Uniped årgang 36, 1/2013, side 26–41

Artikkelen er vitenskapelig vurdert av forskere utenfor redaksjonen

Av Vidar Gynnild

"Kriteriebasert vurdering" – hva innebærer det i praksis?

Vidar Gynnild
Norges teknisk-naturvitenskapelige universitet, PLU, Seksjon for universitets-pedagogikk.
E-post: vidar.gynnild@plu.ntnu.no

Sammendrag
Universitets- og høgskolerådet (UHR) har slått fast at kriteriebasert vurdering skal benyttes i høyere utdanning, men uten å forklare hvordan prinsippet skal omsettes i praksis. Denne artikkelen reiser spørsmål ved prinsipielle og teoretiske sider ved kriteriebasert vurdering, og benytter fire ulike tolkninger av begrepet som teoretisk rammeverk. To tilfeldig valgte emner er valgt for å eksemplifisere hvordan ulike tolkinger av kriteriebasert vurdering påvirker de endelige karakterfordelingene. Mens bruken av ekstern sensor har vært sett på som det viktigste sikringstiltaket ved vurdering, viser denne undersøkelsen av operasjonaliseringen av vurderingsprinsippet kan være enda mer betydningsfullt for de endelige karakterene. Selv om eksemplene er hentet fra en konkret utdanning, er de prinsipielle utfordringene felles for all høyere utdanning. Artikkelen har derfor interesse for alle med ansvar for vurdering.

Abstract
The Norwegian Association of Higher Education Institutions (UHR) has decided that Criterion Based Assessment is to be used in all higher education in Norway; however, without any guidance on its implementation. This study introduces four interpretations of Criterion Based Assessment as an analytical framework, and two courses are randomly selected to provide evidence of the effects of the different interpretations of the assessment principle. Yet the purpose of the study is principal and theoretical, it also brings evidence on the effects of various interpretations of the assessment principle. While the use of an external examiner has been considered the primary quality assurance measure in assessment and grading, this study shows that the interpretation of the selected assessment principle may be of even greater significance. This article should therefore be of interest to all with a stake in assessment.

Innledning

Til tross for at det fra ulike hold hevdes at karakterene ikke er det viktigste ved en utdanning, finnes det godt belegg for at vurdering og vurderingsuttrykk rører ved vesentlige spørsmål i et utdanningsløp.

Citation: *Uniped* 2013, **36**: 20959 - http://dx.doi.org/10.3402/uniped.v36i1.20959

Karakterer fungerer som et seleksjonsfilter ved opptak til videre studier, og benyttes ofte ved søknad på stillinger. Karakterene er derfor også et virkemiddel for å styre studentenes innsats, både tematisk og tidsmessig (Boud, Cohen, & Sampson, 1999; Rust, 2002). Generelt knytter det seg derfor interesse til hva karakterene uttrykker og om vurdering skjer med utgangspunkt i et felles forstått evalueringsteoretisk rammeverk. Spørsmålet er blitt aktualisert under arbeid fram mot et felles europeisk utdanningsmarked, med behov for å rydde opp i jungelen av vurderingsordninger (The European Higher Education Area, 1999).

Denne artikkelen fokuserer først på myndighetsnivåets bidrag til felles institusjonelle praksiser, der Universitets- og høgskolerådets utdanningsutvalg (UU) har spilt en sentral rolle. I forlengelsen av dette rapporterer jeg fra en undersøkelse av vurdering og karaktersetting i to tilfeldig valgte emner innen høyere utdanning. Her dokumenterer jeg hvordan vurdering og karaktersetting i dette tilfellet foregikk og hvordan ulike tolkninger av *kriteriebasert vurdering* i praksis fører til vidt forskjellig karakterfordeling. Hensikten med den empiriske undersøkelsen er å bidra til det John Biggs har kalt «the descent from rhetoric» (Biggs, 1996) ved å undersøke hvordan prinsipper for vurdering på nasjonal basis blir tolket og omsatt i praksis.

Bakgrunn

Hensikten med dagens karaktersystem var å etablere en felles nasjonal ordning for all høyere utdanning. Forut for dette eksisterte det et mangfold av vurderingsordninger med ulik bruk av karakterskalaen innen ulike utdanninger. Dette ble oppfattet som problematisk etter hvert som et økende antall studenter ønsket å kombinere emner avlagt ved flere ulike institusjoner som del av en akademisk grad. Behovet for en mer enhetlig praksis ble dermed maktpåliggende. I samme retning dro ønsket om økt internasjonal mobilitet med innpassing av emner og grader i nasjonale utdanningsløp. Som svar på denne utfordringen, nedsatte Universitetsrådet i 1999 en arbeidsgruppe som skulle utrede forslag til nasjonalt karaktersystem. Resultatet ble nedfelt i en egen rapport (Det norske universitetsråd, 2000). Dagens karaktersystem ble deretter tatt inn i lov om universiteter og høgskoler og innført ved universiteter og høgskoler høsten 2003 (Glasser, 2008). Som oppfølging til dette, kom Utdannings- og forskningsdepartementet i eget brev (10.05.2004) med nærmere retningslinjer om hvordan karaktersystemet var ment å fungere. Her slås det fast at karaktersettingen skal ta utgangspunkt i de verbale beskrivelsene som er gitt for hvert nivå og at kravene for bestått «... ikke skal gjøres avhengig av studentenes forutsetninger for å gjennomføre emnet» (Kunnskapsdepartementet, 2004).

Som de fleste vil være kjent med, består karaktersystemet av en bokstavskala fra A-F og en todelt skala med bestått og ikke bestått. I en rapport fra en arbeidsgruppe oppnevnt av Universitets- og høgskolerådet slås det fast at begge skalaene er absolutte «i den forstand at de er kriteriebaserte. Dersom en prestasjon tilfredsstiller kriteriene for en karakter, så skal man gi denne karakteren uavhengig av hvordan karakterfordelingen av de øvrige karakterene i eksamenskullet er» (Glasser, 2008). Samtidig sies det at karakterfordelingen over tid og for et stort antall kandidater skal være i «rimelig samsvar med den relative ECTS-skalaen» (Glasser, 2008). Bruken av karakterskalaen ligger med andre ord i spenningsfeltet mellom

to ulike prinsipper for vurdering, oftest omtalt som *kriteriebasert* og *normbasert vurdering*. Mens det sistnevnte prinsippet historisk sett har vært mest vanlig, har kravet om kriteriebasert vurdering blitt fremmet med stadig større styrke. Vi utforsker disse begrepene nærmere i det følgende.

