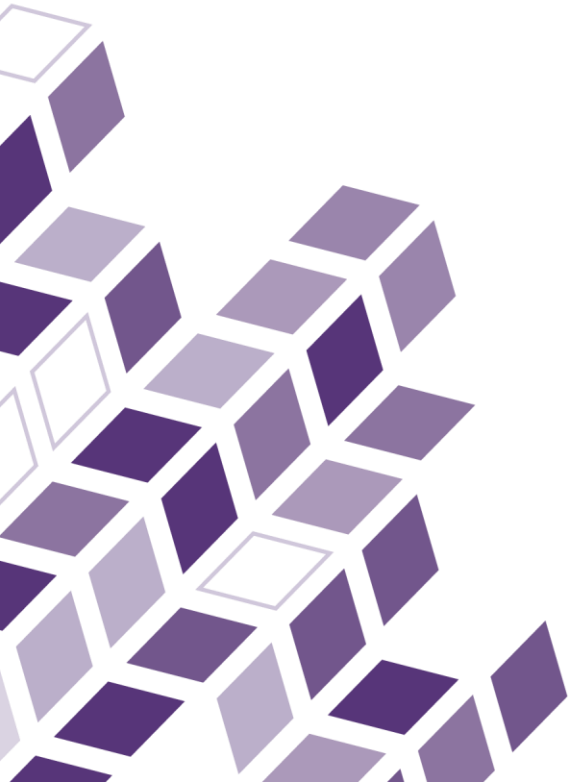
NOKUTS utredninger og analyser

Kriterier og skjønn i evaluering

En kasusstudie i utøvende musikkutdanning

Oktober 2010



NOKUT Nasjonalt organ for kvalitet i utdanningen

Rapporttittel:	Kriterier og skjønn i evaluering: en kasusstudie i utøvende musikkutdanning
Forfatter(e):	Professor Vidar Gynnild
Dato:	05.10.2010
NOKUTs rapporter nr:	2010 - 4

NOKUT Nasjonalt organ for kvalitet i utdanningen

Forord

NOKUTs analyse- og utredningsrapporter har til formål å gi bidrag til økt kunnskap om forhold innenfor høyere utdanning og fagskoleutdanning som har betydning for kvaliteten i studiene. Vi håper at de kan gi nyttige ideer og stimulans til institusjonenes arbeid med å kvalitetssikre og videreutvikle sine studietilbud. Rapportene vil dels formidle analyser av informasjon som NOKUT innhenter gjennom sin evaluerings-, akkrediterings- og godkjenningsvirksomhet, dels også resultater fra særskilte undersøkende prosjekter som NOKUT foretar, ofte i samarbeid med eksterne.

Den foreliggende rapporten formidler resultater fra et prosjekt som professor Vidar Gynnild gjennomførte da han hadde en bistilling i NOKUT i 2008-10. Professor Gynnild har sin hovedstilling ved NTNU. Med utgangspunkt i utøvende musikkutdanning problematiserer han flere forhold rundt vurdering og karaktersetting i høyere utdanning, blant annet et uavklart forhold mellom normativ og kriteriebasert vurdering og motsetningen – og samspillet - mellom analytisk og holistisk vurdering. Selv om disse problemstillingene kan være spesielt påtrengende ved vurderinger av kunstneriske prestasjoner, har de gyldighet og aktualitet også for all annen bedømmelse av læringsresultat. Rapporten er også publisert som vitenskapelig vurdert artikkel i tidsskriftet UNIPED (Nr. 2/2010).

Oslo, 5. oktober 2010

Terje Mørland
Direktør

Sammendrag

Et nytt, felles karaktersystem for høyere utdanning ble innført i 2003, med beskrivelser av hvert trinn på skalaen fra A–E. Karakteren C var ment å uttrykke en middels god prestasjon, med de øvrige karakterene godt fordelt på begge sider av dette midtpunktet. Etter flere års bruk for et stort antall studenter, har denne forventningen ikke slått til i alle utdanninger, spesielt ikke på mastergradsnivå. Undersøkelsen viser at institusjoner for utøvende musikkutdanning nesten uten unntak benytter de tre beste karakterene (A, B og C). Er karakterene blitt for gode, eller speiler de bare et forventet, høyt prestasjonsnivå? Rapporten bygger på data fra et pilotprosjekt innen utøvende sangutdanning, med utvikling av eksplisitte vurderingskriterier som bidrag til en mer transparent og pålitelig vurdering i en kunstnerisk disiplin. Rapporten behandler utfordringene ut fra et evalueringsteoretisk ståsted, og denne prinsipielle tilnærmingen gjør rapporten interessant for alle som har et engasjement i forhold til teoretiske og/eller praktiske spørsmål ved vurdering.

Innhold

1 Innledning

Norges deltakelse i Bologna-prosessen fra 1999 markerer et tidsskille for høyere utdanning. Dette har medført endringer i gradsstruktur, studieplaner, undervisnings- og vurderingsformer, slik at studenter lettere kan bevege seg mellom ulike utdanningsinstitusjoner på tvers av grenser. Standardisering av studiepoeng og arbeidsbelastning samt mer jamførbare akademiske grader og vurderingsuttrykk har hatt økt studentmobilitet som sentralt mål (St.meld. nr. 27 (2000–2001) Gjør din plikt – Krev din rett, 2001). Nasjonalstatene har historisk sett dels benyttet helt ulike karakterskalaer, og dels har ytre sett identiske skalaer vært benyttet på svært ulike måter. Dette har aktualisert behov for større grad av samordning av vurdering og karakteruttrykk. Mens de fleste land har valgt å beholde sine nasjonale karakterskalaer i tillegg til en konverteringsskala, besluttet Norge å benytte konverteringsskalaen som sin nye, felles skala for all høyere utdanning. Skalaen blir oftest omtalt som European Credit Transfer System (ECTS) (Commission, 2009).

