Erfaringer fra strukturert peer review ved bruk av et egetutviklet sensureringsprogram

Omid Mirmotahari og Yngvar Berg

Universitetet i Oslo,

Institutt for Informatikk

Sammendrag

Peer review eller fagfellevurdering er en stadig vanligere metode som kan bidra til å øke læringsutbytte. I denne studien retter sudentene medstudenters besvarelser. Det viser seg at metoden gir studentene økt læringsutbytte og større faglig refleksjon. Studien ble gjennomført ved å bruke siste års eksamen som utgangspunkt for en fagfellevurdering. Et eget dataprogram er utviklet for strukturert gjennomføring av sensureng eller fagfellevurdering. Det viser seg at studenter som deltok i fagfellevurderingen gjennomgående fikk bedre resultat på avsluttende eksamen. Dette tyder på at strukturert fagfellevurdering gjennomført av studenter gir økt læringsutbytte. Resultatene kan også komme til nytte for faglærere.

Keywords: Peer Review, Automatisk tilbakemelding, Automatisk begrunnelse, Formativ vurdering, Digital vudering, Læringsanalyse.

1 Introduksjon

Fagfellevurdering eller ekstern fagvurdering er typiske norske oversettelser for det engelske ordet peer review. I forskningssammenheng er peer review en viktig metode for å sikre kvalitet, aktualitet og relevans i vitenskapelige publikasjoner. Den som fagfellevaluerer vil også kunne øke sin faglige innsikt. Vi antar at studenter kan få økt faglig innsikt ved å utføre fagfelleevalueringer. Studentene får en bedre og dypere forståelse av faget samtidig som de får en personlig utvikling innen generell profesjonell kompetanse [1–6]. Peer review gir dessuten et nyttig læringsutbytte for studenten som retter en oppgave. Dette styrkes ytterligere ved struktureringen av retteprosessen. Studentene lærer hva som verdsettes i en besvarelse, og å se etter vanlige feil og mangler, og det vil bidra til å gi studentene et meta-perspektiv på sin egen forståelse og læring. Derfor er peer review viktig som er læringselement og ikke bare som en kostnadsreduserende faktor. Det å bruke fagfellevurdering som metode blant studenter i et emne kan være tidkrevende og utfordrende for både faglærer og studenter. Det er viktig at den gjennomføres strukturert og godt planlagt av faglærer. Det er i de siste årene kommet flere dataprogrammer og systemer som gjør det enklere å organisere fagfellevurderinger, spesielt online programmer som peergrade[1], Canvas rubrics[2] er nyttige verktøy. Felles

Denne artikkelen ble presentert på konferansen NIK-2017; se http://www.nik.no/.

[1] https://www.peergrade.io/

[2] https://www.canvaslms.com/

for disse programmene er at de håndterer innlevering og utlevering av besvarelser på en effektiv måte, men det å sette opp vurderingskriterier er fortsatt tidskrevende for faglærer. Ved hjelp av deikerte programmer har derfor fokuset i forhold til peer review skiftet mot det å utvikle gode rubrikker (rubrics) eller skjema for vurdering. Forskning på student peer review rubrics viser at mest effektive rubrikker er de som kan bli brukt for alle typer oppgaver i et kurs og som har samme ordbruk og kriterier og som er uavhengige av selve problemløsningsrammeverket [7]. En interessant problemstilling som har dukket opp som følge av student peer review er studentenes tillit til egen evne til å rettferdig kritisere og gi tilbakemelding til medstudenter [8], og tillit til at egen besvarelse blir vurdert rettferdig av medstudenter [9].

Med strukturert peer review mener vi en strukturert metode for studentene når de skal gjennomføre en fagvurdering eller sensurering. Sensor eller peer som skal vurdere en besvarelse blir geleidet ved hjelp av et dataprogram der de skal påpeke eventuelle feil og mangler. Sensureringsprogrammet er designet som en avkrysningsskjema hvor sensor velger feil og mangler. Basert på hvilke valg eller avkrysning sensor gjør vil gi utslag på ulike kriterier og måleparameter som faglærer har etablert. En sensor vil umiddelbart se utslaget for de ulike kriteriene basert på avkrysningen som gjøres. For å kunne gi økt fleksibilitet til sensor, vil det være muligheter for en sensor til å overstyre vekten for utslaget på kriteriene. Tilbakemeldingene som lages er strukturerte i forhold til det sensor har påpekt / valgt.

Det å gi en oppgave med peer review som formativ vurdering er en god måte å gi studenter en dypere forståelse i faget og det kan brukes til å gi studentene tilbakemelding underveis i semesteret. Gjennom en formativ «feedforward» vurdering vil studentenes evne til egen refleksjon øke [10–13]. Slik som Biggs [14] beskriver constructive alignment, samhandling, er i hovedsak at det er et tett samsvar mellom undervisningselementene, det forventede læringsutbytte og vurderingen i et emne. Peer review bidrar til å styrke dette samsvaret og gjøre studentene mer fortrolige med sammenhengen mellom innhold, vurdering og vurderingskritierer. Det å bruke peer review som et verktøy i en eksamenssetting mener vi kan gi en god samhandling. En «prøveeksamen» som blir evaluert av medstudenter med formativ feedforward tilbakemelding gjør at studentene vil kunne stille bedre til avsluttende eksamen.

Vårt forsøk er lagt opp til at studentene kan velge å ha en prøveeksamen, i så eksamensrealistiske omgivelser som mulig, helt i slutten av et emne. Etter innlevert prøveeksamen vil hver kandidat få utdelt fem besvarelser fra medstudenter som de må rette, peer-review, gjennom å bruke et sensureringsprogram. Dette sensureringsprogrammet vil lagre og loggføre alle endringer og bemerkninger studentene/sensorene har og etter endt sensurering vil faglærer kunne hente ut rapporter og analyser over besvarelsene. Vi vil i denne artikkelen vise noen av resultatene som belyser dette.

I denne artikkelen gis først en kort presentasjon i avsnitt 2 av sensureringsprogrammet som er utviklet. I avsnitt 3 beskriver vi metoden og de praktiske elementene for gjennomføringen av dette forsøket. Avsnitt 4 viser oversikten over studentes resultater både i prøveeksamen og i endelig eksamen. Studentenes tilbakemeldinger er presentert i avsnitt 5 med tre utvalgte fokusområder. All datainnsamling som gjøres ved bruk av programmet og en rekke post-analyser, inkludert tilbakemeldinger til faglærer, presenteres i avsnitt 6. Artiklen avsluttes med en diskusjon av de ulike funnene og resultatene. Her besvarer vi forskningsspørsmålene med tilhørende forslag for videre arbeid.

Det er mange aspekter i dette forsøket som kan ha stor akademisk og praktisk nytteverdi for undervisningen. I første omgang ønsker vi å belyse hvordan et slikt

eksperiment som dette kan påvirke studentenes læringsutbytte og resultater, studentenes samhandling og å addressere undervisningsansvarliges erfaringer. Vi ønsker å belyse:

- Vil forsøket bidra til økt samhandling og bedre forståelse av læringsmålene for studentene?
- Vil studentene oppnå bedre resultater på eksamen som følge av strukturert peer-review med prøveeksamen?
- Vil sensureringsprogrammet gi verdifull informasjon til faglærer om studentenes læring, og kan det brukes til å endre/videreutvikle emnet?

