Uniped årgang 37, 2/2014, side 33–43

Artikkelen er vitenskapelig vurdert av forskere utenfor redaksjonen

Av Ole Jakob Bergfjord

Studentevaluering i høyere utdanning – en empirisk studie fra HiB

Ole Jakob Bergfjord
Førsteamanuensis i økonomi ved Høgskolen i Bergen (HiB), Norge.
E-post: ojb@hib.no

Sammendrag
Studien benytter resultater fra studentevaluering av tre kurs ved bachelorutdanningen i økonomi og administrasjon ved Høgskolen i Bergen (HiB) til å undersøke eventuelle sammenhenger mellom kursomfang/vanskelighetsgrad, studentens egen innsats og deres vurdering av kurset. Hovedresultatet er at det er en positiv sammenheng mellom innsats og kursevaluering. Det er imidlertid ikke mulig å si noe om kausalitetsforhold – ytterligere forskning er nødvendig for å få bedre innsikt i disse sammenhengene.

Summary
The study is based on quantitative student evaluations of three undergraduate courses (Econ/Business) from the Bergen University College. Students were asked to rate different aspects of the courses, as well as report their own effort, and how difficult and extensive they perceived the course to be. The main purpose of this study is to look into possible relations between these variables, in particular how students' effort and their perception of the course as difficult or not influence their evaluation of the course. The results are somewhat surprising; it is particularly interesting that a positive correlation is found between effort and evaluation. Although a causal relationship has not been established, some possible implications of this finding are discussed, and some ideas for further research are proposed.

Introduksjon og litteratur

De prinsipielle argumentene for å gjennomføre studentevalueringer av undervisning er åpenbare. For det første gis studentene en mulighet til å komme med tilbakemeldinger. Både de reelle tilbakemeldingene og studentenes oppfatning av medbestemmelse er viktig, ikke minst i en tid hvor studentrollen i stadig større grad kan sammenlignes med kunderollen i et normalt marked for tjenester. For det andre bør systematisk evaluering gi incentiver og motivasjon til å yte sitt beste for den enkelte foreleser. For det tredje kan data fra studentevalueringer brukes som nyttig beslutningsstøtte i forbindelse

Citation: *Uniped* 2014, **37**: 22433 - http://dx.doi.org/10.3402/uniped.v37.22433

med ansettelser og styring av kurs- og studieporteføljer, og til slutt vil det selvsagt være et poeng at resultatet av evalueringene kan brukes til å gjennomføre forbedringer, både hos den enkelte foreleser og på institusjonsnivå.

Disse prinsipielle fordelene er utvilsomt en viktig grunn til at studentevalueringer synes å få en viktigere og viktigere rolle innen høyere utdanning, også i Norge. Viktigst så langt er kanskje effekten for den enkelte foreleser. Rent formelt tillegges studentevalueringer vekt i forbindelse med ansettelser, forfremmelser og lønnsoppgjør, men den uformelle effekten knyttet til status og mestringsfølelse kan også være viktig for mange. Betydningen utover individnivået er nok også til stede. Dersom et kurs over tid får dårlige studentevalueringer, kanskje til og med selv om det har blitt forelest av flere ulike forelesere, vil det være fristende å anta at det er noe galt med kurset, og enten fjerne det fra kursporteføljen eller gjøre dramatiske endringer. Hvis kursene innen et helt studium eller fagfelt får dårlige evalueringer, vil det på samme måte kunne få konsekvenser for hele studiet eller fagfeltet – enten direkte gjennom at institusjonsledelsen reduserer eller fjerner tilbudet, eller mer langsiktig og indirekte gjennom at tilstrømningen av nye studenter til feltet reduseres.

