Varierer eksamensresultater med vanskelighetsgrad/arbeidsmengde?

På grunnlag av nasjonal eksamen i årsregnskap på bachelorstudiet i regnskap og revisjon fra 2015–19 ble det undersøkt om det foreligger en sammenheng mellom eksamensresultat og vanskelighetsgrad/arbeidsmengde. Resultatene av undersøkelsen indikerer at det er en slik sammenheng, men det er vanskelig å underbygge konklusjoner statistisk.

Statsautorisert revisor/
førstelektor
Eddie Flatmo Rekdal
Høgskolen i Molde

Statsautorisert revisor/dosent
Kjell Magne Baksaas
Universitetet i Sørøst-Norge

Oppsummering av undersøkelsen

- å vurdere vanskelighetsgrad og arbeidsmengde for eksamensoppgaver er vanskelig
- det finnes klare indikasjoner på sammenheng mellom eksamensresultater og vanskelighetsgrad/arbeidsmengde
- selv med en nasjonal eksamen arrangert av NOKUT, kan den enkelte student være heldig eller uheldig med oppgavesettets vanskelighet/arbeidsmengde

Undersøkelsen viser også at det å vurdere vanskelighetsgrad og arbeidsomfang i seg selv er en krevende øvelse. Artikkelen belyser og reflekterer rundt spørsmål om hvorvidt studenter kan være «heldige» eller «uheldige» med eksamensoppgaver. Utgangspunktet er en forventning om at på samme måte som at en karakter skal reflektere kunnskapsnivå uavhengig av institusjon, bør den også være uavhengig av eksamensår og variasjon i vanskelighetsgrad og arbeidsmengde.

Bakgrunn

NOKUT gjennomførte på oppdrag fra Kunnskapsdepartementet nasjonale deleksamener på tre bachelorutdanninger. Innen økonomisk-administrative fag valgte NOKUT og NRØA[2] eksamen i faget «Årsregnskap og god regnskapsskikk» på bachelorstudiet i regnskap og revisjon (BRR). Det ble avholdt nasjonal eksamen og felles sensur i regi av NOKUT for årene 2015–2019.

Et av formålene var å gi de forskjellige fagmiljøene mulighet til å sammenligne seg med tilsvarende fagmiljøer ved andre institusjoner (Kunnskapsdepartementet 2014).[3] Oppdragsbrevet påpeker at dette kunne gi miljøene grunnlag for utvikling og bidra til økt tillit til utdanningene i samfunnet.

Det var også et mål at en gitt karakter skal reflektere samme kunnskapsnivå, uavhengig av hvilken institusjon eksamenen er avlagt ved.

Et annet interessant forhold for ovennevnte sensurresultater er å undersøke om det er en sammenheng mellom eksamensresultater og vanskelighetsgrad/omfang ved eksamensoppgavene. Eller med andre ord om en gitt karakter reflekterer samme kunnskapsnivå uavhengig av eksamensår eller vanskelighetsgrad/omfang til eksamen.

At karakter er uavhengig av institusjon og eksamensår, er grunnleggende for studentenes rettssikkerhet. Effekten av ulike resultater fra år til år er til en viss grad kompensert gjennom at årskullets

1 Artikkelforfatterne var involvert i sensureringen av eksamenene som denne artikkelen berører.

2 Universitets- og Høgskolerådet Nasjonalt Råd For Økonomisk-Administrativ Utdanning, nå endret navn til UHR-økonomi og administrasjon

3 Oppdragsbrev fra KD til NOKUT, datert 08.09.14

eksamensresultater i samme emne på samme institusjon fremkommer av studentenes vitnemål.

De svake eksamensresultatene i 2015 førte til oppslag i Dagens Næringsliv. Året etter var det betydelig bedre resultater. Dette vekket nysgjerrigheten for å få svar på spørsmålet i tittelen. Lett tilgjengelige eksamensresultater for årene 2015–19 for årlig rundt 500 studenter gjorde det mulig å foreta en enkel analyse med begrensede ressurser.