2 Dedikert sensureringsprogram

Sensureringsprogrammet som ble utviklet ble først presentert i [15]. Det har siden gjennomgått noen revisjoner og fått nye analytiske verktøy. Selv om programmet i første omgang har vært utviklet for sensurering av eksamen og gi individuell automatiske tilbakemeldinger, har det i tillegg et stort potensiale for å kunne brukes i formativ vurdering. Sett fra sensorerens side vil dette kunne være tidsbesparende og veiledende, mens det for studentene vil bidra til en mer rettferdig vurdering av deres besvarelse og de vil få en skriftlig feedforward tilbakemelding. For faglærer vil programmet kunne gi en rekke analyser og rapporter både om sensorenes retting, studentenes besvarelser og kunnskapsnivå. Et skjermbilde av programmet er vist i Figur 1. Som vist på figuren samles det en rekke data for hver eneste oppgave. Sensors hovedoppgave er å krysse av for hvilke «feil» studenten har gjort i rubrikken betegnet som begrunnelse. Basert på valget av «feil»/begrunnelse vil systemet gi et utslag både i forhold til poeng for den gitte oppgaven, samt utslag for det som vi kaller for kriterier. Graden av utslag for de ulike elementene ligger i kjernen av programmet og er utarbeidet i forkant av faglærer. Sensor kan overstyre utslagene ved behov og ytterligere bruke fritekst området for å gi utdypning eller spesifikke kommentarer. Programmet vil hele tiden følge valgene som sensor gjør og lagre alle endringer og overstyringer. All data blir så behandlet slik at systemet kan kalibrere sensorenes skjønnsmessige vurderinger av oppgavene. Gjennom de kvantifiserte kriteriene og begrunnelsene vil sensureringsprogrammet utføre en rekke beregninger og innhente en tekstlige predefinerte «fraser» som settes sammen til en helhetlig tekstlig tilbakmelding. Denne individuelle tilbakemeldingen er automatisk generert på bakgrunn av ulike terskler for kriteriene som er bestemt på forhånd. Hvert kriterie har underliggende tekstfraser og avhengig av den vektede akkumulerte verdien for kriteriet vil en spesifikk tekstfrase bli plukket ut. Således blir alle kriteriene satt sammen til en helhetlig skriftlig indidivuell tilbakemelding. Programmet kan gi både faglig begrunnelse og individuell tilbakemelding. Dersom sensor velger å gi både faglig begrunnelse og individuell tilbakemelding vil den første delen av tilbakemeldingen være en faglig begrunnelse for hver deloppgave og delkarakter/poengsum. Den andre delen vil være en individuell tilbakemelding basert på feedforward.

3 Metode

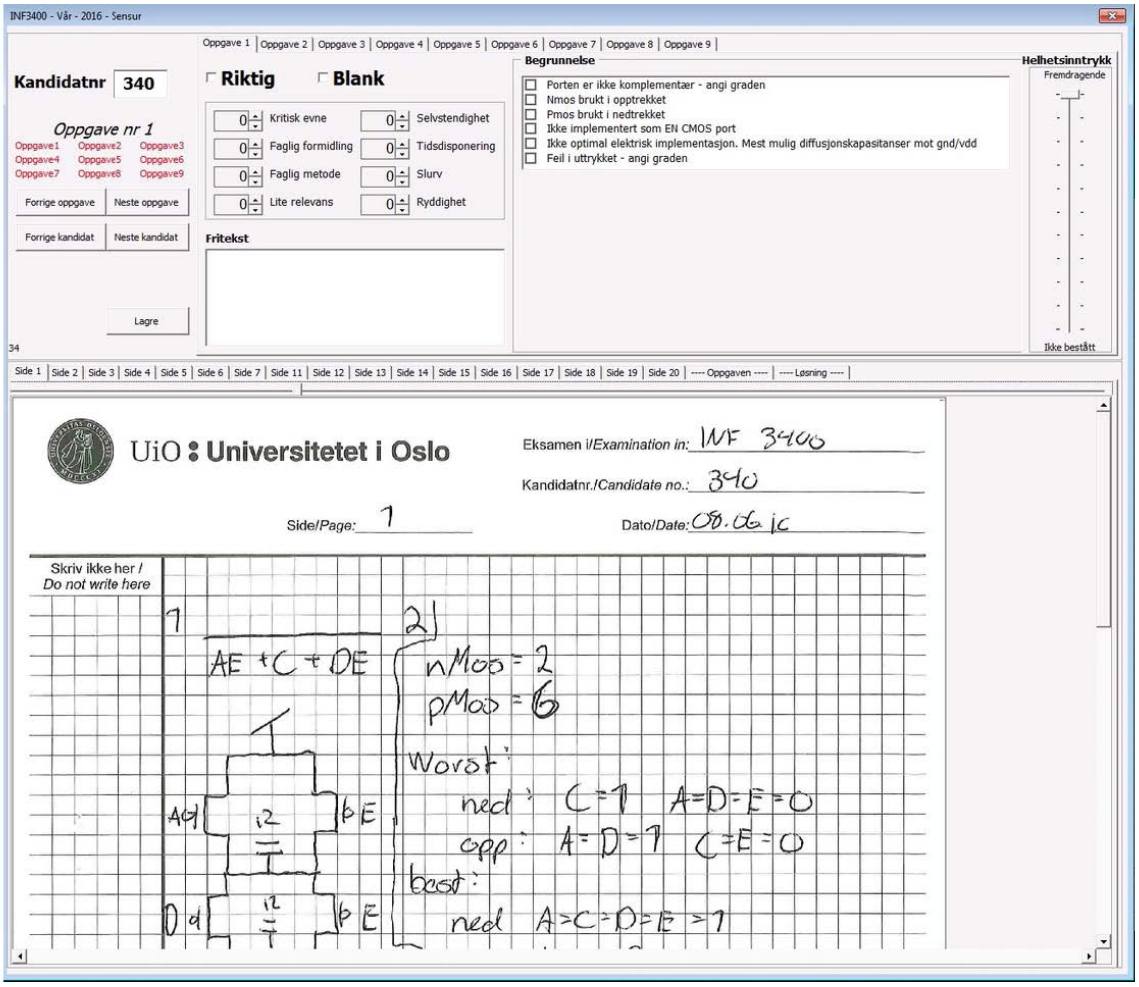
Vi har valgt å gjennomføre forsøket i INF3400 - Digital nanoelektronikk med 10 studiepoeng. Emnet inngår i fjerde semester og det er åpent for studenter tatt opp på

bachelor i studieprogammene I:NOR[3] og ELDAT[4], begge tilhørende Universitetet i Oslo. Emnet består av 2 timer forelesninger, 2 timer lab, 2 timer gruppeundervisning per uke og har totalt 8 obligatoriske innleveringer. Påmeldingen for emnet i 2017 var 46 studenter og antall kvalifisert til eksamen var 42.

I 2017 fikk studentene et tilbud der de kunne velge mellom å gjennomføre en tradisjonell obligatorisk oppgave nummer 8 eller ta prøveeksamen med peer-review. Studentene var ikke på forhånd kjent med at prøveeksamen var forrige års eksamen. Eksamenssettet hadde heller ikke blitt gjort tilgjengelig for studentene tidligere. Pensum for hele emnet ble gjennomgått før prøveeksamen og tidspunktet for prøveeksamen var satt til ca fire uker før avsluttende eksamen. Selve prøveeksamen ble gjennomført i så realistiske omgivelser som mulig. Det innebar at studentene fikk utdelt gjennomslagspapir til sine besvarelser og de fikk fire timer til å løse eksamen i et rom tilsvarende eksamenslokalene. Hver student fikk utdelt et kandidatnummer for å sikre anonymitet og det var bare seminarlærerne som hadde oversikten over kandidatnummer og navnet til studentene. Det var også seminarlærerne som samlet inn besvarelsene, skannet dem og

[3]Informatikk: Robotikk og intelligente systemer

[4]Elektronikk og datateknologi



Figur 1: Skjermbilde av sensureringsprogrammet i emnet INF3400 for eksamen i 2016. Studentenes besvarelser er skannet inn og programmet finner automatisk riktig kandidat og riktig oppgave.