Perspektiver og terminologi

Arbeidet med det nasjonale kvalifikasjonsrammeverket signaliserer en perspektivforskyvning fra undervisning til læring, fra virkemidler til resultat. Dette blir noen ganger omtalt som et paradigmeskifte, altså som en endret virkelighetsoppfatning med følger for målbeskrivelser, kunnskaps- og læringssyn, arbeidsmåter og suksesskriterier (Barr & Tagg, 1995):

> «*In the Instruction Paradigm, the mission of the college is to provide instruction, to teach. The method and the product are one and the same. The means is the end. In the Learning Paradigm, the mission of the college is to produce learning. The method and the product are separate. The end governs the means.*»
> *(Barr & Tagg, 1995, s. 15)*

Et lett synlig uttrykk for instruksjonsparadigmets dominans gjenspeiles i omfattende bruk av studentevaluering av undervisningen, med liten eller ingen vekt på studentenes læring. Under dette perspektivet forble vurdering og karaktersetting lite utviklet, både terminologisk og metodisk. Kandidatenes prestasjoner ble i stor grad vurdert i forhold til hverandre, slik Biggs beskriver:

> «*A common method [. . .] depends on how students compare to each other ('norm-referenced'), rather than on whether an individual's learning meets the intended outcomes ('criterion-referenced'). In the former case, there is no inherent relation between what is taught and what is tested. The aim is to get a spread between students, not to see how well individuals have learned what they were supposed to have learned.*»
> *(Biggs, 2007, s. 61)*

Gjennom dette sitatet blir vi eksponert for to ulike rammeverk for vurdering: *normbasert* og *kriteriebasert vurdering*. Det førstnevnte har en lengre historie å vise til sammenlignet med kriteriebasert vurdering, og er enklere å operasjonalisere. Normbasert vurdering representerer et effektivt våpen mot inflasjon i karaktersystemet, men er ellers langt fra uproblematisk som metode. Vurderingskriteriene er forankret i en annen gruppes prestasjoner, og karakterene blir i dette tilfellet ingen pålitelig indikator på faglig nivå i absolutt betydning. «It is structurally self-adjusting in that it is automatically blind to such factors as the quality of teaching and the design of the assessment program» (Sadler, 2010). Forventninger til en gitt fordeling demper inflasjon i karaktersystemet, men kan oppleves som urettferdig ved at karakterfordelingen blir viktigere enn å dokumentere reelt prestasjonsnivå. Begrepsmessig må *normbasert vurdering* skilles fra *normalfordelt karakterfordeling*, som er et statistisk begrep, slik Sadler beskriver:

> «*It could be symmetrical and bell-shaped; it could be skewed, or even rectangular. There is no intrinsic connection between norm-referencing and the normal probability distribution in statistics, although it often happens that the distribution for large classes turn out to be roughly bell-shaped.*»
> *(*Sadler, 2010*)*

Spørsmålet om hva kriteriebasert vurdering er, eller kan forstås som, har fått særlig aktualitet etter innføring av det nasjonale kvalifikasjonsrammeverket for høyere utdanning med dets vekt på resultatene av læringsarbeidet. Dette perspektivet avdekket et behov for et velfundert begrepsapparat for å skille ulike fenomener fra hverandre. Mens *læringsmål*, eller tilsiktet læringsresultat, uttrykker intensjonalitet, er *bedømt læringsresultat* et evalueringsrelatert begrep som forutsetter kriterier og standarder, enten implisitt eller eksplisitt. Begrepet *læringsutbytte* er også et evalueringsrelatert begrep, men her forvaltes kriterier og standarder av kandidatene selv, for eksempel i utsagn som: «Jeg hadde et godt læringsutbytte i emne x.» I *kriteriebasert vurdering* forvaltes kriterier og standarder av faglærer (og eventuelt sensor), og samlet vurdering skjer ved avslutning av emne, aldri som et forhold mellom før og etter gjennomføring av et læringsforløp. Dette bekreftes i tidsangivelsen for læringsmålene, med vekt på innledninger som: «*Etter å ha gjennomført emnet*, skal studentene være i stand til å . . .»

I motsetning til normbasert vurdering, har *kriteriebasert vurdering* en kortere historie å vise til, samtidig som det ikke finnes entydige retningslinjer for å operasjonalisere begrepet. I denne artikkelen har jeg valgt Sadlers undersøkelse (Sadler, 2005) som teoretisk ramme for analysen. Hensikten er altså ikke å lage en omfattende litteraturstudie, men å gjennomføre en avgrenset empirisk studie i to tilfeldig valgte emner. Hensikten er å undersøke resultatene av ulike tolkninger av *kriteriebasert vurdering*.

Hva er kriteriebasert vurdering?

Evaluering og *vurdering* blir her benyttet synonymt i tråd med følgende definisjon:

> «Assessment *[. . .] refers to the process of forming a judgment about the quality and extent of student achievement or performance, and therefore by inference a judgment about what learning has taken place.*»
> *(Sadler,* 2005*)*

Bruken av *kriteriebasert vurdering* har vært myndighetenes svar på hvordan dette skal oppnås, med vekt på bruk av karakterbeskrivelser i operasjonaliseringen av prinsippet. Noen erfarer likevel at beskrivelsene ikke gjør oppgaven enklere og i noen sammenhenger viser seg utilstrekkelig eller uegnet, slik Sadler beskriver:

> «*For a range of related epistemological reasons, it is often impossible to express quality-based achievement standards in purely propositional or declarative form. They cannot be written down as detailed verbal descriptions, categories or lists which can then be used by students and assessors alike.*»
> *(Sadler,* 2009*, s. 820)*

På norsk brukes begrepet *kriteriebasert vurdering* i en betydning som tildekker det analytiske skillet mellom *kriterier* og *standarder*:

> «*For the purpose of making progress specifically on developing the concept of standards referenced assessment and grading, criteria and standards need to be distinguished, because they play different roles.*»
> *(Sadler, 2009, s. 819)*

Kriteriene angir *hva som skal vurderes*, mens *prestasjonens nivå* bestemmes ut fra standarder beskrevet som «fixed external anchor points» (Sadler, 2009, s. 819). Definisjonen av en standard inneholder, ifølge Sadler (2009), tre viktige elementer: Den er sosialt forankret, og forblir ideelt sett stabil over en viss tidsperiode. Standarder praktiseres ulikt innen ulike disipliner, og ikke minst viktig: «They are the property of the academy as a collective, and are not determined or held privately by individual teachers and course teams» (Sadler, 2009, s. 819). I praksis benyttes kriterier og standarder som del av enten en holistisk (helhetlig) eller analytisk prosess. Sistnevnte metode er mye benyttet i naturvitenskapelig og ingeniørfaglig sammenheng fordi oppgaver relativt lett lar seg stykke opp i mindre deler, som så blir gitt scorer eller poeng.