Om eksamen og sensurering

Eksamen i «Årsregnskap og god regnskapsskikk» er avsluttende eksamen i Finansregnskap på BRR. Eksamen er på seks timer bestående av fire–fem oppgaver, der det er oppgitt anslått tid for hver av oppgavene. Et eksamenssett består i hovedsak av oppgaver basert på regnskapstall og -opplysninger, men også av en del rene teorioppgaver. Store deler av eksamen er kvantitativ, og sensorer baserer seg ofte på en matematisk tilnærming ved at eksamenssettet tildeles 100 %. Karakterbeskrivelsene fra UHR angir en helhetlig tilnærming og vår erfaring er at sensorene i tillegg til eller som en oppsummering foretar en helhetsbedømmelse ved karakterfastsettelsen. I praksis vil vi som erfarne undervisere i høyere utdanning tro at den enkelt sensor bevisst eller ubevisst foretar en nivåjustering av karakternivå, ut fra en vurdering av samlet prestasjon på alle besvarelsene.

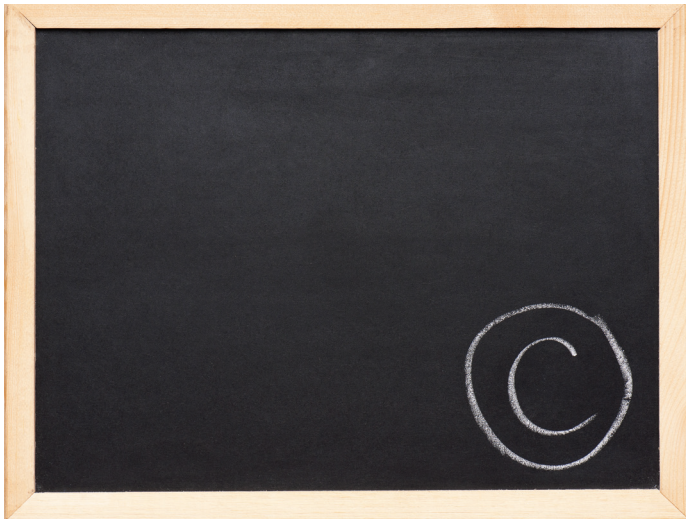
Eksamensoppgavene ble utarbeidet av en komité på tre faglærere fra de universitetene eller høyskolene som tilbyr kurset. Til forskjell fra nasjonale prøver i grunnskolen ble oppgavesettet ikke pilotert eller på annen måte prøvet ut mot studenter eller vurdert av andre enn de tre i oppgavekomiteen. De årlig rundt 500 besvarelsene ble sendt inn til NOKUT, som videresendte dem til seks–sju sensorer. Alle besvarelser ble sensurert av to sensorer slik at hver sensor sensurerte sammen med alle de andre i sensorpar.

Selv med et så stort antall som 500 årlige besvarelser var meldingen fra NOKUT klar på at det ikke skulle skje en nivåvurdering av karakterfordeling. Sensorene ble delt i sensorpar, og hvert par skulle rapportere fastsatt karakter til NOKUT uten samordning med andre par (Hamberg og Tokstad 2015 pkt.1.3.3). I NOKUTs oppsummering fra de nasjonale deleksamenene (Hamberg og Tokstad 2017 pkt.2.1 og 2.3)[4] er problemstillingen om nivåvurdering løftet frem, men i liten grad berørt.

Det er et særtrekk ved denne eksamenen at karakteren C eller bedre må oppnås som et av kravene til tittelen registrert revisor. Det samme gjelder tre andre fag som inngår i BRR (revisjon, rettslære og skatterett).[5] Karakterkravet medfører at flere studenter enn normalt kontinuerer i disse

4 NOKUTs oppsummeringer: Nasjonal deleksamen – et pilotprosjekt og en mulighetsstudie. Oppsummering av erfaringene med å gjennomføre nasjonal deleksamen i tre profesjonsutdanninger. Mars 2017.

5 Karakterkravet må være oppfylt, sammen med blant annet praksiskrav når kandidaten søker Finanstilsynet.



Karakteren C eller bedre må oppnås som et av kravene til tittelen registrert revisor. Det samme gjelder tre andre fag som inngår i BRR (revisjon, rettslære og skatterett).

emnene innen høyere utdanning. Kontinuasjon og ordinær eksamen gjennomføres samtidig, slik at det ikke er mulig å skille den ene kategorien studenter fra den andre.

