En eksempel-studie av automatisk tilbakemelding for programmeringsfag i høyere utdanning.

Omid Mirmotahari, Kristian A. Hiorth og Stein Gjessing
Institutt for Informatikk, Universitetet i Oslo
Ester Fremstad og Crina Damsa
Institutt for Pedagogikk, Universitetet i Oslo

Sammendrag

Denne studien er gjennomført i forbindelse med eksamen i et førsteårs programmeringsemne ved et universitet. Vi har undersøkt hvordan vi ved bruk av et egenutviklet verktøy for sensurering som består av kvantitativ vekting og som samtidig genererer kvalitative tilbakemeldinger til studentene forbedrer kvaliteten på sensur og på eksamensoppgavene, og ikke minst støtter studentenes videre læring. Gjennom analyse av kvantitative data generert av systemet, samt kvalitative data fra involverte parter fremkommer tydelig forbedringer i validitet og reliabilitet i sensur, samt positive erfaringer fra sensorer og faglærere og ikke minst fra studentene som opplever at det å få tilbakemelding på eksamen bidrar til deres læring og utvikling. Vi vil i denne artikkelen beskrive prosessen med å utviklingen dette systemet for sensur og automatisk tilbakemelding, og presentere resultatene sett fra faglærere, sensorenes og studentenes perspektiv.

Keywords: eksamenssensur, automatisk tilbakemelding, førsteårs emne, store programmeringsfag, objektorientert programmering, høyere utdanning.

1 Introduksjon

Studier har vist at det å få tilbakemeldinger har stor påvirkning på læringsutbyttet. Selv om man i løpet av de siste par tiår har sett en økende grad av studentsentrering i høyere utdanning, har vi likevel ikke sett tilsvarende endring i fokus når det gjelder vurdering. Snarere ser det ut til at tilbakemeldingspraksiser i stor grad fremdeles ses som ovenfra og ned overføring av informasjon kontrollert av lærer. I [Tee and Ahmed, 2014] påpeker forfatterene at dette er problematisk fordi det ignorerer måten tilbakemelding interagerer med studentenes selvforståelse og motivasjon, og fremhever betydningen av å aktivisere studenten og å bruke både lærervurdering, selvvurdering og medstudentvurdering. Med relativt enkle grep tidlig i studiet, der studentene får tilbakemelding på sine prestasjoner, kan de både få støtte til egen læring og bli bedre på å formidle sine kunnskaper, ferdigheter og forståelse på eksamen. Nettopp fordi tilbakemelding er viktig for å utvikle metakognisjon og etablere gode studievaner og studie- og eksamensteknikk er det viktig å adressere dette for de aller ferskeste studentene. Vi har derfor i denne studien valgt å gi alle studenter på dette førsteårsemnet individuell tilbakemelding på hvordan de besvarte oppgavene og hva de burde jobbe videre med, og utviklet et digitalt verktøy for dette som automatisk generer tilbakemelding til studentene, basert på vurderingen av besvarelsene.

Forskning tyder på at en helhetlig tilnærming til vurdering og tilbakemelding som bygger på sosio-konstruktivistiske prinsipper er det som gjør tilbakemelding mest produktiv i den forstand at den støtter opp under studentenes læring [Boud and Molloy, 2013, Esterhazy and Damşa, 2017, Juwah et al., 2004]. Dette betyr at vurdering og tilbakemelding ses på som del av en pågående utviklingsprosess, der studenten ses som aktiv deltager, og at summativ og formativ vurdering ses som del av hele læringsprosessen slik at vurdering og tilbakemelding støtter opp under læringsmålene [Rust, 2002]. Dette gjør også at studentene lærer mer av eksamen, dvs. at eksamen ikke bare blir en vurdering *av* læring, for å finne ut i hvilken grad studentene har nådd læringsmålene, men også en vurdering *for* læring, i form av at tilbakemeldingene bidrar til at læring i faget fortsetter. Det å gi tilbakemeldinger har det vært mange ulike forsøk på, også noen få innen automatisering av tilbakemeldingene [Jiménez-gonzález et al., 2008, Malmi and Korhonen, 2004, Mirmotahari and Berg, 2017, Siddiqi et al., 2010, Thelwall, 2000].

Reliabilitet og ulike former for validitet er utfordringer som er sentrale i forbindelse med eksamen og vurdering. Studier viser at sensorer vurderer eksamensbesvarelser svært ulikt [Raaheim, 2000]. Forfatteren [Raaheim, 2000] peker i denne sammenheng på at mangelen på kriterier og sensorveiledning er sentrale for å forklare mangelen på reliabilitet. Sentralt i det sensurverktøyet som er utviklet og tatt i bruk i denne studien [Mirmotahari and Berg, 2017] er utvikling og bruk av klare kriterier og sensorveiledninger. En stor del av arbeidet med å etablere bruken av dette verktøyet i forbindelse med hver eksamen er nettopp utviklingen av kriterier, måleskalaer og vekting. Når studentene får innblikk i disse kriteriene, hjelper det dem å forstå hva som kreves av dem i en hittil ukjent akademisk setting, og bidrar til å skape transparens og et ikke-truende læringsmiljø [Rust, 2002]. En slik transparens er en av de mest sentrale elementene i «constructive alignment» [Biggs and Tang, 2007]. En slik transparens, kombinert med automatiske kvalitative tilbakemeldinger kan bidra til å redusere behovet for begrunnelser og også klager på sensur, fordi studentene gis innsikt i hva de har gjort riktig og hva de har gjort feil.