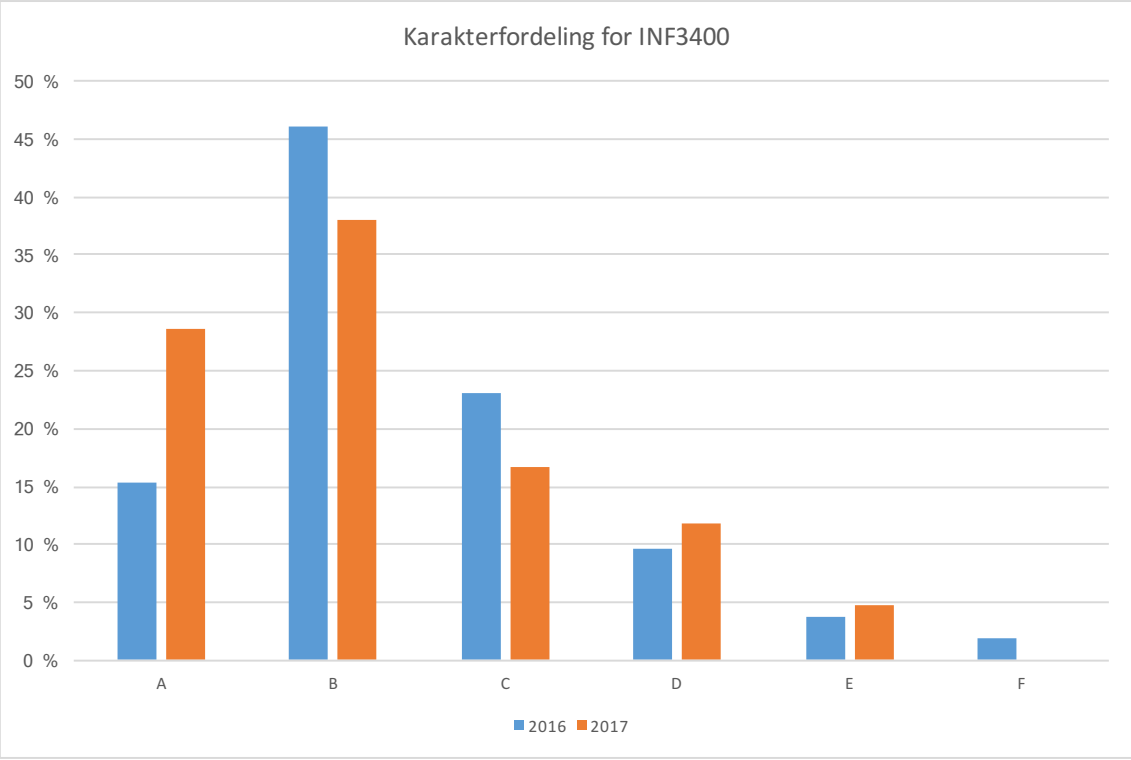
sendte ut for peer-review. Hver student fikk besvarelser fra fem medstudenter umiddelbart etter prøveeksamen, som de skulle rette og gi tilbakemelding på innen en uke. Det var 35 som valgte å ta prøveeksamen, og 7 som ønsket å ta den tradisjonelle obligatoriske oppgaven.

For å sikre at studentene fikk tilbakemelding, rettet faglærere hver besvarelse ved bruk av programmet og sendte individuelle automatisk genererte tilbakemeldinger til studentene. Vi fikk da samlet inn data som er direkte sammenlignbare med forrige års eksamen eksamen, fordi samme sensorer og sensureringsprogram ble brukt for fjorårets kull.

Studentenes erfaringer ble samlet gjennom anonyme nettskjema og det ble tilfeldig valgt 5 studenter for intervju. Resultatene er presentert i de neste avsnittene.

4 Studentenes resultater

Karakterfordeling i 2016 og 2017 er vist i Figur 2. Figuren viser forbedringer i karakterer for 2017 kullet i forhold til 2016 kullet. Spesielt er det verdt å merke seg resultatene for karakteren A som har fått en vesentlig økning med nesten en dobling i forhold til 2016. For karakterene B og C er antallet i 2017 det noe lavere enn i 2016, og noe overraskende er det flere D og E i 2017 enn i 2016. Det kan virke som om det er et «klasseskille» i karakterefordeling i 2017. Det kan være ulike årsaker til en slik fordeling, men først er det interessant å se på hvordan karakterfordelingen har vært for de som tok prøveeksamen i forhold til de som ikke tok prøveeksamen. Resultatet er gitt i Tabell 1 med nøkkeltallene for resultatene fra endelig eksamen 2017 for de to gruppene, altså de som tok prøveeksamen og de som tok tradisjonell obligatorisk oppgave. Tallene viser at snittkarakteren oppnådd på endelig eksamen for gruppen som har tatt prøveeksamen er



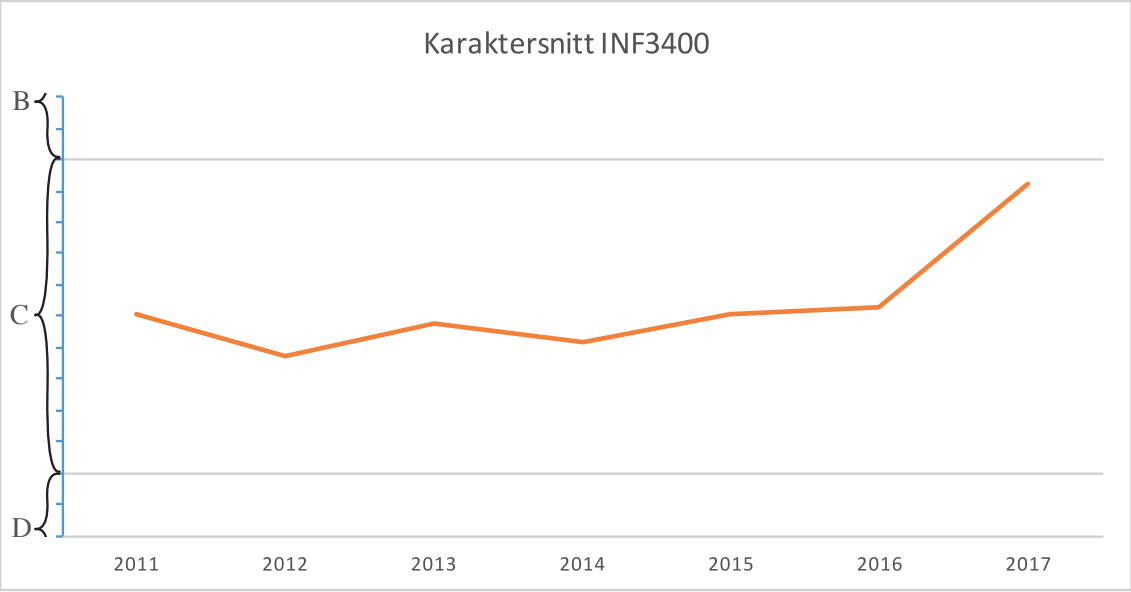
Figur 2: Karakterfordeling for avsluttende eksamen i henholdvis 2016 og 2017.

