Uniped årgang 37, 3/2014, side 63–77

Artikkelen er vitenskapelig vurdert av forskere utenfor redaksjonen

Av Johan Fredrik Rye

Konsistente karakterer? *En undersøkelse av sensorreliabilitet i sosiologifaget*

Johan Fredrik Rye
Institutt for sosiologi og statsvitenskap, Norges teknisk-naturvitenskapelige universitet, Trondheim.
E-post: Johan.Fredrik.Rye@svt.ntnu.no

Sammendrag
I kjølvannet av Bologna-prosessen arbeides det med å standardisere karakterpraksisene ved de høyere lærestedene i Norge. I denne artikkelen presenteres erfaringene fra et vurderingsarbeid innen sosiologifaget, der et seksmanns sensurpanel hver for seg og kollektivt vurderte tolv bachelor-oppgaver fra seks læresteder. Resultatene viser til dels betydelige avvik i hvordan eksamensarbeidene ble vurdert. I seks av ti tilfeller satte panelsensorene en annen karakter enn den opprinnelige. Mellom delta-kerne i sensurpanelet er det også betydelige forskjeller, både i nivå og rangering av oppgavene. Det er likevel sjelden snakk om store avvik, som i de fleste tilfellene dreier seg om ett trinn på karakterskalaene. Erfaringene fra sosiologifaget styrkes av tilsvarende undersøkelser i andre samfunnsvitenskapelige fag. Forskjellene i sensorenes vurderinger knyttes til mangelen på eksplisitte standarder for hva som representerer god akademisk kvalitet, hvordan studentenes arbeider skal vurderes opp mot disse standardene og endelig hva de forskjellige karakterene representerer – om C er en «god» eller «dårlig» karakter. Skal Bologna-prosessens mål om enhetlige nasjonale og europeiske karaktersystemer innfris, er det derfor behov for mer eksplisitte retningslinjer for karaktersetting. Det fordres også for utvikling av uformelle kulturer som bidrar til at forskjellige sensorer faktisk standardiserer sine karakterpraksiser i det daglige sensorarbeidet.

Abstract
In the wake of the Bologna process, the Norwegian higher education system is working to improve grading standardization. This paper reports results from an evaluation of twelve bachelor theses delivered at the six Norwegian universities offering bachelor studies in sociology. A panel of six graders evaluated the theses, both individually and collectively. Results show substantial divergence in grading. In six out of ten cases, the individual graders deviated from the original committee's decision. There were also substantial inconsistencies between the panel member's grading. However, in most cases discrepancies were modest and without any clear pattern. Discrepancies seem to follow from graders' different standards of «good»

Citation: *Uniped* 2014, **37**: 22654 - http://dx.doi.org/10.3402/uniped.v37.22654

science, different readings of students' works, and different interpretations of the grading scale. The findings suggest the need for a stronger standardization of grading practices across institutions to meet the Bologna objectives.

Innledning

Å sette karakterer på studentenes faglige prestasjoner er et grunnleggende element ved virksomheten ved alle universiteter og høyskoler (Sadler, 2009). Studentene trenger tilbakemelding på sine faglige arbeider og samfunnet trenger vurderinger av deres kompetanse. I de senere årene er det også lagt større vekt på karakterenes formative funksjon – det vil si at evalueringsoppleggene inngår som en vesentlig del av studentenes læringsprosesser (Gynnild, 2003). Det legges derfor stor vekt på å utvikle praksiser som fordeler de korrekte karakterene til de riktige studentene. Et anslag fra egen arbeidsplass antyder at omtrent en sjettedel av undervisningsressursene er knyttet til vurderingsarbeid.

Som oftest er resultatet akseptabelt. Karaktersystemet som prinsipp vurderes gjerne som legitimt, både av studentene som mottar karakterene og av de vitenskapelig ansatte som deler dem ut. Svært få klager på karakterene, og i de fleste tilfellene opprettholdes den opprinnelige karakteren (f.eks. Gynnild, 2003; Gabrielsen, By & Randen, 2012). Samtidig vet alle at karaktergivning ikke er noen eksakt vitenskap. Flere empiriske undersøkelser har dokumentert til dels betydelig sprik i forskjellige sensorers vurderinger av ett og samme eksamensarbeid (f.eks. Asmyhr, 2011; Bjølseth, Havnes & Lauvås, 2011; Kristiansen, Nätt & Heide, 2013; Raaheim, 2000).

Hva skal karakterene måle? Hvilke kriterier skal anvendes? Hva er en «god» prestasjon og hva står til stryk? Det finnes en rik litteratur på dette feltet, inkludert et eget internasjonalt tidsskrift (*Assessment and Evaluation in Higher Education*), som problematiserer de ulike prinsippene for fastsettelse av karakterer og de mange utfordringene ved de forskjellige systemene. Ikke minst er det en utfordring at karakterer fastsettes på grunnlag av vurderinger som har et genuint skjønnsmessig preg ved seg, og at det alltid vil være rom for menneskelige feil eller svikt i arbeidet med å sette karakterer. I et sosialkonstruksjonistisk perspektiv er de endelige karakterene alltid et utfall av et samarbeid mellom to parter: studenten som har levert eksamensoppgaven og sensorene som har vurdert den.

Resultatet er at de konkrete karakterene ofte problematiseres, iblant med god grunn. De fleste studenter opplever i løpet av sine studier at de under studietiden får én eller flere «gale» karakterer. Blant de vitenskapelig ansatte er anekdotene utallige om meningsløse karakter som er mer eller mindre vilkårlig delt ut til heldige eller uheldige studenter. Det er likevel de mer systematiske svakheter ved karaktersettingen som er mest utfordrende. Det kan være forskjeller i kriteriene for vurdering av faglige prestasjoner på ett og samme kurs, for eksempel mellom forskjellige sensurkommisjoner (Asmyhr, 2011) eller over tid fra ett semester til det neste. Det kan videre skje at de samme studentene vurderes ulikt, for eksempel i forskjellige eksamensformer (muntlig eksamen, skoleeksamen, mappeevaluering osv.) eller av forskjellige sensorer. Kristiansen et al. (2013) viser for eksempel i sin studie av karakterpraksiser ved Høgskolen i Østfold at mappevurderinger i gjennomsnitt gir 0,86 karaktertrinn bedre karakter enn skriftlig eksamen. Det er

også muligheter for andre systematiske nivåforskjeller, for eksempel på tvers av kurs og studieprogram eller mellom læresteder. Internasjonaliseringen av høyere studier har dessuten aktualisert at ulike nasjonale karakterkulturer skaper problemer på den globale utdanningsarenaen.