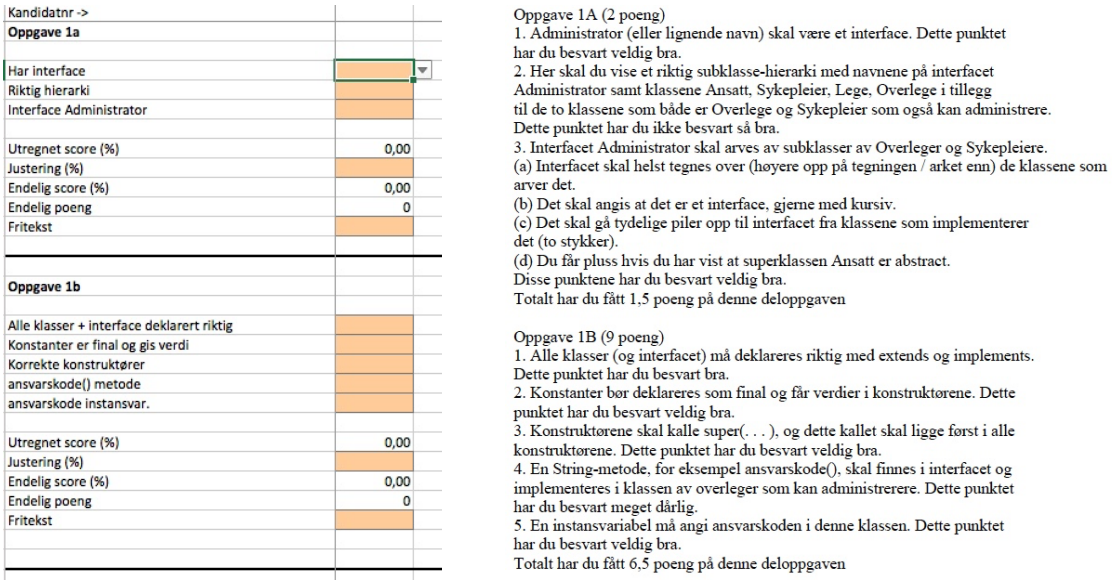
Denne studien belyser, med utgangspunkt i bruk av dette verktøyet i et introduksjonsemne (INF1010) i objektorientert programmering for førsteårsstudenter på institutt for Informatikk ved Universitetet i Oslo, hvordan verktøyet for sensur og automatisk tilbakemelding kan

(a) bidra til et mer samstemt forhold mellom vurdering og undervisning
(b) styrke validitet og reliabilitet av sensur
(c) redusere sensorenes tidsforbruk
(d) generere data som gir faglærer verdifull informasjon

Resten av denne artikkelen er bygget opp som følger: i neste avsnitt, 2, gir vi en kort beskrivelse av det egenutviklede sensureringsprogrammet. Deretter, i avsnitt 3, beskriver vi forskningsmetoden for studien. I avsnitt 4 presenterer vi ulike analyser av innsamlede datamaterialet og resultatene med tilknyttede kvalitative studier og erfaringer fra faglærere, sensorer og studenter. I avsnitt 5 konkluderer vi denne studien.

2 Dedikert sensureringsprogram

Oppgavesettet for eksamen våren 2017 ble utviklet slik at i hver deloppgave kunne studentene vise hvor godt de behersket et eller flere (stort sett relaterte) læringsmål. Hele oppgavesettet ble utviklet slik at flest mulige sentrale læringsmål var dekket. Hovedsakelig basert på arbeidsmengden fikk hver av de 16 deloppgavene en vekt slik



Kandidatnr ->	
Oppgave 1a	
Har interface	
Riktig hierarki	
Interface Administrator	
Utregnet score (%)	0,00
Justering (%)	
Endelig score (%)	0,00
Endelig poeng	0
Fritekst	
Oppgave 1b	
Alle klasser + interface deklarert riktig	
Konstanter er final og gis verdi	
Korrekte konstruktører	
ansvarskode() metode	
ansvarskode instansvar.	
Utregnet score (%)	0,00
Justering (%)	
Endelig score (%)	0,00
Endelig poeng	0
Fritekst	

Oppgave 1A (2 poeng)
1. Administrator (eller lignende navn) skal være et interface. Dette punktet har du besvart veldig bra.
2. Her skal du vise et riktig subklasse-hierarki med navnene på interfacet Administrator samt klassene Ansatt, Sykepleier, Lege, Overlege i tillegg til de to klassene som både er Overlege og Sykepleier som også kan administrere. Dette punktet har du ikke besvart så bra.
3. Interfacet Administrator skal arves av subklasser av Overleger og Sykepleiere.
(a) Interfacet skal helst tegnes over (høyere opp på tegningen / arket enn) de klassene som arver det.
(b) Det skal angis at det er et interface, gjerne med kursiv.
(c) Det skal gå tydelige piler opp til interfacet fra klassene som implementerer det (to stykker).
(d) Du får pluss hvis du har vist at superklassen Ansatt er abstract.
Disse punktene har du besvart veldig bra.
Totalt har du fått 1,5 poeng på denne deloppgaven

Oppgave 1B (9 poeng)
1. Alle klasser (og interfacet) må deklareres riktig med extends og implements. Dette punktet har du besvart bra.
2. Konstanter bør deklareres som final og får verdier i konstruktørene. Dette punktet har du besvart veldig bra.
3. Konstruktørene skal kalle super(. . .), og dette kallet skal ligge først i alle konstruktørene. Dette punktet har du besvart veldig bra.
4. En String-metode, for eksempel ansvarskode(), skal finnes i interfacet og implementeres i klassen av overleger som kan administrerere. Dette punktet har du besvart meget dårlig.
5. En instansvariabel må angi ansvarskoden i denne klassen. Dette punktet har du besvart veldig bra.
Totalt har du fått 6,5 poeng på denne deloppgaven

(a) (b)

Figur 1: *(a) Utsnitt av skjermbilde fra programmet som sensorene bruker for å sensurere. Som det fremkommer er det for hver deloppgave ulikt antall delmål og det er for hver av disse delmålene sensorene gir 0 - 5 poeng (oransje felter). Programmet regner således ut automatisk total poengsum for den gitte oppgaven. Er sensor uenig kan de benytte seg av feltet for justering. (b) Utsnitt av den tekstlige tilbakemeldingen til en tilfeldig student for oppgave 1a og 1b.*

at summen ble 100 poeng. Vektingen mellom oppgavene ble kunngjort til studentene i oppgaveteksten på eksamen.