Gruppe	Snitt	Beste karakter	Laveste karakter
Prøveeksamen	B	A	D
Oblig 8	D	C	E

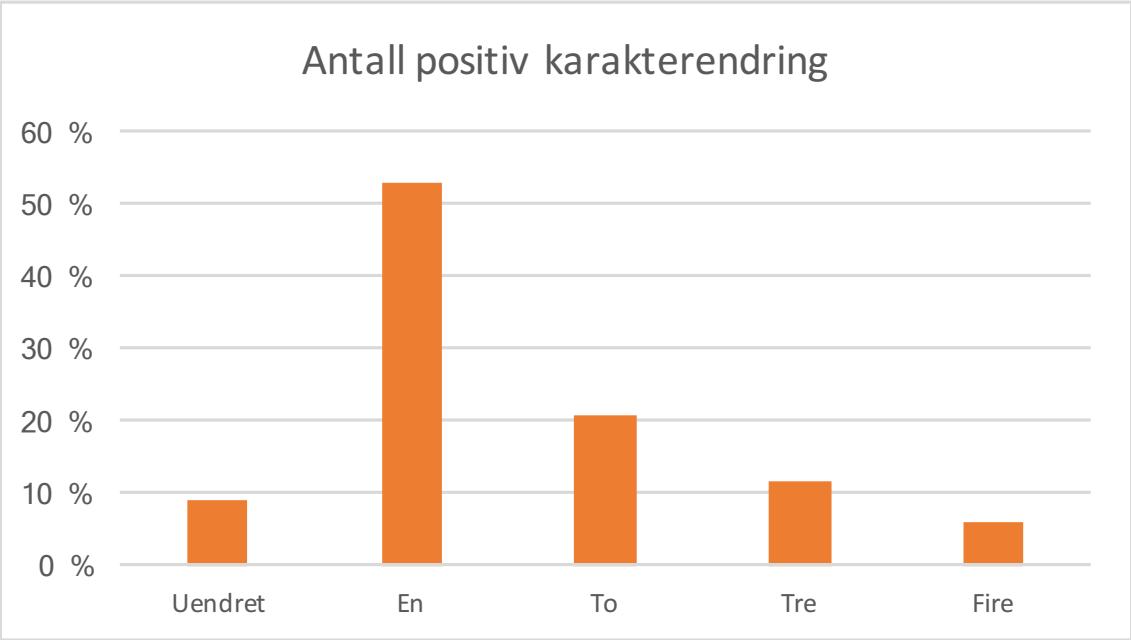
Tabell 1: Oversikt over hvordan de to ulike gruppene av studenter, de som valgte prøveeksamen og de som valgte tradisjonell oblig 8, gjorde det på endelig eksamen i 2017.

vesentlig bedre enn for den andre gruppen. Når vi ser på karakterspennet for de to ulike gruppene så fremkommer det at gruppen som gjennomførte tradisjonell oblig 8 ligger i nedre karaktersjikt. Dette er trolig også hovedgrunnen til at karakterfordelingen, slik vist i Figur 2, fremstår med et «klasseskille».

Vi kan sammenligne resultatet for prøveseksamen i 2017 og resultatet for avsluttende eksamen for 2016. For prøveeksamen i 2017 var studentenes snittkarakter midt på C. Dette var tilsvarende snittet for avsluttende eksamen i 2016 og tilsvarende for resultatene fra 2011 til 2016 som vist i Figur 3. Det har vært samme faglærer i dette emnet de siste 10 årene og således har også utarbeidelsen av eksamen og eksamensnivået vært meget jevnt og egnet for direkte sammenligning. Det er bare i det siste året at snittkarakteren har hatt en vesentlig økning mot B slik som grafen i Figur 3 viser. Vi har sammenlignet karakteren for de kandidatenes prøveeksamen mot endelig eksamenskarakter og funnet endringen. Det viser seg at ingen av studentene har hatt en negativ endring i karakteren mellom prøveeksamen og avsluttende ekasmen. I Figur 4 er det vist en grafisk oversikt over antall kandidater og deres representative karakterendring fra prøveeksamen til avsluttende eksamen. Vi ser at over halvparten av studentene har gått opp minst en karakter, mens i underkant av 10% har uendret karakter. Noen av studentene gikk opp tre eller fire karakterer og ved nærmere analyse viser det seg at dette har vært de som hovedsakelig har fått F eller E på prøveeksamen, mens de som fikk en karakterforbedring på en eller to karakterer i hovedsak har vært D og C kandidater på prøveeksamen. I nettskjema og



Figur 3: Karaktersnitt i INF3400 fra 2011 til 2017. Karaktersnittet har vært midt på C fram til og med 2026, mens det i 2017 var en markant økning opp mot B.



Figur 4: Oversikt over hvor mange som fikk en positiv karakterendring i forhold til prøveeksamen. Det var ingen som hadde negativ karakterendring fra prøveeksamen og til endelig eksamen.

intervjuene har det fremkommet at noen (antageligvis E og F kandidatene) møtte opp meget uforberedt og egentlig ikke anså sin egen besvarelse som verdig for en endelig eksamen.

5 Studentenes tilbakemelding

Etter gjennomføringen av prøveeksamen og i den strukturerte peer-reviewen ble alle deltakende studenter bedt om å fylle ut et nettskjema for evaluering av forsøket. Nettskjemaet hadde en svarprosent på 77% og ble utfylt innen en uke etter forsøk. Vi valgte å ha en inndeling i nettskjemaet for å fokuserer på evalueringen og tilbakemeldingen fra studentene. Nettskjemaets hadde følgende deler:

- Det å ha en prøveeksamen.
- Det å være deltaker av et peer-review fra et students perspektiv.
- Det å være en peer (sensor) for medstudenter.
- Evaluering av det tekniske opplegget (dataprogrammet, brukervennlighet og tidsbruk).
- Personlig profesjonell utvikling for studentene.
- Læringsutbytte for studentene som deltaker og sensor.

Gjennom intervjuene ble nettskjemaet også diskutert. Siden nettskjemaet var anonymisert har vi ikke hatt mulighetene til å koble nettskjema med intervjuene. I de neste avsnittene går vi igjennom tre av hovedgruppene for tilbakemelding: (1) Studentenes tilbakemelding på det å være deltakere i en peer-review forsøk, (2) studentenes tilbakemelding på det

Antall	Type justering
20%	Ingen justering
22%	En besvarelse, noen oppgaver
55%	To-tre besvarelser, noen oppgaver
3%	Alle besvarelser, noen oppgaver

Tabell 2: Studenenes tilbakemelding om hvor mye justering av poengsum som ble gjort i de besvarelsene de har sensurert gjennom programmet.

å være sensor/peer for medstudenter og til slutt (3) studentens personlige utvikling og utbytte i dette forsøket.

Peer-review som student

Det er ulike elementer og faktorer som har betydning for studentenes opplevelse av forsøket, hvorvidt studentene har hatt nok tid til å repetere pensum, sette seg i «eksamensmodus» eller motivere seg er vanskelig å måle. I avsnittet om metode har vi gjennomgått mye av det praktiske rundt prøveeksamen. Vi har forsøkt å evaluere studentenes inntrykk og opplevelse av å gjennomføre en slik prøveeksamen. Majoriteten, 78%, av studentene svarte meget positivt på å ha en slik prøveeksamen med peer-review. Noen ønsket og etterlyste muligheter for å ha flere slike tiltak, jmfr Sitat:1.

