Artikkelen er vitenskapelig vurdert av forskere utenfor redaksjonen

Lavt sensorsamsvar – kan det bedres?

Av Gunnar Bjølseth, Anton Havnes og Per Lauvås

Gunnar Bjølseth
Høgskolelektor ved Høgskolen i Oslo, Avdeling for sykepleierutdanning.
E-post: gunnar.bjolseth@su.hio.no

Anton Havnes
Professor ved Pedagogisk utviklingssenter, Høgskolen i Oslo.
E-post: Anton.havnes@hio.no

Per Lauvås
Pensjonert professor i pedagogikk. Arbeidet tidligere ved Høgskolen i Oslo, i Østfold og ved Pedagogisk forskningsinstitutt, Universitet i Oslo.
E-post: perlauvas@gmail.com

Sammendrag
Etter innføringen av Kvalitetsreformen er kravet om ekstern sensur redusert betydelig i høgre utdanning. Antallet sensorer er også redusert. Dermed forutsettes det at sensureringen holder høy kvalitet, også når den bare gjøres av en intern sensor. Et første krav er høyt sensorsamsvar, det vil si godt samsvar i vurdering og karaktersetting innenfor gruppen av sensorer. Men omfattende forskning viser stor variasjon og ofte dårligere sensorsamsvar enn det man forutsetter. Ved én profesjonsutdanning tydet noen klagesaker på lavt sensorsamsvar. Begrunnelser for gitte karakter var til dels sprikende og lite konsistent. Man besluttet å gjøre et krafttak for å kartlegge situasjonen og bedre kvaliteten i sensureringen. Etter ett år med kvalifisering av hele lærerpersonalet, var resultatene beskjedne. Det viste seg langt vanskeligere å endre sensureringspraksis blant lærerne enn forventet. Dette resultatet gir grunnlag for å stille spørsmål om hvordan man kan heve samsvaret innen et sensorkorps opp til et nivå som er akseptabelt når antallet sensorer er lite.

Introduksjon

Denne artikkelen er ikke resultatet av en kontrollert eksperimentell studie, men dokumentasjon fra et relativt omfattende utviklingsarbeid ved én avdeling.

Det hele startet egentlig med en klagesak. En student som hadde fått karakteren B, klaget og opplevde å bli strøket av klagekommisjonen. Ved avdelingen bestemte man seg for å gjøre noe med situasjonen. Avdelingens ledelse satte i gang et ambisiøst prosjekt der kvaliteten i sensureringen skulle høynes, både med hensyn til karaktersetting og begrunnelse for gitte karakterer. Seks fagdager fordelt over ett år for hele personalet ble viet sensurering. Styringsgruppe ble oppnevnt og to høgskolepedagoger påtok seg oppgaven å gjennomføre utviklingstiltaket. Optimismen var stor. Målet var å utvikle en mer uniform vurdering og en større bevissthet om vurderingskriterier. Men resultatet ble magert. Det viste seg vanskeligere enn antatt å påvirke sensorer og å få kolleger til å utvikle en felles vurderingspraksis.

Sensorsamsvar i høgre utdanning

Det er en vanlig oppfatning blant de som er involvert i sensurering av studentarbeider at samsvaret sensorer imellom er godt. Etter at sensorer har lest besvarelsene man har fått, byr ikke samtalen med annen sensor på store overraskelser. Men stemmer denne oppfatningen med virkeligheten?

Samsvaret mellom sensorene dreier seg både om lik rangering av besvarelsene innbyrdes så vel som anvendelse av karakterskalaen, dvs. strengheten i vurderingen av de enkelte besvarelsene. Vi kan altså tenke oss at selv om to sensorer er enige i at Karis besvarelse er klart bedre enn Mettes, kan den ene gi Kari A og Mette C mens den andre gir Kari B og Mette D. Det er den første formen som er best undersøkt, selv om det finnes unntak (se f.eks. Torrance, 1995, s. 97–98).

Samsvaret er selvsagt lavest ved såkalte «ekspressive prøver» – besvarelser som ikke kan klassifiseres som riktig eller galt, dvs. essays, rapporter, avhandlinger og mange typer eksamensbesvarelser. Ved flervalgstester eller andre prøver der retting nærmest kan automatiseres, er sensorsamsvaret høyt. Der er andre problemstillinger viktigere.

I en norsk studie analyserte Raaheim (2000) sensureringen av essaybesvarelser i psykologi. Han fikk karakterforslagene til de ordinære sensorene og fikk syv nye «sensorer» til å vurdere de samme besvarelsene. Sensorsamsvaret var relativt godt (gjennomsnittlig korrelasjon mellom sensorene, r=.75). Men selv med en såpass høy korrelasjon, var det like fullt betydelig variasjon mellom sensorene. Karakterene på samme besvarelse kunne i verste fall variere fra 2,4 til stryk. Variasjonsbredden i karakterer mellom de syv «sensorene» var en hel karakter (dvs. fra 2,0 til 3,0) eller mer for nesten halvparten av besvarelsene (45,2 %). Raaheim engasjerte deretter det han betegnet som «supersensorer» – noen av de mest erfarne professorene i psykologi i landet – til å vurdere de besvarelsene der det var størst sprik. Også de vurderte besvarelsene ulikt, selv når de fikk tilgang til de karakterforslagene som var gitt. En gjennomsnittlig høy korrelasjon mellom sensorene innebar slett ikke noen konsistens i vurderingen av enkeltbesvarelser. I noen tilfeller ville karakter kunne variere fra 2,7 til stryk også når man satte sammen tenkte «kommisjoner». Så selv med et akseptabelt gjennomsnittlig samsvar, vil enkeltstudenter tjene eller tape avhengig av hvem sensor er. Selv når samsvaret er meget godt (.85–.90), finner Black det riktig å vise til at «[...] the errors in pupils' scores that are implied may mean a significant proportion are given the wrong grade» (Black, 1998, s. 41).