Eksaminator laget en liste på ett til seks læringsmål for hver deloppgave, som ble basis for sensorveiledningen og begrunnelsen til studentene. Basert på studentens besvarelse, skulle sensor bedømme hvor godt studenten hadde nådd disse målene. Her ble det brukt en seksdelt skala hvor 0 = manglende/fraværende, 1 = særdeles svak, 2 = svak, 3 = god, 4 = bra og 5 = særdeles bra. Læringsmålene innad i hver deloppgave ble så vektet, og programmet beregnet en poengsum for den gitte deloppgaven (avrundet til nærmeste halve poeng). Sensor får se denne poengsummen simultant og har mulighet til å justere poengsummen skjønsmessig i en egen post - «justeringsbolken».

Poengsummen per deloppgave (og den totale poengsum) ble oppgitt til kandidatene i tilbakemeldingen. Den tekstlige tilbakemeldingen for hver deloppgave ble hentet ut basert på hva de har oppnådd på hvert delmål. Tekstfrasene som ble satt inn i tilbakemeldingene var basert i all hovedsak på den utfyllende sensorveiledningen utarbeidet av eksaminator i forkant av eksamen. Både selve delmålene og formuleringene brukt for å beskrive disse ble valgt med tanke på at det skulle være naturlig å bedømme kandidatenes oppnåelse av disse på en diskret skala som beskrevet over, samt at bedømmelsen kunne settes inn i frasen og fortsatt gi god mening for leseren. Som vist i figur 1(b) inneholdt tilbakemeldingene en tekstlig beskrivelse av hvert delmål, etterfulgt av en tekstlig beskrivelse av bedømmelsen. Opprinnelig ble rapportverktøyet laget slik at det var mulig å generere forskjellige tekstfraser basert på hvordan en kandidat faktisk hadde oppnådd delmålene. Årsaken var at man oppdaget at noen deloppgaver ga opphav til flere riktige svar, som ikke var sammenfallende med den tiltenkte fasiten og dermed passet dårlig med frasene fra sensorveiledningen. Funksjonen virket slik at sensor kunne velge

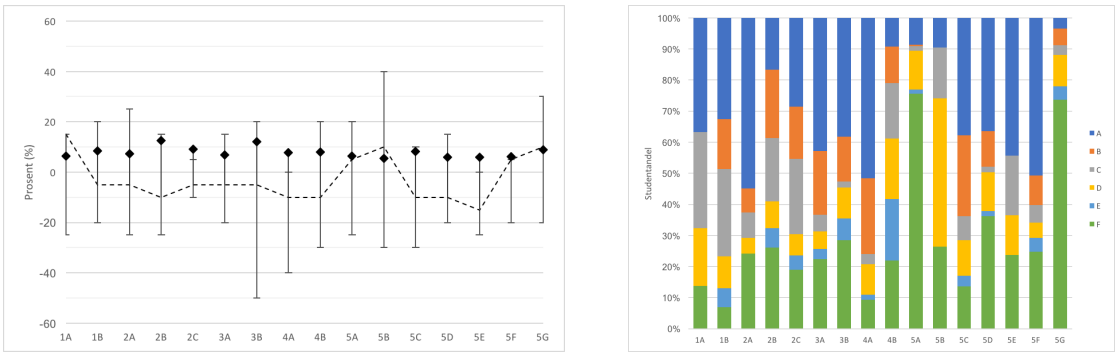
å bedømme én (og kun én) av flere bedømmelseslinjer for å angi hva slags løsning kandidaten hadde valgt. Selv om dette ble innarbeidet under selve sensurarbeidet og datagrunnlaget muliggjorde bruken av funksjonen, valgte vi allikevel å droppe dette i de endelige rapportene. Istedenfor ble de berørte frasene omskrevet for bedre å dekke flere tolkninger av deloppgaven. Årsaken var at beskrivelsene fortsatt stemte for dårlig overens med det studentene hadde besvart, og at det kunne virke forvirrende dersom studentene sammenlignet sine tilbakemeldinger. I virkeligheten var nok disse deloppgavene formulert på en lite egnet måte for denne formen for vurdering.

Studentene fikk således ikke tallverdiene for hva de har oppnådd på delmålene, men heller en skriftlig tilbakemelding som representerer denne verdien. Heller ikke den interne vektingen av delmålene innad i deloppgavene ble gjort kjent for studentene. For hvert delmål hadde sensor en mulighet til å ikke sette noen verdi. Da ville beskrivelsen av hvor godt kandidaten nådde dette målet ikke bli med i tilbakemeldingen. Sensor hadde også en mulighet til å skrive inn en fritekst som ble gitt i tillegg til studentene i tilbakemeldingen.

Selve rapporten ble generert av et egetutviklet, enkelt Python-program som benyttet en LaTeX-mal kombinert med en datamodell av eksamensoppgaven. På denne måten kunne vi enkelt implementere tilpassede funksjoner som å skjule irrelevante tilbakemeldinger ved uteblitt svar. Denslags funksjonalitet er overraskende vanskelig å få til ved bruk av hyllevareløsninger som for eksempel «Mail Merge»-funksjonen i Word, derav valget om å utvikle egen programvare. Studentene fikk til slutt tilsendt sin individuelle rapport til hver sin studentepostadresse. Selve overføringen av data mellom sensureringsverktøyet (basert på Excel) og rapportverktøyet innebar noe manuell sammenstilling av de forskjellige sensorenes data. I den forbindelse erfarte vi at det kan oppstå problemer, særdeles knyttet til fritekstkommentarer. Dessverre viser det seg at Excel 2016 fortsatt enkoder tekst med forskjellige tegnsett på f.eks. Mac og Windows, og at den innebygde dataeksportfunksjonen ikke nødvendigvis tar høyde for at celler i regnearket kan inneholde samme tegn som benyttes som separatortegn i de eksporterte filene.

