En eller to sensorer?

Et eksperiment i sosial interaksjon

BJØRN ERIK RASCH
b.e.rasch@stv.uio.no

SARA KRISTINE ERIKSEN
sara.eriksen@stv.uio.no

SINGLE- OR DOUBLE-MARKING OF STUDENT EXAMINATION PAPERS?
AN EXPERIMENT ON SOCIAL INTERACTION

Over the past few years, the issue of using one or two examiners marking student examination papers has been the subject of much debate in Norway. In this article, we study one aspect of this controversy by conducting an experiment on two groups of examiners marking final examination papers from an introductory course on political science. One group did the grading alone, while the other was assigned a second examiner to work with. Our aim was to look into the extent to which suggestions from the first examiner influence the views of the second. The results indicate that an erroneous judgment by an examiner (for example, the suggested mark is too generous) tends to be «inherited» rather than corrected by the second examiner. We argue that the reason is either social pressure (cf. Solomon Asch) or informal adjustments in a situation which is unstructured and ambiguous (cf. Muzafer Sherif).

Keywords:
- marking of student examination papers
- social interaction
- experiment

TIDSSKRIFT FOR SAMFUNNSFORSKNING
VOL 50, NR 3, 293–315

Kvalitetsreformen har ført til en rekke pedagogiske endringer. Fagtilbudene er blitt modulisert, det er innført flere eksamener, prøver, kvalifiseringsoppgaver og obligatoriske innleveringer, studentene får mer regelmessig oppfølging med hyppigere tilbakemeldinger og vurderingsformene er blitt langt mer varierte (Dysthe 2007; Dysthe et al. 2006). Mange av endringene er svært arbeids- og ressurskrevende, og har ledet til et press i retning av å utvikle enklere former for sensur enn det som var vanlig tidligere. Hovedbildet ved landets universiteter og høyskoler er imidlertid stor grad av variasjon når det gjelder sensurgjennomføring (Solum 2005). Det benyttes ordninger med en eller to sensorer, og varierende innslag av eksterne sensorer eller kontrollører.[1] Der det brukes én (intern) sensor, er det utviklet en eller annen form for tilsynssensorordning med ekstern deltakelse, og gjerne elementer av stikkprøvekontroll, krysslesing eller lignende. Det synes å være en tendens i retning av mer utstrakt bruk av intern sensur i fag som har et stort antall bedømmelser per student, dvs. der det er mange deleksamener og innleveringer (Solum 2005:4).

Spørsmålet om en eller to sensorer har vært gjenstand for offentlig debatt, med ganske steile fronter. Noen legger vekt på at det ut fra rettssikkerhetshensyn bør være to sensorer, mens andre hevder at sammenfallet i vurderingene som ulike sensorer foretar er så stort at flere sensorer er unødvendig (eller at kostnadene med flere sensorer dermed ikke lar seg forsvare). Det er imidlertid lite forskning på området, og vi har ikke vært i stand til å finne analyser fra Norge eller andre land som kaster lys over egenskaper og konsekvenser sensorordninger med henholdsvis en og to sensorer har. Ofte er også ordningene så ulike mellom land at det er begrenset hvilke lærdommer en kan trekke av sammenligninger (se f.eks. Brandt & Stensaker 2005 for ordninger i Sverige, Danmark og England).

Formålet med undersøkelsen er å få en bedre forståelse av samhandlingen i eksamenskommisjoner. Er det slik at karaktersettingen

En stor takk til studiekonsulent Dagfinn Hagen. Hans innsats har vært helt avgjørende for den praktiske gjennomføringen av eksperimentet, og han har deltatt i diskusjonen av alle sider ved opplegget. Professor i psykologi Pål Kraft ga svært nyttige innspill i en innledende fase. Takk til Jon Hovi, Vidar Gynnild, Hans-Kristian Hernes, Arild Raaheim, Anne Julie Semb, Bjørn Stensaker, Karl Halvor Teigen, Jarle Weigård, redaksjonen i TfS og tidsskriftets to anonyme konsulenter for kommentarer. Vi står imidlertid alene ansvarlig for innholdet.

påvirkes av om det er en eller to sensorer som foretar bedømmelsen? I så fall, på hvilken måte påvirkes karakterene? Som et første, prøvende forsøk på å besvare disse spørsmålene, har vi gjennomført et eksperiment der en relativt stor gruppe sensorer vurderer en eksamensbesvarelse på et innføringsemne (bachelor) i statsvitenskap. Det bør understrekes at vi kun tar opp noen få sider ved sensorordninger, og det er flere viktige forhold vi ikke berører. Videre legger vi ikke opp til en normativ diskusjon: vi tar ikke stilling til hvilken sensorordning som alt i alt er «best», men analysen gir kunnskap som kan være av betydning for denne debatten.

Artikkelen er disponert på den måten at vi først diskuterer hvilke forskjeller vi forventer å finne mellom ordninger med en og to sensorer. Deretter redegjør vi for det eksperimentelle designet som er valgt, inkludert mulige etiske betenkeligheter ved undersøkelsen. Neste trinn er å presentere resultatene fra eksperimentet. Til sist drøfter vi relevansen av våre funn i forhold til debatten om antall personer i eksamenskommisjoner.

EN HYPOTESE OM PÅVIRKNING VED KARAKTERSETTING

Det å sette karakter er en beslutning. Hvor kompleks og utfordrende denne beslutningssituasjonen er, vil i høy grad avhenge av eksamensoppgavens art og hvor sterke signaler og føringer sensorene opplever å få i tilknytning til bedømmelsen. Det vil typisk være vanskeligere å vurdere en hjemmeoppgave hvor studenten har valgt tema selv og hvor et eksplisitt sett av vurderingskriterier ikke finnes, enn en eksamensbesvarelse på et område der det kan lages en fasit med riktige svar.

De oppgavene studenter stilles overfor i eksamenssammenheng varierer voldsomt. Det vi har i tankene i denne studien er først og fremst fag og eksamener der det er tale om relativt åpne og generelle spørsmål som inviterer til drøftinger, og som kan angripes og løses på flere måter. Det betyr at det ikke er mulig å lage en fasit i streng forstand, bare veiledninger i hvordan oppgaven bør angripes og hva det vil legges vekt på ved bedømmelsen. Det er mange fag som benytter slike oppgaver. Rommet for skjønn ved vurderingen vil derfor i utgangspunktet være stort. Det vil selvsagt kunne innsnevres ved bruk av detaljerte sensorveiledninger (jf. Baird et al. 2004), men det er lett å tenke seg at ulike sensorer likevel vil ende opp med ulike konklusjoner.

Det foreligger noen tidligere studier av eksamenssensur (Raaheim 2000; Teigen 1986). Det er undersøkelser som gjelder grunnfagsnivået i psykologi, og de er følgelig gjennomført før innføring av bokstavkarakterer – som innebar en dramatisk reduksjon av antallet karaktertrinn – og før oppdelingen i småemner innenfor en ny gradsstruktur. Resultatene kan derfor ikke overføres direkte til dagens studiesituasjon. Raaheim (2000) viser at det til dels er svært dårlig samsvar mellom sensorenes vurderinger (i alt syv sensorer) av de 50 besvarelsene som inngikk i hans analyse (lav inter-bedømmer reliabilitet). I tillegg til at karakterskalaen var annerledes, forelå det ikke sensorveiledning eller andre tilsvarende hjelpemidler for sensorene. Videre var oppgavetekstene svært åpne. Teigen (1986) tegner et bilde med noe mindre avvik mellom sensorer som leser samme besvarelse. Noen spredte analyser fra andre land tyder også på stor spredning ved bedømmelse av mange typer studentarbeider (Brown & Glasner 1999; Rowntree 1987).[2]

Det er grunn til å forvente at sensors kyndighet (ekspertise) og erfaring innenfor det fagfeltet eksamensarbeidet omhandler spiller en viss rolle ved karaktersettingen. Ekspertise er nødvendig for et sikkert faglig skjønn, og gjennom erfaring sosialiseres sensor, tilegner seg de faglige kvalitetsnormene – som ofte kan være både uskrevne og uuttalte – og tilpasser seg bevisst eller ubevist standarder og vurderinger som andre i fagmiljøet over tid observeres å legge til grunn.

