har opklarende spørgsmål til skemaet.

   I udarbejdelsen af en rutine for tilbagemelding fra censor bør man inddrage de erfaringer, som Københavns Kommune – i samarbejde med ministeriet – har indhøstet i et 2-årigt forsøg med at etablere feedback rutiner.

2. Censorrettepålideligheden øges ved at censorerne i højere grad end nu opbygger en fælles standard. Dette kan ske ved at etablere og facilitere strukturerede sociale processer blandt censorerne og ved at der etableres mere præcise kriterier. Vi anbefaler at

   Censorgruppernes interne sparring optimeres. Det kan fx ske ved at alle censorgrupper mødes virtuelt lige efter at prøven er overstået, dvs. inden – eller umiddelbart efter at – bedømmelsen er gået i gang. Formanden for gruppen vælger 3 besvarelser, som inden mødet sendes til alle gruppens medlemmer, så de kan være udgangspunkt for drøftelser på mødet.
   Mødet har to formål.
   A: At give mulighed for en fælles drøftelse af årets opgaver og niveauet for bedømmelse.
   B: At gruppen kan aftale "regler" for sparring, så eventuelle sociale og praktiske barrierer for at henvende sig i forbindelse med sparring reduceres.

   STUK bør eventuelt give anvisninger på hvordan de it-tekniske udfordringer i forbindelse med det virtuelle møde kan reduceres. Det kan eventuelt anbefales at grupperne bruger Skype. Dog vil den bedste løsning nok være, at Skolekom udstyres med mødefaciliteter.

   Man bør overveje, om det er muligt at beholde de samme censorgrupper 3-4 år i træk, så medlemmerne lærer hinanden at kende.

3. Censormøderne i større udstrækning bidrager til en professionalisering af censorerne.

   Det vil styrke censorkorpset at gøre det obligatorisk for censorer at deltage i censormødet.

4. Der strammes op på kriterier for bedømmelser i dansk og sprogfagene.

Det kan fx ske ved at fagdidaktikere i samarbejde med den relevante censorgruppe diskuterer sig frem til i hvilket omfang bedømmelseskriterierne kan præciseres, fx ved at de vejledende karakterbeskrivelser udbygges til at omfatte alle karakterniveauer.

5. Regler, praktiske forhold og logistik vedrørende behandling af prøvebesvarelserne præciseres for at lette censorarbejdet og minimere muligheden for fejl.

Dette gælder fx rutiner og regler for, hvordan skolerne skal agere i forbindelse med indsamling, forsendelse og opbevaring af besvarelser og reglerne for hvordan klagesager håndteres.

6. Der ikke er tilstrækkeligt evidens for skævvridende effekter af den delvist gennemskuelige anonymisering med UNI-login til at anbefale at ændre anonymiseringsmetoden på kort sigt.

Af mere principielle årsager og fordi vi finder mindre forskelle i bedømmelsen mellem drenge/piger og danske/etniske elever, bør det dog på lidt længere sigt overvejes, om der skal indføres et bedømmelsessystem, der kan sikre en fuldt anonym bedømmelse ved folkeskolens prøver.

De behandlede problematikker og de anvendte forskningsmetoder

Følgegruppen har holdt en lang række møder internt og fem møder med STUK. Gennem disse møder blev der et øget fokus mod hvorledes censorgrupperne kunne organisere deres arbejde og herigennem optimere kvaliteten og pålideligheden af rettearbejdet.

Kvaliteten af en test eller prøve bedømmes normalt ved at vurdere gyldigheden (validiteten) og pålideligheden (reliabiliteten) af prøven. Ændringen i bedømmelsesordningen berør i sagens natur ikke prøvernes indhold (og dermed deres validitet), men blot bedømmelsen heraf. Følgegruppen har kun haft til opdrag at undersøge selve bedømmelsen af prøverne, hvorfor vi i denne rapport kun berører validiteten perifert, dvs. vi vurderer ikke i hvilket omfang prøven rent faktisk prøver de færdigheder og kompetencer, som skal opfyldes ifølge Fælles Mål. Validiteten kan ellers opfattes som den vigtigste egenskab ved en test (Messick, 1989), men ingen prøver er fuldt valide, dvs. kan måle alt det, som man er interesseret i at teste. Tid og økonomi og prøveformater sætter grænser for hvor præcist elevens faglighed kan måles. Derimod er det vigtigt ved en high stakes test (dvs. en test hvis resultat har stor betydning for den testede og måske også for lærer og skole) at den er pålidelig, dvs. at den behandler alle elever lige, så denne egenskab vil ofte af systemet blive vægtet højt, måske endda højere end gyldigheden. Det er da også pålideligheden, der er blevet problematiseret ved kun at lade en censor rette prøverne, og det er derfor pålideligheden, denne undersøgelse koncentrerer sig om. Derudover kan der være en vis modsætning mellem pålidelighed og gyldighed. Fx er en multiple choice test billig og pålidelig, den kan scores let og præcist, men er ofte ikke særlig valid, idet den i mange af de i praksis anvendte udformninger kun kan indfange ret simple faglige egenskaber. En god skriftlig prøve med open response svar og en mundlig prøve har mulighed for at indfange mere avancerede svar og kan dermed have en højere gyldighed, ligesom en længerevarende projektprøve kan indfange avancerede kompetencer, men de er svære at score fuldt pålideligt.

En af de vigtigste måder til at øge pålideligheden er at bedømmerne afstemmer deres vurderingskriterier for at opnå en *fælles standard*. Vi har bedt udvalgte censorer om at anvende forskellige måder at afstemme deres rettelser på og give skriftlige tilbagemeldinger på hvorledes de virkede. I forlængelse heraf kan pålideligheden øges systematisk ved at prøveudbyder præciserer og operationaliserer kriterierne for de anvendte karakterer, fx ved at udbygge vejledende karakterbeskrivelser inden for de forskellige prøver.

Et andet kvalitetsmål inden for pålidelighedsfeltet handler om *objektivitet*, dvs. at resultatet af prøven er uafhængigt af den person, der foretager bedømmelsen, og især personens kendskab til eleven, således at karaktergivningen fx ikke er påvirket af bedømmerens opfattelse af forskellige elevgrupper. Vi undersøger i denne sammenhæng hvad det betyder, hvis censor har kendskab til elevens køn og/eller etnicitet.

Ud over pålidelighed og objektivitet kan et prøveresultat påvirkes af eventuelle *fejl* i administrationen af prøven. Det kan fx være i forbindelse med registrering, forsendelse og

opbevaring af prøverne. Det har ligget uden for rapporten at vurdere omfanget af sådanne fejl, men i forbindelse med behandlingen af censorsvarene gives der nogle eksempler på mulige fejlkilder. Desuden kunne det være interessant at undersøge hvorvidt de få eksempler på meget store forskelle mellem prøvekarakter og genretning skyldes administrative fejl.

