

Sig det med sprogløblomster! - censorers kvalitetsnormer og positioneringer ved bedømmelsen af afgangseksamen i 'Dansk, skriftlig fremstilling'

*Solveig Troelsen**

VIA University College

Bedømmelsen af folkeskolens afgangseksamen i skriftlig dansk har de seneste år fået mærkbart større betydning og finder sted i en foranderlig og ustabil kontekst. Simultant med indførelsen af en enebedømmerordning er eksamenskaraktererne i dansk og matematik generelt blevet afgørende for elevernes ret til ungdomsuddannelse. Og for prøven i skriftlig fremstilling bidrager flertydighed i skriveordre og vurderingskriterier til en ustabil kontekst. Artiklen anlægger et sociokulturelt perspektiv på bedømmernes arbejde og undersøger med deltagelse af seks beskikkede censorer bedømmelsen som semiocyt-hændelse i kontekst med særligt fokus på de kvalitetsnormer som danner grundlag for eksamenskaraktererne. I et design bestående af individuelle think-aloud-bedømmelser og fokusgruppesamtaler indkredses censorernes kvalitetsnormer, selvopfattelse og opfattelse af elevskriverne, dels situeret, dels på afstand i et reflektiv dialogisk rum. Fundene viser at skriveordre og vurderingskriterier forventes adresseret i elevteksterne, og at der samtidig findes kvalitetsnormer indlejret i tekstkulturen som tilskrives afgørende betydning i bedømmelsen, men ikke formidles eksplicit til elever og lærere. Der er tegn på et skrivefagligt fortolkningsfællesskab blandt censorerne og samtidig individuelle variationer, og elevteksterne tilgås som metonymiske repræsentationer af elevskriverne. Studiet bestyrker en tidligere undersøgelses kritik af at eksamenskaraktererne trods svag reliabilitet anvendes til elevsortering, og kalder på yderligere forskning i bedømmelse i kontekst samt forskningsinformeret revision af den gældende bedømmelsesordning.

Nøgleord: Solobedømmelse, tekstnormer, fortolkningsfællesskaber, censoridentitet, bedømmervariation, high-stakes-kontekst

* Corresponding author: SOTR@VIA.DK

Introduktion – bedømmelse i en skærpet kontekst

Når censorer ved afgangseksamen i *Dansk, skriftlig fremstilling* bedømmer niendeklasses elevernes tekster og manifesterer deres vurdering indeksikalsk i en karakter, har de magt og ansvar i en grad som på få år er øget dramatisk. Eksamen i skriftlig fremstilling kendetegnes aktuelt ved forandrede rammer og et materiale præget af gennemgribende flertydighed (Troelsen, 2018) som igen afføder stor mangfoldighed og kompleksitet i elevernes besvarelser (Troelsen, 2020a). Karakterer læses i uddannelsessystemet og samfundet i bredere forstand som indikationer på elevs og studerendes faglige kompetencer, og grundskolens eksamenskarakterer har fået afgørende betydning for elevernes muligheder. Med eksamensskolens afskaffelse i 1958 (Indenrigsministeriet, 1958) indførtes i skoleloven betegnelsen ”prøve” frem for ”eksamen” for at understrege at grundskolen var for alle og ikke skulle bestås. Det var derfor et – relativt upåagtet – paradigmeskift i dansk skolepolitik da en erhvervsskolereform i 2015 og en gymnasiereform i 2019 transformerede folkeskolens rituelle prøver til en *high-stakes-eksamen*¹ – mellem grundskole- og ungdomsuddannelsesniveau. 15-16-årige mister herefter deres retskrav på fortsat uddannelse – erhvervsuddannelse såvel som gymnasial uddannelse – og dermed deltagelse i arbejds- og samfundsliv hvis de ikke består afgangseksamen i dansk og matematik. Samtidig med denne skærpelse – i 2016 – ændredes censorordningen så elevernes dansklærer ikke længere fungerer som eksaminator og danner bedømmerpar med en ministerielt beskikket censor. Én censor er nu alene om bedømmelsen og kan suverænt fastsætte karakteren for en eksamensbesvarelse og afgøre om eleven forkastes eller anerkendes som værdig til fortsat uddannelse. Dette skal ses i lyset af at skriveprøver generelt er forbundet med en lav bedømmerreliabilitet (fx Berge, 2005; Cooper & Odell, 1977; Moss, 1994) som mindskes yderligere jo færre bedømmere der er. I den konkrete kontekst blev dette bekræftet da en følgegruppe efter to år med ordningen med én bedømmer konstaterede en gennemsnitskorrelation på 0.55 svarende til omtrent 40 % sandsynlighed for at censorer bedømmer ens (Dolin, Nielsen, & Rangvid, 2018).² Censor har altså suveræn magt til at forkaste eller anerkende elevernes tekster i større eller mindre grad og agerer samtidig på vilkår der må betegnes som usikre. Eller som følgegruppen udtrykker det:

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 elevs prøveresultater. (Dolin et al., 2018, s. 17)

Ikke desto mindre er elevselektion ét blandt flere formål karaktererne tjener i praksis. Det lille, indeksikalske tegn anvendes i multiple sammenhænge og tilskrives betydning alt efter aktører og situation. Karakteren indikerer eksempelvis uddannelsesparathed og faglig kompetence når institutioner, politikere, forskere og tænketanke sorterer og grupperer elever. Og set fra elevens synsvinkel i et lidt bredere perspektiv repræsenterer den desuden graden af anerkendelse eller forkastelse – på ét niveau af elevens specifikke kompetencer, på et andet af eleven som egnet og værdig deltager.

¹ Betegnelsen er udbredt i amerikansk testkultur og stammer fra 1920'ernes gamblemiljø hvor en indsats enten kan belønnes med betydelig gevinst eller medføre tilsvarende tab.

² Korrelationen er beregnet ved at lade 30 beskikkede censorer (ca. hver femte) og en ”ekspertretter” udpeget af fagkonsulenterne kontrolrette 5 udvalgte elevtekster (se Dolin et. al., 2018, p. 11 f.).

Reliabilitetsproblematikken er ikke i forgrund i dette studie, men én blandt flere kontekstuelle betingelser som virker ind på situationen. Forskningsinteressen retter sig mod at eksplicite og forstå hvilken betydning censorer selv tilskriver de karakterer de giver til afgangsprøven som svar på elevernes tekstligt medierede besvarelse af skriveordren, og operationaliseres i forskningsspørgsmålet: *Hvilke kvalitetsnormer danner grundlag for danskens censorers bedømmelse og udmøntede karakterer ved afgangseksamen i skriftlig fremstilling? Herunder: Hvordan forstår censorer deres egen rolle i bedømmelsen? Og hvordan opfatter de elevskriverne?*

Berge (2002) argumenterer for at bedømmelsesstudier bør have større fokus på sammenhængen mellem kvalitative træk ved prøvetekster og censorernes bedømmelser, og citerer Huot for at appellere til et øget fokus på bedømmerne. Forskningen har tendens til at søge svaret på bedømmervariation i elevteksterne (fx Skar & Aasen, 2016, der spørger til bedømmervariationen og svarer med en elevtekstanalyse). Dette studie tilstræber alternativt hertil at spørge direkte til de kvalitetsnormer censorerne møder elevteksterne med og enten finder honoreret eller ikke. Studiets dobbelte afsæt er dels de ovenfor beskrevne empiriske betingelser for eksamen med skærpede vilkår for censorarbejdet, svag reliabilitet og forandringer i tekstkulturen, dels eksisterende forskning. Praksis er evalueret af en ministeriel følge-gruppe med fokus på reliabilitet, effekten af elevanonymisering og censorernes opfattelse af ordningen (Dolin et al., 2018) og VIVE³ med fokus på lederes, læreres og censorers oplevelse af ordningen (Flarup, Bro, & Arendt, 2019). Censorerne er positive og betragter bedømmelsen som mere ”sikker”, ”effektiv” og ”objektiv” blandt andet i kraft af nye vejledende karakterbeskrivelser med baggrund i vurderingskriterierne (STUK, 2016), og de oplever afskaffelsen af forhandlingssituationen som positiv – en mulighed for at fastholde fokus (Dolin et al., 2018). Både lærere, skoleledere og censorer problematiserer dog at lærerne, fordi de ikke læser deres egne elevers besvarelser, går glip af en vigtig evaluering mulighed; noget begge rapporter vinkler som et behov for ”feedback” til lærerne. Lærerne og skolelederne er – modsat censorerne – stærkt bekymrede for ”rettepålideligheden” (Flarup et al., 2019, s. 28).

Forskningsmæssig fundering

Forskningsmæssigt funderes studiet på den ene side i den stadig ekspanderende viden om bedømmelse af elevs skriftsprog, internationalt især repræsenteret ved fx tidsskriftet *Assessing Writing* som har bidraget til at definere og konsolidere feltet,⁴ og i Skandinavien udfoldet i svensk og norsk forskning i skriveprøver og bedømmelse. På den anden side ses et påfaldende fravær af lignende forskning i dansk kontekst der kun tæller to ministerielt initierede evalueringsskemaer af gældende eksamensordning (Dolin et al., 2018; Flarup et al., 2019). Nærværende studie har den dobbelte ambition både at lægge sig i forlængelse af og bidrage til denne forskning og samtidig initiere og invitere til belysning af feltet i en dansk kontekst og bidrage med empiriske fund. I det følgende tages afsæt i forskningsspørgsmålet når en gennemgang af fortrinsvis skandinavisk forskning tegner studiets basis og afsæt – med fokus på især censorers tekstnormer og doxa og deres selvforståelse og positionering i en kompleks kontekst. Hovedvægten lægges på kvalitative studier, og da forskningsinteressen først og fremmest er at forstå censorernes tekstnormer og doxa, inddrages psykometrisk forskning

³ Det Nationale Forsknings- og Analysecenter for Velfærd.

⁴ For en gennemgang heraf se fx Borgström (2014) eller tidsskriftets vol. 42, *Special Issue: Framing the Future of Writing Assessment* som gør status og ser fremad på baggrund af de første 25 år (Slomp & East, 2019).

kun perifert. Indledningsvis vil fænomenet bedømmervariation blive berørt som et essentielt vilkår i den aktuelle kontekst.