3 Metode

Denne studien ble gjennomført for 528 studenter som avla eksamen i emnet INF1010 - Objektorientert programmering, våren 2017. Emnet inngår i andre semester som obligatorisk emne for alle studentene tatt opp til studieprogrammene ved institutt for Informatikk ved Universitetet i Oslo. Emnet har vært undervist av samme faglærer de siste 12 årene og har i grove trekk beholdt samme innhold i denne perioden. Undervisningen i emnet strekker seg over 14 uker med fire timer forelesning, to timer gruppeundervisning og åpent tilbud om to timer programmeringslabratorium med studentassistenter per uke. I forkant av eksamen skal alle studentene ha bestått 7 obligatoriske programeringsoppgaver. Sluttkarakteren er kun basert på avsluttende skriftlig 6-timers eksamen. Eksamensoppgavene i 2017 bestod av 16 deloppgaver med ulik vekt, tilsvarende de foregående årene. Det var 14 sensorer totalt i tillegg til faglærerne. Hver besvarelse ble rettet av to sensorer med overlappende studentmasse mellom en til tre sensorgrupper. Påmeldingen til emnet var 614, hvorav 528 var kvalifisert og avla eksamen våren 2017. Data består av data generert av sensorverktøyet, intervjuer med sensorene, samt et spørreskjema utfylt av studentene (svarprosent på 20%) i etterkant av eksamen og mottatt sensur med automatisk tilbakemelding.



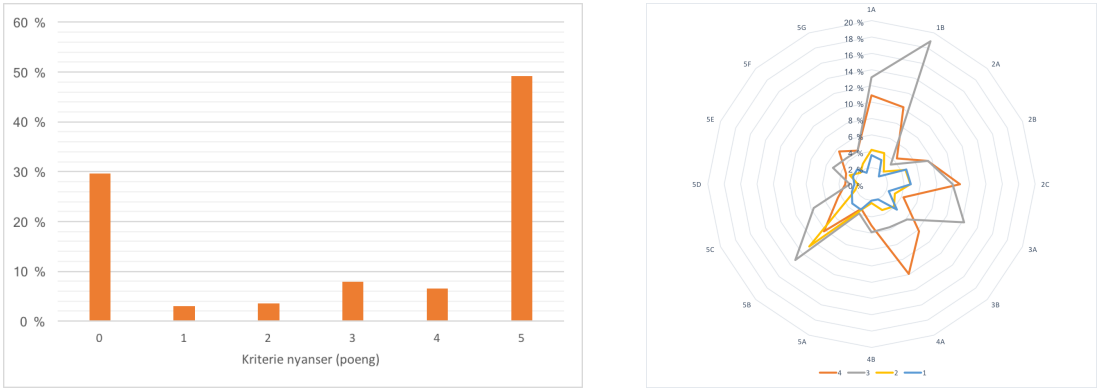
(a) (b)

Figur 2: *(a) Relasjonen mellom alle deloppgavene på eksamen og sensorenes poengjustering. Topp- og bunnstrekene representerer henholdsvis maks. og min. prosentvise poeng relativ til oppgavens totale poeng. Den stiplede linjen representerer medianen av disse poengjusteringene. Diamantpunktene viser antall besvarelser som har blitt justert relativt til hele kullet. (b) Viser karakterdistribusjonen for hele studentkullet fordelt over deloppgavene. Denne oversikten er et nyttig verktøy for faglærer å validere eksamensoppgavene opp mot både sensorveiledningen og vektingen for sluttkarakter.*

4 Resultater

Et av de mest omdiskuterte elementene ved et slik automatisk tilbakemeldingssystem er kriteriene. Disse kriteriene dannes for å gjenspeile læringsmålene i emnet gjennom å kvantifisere dem. I hvor stor grad en slik kvantifisering lykkes henger sammen med faktorer som nyanseringsgrad (her brukt 0-5 poeng), oppgavens utforming, studentenes entydige løsning og sensorenes forståelse av delmålene. Et av tiltakene som er innebygd i sensureringsprogrammet er muligheten for å justere hver deloppgave. I Figur 2(a) ser vi omfanget av justeringer gjort av alle sensorene. Fra diamantpunktene i grafen som representerer den prosentvise mengden av alle besvarelsene som er blitt justert, ser vi at majoriteten av deloppgaver som er blitt justert er under 10% av totalt 528 besvarelser. Hvorav selve mengden av justering for hver deloppgave ligger rundt $\pm$ 20% av oppgavens relative poeng. I hovedsak avviker deloppgave 3B og 5B fra disse resultatene, dog er deres median i justering innenfor 10%. Ved nærmere gjennomgang av de to konkrete punktene viser det seg at det skyldes tre konkrete besvarelser som har hatt «kreative» løsninger som ikke faller godt nok inn under delmålene som er satt. Alle disse kandidatene har også fått tydelige tilbakemeldinger gjennom friteksten i deloppgaven. Leser vi av medianen i Figur 2(a) ser vi at det trendvis er negativ justering. Dette vil tyde på at sensorene ønsker å gi mer trekk enn det de har mulighet for ettersom nyanseringsbildet for kriteriene går fra 0 - 5 og ikke innhar minuspoeng. Vi kan se dette fra intervjuene med sensorene om hva de mener om dette i avsnitt 4. Figur 3(a) viser i hvor stor grad sensorene har benyttet seg av nyanseringsmuligheten for hvert delmål. Ikke helt uventet er det i hovedsak benyttet fullt poeng (5 poeng) om delmålet er oppnådd eller 0 poeng hvis det er fraværende. Siden årets eksamen hadde over 50 delmål, blir derfor også utslaget for bruken av 0 og 5 poeng vesentlig større enn de andre nyansene. I Figure 3(b) ser vi oversikten over hvilke oppgaver sensorene har benyttet seg mest av nyansene (1-4 poeng). Dette kan også sees i sammenheng med hvilke oppgaver som har vært justert mest og hvilke oppgaver som har gitt en størst spredning i form av karakter (Figur 2(b)).