Pålidelighedsproblematikken

Pålidelighed, også kaldet reliabiliteten, forstås i denne sammenhæng som et statistisk begreb, der udtrykker målenøjagtigheden. Det kan enten være som *repeterbarhed*, dvs. i hvilket omfang to målinger giver samme resultat, hvis de gentages, altså en individuel (intern) censornøjagtighed (fx to bedømmelser af samme elev fra samme censor), eller som *reproducerbarhed*, dvs. i hvilket omfang to personer, der uafhængigt af hinanden bedømmer den samme besvarelse, får samme resultat, altså en mellem-censor nøjagtighed. Begge mål kan kvantificeres som sandsynligheden for at de to bedømmelser (foretaget under samme vilkår) afviger en vis størrelse fra hinanden.

Pålideligheden mindskes ved at:

1. En bestemt elevs præstation afhænger af hvilket spørgsmål der trækkes
2. Samme elev præsterer forskelligt til forskellige tider
3. Den enkelte censor bedømmer den samme præstation forskelligt til forskellige tider eller situationer (repeterbarhed).
4. Forskellige bedømmere giver forskellig bedømmelse for samme præstation (reproducerbarhed)

Den første upålidelighedsfaktor "the luck of the draw" kan ud over individuelle effekter også forklare årsvariationer i prøvekarakterer for hele årgange i det omfang prøverne forskellige år har forskellige sværhedsgrader. Denne upålidelighedsfaktor samt den enkelte elevs variation i præstation accepteres af de fleste som et fælles vilkår ved test og eksaminer og indgår ikke i vores vurdering af pålideligheden.

Hvad angår bedømmerne, viser forskning at undervisere generelt er upålidelige bedømmere:

"Et århundredes forskning har konsistent vist at lærere ikke er pålidelige bedømmere af elevers læring, hvis ikke de bruger strategier til at reducere målefejl" (McMillan, 2013, p. 110)

McMillan (2013) refererer et forskningsforsøg hvor 300 essays rettet af 53 forskellige censorer. 94 % af opgaverne fik 7 forskellige karakterer på en 7-trin skala.

Det er derfor vigtigt at anvende forskellige strategier til at reducere vurderingsupålidelighed, og i et censorsystem vil det primært kunne ske ved at

1. Opstille retningslinjer eller skemaer for bedømmelse

2. Danne praksisfællesskaber til at opbygge en fælles forståelse af hvad man kan forvente af elever.

Metoder til undersøgelse af pålidelighed

Såvel repeterbarhed og reproducerbarhed som objektivitet er pålidelighedsproblemer, der i undersøgelsen er analyseret ved at genrette prøver fra Folkeskolens Afgangsprøve, sommer 2016.

Der er udtrukket 30 censorer (ca. hver femte på censorlisten) inden for både dansk og matematik som skal fungere som kontrolrettere.

Der udtrækkes grupper af 5 elever fra hvert af 30 store prøvesæt. Disse elevers besvarelser er rettet af en censor efter de normale censorretningslinjer. Hvert sæt af 5 elever sendes til en af de 30 udtrukne kontrolrettere i hhv. matematik og dansk til genretning.

Det samlede sample af elevbesvarelser (150 stk.) sendes derefter til retning hos en ekspertretter i matematik og en ekspertretter i dansk, udpeget af læringskonsulenterne. For hver elev er der således tre karakterer for dansk og tre karakterer for matematik:

En prøvekarakter, en kontrolkarakter og en ekspertkarakter.

Disse data er analyseret statistisk af professor MSO Helle Sørensen fra Laboratorium for Anvendt Statistik, Institut for Matematiske Fag, Københavns Universitet. Der er opstillet en analysemodel der dels beskriver bedømmelsen for en gennemsnitsdreng eller -pige som går på en gennemsnitsskole, foretaget af en gennemsnitscensor som enten er prøve-, kontrol-, eller ekspertcensor, og dels inkluderer fire kilder til tilfældig variation i modellen: Elev-, skole-, censor-, og residualvariation. Det estimeres hvor fremtrædende de enkelte kilder til variation er:

– Elev-til-elev variationen er udtryk for forskelle mellem bedømmelser af forskellige elever af samme køn og fra samme skole, foretaget af samme censor. Denne variation beskriver altså forskellen mellem elever.
– Skole-til-skole variationen er udtryk for forskelle mellem gennemsnitselever på forskellige skoler.
– Censor-til-censor variation er udtryk for forskelle mellem bedømmelser af samme besvarelse foretaget af forskellige censorer (af samme type). Størrelsen af censorvariationen tillades at afhænge af censortypen, så det fx er muligt at beskrive hvorvidt ekspertcensorer er mere enige end prøvecensorer.
– Residualvariationen er den variation der ikke kan forklares af ovenstående faktorer, specielt forskelle mellem bedømmelser fra samme elev og samme censor; altså hvad vi kan tænke på som "dag-til-dag variation": Hvor forskelligt vil en censor bedømme den samme besvarelse hvis han/hun ser den flere gange.

Estimater fra modellen bruges til at beregne sandsynligheder vedr. pålidelighed af bedømmelserne i to censurregimer:

1. Der er kun en censor, som giver karakteren alene
2. Der er to censorer.

Analysen er bragt som Bilag 1, hvor analysemetoden er gennemgået grundigt.

Anonymiseringsproblematikken

Et af formålene med ændringen af bedømmelsesordningen har været at gøre bedømmelsen af de skriftlige opgaver helt objektivt, således at hverken personligt kendskab til eleven (fra lærerens side) eller generelle forventninger til elevgrupper (fra censors side) får indflydelse på opgavebedømmelsen ved folkeskolens skriftlige prøver.

Til og med sommer 2015 blev folkeskolens skriftlige prøver rettet af både elevens lærer og en ekstern censor, og prøverne blev afleveret ikke-anonyme til bedømmerne, idet elevernes fulde navn var skrevet på opgavebesvarelserne. Den ikke-anonyme bedømmelse kan dog anses for problematisk, fordi international faglitteratur viser, at det kan give anledning til, at bedømmelsen ikke udelukkende baseres på elevens præstation i den specifikke prøve, men kan være influeret af lærerens kendskab til eleven (fx elevens opførsel) eller censors generelle faglige forventninger til den gruppe eleven tilhører, fx drenge/piger, etnicitet (såkaldt statistisk bias). Eksempler på sådanne studier i faglitteraturen er Hanna & Linden (2012), Lavy (2008), og Hinnerich, Höglin & Johannesson (2015).

Ændringen i bedømmelsesordningen skulle bidrage til at gøre bedømmelsen ved folkeskolens prøver objektiv ved at gøre bedømmelsen af opgaverne anonym. Fra sommer 2016 foretages bedømmelsen således for det første udelukkende af en ekstern censor og for det andet er opgaveafleveringen forsøgt anonymiseret. I stedet for elevens navn, skrives nu elevens UNI•Login på opgavebesvarelserne. Hvor man med det første tiltag undgår, at personlige forhold får indflydelse på bedømmelsen, er det ikke klart, at anonymiseringen med UNI•Login fungerer efter hensigten. UNI•Login giver nemlig kun en delvis anonymisering, da elevens køn og etnicitet ofte kan udledes af UNI•Login (jf. boks 1). Følgegruppen har derfor undersøgt, om karakterforskellene mellem elevgrupper (drenge vs. Piger; dansk vs. indvandrerbaggrund) er de samme uanset, om elevernes UNI•Login er skrevet på opgaverne eller et fuldt anonymt prøvenummer.