Vi skal komme tilbake til problemer knyttet til studentevalueringer, men for begrensningens del kan det være nyttig å innse at enkelte av disse problemene er vanskelige å løse. Studentevalueringer vil per definisjon være subjektive og speile den enkelte students oppfatning av kvalitet heller enn den objektive kvaliteten. Det er heller ikke åpenbart at det i det hele tatt gir mening å snakke om objektiv kvalitet for et kurs eller en foreleser, og i hvert fall ikke at det er rimelig å forvente at en slik objektiv kvalitet skal være uavhengig av de forholdene forskningen indikerer at påvirker evalueringen. Det er også verdt å påpeke at selv imperfekte studentevalueringer kan ha verdi. Taylor og Tyler (2012) finner at eksistensen av et formelt evalueringssystem har en positiv effekt på undervisningskvaliteten, og argumenterer for at dette skyldes at evalueringen gir foreleseren nyttig informasjon som gir grunnlag for forbedringer og/eller at evalueringer øker motivasjonen hos foreleseren og i hvert fall gjør det mindre attraktivt å minimere arbeidsinnsatsen knyttet til undervisning. Man kan selvsagt argumentere for at denne effekten i så fall bør være sterkere dess tettere koblingen mellom evalueringsresultater og reell kvalitet er, men det er uansett neppe verken realistisk eller ønskelig at studentevalueringer forsvinner fra høyere utdanning.

Selv om det altså er gode grunner til å benytte studentevaluering i høyere utdanning, eksisterer det en betydelig litteratur som tar opp en rekke problemer knyttet til å benytte slike evalueringer ukritisk. Mye av den viktigste litteraturen er godt oppsummert av Aarstad (2012), og jeg henviser til denne litteraturstudien for en grundig gjennomgang av tidligere arbeider. Det som er felles for de fleste bidragene på dette området, er at de påpeker skjevheter knyttet til at studentevalueringene i virkeligheten måler andre ting enn det de er ment å måle, nemlig kursets ellers foreleserens objektive kvalitet. Forhold som kan se ut til å påvirke studenters evaluering av et kurs eller en foreleser er f.eks. studentkvalitet og forventninger (Seiler, Seiler & Chiang, 1999; McPherson, 2006), kursets vanskelighetsgrad (Carrell & West, 2010), forelesers kjønn (Arbuckle & Williams, 2003), forelesers

personlighet (Radmacher & Martin, 2001), forelesers rase (Haskins, Rose-St. Prix & Elbaum, 1997) og forelesers fysiske attraktivitet (Riniolo, Johnson, Sherman & Misso, 2006).

Poenget er altså – for å sette det på spissen – at dersom studentevalueringene er dårlige, trenger ikke dette nødvendigvis bety at kurset eller foreleseren er dårlig; det kan like gjerne skyldes at foreleseren er kvinnelig, mørkhudet og lite fysisk attraktiv, studentene har lave forventninger og kurset er på et høyt faglig nivå. Hvis dette er tilfelle, blir det vanskelig å bruke studentevalueringene til noe som helst.[1] Det er derfor et poeng med best mulig kunnskap om når man kan stole på at resultatene gir et godt helhetsinntrykk og beslutningsgrunnlag, og hvilke konklusjoner man ikke kan trekke på bakgrunn av slike resultater. I denne artikkelen beskrives noen resultater fra en mindre empirisk studie i Norge, som forhåpentligvis kan gi noen nyttige indikasjoner på dette og også inspirere til debatt og videre studier av et komplekst, men viktig tema.

Datamateriale og metode

Studentevalueringer gjennomføres på mange måter. Man kan skille mellom formell og uformell evaluering, kvalitativ og kvantitativ evaluering, og evaluering underveis og i ettertid. I denne studien benytter vi kvantitative data fra en formell undersøkelse gjennomført via itslearning.[2] Det benyttes data fra underveisevalueringer gjort omtrent midt i semesteret heller enn evalueringer gjort etter at kurset er ferdig. Den viktigste grunnen til dette er pragmatisk; nesten dobbelt så mange studenter har svart på underveisevalueringen som sluttevalueringen. Man kan selvsagt si at studentene har et bedre grunnlag for å uttale seg etter at de har vært gjennom hele kurset, men på den andre siden gjennomføres sluttevalueringen flere måneder etter at undervisningen er ferdig, med påfølgende risiko for upresise erindringer og svar.