> *«Jeg synes man kan erstatte noen obliger med noen prøveeksamener eller eksamenslignende oppgaver til hva man har gjennomgått så langt i semesteret.»*
>
> *(*Student #1803025*)*

Sitat:1

Det var ikke mye som tydet på at følte seg presset eller ukomfortable i forhold til å ha hatt nok tid til forberedelser. Noe overraskende, som også beskriver den positive holdningen til studentene, er hva de synes om at medstudenter skulle rette deres besvarelser. Her kom det frem at 11% liker ikke at medstudenter retter, 22% synes det var greit, mens 67% synes det er bra. En overveiende positiv holdning til et kollektivt læringsløft for kullet.

Peer-review som sensor

Det er ikke alltid at sensurerings programmet klarer å nyansere poeng for detaljer i besvarelsen og det er derfor sensorer har muligheten til å overstyre poenggivningen for hver enkelt deloppgave. Det er interessant å finne ut om studentene var enige i de automatiske poengtrekkene, samt hvor mye de eventuelt justerte poengene. Alle justeringer som sensor gjør blir lagret slik at det kan hentes ut for analyse og kalibrering. På spørsmålet om studentene har justert oppgavene fikk vi fordelingen som vist i Tabell 2, årsakene til hvorfor de endret poeng var hovedsakelig fordi de følte poengtrekkene var litt høye. Dette tyder kanskje på at studentene egentlig er litt «snillere» sensorer enn det faglærer i utgangspunktet ønsket, jmfr Sitat:2.

«Jeg følte at noen gjorde en bra jobb og viste mye kompentanse men hadde noen små feil og da synes jeg den streken gikk altfor langt ned aå jeg dro den litt opp på flere oppgaver jeg rettet, for jeg synes at kandidaten fortjente det.» (Student #1800030)

Sitat:2

Hele 93% mener de har hatt utbytte av å rette medstudentenes prøveeksamen. Fra de utvalgte sitatene, Sitat:3–4, får vi inntrykk av hva studentene fikk mest utbytte av.

«Bra å se fremgangsmetode til andre studenter og se forskjellige måter å løse en oppgave på.» (Student #1799359)

Sitat:3

«Lærer av andre sine feil, lettere å se andre sine feil enn sine egne, må sette seg godt inn i prøven får å kunne sensurere på en god måte.» (Student #1830343)

Sitat:4

For å finne ut av påvirkningen sensureringsprogrammet har hatt på studentenes opplevelse av sensurering har vi vært opptatt av brukervennligheten til programmet, samt stabiliteten. Dette er fordi vi vet at utfordringer på de to nevnte punktene kan redusere det faglig utbytte i dette forsøket. Resultatene er vist i tabell 3.

	Svært bra	Bra	Passe	Dårlig	Svært dårlig
Brukervennlighet	14,8%	44,4%	37,0%	3,7%	0,0%
Hvor stabilt er programmet?	18,5%	44,4%	18,5%	11,1%	7,4%

Tabell 3: Studentenes tilbakemelding på brukervennlighet og stabilitet på sensureringsprogrammet.

Personlig utvikling

Vi stilte spørsmål om studentene etter å ha tatt prøveeksamen og rettet medstudenters besvarelse sitter igjen med en bedre forståelse for læringsmålene for emnet? Til dette svarte 97% at de har fått bedre forståelse enn de hadde. Noen utvalgte sitater 5 og 6 gir et godt bilde av studentenes personlig utvikling i dette forsøket.

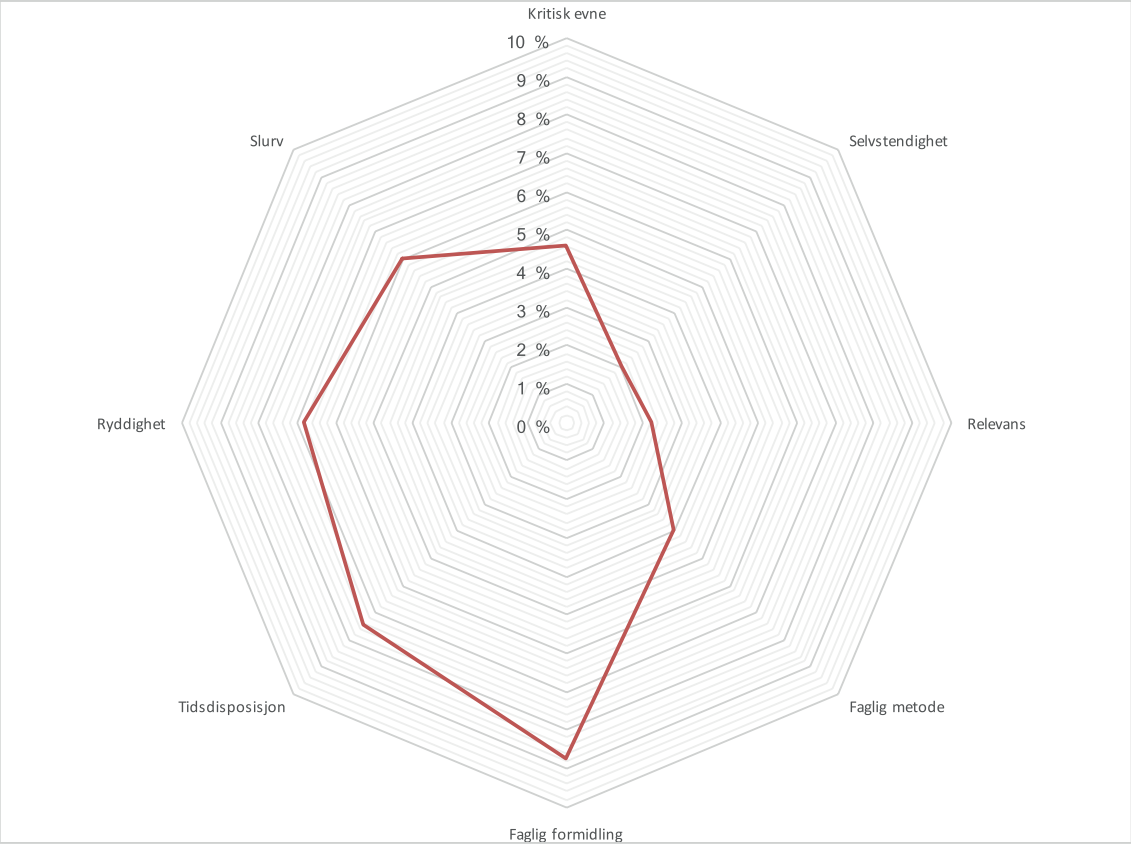
«Prøveeksamen gir meg en oversikt over hva jeg bør øve på og dekke manglende kunnskap i det fagfeltet. Selv om jeg kan det grunnleggende, så ga prøveeksamen et innblikk i at jeg må øve å trekke frem elementer og koble forbindelser mellom dem.» (Student #1799316)

Sitat:5

«Jeg fikk større oversikt over hva som er pensum, og forståelse på grunn av at jeg har regnet på følgefeil osv i andres besvarelser.» (Student #1806936) Sitat:6