Boks 1: Delvis anonymisering af køn og etnicitet med UNI-login

- UNI•Login er et digitalt id for børn, unge og ansatte på institutioner, som giver adgang til nationale tjenester og en lang række pædagogiske services og online læremidler. Fx adgang til skolens intranet og de nationale test.
- UNI•Login består af fire bogstaver efterfulgt af fire tal/cifre. De fire bogstaver er begyndelsen af elevens fornavn og kan derfor bære information om elevens køn og etnicitet. De fire tal bærer ingen information. Når der skrives om UNI•Login i dette afsnit,

refereres typisk kun til bogstavdelen.

- I mange tilfælde vil de første 4 bogstaver af fornavnet afsløre elevens køn og etnicitet. Fx begynder kun mandlige fornavne med *jaco* eller *rasm*, mens kun piger begynder med *caro* eller *sofi*. Andre UNI•Login vil til gengæld ikke afsløre elevens køn, da det kan dække over både drenge- eller pigenavne. Eksempler herpå er fx *nico*, som dækker over både Nicolaj eller Nicoline, og *math*, som kan stå for både Mathias eller Mathilde. Tilsvarende vil UNI•Logins som *moha* eller *fati* afsløre at eleven har indvandrerbaggrund. UNI•Login skjuler derfor ikke køn og/eller etnicitet hos alle elever.

Metoder til undersøgelse af anonymiseringsproblematikken

For at undersøge, om nogle elevgrupper ville være blevet relativt bedre bedømt ved fuld anonymitet sammenlignet med den kun delvise anonymisering ved UNI•Login, har UVM på foranledning af følgegruppen fået genbedømt (kontrolrettet) en række opgavebesvarelser fra sommer 2016, hvor elevens UNI•Login er erstattet med et fuldt anonymt prøvenummer (jf. også boks 2). For hver af disse opgaver foreligger der således både en karakterer ved bedømmelse med UNI•Login og ved bedømmelse med prøvenummer. Ved at sammenligne forskellene i gennemsnitskarakterer ved de to typer bedømmelse mellem elevgrupper undersøges, om der er systematiske forskelle ved anonym versus ikke-anonym bedømmelse. For eksempel ville man på baggrund af resultaterne fra den eksisterende fagliteratur på området forvente, at drenge får bedre karakterer i dansk, skriftlig fremstilling, sammenlignet med pigerne, når bedømmelsen foregår helt anonymt, dvs. med prøvenummer, end når UNI•Login afslører, at opgaven er skrevet af en dreng.

Problemstillingen er endvidere belyst med en anden metode, som udnytter, at nogle UNI•Logins entydigt afslører elevens køn/etnicitet, mens andre UNI•Logins er neutrale. Her sammenlignes hvordan elever, der har et UNI•Login, der afslører elevens køn, klarer sig sammenlignet med elever, hvor UNI•Login ikke afslører, om det er en dreng eller en pige. Hvis der ikke er systematisk forskel i karaktererne, er det et tegn på, at det ikke er problematisk, at UNI•Login ikke anonymiserer køn hos alle elever. Denne undersøgelse er kun kort beskrevet her, men er grundigt dokumenteret i et særskilt arbejdspapir (Rangvid, 2018).

Boks 2: Dataindsamling vedr. kontrolrettelserne

- Tilfældigt udtræk blandt elever som har et UNI•Login, der afslører deres køn og etnicitet.For at understøtte denne udvælgelse empirisk har STIL (Styrelse for it og læring) produceret lister, der for hver UL viser, hvor mange elever med dette UL der er piger/drenge hhv. danske/etniske elever. Fx har elever i gruppen "danske drenge" alle et UNI•LOGIN-login, der entydigt identificerer eleven som dansk dreng, forstået ved at det altovervejende er drenge, der har dette UNI•LOGIN-login.
- Bruttostikprøve: 360 elever i 9. klasse (2015/2016) fordelt ligeligt på 4 kategorier (danske drenge, danske piger, drenge med indvandrerbaggrund, piger med indvandrerbaggrund).

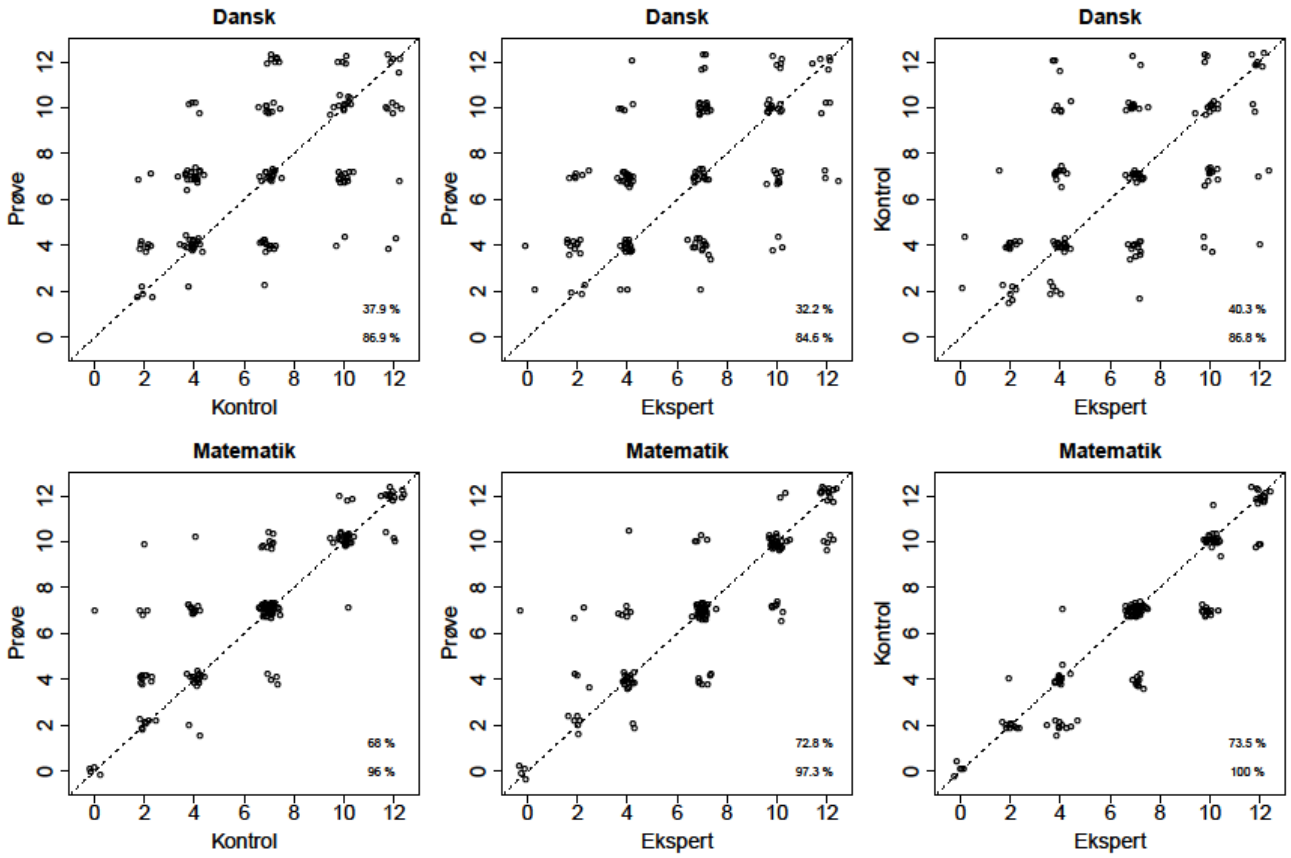
Bruttostikprøven er den samme for dansk og matematik, dvs. opgaverne i dansk og matematik er søgt indhentet for de samme elever.