Respondentene er førstekullsstudenter ved HiBs bachelorstudium i økonomi og administrasjon. Studiet er populært, og opptakskravene har vært høye gjennom hele perioden som studeres. Man må derfor kunne anta at både motivasjon og forventninger hos studentene gjennom hele perioden har ligget på et stabilt og relativt høyt nivå. Samtidig er det verdt å merke seg at undersøkelsen gjennomføres midt i deres første semester som studenter. Dette betyr for eksempel at de har lite sammenligningsgrunnlag når de blir bedt om å uttale seg om kursene. Dette kunne vært unngått ved å benytte data fra andre eller tredje kull på samme studium, men dette ville igjen gitt et betydelig svakere datamateriale – antall studenter er færre, det er flere mulige valgfag, osv. Studentene har fire fag i første semester. I denne analysen er resultatene fra tre av fagene inkludert. Det siste faget er utelatt på grunn av at det ikke er direkte sammenlignbart – det har flere forelesere som får separate evalueringer på enkelte spørsmål, og kullet deles opp i ulike grupper (med ulike forelesere og opplegg) basert på forkunnskaper.

I analysen inngår resultater fra de samme tre kursene i 2010, 2011 og 2012. Hvert kurs har blitt evaluert av knapt 100 studenter, så totalt sett finnes det 758 kursevalueringer. Kursene har grovt sett hatt samme opplegg hvert år. Navn på kurs er her anonymisert av hensyn til de involverte foreleserne.

Siden de fleste studentene i hvert kull har (og evaluerer) alle de tre aktuelle kursene, vil altså antall studenter som er involvert i undersøkelsen være betydelig lavere enn 758. På grunn av anonymitet vet vi ikke hvilke studenter som har evaluert hvilke kurs, men dersom vi antar at hver involverte student i snitt har evaluert 2,5 kurs, skulle dette tilsi at ca. 300 studenter har vært involvert – dvs. ca. 100 per kull. På tidspunktet undersøkelsen ble gjennomført, har kullene bestått av ca. 120–130 studenter, så svarprosenten er altså relativt god. Når det gjelder representativitet, er det vanskelig å si noe sikkert om denne så lenge deltagerne er anonyme. En naturlig antagelse kan være at de studentene som *ikke* deltar i undersøkelsen, har færre sterke meninger om kursene enn de øvrige (både på godt og vondt). Kanskje kan man også tenke seg at denne gruppen har lavere engasjement og legger mindre innsats i studiene, men dette blir kun spekulasjoner. Så lenge en så stor andel av kullet deltar i evalueringen, vil representativiteten uansett neppe være et kritisk problem.

Som det vedlagte skjemaet viser, ble studentene bedt om å vurdere tre egenskaper ved hvert kurs på en skala fra 1 (meget høyt) til 5 (meget lavt). De tre egenskapene var læringsutbytte av forelesningene, læringsutbytte av innleveringer og læringsutbytte av lærebok. I tillegg ble de, på samme skala, bedt om å vurdere kursets vanskelighetsgrad, kursets omfang og sin egen arbeidsinnsats i kurset.[3]

Det er verdt å merke seg at dette datasettet gir grunnlag for en rekke mulige studier. I denne sammenhengen er det derfor et poeng å begrense seg til noen sentrale temaer, og jeg har valgt å se på sammenhengen mellom studentevaluering og studentenes oppfatning av kursomfang, vanskelighetsgrad og egeninnsats. Dette er interessante temaer, fordi de berører aspekter ved tidligere forskningsresultater samtidig som disse variablene, så vidt jeg vet, ikke er direkte studert tidligere.