I vårt tilfelle ble prosentpoeng benyttet på en skala fra 0–100. Scorene uttrykker ikke absolutt verdi, men refererer til sosialt bestemte kriterier og standarder, som omtalt foran. I en studie konkluderte Sadler (2005) med fire ulike tolkninger av *kriteriebasert vurdering*, der han hevdet at begrepsforståelsen varierer både mellom og innad ved institusjoner. Ettersom begrepsdefinisjonene står sentralt i denne analysen, gis en kort introduksjon i det følgende:

Grading Model 1: Achievement of course objectives (Sadler, 2005)

Det sentrale spørsmålet er her i hvilken grad læringsmålene blir oppfylt. Dette nedfelles i rammeverket for evaluering, for eksempel slik at karakterbeskrivelsene systematisk og konsekvent refererer til grad av måloppnåelse. Denne tolkningen gjenfinnes i rapporten fra en arbeidsgruppe nedsatt av Universitets- og høgskolerådet (UHR): «Arbeidsgruppen foreslo videre at prestasjonene innenfor hvert emne vurderes i henhold til etablerte læringsmål for emnet» (Glasser, 2008). Som jeg har redegjort for i annen sammenheng, ble karakterbeskrivelsene ikke konsistent utformet i tråd med det valgte utgangspunktet (Gynnild, 2010). Dels ble karakterbeskrivelsene utformet med læringsmål som referanse, dels med relativ referanse, for eksempel som «[. . .] prestasjon som klart utmerker seg» [i forhold til resten av gruppen] (Glasser, 2008).

Grading Model 2: Overall achievement as measured by score totals (Sadler, 2005)

Denne modellen benyttes ved en rekke institusjoner. Begrepet *kriteriebasert* knyttes her til at bestemte krav, eller terskler, er nådd. De fleste institusjoner deler skalaen i segmenter eller trinn, for eksempel slik det er blitt gjort i sivilingeniørstudiet ved NTNU (se Tabell 1). Dette er altså en institusjonsspesifikk skala som ikke er nedfelt i lovverket for høyere utdanning. Andre institusjoner følger ikke nødvendigvis de samme karakterintervallene, som beskrevet i Tabell 1. I dette tilfellet legger vi spesielt merke til at intervallet for karakteren «C» er dobbelt så stort som de øvrige karakterintervallene.

Tabell 1. Karakterskala benyttet ved NTNU

A	B	C	D	E	F
100 - 90	89 - 80	79 - 60	59 - 50	49 - 40	39 - 0

Denne modellen er kriteriebasert i den grad vurderingen ikke skjer med referanse til andre kandidaters prestasjoner. I mange praktiske settinger finner man likevel hybridvarianter av relativ og absolutt vurdering.

Grading Model 3: Grades reflecting patterns of achievement (Sadler, 2005)

Denne modellen er mindre kjent og lite benyttet sammenlignet med den foregående modellen. Erfarne sensorer vil imidlertid ha lagt merke til at prestasjoner varier, for eksempel i ulike deler av en eksamen. Dette fører noen ganger til at kandidater består med ingen eller marginal innsikt i deler av et emne. Det gis derfor argumenter for å finne prosedyrer for beregning av samlet score som ivaretar minimumskrav til faglig innsikt i ulike deler av et emne.

Grading Model 4: Specified qualitative criteria or attributes (Sadler, 2005)

De fleste institusjoner benytter kvalitative kriterier i såkalte *mission* eller *policy statements*, som gjelder hele institusjonen, eller det kan være kompetanserelaterte kriterier knyttet til konkret emne eller oppgave, for eksempel krav til analytisk tenking, språkbruk og flyt i skriftlig framstilling. På overordnet nivå gir EUs (EQF) kvalifikasjonsrammeverk en rekke eksempler, spesielt knyttet til kategorien *Competence* (European Commision, 2008). I mange tilfeller blir ikke slike kvalitative kriterier benyttet ved vurdering av enkeltoppgaver, men inngår ofte som del av en totalvurdering av en oppgave eller besvarelse. Denne modellen har mindre relevans i forhold til denne artikkelen, og blir derfor ikke undersøkt nærmere i det følgende.

Problemstillinger og metode

Data til denne artikkelen er hentet fra to emner – ett i tredje studieår og ett i fjerde studieår, begge på 7,5 studiepoeng. Vurdering og karaktersetting ble i begge tilfeller gjennomført av respektive faglærere og ekstern sensor, uavhengig av hverandre. Data generert underveis i vurderingsarbeidet danner grunnlaget for analysen, fra råscorer til endelig karakter. Poengskjema for sensor og faglærer ble i begge tilfeller gjort tilgjengelig, og vurderingen ble gjennomført i forhold til en skala fra 0–100 med konvertering til bokstavkarakterer ut fra gitte karaktergrenser, som vist i Tabell 1.

Denne undersøkelsen representerer ingen statistisk test på effekten av ekstern sensor eller av ulike realiseringer av *kriteriebasert vurdering*. En slik test ville kreve tilgang til data fra langt flere enn de to eksamenene i denne undersøkelsen. Undersøkelsen er derfor å regne som en avgrenset kvantitativ kasusstudie innen to tilfeldig valgte emner. Hensikten er derfor ikke å generalisere, men prinsipielt å undersøke effekten av ulike tolkninger av kriteriebasert

vurdering. De to emnene vi undersøker, blir i det følgende omtalt som *Emne X* og *Emne Y*. Metodisk gjennomføres denne undersøkelsen som en kasusstudie, definert som «an empirical enquiry that investigates a contemporary phenomenon within its real-life context» (Yin, 1994, s. 13). Perspektivet er beskrivende, utforskende og analyserende, og er som tidligere nevnt motivert av en empirisk studie av Sadler (2005) der han oppsummerer fire definisjoner av fenomenet *kriteriebasert vurdering*. Problemstillingene i denne artikkelen er som følger:

- Hvordan foregikk vurdering og karaktersetting i de to emnene vi undersøker?
- Hvilken betydning har *ekstern sensor* for karakterfastsettingen i emnene?
- Hvilken forståelse av fenomenet *kriteriebasert vurdering* blir benyttet i de to tilfellene?
- Hvordan påvirker ulike tolkninger av *kriteriebasert vurdering* karakterfordelingene?