- Analysestikprøve: 243 for dansk og 258 for matematik med både prøve- og kontrolkarakterer. En del af eleverne var fritaget. Andre årsager til at eleven ikke indgår i analysestikprøven er fx at skolen ikke kunne levere opgaverne, at eleven ikke gik på den skole, som var oplyst i registrene, og at enkelte censorer ikke afleverede kontrolrettelserne.
- Før udsendelse til kontrolrettelse er elevernes UNI•Login fjernet fra opgaverne og opgaverne har i stedet fået et fuldt anonymt prøvenummer. Opgaverne er sendt til hhv. 26 danskcensorer og 26 matematikcensorer. Disse censorer har også været censor ved folkeskolens prøver i sommer 2016. Hver censor har rettet opgaver fra hver af de fire elevgrupper.
- De samme materialer som ved den almindelig prøvebedømmelse blev medsendt/stillet til rådighed (fx rettevejledninger, karakterbeskrivelser mv.) for kontrolrettelsen.
- Ved at sammenligne gennemsnittet af karakterne på tværs af elever med hhv. UNI•Login og med fuld anonymitet vil man kunne vise, om nogle elevgrupper oplever en (relativ) fordel ved fuld anonym bedømmelse.

Pålideligheden af Folkeskolens prøver

Pålideligheden er beregnet på baggrund af Folkeskolens prøver i dansk og matematik ved sommerprøven 2016. Besvarelserne for 150 elever er rettet på tre forskellige måder, nemlig af en censor under almindelige rettevilkår, af en kontrolretter og af en ekspertretter. Resultaterne nedenfor er alle hentet fra Helle Sørensens udregninger i Bilag 1.

Nedenfor er de tre karakterer tegnet mod hinanden for hvert fag, dvs. kontrol vs prøve, ekspert vs prøve og ekspert vs kontrol. I alle plots er indlagt diagonalen svarende til at de to karakterer er ens, og der er tilføjet lidt støj til målingerne for at illustrere antal hændelser på de enkelte positioner. Desuden er nederst til højre i hvert plot angivet hvor mange procent af eleverne der får samme karakter ved de to retninger (øverst), samt hvor mange procent af eleverne hvor de to karakterer højst afviger en enkelt karakter (nederst).



Det er tydeligt at matematikkarakterne samler sig mere om diagonalen end danskkarakterne, hvilket viser at der er større overensstemmelse mellem matematikkarakterer end mellem danskkarakterer. Det er næppe overraskende.

Der er syv besvarelser i dansk der både får bedømmelsen 12 og 4 svarende til en forskel på tre karakterer. Den tredje karakter for disse besvarelser er 7 i fire tilfælde og 10 i tre tilfælde. En enkelt besvarelse i matematik bliver bedømt til både 02 og 10 (med 4 som tredje karakter). Det

kunne være interessant at kigge kvalitativt på disse besvarelser, fx for at undersøge i hvilket omfang der er tale om egentlige fejl i prøveadministrationen.

Beregningerne på datamaterialet viser følgende.

For dansk:

- Der er ikke evidens for at karakterniveauet givet ved prøve, kontrolretning og ekspertretning afviger fra hinanden på populationsniveau. De forskelle som ses i datasættet kan meget vel blot skyldes den tilfældige udvælgelse af censorer til datasættet.
- Ekspertcensorerne bedømmer lidt mere ens end de almindeligt udpegede censorer, men det er ikke statistisk signifikant og kan således sagtens være udslag af tilfældige udsving.
- Dag-til-dag variationen (altså den enkelte censors forskel i bedømmelse af samme opgave) er større end variationen mellem censorer, eller udtrykt anderledes: en censor er mere internt uenig med sig selv end han/hun i gennemsnit er uenig med en anden censor.
- For censurregime 1 (kun en censor) vil der for to karakterer givet af to forskellige censorer være 40% sandsynlighed for at de to karakterer er ens, og 87% sandsynlighed for at de højst afviger med en karakter.
- For censurregime 2 (to censorer) vil der for to karakterer givet af to forskellige censorpar være 50% sandsynlighed for at de to karakterer er ens, og 95% sandsynlighed for at de højst afviger med en karakter.

For matematik:

- Kontrolkarakterne er signifikant lavere end både prøvekarakterne (p-værdi 0.017) og ekspertkarakterne (p-værdi 0.003). En p-værdi på 0.017 viser at der er 1,7% sandsynlighed for at det er tilfældigt, at kontrolkarakteren er lavere end prøvekarakterne.
- Censor-til-censor variationen er mindst for ekspertcensorerne, hvilket antyder at de bedømmer mere ens end de andre typer censorer, men forskellen er ikke statistisk signifikant (p-værdi 0.08) og kan således sagtens være udslag af tilfældige udsving.
- Dag-til-dag variationen (altså den enkelte censors forskel i bedømmelse af samme opgave) er af samme størrelsesorden som variationen mellem censorer, eller udtryk anderledes: en censor er lige så internt uenig med sig selv som han/hun i gennemsnit er uenig med en anden censor.
- For censurregime 1 (kun en censor) vil der for to karakterer givet af to forskellige censorer være 72% sandsynlighed for at de to karakterer er ens, og 99% sandsynlighed for at de højst afviger med en karakter.
- For censurregime 2 (to censorer) vil der for to karakterer givet af to forskellige censorpar være 79% sandsynlighed for at de to karakterer er ens, og 99% sandsynlighed for at de højst afviger med en karakter.