Hvis det er eksamenskommisjoner med flere medlemmer som skal sette karakter, er det mulig å organisere dette på en slik måte at sensorene verken diskuterer karaktersettingen seg imellom eller trer i kontakt med hverandre. Sensorene kan for eksempel spille inn sine karakterforslag enkeltvis og uavhengig av hverandre, og så fremkommer den endelige karakteren som en ren sammenveiing eller et gjennomsnitt av sensorenes forslag. Dette kan skje automatisk eller rent administrativt.

Det vanligste i eksamenskommisjoner med flere medlemmer er imidlertid at de fungerer som *kollektiver*. Sensorene har kontakt, og de diskuterer besvarelsen seg imellom før karakter bestemmes. Hvor omfattende disse diskusjonene er, vil blant annet avhenge av hva slags eksamensarbeid som er til bedømmelse, om sensorene i utgangspunktet synes å være enige om karakteren og hvor mange arbeider de skal gjennom på den tiden som står til disposisjon. Likevel må interaksjonen i

eksamenskommisjonen bestå i at den ene sensoren først formidler noen av sine synspunkter eller sitt forslag til karakter, og den andre sensoren responderer ved å si seg mer eller mindre enig.

Som nevnt må samhandlingen i en eksamenskommisjon (som opptrer som kollektiv) i praksis være *sekvensiell*; én må tilkjennegi sine synspunkter på besvarelsen som er til vurdering først. Det første innspillet kan bestå i et konkret karakterforslag, men det kan selvsagt også være langt mer prøvende og upresist. Likedan kan det begrense seg til å angi karakter, men det kan også være langt mer fyldig og inneholde en begrunnelse (og i noen tilfeller er det snarere slik at det antydes en begrunnelse uten klar spesifisering av karaktertrinn). Mulighetene er mange. I fortsettelsen forholder vi oss imidlertid til den reneste, og i en viss forstand enkleste, situasjonen: *Den ene sensoren gir presist uttrykk for hvilken karakter hun eller han mener eksamensbesvarelsen fortjener, og den andre sensoren responderer med sitt forslag til karakter – under den forutsetning at uenighet vil lede til nærmere diskusjon*. For enkelhets skyld vil vi noen steder nedenfor betegne kommisjonsmedlemmene som *førstesensor* og *andresensor*, basert på hvem som kommer med første innspill.

Vil så det første innspillet ha selvstendig betydning for karaktersettingen? Resultater fra noen klassiske eksperimenter gir grunn til å forvente at svaret kan være ja. I en serie eksperimenter på 1930-tallet undersøkte Muzafer Sherif (1936, 1937) sider ved sosial innflytelse. Dette var eksperimenter som gjorde bruk av autokinetiske effekter. Forsøkspersoner ble plassert i et helt mørkt rom og skulle betrakte et enkelt lyspunkt på veggen foran seg. Selv om lyspunktet står helt stille, ser det for betrakteren ut til å flytte på seg. Denne optiske illusjonen ble utnyttet i eksperimentet. Forsøkspersonene ble bedt om å uttale seg om hvor mye lyspunktet flyttet seg, og det var svært stor variasjon i de estimatene som ble rapportert av personer som satt alene. Samtidig utviklet hver enkelt en personlig «norm» for hvor stor bevegelse det dreide seg om når forsøket ble gjentatt flere ganger. Det ble også gjort forsøk med de samme personene i grupper på to og tre i samme rom, hvor gruppemedlemmene anga sine estimater høyt i tur og orden. Bare én i hver gruppe var reell forsøksperson uten informasjon om formålet med eksperimentet (og uten kjennskap til autokinetiske effekter). De øvrige medlemmene var instruert til å gi estimater

som systematisk lå over eller under den personlige normen forsøkspersonen i utgangspunktet hadde utviklet. Ved gjentatte forsøk kunne en observere at forsøkspersonens estimater ble justert i retning av den eller de andre deltakernes anslag. Etter ganske få runder var estimatene omtrent like; det skjedde relativt raskt en form for tilpasning og utvikling av en *sosial norm*.

I Sherifs eksperimenter sto forsøkspersonene overfor en oppgave der det ikke fantes noe korrekt svar på hvor langt lyspunktet flyttet seg. Det vil si: Lyspunktet sto i realiteten stille, men opplevelsen av bevegelse var høyst subjektiv, og varierte fra person til person. Til tross for dette ble det over tid utviklet en felles referanseramme, slik at gruppemedlemmene etter hvert «lærte seg» å bedømme situasjonen temmelig likt. Den informasjonen som gruppemedlemmer tilveiebrakte gjennom sine anslag, inngikk som del av beslutningsgrunnlaget for forsøkspersonene og den bedømmelsen de gjorde av den uklare situasjonen de sto i. Tilpasningsprosessen er mer eller mindre ubevisst. Det kan gis forskjellige tolkninger av hva eksperimentene til Sherif har vist. Det er imidlertid vanlig å legge vekt på at de anskueliggjør «how social norms can arise in groups as a collective response to new, unstructured, ambiguous situations, introducing stable and coherent knowledge of the situation» (Turner 1991:10).

I tillegg til normdannelse kan *sosialt press* være en faktor som i prinsippet har relevans også i forbindelse med eksamenskommisjoner. Tidlig på 1950-tallet gjennomførte Solomon Asch (1952) noen eksperimenter som i en viss forstand er beslektet med de som er beskrevet ovenfor. Denne gangen er det en enkel oppgave forsøkspersonene forholder seg til, og den har et utvetydig, riktig svar. Forsøkspersonene ble presentert for to plansjer. På den ene var det én lang strek, og ellers bare en lys flate. På den andre var det tre streker med markante forskjeller i lengde – den lengste nøyaktig like lang som streken på den første plansjen. Forsøkspersonene skulle så ta stilling til hvilken av de tre strekene som var nærmest den på den første plansjen i lengde. Når plansjene settes ved siden av hverandre er det lett å se hva som er riktig svar, og det er uhyre sjelden at forsøkspersoner som opptrer alene plukker ut feil strek.

I det opprinnelige eksperimentet deltok grupper på fem personer, hvorav bare én var uvitende forsøksperson. Gruppemedlemmene ble

plassert ved et bord, og ble bedt om å svare høyt på hvilke to streker som var like, og de ble bedt om å gjøre det i en rekkefølge som tilsynelatende (for forsøkspersonen) var tilfeldig. Forsøkspersonen fikk likevel en plassering som medførte at de skulle gi svar helt til slutt. Eksperimentene ble innledet med et par runder hvor alle gruppemedlemmene ga riktige svar. I den tredje runden begynner førstemann å tvile, og bruker litt tid før galt svar oppgis. Medlem to, tre og fire, som er instruert på forhånd om hva de skal si, oppgir også galt svar. Hva gjør så forsøkspersonen når egne sanser sier én ting, mens alle andre medlemmer i gruppen klart uttrykker noe annet? Det varierte en god del forsøkspersoner imellom, men i litt over 1/3 av gangene eksperimentet ble gjennomført ble gruppens flertallsoppfatning fulgt og galt svar oppgitt også av forsøkspersonen. Selv i en situasjon som i utgangspunktet er svært enkel, finner vi altså en tilpasning til andres (klart feilaktige) oppfatninger. Eksperimentet er gjentatt senere med mange typer variasjoner, og hovedresultatet står ved lag.