Som tidligere nevnt består datamaterialet av 758 kursevalueringer. Det er verdt å kort gjøre rede for hvordan datamaterialet er bearbeidet i forhold til råmaterialet som følger av det vedlagte skjemaet. La oss først se på det vi er mest opptatt av, nemlig målet for studenttilfredshet eller studentens evaluering av kurset. I spørreskjemaet ser vi at studentene er bedt om å komme med tre vurderinger av dette. De er bedt om å vurdere tilfredshet med foreleser, tilfredshet med lærebok og tilfredshet med innleveringsoppgaver. I den videre analysen er (det uvektede) gjennomsnittet av disse tre brukt som et mål på studentenes evaluering av kurset, og denne variabelen kalles videre «evaluering». Det er verdt å merke seg at det er en relativt sterk parvis korrelasjon (rundt 0,35 i gjennomsnitt) mellom scoren studentene gir på disse tre områdene. Dette er for så vidt ikke overraskende, og kan forklares på flere måter. Man kan tenke seg at det finnes en underliggende «objektiv» kvalitetsfaktor som ligger under alle disse målene, man kan tenke seg at underliggende evner, motivasjon og interesse påvirker alle målene i samme retning, eller man kan forestille seg en form for projisering, f.eks. at dersom man liker foreleseren, blir man mer positivt innstilt til læreboken. I hvor stor grad hver enkelt av disse hypotesene stemmer, er et åpent spørsmål, og på siden av denne studien. Poenget er uansett at de tre scorene er korrelerte, og at gjennomsnittet synes å være et godt egnet mål for studentens totale tilfredshet med kurset.

Studentenes rapporterte oppfatning av omfang, vanskelighetsgrad og egen innsats er brukt slik de står. I tillegg er det tatt inn dummy-variabler (dummyX, dummyY, dummyZ) for å kunne se hvilket kurs (X, Y eller Z) vurderingen gjelder. Dette skyldes at det er betydelige forskjeller mellom gjennomsnittsscoren for de ulike kursene, og at det dermed kan være interessant å ha muligheten til å trekke ut dette som en egen effekt. Dette gjøres i praksis ved at en evaluering av kurs X får verdi 1 på variabelen «dummyX» og verdi 0 på variablene «dummyY» og «dummyZ», mens en evaluering av kurs Y får verdi 1 på variabelen «dummyY» og 0 på de to andre osv. Poenget med dette er at man i en regresjonsanalyse kan se hvordan dummy-variablene (som altså er indikatorer på hvilket kurs evalueringen gjelder) påvirker den avhengige variabelen.

Analyse

Korrelasjon mellom ulike variabler

Neste steg blir å undersøke sammenhengen mellom de ulike variablene ved hjelp av statistisk analyse. Hovedpoenget i denne delen er å studere hvordan de andre variablene påvirker studentenes evaluering av kurset, men det kan innledningsvis være interessant å se litt på den interne sammenhengen mellom de øvrige variablene. Delvis fordi dette potensielt kan påvirke regresjonsanalysen via multikollinearitet og delvis fordi sammenhengen kan være interessant i seg selv. Tabell 1 viser den parvise korrelasjonen mellom oppfattet omfang, vanskelighetsgrad og innsats på et kurs.[4]

Tabell 1. Korrelasjon mellom ulike variabler.

	Vanskegrad	Omfang	Innsats
Vanskegrad	1.0000		
Omfang	0.5535*	1.0000	
Innsats	0.1624*	0.3281*	1.0000

Det er ikke overraskende at det er en sterk korrelasjon mellom studentenes oppfatning av et kurs' omfang og det samme kursets vanskelighetsgrad. Det er en klar positiv korrelasjon mellom innsats og de øvrige to variablene også, men denne er betydelig svakere. Særlig korrelasjonen mellom innsats og opplevd vanskelighetsgrad (0,16) kan oppfattes som overraskende svak, noe som kanskje kan oppfattes som betenkelig. Som kursansvarlig vil man typisk ønske en sterk korrelasjon her: Dersom studentene pga. gode forkunnskaper eller evner oppfatter kurset som lett, kan man kanskje akseptere en lavere innsats dersom de uansett lærer det de skal, mens man vil ønske at studentene legger inn en ekstra innsats dersom de opplever kurset eller deler av kurset som vanskelig. Dette synes kun i begrenset grad å stemme. Man kan tenke seg at omfang og vanskelighetsgrad i enkelte sammenhenger virker demotiverende og dermed reduserer innsatsen, men man kan også tenke seg at studenter har begrenset med tid og ressurser slik at de i praksis ikke nødvendigvis har muligheten til å øke innsatsen selv om kursomfang eller vanskelighetsgrad kanskje skulle tilsi det.