Nasjonal bruk av felles karakteruttrykk gir ytre sett inntrykk av at høyere utdanning nå har fått én felles ”valuta”, som antas å være konvertibel mellom institusjoner og land. En viktig premiss for dette var imidlertid felles forståelse og konsistent bruk av skalaen på tvers av emner og nivå i utdanningen. Bruken av karakterskalaen skulle samordnes ved en gitt fordelingsnøkkel for de ulike karakterene. For studenter som består eksamen, la departementet til grunn at rundt 10 prosent av studentene ville oppnå A, 25 prosent B, 30 prosent C, 25 prosent D og 10 prosent E. En slik prosentvis fordeling kan oppnås ved at prestasjoner rangeres og sorteres etter gitte normer. Selv om en slik fordeling ikke var ment som et absolutt krav ved hver prøve eller eksamen, ble forventningen opprettholdt for mange studenter over tid. Ved de utøvende musikkutdanningene har imidlertid ikke denne forventningen slått til. Et typisk eksempel er Norges musikkhøgskole, der de mange gode karakterene bidro til en omfattende intern høring om karakteruttrykk våren 2009 med etterfølgende behandling i styringsorganene. Resultatet ble at graderte karakterer ble beholdt for lavere grad, men sløyfet for mastergradsstudiene i utøvende disipliner. Spenningen mellom relativ og absolutt vurdering er ikke et eksklusivt norsk fenomen, men eksemplet viser forholdet mellom ECTS-skalaens krav til relativ (normbasert) vurdering og det europeiske kvalifikasjonsrammeverkets absolutte (kriteriebaserte) krav til kompetansemål.

I denne rapporten undersøker jeg først karakterfordelinger ved sju utøvende musikkutdanninger i Norge, med spesiell vekt på Norges musikkhøgskole (NMH). De øvrige institusjonene er Barratt Due Musikkinstitutt (BDM), Universitetet i Agder (UiA), Universitetet i Stavanger (UiS), Griegakademiet, Universitetet i Bergen (Grieg), Norges teknisk-naturvitenskapelige universitet (NTNU) og Høgskolen i Tromsø (HiTø). Dokumentasjon av karakterfordelinger danner grunnlag for forsknings- og utviklingsarbeid innen det sangfaglige miljøet ved NMH, med spesiell vekt på vurderingskriterier. Ved NMH er det lang tradisjon for at prestasjonsvurdering innen utøvende disipliner har vært basert på ”taus” kunnskap og implisitte kriterier. Utfordringer knyttet til bedømmelse av kunstneriske prestasjoner kan oppfattes som et spesielt tilfelle dersom man sammenlikner med en rekke andre studier i høyere utdanning. Jeg tror likevel at vesentlige spørsmål om holistisk (helhetlig) versus analytisk vurdering – som står sentralt i høyere utdanning generelt – framstår med enda større tydelighet ettersom prinsipielle spørsmål ved vurdering melder seg med enda større kraft i kunstneriske disipliner. Med Bent Flyvbjergs vitenskapsteoretiske terminologi kan vår studie anses som en ”ekstrem case” (Flyvbjerg, 2010).

Rapporten beskriver og analyserer et utviklingsprosjekt ved NMH som retter oppmerksomhet mot prinsipielt vesentlige spørsmål ved prestasjonsbedømmelse, spesielt bruken av kriterier ved vurdering av sang. Sentrale begreper og prinsipper blir benyttet, med referanse til Royce Sadlers evalueringsteoretiske arbeid. Han har i særlig grad bidratt til innsikt i vurdering som fenomen og begrep, med klargjørende bidrag i forhold til teoretiske og prinsipielle spørsmål på dette feltet.

1.1 Problemstillinger og metode

I 2005 etablerte de utøvende musikkutdanningene en ordning med felles sensor i noen utvalgte instrumenter, blant disse sang. Fellestrekk i bruken av den graderte karakterskalaen ved disse institusjonene dannet grunnlag for et mer omfattende forsknings- og utviklingsarbeid ved NMH. Vurderingskulturen har vært preget av en holistisk og implisitt tilnærming, uten felles kriterier og språklig repertoar for tilbakemelding overfor studentene. Faglærerne uttrykte frustrasjon over denne situasjonen, spesielt på grunnlag av nasjonalt pålagte krav til bruken av karakterskalaen. Det var derfor en gruppe godt motiverte undervisere knyttet til det sangfaglige miljøet ved NMH som sa seg interessert i å delta i et forsknings- og utviklingsarbeid i 2008/2009. Hensikten var først å undersøke bruken av karakterskalaen med hovedvekt på følgende problemstillinger:

- Finnes det karakteristiske mønstre med hensyn til karaktersetting i musikkutdanningene?
- Er karaktersettingen i tråd med nasjonale krav og føringer for bruk av karakterskalaen?
- I hvilken grad, eventuelt hvordan, kan kriterier for prestasjonsvurdering gjøres eksplisitt?

Mens data til de to første problemstillingene var lett tilgjengelig ved NMH, krevde den tredje problemstillingen en annen tilnærming, ofte omtalt som aksjonsforskning eller aksjonslæring (Zuber-Skerritt, 2002). I vårt tilfelle hadde alle faglærerne mange års erfaring med undervisning og vurdering av sangprestasjoner, men langt mindre erfaring med samtale om og tydeliggjøring av prinsipper og kriterier ved vurdering av sangprestasjoner, slik denne faglæreren forteller:

> *«Selv om vi nok i våre kommentarer legger vekt på forskjellige ting, så er vi nok stort sett ganske samstemte. Men vi har jo i liten grad – hvis vi skal være helt ærlig – delt essensen i dette. Hva den egentlig går ut på. […] Det er mye taus kunnskap, som man bærer med seg og som man adopterer gjennom egne studietider og yrkeserfaring. Undervisninga er jo på en måte litt privatisert. Selv om vi ofte blir veldig fort enige i karaktersetting, så er det jo dessverre slik at grunnen til at vi setter den karakteren vi setter kan være litt forskjellig.» (Faglærer, NMH, 22.09.09).*

En av utfordringene for faglærerne var derfor å bli oppmerksom på og kunne verbalisere hvilke kriterier som faktisk ble benyttet ved vurdering, og hvordan spørsmål om faglig nivå ble håndtert ved karaktersetting. Rapportforfatteren fungerte i en aksjonsforskningsrolle som faglig veileder og forsker – en kombinasjon som ikke ble opplevd som spesielt konfliktfylt i vårt tilfelle. Jeg definerte i stor grad min egen rolle i forhold til et utviklingsrettet prosjekt, uten at målet var klart definert på forhånd. Mitt antatt vesentligste bidrag besto i å introdusere og forklare generelle evalueringsfaglige begreper og teori, som så ble forsøkt applisert i en sangfaglig kontekst. Jeg fungerte også i rollen som "djevelens advokat", utfordret ved å stille kritiske spørsmål og bidro i diskusjonene, samtidig som jeg sørget for rammer med hensyn til framdrift. Ettersom jeg faglig sett var å anse som en novise i utøvende sangutdanning, forble dette faglærernes eget ansvar.