Også eksperimentet til Asch kan tolkes på flere vis, men vanligvis ses tilpasning og konformitet som utslag av opplevd gruppepress; forsøkspersonene er egentlig ikke usikre på hva riktig svar skal være, men lar seg lede eller overbevise til tross for at de vet med seg selv hva riktig svar er (Taylor et al. 2000). Shiller (1995:182) foreslår imidlertid en fortolkning som er informasjonsbasert snarere enn fundert på gruppepress. En rasjonell forsøksperson vil kunne tenke at oppgaven ikke er så lett som den ser ut til, og at sannsynligheten for at alle andre skal ta fullstendig feil er svært liten. Dermed vil en lett trekke den slutning at andres oppfatninger synliggjør at en selv ikke er helt i stand til å løse oppgaven på en kompetent måte, og i stedet følges flertallet.

En studie av Koehler og Harvey (1997) kan også nevnes i denne sammenheng. De finnes en tendens til at folk tror mer på andres vurderinger enn sine egne, trolig fordi de er klar over sin egen usikkerhet, men (feilaktig) tror andre er mindre usikre i sine vurderinger enn de i virkeligheten er.

I Asch-eksperimentene er gruppestørrelse en vesentlig variabel. Det er svært få forsøkspersoner som gir opp det egne sanser forteller hvis bare én person til er med i eksperimentet. Typisk dreier det seg om under fem prosent av tilfellene. Med to personer som står mot forsøkspersonen er det en klar økning i tilpasning eller ettergivenhet,

og med fire medlemmer av gruppen totalt gir forsøkspersonen etter i rundt 1/3 av tilfellene. Økning av antallet gruppemedlemmer ut over dette påvirker imidlertid ikke forsøkspersonens adferd i nevneverdig grad (Asch 1955; Insko et al. 1985).

På hvilken måte er disse eksperimentene relevante for forståelsen av eksamenskommisjoners arbeid? Direkte paralleller finnes neppe. Eksamenskommisjonene vi har for øye har kun to medlemmer, slik at en normalt ikke skulle vente noe stort innslag av gruppepress. Kjennskap til medsensors identitet kan imidlertid lede til et mer eller mindre ubevisst sosialt press.

Typisk vil imidlertid sensorer være høyt kompetente, uavhengige fagfolk som skulle være godt trent i å foreta et selvstendig skjønn – og være i stand til å stole på eget skjønn. Dessuten er det å bedømme en eksamensbesvarelse verken en utfordring som er så åpen og tvetydig som i Sherifs eksperimenter, eller en så enkel oppgave som hos Asch. Kanskje er det rimelig å si at utfordringen ligger et sted midt imellom.

Vi vil likevel anta at den sosiale interaksjonen i eksamenskommisjoner ikke nødvendigvis er helt fri for forhold som preger annen samhandling. Vår påstand er derfor:

HYPOTESE: Det første karakterinnspillet fra en sensor i en eksamenskommisjon påvirker *i seg selv* hvilket forslag den andre sensoren fremmer, og dermed indirekte karakterfastsettelsen.

EKSPERIMENTELT DESIGN

I dette avsnittet beskrives det eksperimentelle designet vi har valgt. Vi ønsker altså å undersøke betydningen av første karakterinnspill i eksamenskommisjoner som består av to personer, noe som kun er aller første trinn i interaksjonen sensorer imellom. Vi har prioritert å få så mange kvalifiserte sensorer som mulig til å lese én og samme besvarelse. Det ville vært urealistisk (og utenfor de økonomiske rammene for prosjektet) å få alle sensorene til å bedømme mange besvarelser; for vårt formål er det bedre med mange sensorer som bedømmer én besvarelse, enn et mindre antall sensorer som vurderer flere besvarelser hver.[3]

Vi har allerede vært inne på noen av valgene som er foretatt. Det er plukket ut en eksamensbesvarelse på et innføringsemne i komparativ politikk ved Institutt for statsvitenskap, Universitetet i Oslo (emnet

STV1300 med fire timers skoleeksamen). Vi valgte et semester med et oppgavesett som var av en slik art at det kunne antas å være ganske mange kvalifiserte sensorer. En relativt kort eksamensbesvarelse ble plukket ut for ikke å gjøre arbeidsbyrden så stor. Besvarelsen var også skrevet med godt leselig håndskrift, for å hindre at sensorer skulle bli usikre på hvilke ord og setninger som faktisk sto på sidene.

Eksamensbesvarelsen hadde i utgangspunktet oppnådd karakteren D. Studenten klaget, men klagekommisjonen opprettholdt D-en. Uavhengig av dette har også emneansvarlig, som ga oppgaven og skrev sensorveiledning, konkludert med en D på besvarelsen. Gitt den fagspesifikke beskrivelsen av karaktertrinnene, læringsmålene for emnet, pensum og sensorveiledningen er det liten tvil om at dette er en rimelig vurdering, selv om det ikke er mulig å snakke om en «riktig» karakter i streng forstand. I den nasjonale, fagspesifikke beskrivelsen av karaktertrinnet D benyttes formuleringer som «ufullstendige pensumkunnskaper» og «begreper, teorier og empirisk kunnskap anvendes ujevnt». I sin begrunnelse for at besvarelsen som er valgt ut ligger på D, har emneansvarlig lagt vekt på «kunnskapssvikt» – «det er manglende pensumkunnskaper som i første rekke trekker ned». Dette er det viktig å merke seg. Besvarelser med andre typer svakheter kunne vært valgt ut. I eksperimentet inngår altså en besvarelse som krever et visst kjennskap til relevant pensum for å kunne gi en trygg bedømmelse.

Det er som nevnt vanskelig å komme helt utenom subjektive elementer ved karaktersetting av den typen det her er tale om, hvor det alltid vil være elementer av skjønn til stede. For å bidra til å redusere rommet for skjønn noe, fulgte detaljert sensorveiledning med besvarelsen. Deltakerne i eksperimentet ble instruert om å gå løs på sensuren nøyaktig på samme måte som de gjør ved ordinær sensur, og heller ikke diskutere besvarelsen med andre.

Det er likevel klart at dette oppdraget skiller seg fra vanlig sensur på flere måter. Én ting er at konklusjonen (karakteren) *ikke* har betydning for kandidaten. Et annet forhold er at hver sensor kun leser én oppgave, og dermed ikke får mulighet til å se den i lys av andre besvarelser og foreta justeringer deretter. Nå er dette likevel ikke så viktig i vår sammenheng, hvor hovedformålet ikke er å få satt «riktig» karakter, men heller å sammenligne vurderinger sensorer (og sensorgrupper) imellom.

Det ble ikke utbetalt ordinært sensorhonorar for jobben, men deltakerne fikk et gavekort som takk for innsatsen. Dette ble sendt ut i forkant sammen med oppgavetekst, besvarelse, sensorveiledning og følgebrev med instruksjoner.

Fra en opprinnelig liste på nærmere 60 navn, endte vi opp med 39 sensorer. Potensielle sensorer fikk forespørsel om å lese en eksamensbesvarelse som ledd i instituttets arbeid med «kvalitetssikring av sensuren».[4] Det var så 37 personer som responderte innen fristen, og det er svarene fra disse som utgjør datamaterialet vi bygger på. Noen av sensorene var tilknyttet Institutt for statsvitenskap ved UiO, men flertallet var fra eksterne miljøer. Det var sensorer på alle stillingsnivåer, og fra andre universiteter (NTNU, UiB og UiT).