Regresjonsanalyse av hvordan ulike variabler påvirker evalueringen

Så over til hovedtemaet for denne analysen, nemlig hvordan de øvrige variablene påvirker studentenes evaluering av et kurs. Dette undersøkes her ved hjelp av en lineær regresjon på følgende form:

$$\text{Evaluering} = A1 * \text{Omfang} + A2 * \text{Vanskegrad} + A3 * \text{Innsats} + A4 * \text{DummyX} + A5 * \text{DummyY} + \text{Konstant}$$

Ved hjelp av dataprogrammet STATA finnes de koeffisientene A1–A5 som minimerer summen av de kvadrerte avvikene i modellen, og dermed presumptivt gir den beste beskrivelsen av hvordan de andre variablene påvirker evalueringen. Tabell 2 viser resultatene fra denne regresjonsanalysen[5]:

Tabell 2. Regresjon – hvordan påvirkes studentevalueringen av et kurs av andre variabler? Koeffisienter merket med * signifikant (95 %) forskjellig fra 0.

Source	SS	df	MS	Number of obs = 758 F(5,752) = 144.64 Prob > F = 0.0000 R-squared = 0.4902 Adj R-squared = 0.4869 Root MSE = 0.55093		
Model	219.510128	5	43.9020255			
Residual	228.247165	752	0.303520166			
Total	447.757293	757	0.591489158			
Evaluering	**Coef.**	**Std. Err.**	**t**	**P > |t|**	**[95% Conf. Interval]**	
Vanskegrad	0.0914*	0.0334441	2.73	0.006	0.0257451	0.157055
Omfang	0.0480961	0.0384053	1.25	0.211	−0.0272983	0.1234904
Innsats	0.257489*	0.0297077	8.67	0.000	0.1991691	0.315809
DummyX	−1.051969*	0.0511661	−20.56	0.000	−1.152414	−0.9515238
DummyY	−0.625791*	0.0518735	−12.06	0.000	−0.7276251	−0.5239569
_cons	1.972704*	0.1183374	16.67	0.000	1.740394	2.205015

Denne enkle modellen gir en rekke interessante resultater. Det første vi kan notere oss er at de inkluderte variablene til sammen forklarer knapt halvparten av variasjonen i studentenes evalueringer ($R^2 \approx 0{,}49$). På den ene siden er dette en relativt god forklaringsgrad i forhold til mange andre lignende modeller. På den andre siden er det verdt å merke seg at hvilket kurs som evalueres (via dummy-variablene) inngår blant variablene. Dersom det fantes en «objektiv» kurskvalitet, ville dummy-variablene alene forklart all variasjon i studentevalueringen, mens de i virkeligheten altså ikke en gang forklarer halvparten, selv sammen med tre andre variabler. Hvilke faktorer som forklarer den andre halvparten av variasjonen, er derfor selvsagt et interessant spørsmål; en hypotese kan være at variablene nevnt i innledningen (studentens forventninger, forelesers kjønn/personlighet, osv.) kan være en del av forklaringen her.

Videre er naturlig nok dummy-koeffisientene signifikante – alt annet likt er det altså signifikante og systematiske forskjeller mellom evalueringen av ulike kurs. Fag Z er holdt utenfor regresjonen, og tjener dermed som et utgangspunkt. Tolkningen her er at evalueringen i fag X, alt annet holdt likt, er 1,05 lavere (dvs. bedre) enn i fag Z, mens evalueringen i fag Y, alt annet holdt likt, ligger 0,62 lavere enn i fag Z.[6] Det er naturlig og forventet at disse koeffisientene er store og signifikante – dette betyr jo bare at det faktisk har betydning for evalueringen hvilket kurs som blir evaluert.

Videre ser vi på koeffisientene for vanskelighetsgrad og omfang. Begge disse er relativt små, men positive. Basert på tradisjonelt signifikansnivå (95 %), har opplevd vanskelighetsgrad statistisk signifikant betydning for evalueringen, mens omfang ikke har det. En positiv sammenheng betyr her at når vanskelighetsgraden eller omfanget øker, vil typisk evalueringen bli bedre.