En række studier i bedømmelse af skriveprøver adresserer den – sammenlignet med andre typer af test – store bedømmervariation som bredt anerkendes som et vilkår (Berge, 1996, 2005; Borgström & Ledin, 2014; Skar & Jølle, 2017). Forsøg på at styrke bedømmerreliabiliteten gennem præcisering af vurderingskriterier og instruktion af censorer har vist sig at have begrænset effekt (Jølle & Skar, 2018), ligesom kollegiale vurderingssamtaler kun delvist influerer den enkelte bedømmers karakterer (Blomqvist, 2018b). Skriveprøvers reliabilitet er således generelt udfordret, men kan forbedres gennem forskellige tiltag. For det første kan det helt overordnet gøres *statistisk* ved at øge antallet af bedømmere. Björnsson definerer den ”rigtige karakter” (rätta betyget) som gennemsnitskarakteren ”när ett inte alltför litet antal lärare bedömer en uppsats” (Björnsson, 1960, s. 77). Definitionen bygger på at reliabiliteten øges proportionalt med antallet af bedømmere idet den principielt ”rigtige karakter” udgøres af gennemsnittet for hundrede uafhængige bedømmere. Der er individuelle forskelle i censorers tekstnormer og doksiske forståelser af tekstkvalitet, og samme karakter kan dække over forskelle i rangering, konceptioner af karakterskala og vurderingskriterier samt forskellige kvalitetsnormer (Skar & Jølle, 2017). Ved flere bedømmere⁵ udviskes den lave korrelation mellem bedømmere, og udsving som for enkeltelever kan have stor betydning, kamufleres. For det andet indikerer forskning at bedømmelse kan kvalificeres og reliabiliteten øges gennem udvikling af *fortolkningsfællesskaber*. Projektet *Kvalitets sikring av læringsutbyttet i norske skriftlig* (KAL) undersøger afgangsprøverne i norsk skriftlig (Berge, Evensen, Hertzberg, & Vagle, 2005) – herunder vurderingen som belyses såvel kvantitativt som kvalitativt – og finder en relativt høj reliabilitet sammenlignet med internationale undersøgelser (Berge, 2005). Det forklarer forskerne med tilstedeværelsen af en samtalekultur blandt norsklærerne: ”I denne kulturen udvikles det en refleksjon omkring læring og tekstkvalitet som resulterer i et betydelig tolkningsfællesskap” (Evensen, 2005, s. 206) og en deraf følgende relativ samstemmighed. Den enkelte censor udmønter sin vurdering i en ”forhandlingskarakter” (Berge, 1993, s. 157) som danner udgangspunkt for ”forhandlingsarbeid med kollektive trekk” (Berge, 2005, s. 103) der munder ud i den endelige karakter. Dermed opnås ikke bare et statistisk korrektiv via flere bedømmere, men også en substantiel kvalificering af bedømmelsen gennem argumentation og forhandling blandt deltagere i et fortolkningsfællesskab. I den danske ordning med én bedømmer findes hverken den statistiske validering eller det fortolknings(praksis)fællesskab som manifesteres, udvikles og konsolideres i dialogen mellem deltagere. Og jeg antager at dette – omvendt parallelt med den norske undersøgelse – er medvirkende årsag når følgeforskning til ordningen konstaterer en meget lav reliabilitet og øget risiko for deciderede fejlbedømmelser (Dolin et al., 2018).

Undersøgelser af bedømmervariation har med forskelligt fokus forsøgt at forklare den særlige kompleksitet som knytter sig til bedømmelser af skriveprøver i L1-faget/modersmålsfaget. Ifølge Borgström og Ledin balancerer bedømmelsen i et spændingsfelt mellem *teknisk og hermeneutisk rationalitet* (Borgström & Ledin, 2014). Teknisk rationalitet knytter sig til bestræbelsen på transparens og forudsigelighed med fx udfoldede vurderingskriterier, mens hermeneutisk rationalitet angår censors professionelle, erfaringsbårne og holistiske dømmekraft og skrevne teksters kompleksitet. Blomqvist finder tilsvarende at bedømmere

⁵ Björnsson (1960) foreslår fem som med 95 % sandsynlighed vil nå samme resultat som de 100.

forholder sig enten *nonkompensatorisk* eller *kompensatorisk* til specifikke vurderingskriterier (Blomqvist, 2018a) – i begge tilfælde med fare for at elevteksten som helhed overses. Forstået sådan kan ikke bare skrivebedømmelse generelt, men også konkrete censorers bedømmelser betragtes som positioneringer i et kontinuum mellem det tekniske og det hermeneutiske. I dette spændingsfelt foreslår Borgström og Ledin en balanceret mellemposition som de kalder *tekstkulturel rationalitet* – funderet i såvel klart beskrevne faglige kriterier som et skrivefagligt fortolkningsfællesskab. I praksis er dette forsøgt realiseret i det norske NORM-projekt⁶ som netop tilstræber at imødekomme den nævnte kompleksitet ved at tage afsæt i lærernes fortolkningsfællesskab i udviklingen af vurderingskriterier og standarder til brug for nationale skriveprøver (Evensen, Berge, Thygesen, Matre, & Solheim, 2016).

Også mere individuelle rationaler har indflydelse på bedømmelsen. Jølle og Skar (2018) viser at professionelle bedømmere selv efter års træning i at bedømme efter NORM-projektets vurderingskriterier baserer deres handlinger på forskellige rationaler. Et væsentligt fund i undersøgelsen er at bedømmervariationen ikke bare skyldes de førnævnte generiske problemer for bedømmelse af skriveprøver eller forskelle i lokale kontekster, men hænger sammen med censorers holdninger til og forståelser af formålet med skriveprøverne og deres positionering og selvforståelse som bedømmere (ibid.).

Forskning i bedømmelsespraksis omhandler enten lærere (fx Blomqvist, 2018a, 2018b; Borgström & Ledin, 2014; Evensen et al., 2016; Skar, 2013) eller professionelle bedømmere (fx Berge, 1996; Jølle & Skar, 2018). Reelt drejer det sig om forskellige kontekster og positioneringer for de samme personer idet professionelle bedømmere oftest er lærere med et særligt hverv, ligesom lærere reelt fungerer som professionelle bedømmere når deres slutevaluering fx får high-stakes-status.⁷ Simultant med testforskningens fremvækst vokser også interessen for *teacher assessment literacy* forstået som læreres kompetence til at analysere og anvende såvel formative som summative test, forholde sig kritisk til testformål og agere i overensstemmelse hermed (Xu & Brown, 2016). I en taksonomisk pyramidemodel (ibid., s. 155) placerer Xu og Brown positioneringen *Teacher as assessor* øverst – på et fundament af viden om test og erfaring med medkonstruktion, deltagelse og refleksion. Om det primære formål med en bedømmelsesaktivitet er undervisning, og altså formativt, eller statlig kontrol, og dermed summativt evaluerende, har betydning for bedømmerens opgave og fokus i bedømmelsen (Jølle & Skar, 2018).

Teoretisk ramme – bedømmelse i kontekst

Med afsæt i den ovenfor beskrevne kontekst anskues bedømmelser – indeksikalsk markeret ved karakterer – i den kulturelle, institutionelle og situationelle kontekst hvori de findes, fungerer og betyder som svar på elevernes tekster (Troelsen, 2020a) der selv besvarer skriveordren (Troelsen, 2018). Denne artikel interesserer sig især for den betydningstilskrivning som sker når censorer giver deres vurdering form som tekst (karakter), de tekstnormer og faglige diskurser som er på spil, og bedømmelsen som social handling.

⁶ Developing national standards for the assessment of writing – a tool for teaching and learning.

⁷ Som i svensk gymnasium (Borgström & Ledin, 2014).

At læse som censor

Hvad censorer ser som kvalitet i elevtekster, afhænger af læsemåde (Skar & Aasen, 2016) og ”øynene som ser” (Nilsen, 2005, s. 158). I tråd med dette studie fastslår Igland med Bakhtin at lærere altid læser elevtekster vurderende som forberedelse til et svar (Igland, 2013), og påpeger ”lesingens fundamentale rolle i all vurdering av tekst” (ibid, 2013, s. 137). Den vurderende læsning trækker på mere end formulerede vurderingskriterier: ”Uansett hvor vel fundert et sett av vurderingskriterier er, er det ikke et like smidig vurderingsredskap som *den dyktige elevtekstkjennerens situasjonsavhengige, praktiske og tverrfaglige synteser av relevant kunnskap og skjønn*” (ibid., s. 151 – min fremhævning). Læreres læsning af elevtekster bygger på hvad Eisner betegner ”educational connoisseurship” der favner de indbyrdes forbundne roller *kender* og *kritiker*: ”kjennerens verdsetting [...] forsyner kritikken med faglig substans” (Igland, 2013, s. 143). Censorer betragtes i dette studie som sådanne *kendere* eller *ekspertlæsere* (Shanahan & Shanahan, 2012), og *kenderskab* er tæt forbundet med *tekstnormer* og *doxa* i den feltspecifikke tekstkultur (Berge, 1996). Tekstnormer forstås som ”a cluster of socially developed criteria defining which qualities can be expected from an utterance uttered in a specific situation in a certain culture” (Berge, 2002, s. 459). De kendetegnes ved at være intersubjektive, dynamiske og implicite og udvikles og fungerer i tolkningsfællesskaber som censorerne får adgang til via deltagelse: ”Lærere som utgjør et tolkningsfællesskap med felles vurderings-skjønn, vet hvilke yringer som gjelder som tekster i skolen, og hvilke yringer som ikke gjør det” (Berge, 2009, s. 48). Og mens nogle normer altså er eksplicit direktiviske, er andre fundet i doxa – i deltagernes praksisbaserede hverdagsmeninger, almindelige antagelser og formentligheder (Berge, 1996). Det er disse normer studiet tilstræber at eksplicitere og fremanalysere i dels den situerede bedømmelsespraksis, dels fokusgruppernes dialogiske rum. For overskuelighedens skyld samles de her i termen *kvalitetsnormer* der ikke skelner mellem formulerede og tavse normer.