1.2 Terminologi

I denne rapporten benyttes ”vurdering” og ”evaluering” synonymt. Dette definerer Sadler som «… the process of forming a judgment about the quality and extent of student achievement of performance, and therefore by inference a judgment about the learning that has taken place» (Sadler, 2005). Som grunnlag for vurdering benytter høyere utdanning mange ulike typer oppgaver og/eller prøveformer, tilpasset utdanningens egenart, formål og nivå. Vurdering benyttes som tilbakemelding til studenter underveis (formativ funksjon) og/eller som grunnlag for en endelig karakter, for eksempel uttrykt ved bokstaver eller tall (summativ funksjon). Iblant ivaretar underveisevalueringer en todelt funksjon, både som læringsredskap og som bidrag til sluttkarakter. Teoretisk sett er dette problematisk fordi kvalifikasjonsrammeverket (EQF) vektlegger beskrivelser av tilsiktet læringsresultat etter gjennomført utdanning, noe som oftest vil være annerledes enn det som kan forventes på tidligere stadier i utdanningsforløpet.

Basert på faglig egenart og tradisjon skjer evalueringen med utgagnspunkt i en helhetsvurdering (holistisk vurdering) eller ved mer detaljert vurdering av de enkelte deler (analytisk vurdering). Sistnevnte er vanlig i teknisk-naturvitenskapelige disipliner, mens holistisk vurdering har en sterkere stilling i humanistiske fag og kunstnerisk/utøvende disipliner.

I alle former for evaluering står spørsmål om referanse sentralt, altså hva som prinsipielt styrer det konkrete evalueringsarbeidet. To ulike prinsipper er omtalt som *normbasert* og *kriteriebasert* vurdering (Sadler, 2005): «*In principle, norm-referencing uses the performance of a group of students as the comparative background for grading decisions, the group usually being defined as all of the students enrolled in a course at one time. Technically, it is the group performance that constitutes the norm. Marks or grades reflect not only how student performances are ranked but also their relative separation*» (Sadler, 2010). Karakterfordelingen forventes å følge med en gitt prosentvis fordeling: «*The aim is to ensure that the relative worth of grades is comparable across the institution. This principle depends only on the apportionment of the various grades, so the shape of the frequency distribution of the underlying marks is irrelevant» (Sadler, 2010).*

Gitte krav til fordeling bidrar til å hindre inflasjon i karaktersystemet, men forteller imidlertid lite om nivå. Derfor gis det gode argumenter for bruk av evalueringsordninger som bidrar til bedre mål på prestasjoner i forhold til gitte krav. Problemet er bare at det ikke finnes noen entydig, felles forståelse av hva *kriteriebasert* vurdering betyr. Det eksisterer derfor ulike tolkninger og dermed ulike praksiser med henvisning til samme evalueringsteoretiske begrep. Sadler diskuterer i en artikkel fire ulike tolkninger av kriteriebasert vurdering, og konkluderer med at ingen av disse er «*… fully capable of delivering on the aspirations of criteria-based grading*» (Sadler, 2005). Kriteriebasert vurdering utfordrer likevel forholdet mellom læringsmål, kriterier og faglig nivå. Ved normbasert vurdering framstår læringsmål (i den grad de finnes) mer som et fromt ønske, der graden av måloppnåelse forsvinner fordi karaktersettingen maskerer faglig nivå.

Perspektivskiftet fra en undervisningssentrert til en student- og læringsorientert tilnærming i høyere utdanning har medført større vekt på *læringsresultatene*, hvordan disse kan beskrives som mål og dokumenteres ved eksamen eller på annen måte. Kort sagt, hva skiller gode prestasjoner fra mindre gode, og hvordan kan prestasjoner forbedres? Utfordringer knyttet til vurdering og læring antas å være større ved økende innslag av skjønn, slik tilfellet er i utøvende disipliner. Det kan eksemplifiseres ved ulike disipliner som matematikk og utøvende kunstnerisk virksomhet, som for eksempel i musikk. Bruken av kriterier og standarder har relevans ved alle former for prestasjonsvurdering, men hvordan

dette kan gjennomføres på hensiktsmessig vis varierer, avhengig av disiplinenes egenart og forutsetninger. Både summativ og formativ vurdering krever innsikt i hvilke kriterier og standarder som gjelder. Dette anses som en forutsetning for at studentene selv kan bidra til å oppnå tilsiktede læringsresultater (Price, 2005).

Vurdering av prestasjoner har derfor blitt viktigere både ut fra et formativt (læringsmessig) og et summativt (oppsummerende) ståsted. I Norge betydde Kvalitetsreformen for høyere utdanning (St.meld. nr. 27 (2000–2001) Gjør din plikt – Krev din rett, 2001) et løft for læringsaspektet ved vurdering, blant annet med bruk av varierte vurderingsordninger underveis i studier. Grunnlaget for bruken av den nye, graderte karakterskalaen har imidlertid forblitt et uavklart tema. Dette kan begrunnes ut fra institusjonsspesifikke forutsetninger og behov, men også ut fra det som kan oppfattes som uklare signaler med hensyn til hvilke prinsipper for vurdering som gjelder. Sist, men ikke minst viser forskningslitteraturen at det er mye vi ikke vet på dette feltet, for eksempel hvordan kriterier og standarder blir skapt og delt i fagmiljøene eller i hvilken grad det eksisterer en felles tolkning og forståelse av begrepene (Yorke, Bridges & Woolf, 2000).

