

# Studenters evaluering av hverandres arbeider

Hans Engebretsen og Karoline Moholth

Høgskolen i Buskerud, Kongsberg

[Hans.Engebretsen@hibu.no](mailto:Hans.Engebretsen@hibu.no), [Karoline.Moholth@hibu.no](mailto:Karoline.Moholth@hibu.no)

## Sammendrag

Rapporten presenterer foreløpige resultater fra et prosjekt der en undersøker studenters nytte av, og reaksjoner på å vurdere hverandres innleveringer. Forsøket er gjennomført i vårsemestret 2007 for studenter i 1. år av bachelorutdanningen i datateknikk. Gruppen som er undersøkt består av hele klassen på 16 studenter. Faglærer vurderte også samtlige innleveringer. Sammenfallet mellom faglærers og studentenes karaktersetning økte klart fra øving 1 til øving 2. Vi finner ingen klar sammenheng mellom studenters eksamensprestasjoner og sammenfall med lærers karaktersetning. Studentens svar angående eget utbytte er relativt nøytrale, men alle kommentarer på spørreskjemaene som ble utdelt i etterkant av opplegget er positive til undervisningsformen.

## Innledning

Hensikten med dette prosjektet har vært å danne seg et bilde av hvilken nytte studenten har av å vurdere hverandres arbeider. Nyten kan deles inn i to hovedkomponenter:

1. Det man lærer ved å rette andres arbeide
2. Det man lærer ved å få tilbakemelding fra andre enn faglærer

Det er også av interesse å studere problemstillinger som:

1. Er utbyttet av denne evalueringsformen avhengig av nivået på:
  - a. Den som har levert besvarelsen
  - b. Den som retter besvarelsen
2. Hvordan avhenger studentens rettetferdighet av deres faglige nivå

For å kunne kontrollere nytteverdien av denne arbeidsformen har vi bedt studentene fylle ut et evalueringsskjema for ordningen, og vi har fått alle (med unntak av en) til å oppgi navn på dette skjemaet, slik at vi kan se korrelasjonen mellom prestasjoner på øvinger, eksamensresultater og rettetferdigheter. Gruppen som har deltatt i forsøket er såpass liten (16 studenter) at det blir vanskelig å trekke sikre konklusjoner med høyt signifikansnivå. Vi har derfor konsentrert oss om å se hvilke trender som synes å være tilstede i det foreliggende materiale, og hvordan disse trendene stemmer med allerede eksisterende forskning på området. Undersøkelsen er gjennomført i faget DVVP1021 "Videregående programvareutvikling". Faget undervises i andre semester av HiBu's bachelorutdanning i ingeniørfag. Det ble gitt tre øvingsoppgaver som alle måtte vurderes til bestått for at kandidaten skulle ha rett til å gå opp til eksamen.

## Tidligere arbeider

Brown og Glaser [1] har samlet en rekke artikler om evalueringsformer i en bok utgitt i 1999. Angela Brew har bidratt med en av artiklene [2], og hennes konklusjon er at når undervisere lar studentene bidra i evalueringen øker den profesjonelle vurderingsevnen til begge parter. I den samme samlingen gjennomførte Jordan en studie [3] for å finne fordeler og ulemper med denne evalueringsformen. Hennes konklusjoner er at den gir økt selvtillit, den lærer studentene å vurdere presentasjonsformer, og gir studentene en følelse av kontroll. Ulemper er at studentene kan la seg influere av vennskap, rasespørsmål og tidligere holdninger til hverandre, og at studentene er redde for å vurdere feil fordi de ikke har dyp nok forståelse av emnet.

Falchikov og Goldfinch [4] har gjennomført en studie som analyserer sammenhengen mellom karakterer som blir utdelt av andre studenter og av faglærer ved å analysere 48 uavhengige studier som sammenlikner studenters og læreres karakterer. Det er en diskusjon om validiteten ved å bruke faglærers karakterer som et mål på hvor riktige studentenes karakterer er. Denne studien sammenlikner mange faktorer, spørsmålenes kvalitet, faglig nivå på kurset som har brukt evalueringen, fagområder, antall studenter og oppgaver som er med osv. Konklusjonene fra denne studien er ikke entydige men noen trender viser seg. Studentevaluering som involverer evaluering av mange små dimensjoner synes å være mindre riktige enn evalueringer som krever global evaluering basert på godt forståtte kriterier.