At forstå og positionere sig som bedømmer

Med Borgström og Ledin forstås bedømmelse som en aktivitet der forudsætter censors positionering i et *kontinuum mellem teknisk og hermeneutisk rationalitet*. Denne positionering hænger sammen med censorernes forståelse af formålet med prøven og af deres egen rolle som bedømmere (Skar & Jølle, 2017) og viser sig blandt andet i deres forholdemåde over for elevtekster, skriveordre, karakterskala, bedømmelseskriterier og den elevskriver som de møder gennem dennes tekst. Midt i dette kontinuum af positioneringsmuligheder installerer Borgström og Ledin en balanceret *tekstkulturel rationalitet* som både trækker på formelle kriterier og på det mere eller mindre tavse fortolkningsfællesskab som bedømmerne tager del i. Her mindskes faren for at elevteksten forsvinder i enten instrumentel insisteren på del-elementer i skriveordre og vurderingskriterier eller subjektive fortolkningsprocesser.

Den indskrevne elev

Validiteten i enhver test, men i potenseret grad i skriveprøver (Berge, 2005), udfordres af brugen af ”limited samples of observation to draw general and abstract conclusions about persons” (Kane, 2006 s. 17). Det er dette studies påstand at relationen mellem elevens tekst og den generelle vurdering i lyset af high-stakes-konstruktionen antager form som en *metonymi* hvor teksten bliver tegn for eleven og dennes egnethed til videre uddannelse og deltagelse i arbejds- og samfundsliv. Dermed bliver elevens skrivehandling en *skriftlig selv-fremstilling*

med henblik på at fremstå anerkendelsesværdig i censors øjne, og det iboende identitetsaspekt i det at skrive (Christensen, Elf, & Krogh, 2014) accentueres. Ivanič bruger begrebet *skrivelse* som hun forstår med fem forskellige perspektiver (Ivanič, 2012). Mest relevant for dette studie er *det opfattede skrivelse* som læseren møder gennem ”the impression they [skriverne] convey of themselves through their writing” (ibid. s. 17). At censorerne ikke bare bedømmer tekst, men også elevskrivere, viser sig eksempelvis i hvad de infererer om elevens køn,⁸ og i ”honnørkommentarene ’kreativitet’, ’engasjement’ og ’modenhed’” (Nilsen, 2005, s. 136) som KAL-projektets deltagere noterer i forbindelse med bedømmelsen.

Tekstnormer er kontekstbundne og kan opdeles i produktionsnormer og receptionsnormer (Berge). I et tidligere studie (Troelsen, 2020a) blev elevbesvarelsenerne analyseret som udtryk for elevernes produktionsnormer, og her fandtes en relativt uklar sammenhæng mellem disse og eksamenskarakteren.⁹ I forlængelse heraf interesserer dette studie sig for *censorerens receptionsnormer* som i TA-bedømmelserne viser sig *situeret* mens deltagerne interagerer med elevbesvarelsener, formelle kriterier og karakterskala i en simuleret praksis, og i fokusgrupperne gøres til genstand for mere *generel refleksion og dialog* mellem flere censorer.

Undersøgelingsdesign og -metode

Forskningsspørgsmålet undersøges via empirisk triangulering bestående af 1) ”think-aloud” (herefter TA) bedømmelser med observationer og bedømmelsesark, 2) fokusgruppesamtaler¹⁰ samt 3) karakterer. For at en normkultur kan undersøges, må den rekonstrueres. Det gøres i to trin med deltagelse af seks beskikkede censorer: Først individuelle TA-bedømmelser som videofilmes og transskriberes hurtigt efter udførelsen hvorefter en foreløbig analyse foretages med henblik på at udtrække temaer til to (ligeledes filmende og transskriberede) fokusgrupper (Halkier, 2016) med hver tre af censorerne. Hensigten med totrinsmodellen er at få indsigt i kvalitetsnormer både i det individuelle situerede og det fælles refleksive. Dette begrundes i studiets dobbelte interesse for dels den aktuelle solobedømmelse, dels et formodet bagvedliggende (traditionsforankret) fortolkningsfællesskab blandt censorerne. Det antages at den tidlige afstand mellem de to hændelser bidrager til censorernes refleksive distance til egen praksis, og at foreløbige analyser af TA-bedømmelserne kan kvalificere fokusgruppernes rammesætning. Endelig inddrages de oprindelige eksamenskarakterer samt karaktererne fra TA-bedømmelserne i det omfang de bidrager til belysning af tekstnormer.

Deltagerrekruttering

Et studie af denne karakter afhænger i høj grad af at deltagerne selv ser værdien af undersøgelsen. I det perspektiv har det været en gave for dette studie at Styrelsen for Undervisning og Kvalitet i første omgang misforstod min forespørgsel om adgang til listen over landets 277 beskikkede censorer (mail fra STUK, 16.11.2018) så jeg kunne udvælge en deltagergruppe med en vis variation med hensyn til køn, lærer- og censorerfaring samt geografisk tilhørsforhold. I stedet annoncerede ministeriet efter projektdeltagere hvorefter jeg modtog ti henvendelser – alle fra entusiastiske og dedikerede censorer. Engagemnet var således

⁸ Elevteksterne er anonymiseret for censor.

⁹ Meget omhyggelige og ud fra en bredere forståelse kompetente besvarelsener modtog eksempelvis kun middelskarakterer.

¹⁰ I det følgende betegnes disse samtaler blot ”fokusgrupper” uagtet at der ikke henvises til gruppen, men gruppesamtalen, jævnfør Halkier (2016).

givet på forhånd, og jeg kunne udvælge seks deltagere som i videst muligt omfang tilgodeser ønsket om diversitet.¹¹ Prisen for denne nødløsning var naturligvis en begrænset repræsentativitet og en mulighed for at studiets deltagere som gruppe potentielt kan adskille sig fra baggrundsgruppen netop i kraft af deres entusiasme. Som analysen vil vise, tyder resultaterne af studiet imidlertid på at gruppen på mange måder svarer til hvad tidligere censorundersøgelser har vist.

TA-bedømmelse som literacy-hændelse

Som andre literacy-hændelser er bedømmelse situeret, og det må derfor antages at censorers viden, forståelser og holdninger af betydning for bedømmelsen bedst studeres ved at genskabe situationen – i det omfang dette er muligt. Det indebærer her at fastholde at censor arbejder alene i bedømmelsessituationen, og tilgængeliggøre samme resurser som i den oprindelige bedømmelsessituation – eksempelvis bedømmelsesarket med vurderingskriterier (se bilag 1). Andre kontekstelementer er imidlertid umulige at genskabe. Det gælder fx situationens alvor. I bestræbelsen på at få adgang til censorernes kvalitetsnormer anvendes en såkaldt 'rekonstruerende empirisk-eksplikativ' metode (Berge, 1996; Matre et al., 2011). Analytisk forudsætter rekonstruktion af kvalitetsnormer forekomsten af indekser herpå som så kan fortolkes abduktivt. Det er altså nødvendigt at opsætte et scenarie hvor sådanne indekser bliver til. Det har forskningshistorisk været gjort på forskellig vis via fx en kombination af spørgeskemaer, dokumentanalyser og dybdeinterviews (Berge, 1996), think-aloud-inspirede interviews (Matre et al., 2011) og forskellige former for 'live rating' (fx Skar & Jølle, 2017) og samtalebaseret bedømmelse af tekster (Blomqvist, 2018a, 2018b; Jølle, 2014). I dette studie anvendes Think Aloud Protocols (Ericsson, 1993) fra simulationer af den individuelle bedømmelsessituation. Metoden har potentiale til at explicitere hvad deltagerne *gør*, og ikke blot hvad de *siger de gør* (ibid.), fordi afrapportering af tanker og handlinger sker simultant med disse og i tilknytning til situationen.¹² Dette studie er ikke optaget af kognitive processer, men af mentale repræsentationer – 'mentefakter' – af kvalitetsnormer (Berge, 1996, med reference til Posner) sådan som de får betydning i bedømmelsespraksis, og adskiller sig således fra tidligere TA-bedømmelser. Blandt andet antages metoden at kunne gribe aspekter som censorerne ikke er sig helt bevidst i situationen – jævnfør forestillingen om at en ikke uvæsentlig del af normerne er forankret i doxa og dermed erfaringsbaseret og i nogen udstrækning intuitiv og (umiddelbart) tavs.

TA-bedømmelser har dog også begrænsninger. For det første er det ikke uproblematisk – epistemologisk set – at ville udforske mentale forestillinger via verbal repræsentation. Det man får ud af at studere TA-optagelser, er med andre ord ikke utvetydige indekser for tavs viden. Den må forskeren slutte sig til via abduktion. For det andet fordrer metoden en kunstig genskabelse af bedømmelsessituationen, og det er selvsagt ikke uproblematisk at

¹¹ At det netop blev seks deltagere, er delvist pragmatisk begrundet i at studiet indgår i et større forskningsprojekt hvis tidsramme sætter visse begrænsninger. Det blev dog tydeligt ved observationen af femte og sjette TA-bedømmelse at et mætningspunkt var nået og visse mønstre gentog sig (jævnfør fx Small, 2009).

¹² Til sammenligning vil et interview om hvad censorer tillægger betydning i bedømmelsesprocessen, bære præg af at afstanden i tid og rum i forhold til den situation der ønskes undersøgt, ikke blot svækker hukommelsen og evnen til at genkalde sig situationen, men også bevirker at deltagernes erindring "forurenes" af senere tilkomne refleksioner.

reestablere en situation der per definition er kontekstbunden. Forskellene mellem de to situationer fremgår af tabel 1.

	Sted	Arbejdsopgave	Situation	Modus
Autentisk bedømmelsessituation	Hjemme i vant miljø, typisk alene	Gennemsnitlig 240 elevbesvarelser pr. censor	High stakes-eksamen hvor karakteren har konsekvenser for elever	Tavs proces
TA-bedømmelser	SDU med forskeren som observatør	7 elevbesvarelser pr. censor udvalgt af forskeren	Rekonstruktion uden konsekvens for elever	Verbaliseret

Tabel 1. Oversigt over de to bedømmelsessituationer.