1.3 Karaktersetting i utøvende musikkutdanning

Det er kjent at karakternivået ved hovedinstrumenteksamener på bachelornivå ved de utøvende musikkutdanningene er en god del høyere enn det som var forutsetningen ved innføringen av den graderte karakterskalaen. I et internt notat til Studieutvalget ved NMH skriver studiesjefen:

> *«Ved innføringen la departementet til grunn at man over tid ville få et karakternivå med karakteren C som gjennomsnittskarakter og en normalfordeling der rundt 10 % av studentene skal få A, 25 % B, 30 % C, 25 % D og 10 % E. Det er i tillegg et mål at det skal være en felles nasjonal forståelse og bruk av den graderte karakterskalaen. Det høye karakternivået viser seg også å være gjeldende for andre hovedemner ved Norges musikkhøgskole (NMH), blant annet for emnet masterarbeid i masterprogrammene. Det kan for øvrig stilles spørsmål om bruken av gradert skala for utøvende kunstfag er hensiktsmessig.» (Notat, NMH, 22.05.09).*

Våren 2009 ble spørsmål om bruken av karakteruttrykk ved NMH diskutert med utgangpunkt i et høringsnotat til Studentutvalget, programutvalg og fagseksjoner. Alle høringsinstanser leverte en uttalelse, og saken ble etter dette lagt fram for Studieutvalget. Tallmaterialet i høringsnotatet omfatter karakterfordeling nasjonalt og ved NMH. Ordningen med nasjonal, felles ekstern sensor ble etablert i 2005 og omfatter de seks statlige utøvende musikkutdanningene samt Barratt Due Musikkinstitutt. Siden ordningen ble etablert, er det blitt avlagt 172 karakterer. Tallmaterialet fra NMH omfatter 517 karakterer. Her inngår karakterer for hovedinstrument (2004–2007) og komposisjon (2006–2008) i kandidatstudiene, emnet masterarbeid i mastergradsstudiet i utøving (2006–2007), musikkpedagogikk (2006–2008) og musikkterapi (2006–2008). Studiesjefen ved NMH konkluderer i et notat (datert 22.05.09) med at

- A, B og C utgjør 91 % (156 av 172) av karakterene som er blitt gitt i ordningen med felles ekstern sensor og 95 % (492 av 517) av karakterene for hovedemnene ved NMH
- det er en tilnærmet normalfordeling mellom de tre beste karakterene (A, B og C) både i ordningen med felles ekstern sensor og ved NMH
- gjennomsnittskarakteren ligger nærmere B (4) enn C (3) både i ordningen med nasjonal felles ekstern sensor (3,76) og ved NMH (3,78)

- det er et relativt stort sprik i karakternivå mellom de ulike instrumentene som inngår i ordningen med felles ekstern sensor, og dessuten at de […] kjente forskjellene i opptaksnivå mellom institusjonene ikke i noe vesentlig grad avspeiler seg i karakterene

Fra de nasjonale eksterne sensorrapportene framkommer det at karakterskalaen forstås og brukes på forskjellig vis ved ulike institusjoner og at karakteren C blir regnet som en dårlig karakter både av studenter og lærere som inngår i ordningen. Den graderte karaktersettingen blir generelt satt i et kritisk lys, blant annet fordi studentene ikke opplever ordningen som helt rettferdig.

Høringsinstansene var på sin side delt i sitt syn på spørsmål om karakteruttrykk. Flere gir uttrykk for at den graderte karakterskalaen fortsatt kan brukes for hovedemner på bachelornivå, mens den todelte skalaen anses som mer hensiktsmessig på masternivå (Notat, NMH, 22.05.09). Ordningen med nasjonal fellessensur vurderes som aktuell for hovedemner med gradert karakteruttrykk, men flere høringsinstanser er kritiske til veileders rolle i kommisjonsarbeid. Det ble også ytret ønske om å styrke en kollegial diskurs om vurdering av kunstneriske prestasjoner. Studentutvalget savner konstruktive tilbakemeldinger, og hevder at karakter i seg selv ikke gir en tilfredsstillende tilbakemelding i forbindelse med vurdering av kunstneriske prestasjoner.

2 Analyse

I et notat fra rektorene ved de fire største universitetene (06.11.03) sies det at karaktersettingen «*… skal ta utgangspunkt i den verbale beskrivelse som er gitt av de* [*karakter*]*nivåene*» (http://www.ifi.uio.no/foransatte/skjemaer/notat_rektor.pdf). Karakteren C forventes som typisk prestasjon av "gjennomsnittsstudenten" og som gjennomsnittlig ståkarakter for store grupper over tid. I brev fra departementet (10.05.04) gjentas argumentene om at karaktersetting skal ta utgangspunkt i de verbale karakterbeskrivelsene «*… der karakteren C skal gi uttrykk for en jevnt god prestasjon …*», mens karakteren A skal gi uttrykk for en «*fremragende prestasjon som klart utmerker seg*» (http://matematikkradet.no/dokumenter/2004-05-10_bokstavkar.pdf). Med dette gis en klar referanse til en *normbasert* vurderingspraksis, der prestasjonene til en gruppe studenter benyttes som grunnlag for vurdering av enkeltprestasjoner og for innbyrdes rangering av prestasjoner. I brev av 10.05.04 (som vist til ovenfor) sies det likevel at det finnes «*… ingen forhåndsgitt fordeling av karakterer som noe eksamens- eller studiekull skal "presses" inn i.*» Prestasjoner skal likevel vurderes opp mot hele skalaen for å «*… skjelne gode fra mindre gode prestasjoner og rangere dem innbyrdes.*» Dette representerer en vurderingspraksis fundert i et undervisningsparadigme, som i praksis er forlatt i det europeiske kvalifikasjonsrammeverket. Her står læringsresultatene i sentrum, noe som krever en prinsipielt annen vurderingspraksis:

> *«A common method of determining students' grades 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 learn» (Biggs & Tang, 2007).*

Det eksisterer ingen nødvendig sammenheng mellom normbasert vurdering og normalfordeling i statistisk betydning, selv om fordelingen ofte framstår i en slik form (Sadler, 2010). Stipulert karakterfordeling i ECTS for mange studenter over tid er for eksempel som følgende oversikt viser: A=10 prosent; B=25 prosent; C=30 prosent; D=25 prosent; E=10 prosent (Glasser, 2008).