De siste årene er det lagt ned betydelige ressurser i å skape et mer helhetlig og standardisert karaktersystem. Et viktig utgangspunkt er Bologna-prosessen, som har som mål å etablere et felles europeisk utdanningssystem, «The European Higher Education System», med blant annet et felles karaktersystem (Reinalda & Kulesza, 2006). I Norge er dette arbeidet fulgt opp av Universitets- og høgskolerådets (UHR) arbeid, blant annet gjennom periodiske karakterundersøkelser i de forskjellige fagdisiplinene (UHR, 2013). Det er nedsatt en egen arbeidsgruppe som analyserer og foreslår videre tiltak til koordinering av karakterpraksisene ved lærestedene. I flere dokumenter (bl.a. KD, 2011) understrekes likevel «behov for vidare samordning».

For perioden 2012–2016 har UHR lagt opp til at de forskjellige fagområder skal gjennomføre karakterundersøkelsen hvert femte år i et rullerende system. Karakterene som gis på både masteroppgaver, bacheloroppgaver og enkeltstående skoleeksamener skal analyseres. I tillegg utpekes det flere andre forhold som bør diskuteres: kvalifikasjonsrammeverket, de fagspesifikke karakterbeskrivelsene på masternivå og bestått/ikke bestått. Det åpnes også for at fagene kan legge vekt på egne, mer fagspesifikke problemstillinger (UHR, 2012). Først ute (i 2012) var de samfunnsvitenskapelige fagene. Seks fag deltok: sosiologi, statsvitenskap, samfunnsøkonomi, sosialantropologi, utviklingsforskning og utviklingsgeografi.

I denne artikkelen skal jeg diskutere resultatene fra ett av disse fagene – sosiologi – der jeg selv deltok som leder for faggruppen som gjennomførte evalueringsarbeidet. Mens man her tidligere hadde lagt vekt på masterarbeidene, ble oppmerksomheten denne gangen rettet mot bacheloroppgavene. Utgangspunktet er seks panelsensorers vurderinger av tolv tilfeldig utvalgte bacheloroppgaver som i løpet av 2012 var innlevert ved ett av landets seks læresteder som tilbyr bachelorstudier i sosiologi. Med utgangspunkt i dette empiriske materialet, diskuteres den følgende problemstillingen: *Hvor konsistente er ulike sensorers vurderinger av ett og samme eksamensarbeid?*

Her vil jeg spesielt diskutere i hvilken grad eventuelle inkonsistenser er knyttet til karakter*nivå*, eller om det også er betydelige forskjeller mellom sensorenes innbyrdes *rangering* av eksamensarbeidene. Jeg vil videre drøfte aspekter som kan forklare eventuelle avvik i vurderinger av bacheloroppgavene mellom sensorene. Er det *ulike* og/eller *uklare* kriterier som forklarer avvikene i karaktersettingen, eller er problemet snarere fortolkningen av kriteriene?

Artikkelens struktur

I neste del av artikkelen presenteres datamaterialet som ligger til grunn for analysene, statistikk fra det sosiologiske sensorpanelet og panelsensorenes kvalitative vurderinger av de tolv oppgavene. Deretter presenteres resultatene. I artikkelens diskusjonsdel vurderer jeg resultatene fra det sosiologiske karakterpanelet og antyder noen mer allmenne problemer

knyttet til arbeidet mot standardisering av karakterpraksiser. Jeg drøfter også sentrale muligheter, utfordringer og problemer ved det pågående arbeidet for standardisering av karakterer i høyere utdanning.

Metode og materiale

Det nasjonale fagrådet i sosiologi nedsatte høsten 2012 et sensurpanel[1] med deltakere fra alle de seks høyere lærestedene som tilbyr bachelorstudier i sosiologi, det vil si universitetene i Tromsø, Nordland, Trondheim, Bergen, Oslo og Stavanger. Disse har relativt likeartede bachelorstudier, selv om det er enkelte forskjeller. Blant annet varierer omfanget på bacheloroppgavene fra 10 til 20 studiepoeng. Det er også varierende omfang og form på veiledningen som studentene får underveis i arbeidet med oppgavene. De fleste lærestedene gjennomførte en muntlig eksamen som kunne justere vurderingen av det skriftlige arbeidet.[2]

Materialet for undersøkelsen var to bacheloroppgaver fra hvert av de seks lærestedene. Oppgavene var tilfeldig valgt ut av lærestedenes administrasjoner, uten noen som helst føringer for hvilke oppgaver som skulle velges. De representerer derfor et bredt spekter av sosiologiske arbeider, både med tanke på faglig nivå, tematikk og teoretiske og metodiske innfallsvinkler.

De tolv oppgavene ble sendt ut til og vurdert av seks panelsensorer, én fra hvert deltakende lærested. Jeg var selv representant for eget lærested og har slik bidratt til dataproduksjonen, uten at dette påvirker materialet eller analysene på avgjørende måter. I ett tilfelle hadde panelsensoren tilfeldigvis deltatt i den opprinnelige sensuren av den ene oppgaven fra eget lærested, men heller ikke dette skal kunne påvirke analysene som presenteres i denne artikkelen.

Panelsensorene hadde til sammen godt over hundre års erfaring fra undervisningsarbeid på sine læresteder, og representerte et bredt spekter når det gjaldt både faglig bakgrunn og pedagogisk og forskningsmessig erfaring. De satte først individuelle karakterer på alle oppgavene. I neste omgang møttes gruppen til et dagslangt møte. Panelsensorene måtte her fremme og begrunne forslagene til karakterer, før man diskuterte seg frem til en samlet karakter for hver av oppgavene. Disse diskusjonene ble nedtegnet og danner grunnlaget for artikkelens kvalitative diskusjoner.