For det tredje udfordres TA-bedømmelser af publikumsbevidsthed ("audience awareness", Barkaoui, 2011) hos deltagerne og konsekvenser heraf for rapporteringen. Deltagere oplever typisk at de henvender sig til "nogen" (forskeren) selv om de sidder alene under optagelserne,¹³ og Smagorinsky citeres (ibid., 2011) for at TA-protokoller kan betragtes som én "tur" i en dialog. Med afsæt heri er det muligt at forstå optagelserne i et mere dialogisk perspektiv. Ved at bede deltagerne tænke højt har jeg bedt dem ytre sig, og som Bakhtin påpeger, er enhver ytring dialogisk (Bakhtin, 1986) – også TA-bedømmelser selv om højttænkningen i princippet er selvhenvendt. Informanterne ved godt forskeren lytter med! Frem for at betragte dette som "støj" i undersøgelsen forsøger jeg at omfavne det dialogiske – blandt andet ved selv at være til stede under optagelserne.

Fokusgrupper – et dialogisk refleksionsrum

De transskriberede TA-bedømmelser er efterfølgende analyseret ekstensivt for at identificere temaer til to fokusgrupper (Halkier, 2016): F1 og F2. Hensigten med disse er at bringe deltagerne individuelle forståelser i dialogisk interaktion for derved at få adgang til fortolkningsfællesskabet og grænserne herfor når tekstnormer diskuteres adskilt fra umiddelbare krav om handling (bedømmelse). Tidligere studier af bedømmelsessamtalen udforsker censorers *forhandling* af kvalitetskriterier situeret (fx Blomqvist, 2018b; Skar, 2013). I dette studie er målet et andet – nemlig dataproduktion på kollektivt niveau. Fokusgrupper er en "metode, hvor data produceres via gruppeinteraktion omkring et emne, som forskeren har bestemt", skriver Halkier (2016, s. 10) og fremhæver *kombinationen af gruppeinteraktion og forskerbestemt emnefokus* som fokusgruppens særlige kendetegn ligesom hun betoner netop denne kombination som "særligt velegnet til at producere empiriske data, der siger noget om *betydningsdannelse i grupper*" (ibid., min fremhævelse). Fokusgrupper vurderes derfor velegnede til at tilgå de – delvist tavse – kvalitetsnormer som er på spil i semiosfæren *vurdering af eksamenstekster*, ved at lade deltagerne samtale i konkrete fortolkningsfællesskaber. Til gengæld har fokusgrupper den svaghed at det atypiske formentlig vil være underrapporteret grundet gruppens sociale dynamikker. Det søges i forskningsdesignet opvejet af kombinationen af TA-bedømmelser og fokusgrupper som adresserer henholdsvis det individuelle og det fælles. Det todelte design komplementerer således forståelsen af de kvalitetsnormer som virker i censorernes arbejde, dels gennem deres praksisnære afrapportering i TA-bedømmelserne, dels gennem

¹³ Barkaoui (2011) finder i sin sammenligning af TA-bedømmelser og tavse bedømmelser at deltagerne i TA-bedømmelser – trods tydelig instruktion om at "tale med sig selv" i en monologisk form – reelt syntes at henvende sig til et publikum/en modtager og således fx filtrere det sagte efter dette.

samkonstruktion og forhandling af temaerne i fokusgrupperne. I tabel 2 nedenfor er studiets tre typer af data sat i forhold til hinanden og den funktion de har i analysen.

	Think-aloud-bedømmelse	Fokusgrupper	Karakterer
Ramme	Situert – praksissimulation	Adskilt fra praksis i tid og rum	To typer: Original eksamenskarakter fra 2017 samt karakterer fra TA-bedømmelse
Resurser	Vurderingskriterier Karakterbeskrivelser Eksamensvejledningen Vurderingsark	7 laminerede samtalekort med temaerne: Tekstkvalitet og tekstnormer, Skriveordre og kvalitet, Del og helhed, Tekstmængde og kvalitet, Sprog og kvalitet, Tekstkvalitet og bedømmelseskriterier samt Karakterer og kvalitet	12-skalaen
Relation	Individuel (i forhandling med fortolkningsfællesskabet)	Kollektiv (i forhandling med fortolkningsfællesskabet)	Individuelle og sammenholdt med andre censorer
Genstandsfelt	Udgangspunkt i konkret bedømmelsesarbejde med elevtekster og bedømmelseskriterier	Generelle refleksioner med udgangspunkt i forskerens tematiseringer	Karakteren som tekst/indeks for tekstkvalitet
Formål for deltagerne	Bedømmelse	Dialogisk eksploration og refleksion	Indeksering af bedømmelse
Form	Højtænkning	Samtale	Indeksikalske tegn
Formål for forskeren	Indblik i <i>tekstnormer i praksis</i> (herunder <i>doxa i praksis</i>) med henblik på at identificere temaer til fokusgruppeinterviews	Identifikation af indekser på, hvad der opfattes som tekstkvalitet. Dialogisk afsøgning af <i>fortolkningsfællesskab</i> (og grænser herfor), tekstnormer og <i>doxa</i>	Referenceramme for fokusgrupper samt afdækning af mønstre i karaktergivning
Forskerens rolle	Observatør	Moderator, facilitator	Fortolker, mønsterafdækker
Analysestrategi	Ekstensiv analyse med henblik på identifikation af temaer til brug for fokusgrupper.	Intensiv, hermeneutisk analyse med samspil af lukkede koder baseret på teori og tidligere forskning og åbne koder udledt af materialet.	Afdækning af mønstre, ligheder og forskelle i relation til de kvalitative data. Anvendes i fokusgruppesamtaler til facilitering af diskussion om karakterer
Status i analysen	Baggrund	Forgrund	Referencepunkt

Tabel 2. Skematisk oversigt over kombinationen TA-bedømmelser/fokusgrupper/karakterer

For at fremme adgangen til deltageres kvalitetsnormer (og ikke fx deres fortolkning af hvad jeg som forsker gerne vil høre) tages afsæt i deres praksisforankrede udsagn i TA-bedømmelserne: dels udsagn som der var stor enighed om, dels udsagn som undrede mig som forsker – nogle gange begge dele. Som moderator strukturerede jeg samtalen efter 8 samtalekort (bilag 2) med temaer fra TA-bedømmelserne – flere steder suppleret med direkte citater fra samme eller fra de elevtekster censorerne bedømte. Der er altså tale om relativt strukturerede fokusgrupper med en forholdsvis stram grad af moderation (Halkier, 2016)

begrundet i forskerens ønske om at komme dybere i de udvalgte temaer. Deltagerne er sammensat ud fra hvornår de praktisk kunne mødes, og kender ikke hinanden på forhånd.¹⁴ Dermed er de nødt til at ekspliciterer normer forholdsvis udfoldet, men man kan samtidig forestille sig at der vil være elementer som er indforståede qua deres fælles tilhørsforhold til censorkorpset. Moderators funktion er at sikre en uformel stemning, aktivitet fra deltagerne og fokus på emnet. I overensstemmelse med det socialkonstruktivistiske udgangspunkt lægges ud med en åben, afsøgende tilgang som gradvis fokuseres ind på mere konkrete, men stadig åbne, temaer afledt af TA-bedømmelserne. Der arbejdes altså med en let tragtfornet styring (Halkier, 2016) med henblik på at få deltagerne til at diskutere tekstkvalitet og tekstnormer åbent, men samtidig forankre diskussionen tematisk.

Analysestrategi

Det empiriske materiale er mangfoldigt og inviterer til fortsat forskning. I det aktuelle studie positionerer jeg mig som forsker tæt på censorerne og er primært interesseret i at få adgang til deres meningskonstruktioner i forbindelse med bedømmelsesarbejdet. Jeg er således ikke optaget af om eller i hvilken grad censorerne bedømmer pålideligt, men fx af hvordan de håndterer reliabilitetsspørgsmålet som én blandt flere kontekstuelle betingelser.¹⁵

Fokusgrupperne behandles samlet og danner forgrund i analysen operationaliseret gennem *kodning*, *kategorisering* og *konceptualisering* (Halkier, 2016)¹⁶ – praktisk udført med NVivo som lagrings- og sorteringsmedie for at facilitere en hermeneutisk analyse. De anvendte koder går på tværs af empirityperne således at TA-bedømmelserne inddrages til uddybning og nuancering. Første kodning er samtidig en identifikation af meningsenheder idet ikke alle ytringer betragtes som sådanne (Malterud, 2012). At identificere disse betragtes som en del af analysen (ibid.), og i det aktuelle tilfælde udelades ekskurer om andre prøveformer, undervisning og andet der ikke samtidig har relevans for deltageres kvalitetsnormer. Analytisk går processen fra den første, ekstensive temakondensering af TA-bedømmelserne over fokusgruppesamtalerne på baggrund heraf og tilbage til TA-bedømmelserne – en analysecirkel som gentages flere gange i dialog med forskningsspørgsmålene.¹⁷ *Kodningen* er dels *deduktiv*, drevet af tidligere forskning samt teori og temaer fra tidligere faser i projektet, dels *induktiv*,¹⁸ udledt af materialet. *Kategorier* er opstået undervejs i anden gennemlæsning og erstatter ofte umiddelbare koder – fx ved sammenlægning af induktive koder og eksplicite vurderingskriterier. Således er den induktive kode ”sproglig kvalitet og mangel på samme” sammenlagt med den deduktive kode for vurderingskriteriet ”sprog” – også hvor deltagerne ikke specifikt taler herom. Denne kategori er altså opstået i krydsfeltet mellem vurderingskriterierne og deltageres spontane udsagn om kvalitet. Andre kategorier er skabt via samling af underkoder på højere abstraktionsniveau eller omdøbning af koder. Den fremvoksende

¹⁴ Den første fokusgruppe havde kun to deltagere foruden moderator som følge af en sygdommelding på dagen fra censoren Hanne.

¹⁵ I min indledende korrespondance med censorerne har jeg flere gange understreget at jeg på baggrund af eksisterende forskning *forventer* en stor bedømmervariation, og at dette er et vilkår i den gældende ordning.