6 Tilbakemelding til faglærer etter sensur

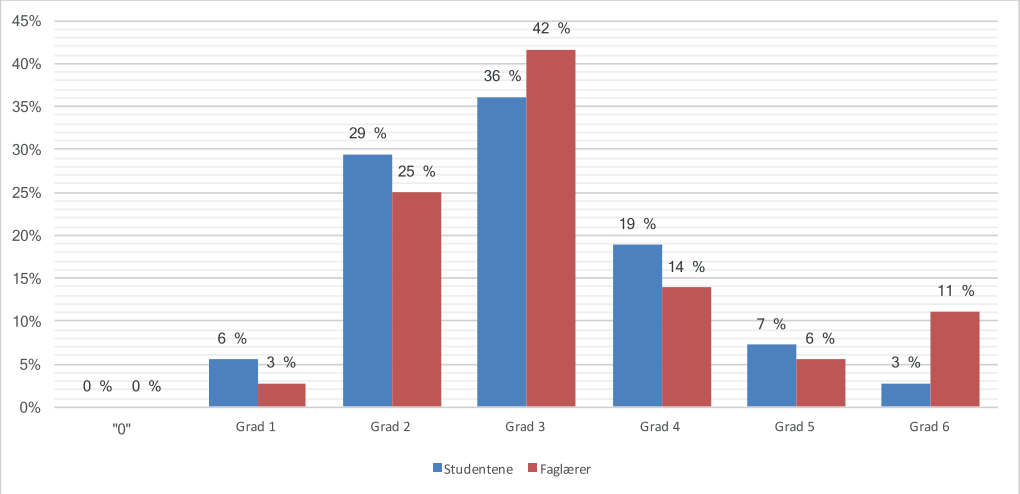
Datagrunnlaget som ble samlet inn var tilsvarende sensureringen av 175 besvarelser. Innsamlingen er ikke normalisert mot sensorer eller kandidater og rådataene som er analysert viser en del interessante trender som er nyttig for faglærer å ta med seg til neste gang emnet skal undervises. Ved å ekstrahere ut kvantifiserte resultater for de ulike kriteriene vil rapporten gi en pekepinn til faglærer om (1) hvordan studentene har gjort det på prøveeksamen og (2) hva har studentene har lært gjennom å være en sensor. Før vi presenterer disse tilbakemeldingene til faglærer er det viktig å se hvordan studentenes sensurering har vært i forhold til faglærerens sensurering, som en form for verifisering av kalibreringen av studentenes sensur. I radardiagrammet vist i Figur 5 ser vi på den prosentvise differansen for de ulike kriteriene mellom faglærers sensurering



Figur 5: Radardiagram som viser prosentvis forskjell mellom studentens sensurering og faglærernes sensurering for de ulike kriteriene.

og studentenes sensurering. Som diagrammet viser er alle kriteriene godt innenfor en differanse på 10%, og for halvparten av kriteriene er den under 5%. Dette er et meget godt resultat som bekrefter at sensureringsprogrammet vil selv med 35 sensorer gi en

rettferdig vurdering av en besvarelse. Det største avviket er i kriterene er faglig formidling og det er ikke uventet. Hovedgrunnen til dette er nok at studentene ikke helt klarer å skille mellom gradene av formidlingen. Vi har valgt ut to kriterier for å illustrere forskjellen i de ulike gradene. I Figur 6 vises fordelingen mellom studentene og faglærerne fordelt på de



Figur 6: Detaljert oversikt over fordelingen av grader av «feil» med hensyn på kriteriet om bruk av faglig metode. Studentenes sensur er summert i blått, mens faglærernes i rødt.

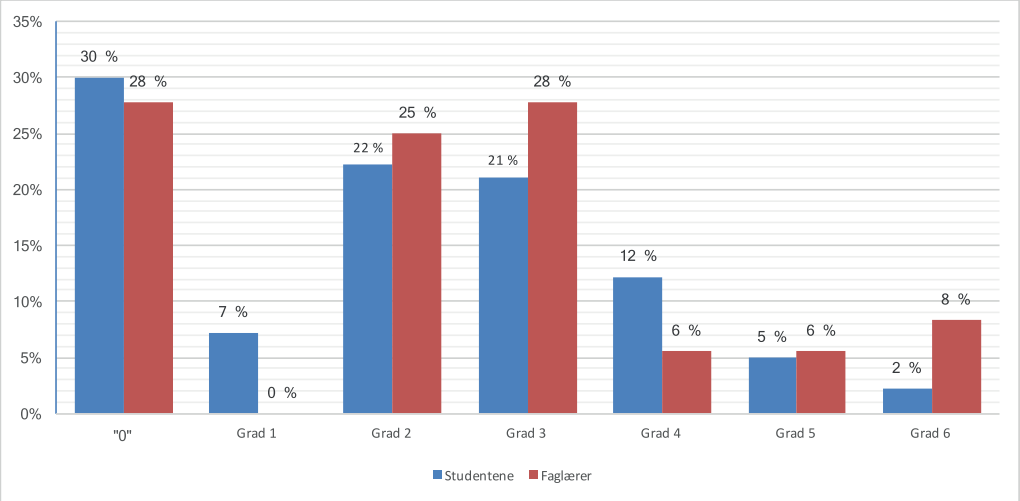
ulike gradene av «feil». Tallene representerer den akkumulerte verdien normalisert mot totalt antall tilbakemeldinger for hele datasettet. Det betyr med andre ord den prosentvis andel av studentene som har fått en gitt grad av tilbakemelding for dette kriteriet. De ulike gradene er basert på en dynamisk terskel som endrer seg basert på kullets relative verdi. Betydningen av «0» er den gitte besvarelse har vist til faglig metode og at det ikke er noe å utsette på bruk eller anvendelsen av den. Grad 1 referer til at det er enkelte mangler ved bruk av faglig metode som forekommer i besvarelsen. Grad 2 har betegnelsen «noe manglende», mens grad 3 har en besvarelse manglende bruk/anvendelse av faglig metode. Grad 4 har vesentlig manglenr, mens grad 5 har grunnleggende mangler og grad 6 har besvarelsen totalt fraværende av faglig metode. I figuren, Figur 6 kan vi se at 36% av studentene har fått en grad 3 i tilbakemelding om faglig metode fra sine medstudenter. Stort sett er det liten forskjell mellom studentene og faglærer på de ulike gradene. Det eneste som skiller seg vesentlig ut er grad 6. Dette henger mye sammen med at studentene gjennomgående vegrer seg mot å være for negativ. Ved gjennomgang av alle kriteriene ser vi at de største forskjellene er mellom bruken av grad 1 og grad 6. Det er nå slik at alle oppgaver blir vurdert for alle kriterier, men naturligvis har noen oppgaver et større utslag for noen kriterier enn andre. I Figur 7 er detaljene og fordelingene vist for kriteriet om kritisk evne. Kritisk evne går ut på studentenes evne til å være kritisk til sin egen besvarelse. Fra denne figuren ser vi at mange studenter har fått «0», som betyr at de har vist kritisk evne. Igjen ser vi de største forskjellene på grad 1 og 6 mellom studentene og faglærne.