De individuelle og kollektive karakterene ble endelig sammenlignet med de opprinnelige karakterene, som i utgangspunktet var kjent kun for sensoren fra lærestedet.

Det statistiske materialet er for lite til å trekke entydige konklusjoner om karaktersetting, verken i sosiologifaget eller innen akademia mer allment. Ulike sensorer, fag, nivåer, disipliner og læresteder har alle sine særtrekk som preger prosessene med å sette karakterer. Det er også problemer knyttet til det kvalitative datamaterialet, panelsensorenes diskusjoner, blant annet fordi også dette representerer et relativt lite materiale. Artikkelens empiriske grunnlag representerer likevel et konstruktivt utgangspunkt for å diskutere sensorreliabilitet innen sosiologifaget og, mer allment, identifisere prosessene som bidrar til forskjellige karakterpraksiser.

Jeg mener derfor at erfaringene som presenteres i denne artikkelen bidrar med viktige innspill til den faglige debatten om karakterer i sosiologifaget, og også i det norske utdanningssystemet mer generelt. Karakterpraksisene som dokumenteres i denne artikkelens case, de norske sosiologiske bacheloroppgavene, er «overførbare» (jf. Lincoln & Guba, 1985) for tilsvarende diskusjoner innen andre fag, både samfunnsvitenskapelige og andre tradisjoner.

Resultater

Denne seksjonen innledes med en deskriptiv presentasjon av panelsensorenes karakterforslag. Deretter analyserer jeg graden av konsistens, både med tanke på *nivå* og *rangering*, og i hvilken grad spriket mellom karakterer og sensorer innebærer en svekkelse av karaktersystemets robusthet. Til slutt analyseres disse resultatene ut fra panelsensorenes kvalitative diskusjoner.

I tabell 1 presenteres panelsensorenes individuelle og sensurpanelets felles vurderinger av oppgavene, de opprinnelige karakterene og sammenfattende statistikk. Tabellens kolonner vises de seks panelsensorenes (a–f) vurderinger av oppgavene (kol. 2–7). De vurderte oppgavene er nummerert fra 1 til 12, med angivelse av lærested (a–f) og om det var den første eller andre oppgaven fra dette lærestedet i parentes. Panelsensor a vurderte for eksempel den fjerde oppgaven, som var den andre oppgaven fra lærested b (dvs. oppgave 4), til karakteren B+.

Panelsensorene ga i utgangspunktet «rene» karakter, men kunne i tvilstilfeller angi plusser eller minuser for å signalisere retning på karakterene. Karakterene i skraverte felter er karakterer som ble gitt til oppgaver ved egen institusjon, for eksempel der representanten fra lærested a vurderte oppgaver fra lærested a, og for øvrig derfor også kjente til de opprinnelige karakterene.

I de to siste kolonene rapporteres de opprinnelige karakterene (kol. 8) og endelig sensurpanelets felles forslag til karakter (kol. 9). Den opprinnelige karakteren for noen av lærestedene (A, C og D) er karakteren på det skriftlige arbeidet *etter* muntlig eksamen. I noen tilfeller kan karakteren ha blitt endret som resultat av den muntlige eksamenen, men panelets medlemmer fra disse lærestedene antar at dette skjer i få tilfeller, og i så fall stort sett ved at karakteren justeres opp etter muntlig eksamen. I teorien kan de faktiske karakterene på oppgavene (før muntlig) derfor være noe lavere enn det som er rapportert i tabell 1.

Panelsensorene forsøkte å enes om én samlet karakter, men dette viste seg umulig for oppgave 2. Her delte panelet seg på midten, med tre hver for henholdsvis C og D.

I tabellens seks nederste rader er forskjellig oppsummerende statistikk beregnet. I raden «Totalt» har man omregnet panelsensorenes bokstavkarakterer til tallverdier (A = 1, B = 2 osv.) og summert disse. I den neste raden er gjennomsnittet regnet ut. De to neste radene angir den *totale* forskjellen mellom panelsensorenes karakter og henholdsvis den opprinnelige karakteren og sensurpanelets felles karakterforslag. Dernest vises prosentandelen A- og B-karakterer som ble gitt av hver av panelsensorene. Den siste raden angir spennet i hvilke karakterer som ble brukt.

Tabell 1. Oversikt over panelsensorenes karakterforslag, sensurpanelets felles forslag og de opprinnelige kommisjonenes karakterer

Oppgave nr.	Panel-sensor a	Panel-sensor b	Panel-sensor c	Panel-sensor d	Panel-sensor e	Panel-sensor f	Oppr. komm.	Sensur-panel
1 (a-I)	D	C	D	C–	D–	C	D	D
2 (a-II)	B–	C–	D	C	D +	D –	B	C/D
3 (b-I)	D–	D–	E	D	F	D	D	D
4 (b-II)	B +	B	C	B	C –	B +	B	B
5 (c-I)	B +	B	B	C	C +	B +	B	B
6 (c-II)	A–	B +	B	A	A	A	B	A
7 (d-I)	A–	B +	B–	C	C	B +	C	B
8 (d-II)	B–	B	C	C	C–	C	C	C
9 (e-I)	A	B +	B	B	B +	B +	A	B
10 (e-II)	C +	C	C	B	C–	D +	D	C
11 (f-I)	C–	C	D +	C	D +	C	B	C
12 (f-II)	B–	B–	C–	C	C	B–	C	C
Totalt (A = 1 osv.)	27	30	37	32	39	32	32	32,5
Gj.snittskarakter	2,3	2,5	3,1	2,7	3,3	2,7	2,7	2,7
Avvik fra oppr. karakter	6	8	7	7	8	6		
Avvik fra felles karakter	5	5	5	5	6	4		
Prosentandel A/B	67%	58%	33%	33%	16%	67%	50%	42%
Spenn i karakterene	A–D	B–D	B–E	A–D	B–F	A–D	A–D	A–D

Manglende konsistens

Resultatene som presenteres i tabell 1 gir et klart svar på artikkelens primære problemstilling – samsvaret i karaktergivningen. Panelsensorenes vurderinger av oppgavene er svært lite konsistente, og tabellen viser at det kan være til dels betydelig sprik i karakterforslagene som forskjellige sensorer setter på bacheloroppgaver. I så mange som halvparten (6/12) av vurderingene ga sensorpanelet en annen felles karakter enn den opprinnelige karakteren. I ett tilfelle (oppgave 2) var avviket på halvannen karakter.