¹⁶ Halkier bruger ”begrebsliggørelse” som for mig (paradoksalt nok) virker endnu mere abstrakt end fremmedordet ”konceptualisering”.

¹⁷ For at sikre loyal gengivelse er alle citater krydstjekket med videooptagelserne grundet den uegale transskription fra fire forskellige studentermedhjælpere.

¹⁸ Men konceptualiseringen og fortolkningen abduktiv – jf. tidligere.

fortolkning er datadreven og afspejles i kategoriseringen. Mange passager har flere koder – eksempelvis for hvilken censor der ytrer sig, eller hvilken elevtekst der tales om, og samtidig for ét eller flere temaer. Endelig udkrystalliseres i fokusgrupperne formuleringer som flere censorer bifalder, eller som på anden vis fremtræder særligt betydningsfulde. De kategoriseres som ”nøgleformulering” – ofte sammen med andre koder. Den *abduktive fortolkning* og konceptualisering anvender *domænefortolkning* til belysning af syv temaer samt analyse af *metaforer* og *selvfølgeligheder* med henblik på at få indblik i deltagernes forståelser (Halkier, 2016). Metaforer betragtes som struktureringer af tanker og forestillinger der viser en kulturs værdier i kraft af hvad de henholdsvis fremviser og skjuler (Lakoff & Johnson, 2002). Selvfølgeligheder ses i metaforerne og fx i sætningsadverbialerne ”selvfølgelig” og ”jo”.¹⁹

Analyse

Nedenstående fremstilling er udformet som en tematisk kondensering og konceptualisering på tværs af de forskellige empiriformer. Temaerne er: 1) at parere skriveordre, 2) tekstkvalitet som noget mere, 3) kvantitet som kvalitet, 4) sproglblomster og litterære normkilder, 5) konstruktion af elevskrivere og 6) censoridentitet – positionering og selvopfattelse. De deltagende censorer er Berit, Carsten, Hanne (F1) og Morten, Rasmus og Vibeke (F2).²⁰ Temaerne præsenteres og eksemplificeres med udsagn fra fortrinsvis fokusgruppesamtalerne, suppleret med enkelte nedslag i TA-bedømmelserne. Nogle steder er ét repræsentativt udsagn anført, andre steder har jeg valgt at lade flere stemmer illustrere kompleksiteten i fortolkningsfællesskabet. Det gælder fx tema 2 hvor censorerne er fælles om at ønske sig ”noget mere” af den gode elevbesvarelse, men har forskellige forsøgsvise benævnelser for hvad dette ”mere” er.

1. At parere skriveordre

Det følger naturligt af eksamenskonstellationen at elevbesvarelserne initieres og rammesættes af den ministerielle skriveordre som danner referencepunkt for både elevskrivere og censorer og kendetegnes ved paradoksalitet og flertydighed som på én gang ordre og invitation (Troelsen, 2018). Et centralt samtalemne i fokusgrupperne er hvorvidt elevskriverne imødekommer kravene. Bedømmelsen er ifølge vejledningen en helhedsvurdering baseret på syv delkriterier (se bilag 1): ”Helhedsvurderingen foretages ved, at man sammenholder besvarelsen med de enkelte områder i vurderingskriterierne og efterfølgende afvejer og giver en samlet vurdering” (STUK, 2016, s. 22). Censorerne arbejder i TA-bedømmelserne mere eller mindre systematisk med vurderingskriterierne operationaliseret i et vurderingsark (bilag 1) som udfyldes undervejs.²¹ Kriteriet ”Opgavens krav” spiller en afgørende rolle i denne praksis.

Fokus er ofte på detailformuleringer i skriveordren og i mindre grad på overordnede forhold som eksempelvis genrekrav. Et eksempel til illustration er censorernes

¹⁹ Disse er fremhævet med fed skrift i empiricitaterne.

²⁰ Se også bilag 3. Hanne blev desværre syg til fokusgruppen og figurerer derfor kun med udsagn fra TA-bedømmelsen.

²¹ Med undtagelse af Morten der slet ikke anvender bedømmelsesarket.

højtænkning om en baggrundsartikel som de møder anonymiseret som ”Tekst 2”.²² Carsten adresserer²³ opgavekravenes opfyldelse allerede inden han læser besvarelsen:

... der står ’fotos’ i flertal, og der er ingen fotos her. (TA-C)

Han skimmer skriveordre og elevtekst og konkluderer på baggrund af enkeltelementer. Hanne refererer til samme formulering og samtidig til en diskussion på et censorkursus:²⁴

... denne her, den kan jeg huske at vi tidligere har diskuteret (...) illustrere den med fotos, faktabokse eller andre grafiske virkemidler. Ja, men så skal der være flere end én. (TA-H, min fremhævning)

Også andre hæfter sig ved flertalsformen og savner mere end tekstens ene faktaboks, og Rasmus googler genren ”baggrundsartikel” som han har svært ved at genkende i teksten. Flere bemærker fraværet af kildehenvisning og googler for at sikre sig mod snyd.²⁵ Generelt fylder skriveordrens detaljerede og samtidig flertydige krav (Troelsen, 2018) meget i vurderingen og viser sig i et stort tidsforbrug i bedømmelsessituationen og en *kontroldiskurs* blandt censorerne der ”tjekker” om alt ”er med”. På ét af samtalekortene om relationen mellem opgavekrav og tekstkvalitet figurerer en pizza-metafor fra Vibekes TA-bedømmelse, og i fokusgruppen beder moderator hende folde denne ud. Bemærk sætningsadverbialt ”jo” som markerer en selvfølgelighed:

Det vil jeg gerne. Øh, altså det der med gode og dårlige tekster... hvis jeg tænker på min rolle som censor, så er der i opgavebeskrivelsen bestilt en pizza, ikk? Den skal indeholde de og de ting. Og dét er der, synes jeg, blevet sådan en tendens hen over de sidste par år: Jo mere tydelig den der opgavebeskrivelse den fremgår, altså stilladset ikk’, jo mere synes jeg også det er det eleverne forsøger at besvare. Men hvis, og så kan man sige at så opfylder de **jo** egentlig opgavekravene, men det behøver **jo** ikke sige noget om tekstkvalitet. Det synes jeg, det er det næste step i det. Altså, ja de kan godt levere den vare der bliver bestilt, men kan de også løfte det ved at putte noget tekstkvalitet ind i det? (F2-V, min fremhævning)

2. Tekstkvalitet som noget mere

Vibekes sidste formulering udpeger et andet fællestræk ved censorernes kvalitetsnormer: Den detaljerede skriveordre har tendens til at stjæle fokus fra noget *mere* som kræves i en god tekst. Hvad dette mere er, forsøges indkredset i fokusgrupperne – stilladseret af samtalekortenes eksempeludsagn. Dialogen cirkler om substansen og tager associative ekskurser til erfaringer fra undervisning, censormøder og andre bedømmelsesopgaver for så at vende tilbage til forsøg på at formulere kvalitetsnormer. Forskellige metaforer foreslås af én og gribes og elaboreres af de øvrige deltagere. Der nikkes forstående blandt de andre når Vibeke taler om at ”kræse for læseren”, og Carsten ønsker sig ”engagement”, Berit ”sprogløblomster” og Morten at blive ”overrasket” og finde ”guldorn”. Her indikeres et relativt veletableret

²² Elevteksten er analyseret andetsteds som *Marikas tekst* (Troelsen, 2020).

²³ Han refererer til en passus i opgave 4: ”illustrer den [baggrundsartiklen] med fotos, faktabokse eller andre grafiske virkemidler” (min understregning). Der er altså ikke noget krav om fotos – hverken ét eller flere – men om grafiske virkemidler.

²⁴ Censorernes uddannelse foregår ved deltagelse på årlige censorkurser afholdt af undervisningsministeriets fagkonsulenter. Her gennemgås rammerne, eksempler på vurdering af elevtekster og ofte et årligt tema eller fokusområde (jf. mail fra fagkonsulenten 11.09.2019).

²⁵ Ved en fejl fra forskerens side var elevtekstens ene kildehenvisning ikke med i den version som deltagerne fik udleveret. Det kan have influeret på bedømmelsen, men her er det mindre relevant.

fortolkningsfællesskab – ud over de formulerede vurderingskriterier. At et sådant fortolkningsfællesskab findes, ækvivalerer fundene i tilsvarende undersøgelser af censorers tekstnormer (Berge, 1996; Berge et al., 2005). I denne specifikke sammenhæng er det værdt at bemærke at det fælles samtidig kan dække over individuelle forskelle bag de tentative og ikke særlig præcise formuleringer. Dog ses også mere eksplicite individuelle forskelle som deltagerne anerkender uden at diskutere – fx når Carsten forholder sig til Berits privilegering af verbalsprog:

Det lægger jeg også vægt på. Men nok ikke i helt så afgørende grad som du gør. Det tror jeg ikke. (F1-C)

Rasmus, der er ny censor, tier når de andre formulerer kvalitetsnormer ud over prøvevejledningens formuleringer. Han anfægter en lagkagemetafor og understreger at

... den kan være pyntet rigtig flot, men den kan smage rigtig dårligt, (F2-R)

og parafraserer vejledningstekstens ”for at en besvarelse kan siges at være dækkende i forhold til opgaveformuleringen, må den være udformet, som om den var et svar på en tilsvarende opgave i ”det virkelige liv” (STUK, 2016, s. 27):

... tekstkvalitet, det er når de øh, laver en opgave der **selvfølgelig** honorerer de krav den skal, men også øh, at man kan se det ligner virkelighedens verden. (F2-R)

Rasmus’ tøven når han formulerer sig eksplicit om tekstkvalitet, deler han dog med de øvrige. Således bekræftes kvalitetsnormernes doksiske karakter som "constitutive, tacit, discursive knowledge" (Berge, 2002, s. 459) som unddrager sig det referentielle sprog, men kan tangeres i metaforer og fungerer som resurse i praksis og tænke måde.