Basert på de ovennevnte grafene og kalibreringen av studentenes sensur og faglærers seunsur kan vi presentere resultatene av de 175 besvarelsene for viktig tilbakmelding til faglærer. Foruten alle kommentarer i fritekst og oversikt over begrunnelser, baserer vi videre presentasjon på kvantifisert datamateriale. I Figur 8 vises et konturplott av mengden med tilbakemelding for de ulike kriteriene for hele kullet akkumulert og fordelt

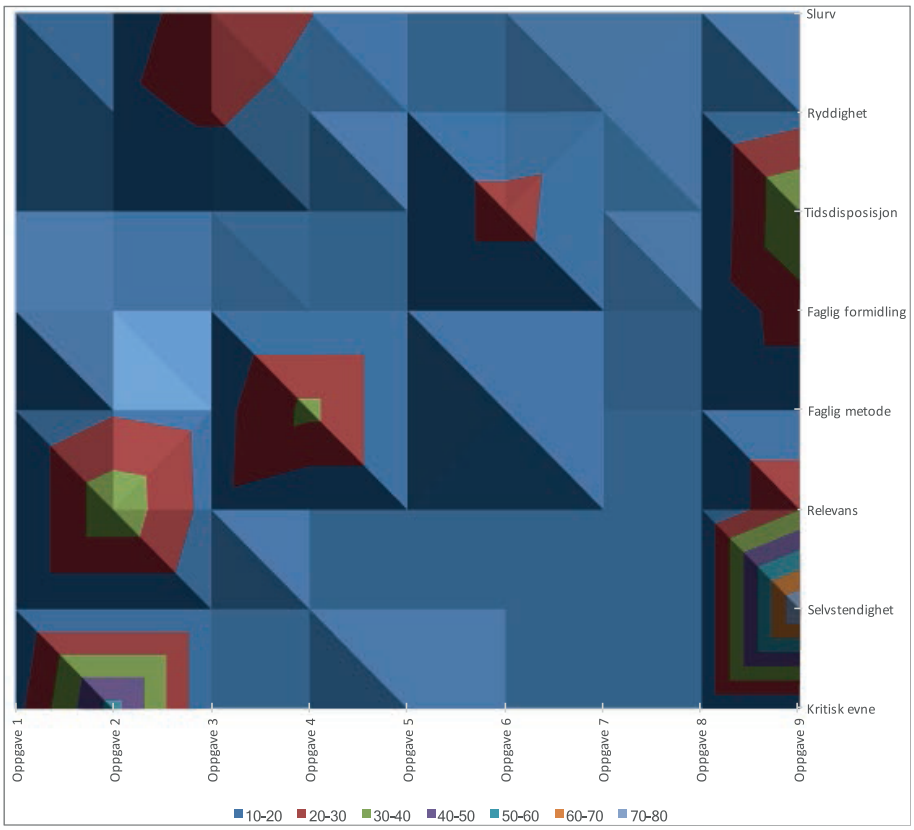
på hver oppgave. Av figuren ser vi at den største toppen i plottet er for oppgave 9 når det gjelder selvstendighet. Oppgave 9 var å skrive en tekst om et faglig tema i pensum. Det er helt tydelig at studentene har vist svært liten selvstendighet fordi de fleste har skrevet rett av foiler og/eller pensumslitteraturen. Dette plottet gir en faglærer rask overblikk over hva studentene trenger å jobbe med i emnet. Fra figuren kan vi lese ut følgende toppunkter: Slurv i oppgave 3, tidsdisposisjon i oppgave 6 og 9, manglende faglig metode i oppgave 4, lite relevans i oppgave 2 og 9, lite selvstendighet i oppgave 9 og kmanglende ritisk evne i oppgave 2. En faglærer som har forfattet eksamenssettet vet hva de ulike oppgavene krever og kan bruke plottet finne ut hvordan adressere de ulike punktene. Selvom noen oppgaver har fått høy tilbakemeldingsprosent betyr det ikke nødvendigvis at oppgaven har vært vanskelig eller at studentene har skåret lavt i poeng. Eksempelvis i oppgave 9 betyr det ikke nødvendigvis at studentene har fått lav karakter for ren avskrift fra foiler. Dette bringer oss over i neste figur som viser karakterfordelingen for hver oppgave. I Figur 9 er karaktertersklene satt på bakgrunn av akkumulert poeng med statisk terskelverdi, i hovedsak for å gruppere inn besvarelsene. Denne informasjonen kan faglærer bruke for raskt å finne ut av de oppgavene hvor studentene har oppnådd høy poeng/karakter og de oppgavene som tilsynelatende fremstår som «vanskelige».

7 Diskusjon

I dette forsøket har vi gitt et tilbud til studentene å velge mellom å ta en prøveeksamen med tilhørende peer review eller en tradisjonell obligatorisk oppgave. Sett i forhold til estimert tidsbruk og faktisk tidsbruk for studentene har dette vært nokså jevnt. Det kunne vært lagt mer vekt på forbredelser tidligere i semesteret for å styrke studentenes evne til å gjennomføre skriflige tilbakemeldinger, spesielt for å isolere utfordringen med usikkerheten for studentene som gjennomførte peer review, slik 11% av studentene påpeker. Det finnes mange emner på universitetet som har obligatoriske skriveøvelser hvor man retter hverandres arbeid, men dessverre er dette lite utbredt i informatikk og spesielt i elektronikk. Forskning viser at det å ha gode spørsmål og veiledninger



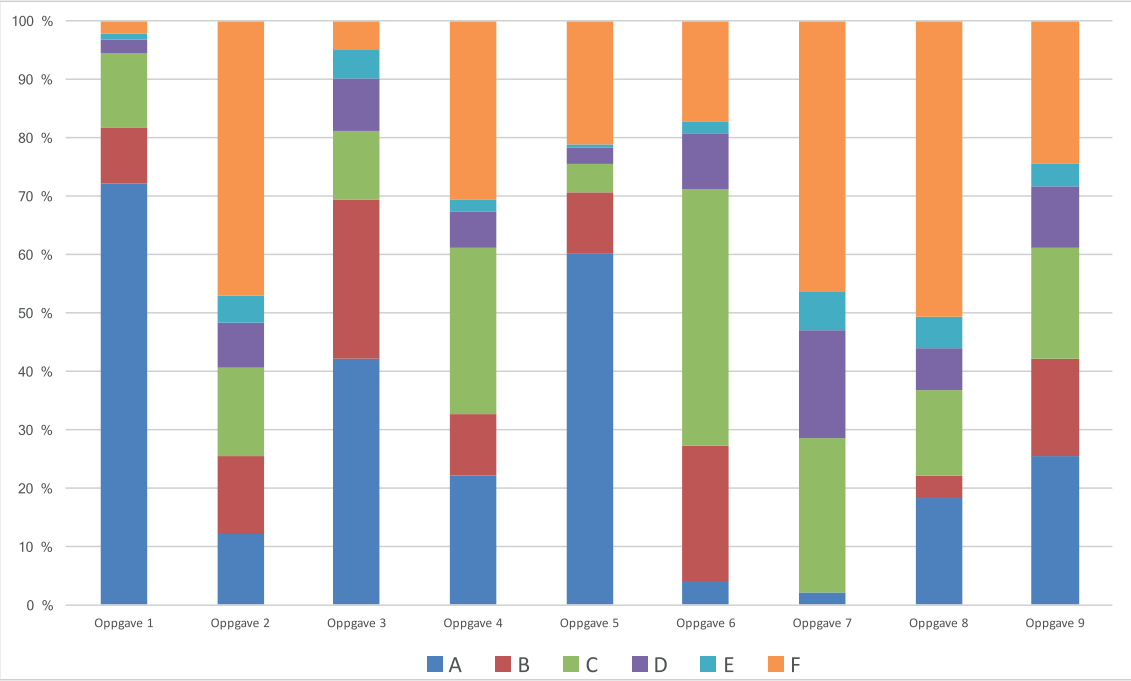
Figur 7: Detaljert oversikt over fordelingen av grader for «feil» med hensyn på kriteriet om bruk av kritisk evne. Studentenes sensur er summert opp i blått, mens faglærernes er i rødt.