Datagrunnlag og analyse

Emne X og Emne Y ble undervist ved samme institutt våren 2011. For Emne X var det totalt 81 besvarelser og for Emne Y 30 besvarelser. Kun data for kandidater som møtte til eksamen er tatt med i analysen. I begge emnene vurderte faglærer og sensor besvarelsene uavhengig av hverandre med utgangspunkt i faglærers løsningsforslag. En skala fra 0–100 ble benyttet ved vurdering av hver deloppgave. En samlet score ble deretter beregnet med utgangspunkt i scorene for hver deloppgave. Eksamensoppgavene var fordelt på følgende vis i de to tilfellene:

Tabell 2. Emne X: Oversikt over oppgaver og deloppgaver

Oppgaver	1	2	3
Deloppgaver	a, b, c, d, e	a, b, c, d, e, f, g, h	a, b, c, d

Emne X besto av tre ulike deler (1, 2 og 3), og hver del ble undervist av tre ulike faglærere som også tok ansvar for respektive oppgaver til eksamen, mens det til eksamen var én felles sensor for alle tre delene som utgjorde kandidatenes besvarelse til eksamen. Emne Y besto av to ulike emneområder med hver sin faglærer, men med én felles sensor for begge delene:

Tabell 3. Emne Y: Forholdet mellom oppgaver og deloppgaver

Oppgaver	1	2
Deloppgaver	a, b, c, d, e, f, g	a, b, c, d, e, f, g

Tellende poengscore til eksamen ble beregnet på litt ulike vis i de to tilfellene. For Emne X ble faglærers totalscore oppjustert med 10 poeng og sensors score oppjustert med 15 poeng. Gjennomsnittet av disse ble brukt som gjeldende totalscore. For Emne Y ble det ikke foretatt noen justering. Gjennomsnittet av faglærers og sensors vurdering ble brukt som gjeldende totalscore. Eksemplene viser at faglærer og sensor i de to emnene benytter ulike

algoritmer for å fastsette endelig score. Begrunnelsen var i vårt tilfelle antatt stor arbeidsmengde, men det gir uansett ingen uttømmende forklaring på hvorfor oppjusteringen ble på henholdsvis 10 og 15 poeng. Et annet mulig motiv kunne være faglærers og sensors engasjement for en bedre karakterfordeling, men dette ble ikke kommentert i sensurmøtet. Spørsmål knyttet til *pålitelighet* i vurderingen står sentralt i undersøkelsen, og vi beregner derfor gjennomsnittsverdier og korrelasjon for å dokumentere sammenfall og avvik i vurderingene. Til slutt undersøker vi effekten av ulike tolkninger av *kriteriebasert vurdering*, spesielt i forhold til vurderingsmodell 3 som beskrevet foran.

Tabell 4. Emne X: Totalscorens gjennomsnitt og standardavvik

	Snitt	Standardavvik
Faglærer, første vurdering	44,4	19,5
Sensor, første vurdering	45,9	20,3
Faglærer og sensor, snitt	45,1	19,7
Faglærer og sensor justert	57,6	19,7

Tabell 4 viser at faglærer og sensor har vurdert besvarelsene svært likt, men at faglærer i dette tilfelle gjennomsnittlig har vært litt strengere enn sensor. Ved å undersøke gjennomsnittsscore for hver av de tre oppgavene, ser vi at vurderingene avviker i litt større grad (Tabell 5). Her ser vi at faglærer og sensor har vurdert oppgave 1 omtrent likt, mens sensor har vært klart snillere for oppgave 2 og faglærer har vært noe snillere enn sensor for oppgave 3.

Tabell 5. Emne X: Gjennomsnittlig score på enkeltoppgavene uten justeringer

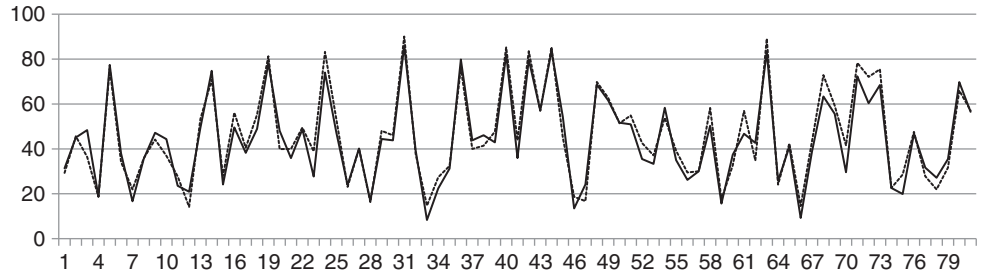
	Oppgave 1	Oppgave 2	Oppgave 3	Total score
Faglærer	47,9	43,3	41,9	44,4
Sensor	48,9	50,8	37,9	45,9

Så langt har vi sett at snittscoren for alle studenter i Emne X er nær den samme for faglærer og sensor, men at forskjellen er større for noen av oppgavene. Vi undersøker etter dette i hvilken grad faglærer og sensor er enig i vurderingen av den enkelte student, altså i hvilken grad scorene korrelerer. Tabell 6 viser korrelasjonskoeffisienter på oppgavenivå.

Tabell 6. Emne X: Korrelasjoner mellom faglærers og sensors vurdering på enkeltoppgaver og totalt

	Oppgave 1	Oppgave 2	Oppgave 3	Total score
Korrelasjon	0,92	0,95	0,95	0,97

Ut fra Tabell 6 ser vi at korrelasjonene på oppgavenivå generelt er høy. Det ser derfor ut til at faglærer og sensor i stor grad har hatt felles syn på hvilke besvarelser som er gode og mindre gode. Dette bekreftes gjennom den grafiske framstillingen i Figur 1, der faglærernes vurdering for oppgave 1, 2 og 3 samlet er sett i lys av sensors vurdering av de samme tre oppgavene. Vi legger også her merke til at mange kandidater faller under strykgrensen på 40 poeng.