3. Kvantitet som kvalitet

Kvalitetsnormer viser sig altså diskursivt som metaforer og selvfølgeligheder (Halkier, 2016) i måden der tales om kvalitet på. Én selvfølgelighed er de talrige eksempler på sammenkædning af tekstlængde og -kvalitet som forekommer på tværs af empirien hos alle censorer – ofte som en glidning fra vurderingskriteriet ”fylde og disponering”.²⁶ Alt andet lige mødes korte tekster med skepsis, og lange forbindes med kvalitet. Udsagn som

... fyldig disponering – jamen det fylder da trods alt en halvanden side, (F1-C)

... der skulle have været mere med for at få den højere op, (TA-B)

og

... der er skrevet lige godt og vel en side, og hvis nu der ikke havde været spalter, ville det ikke fylde lige så meget, (TA-H)

er eksempler på højttenkning om enkelttekster. Temaet udfoldes i fokusgrupperne hvor Rasmus fx reflekterer over kortfattetthed og ender i en ringslutning:

... korte tekster efter tre og en halv times arbejde... Har man ikke nået mere end det man så har, så er det... så er det kortfattet. (F2-R)

Morten taler sig frem til en fortolkning af hvorfor korte tekster ”jo” er et problem:

²⁶ Tekstens længde er ikke omtalt i nogen styredokumenter.

Øh, og dermed vil jeg sige, når du siger (henvender sig til Vibeke), altså en halv side... Men er det nok? Det er det ikke nødvendigvis, men nej, det er ikke. Man skal også gøre sig umage, altså jeg synes ikke på en halv side... Altså man kan jo ikke komme til at vise en eller anden form for altså... engagement (F2-M, min fremhævning),

og Vibeke er én blandt flere der lader tekstlængde afgøre den endelige karakter:

... jeg tror ikke jeg har givet 12 for en besvarelse på trekvart side, det tror jeg ikke jeg har gjort. (F2-V)

Eksemplerne er legio og indikerer altså at en vis mængde verbaltekst – uanset genre – er nødvendig for at honorere eksamenssemiosfærens (Troelsen, 2020a) fordring om at demonstrere både kompetence og engagement, og får hos censorerne forrang frem for enkeltelevers investering i fx multimodalitet, sproglig præcision og minimalistisk genretro stil. Normen figurerer som en ikke-formuleret selvfølgelighed i fortolkningsfællesskabet og suspenderer sammenhængen mellem skriveformål og semiotiske resurser. Carsten har i sin egen undervisning i kortprosa ekspliciteret normen og rådet eleverne til at lade tekstmængde gå forud for genrekrav i eksamenssituationen:

... så skal I altså skrive mere. For I kan ikke vise nok sprog på ti linjer. (F1-C)

4. Sprogblomster og litterære normkilder

Eksamensvejledningen fremhæver ”det skriftsproglige udtryk” i vurderingen, med vægt på sproglig forståelighed, klarhed og variation (STUK, 2016, s. 22). Det forbinder censorerne først og fremmest med et ikke udfoldet, men selvfølgeligt krav om at sproget skal være ”i orden” (fx F2-C).

Det er **jo** også en opgave i ’Dansk, skriftlig fremstilling’ hvor man fremstiller noget sprog. (F2-R)

Desuden hylder de samstemmende et kreativt og litterært sprog – som andre eksamensår især har udfoldet sig i fiktionstekster:

Der er ingen tvivl om, hvis de selvfølgelig kan de andre ting her også og så har sprogblomster med, så ligger man **jo** højt i sproget (...). Jeg er ked af at den der fiktionstekst er væk, for der ligger så meget mulighed for at lege med sproget som ikke ligger på samme måde i en journalistisk genre fx. (F1-B)

Sprogblomsterne hyldes helt bogstaveligt når alle i TA-bedømmelserne fremhæver formuleringen ”Jeg har set det spire hurtigere end klematis” om digital mobning i et elevessay, og de savnes i fx en anden elevs baggrundsartikel som er holdt i nøgtern journalistisk stil. Et andet citat glæder alle i fokusgrupperne, og de forsøger at sætte ord på hvorfor:

Vibeke: Det giver også sådan en ro, altså der. Det er sådan... læsetempoet bliver sådan... øh, hvad skal man sige (kort pause) lidt romanagtigt, altså...

Moderator: Mhm?

Morten: Ja, men det essay... det’ et essay.

Vibeke: Ja.

Morten: Hvordan... Ja, det kan man **selvfølgelig** godt. Men det er sådan roman- eller lidt novelleagtigt, ikk’?

Vibeke: Mhm...

Moderator: Mhm?

Vibeke: Men det er jo yderst beskrivende. (F2)

Intuitivt anerkender Vibeke og Morten her den litterære stil, men funderer også over at den forekommer i et essay. De er enige om at elever der læser skønlitteratur, har lettere ved at skrive i alle genrer. Rasmus forholder sig som den sidste i F2 tavs og supplerer først senere med referencer til vejledningen der faktisk placerer essayet midt mellem fakta og fiktion, og peger på at sproglig kvalitet afhænger af genre.

5. Konstruktion af elevskrivere

Både TA-bedømmelser og fokusgrupper rummer talrige indekser på at censorerne læser teksterne som metonymiske repræsentationer af elever, eller hvad Ivanič kalder *opfattet skrive selv* (Ivanič, 2012). De opfatter elevskriverens personlighed og adfærd ud fra tekstlige træk og sammenholder den diskursive elevskriver med elever de kender. Eksempelvis infererer de sig til den biografiske skribers køn i de ellers anonymiserede tekster.²⁷ Og besvarelsernes mangler og kvaliteter læses som tegn på henholdsvis fravær og tilstedeværelse af personlige kvaliteter – fx ”sjusk” (F2-M), dovenskab (F2-V), ”engagement” (F1-C) og ”har gjort sig umage” (F2-M). Eller man slutter fra tekstlige træk til situeret adfærd: Carsten ytrer at en elevskriver der kun sporadisk besvarer skriveordren, ”ikke har læst opgaven igennem” (TA-C), og Morten at eleven bag en kort tekst uden afsnitsopdeling er ”gået efter tre kvarter” (F2-M) og ”bare [har] læst det af ikke” (F2-M). Vibeke problematiserer denne slutning til konkret adfærd, og Rasmus replicerer med selvfølgelighed:

Men vi kan godt se et engagement. Vi kan godt se engagement i den elev som godt nok staver som en brækket arm, men bare skriver derudad uden et punktum. Det er jo også et engagement. (F2-R)

Generelt værdsættes den engagerede elev der ”har noget på hjerte” (flere censorer), og ét af tegnene på engagement er, som før nævnt, tekstmængden. Metaforikken viser at censorerne oplever at velskrivende elever ”kræser for” dem (F2-V), og at de omvendt bliver sat ”på arbejde” (F2-alle) af mangelfulde tekster der udfordrer deres læsning og besværliggør bedømmelsen. Hanne positionerer sig mere neutralt og fagligt når hun overvejer

... om det er en tosproget elev siden vedkommende måske ikke har fanget at færgen er gået på grund, og vedkommende måske heller ikke sådan for alvor ved hvad en besætning er,

og en anden elev

måske har én eller anden ordblindeproblematik... (TA-H)

Generelt stiller eksamenssemiosfæren, trods forkærligheden for litterært sprog, andre krav til elevskrivere end til fx en prisbelønnet forfatter.²⁸

Moderator: Hvordan ville Helle Helle klare sig?

²⁷ Teksterne er anonymiseret som i den ordinære eksamenssituation. Forskeren kender imidlertid elevernes køn i og med at der er indhentet tilsagn om brug af teksterne til forskning. I fremstillingen er der derfor anvendt elevpseudonymer der indikerer elevens køn og etnicitet.

²⁸ Helle Helle (f. 1965) der ofte læses i danskundervisningen, er kendt for at skrive kort og nøgternt – altså modsat censorernes normer om tekstmængde og blomstrende sprog.

Berit: Ja, hvis vi ikke vidste det var hende... (F1)

6. Censoridentitet – positionering og selvopfattelse

Ifølge prøvebekendtgørelsen skal censor ”medvirke til og påse, at eleverne får en ensartet og retfærdig behandling, og at deres præstationer får en pålidelig bedømmelse” (MBUL, 2016). Denne rolle affilierer censorerne sig med, og de tvivler ikke på at de overordnet set lever op til den – på trods af den dokumenteret lave reliabilitet i den gældende ordning (Dolin et al., 2018; Flarup et al., 2019). Som tidligere nævnt er der tillige tegn på at de påtager sig en kontrolfunktion og fx vælger at ”ringe til næsten alle skoler” (F2-C) for at påpege mangler i elevtekster eller mistanke om snyd. De ser sig selv som ”professionelle” (F2-R) der i bedømmelsesarbejdet må lægge afstand til omsorgsfølelse:

... at man gerne vil være gode mennesker ved andre og gøre – og behandle dem pænt – altså, og det ved jeg ikke... men... men det har jo enormt meget betydning om det er 00 eller 02. (F2-R)

Direkte adspurgt af moderator svarer alle i F2 at de ikke er særskilt bevidste om elevskriveren når de overvejer den afgørende grænse mellem bestået og ikke bestået (00/02), og netop dér tvivler de ikke:

Morten: ... man er sgu aldrig i tvivl når det er 00. Det er jeg altså ikke.

Vibeke: Det er jeg heller ikke.

Morten: Men 02 og 4 og alt det der, og 4 og 7... Alt det, dér kan være masser af tvivl... Men lige præcis dét der – altså jeg kan også være i tvivl om 10 og 12... (F2)

En nærliggende fortolkning af denne sikkerhed om hvorvidt en tekst skal anerkendes i kulturen eller ej (Troelsen, 2020a), er at tvivl herom fortrænges fordi den ville indebære risiko for ansigtstab og krakeleret selvforståelse for censorerne.²⁹ Selvsikkerheden genfindes som nævnt i den nationale evaluering af ordningen (se ovenfor).

Den selvsikre positionering rystes imidlertid med det sidste samtalekort i fokusgrupperne (figur 1):

Karakterer og kvalitet

Tabellen viser jeres bedømmelser og den karakter, teksterne fik ved den ordinære censur. Yderst til højre finder I den gennemsnitlige karakter.

Tal sammen om jeres tanker om karakterer og kvalitet.