I hennes gjennomgang av forskningsresultatene på området, konkluderer Bloxham (2009) med at tiltroen til at eksamensbesvarelser kan vurderes og karaktersettes nøyaktig og likt ikke har noen sterk basis i de empiriske studier som foreligger. Denne tiltroen holdes likevel ved like, som antydet av Price (2005), fordi det er for ubehagelig å ta innover seg at holdbarheten i våre sensurordninger er svak. Studenter flest ser også ut til å beholde tiltroen til sensureringen, selv når de kjenner til at sensorene vektlegger ulikt (Crook, Gross & Dymott, 2006). Bloxham (2009) viser til de vanligste årsakene til dårlig sensorsamsvar:

- Komplekse læringsresultater kan ikke reduseres til noe som er enkelt å vurdere endimensjonalt og liketil. Samsvaret er størst når vurderingen gjelder enkle forhold og dårligere

når det er mer sentrale – og viktigere – aspekter som vurderes. Men selv på de enkle forholdene kan det være stor variasjon. Selv i de tilfellene hvor kriteriene for vurdering er gjort eksplisitte, risikerer studentene at ulike sensorer vurderer deres arbeider ulikt.

- Vurderingen må i stor grad baseres på faglig skjønn, og skjønnet utvikles over tid som ett aspekt av sosialiseringen inn i fag-, disiplin- og profesjonskonteksten, og utgjør de individuelle «locally constructed and tacit standards» (ibid., s. 212) som legges til grunn ved vurderingen. Sensorene betrakter seg etter hvert som eksperter i sensurering. Deres vurdering blir mer intuitiv enn systematisk og bevisst, og noen studier tyder på at erfarne sensorer ikke er mer konsistente i sine vurderinger enn mer uerfarne, bl.a. fordi de legger mindre vekt på vedtatte vurderingskriterier, -prosedyrer og andre støttetiltak i sensureringen. De baserer seg på sitt eget skjønn.
- Kombinasjonen av intuitiv og privat preget vurdering, mangelfull profesjonalitet i vurderingen og ingen anledninger til å kalibrere eget skjønn med andres svekker sensorsamsvaret.
- Det ser ikke ut til at det hjelper særlig å lage eksplisitte vurderingskriterier, vurderingsskjemaer («marking grids») eller lignende. Anvendelsen av slike hjelpemidler forutsetter at sensorene anerkjenner dem og akter å følge dem. Det ser ikke ut til å komme av seg selv, særlig når etableringen av slikt har et byråkratisk tilsnitt. Derimot ser det ut til at felles standarder utvikles innenfor velfungerende lærergrupper (se f.eks. Elwood & Klenowski, 2002).

Bloxham etterlyser også kunnskap om effekten av ulike former for «moderation» av sensureringen, dvs. prosesser som skal sikre «valid, fair and reliable» sensurering og at vurderingskriteriene brukes systematisk. Slik «oversensurering» foregår litt mer variert i Storbritannia enn i vårt land, men også hos oss er utviklingen i gang, f.eks. med «kalibreringsmøter» ved starten av sensureringen. Effekten av slike tiltak er lite undersøkt empirisk, og resultatene fra de studiene som finnes gir et uklart bilde. Det ser f.eks. ut til at «second or double marking» har begrenset verdi.

Samlet sett ser det ut til at det i Storbritannia finnes belegg for å hevde at sensorsamsvaret i mange tilfeller representerer et problem (Harlen, 2005). Knight (2001) bruker sterke ord – «ramshackle» skulle vel kunne oversettes med «falleferdig» – mens andre (Murphy, 2006; Elton & Johnson, 2002; Knight & Yorke, 2003) er litt mildere i språkbruken. Det vi for all del vil unngå, er likevel ikke helt utenkelig, nemlig at «[...] assessment by different examiners produces marks with considerable variability such that in determination of these marks the part played by the examiner can be greater than that of the performance of the examinee» (Pieron, i Light & Cox, 2007, s. 300).

Betydningen av vurderingskriterier og -standarder

I senere tid er formulering av læringsmål og kriterier blitt vanlig. De tjener både til å styre studenters læring og gi grunnlag for vurdering. De skal skape motivasjon og retning i læringen, og «make assessment more amenable to moderation and standardisation between markers, thereby enhancing reliability» (Ecclestone, 2001, s. 302). De skal vise hva som vektleg-

ges og som representerer gyldig kompetanse innen et kunnskapsområde, en disiplin eller et profesjonsfelt (Kvale, 1996).

Men en sak er å definere og klargjøre vurderingskriterier, noe annet er å bruke dem likt i praksis. Kriterier er – på den ene siden – allmenne og ment å være gyldige i et spekter av ulike oppgavetemaer, samtidig som de i praktisk anvendelse er spesifikt knyttet til kontekster, situasjoner, temaer og kasuser. De er «subject to multiple interpretations by both individual staff members and students» (Rust et al., 2003, s. 327). «Statements of expected standards, curriculum objectives or learning outcomes are generally insufficient to convey the richness of meaning that is wrapped up in them» (Yorke, 2003, s. 480).