Alle besvarelser er tilfeldig utdelt og mengden av besvarelser per sensor varierer fra 6% - 22%. Figur 4 viser sensorenes individuelle karaktergivning fordelt på besvarelsene



(a) (b)

Figur 3: *(a) Denne grafen viser hvordan sensorene har benyttet seg av de 6 nivåene for nyansering av delmålene. På x-aksen er nyanseringspunktene 0 - 5 poeng og y-aksen viser den relative utslaget for hver nyanse basert på den akkumulerte summen av alle delmål for alle besvarelser. (b) Radarplottet viser i hvor stor grad de ulike nyanseringspoengene er benyttet fordelt på alle deloppgavene. For å gjøre dette plottet mer lesbart har vi ekskludert poengene 0 og 5. De stedene det er lavest bruk av nyansering er også enstydende med at det er de samme som har høyest utslag for poeng 0 og 5.*

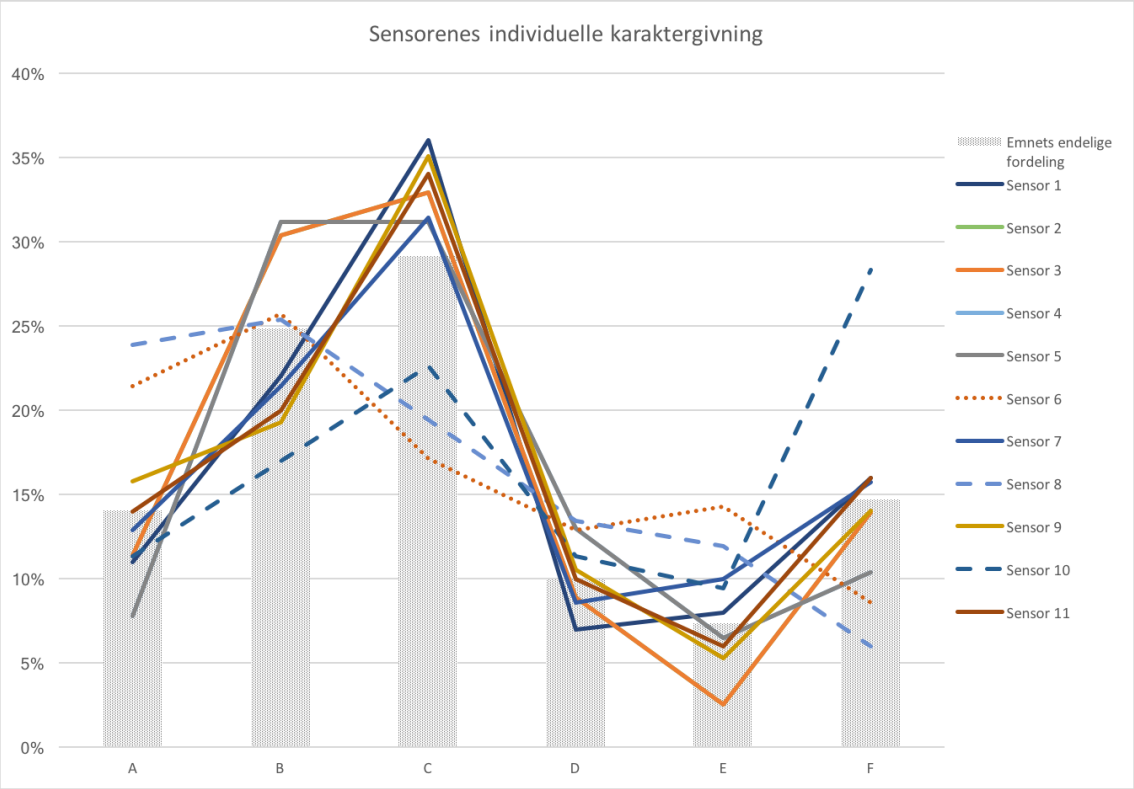
de har fått utdelt, de utvalgte sensorene er de som har fått utdelt totalt antall besvarelser som er mer enn 10% (av 528 besvarelser). Majoriteten av sensorene ser ut til å ha en normalfordeling omkring en sterk C, tilsvarende hele kullets karakterfordeling (grå stolpene). Det er tre sensorer som har et større avvik fra denne trenden, de er merket med stiplede linjer. Det viser seg at sensor 10 er generelt «strengere» i sensurering enn gjennomsnittet, mens sensor 6 og 8 er «snillere» sensorer.

Erfaringer fra faglærerne

Faglærernes hovedmotivasjon har vært å utvikle bedre eksamensoppgaver, styrke reliabiliteten på sensur, og å gi tilbakemelding til studentene for å øke deres læringsutbytte. Sekundært er det mulighetene for å unngå at sensorene må sensurere på nytt for de som ber om en begrunnelse en stund etter sensuren. Det har dermed også gått med en del tid som kan betegnes som første-gangs investering. Faglærer har estimert et tillegg på ca. 10 dagsverk som følge av dette. Utover disse timene til faglærer har det også vært studentassistenter som har administrert programmet og dataflyten.