Tekst	Rasmus	Carsten	Vibeke	Morten	Hanne	Berit	FP9	Gennemsnit
Emre	02	02	4	02	02	02	02	2,3
Marika	7	4	10	4	7	7	7	6,6
Veronica	10	7	7	7	10	10	7	8,3
Samira	7	10	10	4	10	7	12	8,6
Ester	12	7	12	10	12	12	12	11
Patrick	02	02	4	00	00	02	02	1,7
Anna	10	12	12	10	10	7	7	9,7

Figur 1: Sidste samtalekort anvendt i fokusgrupperne.

²⁹ Tak til Michael Tengberg for inspiration til denne pointe under en samtale i november 2019.

Dramaturgisk er oversigten over karakterer placeret sidst i fokusgruppesamtalerne, og den vækker forbløffelse selv om jeg fra start har forberedt deltagerne på hvad de kan forvente. Gennemsnitskorrelationen i det meget lille udvalg er i virkeligheden påfaldende høj sammenlignet med kendte studier: 0.82 mod 0.55 i den nationale evaluering af ordningen.³⁰ Alligevel skaber oversigten forundring og frustration – udtrykt som skiftevis tavshed, selvfølgelig og løsrevne ytringer.

Vibeke: Ej, hvor sjovt, ikk'...

Moderator: Mhm?

Vibeke: ... at vi er, altså helt...

Morten: ... altså uenighed, det er der jo altså... (F2)

Den anden gruppe diskuterer om de har ”ramt skævt” eller ”rigtigt” (F1-B&C) sammenlignet med den censor som bedømte teksterne oprindeligt (FP9 i figur 1), og viser altså metaforisk at de tillægger den oprindelige bedømmelse autoritet og status af mål uagtet at denne autoritet er meningsløs qua den indbyrdes variation.

Karakteroversigten skaber usikkerhed om bedømmerpositionen. Morten stiver sig selv af – han har givet 4 for en besvarelse som en anden har givet 12:

(...) det kan være jeg har haft et blackout. Men nu vil jeg til hver en tid kunne pille de der 12-taller fra hinanden og i hvert fald få det ned på et 10-tal. Det der, det er simpelthen... det der 12-tal... (kigger på papiret på bordet), ja, det er helt langt væk. Det er helt langt væk fra skiven... (F2-M)

Berit undres:

(...) altså, jeg tænker **jo**, at da jeg sad der, da ramte jeg det jeg syntes de skulle have, ikke også. Og det er vi **jo** alle sammen enige om at vi har gjort. Men jeg kan mere sige at... er vi uddannet godt nok? (F1-B)

Og Rasmus er frustreret:

... jeg troede jeg var så omhyggelig og så præcis. Det ved jeg godt, det er jeg ikke, men alligevel... (F2-R)

Alles selvforståelse er tydeligt udfordret, og der er ansats til rehabilitering af den tidligere bedømmelsesordning hvor censor voterede med dansklæreren. Den har censorerne ellers fundet besværlig og antydte at elevernes lærere havde tendens til at ville trække karaktererne op (Dolin et al., 2018; Flarup et al., 2019).³¹ Nu forbindes forhandlingssituationen med mulig kvalitet:

Man kunne bruge den [forhandlingen] til... og så sige at: du har givet 7, og jeg har givet 12. Hvorfor er vi så forskellige? Og så landede man et eller andet sted, og det kunne både have været 7, 10 eller 12. Hvad der var af kvalitet i teksten. (F1-B)

³⁰ Kjell Lars Berge har venligst foretaget diverse statistiske beregninger på materialet, og det skal siges at dette tal er usikkert grundet det meget lille materiale. De mere kvalitative beregninger er mere valide og viser interessante forskelle mellem censorerne hvad angår rangering og brug af skala. Disse gemmes dog til et senere studie idet fokus her fastholdes på censorernes tekstnormer.

³¹ Der er ikke belæg for dette. Evalueringen viser at karakterniveauet ikke er forandret med den nye ordning (Flarup et al., 2019).

Således bliver Berit den der – foruden bedre uddannelse af censorerne (se ovenfor) – foreslår en tilbagevenden til den dialogiske forhandling om karakterer og peger på at fortolkningsfællesskabets kvalitetsnormer kan være en mulig kilde til forbedring.

Sammenfatning og diskussion

Jeg har i denne artikel undersøgt 6 beskikkede censorers kvalitetsnormer og positioneringer i en high-stakes-kontekst hvor de som enebedømmere er tildelt betydelig magt over elevers fremtidige uddannelsesmuligheder og skal agere i en censurordning med meget lav reliabilitet. Jeg har således spurgt til den betydning de selv tillægger deres bedømmelse og udmøntede karakterer – herunder hvordan de forstår deres egen rolle i bedømmelsen, og hvordan de opfatter elevskriverne. Det er vigtigt for mig at understrege at hensigten med studiet ikke har været at problematisere censorernes praksis. De agerer som de kan, under de givne rammebetingelser skabt af det politiske niveau og administreret af Undervisningsministeriet. Min erkendelsesinteresse har i stedet været rettet mod de kvalitetsnormer og -opfattelser som i denne kontekst ligger bag elevernes eksamenskarakterer. Fundene skal ses i lyset af dette.

Studiet betragter censorerne som eksamenstekstkulturens kendere. Deres *kvalitetsnormer* viser sig i *eksplicite ytringer* om kvalitet og *indlejrede forståelser* i fokusgruppesamtalernes metaforiske tegning af et ”sammenhængende system” (Lakoff & Johnson, 2002, s. 28) der betragter kvalitet som *noget mere*. Det er således ”en del af [kvalitets]begrebet” (ibid., s. 29) at det indebærer tekstligt og sprogligt overskud både kvantitativt og kvalitativt – også ud over de formulerede krav. Skønlitterære normkilder og diskursivt signaleret engagement er indskrevet i tekstkulturens dna – uanset genre. Hansen karakteriserer i forbindelse med KiDM-projektet³² (Hansen, Elf, Misfeldt, Gissel, & Lindhardt, 2020) ”kvalitet som et konkret, sanseligt og i sidste instans ubestemmeligt fænomen. Kort fortalt skal kvalitet opleves” (Hansen, 2019) og altså opleves *af nogen*. Det er da også tydeligt at censorerne forholder sig æstetisk-fænomenologisk i deres metaforiske indkredsning af fænomenet, og de bruger i høj grad sig selv i vurderingsprocessen.

Censorernes positionering er i nogen grad ambivalent. Vurderingskriterierne og skriveordrens fordringer ses som minimumskrav og kontrolpunkter med stor betydning for bedømmelsen. Samtidig forudsætter fuld anerkendelse at elevteksten rummer fornævnte ”mere”.³³ Konkrete kvalitetsnormer for fx teksters indhold, indre struktur eller kommunikative funktion nævner censorerne ikke, og i spændingsfeltet mellem teknisk og hermeneutisk rationalitet er der tendens til at *elevteksten forsvinder* til fordel for instrumentelt fokus på minutøse krav på den ene side og (mere eller mindre subjektive) tavse kvalitetsnormer på den anden. Den afbalancerede tekstkulturelle rationalitet funderet i fortolkningsfællesskabet bliver klemmt i den monologiske solobedømmelse hvor kvalitet ikke formuleres og forhandles.

Censorernes *sehforståelse* er præget af tiltro til egen dømmekraft og professionalisme, og de opfatter sig både som kontrollører af at kravene overholdes, og garanter for tekstkulturens videreførelse. Selvsikkerheden udfordres dog i konfrontationen med den faktiske bedømmervariation, og fokusgruppernes spontane forbedringsforslag er videreuddannelse og revitalisering af censorpar bestående af professionelle bedømmere og elevernes faglærere.

³² *Kvalitet i Dansk og Matematik*.

³³ Dette kan delvist forklare at elevtekster der håndterer skriveordrens krav minutøst og lever op til vurderingskriterierne, ikke modtager fuld anerkendelse af bedømmerne (Troelsen, 2020a).

En praksisimplikation af studiet kunne være at lytte til censorerne og genoverveje en opdateret version af denne tidligere praksis – af flere grunde: Det kan muligvis *styrke reliabiliteten*. At følgegruppens evaluering (se note 3) kun finder svag reliabilitetsforbedring ved to bedømmere (Dolin et al., 2018), kan skyldes at undersøgelsesdesignet alene baseres på en statistisk parkonstruktion uden forhandling med basis i et fortolkningsfællesskab af den type som KAL-projektet udpeger som nøglen til en relativt høj reliabilitet.³⁴ Desuden kan en forhandlingsmodel forventes at *styrke den faglige tekstekultur* som danner grundlag for både undervisning og eksamen. Fortolkningsfællesskaber forudsætter dialogisk interaktion, og i dialogen mellem professionelle bedømmere og elevernes faglærere udvikles og konsolideres fælles kvalitetsnormer for undervisning og evaluering. Dette sidste vil endvidere løse det problem at også lærerne savner sparring og selvevaluering (Flarup et al., 2019).³⁵ Endelig kan man overveje at lade sig inspirere af NORM-projektets udvikling af vurderingskriterier i dialog med og funderet i fortolkningsfællesskabet – eventuelt operationaliseret som interventionsstudie.

Tak til...

Professor Kjell Lars Berge for uvurderlig inspiration til dette studie og til professor Nikolaj Elf for udfordring af og idéer til tidligere udkast af artiklen.

NB. En tidligere version af artiklen findes som kapitel 6 i min ph.d.-afhandling *Forkastet eller anerkendt? Et eksplorativt studie i folkeskoleelevers afgangseksamen i 'Dansk, skriftlig fremstilling'* (Troelsen, 2020b).

Om forfatterne

Solveig Troelsen er ph.d. og lektor ved VIA University College, læreruddannelsen i Årus. Hun underviser og forsker i danskfagets didaktik, skriveidaktik og skriveprøver.