De seks panelsensorene var videre aldri internt helt enige i fastsettelsen av karakteren. I ett tilfelle (oppgave 9) var fem av seks panelsensorer enige, men ellers var spriket mellom panelsensorene alltid større. I fire tilfeller ble tre forskjellige karakterer foreslått (oppgavene 2, 3, 7 og 10). Det mest markante spriket gjaldt oppgave 2. Her ble det foreslått både en svak B og en svak D.

Det er videre betydelige forskjeller mellom panelsensorenes individuelle karakterer og de som ble gitt av sensurpanelet i fellesskap samt mellom panelsensorenes karakter og

Tabell 2. Forskjeller mellom panelsensorenes karakterforslag og sensurpanelets felles forslag og de opprinnelige kommisjonenes karakter

Avvikene størrelse	Avvik mellom panelsensorer og opprinnelig karakter	Avvik mellom panelsensorer og sensurpanelets felles karakter
0 karaktertrinn	29	42
0,5 karaktertrinn		5
1 karaktertrinn	35	22
1,5 karaktertrinn		1
2 karaktertrinn	8	1
3 karaktertrinn	0	0
4 karaktertrinn	0	0
5 karaktertrinn	0	0
Totalt	72	72

karakterene som ble gitt av de originale kommisjonene. Dette vises i tabell 2. Her presenteres forskjellene mellom de 72 enkeltkarakterene som ble gitt av panelsensorene (6 sensorer, 12 oppgaver), sensurpanelets felles karakter og de opprinnelige karakterene.

I 40 prosent av tilfellene (29 karakterer) endte panelsensorene på samme resultat som de opprinnelige kommisjonene. Majoriteten av panelsensorenes vurderinger var imidlertid forskjellige fra de opprinnelige karakterene. I nesten halvparten av vurderingene (49 prosent, 35 karakterer) forslo panelsensorene en karakter som lå ett trinn unna den opprinnelige kommisjonens karakter. I hvert tiende tilfelle (11 prosent, 8 karakterer) avvek de alternative forslagene to trinn på karakterskalaen. To panelsensorer avvek fra de opprinnelige karakterene på åtte av de tolv oppgavene. Også de andre panelsensorene avvek på minst seks av oppgavene.

Klare nivåforskjeller

Det er videre klare *nivå*forskjeller mellom panelsensorene. Den «strengeste» panelsensoren ga karakteren 3,3 i gjennomsnitt (som tilsvarer en svak C) og vurderte oppgavene til A eller B i kun 16 prosent av tilfellene. Den «snilleste» panelsensoren lå jevnt over én karakter høyere (2,3, som tilsvarer en svak B) og ga karakterene A eller B i 67 prosent av tilfellene.

Dette kan skyldes individuelle forhold ved panelsensorene. Blant annet ser man en mulig tendens til at de deltakerne i sensurpanelet som hadde lengst undervisningserfaring, om enn ikke nødvendigvis med veiledning/evaluering av bacheloroppgaver eller lignende undervisningsarbeid, ga litt bedre karakterer enn de andre deltakerne i panelet. Alternativt kan man se panelsensorene som representanter for lokale karakterkulturer, slik at deres karakternivå reflekterer nivået ved deres læresteder. I så fall er det klare forskjeller i hvordan karakterskalaen praktiseres omkring i landet.

Det var naturlig nok mindre sprik mellom karakterforslagene til de enkelte panelsensorene og sensurpanelets kollektive forslag. Her var det bare i ett tilfelle et sprik på to karaktertrinn (oppgave 3), for øvrig den eneste oppgaven der man diskuterte en mulig strykkarakter.

Ulik rangering

Analysen ovenfor viser at panelsensorene ofte setter forskjellig karakter på én og samme oppgave. Det kan skyldes enten mer generelle nivåforskjeller (sensor x gir systematisk bedre/svakere karakterer) eller innbyrdes ulike vurderinger av oppgavene (sensorene x og y vurderer forholdet mellom oppgavene z og w forskjellig) (Bjølseth et al., 2011). Jeg skal kort diskutere dette spørsmålet med utgangspunkt i tabell 3. Her er de absolutte karakterene erstattet med rangerte karakterer: Karakterene fra hver panelsensor, sensurpanelet og de opprinnelige karakterene er erstattet med rangerte plasseringer fra 1 til 12 for hver enkelt oppgave. Oppgaven som fikk den beste karakteren har fått verdien 1, og den dårligste har fått verdien 12. For eksempel ser man at den første panelsensoren (panelsensor a) ga sin aller beste karakter (ren A) til oppgave 9, rangerte oppgavene 6 og 7 som de nest beste (A–) og så videre.

Fordelen med en slik presentasjonsform er at man tar bort effekten av *nivå*forskjeller mellom panelsensorenes vurderinger, sensurpanelets felles forslag og de opprinnelige kommisjonenes karakterer.