Faglærerne sensurerte 30 besvarelser som en kontrollgruppe for å teste ut dette programmet. Erfaringene deres tilsier at målene som var skrevet opp virket mer som en sjekkliste og de har grunn til å tro at det vil bli en mer rettferdig vurdering. Spesielt trekker de frem de som har løst oppgaven mest lik «fasiten» og dermed ha fått full poengpott på alle punktene. For de mer kreative (mer eller mindre gode) løsningene antar de at det vil være noe vanskeligere å bruke disse punktene og deres representative vekting, det er derfor mulig å bruke justeringssatsen på hver deloppgave som et balanserende element. Denne justeringssatsen ble også brukt av faglærerne. Mer kreative (men ofte riktige) besvarelser fikk nok ikke så gode begrunnelser som de som fulgte fasiten.

Det å generere tilbakemeldinger til studentene gir også muligheten til å ekstrahere ut nyttig informasjon tilbake til faglærer. Spesielt har det vært nyttig med karakterdistribusjonen fordelt over deloppgavene, jmfr Figur 2(b). Med en slik oversikt kan man fatte datadrevne beslutninger for vekting mellom deloppgaver, eventuelt også å eliminere opp-



Figur 4: *Denne grafen viser «sensorenes profil» og karakterfordeling basert på deres utvalg av besvarelser. De grå stolpene viser karakterfordelingen for den endelige karakteren i emnet.*

gaver som har vært «feil», spesielt tydelig vil det være fra Figur 2(b). Som et eksempel viser oppgave 4A at over 75% av studentene fikk toppkarakteren A og B, burde da denne oppgaven telle like mye som oppgave 5B som har 75% av studentene fikk bunnkarakterene E og F? I så fall vil dette bidra inn i refleksjonen faglærer gjør for utarbeidelsen av neste års eksamen.

Erfaringer fra sensorene

I forkant av sensureringen har alle sensorene fått utdelt et informasjonsskriv som forklarer prosedyren og hvordan de skal bruke sensureringsprogrammet. Før selve sensuren mente noen av sensorene at innføringen av automatisk tilbakemelding på denne måten ville gjøre at sensuren kom til å bli mer arbeidskrevende, og en sensor trakk seg på grunn av dette. Eksamen var fredag 16.juni 2017. Sensorene fikk besvarelsene sent lørdag 17.juni 2017 og sensuren var ferdig onsdag 28.juni 2017, en dag før fristen.

I etterkant av sensureringen er det blitt gjennomført både kvalitative intervjuer og spørreskjema av samtlige sensorer. To av sensorene var førstegangssensor, mens resten har vært sensor i dette emnet tidligere eller andre tilsvarende emner. Ingen av sensorene har systematisk foretatt tidsavregning ettersom godtgjørelsen for arbeidet er akkordbasert for antall besvarelser. Allikevel har sensorene gitt uttrykk for hva de har brukt av tid. Naturligvis tar de første besvarelsene lengre tid og samtlige sensorer mener de har brukt mellom 45-60 minutter for de aller første. Etterhvert har de brukt jevnt over 20-25 minutter per besvarelse. Med tanke på sensureringsmetode har de en god fordeling mellom å rette horisontalt eller vertikalt. Vanlig sensureringsprosedyre er å sensurere

«vertikalt», det vil si å rette samme deloppgave for alle studentene etter hverandre før man sensurerer neste deloppgave. «Horisontal sensurering» betyr å sensurere en hel besvarelse for en kandidat før man sensurerer neste kandidat. Det er både fordeler og ulemper med disse to måtene å sensurere på. For «vertikalt» argumenteres det at det gir en mer rettferdig vurdering, mens for «horisontalt» argumenteres det for et helhetlig inntrykk for den gitte kandidat som kan gi en mer presis vurdering. Felles for begge sensorgrupperingene er de fremhever delmålene som gode ankerpunkter for å gjøre en mest mulig rettferdig vurdering. Noen sensorer har brukt fritekstfeltet flittig til å begynne med, men etterhvert gitt det opp av hovedsakelig to årsaker. Den ene er åpenbart tidsforbruket - her viser de til at de ikke helt hadde fått med seg at studentene ville få automatisk tilbakemelding basert på poenggivningen. Den andre årsaken beror på sensorenes kritiske tilbakeblikk på kommentarene, hvor de påpeker at det fort ble en stikkordsform i friteksten.

Når det gjelder justeringene de har foretatt er de samlet sett enige om at delmålene var beskrivende nok og at justeringstilfellene var isolerte og distinktive. For de aller fleste delmålene kunne det ha vært godt nok med tre-nivåer for nyansering. Dog påpekte de at de heller vil ha muligheten for en seks-nivås skala på alle oppgavene i fremtiden, selv om de ikke brukes for alle. Utdypende om justeringene de har foretatt, var det ikke homogent om justeringen ble brukt for å trekke ned eller å trekke opp. Der de trakk ned skyldte oftest at studentene hadde klart delmålene, men at det var tydelig at de manglet forståelse eller at de hadde med unødvendig mye kode. Spesielt tydelig blir det når noen kandidater åpenbart har skrevet av programsnutter fra hjelpemidlene (alle trykte og skrevne hjelpemidler var tillatt å ha med på eksamen). Det ble også diskutert forholdet mellom å ha minuspoeng slik at sensorene kunne gi ekstra trekk for enkelte delmål. Det er fordeler og ulemper med å bruke minuspoeng, men vi tror at justeringsbolken gir mulighet for å indirekte kunne bruke minuspoeng og dermed unngå de ulemper minuspoeng vil ha for vekting og sluttsum. For tilfelle med å gi positiv justering var det oftest følgefeil som gjorde at utslaget på delmålene var lavere enn det studentene viste av forståelse og dermed valgte sensorene å justere opp. Det var felles konsensus om bruken av justeringsbolken fremfor å justere delmålpoengene, dette anseer vi som særdeles positivt for vårt innsamlede datagrunnlag.