Universitets- og høgskolerådets analysegruppe fant store forskjeller mellom ulike utdanninger når det gjaldt karakterfordelinger, blant annet for masterarbeider: «*Det kan tyde på at det er noe særegent ved større selvstendige arbeid som gjør dem vanskelige å vurdere på tilsvarende måte som for eksempel skoleeksamener, og som fører til en annen bruk av karakterskalaen*» (Glasser, 2008, s. 6). UHR fant videre at enkelte fagmiljøer har en svært høy andel A og B på masterarbeider, og at disse bør vurdere å benytte bestått/ikke bestått som karakteruttrykk hvis man ellers ikke oppnår større spredning av karakterene (Glasser, 2008, s. 9).

I det samme notatet blir gjeldende karakterbeskrivelser kritisert, og rapporten konkluderer med at benyttede begreper ikke er universelle og at vektleggingen varierer både med hensyn til fag og nivå (Glasser, 2008, s. 11). Den peker samtidig på en annen metode for prestasjonsvurdering ved å knytte vurderingen opp mot *læringsmål*, altså en form for kriteriebasert vurdering. UHRs analysegruppe foreslår at dette prinsippet legges til grunn også ved utforming av de generelle karakterbeskrivelsene, noe som betyr at disse får en annerledes utforming (Glasser, 2008, s. 12). Tabell 1 viser at beskrivelsene er beholdt i sin opprinnelige form for karakterene A, E og F, mens beskrivelsene er endret i større eller mindre grad for B, C og D.

Tabell 1: Karakterbeskrivelser fra 2003, 2008 og 2009 med forsøksvis kategorisering som normbasert eller kriteriebasert.

Karakter	Generell, ikke fagspesifikk omtale av vurderingskriterier	Norm	Krit
A	Fremragende prestasjon som klart utmerker seg (2003)	x	
	Fremragende prestasjon som klart utmerker seg (UHR, 10.12.08)	x	
	Fremragende prestasjon som klart utmerker seg (UHR, 10.07.09)	x	
B	Meget god prestasjon som ligger over gjennomsnittet	x	
	Meget god prestasjon som ligger klart over det forventede nivå	x	x
	Meget god prestasjon som ligger klart over det forventede nivå	x	x
C	Gjennomsnittlig prestasjon som er tilfredsstillende på de fleste områder	x	x
	God prestasjon som oppfyller forventningene	x	x
	Prestasjon som oppfyller læringsmålene på en god måte		x
D	Prestasjon under gjennomsnittet med en del vesentlige mangler	x	x
	Akseptabel prestasjon som likevel ikke oppfyller forventningene fullt ut		x x
	Akseptabel prestasjon som ligger over minimumskravene		
E	Prestasjon som tilfredsstiller minimumskravene, men heller ikke mer		x
	Prestasjon som tilfredsstiller minimumskravene, men heller ikke mer		x
	Prestasjon som tilfredsstiller minimumskravene, men heller ikke mer		x
F	Prestasjon som ikke tilfredsstiller minimumskravene		x
	Prestasjon som ikke tilfredsstiller minimumskravene		x
	Prestasjon som ikke tilfredsstiller minimumskravene		x

Noen av formuleringene er klart normbasert (A), i andre tilfeller er et kriteriebasert prinsipp lagt til grunn (E og F). For B, C og D kan beskrivelsene tolkes på ulikt vis, for eksempel: «*Meget god prestasjon som ligger over gjennomsnittet*» har gruppens prestasjon som referanse. I UHRs forslag (10.12.08) er den opprinnelige formuleringen blitt til: «*Meget god prestasjon som ligger klart over det forventede nivå*» [til gruppen, eller i forhold til læringsmål?]. Ettersom beskrivelsen er ufullstendig, framstår tolkningen som usikker. På prinsipielt grunnlag er det naturlig å spørre om en kombinasjon av absolutt og relativ vurdering er ønskelig eller mulig.

Tilsvarende eksempler finner vi for C og D. Dette skaper uklarhet om vurderingsgrunnlaget, med fare for inkonsistent vurdering, spesielt ved bruk av flere sensorgrupper. Et eksempel på en alternativ løsning foreligger i UHRs foreløpig siste beskrivelse av karakteren C (datert 10.07.09): «*Prestasjon som oppfyller læringsmålene på en god måte*» (Glasser, 2009). Her benyttes begrepet "læringsmål" første gang, noe som muligens indikerer en bevegelse i kriteriebasert retning. Pålitelig vurdering forutsetter likevel fellesforståelse av "læringsmålene", og hva det i praksis betyr å oppfylle disse "på en god måte". Ifølge UHRs analysegruppe (Glasser, 2009) oppfattes beskrivelsene for A, E og F som "uproblematiske", til tross for at omtalen av A og E/F evalueringsteoretisk bygger på ulike prinsipper. En prestasjon som "utmerker seg" (A) har gruppens resultat som referanse, mens henvisning til "minimumskravene" mest sannsynlig har en annen referanse. De nyeste karakterbeskrivelsene representerer en dreining i kriteriebasert retning, men oppleves fortsatt som problematiske. En faglærer utdyper dette slik:

> *«Det som er problematisk innenfor sang, er at du kanskje må sammenligne en sopran med en bass. Og det som er standarder for en stemmetype, er ikke nødvendigvis standarder for en annen stemmetype, altså i forhold til å vurdere hvor det ligger i nivå. Og en dramatisk stemme, der må man sette opp andre standarder enn for en lyrisk stemme. Og det er skrevet repertoar til alle disse stemmetypene, men repertoaret kan veldig vanskelig sammenlignes» (Faglærer, NMH, 22.09.09).*

En premiss for vårt arbeid var troen på at nytteverdien av generiske karakterbeskrivelser kan økes dersom fagmiljøene presiserer hva beskrivelsene betyr i praksis: «*Karakterbeskrivelsene sier jo noe selvsagt, men hva er konkretiseringen, for eksempel "fremragende", hva betyr det? Hva betyr det at man klart utmerker seg?*» (Faglærer, NMH, 22.09.09). På et kriteriebasert grunnlag representerer ikke etablerte mønstre, med vekt på karakterene A, B og C, nødvendigvis et problem. Høye opptakskrav betyr at rekrutteringen representerer en streng seleksjon, noe som i sin tur også må forventes å komme til uttrykk i forhold til de resultatene som oppnås. Ut fra en normbasert vurderingspraksis framstår en systematisk skeivfordeling av karakterene over tid som mer bekymringsfull. Normbasering krever teoretisk sett at denne typen mønstre ikke oppstår.