Sensorene ble delt *tilfeldig* inn i to grupper – gruppe B (eksperimentgruppe) og gruppe A (kontrollgruppe) – men slik at det ble omtrent like mange eksterne og interne i hver gruppe, omtrent like mange på de ulike formelle kompetansenivåene (førstestilling, mellomstilling, rekrutteringsstilling) og et nokså likt antall sensorer som tilhørte fagområdet komparativ politikk. Formålet med randomisering er som kjent å forsøke å få gruppene så like som mulig, slik at verdiene på andre relevante variable enn eksperimentvariabelen holdes konstant (jf. Morton & Williams 2008; McDermott 2002). Utfordringen er imidlertid at antallet sensorer ikke er større. Dette er noe av bakgrunnen for at vi har valgt *matchet randomisering* (Svartdal 2004:kap. 7). Ekstern–intern, stillingsnivå og subdisiplin er variable som vil kunne ha betydning ved karaktersettingen, og som det derfor er tatt hensyn til ved fordelingen på grupper.

Begge gruppene A og B fikk samme oppdrag, nemlig å lese og bedømme eksamensbesvarelsen som er beskrevet ovenfor. Sensorene i gruppe A skulle foreta bedømmelsen enkeltvis, og melde tilbake et karakterforslag (en ren karakter) uten å konferere med andre. Det var 19 av de 20 sensorene i gruppe A som rapporterte en karakter innen fristen som var satt. Én sensor rapporterte noe senere, kan i prinsippet ha fått kjennskap til eksperimentinformasjon og er følgelig ikke tatt med i materialet som analyseres.

Sensorene i gruppe B skulle også lese eksamensbesvarelsen uten å konferere med andre. Men de fikk oppgitt at de *ikke* bare skulle sende inn karakterforslaget sitt. De fikk oppgitt en *medsensor*, og de ble

bedt om å være klare til å ta sensuren på et bestemt tidspunkt. På dette tidspunktet, da de ventet en telefon, fikk de i stedet medsensors forslag til karakter oversendt på e-post.[5]

Det er altså lagt opp til at sensorene leser og gjør seg opp en mening om besvarelsen *før* de får det første karakterforslaget fra medsensor. Dersom opplegget hadde vært motsatt, med karakterinnspillet først (slik praksis ofte er ved klagesensur), kunne en «ankereffekt» ha gjort seg gjeldende i tillegg til eventuell sosial påvirkning (se for eksempel Tversky & Kahneman 1974).[6] Ankereffekten er søkt unngått, men vi har ingen garanti for at sensorene virkelig har gjort seg opp en selvstendig mening om besvarelsen før de fikk informasjon om medsensors karakterforslag.

Sensorene fikk ellers formidlet medsensors karakterforslag via den vitenskapelige assistenten på prosjektet, med følgende formulering:

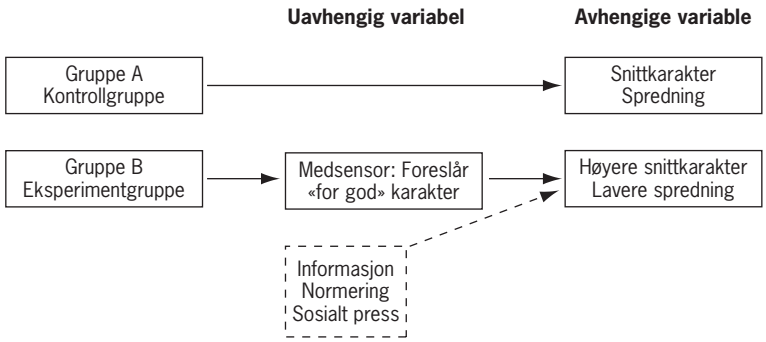
> Jeg har nå gått gjennom den ene oppgaven som ble sendt ut til deg for en tid tilbake. Jeg synes ikke at det var så lett å vurdere den, men har til slutt endt opp med å sette karakteren B.

Det ble altså tilkjennegitt noe usikkerhet. Så foreslås en ren karakter som i en viss forstand var «feil».[7] Karakteren lå to trinn unna det ordinær sensur og klagesensur hadde endt opp med. Sensorene (gruppe B) ble så bedt om å melde tilbake sitt karakterforslag, men altså *ikke* direkte til medsensor.[8] Det ble understreket at samme fremgangsmåte skulle følges også hvis en sensor i gruppe B var uenig i det første karakterinnspillet:

> Det gjelder også dersom du i utgangspunktet vil gå inn for en annen karakter enn den som er angitt ovenfor, og ønsker en nærmere diskusjon før karakteren fastsettes.

Innenfor det eksperimentelle designet er «feilkarakteren» fra medsensor den uavhengige variabelen som bare eksperimentgruppen (gruppe B) utsettes for; kontrollgruppen får ingen informasjon av denne typen. Figur 1 gir oversikt over det eksperimentelle designet. De antatt underliggende påvirkningsfaktorene vises også.

Vi vurderte å velge en B-besvarelse heller enn en D-besvarelse. Det hadde gitt sensorene litt mer å lese, men vi har ingen grunn til å tro at dette i seg selv ville ha endret resultatene vesentlig. Men det er selv-



FIGUR 1. *Oversikt over det eksperimentelle designet.*

sagt mulig at det er annerledes å forholde seg til en «feil» som går i positiv retning (bedre karakter enn besvarelsen fortjener) enn i negativ retning (dårligere karakter enn besvarelsen fortjener).

Eksperimentelle studier der deltakerne blir forledet på en eller annen måte (som i eksperimentene til Asch og Sherif), kan reise vanskelige etiske problemstillinger. I vår undersøkelse var *formålet* som sensorene fikk oppgitt i og for seg oppriktig («kvalitetssikring av sensuren»), men opplysningene var generelle og lite informative. Ingen deltakere ble fortalt om hypotesen om sosial påvirkning som lå bak eksperimentet – da ville det ikke hatt noen hensikt å gjennomføre det. Videre ble de som tilhørte eksperimentgruppen meddelt en karakter (via tredjeperson) som vi antok var et par karaktertrinn for god. Sensorer som responderte på dette innspillet ved å justere egen vurdering kan i ettertid ha følt seg lurt, og det er den eventuelle psykologiske belastning disse er påført som skaper etiske betenkeligheter. Det at sensorene i eksperimentgruppen ikke var i direkte kontakt med medsensor, men kommuniserte via en tredjeperson med taushetsplikt (som det ble opplyst om i følgebrevet), ble gjort for å redusere den mulige belastningen forsøkspersonene kunne føle. Dernest fikk alle deltakerne en tilbakemelding om eksperimentet – inkludert hypotese og hovedresultater – straks etter at dataene forelå (jf. Morton & Williams 2009:kap. 12.5 om «debriefing»). Innslag av uoppriktighet i eksperimentelle studier er langt fra uvanlig (Hertwig & Ortmann 2001), og informert samtykke fra forsøkspersonene er ikke alltid en reell mulighet. Det avgjørende blir da å avveie de faglige gevinstene