Med tanke på interrater-reliabiliteten for sensureringen viser det seg at sensorparene, to og to sensorer for hver besvarelse, har hatt overraskende små forskjeller, altså relativ differanse i prosentpoeng. Sensorene forteller at der det var forskjeller mellom dem var det utelukkende i karaktertersklene. De resonnerte seg frem til at årsaken ligger i kvantifiseringen av delmålene opp mot statiske verdier for karakter. Dette gjør at 0.5 poengs forskjell mellom to sensorer faktisk kan utgjøre en hel karakter i forskjell. Validiteten av sensureringen har blitt ivaretatt gjennom det at alle sensorparene har gått igjennom avvikene og avgitt en felles forent poengsum. Ytterligere validering er også foretatt hvor alle studenter i grensesone, definert av faglærer på $\pm$ 2 poeng, er blitt nøye gjennomgått og etter en helhetsvurdering av begge sensorer fått fastsatt endelig karakter.

Slik det også fremkommer av avsnitt 2 så er delmålene bestemt ut fra hva eksaminator har tenkt om løsningen for oppgavene. Disse delmålene har vært til god veiledning for sensorene, spesielt har førstegangssensorene ansett dem som et bedre hjelpemiddel enn sensorveiledningen. Sensorene har også hatt muligheten til å diskutere detaljer med hverandre og faglærer, de ser ikke behovet for mer oppfølgning enn det som har vært gitt over mail.

Erfaringer fra studentene

Over 67% av studentene sier at den automatiske tilbakemeldingen samstemmer veldig mye med deres eget inntrykk av sin prestasjon på eksamen og at de har hatt veldig mye utbytte av tilbakemeldingen.

> *«Automatisk tilbakemelding var veldig nyttig for meg, og ga med innsikt i egen kapasitet og områder for forbedring. Jeg ville ikke bedt om begrunnelse selv, men lærte mye av tilbakemeldingen. Det har motivert meg til å be om begrunnelse, selv om man ikke er uenig med karakter. Automatisk tilbakemelding burde være del av alle fag der det er mulig!»* (Student #2978171)

Sitat:1

I forhold til studiemiljø svarer 43% at tilbakemeldingen har oppfordret dem til å ta kontakt med andre studenter og at det har tilbakemeldingen har i stor grad vært brukt i diskusjon. Dette illustrerer at studentene erfarer at tilbakemeldingene er nyttig for videre læring og over 45% av studentene sier at de har lest og brukt tilbakemeldingen flere ganger etter sensuren og i videre studie. De følgende sitatene viser også at det tilbakemeldingene og transparent vurdering gir studentene tillit til vurderingene.

> *«Det var veldig betryggende å få automatisk begrunnelse, og jeg følte meg mye tryggere på vurderingen som hadde blitt gjort i emnet.»* (Student #2977287)

Sitat:2

> *«Ordningen er god og bør videreføres. Jeg bommet på en av karaktermarginene med 0.5p, men følte fremdeles at tilbakemeldingene var tydelige nok og poengsettingen godt nok begrunnet til at det neppe ville føre fram å klage. Det gav også innsikt i hvilke ting jeg burde ha tenkt på eller gjort bedre, som er svært nyttig.»* (Student #2977450)

Sitat:3

> *«Automatisk begrunnelse var en fantastisk overraskelse - jeg skulle ønske alle fag gjorde dette. Jeg fikk svar på alt jeg lurte på angående min karakter – før jeg i det hele tatt kom på at det var noe jeg lurte på. Jeg følte også at karakteren var mer rettferdig etter å ha mottatt en så grundig og gjennomført automatisk tilbakemelding. Denne tilbakemeldingen er MYE BEDRE enn tilbakemeldingene jeg faktisk har bedt om. Keep it up!»*
> (Student #3073951)

Sitat:4

I forhold til rollen tilbakemeldingen har spilt for om studentene har valgt å klage eller ikke, svarer 54% av de som ikke klaget at tilbakemeldingen gav stor nok innsikt i bedømmelsen for deres valg. Mens, 75% av de som klaget har basert sin klage på tilbakemeldingen.

Klageandelen for emnet var på 5%.

5 Diskusjon

I denne artikkelen har vi presentert et studie gjort for et stort informatikk emne, n > 500 studenter, om automatisk tilbakemelding på eksamen. Som resultatene viser er et godt forarbeid viktig for å lykkes med gode kriterier og nyanseringsmuligheten for disse kriteriene. Selvom vi kan påstå at automatisk tilbakemelding reduserer sensoreringstid, så blir den egentlige total timesregnskapet nesten likt tidligere. Mye av den innsparte tiden for sensorene går til faglærer som investerer i utviklingen av sensorveiledning og etablering av kriterier. Det kan påberopes at tiden faglærer bruker på kriteriene også i stor grad vil føre til samstemt vurdering («constructive alignment») og på den måte vil tidsinvesteringen også bidra til økt kvalitet på spørsmålene. Som også Raaheim fremhever [Raaheim, 2000], viser denne studien at det er sterk sammenheng mellom tydelige kriterier og sensorveiledning, og reliabiliteten i vurderingene. Erfaringen fra sensorene viser en høy interratereabilitet, spesielt gjennom mengde av justeringer de har foretatt på eksamenssensuren, Figur 2(a). Disse justeringene har hovedsakelig vært foretatt for løsninger som har vært riktige, men som har vært veldig annerledes enn «fasiten». I diskusjonen videre for neste studie vil være å kunne se på relasjonen mellom nyanseringsintervallet, her brukt 0-5 poeng, og justeringsmengden. Vi ser fra Figur 3(a) at bruken av ytterpunktene i nyanseringsskalaen er signifikant, men det kan også argumenteres med mengden av delkriterier. Det var totalt 50 delmål og 528 besvarelser hvilket gir 26.400 datainnsamlingspunkter, da vil det naturligvis gi stort utslag om noen av delmålene har vært for konkret binært kvantifiserende. Med Figur 3(b) kan vi se hvilke delmål som har vært best egnet til kvantifisering. Dette er dog ikke ensbetydende med at de oppgavene som ikke slår ut her ikke kan eller bør kvantifiseres, men heller at det ikke har vært behov for en skala fra 0-5 poeng.