2.1 Normbasert eller kriteriebasert vurdering?

I dagligtale brukes begrepene "kriterier" og "standarder" ofte synonymt. Royce Sadler har i et konferanseinnlegg beskrevet begrepsforvirringen og konsekvensene av den på følgende vis:

> *«Criteria-based assessment is often embraced as the educationally most defensible approach to assessing student learning for both formative and grading purposes. Historically, criteria-based assessment arose as an alternative to the explicit use of inter-student comparisons as the basis for grading decisions. Criteria, however, is often confused with standards, leading to inappropriate practice. Where standards are acknowledged as the key issue, common ways of specifying them are fundamentally inadequate» (Sadler, 2003).*

I denne rapporten opprettholder vi et begrepsmessig skille, slik at *kriterium* betyr kvalitet eller art, hva som skal inkluderes og vektlegges, mens *standard* viser til nivå. Begge dimensjonene ligger som uuttalte forutsetninger ved all vurdering, men konkretisering av begrepene er ofte utilstrekkelig, noe som hemmer diskurser om vurderingens forutsetninger og grunnlag.

Vurdering av kunstneriske prestasjoner er en utfordring fordi den til sist bygger på den helhetlige opplevelsen, som oftest er noe annet enn summen av de enkelte delene: «*Det er en kunstnerisk form, og det må være en vesentlig porsjon opplevelse. Jeg må på en måte som vurderingsperson være i stand*

til å la meg begeistre» (Faglærer, NMH, 22.09.09). Slik holistisk vurdering er integrert, og "finstemt" i forhold til et stort register av kriterier som er skapt via utdanning og erfaring. De involverte faglærerne mente likevel at eksplisitte kriterier kunne være en god støtte og et nyttig bidrag for bedre vurderingspraksis. I tillegg til søk og oppdatering i aktuell litteratur, gjennomførte jeg mer "åpne" undersøkelser på Internett med relevans for vurdering i utøvende disipliner. I den anledning kom jeg over en doktorgradsavhandling i Australia (Wrigley, 2005), med teoretisk og metodisk interesse for arbeidet vårt ved NMH. Fordi kriteriene var utviklet i en noe annen kontekst, ble disse bearbeidet og forsøkt tilpasset behovene ved NMH (tabell 2).

Tabell 2: Kriteriesett for vurdering av sangprestasjoner (Bjørkøy, NMH, 15.01.10)

Karakterer	A 5	B 4	C 3	D 2	E 1	F 0	
Teknikk (ønskede kvaliteter)							**Kommentarer:**
Kroppsholdning, avspenthet							
Pust og støtte (energi, kompresjon)							
Ansats, luftstrøm							
Vokalegalisering (klangjevnhet)							
Intonasjon							
Registeregalisering (balanse)							
Stemmens fokusering							
Stemmeomfang							
Klang/timbre							
Koloratur, triller, forsikringer							
Språk/tekst (ønskede kvaliteter)							**Kommentarer:**
Artikulasjon og tekstuttale (språklig nøyaktighet)							
Teksttolkning (forståelse, selvstendighet og deklamasjon)							
Språkbredde (for eksamenskonsert)							
Musikalitet (ønskede kvaliteter)							**Kommentarer:**
Stilforståelse (stilistisk differensiering)							
Frasering /legato							
Dynamikk (variasjon, kontraster)							
Tempo (tempovalg, tempostabilitet)							
Rytme (stringens og variasjon)							
Formidling (ønskede kvaliteter)							**Kommentarer:**
Presentasjon (muntlige introduksjoner)							
Kommunikasjon (innlevelse, utstråling, uttrykk)							
Scenisk fremstilling og iscenesettelse							
Repertoarvalg (for eksamenskonsert)							**Kommentarer:**

Å utvikle et komplett sett av eksplisitte vurderingskriterier, er en nesten uoverkommelig oppgave. Utfordringen består derfor i å optimalisere forholdet mellom presisjon og nytteverdi. En tidlig versjon av kriteriesettet ble brukt for utprøving internt ved NMH i 2009, som i sin tur ble revidert fire-fem ganger i løpet av våren og høsten 2009 (tabell 2). Dels besto utfordringen i å etablere overordnede kategorier, dels å gjøre kriterier for hver kategori mest mulig entydig og tydelig for brukerne. Følgende sitat eksemplifiserer utfordringer ved valg av vurderingskriterier:

> *«Och så saknade jag två saker när jag hörde så många luftiga stämmor. Då saknar jag rubriken* fokus *och rubriken* legato*, som jag skulle vilja få in. […] Det är så här vi måste göra att man liksom gör ett förslag och så arbetar man på det. Och då när man jobbar med det och man ser bristarna, och vad man kan göra vidare» (Faglærer, 30.04.09).*

Følgende to sitater eksemplifiserer utfordringer knyttet til valg og kategorisering av kriterier:

> *«Jeg synes at "innlevelse" under bolken "språk/tekst" egentlig ikke hører hjemme der. Synes også at "innlevelse" overlapper med "kommunikasjon" og "utstråling". Kanskje kunne vi gjøre et eneste punkt av dem alle og putte dem under "formidling"?» (Faglærer, 23.09.09).*

> *«Och presentation är en intressant sak som vi borde oppmuntra våra studenter att lägga mera vikt vid. Nu visste dom inte att vi skulle bedömma det. Hade dom vetat det, hade dom kanske presenterat sina sånger lite mer poetiskt, och gett lite mer introduktion till vad det är dom faktiskt sjunger om och inte bara ramsat upp en titel» (Faglærer, 30.04.09).*

Selv om kriteriene ble utviklet for bruk ved sluttvurdering av sangprestasjoner, viser sitatet foran at valg av kriterier også har læringsmessige implikasjoner. Kriterier knytter forbindelsen tilbake til læringsmål og til spørsmål om *hva* som skal læres, slik dette sitatet viser: «*Jeg savner titler som "klang/timbre" […] Burde vi ikke ha en linje for "koloratur/staccato", da dette er spesielle sangteknikker som krever ganske mye å lære?*» (Faglærer, NMH, 23.09.09).