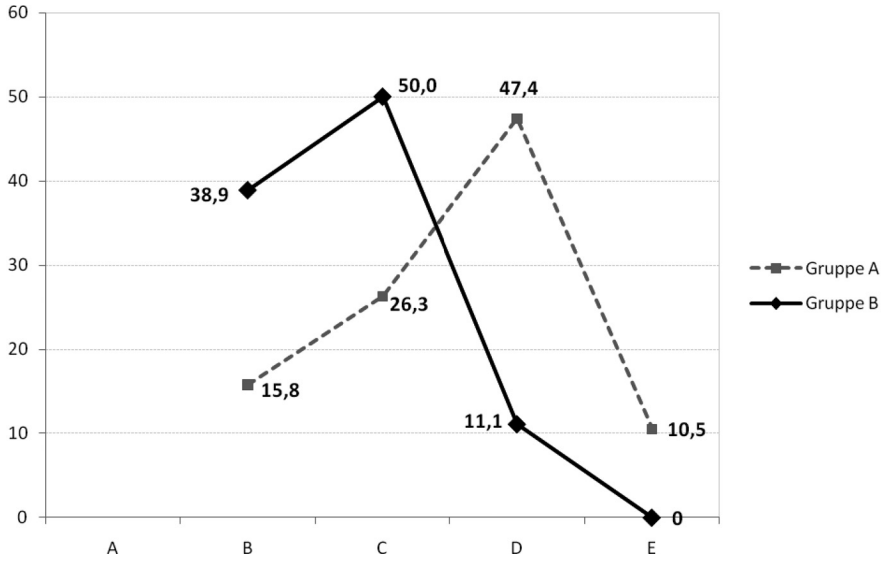
(økt kunnskap) opp mot antatte, kortsiktige belastninger av følelsesmessig art for noen av deltakerne. Noen vil si at uoppriktighet i eksperimenter alltid er uetisk (for eksempel Bok 1978:194). I eksperimentet vi har gjennomført er et begrenset element av uoppriktighet helt nødvendig for å få den kunnskapen vi er ute etter, men det er valgt prosedyrer som gjør den potensielle belastningen (enkelte) forsøkspersoner utsettes for meget begrenset. Eksperimentet antas derfor – med god margin – å være etisk forsvarlig.

Det er selvsagt flere forhold knyttet til det eksperimentelle designet som kan diskuteres. Selv om vi ønsker å si noe som den innledende interaksjonen i eksamenskommisjoner som opptrer som kollektiver, er det valgt løsninger som i noen grad skiller seg fra ordinær sensur. Det kan selvsagt ha betydning ved vurderingen av hvor generelle og relevante resultatene er for praktisk sensur.

ANALYSEMETODE OG RESULTATER

Vi har formulert en hypotese om effekten av førstesensors forslag til karakter på andresensors respons, og skissert det eksperimentelle designet som er fulgt for å teste hypotesen. Analysen gjennomføres på enklest mulig måte, ved å sammenligne sentraltendens og spredning i eksperiment- og kontrollgruppe – grupper som i utgangspunktet ble satt sammen ved matchet randomisering for blant annet å nøytralisere eventuelle effekter av bakgrunnsfaktorer. Dersom det ikke er noen påvirkning mellom sensorene, skulle en ikke vente systematiske forskjeller i karakterfordeling gruppene imellom. Dersom førstesensors forslag har en kausal effekt, enten innflytelsen er bevisst og erkjent eller ikke, vil det gi seg utslag i at gjennomsnittskarakteren i eksperimentgruppen – sammenlignet med kontrollgruppen – flyttes i retning av førstesensors forslag, og spredningen (målt ved standardavviket) reduseres. De to sensorgruppene er veldig små, noe som gir lav teststyrke. Implikasjonen av dette er at bare store effekter kan gi statistisk signifikant utfall. Det er et problem, men innebærer samtidig at det er høy risiko knyttet til eksperimentet.

Hovedresultatene fra undersøkelsen er gjengitt i figur 2.[9] Figuren viser karakterfordelingen innenfor begge sensorgruppene, med gjennomsnitt og standardavvik. Bokstavkarakterene er oversatt til en tall-



FIGUR 2. *Fordelingen av karakterer i kontrollgruppe (A) og eksperimentgruppe (B). Prosent.*

N=37 samlet.
Gruppe A (N=19): gjennomsnittet 2,47 (en god D) og standardavvik 0,90.
Gruppe B (N=18), eksperimentgruppen: gjennomsnittet 3,27 (en god C) og standardavviket 0,67.
Forskjellen mellom gjennomsnittene (0,80) er signifikant på 5 %-nivå (t-verdi=3,05).

skala med E=1, D=2, C=3, B=4 og A=5 ved beregningene. Der det er behov for det er ordinære avrundingsregler benyttet.

I gruppe A foreslår 47,4 prosent av sensorene (9 av 19) karakteren D (modus). Så vidt over halvparten av sensorene i denne gruppen foreslår en annen karakter enn den besvarelsen opprinnelig hadde fått (og som var ukjent for sensorene). Vel en fjerdedel satte C (fem sensorer). Et par sensorer ville ha svakeste ståkarakter (E), mens i overkant av 15 prosent ville gå helt opp til B (tre stykker). Tyngdepunktet i fordelingen ligger imidlertid på D, i og med et gjennomsnitt på 2,47 (en D nær grensen til C). Karakteren D er også median – et tredje mål på sentraltendens – i fordelingen.

I gruppe B responderer halvparten av sensorene med karakteren C, og nesten 40 prosent av dem går inn for en B. I denne gruppen ligger

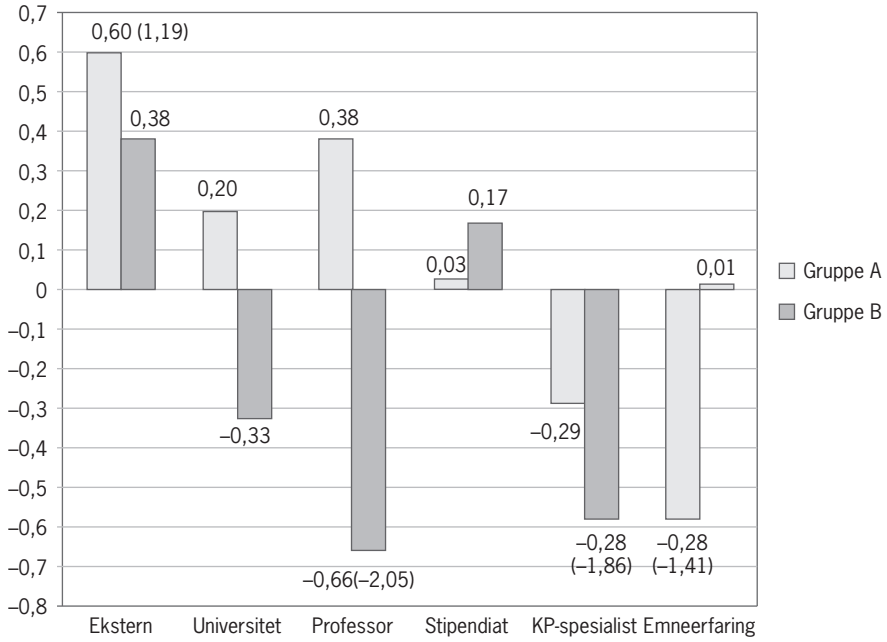
modus, median og gjennomsnitt (3,27) på C. Karakteren E foreslås ikke av noen i eksperimentgruppen.

Både når det gjelder sentraltendens og spredning er mønsteret som en kan forvente ut fra hypotesen. Ser vi på gjennomsnittkarakteren, er det en forskjell mellom gruppe A og B på 0,80 – det vil si nesten et helt karaktertrinn. Denne forskjellen er klart signifikant (t-verdi på 3,05). Likedan reduseres spredningen fra et standardavvik på 0,90 i gruppe A til 0,67 i eksperimentgruppen.[10]

Vi merker oss altså at andresensors respons farges av førstesensors forslag *før* en eventuell diskusjon av karakteren finner sted; hvilken karakter andresensor sier at besvarelsen *i utgangspunktet* fortjener, ser ut til å bli preget av det første innspillet. Vi observerer med andre ord at «feilkarakteren» som oppgis av medsensor som første innspill «driver» de øvrige sensorenes respons i forventet retning, og skaper et avvik i gjennomsnittskarakter mellom eksperimentgruppe og kontrollgruppe som er for markant til å skyldes rene tilfeldigheter.