Tabell 3. Oversikt over rangerte karakterforslag: panelsensorenes forslag, sensurpanelets felles forslag og de opprinnelige kommisjonenes karakterer

	Panel-sensor a	Panel-sensor b	Panel-sensor c	Panel-sensor d	Panel-sensor e	Panel-sensor f	Oppr. komm.	Sensur-panel
9 (e-I)	1	1	1	2	2	2	1	2
2 (a-II)	6	11	10	5	9	11	2	10
4 (b-II)	4	4	5	2	6	2		2
5 (c-I)	4	4	1	5	3	2		2
6 (c-II)	2	1	1	1	1	1		1
11 (f-I)	10	8	9	5	10	7		6
7 (d-I)	2	1	4	5	4	2	7	2
8 (d-II)	6	4	5	5	6	7		6
12 (f-II)	6	7	8	5	4	6		6
1 (a-I)	11	8	10	11	11	7	10	11
3 (b-I)	12	12	12	12	12	11		11
10 (e-II)	9	8	5	2	6	10		6

Resultatene er tvetydige. Oppgaven som opprinnelig fikk den beste karakteren (oppgave 9), rangeres også svært høyt (som nr. 1 eller 2) av alle sensorene. Det samme gjelder i motsatt ende av skalaen: Alle panelsensorene rangerer oppgave 3 som den dårligste (ev. på delt sisteplass).

Det er større sprik for enkelte av de andre oppgavene. Oppgave 2, som fikk den nest beste karakteren i den opprinnelige vurderingen, vurderes som blant de aller svakeste av flere av panelsensorene. Oppgave 10 ble av én panelsensor vurdert som den nest beste besvarelsen, men rangeres på delt sisteplass ut fra de opprinnelige karakterene.

Dette kan indikere at forskjellige sensorer/læresteder bruker ulike kriterier i vurderingene av bacheloroppgavene. Et godt eksempel er oppgave 2, der kandidatens språklige fremstilling fikk høyst forskjellige konsekvenser for karaktersettingen blant sensorene.

Inkonsistente, men robuste

Resultatene så langt gir et klart svar på problemstillingen. Det er betydelige inkonsistenser i sensorenes karakterpraksiser, og disse er knyttet både til nivå og rangering. Det er derfor grunn til å understreke at panelsensorens karaktersetting, tross det klare spriket mellom de konkrete vurderingene av hver enkelt av de tolv oppgavene, fremstår som rimelig robust på et mer overordnet nivå.

Det er blant annet ingen synlig systematikk i forskjellene mellom de opprinnelige karakterene og sensorenes forslag. I tre oppgavers tilfelle (oppgavene 6, 7 og 10) ga panelsensorene *bedre* karakter enn de opprinnelige kommisjonene. I tre andre tilfeller (oppgavene 2, 9 og 11) ga panelsensorene *dårligere* karakter. Det samlede karaktersnittet til sensurpanelet og de opprinnelige kommisjonene er derfor identisk (2,7, som tilsvarer en sterk C).

En sammenligning mellom de opprinnelige karakterene og sensorenes karakterer viser ikke at noen av lærestedene peker seg ut som for «strenge» eller for «snille». I ett læresteds tilfelle var det samsvar mellom panelsensorenes karakterer og de opprinnelige karakterene på begge de vurderte oppgavene. I et annet læresteds tilfelle var det avvik i karakterene på begge oppgavene. Her fikk én oppgave bedre karakter, den andre dårligere karakter. I de fire øvrige lærestedenes tilfelle var det avvik i karakteren på den ene av oppgavene.

Tabell 4. Fordelingen av panelsensorenes karakterer

	Antall	Prosent
A	6	8,3
B	25	34,7
C	26	36,1
D	13	18,1
E	1	1,4
F	1	1,4
Totalt	72	100,0

I tabell 4 vises fordelingen av de totalt 72 karakterene som panelsensorene ga på oppgavene (12 oppgaver, 6 sensorer). Materialet har en overvekt av gode karakterer, og i hele 43 prosent av vurderingene gis karakterene A eller B.

Halvparten av sensorene i panelet foreslo aldri å bruke karakteren A på noen av oppgavene, og bare to panelsensorer foreslo å bruke E og F (samme oppgave, 3). Dette kan indikere at man ikke bruker hele karakterskalaen.

Kvalitative vurderinger

I den videre diskusjonen skal jeg ta for meg noen mekanismer som kan bidra til å forklare forskjellene i karaktersettingen, jf. problemstillingens siste spørsmål om hvorvidt det er ulike og/eller uklare kriterier som forklarer avvikene i karaktersettingen eller om problemet er knyttet til panelsensorenes fortolkninger av disse kriteriene. Utgangspunktet for denne drøftingen er det sosiologiske sensurpanelets plenumsdiskusjoner, der deltakerne argumenterte seg frem til omforente karakterer.

Hovedfunnet fra disse diskusjonene er at panelsensorene både bruker *uklare*, og til dels også *ulike*, kriterier for hvilke tekstlige fremstillinger som representerer akademiske kvalitet. I sensurpanelets diskusjoner ble det gjerne brukt argumenter som viste til aspekter ved bacheloroppgavene som ikke var klart objektivt definerte, og som vanskelig lar seg forklare eksplisitt og dermed også blir vanskelig tilgjengelig og etterprøvbare. Oppgavene kunne for eksempel bli omtalt som «spenstige», eller de ble priset for å diskutere et spesielt «aktuelt» eller «spennende» tema. I andre tilfeller ble tekstene omtalt som «modne» eller «umodne». Problemstillingene kunne være «utfordrende» eller «lite kreative». Det ble også lagt vekt på om studentene viste «originalitet», «selvstendighet» og evne til «kritisk refleksjon» i oppgaveteksten.

Selv om alle sensorene syntes å være fortrolige med disse vendingene, er det vanskelig å klargjøre hva som ligger i slike uttrykk. Hva er kriteriene for «sosiologisk spenst», hva er det som viser tekstens «modenhet» eller «kreativitet»? Hvordan skal de forskjellige aspektene vektlegges?

Dette er problemer som er uløselig knyttet til samfunnsfagenes karakter. Det er sjelden studentenes utfordring å finne frem til korrekte, eksakte svar i eksamensarbeidene. Det er snarere evnene til fortolkninger, vurderinger og diskusjon som representerer læringsmålene. Det er likevel interessant å se hvor lite eksplisitte slike vurderinger synes å være.