Fra litteraturen er det kjent at det er et spenningsfelt mellom læringsutbyttebeskrivelsene og eksamensvurdering. Vi har i denne undersøkelsen valgt å ikke fokusere på det spørsmålet.

Eksamensresultatene

Det foreligger omfattende rapporter utgitt av NOKUT for de fem eksamensårene.[6] Vi begrenser oss her til følgende tall, der gjennomsnittskarakteren er beregnet på en skala fra 1 til 5, der A=5, B=4, C=3, D=2, E=1 og F=0:

	2015	2016	2017	2018	2019
Gjennomsnittskarakter	1,9	2,5	2,1	2,2	2,1
Andel bestått	0,76	0,85	0,83	0,85	0,85
Andel C eller bedre	0,37	0,52	0,39	0,46	0,39
Antall studenter	535	545	532	544	409

Vi ser at gjennomsnittskarakteren varierte mye fra 2015–17, men var stabil de tre siste årene. Andel bestått var lav det første året, men holdt seg meget stabil siden. Andelen som fikk C eller bedre viser samme utvikling. Som vi skal komme tilbake til, er fem år kort tid som grunnlag for statiske beregninger, men korrelasjonskoeffisienten mellom gjennomsnittskarakteren og andelen som fikk C eller bedre var på hele 0,95.

Vurdering av vanskelighetsgrad/arbeidsmengde

De årlige eksamenssettene er i overveiende grad basert på caseoppgaver og i mindre utstrekning åpne teorioppgaver.

6 NOKUTs rapporter for det enkelte år er tilgjengelig på NOKUT.no

Caseoppgavene gjør det mer krevende å vurdere studentenes opplevelse av vanskelighetsgrad og arbeidsmengde.

Vurderingene ble foretatt av fem fagpersoner inkludert forfatterne som alle har undervist i kurset i en årrekke.[7] Fire var med i sensorkorpset, varierende fra ett til alle årene. En var med i eksamenskomiteen alle årene. Det kan selvsagt stilles spørsmål ved habiliteten til de som vurderte oppgavene, men fagmiljøet er lite, så det var ikke så mange å velge blant. Fordelen med dem som ble valgt ut, er at de kjenner kurset og dets innhold og nivå godt. Det er også vanskelig å finne grunner til at de ikke skulle kunne foreta en uavhengig og faglig basert vurdering.

Undersøkelsen ble gjort i form av et spørreskjema der de fem ble bedt om å vurdere vanskelighetsgrad på hver oppgave og for eksamenssettet samlet sett og oppgavesettets arbeidsmengde.

Eksamensresultatene forventes å bli dårligere ved økt vanskelighetsgrad og arbeidsmengde. For at vurderingene skal ha samme retning som eksamensresultatene, presenteres resultatet av vurderingene slik:

For vanskelighetsgrad:
5=Enkel
4=Under middels vanskelig
3=Middels
2=Over middels vanskelig
1=Vanskelig

For arbeidsmengde:
5=For lite
4=Under middels
3=Middels (passe)
2=Over middels
1=For mye

Veid vanskelighetsgrad

Det ble her satt poeng som beskrevet ovenfor på hver oppgave. Deretter ble denne veid med anslått tid pr. oppgave for å komme frem til vanskelighetsgrad på eksamenssettet. Det var overraskende stor spredning i vurderingene her:

Vanskelighetsgrad – OPPGAVE FOR OPPGAVE	Fordeling	Fordeling i %
Vurdering fordelt på ett poeng	2	9 %
Vurdering fordelt på to poeng	10	43 %
Vurdering fordelt på tre poeng	9	39 %
Vurdering fordelt på fire poeng	2	9 %
Antall oppgaver 2015-2019	23	100 %

Fordeling på to poeng er akseptabelt. Fordeling på tre og fire poeng vurderer vi som noe høyt.