Sadler (1987, 2005) understreker at fastsettingen av kriterier ikke er tilstrekkelig for at disse kriteriene faktisk slår inn i vurderingen. For det første kan studenter tilfredsstille et kriterium på ulikt nivå, f.eks. at en besvarelse kan innfri et kriterium godt eller dårlig. O'Donovan et al. (2001) forsøkte å løse dette ved hjelp av et såkalt "criterion-specific assessment grid», der det for hvert kriterium er definert flere nivåer (A–E). Nivåene gir imidlertid ikke mening «without the benefit of explanation, exemplars and the opportunity for discussion» (ibid., s. 83). Gibbs & Dunbar-Goddet er enda mer kritiske til generelle kriteriebeskrivelser: «Clarity was achieved implicitly as a by-product of many cycles of experience of the same kind of 'performance of understanding' within a community of practice, rather than by explicitness» (2007, s. 25). Felles vurderingspraksis ser altså ut til å være avhengig av felles diskusjon om kriteriene satt opp mot anvendelsen av dem på konkrete besvarelser.

Hvordan kan sensorsamsvaret forbedres?

Resultatene fra flere studier (Nystrand et al., 1993; leMahieu et al., 1995; Supowitz et al., 1997; Baume and Yorke, 2002) tyder på at opplæring («training») er et viktig, kanskje til og med nødvendig tiltak for å sikre at sensorer vurderer samme besvarelse på nær samme måte. Med «opplæring» refereres det til ganske forskjellige tiltak, men først og fremst kollegiale diskusjoner om vurderingen av konkrete besvarelser. Andre studier tyder på at opplæring ikke har noen særlig betydning. Pitts et al. (2002) fant f.eks. bare en beskjeden økning i sensorsamsvaret ved vurdering av mapper etter opplæring.

Andre (f.eks. Elwood & Klenowski, 2002; Goos & Moni, 2001; Baird et al., 2004) viser til betydningen av å innlemme studenter i vurderingspraksis og ikke bare la dem være objekter for andres vurdering.

> *«In pursuing the goals of effective assessment for learning, it is fundamental for teachers and students to grow in a community of practice where nothing in the assessment process is hidden and all hurdles are understood clearly and explicitly.»*
>
> *(Elwood & Klenowski, 2002, s. 255)*

Selv om det er noe uklart hvor stor betydning opplæring av sensorer og andre sensureringstiltak har, ser det ut til å være relativt stor enighet om at kollektivitet er viktig. Lærere/sensorer må samtale om vurderingen når den foregår; på den måten kan det utvikles en felles forståelse og felles praksis. Studentene bør også innlemmes i vurderingspraksis og samtalene om læringsmål og kriterier.

Mens det tidligere var flere sensorer i Norge og krav om ekstern sensor ved alle eksamenselementer, er både antallet sensorer og andelen av ekstern sensur nå redusert. Med flere sensorer blir i hvert fall mulighetene for dramatiske utslag av lavt sensorsamsvar redusert. Et system med flere sensorer jevner ut forskjellene mellom ulike sensorers vurdering av et gitt studentarbeid. Det forhold at fagpersoner innen ett og samme fagmiljø har til dels svært ulik vektlegging av hva som kjennetegner faglig kvalitet, er også en utfordring for fagmiljøet. Arbeidet med å kartlegge bakgrunnen for ulik vurdering og undersøke mulighetene for å øke samsvaret i bedømmingen av studentarbeider, er vesentlig både med hensyn til studentene og det faglige kollegiet.

Det teoretiske grunnlaget vi hadde for å gå i gang med et relativt omfattende utviklingsprosjekt, er nå trukket frem i sterkt konsentrert form. Vi måtte gå ut fra at sensorsamsvaret innen lærergruppen var relativt lavt, slik det er dokumentert andre steder i høgre utdanning, gitt karakteristika ved eksamensformen og de foreliggende indikasjoner på problemer i sensureringsarbeidet. Vi regnet med at det beste tiltaket var å sette i gang samtaler omkring sensurering så nær opp til den virkelige sensursituasjonen som mulig. Selv om litteraturen gir et noe variert bilde på dette punktet, mente vi at tiltaket burde gi et vesentlig løft i sensureringsarbeidet.

Bedre sensorsamsvar – et utviklingsprosjekt: Metode

Avdelingen besluttet å gjennomføre et krafttak for å utvikle en mest mulig felles forståelse og sensureringspraksis. Seks fagdager ble plottet inn på kalenderen over ett år. Styringsgruppe ble etablert og prosjektet satt i gang. Høgskolepedagogene (2.- og 3.-forfatter) fikk ansvar for opplegg og gjennomføring av fagdagene i samarbeid med styringsgruppen.

Det må nevnes at den aktuelle avdelingen nylig hadde startet med mappeeksamen som besto av hjemmeeksamen basert på studieoppgaver fra tidligere semester. Bedømmelsen av disse bød på nye utfordringer sammenlignet med ordinære eksamensformer.

Før hver fagdag ble det plukket ut et antall besvarelser som ble distribuert til deltakerne med litt ulikt oppdrag fra gang til gang. Vanligst skulle de lese besvarelsen, sette karakter og begrunne den. Styringsgruppen valgte ut besvarelser som dekket alle årene på bachelor og hele karakterspennet (A–F). Totalt ble 25 besvarelser anvendt. Karakteren som opprinnelig var satt på besvarelsene ble ikke oppgitt.