Dette er et interessant resultat, fordi det til en viss grad motsier resultatene fra enkelte tidligere studier (se for eksempel Carrell & West, 2010), hvor det kan se ut som om kurs som oppfattes som vanskelige, typisk får dårligere evalueringer. Begge koeffisientene er små, og koeffisienten for omfang er ikke statistisk signifikant. Det er derfor ikke grunnlag for å trekke noen klare konklusjoner på grunnlag av denne studien, men det er likevel ikke åpenbart hvordan en slik sammenheng eventuelt kan forklares. En mulig forklaring – og selvsagt en hyggelig forklaring fra foreleserens ståsted – er at studenter faktisk liker å bli utfordret, og dermed setter pris på kurs som er litt vanskelige og/eller omfattende.

Om sammenhengen mellom egeninnsats og kursevaluering

Mest interessant er likevel sammenhengen mellom innsats og evaluering. Denne er positiv, signifikant og klart sterkere enn effektene fra omfang og vanskelighetsgrad. Dette betyr altså at dess større innsats som legges ned, dess bedre evaluering vil man typisk gi – alt annet holdt likt. Dette er igjen et resultat som lett kan tolkes som positivt fra en forelesers ståsted: Hvis man får/tvinger studentene til å legge ned mer innsats i kurset, vil dette også gjøre dem mer fornøyde med kurset.

Dette er selvsagt én mulig tolkning. Men det finnes flere mulige tolkninger, og ut fra den statistiske analysen alene er det vanskelig å komme frem til en endelig konklusjon. Det mest åpenbare problemet er at korrelasjon ikke betyr kausalitet – vi kan ikke ut fra denne analysen si at høy innsats leder til høyere tilfredshet. En slik tolkning er ikke urimelig; det er lett å se for seg at høy innsats bidrar til mestringsfølelse, som igjen bidrar til økt tilfredshet og bedre evaluering av kurset. Det er imidlertid minst like lett å se for seg en kausalitet som går motsatt vei, nemlig at et kurs og en foreleser man liker gir økt motivasjon for kurset, noe som igjen bidrar til at man øker innsatsen i kurset.

I stedet for eller i tillegg til disse sammenhengene, kan man selvsagt tenke seg at det finnes faktorer utenfor modellen som påvirker både innsats og evaluering, og dermed skaper inntrykk av en spuriøs sammenheng mellom variablene. Én slik faktor er den enkelte student. Man kan tenke seg at personlighet, studiesituasjon eller andre individuelle forhold påvirker både innsatsen og tilfredsheten i samme retning. Dette kunne vært kontrollert for ved å innføre dummy-variabler for hver enkelt student, men siden hver

student maksimalt vurderer tre kurs (og ofte bare ett eller to), ville dette gjort modellen vanskelig å håndtere.

Implikasjoner og konklusjon

Det er selvsagt umulig å trekke bastante konklusjoner på bakgrunn av en såpass begrenset studie. Enkelte implikasjoner kan man likevel se for seg. En mulig og oppløftende tolkning er at vi som kursansvarlige ikke bør være redde for å utfordre studentene. Det burde være et selvsagt mål for oss at studentenes utbytte skal være så bredt og dypt som mulig – noe som stort sett også krever at de legger ned mye innsats. Man opplever tidvis et press fra studenter for å senke kravene («pensum er for stort», «kurset krever for mye arbeid – vi har andre fag også»), men denne studien kan peke i retning av at et ambisiøst og krevende opplegg i seg selv ikke gjør studentene mindre tilfredse.