Referencer

- Bakhtin, M. M. (1986). The Problem of Speech Genres. In C. Emerson & M. Holquist (Eds.), *Speech genres and other late essays* Austin: University of Texas Press.
- Barkaoui, K. (2011). Think-aloud protocols in research on essay rating: An empirical study of their veridicality and reactivity. *Language Testing*, 28(1), 51-75.
- Barton, D., Hamilton, M., & Ivanič, R. (2000). *Situated literacies: Reading and writing in context*. Psychology Press.
- Berge, K. L. (1993). Hvorfor er det så vanskelig å få til en pålitelig bedømming av elevtekster? . In P. A. Michelsen (Ed.), *Sjangeroppbrudd* (s. 151–190). Oslo: Ad Notam.

³⁴ Om reliabiliteten i den tidligere danske ordning var højere, er mig bekendt aldrig undersøgt.

³⁵ Begge evalueringer peger på alvorlige problemer som følge af at faglærerne savner den sparring og selvevaluering som tidligere var indlejret i voringssituationen (Dolin et al., 2018; Flarup et al., 2019). Og da karakterniveauet ikke er faldet med overgangen til den anonyme enebedømmerordning (Flarup et al., 2019), er der ikke belæg for censorernes oplevelse af at faglærerne påvirker karaktererne unødigt i positiv retning fordi de kender eleverne.

- Berge, K. L. (1996). *Norsksensorenes tekstnormer og doxa: en kultursemiotisk og sosiotekstologisk analyse*. Institutt for anvendt språkvitenskap det historisk-filosofiske fakultet NTNU, Trondheim.
- Berge, K. L. (2002). Hidden Norms in Assessment of Students' Exam Essays in Norwegian Upper Secondary Schools. *Written Communication*, 19(4), 458-492.
- Berge, K. L. (2005). Skriveprøvenes pålitelighet. In K. Berge, L. Evensen, F. Hertzberg, & W. Vagle (Eds.), *Ungdommers skrivekompetanse Bind I*.
- Berge, K. L. (2009). Er tolkningsfellesskap mulig å oppnå i skriveprøver? In O. K. Haugaløkken, L. S. Evensen, F. Hertzberg, & H. Otnes (Eds.), *Tekstvurdering som didaktisk utfordring* (s. 44-54). Oslo: Universitetsforlaget.
- Berge, K. L., Evensen, L. S., Hertzberg, F., & Vagle, W. (2005). *Ungdommers skrivekompetanse (Bind I & II)*. Oslo: Universitetsforlaget.
- Björnsson, C. H. (1960). *Uppsatsbedömning och uppsatsskrivning*. Stockholm: Almqvist och Wiksell.
- Blomqvist, P. (2018a). Vad uppmärksammar lärare i samtal om skrivbedömning? Svensklärares normer för beslut om summativ bedömning. *Nordisk tidskrift för allmän didaktik*, 4(1), 34-55.
- Blomqvist, P. (2018b). Är vi överens om det? Samspel och samstämmighet i svensklärares bedömningsamtal om gymnasieelevers skrivande. *Nordic Journal of Literacy Research*, 4(1).
- Borgström, E. (2014). *Skrivbedömning: om uppgifter, texter och bedömningsanvisningar i svenskämnet nationella prov*. (PhD). Örebro university,
- Borgström, E., & Ledin, P. (2014). Bedömarvariation: Balansen mellan teknisk och hermeneutisk rationalitet vid bedömning av skrivprov. *Språk & stil*, 24, 133-165.
- Christensen, T. S., Elf, N. F., & Krogh, E. (2014). *Skrivekulturer i folkeskolens niende klasse*. Odense: Syddansk Universitetsforlag.
- Cooper, C. R., & Odell, L. (1977). Evaluating writing : describing, measuring, judging. In C. R. Cooper & L. Odell (Eds.). Urbana, Ill: NCTE.
- Dolin, J., Nielsen, K., & Rangvid, B. S. (2018). *Rapport fra Følgegruppen for en bedømmer ved folkeskolens prøver*.
- Elf, N. (2009). *Towards semiocy. Exploring a New Rationale for Teaching Modes and Media of Hans Christian Andersen Fairytales in Four Commercial Upper-Secondary "Danish" Classes: A Design-Based Educational Intervention*. (PhD). SDU,
- Ericsson, K. A. (1993). *Protocol analysis: verbal reports as data* (Rev. ed. ed.). Cambridge, Mass: MIT Press.
- Evensen, L. S. (2005). Pålitelighet og betydning. Etterord. In K. L. Berge, L. S. Evensen, F. Hertzberg, & W. Vagle (Eds.), *Ungdommers skrivekompetanse. Bind I: Norsksensuren som kvalitetsvurdering* (s. 205-211).
- Evensen, L. S., Berge, K. L., Thygesen, R., Matre, S., & Solheim, R. (2016). Standards as a tool for teaching and assessing cross-curricular writing. *The Curriculum Journal*, 27(2), 229-245.
- Flarup, L. H., Bro, L. L., & Arendt, K. S. (2019). *Evaluering af ordningen med én bedømmer ved folkeskolens skriftlige prøver*.
- Halkier, B. (2016). *Fokusgrupper* (3rd ed.). København: Samfundslitteratur.

- Hansen, T. I. (2019). Perspektiver på god kvalitet i undervisningen. Retrieved from <https://www.folkeskolen.dk/654223/perspektiver-paa-god-kvalitet-i-undervisningen>
- Hansen, T. I., Elf, N., Misfeldt, M., Gissel, S. T., & Lindhardt, B. (2020). *Kvalitet i dansk og matematik. Et lodtrækningsforsøg med fokus på undersøgelsesorienteret dansk- og matematikundervisning. Slutrapport*. Retrieved from file:///C:/Users/sotr/Downloads/200115-Evalueringssrapporten-Kvalitet-i-dansk-og-matematik%20(1).pdf
- Igland, M-A. (2013). Kunsten å lese elevtekster. In G. Skar & M. Tengberg (Eds.), *Läsning! Svensk lärarförningens Årsskrift*: Svensk lärarförningens.
- Indenrigsministeriet. (1958). Lov nr. 220 af 18. juni: Bekendtgørelse af Lov om Folkeskolen. Retrieved from <http://danmarkshistorien.dk/leksikon-og-kilder/vis/materiale/bekendtgørelse-af-lov-om-folkeskolen-18-juni-1958/>
- Ivanič, R. (2012). Writing the self: the discursual construction of identity on intersecting timescales. In S. Matre, D. K. Sjøhelle, & R. Solheim (Eds.), *Teorier om tekst i møte med skolens lese- og skrivepraksiser*. Oslo: Universitetsforlaget.
- Jølle, L. (2014). Pair assessment of pupil writing: A dialogic approach for studying the development of rater competence. *Assessing Writing*, 20, 37-52.
- Jølle, L., & Skar, G. B. (2018). "Digging for Gold" or "Sticking to the Criteria": Teachers' Rationales When Serving as Professional Raters. *Scandinavian Journal of Educational Research*, 1-17.
- Kane, M. (2006). Validation. In *Educational Measurement*.
- Karlsson, A-M. (1997). Textnormer i och utanför skolan – att skriva insändare på riktigt och på låtsas. *Svenskans beskrivning*, 22.
- Krogh, E. (2015). Faglighed og skriftlighed - teori, metode og analyseramme. In E. Krogh, T. S. Christensen, & K. S. Jakobsen (Eds.), *Elevskrivere i gymnasiefag*. Odense: Syddansk Universitetsforlag.
- Lakoff, G., & Johnson, M. (2002). *Hverdagens metaforer*. Kbh: Hans Reitzel.
- Malterud, K. (2012). Systematic text condensation: a strategy for qualitative analysis. *Scandinavian journal of public health*, 40(8), 795-805.
- Matre, S., Berge, K. L., Evensen, L. S., Fasting, R., Solheim, R., & Thygesen, R. (2011). *Developing National Standards for the Teaching and Assessment of Writing. Rapport frå forprosjekt*. Retrieved from
- MBUL. (2016). *Bekendtgørelse om folkeskolens prøver. BEK nr 1132 af 25/08/2016*. København: Lovtidende A
- Moss, P. A. (1994). Can There Be Validity Without Reliability? *Educational researcher*, 23(2), 5-12.
- Nilsen, H. (2005). Hva sier sensor? Sensor kommentarer om tekstkvalitet. In K. Berge, L. Evensen, F. Hertzberg, & W. Vagle (Eds.), *Ungdommers skrivekompetanse Bind I*.
- Shanahan, T., & Shanahan, C. (2012). What Is Disciplinary Literacy and Why Does It Matter? *Topics in Language Disorders*, 32(1), 7-18.
- Skar, G. B. (2013). *Skrivbedømming och validitet: fallstudier av skrivbedømming i svenskundervisning på gymnasiet*. Institutionen för språkdidaktik, Stockholms universitet,
- Skar, G. B., & Jølle, L. J. (2017). Teachers as raters: An investigation of a long-term writing assessment program. *L1-Educational Studies in Language and Literature*, 17, 1-30.

- Skar, G. B., & Aasen, A. J. (2016). Risikotrekk og skjulte kvalitetar i elevtekstar. *Nordic Journal of Literacy Research*, 2(0).
- Slomp, D. H., & East, M. (Eds.). (2019). *Special issue: Framing the Future of Writing Assessment* (Vol. 42).
- Small, M. L. (2009). 'How many cases do I need?': On science and the logic of case selection in field-based research. *Ethnography*, 10(1), 5-38.
- Vejledning til folkeskolens prøver i faget dansk – 9. klasse, (2016).
- Troelsen, S. (2018). En invitation man ikke kan afslå – analyse af afgangsprøven i skriftlig fremstilling med særligt fokus på skriveordren. *Nordic Journal of Literacy Research*, 4(1), 142-166.
- Troelsen, S. (2020a). At håndtere flertydighed under pres – elevskriveres tekstnormer og skri-verselv ved folkeskolens afgangseksamen i skriftlig fremstilling. *Nordic Journal of Literacy Research*, 6(1), 126-183.
- Troelsen, S. (2020b). *Forkastet eller anerkendt? Et eksplorativt studie i folkeskoleelevers afgangseksamen i 'Dansk, skriftlig fremstilling'*. (Ph.d.-afhandling). Syddansk Universitet, Odense.
- Xu, Y., & Brown, G. T. L. (2016). Teacher assessment literacy in practice: A reconceptualization. *Teaching and Teacher Education*, 58, 149-162.