Arbeidsseminarene fulgte stort sett samme mønster. Vi presenterte først resultatene fra forrige samling og forskning som ble ansett relevant for dagens tema. Hovedsaken var at deltakerne i grupper med utgangspunkt i deres individuelle vurdering av besvarelsene skulle diskutere seg frem til en felles karakter og begrunnelse. Karakterforslagene og begrunnelsene ble samlet, både fra enkeltpersoner og gruppene som hadde arbeidet sammen på fagdagen. Dette materialet (som ikke ble komplett pga. alt det sydende liv som oppsto på seminarene!) ble analysert, og resultatene ble fôret tilbake til deltakerne på neste fagdag. Selv om fagdagene ble gjort obligatoriske av ledelsen, varierte oppmøtet. Noe «komplett datasett» finnes dermed ikke. Materialet gir likevel et rikt bilde av forskjellene i vurderingspraksis ved utdanningen og grunnlaget for det svake sensorsamsvaret.

I hele prosessen ble det lagt vekt på at det var spredningen i karakter mellom alle som

hadde lest besvarelsen som var av interesse, ikke å etterprøve den karakter som var gitt på angjeldende oppgave ved ordinær sensur eller å se på den som «korrekt» karakter.

I tillegg til at dokumentasjonen la grunnlaget for planleggingen av neste arbeidsseminar, ligger den også til grunn for denne artikkelen. Metoden har altså vært utviklingsrettet ved at analysene har bidratt inn i en lokal utvikling av praksis. Det har vært en kartleggingsstudie ved at vi systematisk har undersøkt vurdering av ulike oppgavetyper, bruk av ulike kriterieesett, vurdering av oppgaver som har fått ulik karakter og ved at lærere har gått sammen om vurdering innad og på tvers av team, faglig spesialisering, studienivå og erfaring. Både kvantitative og kvalitative data ble brukt i analysen underveis og etter gjennomføringen av utviklingsprosjektet. Vi rekker ikke å ta opp alle variablene i denne artikkelen, men vil ha søkelyset på sensorsamsvar i karaktersetting, bruken av kriterier og begrunnelser av karakter.

Resultater

Sensorsamsvar

Vi ser først på utviklingen av sensorsamsvar. På det første arbeidsseminaret kom deltakerne med karakterforslag og skriftlig begrunnelse for tre besvarelser. I grupper (som altså hadde lest den samme besvarelsen) diskuterte de sine forhåndssatte karakterer og begrunnelser, og kom deretter frem til et felles karakterforslag og begrunnelse. Tabell 1 viser spredningen i de individuelle karakterforslagene som deltakerne hadde med seg og som enda ikke var diskutert i gruppene, her presentert som avviket fra opprinnelig karakter.

Tabell 1. Avvik mellom ordinær sensur og forslag fra individuelle lærere, første seminar (sept 06)

	Samme karakter	Avvik = 1 karakter	Avvik = 2 karakterer	Avvik = 3 karakterer	Avvik ≥ 4 karakterer	N
1. Studieår besvarelse gitt A	6%	6%	6%	50%	32%	16
2.Studieår besvarelse gitt C	30%	40%	30%	–	–	10
3. Studieår besvarelse gitt E	15%	69%	15%	–	–	13
Total	15%	36%	15%	21%	13%	39

Halvparten av de individuelle karakterforslagene (49 %) avvek to trinn eller mer fra den karakteren som var gitt ved ordinær sensur. De største avvikene fant vi der besvarelsen hadde blitt belønnet med A ved ordinær sensur. Det fantes til og med eksempel på at en deltaker på seminaret kom med forslag på strykkarakter for denne besvarelsen. Det verst tenkelige scenarioet materialiserte seg her: Dersom en kandidat kunne ha funnet på å klage på sin A (noe som selvsagt er utenkelig i praksis), kunne vedkommende – med maksimal uflaks – oppleve å stryke.

Karakterforslagene varierte mindre mellom grupper enn mellom enkeltpersoner, slik man måtte kunne forvente. Like fullt var mange deltakere overrasket over at det var så stor

variasjon innad i lærergruppen, noe som stemte dårlig med tidligere erfaringer fra ordinær sensurering.

Ikke bare var det stor variasjon på individnivå. På fagdagene kunne vi også spore tendenser til mer systematiske forskjeller. Etter ett arbeidsseminar undersøkte vi om det var systematiske forskjeller mellom lærere som underviste på ulike årstrinn og mellom personer med ulik erfaring med sensurering. Noe overraskende var det større spredning i karakterforslag fra erfarne sensorer enn de med liten erfaring. Selv om dette ikke er godt empirisk dokumentert, kunne det virke som om lang sensorerfaring gir større sikkerhet i vurderingen, men også sterkere tendens til å insistere på egen vurdering og mindre vilje til revisjon og forandring.

I det siste arbeidsseminaret ett år senere, ble alle bedt om å vurdere én bestemt besvarelse. Den hadde oppnådd karakteren A ved ordinær sensur året før, men var også i dette tilfellet en besvarelse som studieledelsen karakteriserte som vanskelig å vurdere. Deltakerne var heller ikke denne gangen informert om hvilken karakter som var oppnådd for besvarelsen.