Studentevaluering av akademisk natur ser ut til å gi et mer entydig resultat enn evaluering av praktiske oppgaver. Det synes å gi like bra resultat å la enkeltstudenter vurdere hverandres oppgaver som å dele dem opp i større grupper. Det synes ikke å være noen forskjell på validiteten på videregående kurs og begynnerkurs. MacPherson [5] gjennomførte en studie hvor målet var å finne ut om effekten av studentevaluering i utviklingen av evnene til å tenke kritisk. Det viste seg der at forskjellen mellom lærer og studentgitte karakterer ble minsket over tid og at etter ett semester er der ingen varians på disse. Denne studien konkluderte også med at det var sannsynlig at instruksjon i kritisk tenkning sannsynligvis øker evnen for studenter til å evaluere hverandre raskt og entydig. Eburian, Holden og Abarbanel [6] gjennomførte en studie spesifikt på fagområdet Java-programmering. Denne studien viser at ved å gi varierte læreopplevelser som inkluderer at studentene instruerer hverandre oppnår man en kunnskap og tiltro som gjør det enklere for studentene å lære avansert Java-programmering.

Gyanani [7] gjennomførte en studie som vurderer effekten av at studenter instruerer og hjelper hverandre. Der synes det at begge sider av denne kunnskapsoverføringen har nytte av det. Studenten som blir instruert får en mindre truende person å forholde seg til og studenten som instruerer får en bedre forståelse av læringen. Konklusjonen her er at å minske tiden en underviser prater til en klasse og øke tiden klassen prater sammen skaper et bedre læringsmiljø.

## Beskrivelse av opplegget

Studentene som deltok i forsøket er studenter i 2. semester av sin dataingeniørutdanning ved HiBu. I faget Videregående programvareutvikling har de tre obligatoriske innleveringer. Disse må beståes for å få gå opp til eksamen. Faglig tema for disse tre øvingene var som følger:

- Binærtre
- Sortering
- Utvidelse av tegneprogram

I tillegg til kildekode leveres det en rapport, vanligvis i Word-format. Ved fordeling av oppgavene var det et prinsipp at dersom person A vurderte person B sin besvarelse, så skulle ikke B vurdere A sin besvarelse. Studentenes vurdering ble gjennomført ved hjelp av et skjema der en skulle kommentere og sette karakter (A-F) på følgende tema og deltema:

- Java-koden
  - Lesbarhet/struktur
  - Kommentarer
  - Teknisk løsning (virker det, hvor mye er ferdig)

- Rapporten
  - Oversiktligheit (layout iht standard/ innholdsfortegnelse)
  - Dokumentasjon av testing

I tillegg skulle en sette en total karakter ved hjelp av den samme skalaen. Øvingsopplegget vårt ble forsøkt gjennomført på alle tre øvingene, men det var en del som ikke leverte vurderingsskjema for den siste innlevering fordi fristen for å levere dette var satt til en dato etter undervisningsslutt, og da var mange studenter fokusert på eksamenslesing. Vi valgte derfor å se bort fra denne i det videre arbeide. Studentene fikk vite om lærers vurdering av en innlevering rett etter at de hadde levert inn sin egen vurdering. Dette innebærer at før de vurderte øving 2 kjente de til avviket fra øving 1.

Vi forsøkte så å se på følgende mulige relasjoner:

- Blir studentene bedre til å gi sammenfallende karakter med lærer, fra den første til den andre innleveringen?
- Hvordan er sammenhengen mellom eksamens karakter og hvor sammenfallende karakter med lærer de gir?
- Hvordan er sammenhengen mellom eksamens karakter og opplevd nytteverdi?
- Hvordan er sammenhengen mellom eksamens karakter og karakter på innleveringer?