Figur 8: Konturplott som viser mengden av tilbakemelding etter sensur fordelt på kriterier og oppgavene i eksamenssettet.

for prosessen igjennom en peer review prosess er avgjørende suksessfaktorer. Det er derfor vi har utviklet et eget program som guider sensor igjennom en retteprosess og tilrettelegger for egen refleksjon. Ved bruk av programmet påpeker studentene økt forståelse av læringsmålene i emnet, fordi disse tydelig fremkommer som kategoriske elementer i begrunnelsesrubrikken. Videre ser de også direkte sammenhengen og utslaget med hensyn på poeng/grad av «feil». Studentene rapporterer at de også nå har en større innsikt i hva de blir vurdert på, som igjen har gitt dem et perspektiv og «rød tråd» i pensum. Gjennom samtalene med studentene og deres tilbakemeldinger på nettskjema sitter vi igjen med en klar formening om at vi, gjennom peer review av en prøveeksamen rettet med sensureringsprogrammet, har en høyere grad av samhandling, som igjen har ført til at studentene har en dypere forståelse av læringsmålene. Gjennom samtaler med studentene fremkom det også en økt interesse for faget som et biprodukt av denne øvelsen. Dette i seg selv er en høy motivasjonsfaktor for studentene å jobbe mer med faget i forkant av en eksamen.

Gjennom de kvalitative analysene om studentenes resultater ser vi at den største effekten har vært på eksamensresultatene. Emnet har jevnt over de siste årene hatt en snittkarakter midt på C, mens det i år har hatt et vesentlig økning mot B. Videre ser vi av karakterfordelingen på endelig eksamen og karakterfordelingen på prøveeksamen at den økte snittkarakteren hovedsakelig skyldes gruppen som har hatt prøveeksamen. Vi viser til en «klasseskille» i karakter mellom grupperingen som har hatt prøveeksamen og de som valgte obligatorisk oppgave. Sensureringsprogrammet for prøveeksamen har vært den samme som for det forrige års eksamenssensur og ble også brukt for årets sensur.



Figur 9: Karakterfordelingen, for hele studentmassen, for hver oppgave i eksamensettet.

Ved sammenkobling av disse tre datasettene finner vi ut at studentenes besvarelser av prøveeksamen har vært noe lavere enn fjorårets kull med hensyn på karakter, men i stor sett tilsvarende når det kommer til gradering og fordeling av kriteriene. Sammenligner vi grad og fordeling av kriteriene mellom prøveeksamen og endelig eksamen i 2017, finner vi en forbedring spesielt med hensyn til de «toppene» som er vist i Figur 8.

Forskning viser at det er mest hensiktsmessig å unngå at studentene gir hverandre karakter. Sensureringsprogrammet har derfor ikke noe antydning til karakterer som A, B, C, D, E eller F. Det eneste studentene kan gjøre er å gi en verdi på en skala fra ikke bestått til fremragende for hver oppgave. Det er ikke studentene som setter karakter, de angir bare verdien og mengden av «feil». Det er nettopp denne måten å sensurere på som tilgjengeliggjør en kvantifisering av sensurering og programmet kan ekstrahere sensorenes skjønnsmessige vurdering av oppgaven. Siden dette har vært et forsøk gjennomført sent i emnet valgte vi å ikke sende ut medstudentenes tilbakemeldinger til hverandre, fordi vi var usikre på hvordan systemet ville kalibrere et stort antall sensureringer. Spesielt ønsket vi å hindre at studentene fikk feilaktige tilbakemeldinger. Kalibreringen viste seg å bli veldig god som vist i Figur 5. Med et utfyllende datagrunnlag, 175 besvarelser, vil faglærer få innblikk i hvilke deler av pensum studentene har utfordringer i og hva dette skyldes. Faglærer får gjennom henholdsvis Figur 8 og Figur 9 et oppslagsverk som han/hun kan bruke for å få en god oversikt over studentenes kunnskapsmangler.

Avslutningsvis mener vi at det dedikerte sensureringsprogrammet har vist seg å ha et godt brukergrensenitt, stabilt og ikke minst godt kalibrert på tvers av 35 sensorer. Som faglærer gir programmet også innblikk i et datamateriale som ikke har vært tilgjenglig tidligere og som kan brukes til å forbedre emnet for fremtidige studenter.

Videre arbeid

I vårt forsøk har vi i første omgang fokusert på å gi studentene en alternativ læringsmetode gjennom å gi formative tilbakemeldinger. For dette har vi brukt et sensureringsprogram som geleider studentene gjennom en vurderingsprosess bestemt av faglærer. Det ble samlet store mengder med data i dette forsøket. Ved å arbeide videre med denne datamengden kan vi få informasjon om hva studentene kan og deres forståelse av pensum når de gjennomførte sensur og krysskoble det med deres egen besvarelse på endelig eksamen. På den måten kan vi finne ut av studentenes progresjon og kunne tallfeste effekten av en slik formativ vurdering og tilbakemelding. Det er også viktig å gjennomføre flere slike forsøk både i flere emner og over flere år for å kunne validere datasettene og filtrere kullspesifikke påvirkninger.

Referanser

[1] D. Hounsell, V. McCune, J. Hounsell, and J. Litjens. The quality of guidance and feedback to students. *Higher Education Research & Development*, 27(1):55–67, 2008.

[2] F. Dochy, Mien Segers, and Dominique Sluijsmans. The Use of Self-, Peer and Co-assessment in Higher Education: a review. *Studies in Higher Education*, 24(3):331–350, 1999.

[3] David Boud. Sustainable Assessment: Rethinking assessment for the learning society. *Studies in Continuing Education*, 22(2):151–167, 2000.

[4] Ngar-Fun Liu and David Carless. Peer feedback: the learning element of peer assessment. *Teaching in Higher Education*, 11(3):279–290, 2006.

[5] Edward F. Gehringer. Electronic peer review and peer grading in computer-science courses. *ACM SIGCSE Bulletin*, 33(1):139–143, 2001.

[6] K.J. Topping. Peer assessment between students in colleges and universities. *Review of Educational Research*, 68(3):249–276, 1998.

[7] Jennifer L. Docktor, Jay Dornfeld, Evan Frodermann, Kenneth Heller, Leonardo Hsu, Koblar Alan Jackson, Andrew Mason, Qing X. Ryan, and Jie Yang. Assessing student written problem solutions: A problem-solving rubric with application to introductory physics. *Physical Review Physics Education Research*, 12(1):010130, 2016.

[8] Raoul Mulder, Chi Baik, Ryan Naylor, and Jon Pearce. How does student peer review influence perceptions, engagement and academic outcomes? A case study. *Assessment & Evaluation in Higher Education*, 39(6):1–21, 2013.

[9] Winnie Cheng and Martin Warren. Having second thoughts: Student perceptions before and after a peer assessment exercise. *Studies in Higher Education*, 22(2):233–239, 1997.

[10] Richard Higgins, Peter Hartley, and Alan Skelton. Getting the Message Across: The problem of communicating assessment feedback. *Teaching in Higher Education*, 6(2):269–274, 2001.

[11] P. Conaghan and A. Lockey. Feedback to feedforward. *Notfall + Rettungsmedizin*, 12(S2):45–48, 2009.

[12] D. Royce Sadler. Beyond feedback: developing student capability in complex appraisal. *Assessment & Evaluation in Higher Education*, 35(5):535–550, 2010.

[13] David John Baker and Danielle Zuvela. Feedforward strategies in the first-year experience of online and distributed learning environments. *Assessment & Evaluation in Higher Education*, 2938(March 2015):1–11, 2012.

[14] John Biggs and Cathrine Tang. *Teaching for Quality Learning at University Third Edition Teaching for Quality Learning at University*, volume 3th edition (1th edition 1999). Open University Press, 2007.

[15] Omid Mirmotahari and Yngvar Berg. Individuell «automagisk» tilbakemelding på skriftlig eksamen. In *Artikkelsamling MNT-konferansen*, pages 109–114, mars 2017.