Tabell 2. Avvik mellom ordinær sensur og forslag fra individuelle lærere, siste seminar (Okt 2007)

	Samme karakter	Avvik = 1 karakter	Avvik = 2 karakterer	Avvik = 3 karakterer	Avvik > 4 karakterer	Total N
3 studieår besvarelse gitt A	6%	13%	44%	22%	16%	32

Samsvaret mellom «sensorene» var fremdeles lavt. Den bedringen vi hadde håpet på, kanskje til og med regnet med, var vanskelig å spore. Mer enn 80 % av deltakerne mente det var riktig å gi C eller dårligere til en besvarelse som ved ordinær sensur var blitt belønnet med A. En tredel av deltakerne foreslo karakteren D eller enda dårligere.

I vurderingen av dette nedslående resultatet, skal noen momenter tas i betraktning:

- Besvarelsen som ble valgt ut til den avsluttende «testen», var en besvarelse som de faglig ansvarlige for studiet antok kunne være noe vanskelig å vurdere. På samme måte som «supersensorene» til dels var ganske kritiske til en del av de oppgavene som var gitt til eksamen i Raaheims studie (2000), kunne det samme være tilfelle her.
- Erfaringen fra fagdagene var at det var en systematisk tendens til strengere vurdering i arbeidsseminarene enn i den reelle vurderingen. Det gjaldt særlig for de oppgavene som hadde fått god karakter.
- Tidspunktet for avsluttende «test» kan ha vært for tidlig til å registrere endring i vurderingspraksis.

Men det generelle bildet er ikke til å komme bort fra: Alt arbeidet som var lagt ned i å utvikle kvaliteten i sensureringen, hadde ikke gitt de resultatene vi trodde og håpet på.

Bruken av vurderingskriterier

Ved avdelingen er det utviklet et kriteriesett som brukes relativt systematisk. Det var stadig diskusjon om bruken av disse vurderingskriteriene på fagdagene. Et gjennomgående synspunkt var at det både var for mange kriterier og at de hadde ulik relevans. En sterkere vektlegging av fagkunnskap ble hyppig etterlyst. Vi ville undersøke dette forholdet litt nærmere ved at deltakerne gjennomførte en mer analytisk form for vurdering enn den holistiske formen som mange foretrekker. Ville vurderingssamsvaret bli høyere når deltakerne konsentrerte vurderingen om ett kriterium? Vi valgte kriteriet «faglig kunnskap». Dette fant sted ved det nest siste seminaret, og var en vurderingsmetode som de ikke hadde erfaring med.

Forsøket ble gjennomført slik: Vi utarbeidet et relativt enkelt vurderingsskjema («scoring rubric») som kun fokuserte på ulike aspekter av fagkunnskap. Vurderingsskjemaet besto av åtte underordnede kriterier som omhandlet ulike aspekter av fagkunnskap. Ved vurderingen skulle det gis ulik vekt på de ulike aspektene, idet det var angitt ulike poengintervaller på de ulike aspektene. Deltakerne skulle gi poeng på hvert delkriterium, summere dem, og så bruke en omregningskala for å komme frem til en poengsum som igjen ga en karakter. Deretter skulle de vurdere hvorvidt den karakteren man hadde kommet frem til på denne måten var et «riktig» resultat ut fra deres vanlige, holistiske vurdering.

Tabell 3. Observert spredning blant lærerne på vurderingskriteriet: Viser faglig innsikt

	Forskjellige aspekter ved vurderingskriteriet: Viser faglig innsikt	Poengskala	Observert Variasjon	
1.	Studenten identifiserer relevant faglig kunnskap i forhold til den situasjonen som er beskrevet i oppgaven.	1–20	48/120	40%
2.	Studenten gir vektige begrunnelser for valg av den teori som trekkes inn.	1–10	41/60	68%
3.	Studenten anvender kunnskap i forhold til beskrevet praksissituasjon på en holdbar måte.	1–15	50/90	56%
4.	Studenten redegjør fokusert, presist og dekkende for relevant teori.	1–30	76/180	42%
5.	Studenten dokumenterer oversikt og kjennskap til det kunnskapsfeltet praksissituasjonen er hentet fra	1–5	16/30	53%
6.	Studenten tolker og bruker data på kvalifisert vis.	1–10	31/60	52%
7.	Studenten identifiserer evt. manglende data i oppgaven og begrunner nødvendighet og relevans av dette.	1–5	19/30	63%
8.	Studenten trekker kun inn relevant kunnskap og ikke noe som er irrelevant.	1–5	16/30	53%
Uenighetsrom benyttet av deltakerne			267/600	45%

Resultatene i tabell 3 må forklares. Antall poeng til disposisjon for hvert kriterium er gitt i tabellen. Den enkle analysen vi har gjennomført gikk ut på å bare ta hensyn til ekstremverdiene. Dersom det er mulig å gi fra 0 til 20 poeng på ett kriterium, vil en score på 20 bety at minst én vurderer har gitt 20 poeng på dette kriteriet mens minst en annen vurderer har gitt

0 poeng. En score på 0 betyr at alle deltakerne hadde gitt samme verdi på besvarelsen på dette kriteriet, uansett hvilken verdi de hadde samlet seg om. På denne måten får vi frem hvor stor del av et teoretisk «uenighetsrom» som er blitt utnyttet. Scoren på 20 ville i så fall bety at hele det teoretiske uenighetsrommet er blitt brukt av deltakerne for dette kriteriet, mens score 0 ville innebære at ikke noe av uenighetsrommet som er til disposisjon er blitt utnyttet, dvs. at alle hadde gitt like mange poeng for den aktuelle besvarelsen på dette delkriteriet.