I forhold til validitet for sensureringen kan vi trekke frem mengden av justeringer som sensorene har foretatt. Justeringene kan også sees på som den subjektive vurderingen og direkte den skjønnsmessige vurderingen sensorene gjør. Dette sett i sammenheng med sensorenes profil (Figur 4) kan vi ekstrahere ut og til en viss grad tallfeste avviket fra de andre. Her kan det på sikt automatiseres slik at sensorenes vurderinger normaliseres i forhold til hverandre og oppnå enda høyere reliabilitet. Et aspekt for validitet i sensureringen er den initielle kalibreringen av sensorene i forhold til faglærer. Gjennom denne studien har vi, basert på faglærers og sensorenes tilbakemelding forstått det slik at delmålene med delkriteriene på mange måter har bidratt til en slags sjekkliste for hva man skal se etter. Ettersom denne «listen» har vært så detaljert og muligheten for å angi en grad (0-5p) har det ikke vært etterspurt om ytterligere informasjonsmøter eller skriv. På mange måter har sensorene selv-kalibrert seg gjennom å bruke dette programmet. Emnets karakterfordeling for årets eksamen følger nært fordelingen historisk sett i emnet og som Figur 4 viser at majoriteten av sensorene også følger samme fordeling, noe som igjen styrker inntrykket at at programmet styrker reliabilitet og validitet.

Studentenes tilbakemeldinger tyder på at de automatiske kvalitative tilbakemeldingene oppleves som positivt både i den forstand at de styrker forståelsen av og tilliten til vurderingen som ligger bak den endelige karakteren, og at tilbakemeldingene oppleves som noe som bidrar til deres faglige utvikling. Både forståelsen av hva de har fått til og hvor de har feilet, og kommentarer som gir en retning til videre arbeid oppleves som verdifulle. Dette korresponderer med litteraturen om formativ vurdering, som fremhever betydningen av 'feedforward' [Nicol and MacFarlane-Dick, 2006]. Tilbakemeldingen ty-

der også på at studentene tar kommentarene aktiv i bruk i videre faglig arbeid, ikke minst i samarbeid med medstudenter. Dermed ser det ut til at denne formen for tilbakemelding ikke bare støtter den faglige læringen, men også studentenes metakognisjon og selvregulering - kompetanser som er avgjørende for å lykkes i høyere utdanning [Bransford et al., 2000].

Referanser

[Biggs and Tang, 2007] Biggs, J. and Tang, C. (2007). *Teaching for Quality Learning at University Third Edition Teaching for Quality Learning at University*, volume 3th ed. Open University Press.

[Boud and Molloy, 2013] Boud, D. and Molloy, E. (2013). *Feedback in Higher and Professional Education: Understanding it and doing it well*. Routledge.

[Bransford et al., 2000] Bransford, J. D., Brown, A. L., and Cocking, R. R. (2000). *How People Learn*.

[Esterhazy and Damşa, 2017] Esterhazy, R. and Damşa, C. (2017). Unpacking the feedback process: an analysis of undergraduate students' interactional meaning-making of feedback comments. *Studies in Higher Education*, 5079:1–15.

[Jiménez-gonzález et al., 2008] Jiménez-gonzález, D., Álvarez, C., López, D., Parcerisa, J.-m., Alonso, J., Pérez, C., Tous, R., Barlet, P., Fernández, M., and Tubella, J. (2008). Work in Progress – Improving Feedback Using an Automatic Assessment Tool. *ASEE/IEEE Frontiers in Education Conference*, pages 9–10.

[Juwah et al., 2004] Juwah, C., Macfarlane-dick, D., Matthew, B., Nicol, D., Ross, D., and Smith, B. (2004). Enhancing student learning through effective formative feedback. *The Higher Education Academy Generic Centre Enhancing*, 1(68):1–41.

[Malmi and Korhonen, 2004] Malmi, L. and Korhonen, A. (2004). Automatic feedback and resubmissions as learning aid. *Proceedings - IEEE International Conference on Advanced Learning Technologies, ICALT 2004*, pages 186–190.

[Mirmotahari and Berg, 2017] Mirmotahari, O. and Berg, Y. (2017). Individuell «automagisk» tilbakemelding på skriftlig eksamen. *Nordic Journal of STEM Education*, 1(1):287–293.

[Nicol and MacFarlane-Dick, 2006] Nicol, D. and MacFarlane-Dick, D. (2006). Formative assessment and selfregulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2):199–218.

[Raaheim, 2000] Raaheim, A. (2000). En studie av inter-bedømmer reliabilitet ved eksamen på psykologi grunnfag. *Tidskrift for Norsk Psykologiforening*, 37:203–213.

[Rust, 2002] Rust, C. (2002). The Impact of Assessment on Student Learning: How Can the Research Literature Practically Help to Inform the Development of Departmental Assessment Strategies and Learner-Centred Assessment Practices? *Active Learning in Higher Education*, 3(2):145–158.

[Siddiqi et al., 2010] Siddiqi, R., Harrison, C. J., and Siddiqi, R. (2010). Improving teaching and learning through automated short-answer marking. *IEEE Transactions on Learning Technologies*, 3(3):237–249.

[Tee and Ahmed, 2014] Tee, D. D. and Ahmed, P. K. (2014). 360 degree feedback: An integrative framework for learning and assessment. *Teaching in Higher Education*, 19(6):579–591.

[Thelwall, 2000] Thelwall, M. (2000). Computer-based assessment: a versatile educational tool. *Computers & Education*, 34(1):37–49.