Figur 1. Emne X: Grafen viser råscorer for de tre faglærernes vurdering av oppgave 1, 2 og 3 (stiplet linje) og sensors vurdering av de samme oppgavene (heltrukket linje). Y-aksen viser score fra 0–100, X-aksen referer til hver enkelt student, n = 81.

Vi undersøker så sammenhengen mellom opprinnelige scorer med karakterfordeling, samt justert karaktersnitt, som omtalt foran, og endelig karakterfordeling i Emne X. I Tabell 7 ser vi at oppjusteringen på 12,5 prosentpoeng i forhold til opprinnelig score har en dramatisk effekt på karakterfordelingen. Strykprosenten blir for eksempel halvert på denne måten.

Tabell 7. Emne X: Prosentvis karakterfordeling ut fra første vurdering og justert snitt

	A	B	C	D	E	F
Faglærer, første vurdering	0	4,9	16,0	9,9	24,7	**44,4**
Sensor, første vurdering	1,2	7,4	12,3	14,8	22,2	**42,0**
Justert snitt (tellende)	9,9	8,6	19,8	24,7	16,0	**21,0**

Dersom vi så benytter en algoritme for karaktersetting med utgangspunkt i Modell 3 (med krav om bestått på alle oppgavene), blir fordelingen annerledes enn verdiene i Tabell 7. Vi holder poengjusteringen utenfor, slik at sammenligningen knyttes til karakterfordelinger basert på faglærers og sensors opprinnelige vurdering (se Tabell 7). Ut fra Tabell 8 ser vi at karakterfordelingen ikke endrer seg for de beste kandidatene, men endrer seg mye for de svakere prestasjonene. Dette virker logisk ettersom det ikke harmonerer å få færre enn 40 poeng på minst én av oppgavene og likevel oppnå en god sluttkarakter. Enkel regning viser at det er umulig å oppnå karakteren «A» eller «B» ved å prestere til stryk på én av oppgavene.

Tabell 8. Emne X: Prosentvis karakterfordeling med krav om bestått på alle tre oppgaver (uten justeringer)

	A	B	C	D	E	F
Faglærer, ekstrakrav	0,0	4,9	16,0	7,4	6,2	**65,4**
Sensor, ekstrakrav	1,2	7,4	12,3	8,6	1,2	**69,1**

Vi gjennomfører nå tilsvarende analyse for Emne Y for om mulig å spore likheter og forskjeller mellom de to emnene. Det ble også benyttet ekstern sensor i begge tilfeller.

Tabell 9. Emne Y: Totalscorens gjennomsnitt og standardavvik

	Snitt	Standardavvik
Faglærer, første vurdering	59,50	16,14
Sensor, første vurdering	48,36	18,86
Faglærer og sensor (reelt snitt)	53,93	17,35

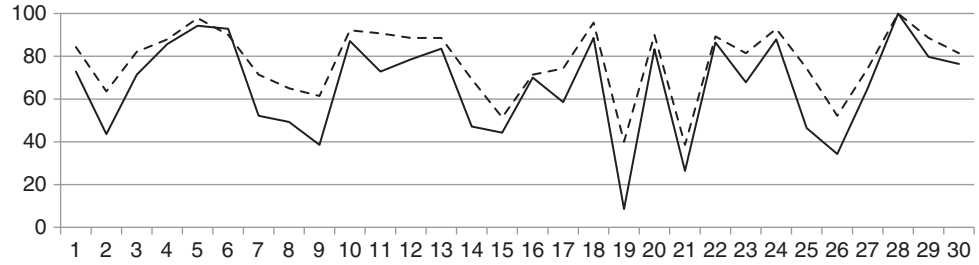
Vi ser at det i dette tilfellet er stor forskjell på faglærers og sensors vurdering. Sensor har i gjennomsnitt vurdert besvarelsene 11 prosentpoeng strengere enn faglærerne. Tabell 10 viser at resultatene er mye bedre på oppgave 1 sammenlignet med oppgave 2. I gjennomsnitt får kandidatene nesten dobbelt så høy score på den første sammenlignet med den andre oppgaven.

Tabell 10. Emne Y: Gjennomsnittlig score på enkeltoppgavene

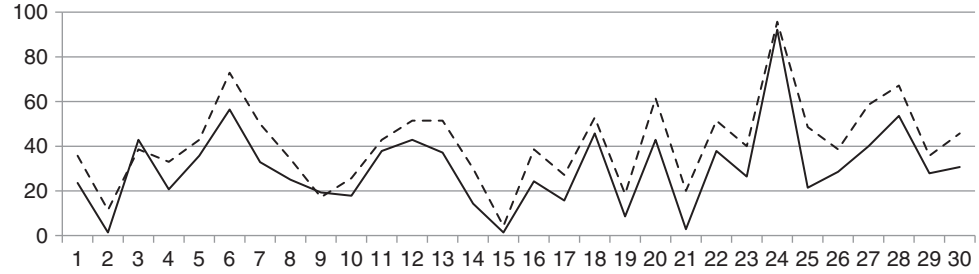
	Oppgave 1	Oppgave 2	Total score
Faglærer	77,6	41,4	59,5
Sensor	66,5	30,3	48,4

De grafiske framstillingene i Figur 2 og Figur 3 visualiserer poengfordelingen for de to oppgavene på individnivå. Vi legger merke til at det er en viss korrelasjon med hensyn til prestasjon for de to oppgavene, men at poenggivningen for oppgave 1 generelt ligger på et høyere nivå sammenlignet med oppgave 2. Dette kan ha sammenheng med at eksamens del 2 adresserer temaer som i stor grad var nye for studentene.

Sammenligner vi poeng gitt av faglærer og sensor for Emne Y, er forskjellene i noen tilfeller betydelige. Vi har derfor interesse av å undersøke korrelasjonen mellom faglærers og sensors vurdering i Emne Y. I Tabell 11 ser vi at korrelasjonene er høye for begge oppgavene. Dette er ikke overraskende ettersom korrelasjonstallet måler i hvilken grad faglærer og sensor er enige i hvilke besvarelser som er bedre eller dårligere enn gjennomsnittet. Ved å sjekke korrelasjon på deloppgavenivå finner vi moderat høye korrelasjoner, med unntak av to tilfeller der korrelasjonen er lav: 0,28 for oppgave 1a og 0,69 for oppgave 2a.