Spesielt balansen mellom «håndverksmessig mestring» og «akademisk kreativitet» ble vurdert på til dels svært forskjellig vis av panelsensorene. Noen eksamensoppgaver holdt ifølge alle panelsensorene et høyt metodeteknisk nivå, men kandidatene anvendte det metodiske grepet til å diskutere relativt enkle forhold. Andre oppgaver utmerket seg med gode teoretiske diskusjoner, men falt igjennom metodisk. Hva er viktigst, og hva bør premieres på karakterskalaen?

Et annet interessant eksempel er spørsmålet om utenlandske studenters prestasjoner. I det tilfellet der det var størst sprik i vurderingene (oppgave 2), hadde den åpenbart utenlandske studenten skrevet en oppgave om utdanningssystemet i eget hjemland. Sensorene var for så vidt enige om at det var en interessant oppgave, i alle fall vurdert fra et norsk perspektiv og av

sosiologer som hadde lite kjennskap til dette landets utdanningssystem. Det var også enighet om at den språklige fremstillingen var svak. Man hadde derimot store problemer med å veie betydningen av disse momentene opp mot hverandre, og det var ingen eksterne kriterier som man kunne enes om å bruke for å vurdere disse aspektene.

Det var også diskusjoner om hvordan studentenes håndtering av forskjellige etiske aspekter skulle vurderes. Én kandidat hadde for eksempel lovet sine informanter anonymitet, mens det i teksten fremkom informasjon som indirekte identifiserte informantene. Hvor alvorlig er slike mangler? I en annen besvarelse hadde studenten gjennomført intervjuer med lekfolk om et tema som enkelte av panelsensorene fant sensitivt (bruk av menneskelig organisk materiale som drivstoff), og de mente studentens intervjuer kunne ha blitt oppfattet som støtende av noen av informantene. Andre vurderte det etiske aspektet annerledes, og la vekt på at det er få begrensninger når det gjelder hvilke tema som kan behandles i en sosiologisk tekst, også bacheloroppgaver.

Andre spørsmål dreide seg om problemstillingenes relevans og deres forankring i den sosiologiske forskningslitteraturen, bruk av faglitteratur i diskusjoner og referanseføringer og andre tekstlige formaliteter.

I sensurpanelets diskusjoner var det ofte enighet om hva som representerer «god sosiologi» og gode studentarbeider. Diskusjonene ledet som regel frem til omforente konklusjoner (dvs. karakterer), men ikke alltid. Som vanlig er i eksamenskommisjoner, bar mange av konklusjonene dessuten preg av kompromisser, der man møttes på midten like mye som man utviklet en felles vurdering av arbeidene. Som nevnt var det ett tilfelle der man heller ikke lyktes med å komme frem til et kompromiss, men endte på splittet karakter etter lange diskusjoner.

Det er derfor interessant at det bare unntaksvis ble referert til de nasjonale beskrivelsene av karakterer. Heller ikke ble det vist til andre eksterne, eksplisitte og autoritative kilder når man argumenterte for hvordan oppgavene skulle bedømmes. I vurderingsarbeidet må sensorene stole på sin egen individualiserte og ofte uartikulerte forståelse av hva som representerer «god sosiologi», fortolke det gitte eksamensarbeidet i lys av denne standarden og endelig overføre resultatet til en bokstavkarakter. Dette reflekterer Asmyhrs (2011) observasjon om at «sensorer anvender sine personlige standarder i vurderingsprosessen» og det dermed uunngåelig oppstår variasjoner i karaktersettingen mellom sensorene.

Merk at også det siste steget er problematisk, ettersom karakterene åpenbart har forskjellig valør ved lærestedene. Noen av panelsensorene formidlet at en C gjerne ble oppfattet som en dårlig karakter i deres hjemlige fagmiljøer, mens andre mente at C var en «grei» karakter. Ved noen læresteder er oppfatningen at C gjerne gis til studenter som ikke skal/kan/bør fortsette med en forskerkarriere, mens C på masteroppgaven ikke er en absolutt hindring for å tas opp på ph.d.-programmet andre steder.

Diskusjon

Resultatene viser at det er betydelige variasjoner i hvordan panelsensorene i sosiologifaget vurderer bacheloroppgavene. Inntrykket styrkes av resultatene fra de andre samfunnsfagene som gjennomførte tilsvarende, men mindre omfattende vurderinger av karaktersettingen

i sine fag. I sosialantropologifaget vurderte to tomannskommisjoner seks oppgaver hver. Når det gjaldt åtte av oppgavene, var det sprik mellom sensurpanelets felles karakter og de opprinnelige karakterene (Wikan, Olsen, Longva & Johansen, udat.). Innen utviklingsforskning ble tre typer oppgaver vurdert: et enkeltstående kurs, bacheloroppgaver og masteroppgaver. Når det gjaldt alle tre oppgavetypene, var det klare forskjeller mellom sensurpanelets vurderinger og de opprinnelige karakterene, men i varierende grad (Stølen, udat.). Samme tendens, men svakere, vises også innen utviklingsgeografi. Her ble 15 skoleeksamener fra bachelorstudiet ved fem læresteder vurdert. I 12 av 15 tilfeller var det samsvar mellom de tre panelsensorenes og de opprinnelige kommisjonenes karakterer. I ett tilfelle var det et sprik på to karakterer mellom den opprinnelige karakteren (D) og panelsensorenes vurdering (Fløysand, 2013). Sensurpanelet i utviklingsgeografi forklarer divergensen med at det hersker uklarhet om hva som kvalifiserer til stryk: Den aktuelle oppgaven var ikke uten faglige kvaliteter, men manglet etter panelsensorenes vurdering relevans med tanke på oppgaveteksten. Selv om eksamensformen her er annerledes (skoleeksamen vs. bacheloroppgave), viser resultatet at problemene knyttet til vurderingsreliabilitet ikke er knyttet til eksamensform, men er mer generiske. Dette understrekes også av samfunnsøkonomenes vurderinger. De konsentrerte seg om oppgaver som hadde fått karakteren E av de opprinnelige kommisjonene. Tolv oppgaver fra seks læresteder ble vurdert av to panelsensorer hver; det var altså til sammen 24 vurderinger. I 10 av disse vurderingene (42 prosent) var konklusjonen at oppgavene kvalifiserte til strykkarakter, og i sensurpanelets rapport antydes det at det kan «være systematisk forskjell mellom institusjoner i hvor en lar grensen mellom E og F gå» (Moilanen, udat.).