Kandidatnr	Oppgave 1				Oppgave 2				Sum abs. avvik	Eks.	
	Student	Faglærer	Avvik	Avvik abs.	Student	Faglærer	Avvik	Avvik abs.		Student	Sum avvik
1	C	C	0	0	C	D	-1	1	1	E	1
2	C	C	0	0	D	B	2	2	2	C	2
3	C	C	0	0	A	A	0	0	0	F	0
4	B	C	-1	1	C	C	0	0	1	E	1
5	A	C	-2	2	D	C	1	1	3	C	3
6	B	B	0	0	B	B	0	0	0	A	0
7	C	E	-2	2	B	B	0	0	2	B	2
8	B	C	-1	1	E	E	0	0	1	C	1
9	B	C	-1	1	B	D	-2	2	3	F	3
10	C	C	0	0	C	C	0	0	0	B	0
11	B	C	-1	1	B	B	0	0	1	D	1
12	D	E	-1	1	C	C	0	0	1	C	1
13	B	C	-1	1	B	B	0	0	1	D	1
14	E	E	0	0	B	B	0	0	0	F	0
15	A	B	-1	1	B	B	0	0	1	D	1
16	B	C	-1	1	E	E	0	0	1	E	1

Figur 1 Sammenligning av studenters og faglærers bedømming

Abs. avvik	Antall	
	Øving 1	Øving 2
0	6	12
1	8	2
2	2	2

Figur 2

Antall studenter med fordelt på absolutt avvik for øving 1 og 2

For å belyse de to første problemstillingene benyttes sammenstillingen i figur 1. Vi ser en klar utvikling i retning av bedre sammenfall mellom student og lærer karaktersetning fra øving 1 til øving 2. Dette gjelder både absoluttverdi av samlet avvik og antall som får forskjellige grad av avvik fra faglærers karaktersetning. For øving 2

har altså samsvaret mellom studentenes og lærers karaktersetting blitt så mye bedre at hele  $\frac{3}{4}$  av studentene treffer samme karakter som lærer. Dette er kort oppsummert i figur 2. Figur 1 viser derimot at det ikke er mulig å se noen klar sammenheng mellom hvor godt studentene treffer på sin karaktersetting, og hvor bra de gjør det på eksamen.

På påstandene nedenfor skal du gi en vurdering på en skala fra 1-5, der 1 betyr svært uenig, 2 litt uenig, 3 nøytral, 4 enig, 5 svært enig							Snitt	Standardavvik	
Spm.	Påstand	1	2	3	4	5			U
1	Jeg føler jeg har god kontroll på pensum i faget	0	2	5	7	2		3,56	0,97
2	Jeg føler at jeg har lært noe faglig av å rette andres innleveringer	1	3	8	3	1		2,94	1,18
3	Det var vanskeligst å rette den første innleveringen	2	4	4	5	1		2,81	1,37
4	Det var ubehagelig at en annen i klassen rettet min oppgave	9	3	2	2	0		1,25	1,60
5	Det var ubehagelig å rette en annens oppgave	7	3	3	3	0		1,69	1,65
6	Tilbakemeldingene jeg fikk var nyttige	1	3	3	9	0		3,19	1,26
7	Tilbakemeldingene jeg fikk var bedre enn de jeg vanligvis får fra faglærer	1	6	5	2	1	1	2,67	1,27
8	Jeg ønsker at samme metode kan benyttes i andre fag	1	5	7	2	1		2,75	1,21
9	Jeg lærte mer av å rette en annens innlevering enn av å løse oppgaven selv	5	6	4	1	0		1,75	1,44
Kommentarer til opplegget:									
Det har vært et bra fag									
Grei ordning									
Synes det er et helt ok opplegg									
Kunne hatt fler, fikk noe overfladiske kommentarer									
Man lærer jo mye av å se hva andre har gjort, ser også andre måter å løse samme problem på									
Lærer mye mer av å gjøre oppgavene enn å rett en annens.									

Figur 3 Spørreskjema med svarfordeling. For hvert spørsmål er beregnet Snitt og standardavvik. Kommentarer er alle enkeltkommentarer fra de av studentene (6 av 16) som har valgt å kommentere opplegget

Rett i etterkant av skriftlig eksamen i faget ble studentene bedt om å fylle ut spørreskjemaet som er vist i figur 3. Hensikten med dette skjemaet var å få studentenes reaksjoner på, og opplevd læringsutbytte av opplegget. Som det framgår av dataene er det stor spredning i meningene om opplegget. Det spørsmålet som faktisk ble mest entydig besvart i hele undersøkelsen, var spørsmål 1. Dette står i en viss kontrast til eksamensresultatene. Karakterfordelingen ved eksamen ble som følger (antall angitt i parentes: A(1), B(2), C(4), D(3), E(3) og F(3)).