Et annet poeng er at denne studien viser nytten av å samle inn mer informasjon enn å kun be studentene gi kurset en rating. Egenvurderingen av innsats, kursomfang og vanskelighetsgrad som er brukt i evalueringsskjemaet ved HiB, er et eksempel på slik informasjon som kan være nyttig for å danne seg et bedre bilde av situasjonen. Det finnes selvsagt også en rekke andre variabler man kunne være interessert i å studere for å kunne gi et bedre svar på om undervisningen faktisk har vært vellykket. Her kan det særlig pekes på eksamensresultater. På grunn av anonymitetskrav, har det ikke vært mulig å koble den enkelte students vurdering av et kurs til den samme studentens endelige karakter i kurset, men dette ville åpenbart vært nyttig. I den grad eksamensopplegg og karakterkrav er stabile og pålitelige, burde jo eksamenskarakterer være en nyttig indikator på hvor stort utbytte studentene faktisk fikk av kurset.

Et tredje poeng går på bruken av evalueringene. Tidligere forskning (se Aarstad, 2012 for en utfyllende oversikt) viser altså at det finnes en rekke faktorer utenfor faglærers kontroll som påvirker kursevalueringen (for eksempel forelesers kjønn, rase, alder osv.). Det er derfor klart at evalueringen som sådan alltid må tas med en klype salt. Et bedre datamateriale (og økt kunnskap om de ulike variablenes innhold og sammenheng) gjør det lettere å bruke evalueringene på en konstruktiv måte, både for å vurdere kurs og foreleser og for å gjennomføre forbedringer. Denne studien er i så måte et eksempel på hvordan mer utfyllende data kan analyseres og påvirke hvordan studentevalueringene brukes. Ved å koble studentevalueringene til innsats (eller andre faktorer), vil man kunne få et mer helhetlig perspektiv på resultatene fra evalueringen. Spørsmålet blir ikke lenger bare «hva bør foreleseren endre på?», men også (for eksempel) «hvordan kan vi legge til rette for økt motivasjon og innsats i dette kurset?» Sagt på en annen måte: Hvis dårlige studentevalueringer er et symptom på problemer, vil mer utfyllende data gjøre det lettere å se hvorfor evalueringen er dårlig slik at man kan bruke evalueringen til å gjøre noe med de underliggende problemene.

Til slutt er det på sin plass å si litt om undersøkelsens svakheter og peke på noen områder for videre studier. En åpenbar svakhet er at undersøkelsen er basert på egenrapporterte data. Man husker kanskje best de siste to dagers skippertak med innleveringsoppgaven i fag X og rapporterer stor innsats i dette faget, selv om man reelt sett har brukt mer tid på fag Y.

Når det gjelder variabelen «innsats», er det heller ikke åpenbart at dette betyr det samme for alle eller at alle typer innsats kan sidestilles. Innsats knyttet til å få klarhet i et ustrukturert undervisningsopplegg vil for eksempel være mindre ønskelig enn innsats knyttet til faglig fordypning, men i vår sammenheng skilles det ikke mellom de to.

Av mer statistiske problemer er det som nevnt vanskelig å si noe om eventuelle kausale sammenhenger basert på denne studien. Det fremkommer heller ikke hvilke andre forhold (utenfor vår modell) som eventuelt påvirker evalueringen fra disse studentene. Det er selvsagt også – som alltid – skjevheter i utvalget vårt; vi vet ikke om det er systematiske forskjeller mellom de som velger å fylle ut et slikt evalueringsskjema og de som ikke gjør det.

Når det gjelder videre studier, er det mange interessante vinklinger man kan tenke seg. Som tidligere nevnt finnes det en rekke andre variabler som kan studeres i en slik sammenheng. Det er også klart at førsteårsstudenter ved HiBs bachelorstudium i økonomi og administrasjon ikke nødvendigvis er representative for alle andre studenter. Man kan tenke seg at studenter med lengre erfaring, studenter med andre forkunnskaper og forventninger og studenter innenfor andre fagområder vil vurdere kurs på en annen måte enn disse studentene ser ut til å gjøre.

Til slutt kunne det også være interessant å gjennomføre mer kvalitative studier – for eksempel å gjennomføre dybdeintervjuer med studenter for å prøve å finne ut mer om hvordan de resonnerer når de evaluerer et kurs og hvilke faktorer de selv mener er relevante.