Erfaringene fra de andre fagenes evalueringer reflekterer i høy grad de rapporterte resultatene fra sosiologi. I undersøkelsene er mønsteret at de enkelte panelsensorenes vurderinger av oppgavene i mange tilfeller varierer, og at deres felles konklusjon ofte også avviker fra de opprinnelige kommisjonenes karakterer. Det er samtidig små variasjoner, sjelden mer enn ett trinn på karakterskalaen. Dette gir grunn til å anta at funnene i denne studien har gyldighet utover sosiologifaget.

Det kan være mange forhold som forklarer sprikene i sensorenes vurderinger, både i den regulære, daglige praksisen omkring på lærestedene og i de refererte sensurpanelene. Spesielt når det gjelder sensurpanelene, er det et viktig moment at de ofte har vurdert oppgaver uten inngående kjennskap til den faglige konteksten for studentenes eksamensarbeider: læringsopplegg, kunnskaps- og ferdighetsmål, faglig nivå på studentgruppen som helhet, osv. Panelsensorene representerte institusjoner som praktiserer forskjellige opplegg for bacheloroppgavene, som nevnt med tanke på omfanget (antall studiepoeng) og undervisningsform (f.eks. veiledning). Dette fører trolig til ulike forventninger til hva en bacheloroppgave «er» og hva som er adekvate vurderinger av oppgavenes kvalitet.

Til dels har dette vært et tilsiktet mål med karakterundersøkelsene: Universitets- og høgskolerådet (UHR) har ønsket å vurdere studentenes prestasjoner ut fra nasjonale standarder, løsrevet fra de lokale forhold ved institusjonen. Dette er naturlig siden UHR arbeider med å utforme og spre klare beskrivelser av betingelsene for å gi de forskjellige karakterene. Erfaringene fra sensurpanelene viser imidlertid at hver enkelt deltaker likevel

opererer med individuelle standarder i vurderingene av eksamensarbeidene og hvilke karakterer disse fortjener. Disse standardene synes i høy grad å være implisitte, ved at deltakerne i sensurplanelet synes å ha til dels divergerende oppfatninger av hva som representerer «god» sosiologi. Det kan med andre ord synes som om den eksplisitte, sentraliserte og formaliserte kunnskapen om karakterskalaen ennå ikke har forankret seg i de implisitte, lokale og tause karakterkulturene omkring på lærestedene.

I praksis møter de regulære eksterne sensorene de samme utfordringene, ikke minst den første gangen man sensurerer et kurs ved et annet lærested. Det er derfor trolig at spriket i sensurpanelenes vurderinger speiler situasjonen i det akademiske hverdagslivet.

Konklusjoner

I artikkelen har jeg diskutert sensorreliabiliteten i sensorers vurderinger av eksamensarbeider. Den sentrale konklusjonen fra studien er at karaktersettingen på bacheloroppgavene i sosiologifaget – og trolig også på oppgaver på andre nivåer og i andre fag – preges av tilfeldigheter, både i utfall (karakter) og i bruken av kriterier og vektlegginger som fører frem til sensorenes konklusjoner.

Problemstillingen aktualiserte spesielt om hvorvidt inkonsistente karakterer kan knyttes til sensorenes forståelse av hva som bør være det generelle karakter*nivået* eller forskjeller i deres interne *rangering* av eksamensarbeidene. Også her gir studien klare resultater. Analysene viser for det første systematiske forskjeller i karakternivået på tvers av panelsensorer, om enn ikke i samme grad på tvers av institusjonene. Noen av panelsensorene lå markant høyere på skalaen enn de andre. For det andre viser analysene at det også er klare avvik i den interne rangeringen av eksamensarbeidene.

Inntrykket bekreftes av andre datakilder, både de andre disiplinenes sensurpaneler og øvrige caseundersøkelser av karaktersettingen på de høyere utdanningsinstitusjonene i Norge (Asmyhr, 2011; Bjølseth et al., 2011; Kristiansen et al., 2013).

Videre peker denne studien på flere viktige forklaringsfaktorer bak de ulike karakterpraksisene. Panelsensorene hadde ulike vurderinger av hva som konstituerer en «god» sosiologisk tekst, både med tanke på substansielle, metodiske, etiske og formale aspekter. Det som noen panelsensorer gjenkjenner som «spenstig» sosiologi, oppfattes av andre som «umodent». Artikkelens drøftinger identifiserer også andre faktorer som trekkes inn i vurderingene, så som betydningen av studentens livsbiografi.

Dette aktualiserer utfordringene ved at man ikke har utviklet klare nasjonale retningslinjer for bacheloroppgavene, verken innenfor sosiologifaget eller mer allment innen det samfunnsvitenskapelige utdanningsfeltet. Man har derfor heller ingen eksplisitte, nasjonale standarder for hva som er en god fagtekst.

Det blir da selvsagt også vanskelig å anvende den Bologna-initierte nasjonale karakterskalaen, trass i dens eksplisitte utforming. Som Gynnild (2013) påpeker er det uansett innebygde problemer med det prinsippet om kriteriebaserte vurderinger som ligger til grunn for UHRs arbeid med utvikling av mer sammenfallende karakterkulturer ved landets høyere læresteder. Gynnild viser at ulike tolkninger av kriteriebasert vurdering i høy grad vil påvirke karakterfordelingene, og hevder at så lenge prinsippene for karaktervurderingene «ikke

nedfelles i felles forståelige retningslinjer for praksis, kan man forvente store sprik på handlingsnivå» (Gynnild, 2013, s. 40).