Sammenligning mellem dansk og matematik:

- Selv om det ligger uden for opdraget, gav datamaterialet mulighed for at sammenligne drenge og piger, og der er signifikant populationsforskel på drenge og piger i dansk, men ikke matematik. Piger får i dansk i snit mellem et halvt og et helt trin højere på 12-trinsskalaen end drenge.
- Bedømmelsen afhænger mindre af den enkelte censor og tilfældigheder for matematik end for dansk.
- Skoleforskelle er mere betydelige i matematik end i dansk.
- De mere ekstreme karakterer bruges i større omfang i matematik end i dansk. Blandt samtlige karakterer i datasættet var der fx kun to 00'er og 41 12-taller i dansk men 14 00'er og 48 12-taller i matematik.
- For begge censurregimer er sandsynligheden for at to bedømmelser er ens eller højst afviger en enkelt karakter større for matematik end for dansk.

Disse resultater kan tolkes og vurderes på mange måder, men tre aspekter falder i øjnene:

1. Pålidelighedsniveauet af censorer

Censorerne har under de givne betingelser (dvs. under anvendelse af de ministerielle karakterniveaubeskrivelser og de anvendte censorgruppesamarbejder) en generel pålidelighed – målt som sandsynligheden for at to karakterer givet af to forskellige censorer er ens - på 0,40 i dansk og 0,72 i matematik. Hvorvidt dette er godt eller dårligt afhænger af hvilket niveau man ønsker.

Det er imidlertid til en vis grad i overensstemmelse med internationale erfaringer. Et review over forskning i censorpålidelighed bestilt af det engelske kontor for kvalitet og eksamination, Ofquel, finder fx (s. 22): "Estimates of marker agreement from blind double-marking studies in O level, A level and GCSE range from a correlation between markers of 0.73 in English A level (Murphy, 1978) to 0.997 in GCSE Mathematics (Newton, 1996)" (Tisi, 2015). At anvende korrelationer som mål for pålidelighed er en anden metode end den i indeværende rapport brugte (andelen hvor to censorer giver samme bedømmelse), og de to mål er derfor ikke direkte sammenlignelige. Korrelationsmålene for figuren s. 14 er for dansk: 0.55, 0.55, 0.56 og for matematik: 0.85, 0.86, 0.93. De danske værdier er lidt under de engelske, især i dansk, hvilket bl.a. kan skyldes de meget præcise engelske eksamenskrav.

Om ikke andet viser tallene med stor tydelighed hvor forsigtig man skal være med at lægge prøvekarakterer til grund for elevselektion. Det er simpelthen ikke muligt at give fuldt pålidelige bedømmelser af elevers prøveresultater.

Der er indikationer på at censorer med stor erfaring i at censorere har større pålidelighed. Dette viser nytten af at opbygge et stærkt censorkorps med stor fastholdelse.

2. De store forskelle i pålidelighed mellem dansk og matematik

Da censorgrupperne fungerer på samme måde i dansk og matematik, må forskellene sandsynligvis tilskrives de forskellige betingelser og kriterier for bedømmelse i de to fag. I matematik er det lettere at opsætte entydige kriterier for de enkelte karakterer, og censorerne har i højere grad sådanne fælles standarder på grund af fagets egenart, således at det er lettere at opnå en høj pålidelighed (interrater reliability) og at anvende hele karakterskalaen. Dette kunne friste til at gøre kriterierne mere præcise i dansk og engelsk, således at den enkelte censor i højere grad kan følge nogle givne retningslinjer for bedømmelserne.

At mere præcise generelle kriterier ikke nødvendigvis er svaret, skyldes dels de faglige konsekvenser og dels generelle kriteriers begrænsninger. Rent fagligt vil et detaljeret sæt kriteriet i et fag som dansk risikere at fremme en utilsigtet instrumentalisering af faget. Detaljeringsgraden skal således balancere mellem at give rum til variation og kreativitet på den ene side og på den anden side hensynet til især svagere elever, der har behov for ret faste retningslinjer for den gode besvarelse. At generelle kriterier har begrænset effekt på censorpålideligheden skyldes en kæde af forhold, som forskning viser gør sig gældende for censorer (Bloxham, 2015):

- Bedømmere har forskellig forståelse af givne standarder
- Selv nok så konkrete kriterier er komplekse og indeholder skjulte underkriterier
- generelle kriterier er ikke konsistente på tværs af fagområder.
- Bedømmere vurderer og vægter kriterier forskelligt i deres bedømmelser.
- Bedømmere har ofte andre kriterier end de officielt givne, som de inddrager i deres bedømmelse.
- sikring af erfaringsudveksling på organisatorisk niveau

Som det udtrykkes af Tisi (2013), så er det nødvendigt at acceptere en vis upålidelighed for narrative besvarelser:

”… lower levels of marker agreement on essay questions may be a result of legitimate differences in opinion between markers. There is a large body of literature that researches this area and argues that the use of questions with longer responses is an important part of the assessment process and so an educational system may choose to accept the lower levels of reliability.” (s. 2)

For at finde et rimeligt niveau af pålidelighed i dansk uden at det går ud over fagets udfoldelse synes det konstruktivt at danskdidaktikere i samarbejde med censorgruppen diskuterer sig frem til

i hvilket omfang bedømmelseskriterierne kan og bør præciseres. Det samme gør sig naturligvis gældende for andre fag med prøver der indebærer længere skriftlige svar.

3. Forskellen i pålidelighed mellem de to censorregimer – altså en eller to censorer

Både i dansk og matematik giver to censorer (et censorpar giver en fælles karakter) en højere pålidelighed end én. Det er især påfaldende i dansk, hvor sandsynligheden for at to forskellige *censorpar* giver samme karakter er 10%-point højere end sandsynligheden for at to *enkeltcensorer* giver samme karakter (det tilsvarende tal for matematik er 7%-point). Desuden: I dansk vil bedømmelserne fra to enkeltcensorer afvige med højst en karakter i 87% af tilfældene, mens afvigelsen mellem to censorpar vil være højst en karakter i 95% af tilfældene.

Uanset om der er én eller to censorer i matematik, er der 99% sandsynlighed for at to bedømmelser højst afviger med en karakter.

Det skal pointeres, at dette ikke siger noget om pålideligheden for situationen én censor og én lærer sammenlignet med situationen én censor. Men det kan påpeges, at i de tilfælde, hvor censor er i tvivl om en karakter, vil censor kunne kontakte en anden censor fra sin sparregruppe, og vi har så en situation med to censorer.

Anonymisering af Folkeskolens skriftlige prøver

Er anonymisering ved UNI•Login lige så god som en fuld anonym bedømmelse? Følgegruppen har undersøgt, om der er systematisk forskel i bedømmelsen for drenge og piger (hhv. danske og indvandrerelever) når elevens UNI•Login fremgår af opgavebesvarelsen sammenlignet med når der er skrevet et prøvenummer på opgaven ved at sammenligne karakterer fra sommerprøverne med karakterer fra genbedømmelser af de samme opgaver, hvor de har været fuldt anonymiserede.