Resultatene tyder på at deltakerne, totalt sett, tok i bruk halvparten (45 %) av det rommet for uenighet som var til disposisjon. Det ser ikke ut til å være noen særlige forskjeller mellom delkriteriene.

Samsvaret mellom kollegene var ikke – slik mange deltakere syntes å tro – høyere når det var fagkunnskap som ble vurdert, sammenlignet med andre kriterier som finnes på listen av vurderingskriterier ved avdelingen. Det syntes ikke å være lettere å vurdere «fagkunnskap» mer likeartet enn andre vurderingskriterier.

Ideen om å foreta analytisk vurdering fikk heller ikke særlig støtte blant deltakerne. Det var et ganske samstemmig syn som kom frem om at de helst ville fortsette med den form for holistisk vurdering som de var vant med.

Forskjeller i begrunnelse for karakterer

Karakterene, begrunnelsene og diskusjonene under seminarene synliggjorde forskjeller i vurderingspraksis som var ukjent for lærerne. Som de kvantitative dataene viser, endte vi likevel opp med omtrent like dårlig samsvar mellom sensorene som året før. De kvalitative dataene bekrefter forskjeller i forståelsen av kriterier og vurderingspraksis som hadde vært fremtredende gjennom prosjektet. Analysen av begrunnelsene viser noe av kompleksiteten i vurdering og grunnlaget for det lave sensorsamsvaret.

En student beskriver en sykepleier som gir en pasient medisiner i en øyeblikkelig hjelpsituasjon der lege ikke er til stede. Studenten er tydelig på at dette er i strid med loven, og diskuterer legale, etiske og sykepleiefaglige sider ved situasjonen. Dette var en selvopplevd episode, og det hører med til historien at legen hadde støttet sykepleieren i at hun gjorde rett i å intervenere. En sensor aksepterer imidlertid ikke diskusjonen om det moralske dilemmaet fordi sykepleieren handlet i strid med sykepleiernes legale ansvar. En annen sensor mente at sykepleieren ikke brøt noen lov fordi medikasjonen var en fortsettelse av tidligere behandling gitt av legen. Begge mente imidlertid at studenten kunne stryke, men med ulike begrunnelser. En tredje sensor mente situasjonen ikke var reell eller var beskrevet feil. En fjerde sensor mente studenten viste profesjonell innsikt i beskrivelsen av situasjonen og fremragende refleksjon over sykepleierens profesjonelle rolle. Disse uenighetene mellom sykepleielærere synliggjør grunnleggende forskjellige normative posisjoner innen spektret av faglig vurdering.

I denne type oppgave beskriver studentene en selvvalgt pasientsituasjon. Forskjellene i vurderingene viser til kvalitet på flere forhold: situasjonsbeskrivelsen, relevans av tema som studentene velger å fokusere på, studentens anvendelse av kunnskap eller de foreslåtte sykepleieintervensjonene. Det er med andre ord minst fire steg hvor uenighet kan oppstå. Vi kan forvente at med flere muligheter for faglighet, vil også mulighetene for ulik vurdering øke.

I noen av de tilfellene der spriket var størst, var uenigheten lærerne imellom rettet mot kasusbeskrivelsen, altså beskrivelsen av pasientsituasjonen og sykdomstilfellet, som skulle danne utgangspunktet for studentenes dokumentasjon av sykepleiefaglig kunnskap. Noen lærere mente kasusbeskrivelsen var feil eller at den sykepleiefaglige intervensjonen brøt med prinsipper for hva sykepleiere har ansvar for. Andre lærere la lite eller mindre vekt på det, og mer vekt på den kunnskapen studenten viste om sykdomsbildet og pasientsituasjonen som hun definerte. Unøyaktigheter i kasusbeskrivelsen la de mindre vekt på.

For å undersøke begrunnelsene nærmere, ble også alle A/B- og D/E-begrunnelser analysert og sammenlignet. Det viste seg at disse begrunnelsene for oppgaver av hhv. høy (A/B) og lav (D/E) kvalitet hadde høy grad av indre konsistens. A/B-begrunnelsene var relativt like og D/E-begrunnelsene var relativt like. Begrunnelsene ble også analysert språklig med vekt på substantiver, verb og adjektiv.

I A/B-begrunnelsene ble det hyppig brukt adjektiver som *selvstendig, konkret, presis, sammenheng, utfordrende, strukturert og relevant*. D/E-begrunnelsene var annerledes: *klipp og lim fra læreboka, generell, unøyaktig, mangelfull, overfladisk og elementær*. Analysen av verb viste hva lærerne forventet at studentene skulle gjøre og substantivene hva besvarelser av god kvalitet skulle handle om. For å få A eller B bør de *definere, beskrive, identifisere, drøfte, holde fokus, diskutere, reflektere, vurdere, sitere, begrunne og integrere*, mens dette manglet i D/E-begrunnelsene. Temaer som ble lagt vekt på var *relevant kunnskap, pasientens situasjon, hovedproblemet, helheten, observasjoner, sykepleieperspektivet, problemområdet, pasientrollen, fagplanen, sammenheng mellom pasientens situasjon og problem, sykepleiers ansvarsområde, forebygging, tverrfaglig samarbeid, behandling, litteratur og forskning*.