Figur 2. Emne Y: Poengfordeling for oppgave 1 (faglærer i stiplet og sensor i heltrukket linje). Y-aksen markerer score fra 0–100, X-aksen refererer til hver enkelt student, n = 30.



Figur 3. Emne Y: Poengfordeling for oppgave 2 (faglærer i stiplet og sensor i heltrukket linje). Y-aksen markerer score fra 0–100, X-aksen refererer til hver enkelt student, n = 30.

Hovedforskjellen mellom faglærers og sensors vurdering i Emne Y ligger likevel ikke i hvilke studenter som presterer best eller dårligst, men på det helhetlige nivået, som visualisert i Figur 2 og Figur 3.

Tabell 11. Emne X: Korrelasjoner mellom faglærers og sensors vurdering på enkeltoppgaver og totalt

	Oppgave 1	Oppgave 2	Total score
Korrelasjon	0,96	0,94	0,96

For Emne Y resulterte faglærers og sensors vurdering i en karakterfordeling som vist i Tabell 12, der også tellende karakterfordeling basert på snittet av vurderingene er tatt med. Vi ser at det i dette tilfellet er betydelig forskjell på karakterfordelingene, noe som ikke overrasker ut fra forskjeller i gjennomsnittsscore. Interessant er det også å se at både faglærer og sensor har gitt like stor prosentandel av karakteren A. Dette kan ha sammenheng med at partene opererte med et sett felles kriterier forstått som riktig (eller nesten riktig) svar på de to oppgavene.

I Tabell 13 ser vi karakterfordelingen vi får når vi legger kravet om bestått på begge oppgaver. Vi ser at dette kravet ikke gir noen forskjell i den prosentvise fordelingen på

Tabell 12. Emne Y: Prosentvis karakterfordeling ved første vurdering og tellende snitt

	A	B	C	D	E	F
Faglærer, første vurdering	3,3	6,7	53,3	10,0	10,0	**16,7**
Sensor, første vurdering	3,3	0,0	26,7	23,3	13,3	**33,3**
Snitt (tellende)	3,3	3,3	26,7	33,3	23,3	**20,0**

karakterene A og B, siden dette er umulig. Når vi sammenligner prosentvis fordeling ved første vurdering i Tabell 12 med fordelingen i Tabell 13, ser vi store forskjeller for karakterene C, D og E. Spesielt iøynefallende er en strykprosent på 73,3 for sensors vurdering med krav om bestått på begge oppgavene. Ved denne eksamensprøven besto mange kandidater ved å gjøre det bra på den ene oppgaven, mens resultatet framstår som katastrofalt med det ekstrakravet vi benytter.

Tabell 13. Emne Y: Prosentvis karakterfordeling med krav om bestått på begge oppgavene (uten justeringer)

	A	B	C	D	E	F
Faglærer, ekstrakrav	3,3	6,7	40,0	0,0	0,0	**50,0**
Sensor, ekstrakrav	3,3	0,0	16,7	6,7	0,0	**73,3**

Diskusjon

I våre to tilfeller ble ikke *læringsmål* benyttet ved planlegging og gjennomføring av eksamen, men det faller naturlig å tro at faglærere dro veksler på dette som en implisitt, underliggende forståelse av faglige krav. *Kriterier* (hva skal vurderes) og *standarder* (som viser til nivå) er de facto nødvendige betingelser i all prestasjonsvurdering – det egentlige spørsmålet er bare om disse variablene blir benyttet implisitt eller eksplisitt. Vurderingen ble i begge tilfeller gjennomført av faglærer og sensor hver for seg med utgangspunkt i faglærers løsningsforslag uten forutgående hjelpemidler, for eksempel ved bruk av vurderingsrubrikker (Stevens & Levi, 2004). Råscorene for hver deloppgave ble slått sammen til tellende score, etter prinsipp fra *Grading Model 2: Overall achievement as measured by score totals* (Sadler, 2005).

Bruken av en fingradert skala fra 0–100 gir ytre sett inntrykk av høy grad av presisjon, men ulikt fysiske måleinstrumenter skjer prestasjonsvurdering uten felles, unik referanse, slik tilfellet er ved temperaturmål, høyde og vekt. Bruken av ekstern sensor har vært begrunnet med dette som bakteppe. I denne undersøkelsen har vi sett at betydningen av ekstern sensor for *Emne X* var svært liten i forhold til de oppjusteringene som ble foretatt av opprinnelige score. For dette emnet var også effekten av ekstrakrav om bestått på alle oppgavene hver for seg langt større enn betydningen av ekstern sensor. I *Emne Y* var det derimot store forskjeller mellom faglærers og sensors vurdering (gjennomsnittlig forskjell i totalscore på 11 poeng). Effekten av å ta snittet av begge vurderingene sammenlignet med bare å bruke faglærers vurdering, tilsvarer en justering på fem poeng eller ekstrakrav om bestått på begge oppgavene.

For Emne X har vi allerede kommentert en betydelig oppjustering av poengscorer med stor arbeidsmengde som begrunnelse. Samtidig vet vi at «. . . det over tid og for et stort antall kandidater, er forventet at fordelingen av bestått karakterer skal være i rimelig samsvar med den relative ECTS-skalaen» (Glasser, 2009, s. 3). Dersom dette etterleves, kan man spørre om verdien av ekstern sensor dersom resultatene uansett oppjusteres til antatt akseptabelt nivå. Når slike justeringer blir foretatt, har bruken av ekstern sensor liten verdi som korreksjon av *bias* (systematisk feil) i vurderingen, for eksempel ved at sensor generelt gir en strengere vurdering enn faglærer. Men som vi har sett for *Emne Y*, har bruken av sensor betydning for korreksjon av *error* (tilfeldig, ikke gjennomgående feil). Korrelasjonen mellom faglærers og sensors vurdering var i begge tilfelle stor. Tidvis betydelige avvik, slik vi har sett for Emne Y, tyder på at faglærer og sensor hadde ulik vurdering av besvarelsenes nivå. Dette er et velkjent fenomen, som også er vel dokumentert og omtalt i forskningslitteraturen (Sadler, 2012, s. 1).