Med udgangspunkt i resultaterne fra den international faglitteratur er det vores hypotese, at drenge og indvandrerelever får dårligere karakterer sammenlignet med piger og danske elever, når opgaven er påført et UNI•Login, der afslører elevens køn/etnicitet sammenlignet med den fuldt anonyme bedømmelse ved kontrolrettelse med prøvenummer. Det formodes desuden, at det spiller en større rolle ved opgavebedømmelsen i dansk (skriftlig fremstilling) end ved matematik (med hjælpemidler), fordi der her formodes at være mere spillerum for skøn ved karaktergivningen. Det undersøges i første omgang ved at sammenligne gennemsnitskarakterer hos forskellige elevgrupper (drenge/piger; danske/indvandrerbaggrund), efterfulgt af en række analyser med statistiske regressionsmetoder.

Tabel 1, øvre panel, viser de gennemsnitlige karakter, som elever i fire elevgrupper opnår i dansk stil. De fire elevgrupper er hhv. danske drenge, danske piger, drenge med indvandrerbaggrund og piger med indvandrerbaggrund. Ligesom ved analyserne vedr. pålidelighed i kapitel 3a, viser det sig også her, at prøvekarakterer generelt er højere end kontrolkarakterer. Det gælder alle fire elevgrupper. Tallene viser desuden, at piger generelt får højere karakterer end drengene i dansk stil, både i prøvekarakter og ved kontrolrettelsen. Det gælder både danske og indvandrerelever.

Det er dog forskellene i de gennemsnitlige prøve- og kontrolkaraktererne på tværs af elevgrupperne, der er relevante for problemstillingen mht. anonymitet. Så at sige: forskelle i forskellene. Forskellene i de gennemsnitlige prøve- og kontrolkaraktererne for de fire grupper er også vist i tabellen. De angiver, hvor meget bedre grupperne står sig, når prøverne bedømmes ikke-anonyme (prøvekarakter) i forhold til, når de bedømmes anonyme (kontrolkarakter). Her ser vi, at forskellen mellem prøve- og kontrolbedømmelsen er næsten lige stor for drenge og piger hos de danske elever. Her er der således ikke tegn på, at drengene står betydeligt dårligere relativt til pigerne ved en ikke-anonym bedømmelse, hvilket vi i udgangspunktet havde en formodning om. Hos eleverne med indvandrerbaggrund er forskellen mellem kønnene lidt større. Her tyder tallene på, at drengene ville have en lille fordel relativt til pigerne ved en anonym bedømmelse. Fordelen er 0,59 for danske drenge mod 0,66 karakterpoint for danske piger. Den tilsvarende forskel mellem kønnene for indvandrerelever er 0,85 og 1,12 karakterpoint.

Med hensyn til den etniske dimension, viser tallene i Tabel 1, øvre panel, at forskellene mellem prøve- og kontrolkaraktererne er lidt større når man sammenligner danske elever og elever med indvandrerbaggrund. Hos indvandrereleverne er fordelen ved at blive bedømt med UNI•Login

(ikke-anonym, prøvekarakter) nemlig større end hos danske elever. Hvor fordelen er på 0,59 karakterpoint hos danske drenge, så er den på 0,85 karakterpoint hos indvandrerdrenge, dvs. en forskel på 0,26 karakterpoint. Forskellen er af samme størrelsesorden for danske piger versus indvandrerpiger (0,66 vs. 1,12), dvs. en forskel på 0,46 karakterpoint. Modsat vores udgangshypotese lader det således til, at karakterne i dansk, skriftlig fremstilling, får et ekstra nøk opad, når censor via UNI•Login kan se, at eleven har indvandrerbaggrund, særligt hos pigerne. Det skal dog allerede nævnes her, at problemstillingen er undersøgt nærmere vha. regressionsanalyser, som beskrives lidt længere nede i dette afsnit. Disse resultater viser, at de her fundne forskelle ikke er statistisk signifikante, dvs. at der er betydelig statistisk usikkerhed omkring, hvorvidt denne forskel er reel eller kan skyldes tilfældigheder.

Tabel 1: Gennemsnitlige karakterer for de fire elevgrupper for hhv. prøvekarakterer og kontrolkarakterer

Dansk stil	Drenge, dansk	Piger, dansk	Drenge, indvandrer	Piger, indvandrer
(1) Prøvekarakter, ikke-anonym (UNI•Login)	5,68	7,41	4,39	5,93
(2)Kontrolkarakter, helt anonym	5,09	6,75	3,54	4,81
Forskel (1)-(2)	0,59	0,66	0,85	1,12

Matematik	Drenge, dansk	Piger, dansk	Drenge, indvandrer	Piger, indvandrer
(1)Prøvekarakter, ikke-anonym (UNI•Login)	7,22	6,70	4,75	5,53
(2)Kontrolkarakter, helt anonym	7,03	6,64	4,29	5,04
Forskel (1)-(2)	0,19	0,06	0,46	0,49

Figur 1, nedre panel, viser de tilsvarende resultater for matematik (med hjælpemidler). For alle elevgrupper gælder, at forskellen mellem prøve- og kontrolkarakterer er klart mindre i matematik end i dansk. Det samme overordnede mønster blev også dokumenteret ved kontrolrettelserne vedrørende pålidelighed i afsnit 3a. Hvad der er afgørende for nærværende analyserne vedr. anonymitet er dog ikke forskelle, der er *fælles* for grupperne, men udelukkende forskelle på tværs af grupperne mht. de to dimensioner: køn og etnicitet. Her finder vi præcist samme mønster som for dansk, nemlig at der næsten ikke kan spores meget forskel mellem drenge og piger hvad angår hvor meget prøvekarakteren i gennemsnit afviger fra kontrolkarakteren, mens den ikke-anonyme bedømmelse med UNI•Login lader til at være en større fordel for elever med indvandrerbaggrund end for elever med dansk baggrund. Også her vil resultaterne fra regressionsanalysen herunder dog vise, at denne forskel ikke er statistisk sikker.

Udover den deskriptive undersøgelse har følgegruppen også fået gennemført en række statistiske regressionsanalyser for at undersøge de fundne forskelle mere grundigt. De statistiske analyser giver bedre mulighed for at tage højde for datastrukturen. For eksempel kan være effekter på censorniveau, da hver censor har rettet et antal opgaver (klynger). De detaljerede regressionsresultater kan ses i et særskilt notat (Sørensen, 2017b).

Når man på baggrund af regressionsresultaterne udfører test af regressionsresultaternes statistiske signifikans for at belyse, om de forskelle der er statistisk sikre eller kan være resultat af den usikkerhed der er omkring data, så finder vi, at der er en tendens til, at drenge i dansk har en lille fordel ved anonym bedømmelse sammenlignet med pigerne, og omvendt i matematik. Desuden ser vi en tendens til at indvandrerelever har en mindre fordel ved ikke-anonym bedømmelse i begge fag. Ingen af disse resultater er dog statistisk sikre (signifikante). Den statistiske analyse giver således ikke evidens for at sige, at de fire grupper adskiller sig i forhold til, hvor meget højere prøvekarakteren er end kontrolkarakteren. Rangvid (2018) har desuden udført regressioner, som særskilt belyser de to dimensioner køn og etnicitet, dvs. hvor der kun opdeles i to grupper ad gangen (hhv. drenge og piger; danske og indvandrerelever) for at få mere statistisk styrke (power). Disse analyser bekræfter de her fundne resultater, idet de viser, at der ikke er belæg for at konkludere, at drenge eller indvandrerelever får bedre karakterer ved anonym bedømmelse (kontrolkarakteren). Med andre ord: det lader ikke til, at en hel anonym bedømmelse gør en forskel for kønsgabet eller det etniske gab i prøvekaraktererne. Der er endvidere udført en række følsomhedsanalyser i henhold til hovedspecifikationen af den statistiske model, som viser, at hovedkonklusionen er robust.