Vi har registrert noen bakgrunnskjennetegn ved sensorene, og figur 3 viser størrelsen på avviket i gjennomsnittskarakter avhengig av disse kjennetegnene (kan betraktes som en serie med t-tester). Gruppe A er de mørke søylene. Det er kun én av forskjellene som er signifikant på konvensjonelt nivå, og det er den som er gjelder professorer versus andre i gruppe B: Professorene er i gjennomsnitt mer enn en halv karakter strengere enn de andre i gruppe B (og dette snittet ligger bare en tiendedel over snittet for professorene i gruppe A). Dette tyder på at professorer i eksperimentgruppen i noe mindre grad enn de øvrige er blitt påvirket av medsensors opprinnelige forslag til karakter.

Sensorene i gruppe B fikk oppgitt medsensor, og har dermed forventet at de opererte som del av en kommisjon med to medlemmer. Tar vi dette et skritt videre, kan vi beregne karaktergjennomsnittet mellom førstesensor og andresensor for hver besvarelse. Dette er ikke nødvendigvis dekkende for hva som vil skje under en virkelig bedømmelse, fordi sensorer ofte vil diskutere nærmere hvis det er stort avvik mellom vurderingene. På den annen side er det jevnt over ikke noe stort avvik mellom første innspill og respons fra sensorene i gruppe B – samme vurdering i 7 av 18 tilfeller, ett karaktertrinns forskjell i ni tilfeller og to karaktertrinns avvik kun i to tilfeller. Ved små avvik (for eksempel ett karaktertrinn) er det ikke uvanlig at sensorer tar et gjen-



FIGUR 3. *Forskjeller i gjennomsnittskarakter innenfor hver eksperimentgruppe, avhengig av ulike kjennetegn ved sensorene.*

Tallene ved hver søyle viser avvik fra gjennomsnitet for hver gruppe sensorer. Negative tall betyr «strengere» bedømmelse i gjennomsnitt. T-verdier står i parentes på de største avvikene.

nomsnitt. Foretas slik beregning, ender vi opp med 15 B-er og 3 C-er. Det gir et karaktersnitt på 3,83 (svært nær en ren B) og et lite standardavvik på 0,38. Modus og median i fordeling er selvsagt også karakteren B. Betraktet på denne måten har feilkarakteren til førstesensor produsert svært forskjellige resultater fra de som observeres i kontrollgruppen.

Det er altså klare forskjeller på karakterforslagene i gruppe A og gruppe B, helt i tråd med forventningen bak hypotesen som er formulert. Ut fra eksperimentet kan vi ikke si entydig hvilke underliggende faktorer som gjør at eksperimentvariabelen slår ut. Trolig er det flere forhold som ligger til grunn, det vil si både en informasjons- eller normeringseffekt (jf. Sherif) og en ren påvirkningseffekt eller sosialt press (jf. Asch) som følge av første karakterinnspill fra medsensor. Eksperi-

mentet er imidlertid designet slik at en ankereffekt ikke skal gjøre seg gjeldende i nevneverdig grad.

Selv om vi har utformet eksperimentet på en slik måte at vi skal kunne ha tillit til resultatene, kan det selvsagt reises innvendinger. Vi vil nevne noen utfordringer i forhold til indre validitet og representativitet.

Kan det være andre variable enn eksperimentvariabelen som slår ut og skaper forskjellen i karaktersetting mellom gruppe A og B? I prinsippet kan en tenke seg at forskjeller i for eksempel alderssammensetning, utdanningsbakgrunn, forskningsorientering eller lignende ligger bak og produserer effekten som observeres. Siden vi har med så små grupper å gjøre, kan det ikke utelukkes at gruppene (uten vår vitende og vilje) er blitt forskjellige i viktige henseender selv om randomisering er benyttet.[11] Vi har imidlertid ingen teoretiske antakelser om slike bakenforliggende forhold som skulle kunne ha effekt. Den mest realistiske kandidaten i så måte er at noen sensorer i gruppe B kan ha lest besvarelsen *etter* at de fikk karakterforslaget fra medsensor, og at forskjellene i karakterer mellom gruppe A og B blir skapt – eller forsterket – som følge av dette. Ankereffekten er som nevnt interessant i seg selv, men det er ikke den vi har intendert å måle. Vi tror imidlertid ikke dette er noe stort problem, fordi de fleste sensorene i gruppe B ga beskjed om at de var klare til å ta sensuren *før* medsensors karakterforslag ble formidlet.

Har vi lokket frem en bestemt respons hos sensorene i gruppe B ved å utforme eksperimentsituasjonen slik vi har gjort? Er det eksperimentsituasjonen som skaper effekten, heller enn de kausale samhandlingsmekanismene vi er opptatt av? I så fall ville dette også være et problem i forhold til eksperimentets *indre validitet*. Vi tror imidlertid ikke at dette kan forklare den observerte forskjellen i karaktersetting. Det ble lagt vekt på å utforme korte, nøytrale tekster i henvendelser til sensorer og i følgebrevet som gikk ut sammen med besvarelse og sensorveiledning. Det er kun et lite avsnitt om oppnevnt medsensor som er forskjellig, og det er valgt formuleringer som ikke på noen måte er ledende; gitt at bruk av to sensorer er det vanlige ved institutt-tet, fremstår opplegget som helt ordinært.

Selvseleksjon er et problem i mange empiriske analyser. Også her er det et innslag av selvseleksjon i den forstand at det var en større gruppe sensorer som fikk henvendelse om å delta i bedømmelsen

enn det antallet som faktisk deltok. Deltakelse krevde at den enkelte sa seg villig til å delta. Det betyr at medlemmene av gruppe A og B samlet sett ikke er helt representativt for sensorkorpset i faget, men de er trolig rimelig representative for den gruppen vitenskapelig ansatte som fra tid til annen påtar seg litt ekstrasensur. Uansett føres elementet av selvseleksjon ikke over i en forskjell i sammensetning av gruppene, fordi randomiseringen – den tilfeldige inndelingen i gruppe A og gruppe B – fant sted *etter* at det var klart hvem som ville delta.

BETYDNING FOR VURDERING AV SENSORORDNINGER

Det å vurdere om bruk av en eller to sensorer er *best*, er en normativ problemstilling som vi ikke tar stilling til. Det ville kreve en systematisk avklaring av hva som er «god» sensur, og en mye bredere diskusjon enn det vi har lagt opp til her. For eksempel måtte ressursspørsmål kommet inn, og slike ting som antall eksamener som inngår i en grad. Vårt formål har vært å belyse hvordan den innledende del av samhandlingen i eksamenskommisjoner faktisk fungerer. Funnene våre kan forbedre grunnlaget for å vurdere ulike sensorordninger opp mot hverandre.

Liten spredning er en påstand som har vært trukket inn som støtte for bruk av bare *én sensor*; hvis sensorer likevel tenker likt, er det lite å oppnå ved å bringe inn flere sensorer. Sensorer har hevdet at de opplever at det ikke er særlig store forskjeller i vurderingen av eksamensbesvarelser i kommisjoner der de deltar sammen med andre. Selv om dette aspektet ikke har vært sentralt i vår analyse, bærer tallene i figur 1 bud om at det finnes forskjeller i sensorenes vurderinger som ikke er ubetydelige (jf. at hele fire karaktertrinn ble foreslått blant sensorene i gruppe A).