Litteratur

Aarstad, J. (2012). Studentevalueringer i høyere utdanning: Hva kan den internasjonale forskningslitteraturen lære oss? *Uniped*, *35*(1), 34–45.

Arbuckle, J. & Williams, B. D. (2003). Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. *Sex Roles*, *49*(9/10), 507–516.

Carrell, S. E. & West, J. E. (2010). Does Professor Quality Matter? Evidence from Random Assignment of Students to Professors. *Journal of Political Economy*, *118*(3), 409–432.

Haskins, A. R., Rose-St. Prix, C. & Elbaum, L. (1997). Covert bias in evaluation of physical therapist students' clinical performance. *Physical Therapy*, *77*(2), 155–163.

McPherson, M. A. (2006). Determinants of how student evaluate teachers. *The Journal of Economic Education*, *37*(1), 3–20.

Radmacher, S. A. & Martin, D. J. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *The Journal of Psychology*, *135*(3), 259–268.

Riniolo, T. C., Johnson, K. C., Sherman, T. R. & Misso, J. A. (2006). Hot or not: Do professors perceived as physically attractive receive higher student evaluations? *The Journal of General Psychology*, *133*(1), 19–35.

Seiler, M. J., Seiler, V. L. & Chiang, D. (1999). Professor, students, and course attributes that contribute to successful teaching evaluations. *Financial Practice and Education*, *9*(2), 91–99.

Taylor, E. S. & Tyler, J. H. (2012). The Effect of Evaluation on Teacher Performance: Evidence from Longitudinal Student Achievement Data of Mid-career Teachers. *American Economic Review*, *102*(7), 3628–3651.

Noter

1 Carrell og West (2010) observerer til og med negativ korrelasjon mellom studentevaluering og karakterer i påfølgende kurs, som man skulle tro var en brukbar indikasjon på reell kvalitet. I så fall er problemet selvsagt enda større: Ved å belønne forelesere som får gode studentevalueringer, vil man da systematisk belønne dårlig kvalitet.

2 Et utdrag av spørreskjemaet studentene ble invitert til å fylle ut er vedlagt.

3 På alle disse spørsmålene var det også et øvrig svaralternativ: «Hadde ikke kurset». Respondenter som svarte dette eller svarte blankt, har blitt fjernet fra datasettet.

4 Dummy-variablene er her holdt utenfor. Ved å finne variablenes «variance influence factors» i STATA, er det testet for multikollinearitet, og dette ser ikke ut til å være et problem av betydning her. Stjerne markerer at korrelasjonen er signifikant (95 % nivå) forskjellig fra 0.

5 Tester viser noe tegn til heteroskedastisitet i datamaterialet. Analysen er kjørt på nytt med robuste standardavvik (White). Dette gir kun små endringer i resultatene og ingen endringer i konklusjonene – har derfor valgt å rapportere OLS-analysen i sin enkleste form.

6 At fortegnet for disse koeffisientene er negativt og at evalueringen er lavere, er her en god ting. Slik skjemaet er formulert og kodet, er en vurdering på 1 det beste man kan oppnå.

Appendiks – eksempel på formulering, spørreskjema

1. Matrisespørsmål
Her ønsker vi at du skal krysse av for hvordan du mener faget X har fungert dette semesteret

	Meget høyt (1)	Høyt (2)	Ok (3)	Lavt (4)	Meget lavt (5)
Læringsutbyttet av forelesningene var	☐	☐	☐	☐	☐
Læringsutbyttet av innleveringene var	☐	☐	☐	☐	☐
Utbyttet av læreboken var	☐	☐	☐	☐	☐
Vanskelighetsgraden i faget var	☐	☐	☐	☐	☐
Arbeidsomfanget i faget var	☐	☐	☐	☐	☐
Min innsats i faget var	☐	☐	☐	☐	☐

2. Åpent spørsmål
Har du kommentarer til faget X? Kommenter gjerne hvorfor du mener det var godt eller dårlig, hva som var spesielt bra og forslag til forbedringer.

(De samme spørsmålene ble stilt for alle kursene studentene hadde i det aktuelle semesteret.)