Udover hovedanalysen er der kørt 3 andre modeller: en ordinal regression på fortegnet af forskellen (plus/nul/minus), en todimensional analyse af eksamens- og kontrolkarakter baseret på normalfordelingsmodel (dif-in-dif), og en todimensional analyse af eksamens- og kontrolkarakter baseret på ordinal regression. Beslægtede analyser af samme problemstilling, som er dokumenteret i et working paper (Rangvid, 2018), bekræfter de her fundne konklusioner.

Analysen med anonymiserede kontrolrettelser er udført for et begrænset antal elever. Problemstillingen er dog også blevet undersøgt på anden vis, som giver mulighed for at tage udgangspunkt i hele elevårgangen, der gik op til folkeskolens prøver i 2016. Resultaterne fra denne analyse er dokumenteret i et særskilt arbejdspapir (Rangvid, 2017, op.cit.). Ved metoden ovenfor belyses problemstillingen ved at sammenligne prøvekarakterer hos elever, hvor UNI•Login afslører deres køn (=ikke-anonym bedømmelse), med bedømmelsen fra en fuldt anonymiseret kontrolrettelse af opgaverne (=anonym bedømmelse). Den anden metode adskiller sig ved at det anonyme sammenligningsgrundlag ikke er kontrolrettelserne, men prøvekarakterer hos elever, hvis køn ikke afsløres af UNI•Login. Metoden hviler på den antagelse, at det er tilnærmelsesvist tilfældigt, om eleverne har et fornavn, der afslører deres køn eller ej. Antagelsen forekommer rimeligt, idet den kun kræver, at forældrene ikke har navngivet børnene systematisk mht. om de første fire bogstaver afslører køn eller ej. (Den tilsvarende antagelse for etnicitet er svagere.) Vi

har dog, for at sikre, at de to elevgrupper ligner hinanden mest muligt, medtaget en række kontrolvariabler i regressionen. Af særlig betydning er, at eleverne ligner hinanden mht. deres boglige færdigheder. Kontrollerne inkluderer derfor resultater fra elevernes præstationer ved de nationale test på tidligere klassetrin. Som et eksempel vil sammenligningen være mellem drenge, der har et afslørende UNI•Login - såsom Jacob eller Rasmus - og drenge, der har UNI•Logins, der bruges af både drenge og piger, som fx Mathias (som også bruges af piger, der hedder Mathilde; jf. boks 1). Sammenligningen vil foregå på tilsvarende vis hos pigerne mellem piger, der fx hedder Caroline eller Sofie med piger, som hedder Nicoline. Resultaterne viser, at der er tendens til at drenge har en lille fordel ved anonym bedømmelse sammenlignet med pigerne, mens det omvendte gør sig gældende i matematik. Ingen af disse tendenser er dog statistisk signifikante.

Konklusionen her bekræfter således resultaterne fra kontrolrettelserne, idet der ikke findes statistisk signifikante effekter af fuldt anonym bedømmelse, hverken mht. køn eller etnicitet.

Analyserne vedr. anonymisering ved UNI•Login viste, at der er en tendens til, at anonym bedømmelse gavner drengene i dansk, pigerne i matematik og indvandrerelever i begge fag. Disse tendenser er dog både små og ikke statistisk sikre. Den overordnede konklusion er derfor, at der måske er små effekter af anonym bedømmelse, men at vi ikke har kunnet finde entydig og sikker evidens for, at karaktergivningen på tværs af køn og etnicitet er systematisk anderledes med UNI•login end ved en fuldt anonym bedømmelse.

Censorernes egne vurderinger af retningerne af prøverne

Gennem to spørgeskemaundersøgelser er indhentet en lang række data, såvel skalabesvarelse som narrative udsagn, fra 2016 og 2017, vedrørende praktiske forhold i den nye ordning, som dels er positive og dels bekymrende. Der gives her et resumé af censorernes svar.

Censorernes vurdering efter de skriftlige prøver i 2016:

Efter at bedømmelsen i forbindelse med 9. klasses prøver 2016 i skriftlig engelsk, dansk og matematik var afsluttet, deltog 428 censorer i en spørgeskemaundersøgelse om deres erfaringer med den nye ordning – 46 fra engelsk, 161 fra dansk og 221 fra matematik.

Nogle spørgsmål skulle besvares ved afkrydsning af allerede formulerede, "lukkede" svar. Andre svar kunne gives i åbne kommentarfelter. Svarelysten og seriøsiteten var stor. Følgegruppen takker de deltagende censorer for de mange svar.

Det generelle billede på baggrund af prøverne i maj/juni 2016 er, at censorerne synes, at den nye ordning fungerer. Trods en række praktiske vanskeligheder, samt at nogle savner den gamle ordning.

357 censorer svarede på spørgsmål om, hvordan de synes den nye ordning fungerer alt i alt. 50% af dem svarede, at den nye ordning fungerer "godt" eller "særdeles godt", mens ca. 30% svarede at den fungerer "nogenlunde".

349 censorer, der også har været bedømmere under den gamle ordning, svarede på spørgsmål om, hvordan de vurderer deres bedømmelsers sikkerhed i den nye ordning i forhold til den gamle. 69% af dem vurderede, at de i den nye ordning havde givet mere sikre bedømmelser end før.

Et stort antal censorer (335) rapporterede, at de brugte de vejledende karakterbeskrivelser.

Censorernes tilbagemeldinger viser, at den tilstræbte anonymisering af elevernes besvarelser gennem brug af uni-login ikke fungerer. 359 censorer (det er 84% af de adspurgte) har svaret på, om de på grundlag af uni-login kan gennemskue elevens køn i mere end hver fjerde af de rettede besvarelser. Det svarer 74% af dem "ja" til.

Censorerne blev også spurgt, om de kan gennemskue elevens etnicitet. 74% af dem svarer "ja" til, at de ved mere end hver fjerde besvarelse kan se, at der er tale om en etnisk dansk elev. 42% af dem svarer "ja" til, at de ved mere end hver fjerde besvarelse kan se, at der er tale om en elev med anden etnisk baggrund end dansk.

Censorerne fremhæver følgende som positivt ved den nye ordning:

- Bedømmelsesproceduren er mere effektiv blandt andet fordi censor ikke skal vente på sin medcensor og fordi der ikke skal bruges tid på kommunikation med denne

- Bedømmelsen er mere sikker, fordi censor har mere ro og tid til bedømmelsen og kan være mere systematisk, fordi rettearbejdet ikke skal koordineres med en eller flere medcensorer.
- Bedømmelsen er mere objektiv, fordi censor ikke er udsat for pres fra klassens lærer.