Sammenliknet med tidligere versjoner, har den foreløpig siste utgaven av kriterieoppsettet (som vist i tabell 2) mindre vekt på teknikk til fordel for musikalske ferdigheter. Dette gjør imidlertid ikke utfordringene knyttet til vurdering enklere, fordi en kunstnerisk prestasjon forutsetter både talent (medfødt evne), ferdigheter og egenskaper, som for eksempel "utstråling". Ved vurdering inngår ikke bare talentet som en forutsetning for prestasjonen, men som en integrert del av den, og selv om kriteriesettet framstår som klart for bruk, gjenstår tolkningsmessige utfordringer. Begreper må forklares for å kunne fungere som analytiske verktøy. Derfor ble i alt 24 sentrale ord og begreper fra kriteriesettet definert/forklart i en "ordliste" som inneholder definisjoner og beskrivelser av ønskede kvaliteter i teknikk, musikalitet og formidling. Ordlista er utviklet av sanglærerkollegiet ved NMH studieåret 2008/2009. Følgende sitater illustrerer dette arbeidet:

> ***«Vokalegalitet***
> *I klassisk akustisk sang er jevnhet i klang et viktig kvalitetskriterium. I det sangtekniske arbeidet blir dette tatt opp i flere sammenhenger, særlig i forbindelse med egalisering av vokalene og i egalisering av stemmens registre. Kravet til vokalegalitet, jevnhet i klangstyrke og på den annen side kravet til tydelige vokaler, er ikke lett å forene. I sangpedagogikken er det litt ulike meninger om hvordan man teknisk skal løse vokalegaliseringen, bl.a. tungas bevegelsesmønster, hvordan ganeseilet skal innstilles, størrelse på kjeveåpning osv.*
>
> *Vokalegalitet er et svært viktig element i klassisk sang. Man skal etterstrebe en klang som både er klangrik og jevn, og samtidig slik at vokalfargene er tydelig nok til at tilhøreren oppfatter teksten som synges og at det blir korrekt språklig uttale.» (Bjørkøy, NMH, 15.01.10).*

«Registeregalisering/registeregalitet
Stemmen har fra naturens side et omfang på to til tre oktaver, og den har ulike klanglige kvaliteter i høyt og lavt leie og på svake og sterke toner. Synger man en tonerekke fra dypeste tone og oppover eller motsatt vei, vil stemmelyden i noen områder være stabil, men mellom disse stabile områdene vil stemmen lett "briste". Dette skyldes anatomiske og fysiologiske forhold i strupehodet, og bl.a. samspillet mellom strupen og pustefunksjonen. De stabile områdene i stemmen kalles register, områdene imellom kalles passasjer. Registrene benevnt nedenfra og oppover er brystregister, mellomregister, randregister og fløyteregister. Hvor registerovergangene er, hersker det en viss uenighet om, men det er samstemmighet om at registeregalisering er et viktig estetisk kjennetegn ved den klassiske sangstilen. En god registeregalisering innebærer minimale forskjeller på klangen i de ulike delene av stemmen.» (Bjørkøy, NMH, 15.01.10).

Selv om vi tror at analytisk vurdering kan benyttes i utøvende og kunstneriske disipliner, viser erfaringene fra dette prosjektet en kompleks utfordring. En faglærer ved NMH beskriver dette på følgende vis: «… veldig ofte så vil man jo oppleve i en vurdering at teknikken henger sammen med musikaliteten, og språk henger kanskje sammen med teknikken også, og med formidling.» Vurderingen kompliseres ytterligere ved at spørsmål om *nivå* framstår som essensielt for hvilken karakter kandidaten får. På masternivå blir "profesjonelle krav" lagt til grunn ved NMH, men i vårt tilfelle foreligger dette fortsatt kun som uartikulert, taus kunnskap.

Heller ikke for studenter i grunnutdanningen ble krav til nivå forsøkt presisert på annen måte enn karakterbeskrivelsene. Rent intuitivt kan man tenke seg at krav til nivå kan beskrives verbalt, men det spørs hvor langt strategien rekker. Et menneskeansikt lar seg beskrive ganske detaljert verbalt, men gjenkjennelse av det samme ansiktet for fremmede vil være svært krevende. En tegning eller et foto vil derimot lette gjenkjennelsen. På samme vis tror vi at gjenkjennelse av nivå (og kvalitet) forutsetter eksempler. Slik innsikt kan forstås som en evne til gjenkjennelse. Taus kunnskap er derfor vesentlig ved vurdering av kunstneriske prestasjoner, slik Polanyi har omtalt (Polanyi, 1958). En grunntanke hos Polanyi er at vi vet mer enn vi kan uttrykke med ord.

2.2 Diskusjon

Sangprestasjoner kan forstås som "a complex learning outcome" (Sadler, 1989), på samme vis som prestasjoner i språk, samfunnsfag, kunstfaglig utdanning m.fl. Et fellestrekk ved denne typen disipliner er *kompleksitet*, det vil si at verken læring eller vurdering lar seg sette på formel. Sadler (1989) hevder at vurdering her bygger på "qualitative judgments" med utgangspunkt i et stort antall implisitte kriterier som blir benyttet ut fra situasjon og behov. Kriteriene kan være "fuzzy" eller "sharp", og jo større behov for profesjonelt skjønn, desto høyere krav til erfaring.

Brødrene Dreyfus har beskrevet læringsprosessen fra "novise" til "ekspert" i fem stadier (Dreyfus & Dreyfus, 1991, 1999). I første stadium spiller regler og retningslinjer for arbeidet en helt sentral rolle, mens regelstyringen etter hvert avtar på mer avanserte stadier med stadig større vekt på intuitive beslutninger på bakgrunn av erfaring. Kompetanseutviklingen beveger seg altså fra et skille mellom teori og praksis for nybegynneren til en integrert forståelse av forholdet mellom teori/prinsipper og praksis hos eksperten. Dette representerer overgangen fra et strukturert til et ustrukturert felt, der handlingskompetansen endrer karakter med erfaring. Begrepet "teoriløs ekspertkompetanse" har vært brukt som forklaring på at eksperten kan være en mindre dyktig underviser sammenliknet med "den kyndige utøver", som kompetansemessig sett antas å befinne seg på et lavere nivå. Ekspertatferd er "intuitiv" i den forstand at individet handler ut fra "taus" kunnskap, altså egentlig uten at individet kan

redegjøre hva man gjorde og hvorfor. Ekspertens ekspertise består av «*… en begribelse af og ageren i forhold til situationens holistiske gestalt, genkendt på bakgrund av dennes likhed med gestalten i andre paradigmatiske situationer, som eksperten tideligere har været i*» (Bonderup Dohn, 2006, s. 41).