Den indre konsistensen i begrunnelsene for høy og lav måloppnåelse tyder på at lærerne er relativt enige på et generelt plan om hva de forventer av studentene. Det var generelt forskjell mellom det å sette ord på hva som skal til for å få bestemte karakterer – for eksempel hva som kjennetegner en B eller C – og det å gi karakteren B eller C på en bestemt besvarelse. Det er i vurderingen av besvarelsene at spriket oppstår. Det å utvikle en felles forståelse for kriterier på generell basis, og slik sett være enige på et allment plan om hva som kjennetegner kvalitet på ulike nivåer, sikrer ikke lik anvendelse av disse kriteriene i vurderingen av aktuelle besvarelser.

Konklusjon og implikasjoner

Ved avslutningen av prosjektet var karakterene i relativt stor grad avhengig av sensorenes ulike vurderinger. Konklusjonene er rimelig klare:

1. Samsvaret mellom sensorer i avdelingen ble vurdert til å være under den standard man synes det er rimelig å akseptere.
2. Et omfattende prosjekt med systematiske diskusjoner i kollegiet om vurdering av studentarbeider knyttet til vurderingskriterier og det å skrive begrunnelser til studentene, ga ingen merkbar bedring i sensorsamsvar. Fremdeles ville en student som fikk B ved eksamen og som klaget, kunne risikere å få en dramatisk lavere karakter av klagekommisjonen.

Lærerne som deltok i prosjektet, vurderte det som interessant, nyttig og relevant. Det ble klart at det er viktig å ha klare kriterier, men at dette på ingen måte er tilstrekkelig. Kriteriene vil alltid kunne forstås ulikt, vektingen mellom kriteriene innbyrdes kan variere og enighet om kriteriene på det allmenne plan garanterer ikke lik anvendelse av kriteriene i vurderingen av konkrete besvarelser. Det ble også tydelig at de færreste finner det lett å endre egen praksis. Særlig for erfarne sensorer synes det vanskelig å tilpasse sin egen vurdering til det som kommer frem i kollegiale overlegninger. Utvikling av vurderingskompetanse ser ut til å være en integrert del av sosialiseringen inn i fagfellekulturen. I denne prosessen slipes en vurderingsstrategi til med standarder som etter hvert får et absolutt preg – et viktig element i ens faglige og profesjonelle identitet. Faktisk kan det se ut til at faglig styrke kan være en hindring for omforent praksis. Vi erfarte også at det er vanskelig å formulere metoder for utvikling av en felles praksis.

Det magre utbyttet av sensurprosjektet kan forstås som uenighet om vurderingsordning. Ved denne avdelingen er det som er kalt holistisk vurdering den vanlige. I samtidig litteratur og politiske beslutninger er det alternativet – den analytiske vurdering – som blir fremhevet som en bedre form. Den innebærer å dele opp vurderingen i flere, mindre elementer/aspekter og sammenholde prestasjon med et sett med klare kriterier for deretter å komme frem til en konklusjon. Sadler har vært en av de tydeligste kritikerne av slik analytisk vurdering, og mener at en holistisk vurdering har mange viktige kvaliteter. Andre forskere, som Gibbs og Dunbar-Godett (2007), støtter en grunnleggende kritikk av det generelle vurderingsregime som er blitt utviklet som det (antatt) optimale de senere år, der det antas at å vurdere besvarelser analytisk i forhold til eksplisitte og tydelige vurderingskriterier er det beste.

I det foreliggende tilfellet kolliderte tanken om å innføre en mer analytisk vurdering med deltakernes syn. Deres forsøk på å prøve ut en form for analytisk vurdering ga da heller ikke resultater. Men spørsmålet består: Holistisk vurdering innebærer en helhetlig vurdering i forhold til nærmest personlige krav, standarder og verdier. Og vår tanke om å nærmest tvinge deltakerne til å uttrykke disse personlige standardene og gjøre dem til gjenstand for analyse, gjennomtenkning og drøfting knyttet til konkret anvendelse, førte altså ikke frem. Vår tanke var også at en analytisk vurdering ville kunne få frem enda mer av de private standardene til gjennomdrøfting. Det fungerte heller ikke – avvisningen av et slikt vurderingsregime ble viktigere enn slike kollegiale drøftinger.

Ambisjonen om å komme frem til noe nær perfekt vurderersamsvar er like mye et fata morgana nå som det var før prosjektet ble satt i gang. Men er en viss kynisk resignasjon den eneste konsekvens man kan trekke – at vi fortsatt må leve med den situasjonen at det kan bety vel så mye *hvem* som setter karakteren som *hva* studentene har prestert? I det minste må man opprettholde ambisjonen om å holde samsvaret innen visse definerte grenser. Og fortsatt står vi igjen med problemet at noen oppgaver kan være svært vanskelig å vurdere, at ulike sensorer vil legge ulike premisser til grunn for vurderingen og at tilliten til vurderingsordningene eroderes.

Man kan tenke flere praktiske metoder for å maskere et lavt sensorsamsvar, å skjule det eller kun sørge for at det ikke kan undersøkes. Men vi står fortsatt fast i vår forhåpning om at det ikke bare er mulig – det er faktisk også nødvendig å utvikle sensureringskompetan-

sen innen et hvilket som helst sensorkorps gjennom samtale og refleksjon i tilknytning til konkret sensurering. Kanskje det springende punktet er knyttet til hvem som skal delta og hvordan det skal bli en integrert del i samarbeidet lærere, sensorer og studenter? De positive resultatene som er rapportert fra velfungerende lærergrupper og med samarbeid med studentene, kunne tyde på det.