Bruken av ekstern sensor har lange tradisjoner som kvalitetssikringstiltak i høyere utdanning, men blir nå benyttet sjeldnere enn før – ofte bare hvert fjerde år. Begrunnelsen for ekstern sensor har vært krav om uhildet og objektiv vurdering og et ønske om kalibrering av krav på tvers av emner og institusjoner. Denne læringsmodellen er kjent fra håndverkspraksis, som forholdet mellom mester og svenn, også omtalt som *communities of practice* i litteraturen (Wenger, 2002). Slike læringsfellesskap fungerer godt under relativ stabilitet; kunnskapen er taus eller implisitt og teorifri forstått som uttalt teori. I denne undersøkelsen har vi sett at vurderingspraksis forble upåvirket av utdanningsmyndighetenes rundskriv, men fulgte etablert praksis ut fra sensorenes egen tolkning av kriteriebasert vurdering med justeringer.

La oss etter dette diskutere mulige forbedringstiltak i våre emner med utgangspunkt i Sadlers definisjon av en *karakter*: «A grade is essentially a symbolic representation of the level of achievement attained by a student» (Sadler, 2009, s. 807). Sluttkarakteren skal speile kandidatens kompetanse ved prøvetidspunktet, upåvirket av andre forhold. Dette blir omtalt som karakterens integritet (*grade integrity*) (Sadler, 2009), og Sadler nevner fire forhold som kan true denne: Tilfeldig feil (1) og systematisk feil i vurdering (2), sammenblanding av hva som er objekt for vurderingen (3) og uavklart vurderingsprinsipp (4) (Sadler, 2009, s. 812–814). Punkt 1, 2 og 4 har spesiell relevans her, og punkt 4 forklarer Sadler på følgende vis:

> «*By a grading principle is meant a theoretically coherent, explicitly articulated set of ideas that forms a bridge between fundamental educational values (such as fairness and authenticity) and the techniques and methods that can be legitimately applied to raw assessment evidence to produce grades.*»
> *(Sadler, 2009, s. 814)*

Til tross for at korrelasjonen mellom faglærers og sensors poenggivning i begge emnene var høy, har vi – spesielt for Emne Y – dokumentasjon på *bias* (systematisk feil) og *error* (tilfeldig feil). Dette ble løst gjennom en omforent forståelse om endelig score (*consensus moderation*) (Sadler, 2012), uten forutgående eksplisitt *kalibrering* av krav til faglig nivå. Påliteligheten i vurderingen ble også, som vi har sett, utfordret gjennom metoden for poengberegning. Denne mulige feilkilden er av en annen art sammenlignet med punkt 1 og 2, som beskrevet

over. Her handler det ikke om feil eller unøyaktighet i vurderingsarbeidet, men om *policy* for vurdering.

Et tilleggskrav om bestått på alle oppgavene øker strykprosenten betydelig i Emne X og mer i Emne Y, der den ville økt med 30–40 % ved et slikt krav. Det betyr at for et gitt antall studenter ender resultatet med stryk eller bestått avhengig av vurderingsmåten som benyttes.

Man kan spørre i hvilken grad karakterene speiler samlet kompetanse, slik intensjonen ved *kriteriebasert vurdering* er. En grunn til at dette har blitt et tema, er det faktum at noen emner er et resultat av sammenslåing av tidligere mindre, separate emner med egen faglærer, litteratur, øvinger og eksamenspraksis. I enkelte tilfeller er det lite samarbeid mellom de involverte faglærerne, uten felles, integrerte øvings- og eksamensoppgaver, og det kan oppstå betydelige forskjeller med hensyn til faglige krav og nivå. Et mulig negativt resultat er at studenter fristes til å spekulere i systemet for å bestå, med store faglige «hull» som uønsket resultat. Et par eksempler vil illustrere dette: La oss si at én kandidat oppnådde følgende poengfordeling på de tre oppgavene i Emne X: (20 + 30 + 90) poeng = 140 poeng, med tellende score 140/3 = 47 poeng, som altså holder til bestått (E) til tross for at kandidaten har strøket med god margin på to av oppgavene. Gitt følgende hypotetiske poengfordeling på de to oppgavene i Emne Y: (95 + 30) poeng = 125 poeng, med tellende totalscore som følger: 125/2 poeng = 63 poeng, som holder til en «C» ifølge NTNUs konverteringsregler (Tabell 1). Etter karakterbeskrivelsen defineres denne prestasjonen på følgende vis: «Prestasjon som oppfyller læringsmålene på en god måte» (Glasser, 2009). Dette til tross for at kandidaten er langt fra å stå i forhold til halvparten av emnets læringsmål.

Ut fra et kvalitetssikringsståsted, vil foregående eksempler trolig vekke en viss undring, og kan tjene som innspill til institusjonens system for kvalitetssikring av utdanning der blant annet følgende spørsmål blir stilt: «Gir sensur og karaktersetting et riktig bilde av prestasjoner og læringsutbytte?» (Styret, 2012). Dersom vi ser bort fra den begrepsmessige forvirringen i spørsmålet rundt *læringsresultat* og *læringsbytte*, er dette et relevant spørsmål. Mens kriterier og standarder forvaltes av faglærer og sensor, relaterer spørsmål om *læringsutbytte* til eksaminandens opplevelse med sammenligning av status før og etter gjennomført emne. I Universitets- og høgskolerådets karakterbeskrivelser (2008) slås det fast at hvis «en prestasjon tilfredsstiller kriteriene for en karakter, skal man gi denne karakteren uavhengig av hvordan fordelingen av de øvrige karakterene i eksamenskullet er» (Glasser, 2008, s. 3). Problemet er at vurdering aldri kan ta utgangspunkt i *kriteriene for en karakter*, slik befalingen er (Glasser, 2008). Vurdering skjer alltid med utgangspunkt i *faglige* kriterier og standarder, og dernest må vurderingsordningen være slik at karakteruttrykket til sist speiler studentenes kompetanse på en tilfredsstillende måte. Et problem er også at de karakterbeskrivelsene som UHR anbefaler dels er formulert ut fra *normbasert*, dels ut fra *kriteriebasert* grunnlag (Gynnild, 2010).