Samlet vurdert vanskelighetsgrad

Når det gjaldt vurdering av vanskelighetsgrad for hele eksamenssettet samlet under ett, var det liten spredning:

Vanskelighetsgrad – SAMLET					
	5-4-3-2-1 osv	5= enkel, 4= under middels, 3= middels osv			
Eksamensår	**2015**	**2016**	**2017**	**2018**	**2019**
	0-0-4-1-0	0-1-4-0-0	0-0-2-3-0	0-1-4-0-0	0-0-3-2-0
Fordelt på antall poeng	2	2	2	2	2
					C/D er plassert i C

En vurderte vanskelighetsgraden for 2019 til 2 til 3. Denne er satt til 3 her, men til 2,5 i gjennomsnittsberegningen nedenfor.

Arbeidsmengde (omfang)

Det var her en «slenger» for 2017, men ellers var resultatene samlet:

Omfang					
	5-4-3-2-1 osv	5= for lite, 4= under middels, 3= middels (passe) osv			
Eksamensår	**2015**	**2016**	**2017**	**2018**	**2019**
	0-0-1-4-0	0-0-3-2-0	0-1-0-4-0	0-1-4-0-0	0-0-4-1-0
Fordelt på antall poeng	2	2	3	2	2
				C/B er plassert i C	C/B er plassert i C

En vurderte arbeidsmengden for 2018 og 2019 til 2 til 3. Disse er satt til 3 her, men til 2,5 i gjennomsnittsberegningen nedenfor.

Oppsummering

Resultatet av vurderingene kan oppsummeres slik (jo lavere tall, jo vanskeligere og arbeidskrevende eksamenssett):

Gjennomsnitt av 5 vurderinger	2015	2016	2017	2018	2019
Vanskelighetsgrad veid	2,55	2,97	2,40	2,99	2,48
Vanskelighetsgrad samlet	2,80	3,20	2,40	3,20	2,50
Arbeidsmengde	2,20	2,60	2,40	3,10	2,70

Vi ser at det er store variasjoner fra år til år. Resultatet av de to måtene å vurdere vanskelighetsgrad på er relativt sammenfallende, men oppgavene er vurdert litt vanskeligere ved vurdering oppgave for oppgave i forhold til samlet.

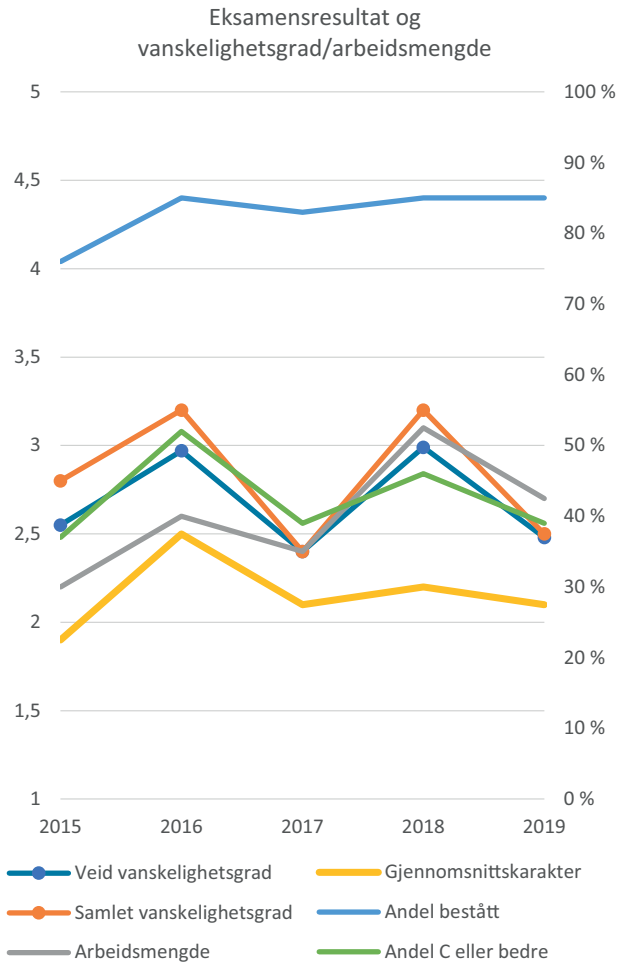
7 En stor takk til våre tre kolleger fra andre institusjoner for at de bidro med sine vurderinger.