For å se på sammenhengen mellom eksamenskarakter og opplevd nytteverdi er det gjort en sammenstilling i figur 4. Denne sammenstillingen belyser også sammenhengen mellom eksamenskarakter og karakter på innleveringene. Figuren viser at det ikke er noen klar sammenheng mellom eksamenskarakter og opplevd nytteverdi. Den viser derimot en klar sammenheng mellom karakter på innleveringene og eksamenskarakter. Begge de som har stort avvik her, har strøket på eksamen. Dette kan skyldes taktiske hensyn. Ved å stryke får de krav på en kontinuasjonmulighet som de ellers kunne ha gått glipp av. Ser vi bort fra disse to er det fra 0 til 3 i samlet avvik (se figur 4).

Vi hadde også noen spørsmål på spørreskjemaet som vi til nå ikke har analysert svarene på. Spørsmål 4 og 5 (se figur 3) er stilt for å undersøke om studentene føler det ubehagelig å rette retten en annen students arbeider, eller om de følte det ubehagelig at deres eget ble vurdert av en medstudent. Før vi går inn på resultatene her er det viktig å slå fast at vi med vårt kjennskap til den enkelte student tok individuelle hensyn ved valg av hvem som skulle rette for hvem. Vi ser at begge disse spørsmålene avslørte lite ubehag, men at graden av ubehag var noe større ved å bli rettet enn ved selv å rette. Vi er meget fornøyd med at ingen sa seg ”svært enig” på disse spørsmålene.

Kandidat nr.	Spørsmål 2	Spørsmål 6	Spørsmål 7	Eksamen	Øving 1	Øving 2	Samlet avvik
1	2	2	1	F	C	B	-7
2	3	2	2	D	C	C	-2
3	3	4	4	C	C	A	2
4	3	4	3	B	C	B	1
5	3	4	3	B	C	C	2
6	4	5	4	F	C	B	-7
7	1	3	3	D	B	C	-3
8	2	3	2	A	B	B	2
9	3	3	2	C	C	B	-1
10	3	4	3	C	C	C	0
11	3	1	2	E	C	D	-3
12	3	2	2	C	C	B	-1
13	5	4	5	E	E	E	0
14	4	4	2	E	C	D	-3
15	2	4	4	D	E	B	-1
16	4	4	3	F	E	E	-2

Figur 4  
Sammenheng mellom opplevd nytteverdi, karakter på innleveringer og eksamenskarakter

Svarene på spørsmål 6, 7 og 9 angir nytteverdi av denne arbeidsmetoden. Her kommer vi klart ut på den positive siden ved at bare 1 av de 16 studentene mener at de har hatt liten nytte av tilbakemeldingene. Når det gjelder sammenligning med den feedback faglærer gir mener bare 3 av studentene at den er dårligere enn den de fikk fra medstudenten som rettet deres oppgave. Bare en student svarer at vedkommende lærte mer av å vurdere andres arbeide enn å selv å bli vurdert. Når vi går inn på denne studentens spørreskjema viser det seg å være en av de tre studentene som strøk til eksamen. Entusiasmen for å bruke denne undervisningsformen i andre fag er i liten grad til stede. Dette kan forklares med at studentene brukte mye tid på å rette, særlig første øving.

## Konklusjoner

Som påpekt i forbindelse med tidligere arbeider [4] kan men trekke i tvil validiteten av å sammenligne studentenes og faglærers karakterer. Vår konklusjon er at de kriterier vi ga for bedømmelse (lesbarhet og struktur for Java-kode samt ryddig rapport med dokumentasjon av testing) bidrar til å gjøre en slik sammenligning relevant. I figur 1 og 2 ser vi at studentenes og lærerens karaktersetting sammenfaller vesentlig bedre for øving 2 enn for øving 1. Dette stemmer med MacPherson's funn [5].