Litteratur

Baird, J.-A, Greatorex, J. & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education*, *11*, 3, s. 331–348.

Baume, D. & Yorke M. (2002). The reliability of assessment by portfolio on a course to develop and accredit teachers in higher education. *Studies in Higher Education*, *27*, 1, s. 7–25.

Black, P. (1998). *Testing: friend or foe?* London: Falmer Press.

Bloxham, S. (2009). Marking and Moderation in the UK: False Assumptions and Wasted Resources. *Assessment & Evaluation in Higher Education, 34*, 2, s. 209–220.

Cox, R. (2007). Examinations and higher education: a survey of the literature. *Higher Education Quarterly*, *21*, s. 292–340.

Crook, C., Gross, H. & Dymott, R. (2006). Assessment Relationships in Higher Education: The Tension of Process and Practice. *British Educational Research Journal,* 32, 1, s. 95–114.

Ecclestone, K. (2001). 'I know a 2:1 when I see it': understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education*, *25*, 3, s. 301–313.

Elton, L. & Johnson, B. (2002). *Assessment in universities: A critical review of research.* York: Higher Education Academy.

Elwood, J. & Klenowski, V. (2002). Creating Communities of Shared Practice: the challenges of assessment use in learning and teaching. *Assessment & Evaluation in Higher Education,* 27, 3, s. 243–256.

Gibbs, G. & Dunbar-Goddet, H. (2007). *The effects of programme assessment environments on student learning.* Oxford Learning Institute. University of Oxford. Report submitted to the Higher Education Academy.

Goos, M. & Moni, K. (2001). Modelling professional practice: a collaborative approach to developing criteria and standards-based assessment in pre-service teacher education courses. *Assessment & Evaluation in Higher Education*, *26*, 1, s. 73–88.

Harlen, W. (2005). «Trusting Teachers' Judgement: Research Evidence of the Reliability and Validity of Teachers' Assessment Used for Summative Purposes.» *Research Papers in Education 20*, 3, s. 245–270.

Knight, P.T. (2001). *A briefing of key concepts.* York: Learning and Teaching Support Network.

Knight, P.T. & Yorke, M. (2003). *Assessment, Learning and Employability.* Maidenhead: Open University Press.

Kvale, S. (1996). Assessment as construction of knowledge. I: R. Hayhoe & J. Pan (Eds.): *East-West Dialogue Knowledge and Higher Education*, s. 117–140. New York: Sharpe.

LeMahieu, P.G., Gitomer D.H. & Eresh, J.A. (1995). Portfolios in large-scale assessment: difficult but not impossible. *Educational Measurement: issues and practice, 14*, s. 11–16, 25–28.

Murphy, R. (2006). Evaluating new priorities for assessment in higher education. I: C. Bryan & K. Clegg (Eds.): *Innovative Assesment in Higher Education*, s. 37–47. London: Routledge.

Light, G. & Cox, R. (2001) *Learning and teaching in higher education: The reflective professional.* London: Sage.

Nystrand, M., Cohen A.S. & Dowling, M. (1993). Addressing reliability problems in the portfolio assessment of college writing. *Educational Assessment*, 1, s. 53–70.

O'Donovan, B., Price, M. & Rust, C. (2001). The student experience of criterion-referenced assessment through the use of a common criteria assessment grid. *Innovations in Learning and Teaching International*, *38*, s. 74–85.

Pitts, J., Coles C., Thomas, P. & Smith, F. (2002). Enhancing reliability in portfolio assessment: discussions between assessors. *Medical Teacher, 24*, 2, s. 197–201.

Price, M. (2005). Assessment Standards: The Role of Communities of Practice and the Scholarship of Assessment. *Assessment and Evaluation in Higher Education*, 30, 3, s. 215–230.

Price, M. & O'Donovan, B. (2006). Clarifying Assessment Criteria and Standards. I: C. Bryan & S. Clegg (Eds.): *Innovative Assessment in Higher Education*, s. 100–109. New York: Routledge.

Rust, C., Price, M. & O'Donovan, B. (2003). Improving students' learning by developing their understanding of assessment criteria and processes. *Assessment and Evaluation in Higher Education*, *28*, 2, s. 147–164.

Raaheim, A. (2000). En studie av inter-bedømmerreliabilitet ved eksamen på psykologi grunnfag (PS101). *Tidsskrift for den norske psykologforening 37*, 3, s. 203–213.

Sadler, D. R. (1987). Specifying and promulgating achievement standards. *Oxford Review of Education, 13*, 2, s. 191–209.

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

Sadler, D. R. (2007). Perils in the meticulous specification of goals and assessment criteria. *Assessment in Education*, *14,* 3, s. 387–392.

Supowitz, J., Macgowan, A. & Slattery, V. (1997). Assessing agreement: an examination of the interrater reliability of portfolio assessment in Rochester. *Educational Assessment*, *4*, s. 237–259.

Torrance, H. (1995). *Evaluating authentic assessment.* Buckingham: Open University Press.

Torrance, H. (2007). Assessment *as* learning? How the use of explicit learning objectives, assessment criteria and feedback in post-secondary education and training can come to dominate learning. *Assessment in Education, 14*, 3, s. 281–294.

Yorke, M. (2003). Formative assessment in higher education: moves towards theory and the enhancement of pedagogic practice. *Higher Education, 45*, 4, s. 477–501.