Censorernes bekymringer i forbindelse med den nye ordning omhandler:

- Elevernes manglende retssikkerhed, herunder faren for at (uundgåelige) fejl i forbindelse med bedømmelsesarbejdet ikke bliver opdaget, når der kun er én bedømmer
- Uklarhed på skolerne omkring klageprocedurer
- Faglige konsekvenser af at censor ikke mere giver fagligt feedback til klassens lærer
- Uklarhed omkring praktiske forhold som deadlines, kopiering og forsendelse af opgaver.

Det hører også med til det generelle billede, at en del censorer er glade for muligheden for at sparre gennem sparregrupper, også selv om de ikke i 2016 gjorde brug af muligheden. Men forslaget om et fast tidspunkt til sparring ser ikke ud til at være en god løsning. Mange efterlyser en mere fleksibel løsning, eventuelt med mulighed for at mødes fysisk med en rettegruppe.

(Se også Bilag 3)

Udvalgte censorgruppers svar på en række spørgsmål i forbindelse med de skriftlige prøver i 2017:

I forbindelse med bedømmelse af 9. klasses prøver 2017 i skriftlig engelsk, dansk og matematik besvarede 16 udvalgte (”håndholdte”) censorgrupper en række spørgsmål vedrørende mødeformen og mødernes indflydelse på gruppens sparring og bedømmelse. Nogle af grupperne blev desuden bedt om at komme med forslag til, hvordan man kan bruge censors indsigt i klassens resultater som feedback til læreren om klassens niveau og resultater, og dermed som formativ feedback til lærerens undervisning.

Også denne undersøgelse bekræfter, at censorernes største bekymring i forbindelse med den nye ordning ikke er usikkerhed i bedømmelsen. Censorernes største bekymring er, at læreren ikke mere modtager feedback på sin undervisning gennem en diskussion med censor på grundlag af klassens besvarelser. Dermed går nyttig viden tabt.

Undersøgelsen havde to formål:

1.: Otte grupper blev bedt om at mødes før bedømmelsen gik i gang for at diskutere deres vurdering af tre udvalgte opgaver, men på grund af en misforståelse endte ni grupper med at mødes virtuelt.

Formålet var at forbedre grundlaget for deres senere bedømmelse. Syv grupper vurderer at et sådant møde har en positiv effekt på sikkerheden af deres bedømmelse. En gruppe finder at mødet er uden denne effekt. Tanken var, at STUK hurtigt efter prøven skulle sende udvælge og tre opgaver og sende dem til gruppernes medlemmer inden mødet. Det viste sig dog ikke at være muligt. I nogle grupper valgte formanden selv tre opgaver og sendte dem rundt til gruppens medlemmer. Andre grupper holdt mødet efter at de var gået i gang med retteprocessen.

Otte grupper finder, at et virtuelt møde har en positiv effekt på sikkerheden af deres bedømmelse. Grupperne fremhæver, at møderne har en social funktion (”man lærer hinanden at kende).

Otte grupper fremhæver fordelene ved at mødes virtuelt i forbindelse bedømmelsesprocessen. Frem for alt, at er det logistisk nemmere for en gruppe at mødes virtuelt frem for fysisk, og at denne mødeform sparer tid. Nogle grupper (5) havde tekniske udfordringer med at få den virtuelle mødeform til at fungere. Et flertal af grupperne (6) brugte Skype.

2.: Otte andre grupper blev bedt om at mødes efter at bedømmelsen var overstået. Formålet var at få mere information om, hvordan gruppen fungerede i forbindelse med bedømmelsen. En af disse grupper deltog ikke i undersøgelsen, som derfor omfatter syv grupper.

Ingen af grupperne havde benyttet sig af muligheden for at sparre. Fem grupper mener, at det vil være en forbedring, at lade gruppen mødes fysisk før bedømmelsen går i gang.

Grupperne peger ret enstemmigt på, at problemer som ombytning af besvarelser, fejl i forbindelse med afholdelse af prøven, snyd (herunder kopiering fra internettet) etc. ikke kan håndteres i forbindelse med censorgrupperne. Rutiner og instruktioner for hvordan denne type af fejl håndteres af censorerne bør derfor præciseres.

Nogle af grupperne blev bedt om at formulere forslag til hvordan en tilbagemelding fra censor til klassens lærere kan gøres præcis og overkommelig. De foreslå, at der indføres en rutine, hvor STUK hvert år (og for hvert fag) udarbejder et skema, som censor skal udfylde på grundlag af klassens besvarelser og derefter sender til læreren. Skemaet kan med fordel gøres elektronisk. Det bør eventuelt gøres muligt for læreren at ringe tilbage til censor, hvis læreren har opklarende spørgsmål til skemaet.

En gruppe henviser til at forsøg som Københavns Kommune forsøg, i samarbejde med ministeriet laver for at sikre, at den viden, censor får om en klasses læring, kommer tilbage til læreren og lærerens fagteam.

(Se også Bilag 4)

Litteratur

Bloxham, S. (2015). *The multiple limitations of assessment criteria*. Transforming Assessment Webinar Series, 4. November 2015.

Hanna, R. N. & L. L. Linden (2012): Discrimination in grading, *American Economic Journal: Economic Policy*, 4(4): 146-168;

Hinnerich , B. T., Höglin, E. & M. Johannesson (2015): Discrimination against students with foreign backgrounds: evidence from grading in Swedish public high schools, *Education Economics* , Volume 23, Issue 6: 660-676.

Lavy, V. (2008). Do gender stereotypes reduce girls' human capital out-comes? Evidence from a natural experiment. *Journal of Public Economics*, 92, 2083–2105;

McMillan (2013). Why we need Research on Classroom Assessment. In McMillan (ed.). *SAGE Handbook of Research on Classroom Assessment*. Los Angeles: SAGE.

Messick, S. (1989). Validity. In R. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–100). Washington, DC: American Council on Education.

Rangvid, B. S. (2018): Gender discrimination in exam grading? Double evidence from a grading reform and a field experiment; VIVE working paper x:2018

Tisi, J., Whitehouse, G., Maughan S. and Burdett, N. (2013). *A Review of Literature on Marking Reliability Research (Report for Ofqual).* Slough: NFER.

Bilag med tekniske rapporter

Bilag 1. Analyse af karakterdata fra FP-2016 vedr. pålidelighed

Bilag 2. Analyser vedr. effekt af censors kendskab til køn og oprindelse ved folkeskolens skriftlige prøver

Bilag 3. Delrapport 1 fra følgegruppen i forbindelse med overgang til én bedømmer ved de skriftlige prøver, februar 2017

Bilag 4. Delrapport 2 fra følgegruppen i forbindelse med overgang til én bedømmer ved de skriftlige prøver, oktober 2017