Et argument som gjerne trekkes inn til støtte for å benytte *to sensorer* er dette: Dersom den ene sensoren vurderer en besvarelse feil – og gir en karakter som er for dårlig eller for god – kan feilen bli rettet opp før karakteren protokollføres. Slik sett representerer ordninger med to sensorer på hver besvarelse en «rettssikkerhetsgaranti». Vi ser ingen grunn til å avvise dette argumentet på grunnlag av vår undersøkelse, og feil blir sikkert i mange tilfeller korrigert eller dempet ved at to sensorer deltar. Noen garanti for at så skjer finnes imidlertid

ikke. I vårt eksperiment påvirker feilen til førstesensor hva mange av andresensorene mener om besvarelsen. Slik påvirkning gjør det i utgangspunktet vanskelig å få korrigert feil førstesensor måtte gjøre seg skyldig i. Vi har vist eksperimentelt at førstesensors feil ikke nødvendigvis vil bli korrigert, og at slike feil i større eller mindre grad vil bli bevart.

I eksamenskommisjoner med to personer er det derfor feil som gjøres av den sensoren som kommer med andre innspill (respons) som vil ha størst sannsynlighet for å kunne bli korrigert, men da kanskje først og fremst i form av at andresensor tilpasser sin respons til det forslaget førstesensor har kommet med. Rettssikkerhetsargumentet synes med andre ord å være relativt svakt.

AVSLUTNING

Det finnes få studier av arbeidet til eksamenskommisjoner, og i det hele tatt lite forskningsbasert kunnskap om karaktersetting. Vi har konsentrert oss om et utvalgt aspekt ved samhandlingen mellom sensorer som setter karakter på eksamensarbeider i fellesskap, og søkt å få en bedre forståelse av i hvilken grad sensorer påvirker hverandre. Nærmere bestemt har vi konsentrert oss om det forhold at én av sensorene i en kommisjon med to medlemmer nødvendigvis må uttale seg først. Er det slik at dette første innspillet i seg selv påvirker hva den andre sensoren mener om den aktuelle besvarelsen? En rekke sosialpsykologiske eksperimenter gir grunnlag for å tro at svaret er ja, for eksempel på grunn av sosialt press eller andre former for sosial tilpasning.

For å undersøke dette nærmere har vi gjennomført et eksperiment med to grupper på i alt nærmere førti sensorer. Den ene gruppen vurderte én eksamensbesvarelse og sendte inn forslag til karakter individuelt. Den andre gruppen fikk oppgitt en medsensor, og medsensoren foreslo en karakter som i en viss forstand var for god (altså en «feil»). Gruppen av sensorer som ble utsatt for manipulasjon gjennom å få oppgitt en svært god karakter, endte opp med en klart bedre gjennomsnittskarakter enn det kontrollgruppen gjorde. Dette tyder på en påvirkning i tråd med hva en skulle forvente på teoretisk grunnlag.

Et argument som ofte trekkes frem til forsvar for bruk av to sensorer, er at det gir gode muligheter for å rette opp feilvurderinger som den ene av sensorene måtte gjøre seg skyldig i. Vi antar i likhet med

mange andre at dette argumentet har noe for seg. Vi finner imidlertid liten støtte for det i eksperimentet som er gjennomført. Her kan feilen til sensoren som foreslår karakter først ikke sies å bli rettet opp, men den «smitter» i stedet over på mange av de andre sensorene og blir i stor grad bevart.

Referanser

Asch, Solomon E. (1952), *Social Psychology*. Englewood Cliffs, NJ: Prentice-Hall.

Asch, Solmon E. (1955), «Opinions and Social Pressure». *Scientific American*, 193 (5):31–35.

Baird, Jo-Anne, Jackie Greatorex & John F. Bell (2004), «What makes marking reliable? Experiments with UK examinations». *Assessment in Education*, 11 (3):331–348.

Bok, Derek (1978), *Lying. Moral Choices in Public and Private Life*. New York: Pantheon Books.

Brandt, Ellen & Bjørn Stensaker (2005), Internasjonale sensorordninger i høyere utdanning. Er erfaringer fra Sverige, Danmark og England relevante for Norge? Arbeidsnotat 39/2005. Oslo: NIFU STEP.

Brown, Sally & Angela Glasner, red. (1999), *Assessment Matters in Higher Education. Choosing and Using Diverse Approaches*. Buckingham: Open University Press.

Dysthe, Olga (2007), «Pedagogiske endringar etter Kvalitetsreforma og konsekvensar for læring. Utfordringar og strategiar vidare». *Uniped*, 30 (3):29–44.

Dysthe, Olga, Arild Raaheim, Ivar Lima & Arne Bygstad (2006), Undervisnings- og vurderingsformer. Pedagogiske konsekvenser av Kvalitetsreformen. Evaluering av Kvalitetsreformen, delrapport 7. Bergen: Rokkansenteret.

Gjølberg, Ole & Kolbjørn Christoffersen (2008), Råvarekvalitet i utdanningen av økonomer ved norske læresteder: En empirisk analyse av studentopptaket 2007. Notat. Institutt for økonomi og ressursforvaltning, UMB.

Hertwig, Ralph & Andreas Ortmann (2001), «Experimental practices in economics: A methodological challenge for psychologists?». *Behavioral and Brain Sciences*, 24:383–451.

Insko, Chester A., Richard H. Smith, Mark D. Alicke, Joel Wade & Sylvester Taylor (1985), «Conformity and Group Size: The Concern of Being Right and the Concern with Being Liked». *Personality and Social Psychology Bulletin*, 11:41–50.

Koehler, Derek J. & Nigel Harvey (1997), «Confidence judgements by actors and observers». *Journal of Behavioral Decision Making*, 10:221–242.

Leyens, Jacques-Philippe & Olivier Corneille (1999), «Asch's Social Psychology: Not as Social as You May Think». *Personality and Social Psychology Review*, 3 (4):345–357.

McDermott, Rose (2002), «Experimental Methods in Political Science». *Annual Review of Political Science*, 5:31–61.

Mohr, Lawrence B. (1990), *Understanding Significance Testing*. Quantitative Applications in the Social Sciences 73. London: SAGE.

Morton, Rebecca B. & Kenneth C. Williams (2008), «Experimentation in Political Science». I: Janet M. Box-Steffensmeier, Henry F. Brady & David Collier red., *The Oxford Handbook of Political Methodology*. Oxford: Oxford University Press.

Morton, Rebecca B. & Kenneth C. Williams (2009), From Nature to the Lab: Experimental Political Science and the Study of Causality. Manuskript. New York University.
Møen, Jarle & Martin Tjelta (2005), «Bruker ulike høgskoler karakterskalaen ulikt? En analyse av sammenhengen mellom skolebakgrunn og faglig suksess ved NHH». *Økonomisk Forum*, nr. 6:1–13.
Partington, John (1994), «Double-Marking Students' Work». *Assessment & Evaluation in Higher Education*, 19:57–60.
Raaheim, Arild (2000), «En studie av inter-bedømmer reliabilitet ved eksamen på psykologi grunnfag». *Tidsskrift for Norsk Psykologforening*, 37:203–213.
Rozin, Paul (2001), «Social Psychology and Science: Some Lessons from Solomon Asch». *Personality and Social Psychology Review*, 5 (1):2–14.
Rowntree, Derek (1987), *Assessing students: How shall we know them?*. London: Kogan Page.
Sherif, Muzafer (1936), *The Psychology of Social Norms*. New York: Harper and Brothers.
Sherif, Muzafer (1937), «An Experimental Approach to the Study of Attitudes». *Sociometry*, 1:90–98.
Shiller, Robert J. (1995), «Conversation, Information, and Herd Behavior». *American Economic Review*, 85 (2):181–185.
Svartdal, Frode (2004), *Psykologiens forskningsmetoder – en introduksjon*. 2. utg. Bergen: Fagbokforlaget.
Solum, Nils Henrik (2005), Sensorordninger i høyere utdanning – kartlegging av status og utviklingstrekk ved 10 institusjoner. Arbeidsnotat 17/2005. Oslo: NIFU STEP.
Turner, John C. (1991), *Social Influence*. Milton Keynes: Open University Press.
Taylor, Shelley E., Letitia Anne Peplau & David O. Sears (2000), *Social Psychology*. 10th ed. Upper Saddle River: Prentice-Hall.
Teigen, Karl Halvor (1986), Karaktersetting ved grunnfagseksamen i psykologi. Upublisert. Det psykologiske fakultet, Universitetet i Bergen.
Tversky, Amos & Daniel Kahneman (1974), «Judgment under uncertainty: Heuristics and biases». *Science*, 185:1124–1131.