Konklusjon

Bruken av kriteriebasert vurdering i høyere utdanning er nå et uomtvistelig krav på nasjonal basis, men denne undersøkelsen viser at operasjonaliseringen av dette prinsippet kan være en krevende øvelse å gjennomføre. I fravær av felles institusjonell praksis, er det opp til faglærer (og eventuelt sensor) å definere hva kriteriebasert vurdering betyr. Når alt kommer til alt, kan

man få inntrykk av at *karakterfordelingen* er en større bekymring enn det *kunnskapsgrunnlaget* karakterene forventes å representere. Dette er forståelig ut fra menneskelige og praktiske hensyn, så som frykten for høy strykprosent med påfølgende sensurklagesaker. Men ut fra et evalueringsteoretisk ståsted, og ut fra forestillingen om Norge som kunnskapsnasjon, virker situasjonen mer bekymringsfull. Sertifiserer vi kandidater med alvorlige kunnskapshull?

Både faglærer og sensor framsto i de to tilfellene som representanter for et kollektiv, som i praksis var ganske upåvirket av verbale karakterbeskrivelser, som uansett ikke bidrar til en tydeligere forståelse av hvordan kriteriebasert vurdering skal forstås. All vurdering har, i tillegg til praksis, også en prinsipiell og teoretisk side. Så lenge prinsippene ikke nedfelles i felles forståelige retningslinjer for praksis, kan man forvente store sprik på handlingsnivå.

Oppsummert, og med referanse til Sadlers empiriske studie av kriteriebasert vurdering, ser vi at læringsmål ikke har vært benyttet ved vurdering i vårt tilfelle. Vurderingen skjer ved bruk av en skala fra 0–100, der scorene i sin tur omsettes til karakterer. Denne undersøkelsen viser at med en annen definisjon, der endelig karakter fastsettes med krav om bestått i ulike deler av et emne, gir en dramatisk endret karakterfordeling. Institusjonens omregningstabell fra råscorer til karakterfordeling virker i tillegg sterkt inspirert av tanken om normalfordeling som prinsipp med karakteren «C» som en middels god prestasjon. Selv om den endelige karakterjusteringen i Emne X ble begrunnet med antatt stor arbeidsbelastning, kan man neppe utelukke at også ønsket om en lavere strykprosent ubevisst kan ha spilt inn.

Mens mye av diskusjonen knyttet til pålitelighet i vurdering har vært rettet mot ekstern sensor, viser denne undersøkelsen at stringens i *tolkningen* av kriteriebasert vurdering kan være vel så viktig for karakterene. Ettersom kriteriebasert vurdering er knesatt som et grunnleggende prinsipp for høyere utdanning, reiser dette en rekke spørsmål av prinsipiell og praktisk art. En spesiell utfordring ligger i at reformarbeid på dette feltet har vært sterkt administrativt drevet og svakt evalueringsfaglig forankret. Samtidig preges kulturbærerne innen de enkelte emner av laugstradisjoner med sterk indre lojalitet og et relativt fritt forhold til administrative krav. Denne undersøkelsen er prinsipiell og analytisk i sin tilnærming uten normative ambisjoner, men kanskje kan den legge grunnlag for dypere refleksjon om hva kriteriebasert vurdering er, eller kan, være. Hvilken praksis som blir valgt, er ikke alene et spørsmål om rett eller galt, men om valg og beslutninger, og et felles forstått rammeverk vil være til hjelp for alle parter.

Forfatteren takker student Christian Preben Bang for bistand til statistiske beregninger.

Referanser

Barr, R. B. & Tagg, J. (1995). From Teaching to Learning: A New Paradigm for Undergraduate Education. *Change*, *27*(6), 12–25.

Biggs, J. (1996). Enhancing teaching through constructive alignment. *Higher Education*, *32*(3), 347–364. doi: 10.1007/bf00138871.

Biggs, J. (2007). *Teaching for quality learning at university: what the student does*. Maidenhead: McGraw-Hill/ Society for Research into Higher Education & Open University Press.

Boud, D., Cohen, R. & Sampson, J. (1999). Peer Learning and Assessment. *Assessment & Evaluation in Higher Education*, *24*(4), 413–426.

Det norske universitetsråd (2000). Mål med mening: En utredning om etablering av en felles nasjonal karakterskala (pp. 74). Agder, Bergen, Oslo, Tromsø, Trondheim.

European Commision (2008). The European Qualifications Framework for Lifelong Learning (EQF) (Publication no. 10.2766/14352). From Office for the Official Publications of the European Communities http://ec.europa.eu/education/pub/pdf/general/eqf/broch_en.pdf

Glasser, R. (2008). *Generelle karakterbeskrivelser for UH-sektoren*. Oslo: Universitets- og høgskolerådet.

Glasser, R. (2009). *Tilleggsrapport for arbeidsgruppe for å se nærmere på UH-sektorens generelle karakterbeskrivelser*. Oslo: Universitets- og høgskolerådet.

Gynnild, V. (2010). Kriterier og skjønn i evaluering: En kasusstudie. *Uniped*, *33*(2), 6–24.

Kunnskapsdepartementet (2004). Retningslinjer for bruk av det nasjonale karaktersystemet. Retrieved 26.10.12 from http://www.regjeringen.no/nb/dep/kd/dok/andre/brev/utvalgte_brev/2004/retningslinjer-for-bruk-at-det-nasjonale.html?id=91189

Rust, C. (2002). The impact of assessment on student learning. *Active Learning in Higher Education*, *3*(3), 145–158.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175–194.

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

Sadler, D. R. (2010). Assessment in higher education. In P.Peterson, E.Baker & B.McGaw (Eds.): *The international encyclopedia of education* (Vol. 3, pp. 249–255). Oxford: Elsevier.

Sadler, D. R. (2012). Assuring academic achievement standards: from moderation to calibration. *Assessment in Education: Principles, Policy & Practice*, s. 1–15. doi: 10.1080/0969594x.2012.714742.

Stevens, D. & Levi, A. J. (2004). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback and promote student learning*. Sterling, Virginia: Stylus Publishing.

Styret (2012). NTNUs system for kvalitetssikring av utdanning. Retrieved 09.11.2012 from http://www.ntnu.no/utdanningskvalitet

The European Higher Education Area (1999). The Bologna Declaration of 19 June 1999. Retrieved from http://www.ond.vlaanderen.be/hogeronderwijs/bologna/documents/MDC/BOLOGNA_DECLARATION1.pdf

Wenger, E. (2002). *Cultivating communities of practice: a guide to managing knowledge*. Boston: Harvard Business School Press.

Yin, R. K. (1994). *Case study research: design and methods*. Thousand Oaks, Calif.: Sage.