Videre forskning på feltet bør derfor utforske de sosiale prosessene som «oversetter», mer eller mindre vellykket, de eksplisitte og formelle nasjonale standardene for karakterfastsetting til faktiske karakterpraksiser omkring på de lokale lærestedene. Dette inviterer først og fremst til mer inngående kvantitative analyser av statistisk materiale: Hvor store er sprikene i karakterene mellom sensorer, kommisjoner, fag, disipliner og læresteder? I hvilken grad er inkonsistens knyttet til nivå eller rangering? Her foreligger det et betydelig datamateriale gjennom UHRs databaser for karakterstatistikk. Trolig vil det også være fruktbart med analyser av mer spissete kvantitative data, for eksempel mer systematiske sammenligninger mellom individuelle sensorers karakterforslag og kommisjonenes endelige konklusjoner, slik jeg har vist i denne artikkelen. Det vil imidlertid også være fruktbart med flere kvalitative analyser av karaktersettingen, slik som demonstrert i denne artikkelen. Slik vil man kunne få en bedre forståelse av mekanismene som ligger til grunn for de individuelle sensorene og sensorkommisjonenes konklusjoner.

Om artikkelen

Artikkelen er skrevet som en oppfølging til forfatterens arbeid med den UHR-pålagte evalueringen av bacheloroppgaver i sosiologifaget i 2012. Deler av teksten er en utvidet og revidert versjon av sensurpanelets rapport, som ble utarbeidet av undertegnende (Rye, 2012). Jeg takker de øvrige medlemmene av sensurpanelet for tillatelse til å referere fra vårt arbeid og for gode innspill i prosessen, både i arbeidet med den opprinnelige rapporten og i arbeidet med det foreliggende manuskriptet. Analysene som presenteres her, inkludert eventuelle feil, står uansett for forfatterens regning.

Litteratur

Asmyhr, M. (2011). Om vurdering av essaybesvarelser i høyere utdanning – en studie av vurderer-reliabilitet. *Uniped*, *34*(4), 17–33.

Bjølseth, G., Havnes, A. & Lauvås, P. (2011). Lavt sensorsamsvar – kan det bedres? *Uniped*, *34*(4), 4–16.

Black, P. (1998). *Testing: Friend or Foe?* London: Falmer Press.

Fløysand, A. (2013). Rapport: karakterundersøkelse fra tre kurs i utviklingsgeografi på bachelornivå ved NTNU, UiO og UiB. Upublisert.

Gabrielsen, E., By, I. Å. & Randen, M. (2012). Klagesensurordningene – hvordan fungerer de? *Uniped*, *35*(1), 5–21.

Gynnild, V. (2003). *Når eksamen endrer karakter. Evaluering for læring i høyere utdanning*. Oslo: Cappelen Akademisk Forlag.

Gynnild, V. (2013). «Kriteriebasert vurdering» – hva innebærer det i praksis? *Uniped*, *36*(1), upaginert.

Kristiansen, M., Nätt, T. H. & Heide, C. F. (2013). Kvantitativ undersøkelse av mulige sammenhenger mellom vurderingsform og karakterer i høyere utdanning. *Uniped*, *36*(2), upaginert.

Kunnskapsdepartementet (KD) (2011). Retningslinjer for det nasjonale karaktersystemet. Brev til universitet og høgskolar, 13.12.2011.

Larsen, S., Johnsen, B. H. & Pallesen, S. (2006). Er opptaket til profesjonsstudiet i psykologi reliabelt? *Tidsskrift for Norsk Psykologiforening*, *43*(3), 221–225.

Lincoln, Y. S. & Guba, E. G. (1985). *Naturalistic Inquiry*. Newbury Park: Sage.

Moilanen, M. (udat.). Karakterundersøking 2012. Samfunnsøkonomi. Upublisert.

Norges teknisk-naturvitenskapelig universitet (udat.). Karakterskalaen (webadresse: http://www.ntnu.no/studier/eksamen/karakterskala, lest 17.08.2011).

Raaheim, A. (2000). En studie av inter-bedømmer reliabilitet ved eksamen på psykologi grunnfag. *Tidsskrift for Norsk Psykologiforening*, *37*(3), 203–213.

Reinalda, B. & Kulesza, E. (2006). *The Bologna Process – Harmonizing Europe's Higher Education: Including the Essential Original Texts*. Opladen: Barbara Budrich Publishers.

Ritzer, G. (1996). *Sociological Theory*. London: McGraw-Hill.

Rye, J. F. (2012). Rapport fra karakterkommisjon i sosiologi 2012. Upublisert.

Sadler, R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education*, *34*(7), 807–826.

Stølen, A. (udat.). Karakterundersøkelsen 2012–2016: Rapport fra Nasjonalt fagråd for utviklingsforskning. Upublisert.

Universitets- og høgskolerådet (UHR) (2012). Karakterundersøkingar 2012–2016. *Brev til nasjonale fakultetsmøte, nasjonale råd, nasjonale profesjonsråd og nasjonale fagråd*, 29.02.2012.

Universitets- og høgskolerådet (UHR) (2013). Karaktersystemet (webadresse: http://www.uhr.no/ressurser/temasider/karaktersystemet_1, lest 27.07.2013).

Wikan, U., Olsen, B., Longva, A. N. & Johansen, S. (udat.). Karakterundersøkelse sosialantropologi. Kommisjonens rapport. Upublisert.

Forfatterpresentasjon

Johan Fredrik Rye er professor i sosiologi ved Institutt for sosiologi og statsvitenskap, Norges teknisk-naturvitenskapelige universitet. Ryes forskningsfelt er arbeidsinnvandring, innenlandsk migrasjon og sosial ulikhet.

Noter

1 Av hensyn til språklig klarhet brukes termen *sensurpanel* her synonymt med Universitets- og høgskolerådets term *karakterkommisjon*. Medlemmene av sensurpanelene benevnes tilsvarende som *panelsensorer*.

2 Universitetet i Stavanger (UiS) deltok med bacheloroppgaver fra sitt tidligere bachelorstudium i personalledelse fordi deres nye sosiologistudium ennå ikke hadde resultert i ferdige bacheloroppgaver. UiS vurderer disse oppgavene som representative for det nye sosiologiske bachelorprogrammet. Ved NTNU var det ikke muntlig eksamen for bacheloroppgavene i 2012, men dette er siden gjeninnført som eksamensform.