Vi finner ingen entydig sammenheng mellom studentenes eksamensresultat, og deres evne til å sette samme karakter som læreren, se figur 1. Det er ikke mulig å se noen klar sammenhengen mellom eksamenskarakter og opplevd nytteverdi. Derimot er det med få unntak, en ganske klar sammenheng mellom karakter på innleveringene og eksamenskarakter. Studentene var entydig på at de ikke følte noe stort ubehag verken ved å rette andres innleveringer eller ved at andre rettet deres innleveringer. Dette er påpekt som et mulig problem av Brown og Glasner [1]. Vi mener vår metode for utvelgelse av hvem som rettet for hvem bidro til å minimalisere dette problemet. Studentenes vurdering av nytteverdien av denne arbeidsformen er positive. De fleste opplever nytteverdien av faglærers tilbakemelding som større enn den de får fra medstudenter. Studentene oppgir at de lærer mer av å løse oppgaven selv enn å rette en annen students innlevering. Studentenes manglende entusiasme for å benytte tilsvarende

vurderingsform i andre fag tilskriver vi arbeidsmengden ved å rette andres innleveringer.

## Videre arbeide

Det vil i framtida være interessant å undersøke om de indikasjoner vi fikk her vil være de samme om vi tester opplegget på en annen gruppe studenter. I tråd med det som står i litteraturen var vi bevisst på at vi ikke skulle utsette helt ferske studenter for denne undervisningsformen. I vårt nye studieopplegg som settes i verk f.o.m. høsten 2007 får vi et videregående programmeringsfag som inneholder en del av de samme tema som i det faget vi nå har undersøkt. Dette faget vil imidlertid undervises i 3. semester, og det er av interesse å se om dette slår positivt ut. Vi vil kommende studieår ha betydelig flere studenter i første årskurs. Dette vil gi oss muligheten til å få et bredere statistisk materiale. Vi er også interessert i å teste ut andre faktorer i tillegg som:

- Kjønn
- Studentenes sosiale relasjoner
- Hvordan påvirker denne undervisningsformen læringsmiljøet i klassen over tid

Den siste problemstillingen ønsker vi også å teste for klassen som er gjenstand for denne studien, og det ønsker vi å gjøre mot slutten av høstsemesteret 2007, altså et omtrent et halvt år etter at de var ferdig med dette faget. Vi ser også et det kan være av interesse å spørre studentene om de opplever sin mangel på faglig dybde som en begrensende faktor ved vurdering av medstudentenes innleveringer. Det kan også være av interesse å undersøke studentenes evne til å vurdere egen innsats på eksamen for å sammenligne denne med deres evne til å vurdere andres innleveringer. Siden spørreskjemaet ble utfylt rett etter eksamen kan dette gjøres ved at studentene på spørreskjemaet (figur 3) blir bedt om å tippe hvilken karakter de ender opp med.

## Referanser

1. Brown, Sally and Glasner, Angela "Assessment matters in higher education", Open University Press. ISBN 0-335-20243-8, 0-335-20242-x, 1999.
2. Brew, Angela "Towards Autonomus Assesment" . Brown, Sally and Glasner, Angela "Assessment matters in higher education", Open University Press. ISBN 0-335-20243-8, 0-335-20242-x, 1999. pp. 159 - 171
3. Shirly Jordan "Self-Assesment and Peer Assessment". Brown, Sally and Glasner, Angela "Assessment matters in higher education", Open University Press. ISBN 0-335-20243-8, 0-335-20242-x, 1999. pp. 172-182.
4. Falchikov, Nancy and Goldfinch, Judy, "Student Peer Assessment in Higher Education: A Meta-Analysis Comparing Peer and Teacher Marks", *Review of Educational Research; Fall2000*, Vol. 70 Issue 3, date, pp.
5. MacPherson, Karen, " The Development of Critical Thinking Skills in Undergraduate Supervisory Management Units: Efficacy of Student Peer Assessment", *Assessment & Evaluation in Higher Education*, v24 n3, Sep 1999, Pages 273-84.
6. Emurian Henry, H., Holden, Heather, K. and Abarbanel, Rachel, A., "Managing programmed instruction and collaborative peer tutoring in the classroom: Applications in teaching Java™", *Computers in Human Behavior, In Press, Corrected Proof*, Available online 28 March 2007
7. Gyanani , T, C and Pahuja ,Premlata, " Effects of Peer Tutoring on Abilities and Achievement *Contemporary Educational Psychology* Volume 20, Issue 4, October 1995, Pages 469-47