Ovennevnte viser hvor krevende det er å vurdere vanskelighetsgrad. Her ble det valgt en enkel og lite ressurskrevende måte å foreta vurderingen på. Vurderingen kunne nok vært utvidet til å omfatte flere som vurderte oppgavene, eller en annen form der for eksempel en gruppe fagpersoner diskuterte seg frem til vurderinger. Det kreves nok en grundigere vurdering fra flere involverte, også studentrepresentanter, for å foreta en bedre vurdering.

Sammenheng mellom eksamensresultater og vanskelighetsgrad/arbeidsmengde

Resultatene foran kan fremstilles slik når andel bestått og andel karakteren C eller bedre er skilt ut fra de øvrige grunnet ulike skalaer:

NB! For at vurderingene skal ha samme retning som eksamensresultatene, viser høye tall lav vanskelighetsgrad/liten arbeidsmengde og lave tall høy vanskelighetsgrad/stor arbeidsmengde i fremstillingen nedenfor.



Vanskelighetsgrad, arbeidsmengde og gjennomsnittskarakterer er fremstilt med hel linje på en skala fra 5 (enkel/for lite) til 1 (vanskelig/for mye). Andel C eller bedre og andel bestått er fremstilt med stiplet linje på en skala der 1=100 %.

Grafene indikerer sammenhenger mellom eksamensresultat og vanskelighetsgrad/ arbeidsmengde. Figuren viser at oppgavesettet for 2015 ble vurdert å være både vanskeligere og ha høyere arbeidsmengde enn oppgavesettet for 2016. Samtidig var karakterene lavere i 2015 enn i 2016 både målt i form av gjennomsnittskarakter og andel kandidater med C eller bedre. Fem år er kort tid som grunnlag for statistiske beregninger. De kan derfor ikke tillegges særlig vekt. Det er likevel interessant å merke seg at vi fikk en såpass høy korrelasjonskoeffisient som 0,88 og 0,81 for sammenhengen mellom «bedre enn C» og henholdsvis veid og samlet vanskelighetsgrad. Det betyr at i oppgavesett som ble ansett som vanskelig, var det færre som oppnådde C eller bedre, og ved oppgavesett som ble ansett som lettere, var karakterene bedre.

Konklusjoner

Undersøkelsen viste at det er vanskelig å vurdere vanskelighetsgrad og arbeidsmengde for eksamensoppgaver. Det var stor spredning ved vurdering av eksamensoppgavene enkeltvis, men bedre samling for hele eksamenssettenes vanskelighetsgrad og arbeidsmengde. Metodisk kunne vi som grunnlag for denne artikkelen ha anvendt andre og mer ressurskrevende måter å vurdere på.

Grafisk fremstilling gir klare indikasjoner på sammenheng mellom eksamensresultater og vanskelighetsgrad/arbeidsmengde. Selv med en nasjonal eksamen arrangert av NOKUT, kan den enkelte student være heldig eller uheldig med oppgavesettets vanskelighet.

Fem års tidshorisont er også en relativt kort periode til å konkludere noe utover de indikasjoner som en grafisk fremstilling gir.

Sensur er en viktig, men komplisert prosess. Andre forhold enn dem vi har nevnt foran, som for eksempel offentlig omtale, kan ha spilt inn. Hvor grensen for de ulike karakterene skal settes, både kvantitativ på en skala fra 0–100 % og kvalitativt ved en helhetlig tilnærming, vil bestandig være der. Vi mener som et utgangspunkt at en overordnet nivåjustering burde vært foretatt.

Vi konstaterer at det selv med nasjonale eksamener arrangert av NOKUT er mulig å stille spørsmål ved om studentenes rettssikkerhet er ivaretatt og metodisk robust nok.

Metodisk mener vi avviklingen av de nasjonale eksamenene kunne vært gjennomført på en bedre og mer robust måte, men det hadde krevd et annet og mer omfattende apparat. En nærmere analyse ville vært nyttig ved praktisering av nasjonale eksamener fremover.