Noter

1. Lov om universiteter og høyskoler (2005-04-01-15, spesielt § 3-9) har ikke noe generelt krav om to sensorer eller bruk av ekstern sensor på annet enn ved bedømmelse av større, selvstendige arbeider innenfor høyere grad (masteroppgaver). I § 3-9 pkt. 1 sies det: «Universiteter og høyskoler skal sørge for at studentenes kunnskaper og ferdigheter blir prøvet og vurdert på en upartisk og faglig betryggende måte. Vurderingen skal også sikre det faglige nivå ved vedkommende studium. Det skal være ekstern evaluering av vurderingen eller vurderingsordningene.»
2. En undersøkelse i regi av Nasjonalt fagråd i statsvitenskap høsten 2007 kan også være relevant i denne sammenheng. Her ble ti masteroppgaver fra hele landet gjennomgått i detalj av en gruppe på syv sensorer. Alle statsvitenskapelige miljøer som har masterutdanning i faget var representert både med oppgaver og sensorer. Oppgavene var trukket ut tilfeldig blant alle som våren 2007 hadde oppnådd å få karakteren B. Sensorgruppen konkluderte med at bare to av de ti B-oppgavene etter deres mening var sikre B-er. En av oppgavene satte et flertall av sensorene til D, men D-en ble trukket frem av enkelte sensorer (ikke alltid de samme) i fire av

de ti tilfellene. Det var en viss spredning i sensorenes vurderinger på den måten at det ikke ble foreslått samme karakter av alle sensorene på noen av de ti masteroppgavene til vurdering. Likevel var det ikke i noe tilfelle mer enn to karaktertrinn som ble foreslått på en og samme oppgave, men det er godt mulig at spredningen ville blitt noe større dersom sensorene hadde jobbet helt uavhengig av hverandre. Hovedkonklusjonen fra arbeidet var at det foreløpig ikke kan sies å være utviklet en felles nasjonal standard for vurdering av masteroppgaver. Se http://folk.uio.no/berasch/NFRSt-KarakterRapport-2007.pdf. Et par studier som indikerer andre typer av ulikheter i bruken av karakterskalaen er Møen & Tjelta (2005) og Gjølberg & Christoffersen (2008).

3. Hvis det var flere besvarelser, kunne også rekkefølgen oppgavene ble lest i tenkes å påvirke karaktersettingen (jf. Raaheim 2000:205–206). Selv om vi ikke helt kan utelukke at idiosynkratiske rettestrategier hos sensorene har betydning også i vårt tilfelle, er det vanskelig å se at slike effekter skulle gjøre seg gjeldende.
4. Henvendelsen kom fra den studiekonsulenten som de siste semestrene har hatt med den praktiske gjennomføringen av eksamener på bachelornivå å gjøre. Sensorer som ønsket utfyllende informasjon om hvorfor kvalitetssikringen ble gjort på denne måten, fikk ikke annet enn de samme generelle opplysningene med litt andre ord. Ingen fikk vite om inndelingen i eksperimentgrupper. Etiske sider ved eksperimentet ble vurdert nøye i forkant, som vi kommer tilbake til. Etiske betenkeligheter gjorde at vi for eksempel ikke valgte å si (feilaktig) at det dreide seg om reell klagesensur.
5. Medsensoren som ble oppgitt var den ene av oss (Rasch). Han er professor og undervisningsleder ved Institutt for statsvitenskap, men tilhører ikke formelt fagområdet komparativ politikk. Han har blant annet deltatt i komiteer i regi av Nasjonalt fagråd i statsvitenskap som har tatt opp spørsmål knyttet til karaktersetting. Mange av de som har deltatt i eksperimentet vil være kjent med dette. Vi vurderte andre muligheter, som for eksempel å oppgi en helt uerfaren medsensor eller å la medsensor være anonym. Det kan gis gode argumenter for at vi burde ha satset på en annen medsensor, men vi valgte heller å opplyse sensorene om at medsensor ikke ville få kjennskap til hvilken karakter de foreslo (se nedenfor). Etiske overveielser spilte også inn, i og med at medsensor med vitende og vilje oppga en karakter som klart avviker fra ordinær sensur.
6. Vi antar at det er forskjell på å lese en besvarelse før karakter fra medsensor er oppgitt, og å lese den etter at medsensor har oppgitt sitt karakterforslag. I det sistnevnte tilfellet starter sensor lesningen med et anker kastet ut, det vil si med kjennskap til en foreløpig konklusjon. En rekke eksperimenter underbygger at ankereffekter lett gjør seg gjeldende.
7. Dette er altså et bevisst valg av en karakter vi antok var for god. Samtidig forekommer denne karakteren (B) blant sensorene i kontrollgruppen, som vi skal se senere, noe som kan ha relevans i forbindelse med den etiske vurderingen av eksperimentet.
8. Bakgrunnen for dette var et ønske om å redusere eventuelle innslag av sosialt press eller lignende. Sensorene i gruppe B skulle ha liten grunn til å føle slikt press fra medsensor all den stund det ble opplyst at denne ikke skulle få kjennskap til karakteren som ble spilt inn (responsen); alt skulle gå via den vitenskapelige assistenten, og det ble understreket i følgebrevet til sensorene at den vitenskapelige assistenten hadde undertegnet en taushetserklæring, og at alle data ville bli anonymisert før de ble analysert nærmere.

9. Hvilken gruppe sensorene tilhører har signifikant effekt på karakterfordelingen også i enveis variansanalyse ($F [1,35] = 9,36$; $p = 0,004$). Effekten opprettholdes ved introduksjon av bl.a. de dikotome variablene professorat og KP-fagspesialisering. Ingen av variablene som vises i figur 3 har selvstendig effekt.
10. Ved tolkning av standardavviket er det viktig å huske tre ting. For det første er alle datapunkter hele tall (hele karakterer uten plusser, minuser eller desimaler). For det andre har vi med små sensorgrupper å gjøre, og standardavviket påvirkes klart av endringer i enkeltsensorers karaktersetting. For det tredje er standardavviket uttrykk for en form for gjennomsnittsbetraktning, dvs. en gjennomsnittlig spredning rundt gjennomsnittet i fordelingen (definert som kvadratroten av variansen – som er snittet av kvadrerte avvik fra gjennomsnittet). Typisk er standardavviket litt høyere enn det (absolutte) gjennomsnittlige avviket fra gjennomsnittet i fordelingen (Mohr 1990:11). Gir alle sensorer samme karakter, er standardavviket null. Hva som normativt sett er en akseptabel spredning i vår kontekst er vanskelig å si, blant annet fordi det er mange fordelinger som kan gi det samme standardavviket. La oss ta et eksempel med en gruppe på 20 sensorer. Vi får det samme standardavviket på en halv karakter (0,51) i en fordeling der den ene halvparten gir D og den andre halvparten C, som i en fordeling der 18 stykker gir D, én gir B og én gir E. Dette er svært ulike fordelinger.
11. Det er følgelig heller ingen signifikante korrelasjoner mellom gruppevariabelen og de bakgrunnsvariablene som fremgår av figur 3.