Representerer så den type ferdighetstilegnelse som fenomenologisk er beskrevet i fem trinn i Dreyfus' modell en adekvat form for kompetanse ved evaluering av studentprestasjoner? Den vekt som nå blir lagt på veldefinerte mål og kriterier kan synes som et paradoks dersom dette i liten eller ikke i det hele tatt blir brukt ved vurdering av studentprestasjoner. I forhold til ferske bedømmere/sensorer påpeker Bonderup Dohn at «*… novicens problem netop [er], at han/hun mangler den erfaringbaserede forståelse af,* hvad *reglerne utsiger, og* hvordan *man følger dem – en forståelse, som eksperten til gengæld har*» (Bonderup Dohn, 2006, s. 42). Noe tyder altså på at de verbale beskrivelsene av vurderingskriterier i seg selv ikke er tilstrekkelig. Som alternativ til brødrene Dreyfus sin modell, foreslår Bonderup Dohn det hun kaller "viden i praksis" som en kombinasjon av språklig formidlet viten og erfaringer og evne til å skille gode fra dårlige prestasjoner med utgangspunkt i erfaring. Evalueringsekspertise framstår etter dette som en annerledes type kompetanse sammenliknet med ferdigheter innen svømming eller sykling:

> *«Men i modsætning til Dreyfus skal intuitionen ikke sees som helt uden relation til sprogligt medieret viden. I stedet er sprogligt artikuleret viden en del af perspektivets baggrund. […] De ikke-sproglige aspekter af ekspertisen er en tavs resonansbund og meningsramme for det sproglige aspekt, der på sin side virker som fokuserende faktor» (Bonderup Dohn, 2006, s. 42).*

Fravær av eksplisitte kriterier ved vurdering av kunstneriske prestasjoner er ikke nødvendigvis å anse som en mangel ettersom kvalitativ vurdering (i hvert fall ideelt sett) spiller på et mer nyansert repertoar enn det som kan uttrykkes skriftlig. På den annen side bidrar språklig artikulert viten til fokusering mot sentrale aspekter i en kvalitativ prestasjonsvurdering.

På hvilken måte kan vi da begrunne at vårt arbeid med vurderingskriterier har verdi eller kan være til nytte? Sunn fornuft tilsier at bruken av kriterier først og fremst har verdi som bidrag til en pålitelig (reliabel) vurdering, altså at vurderingen blir den samme eller nesten den samme dersom den blir gjennomført av ulike, men kyndige personer. Til dette hevder Sadler (1989) at i læringsarbeid er den viktigste dimensjonen ved kriteriene knyttet til spørsmål om gyldighet (validitet), og at pålitelighet med nødvendighet følger av dette. Sett fra studentenes ståsted er det nærliggende å tenke at tydelige evalueringskriterier kunne være nyttig for egenvurdering og læring, men denne utfordringen er mer sammensatt enn som så. Som profesjonelle bedømmere trekker veksler på erfaringer og eksempler i kombinasjon med språklig formidlet viten, gis det grunn til å tro at ordene i seg selv ikke rekker langt for studentene. Kriteriene trenger å omsettes i eksempler, eller gestaltninger, som i sin tur kan fungere som referanse ved senere anledninger. Vurdering har også en erfaringsdimensjon som betyr at tid er en viktig forutsetning for læring.

I vårt tilfelle har vi sett at arbeidet med vurderingskriteriene har skapt en større felles plattform for hvilke viktige aspekter som ønskes vektlagt ved vurderingen. Selv om hvert kriterium ikke ble like aktivt brukt for hver kandidat, fungerte de i vårt tilfelle som en referanseramme for hvilke aspekter ved prestasjonen som skulle tas hensyn til. Faglærernes erfaringer ved NMH tyder på at en altfor "nærsynt" bruk av vurderingskriteriene neppe bidrar til en mer gyldig eller pålitelig vurdering. Å finne en god balanse mellom deler og helhet framsto som krevende.

> «*As most academics can attest, working systematically through the criteria one at a time would be highly labour intensive. Apart from the labour aspect, a significant reason for not operating in this way is that assessors are initially more interested in how the work comes together as a whole than in performance on individual criteria*» *(Sadler, 2009b, p. 165).*

Sammenhengen mellom analytisk og holistisk vurdering er komplisert. Sadler (2009) viser til at en prestasjon som samlet framstår som briljant, ikke nødvendigvis får topp score på hvert enkelt kriterium, hvilket logisk sett ville være en nødvendig betingelse. Og motsatt, at en prestasjon som får høy score på hvert kriterium, samlet sett framstår som middelmådig. I andre tilfeller kan det være problematisk å forklare forskjeller i prestasjon med utgangspunkt i de gitte kriteriene, altså at eksplisitte kriterier er utilstrekkelige til å sannsynliggjøre forskjeller i vurderingsuttrykk: «*… it lies in a positively identified criterion which was not included in the preset list distributed to students but turns out to be crucial to the judgment*» (Sadler, 2009b, p. 166).

Valg av kriterier og innbyrdes vekting er spørsmål som savner fasitsvar. Å velge betyr imidlertid også å velge bort, og sannsynligheten er liten for at nøyaktig de samme kriteriene blir valgt av ulike faglærere og sensorer. Eller – i den grad kriteriene ytre sett er (nesten) identiske – kan man neppe forvente verken helt sammenfallende forståelse eller identisk bruk av dem. Vi spør derfor hvilken effekt eksplisitte kriterier har, eller kan ha, i praksis. Ingen vurdering har forutsetninger for å bli fullstendig objektiv i en streng betydning av ordet, men kanskje har analytisk vurdering forutsetninger for å bli mer konsistent og dermed mer pålitelig sammenliknet med implisitt vurdering? Nå kan imidlertid vurderingen ytre sett framstå som reliabel i den forstand at sensorer gir samme karakter, men det betyr ikke at de har gitt identiske scorer for hvert kriterium. Som indikator på veiledningsbehov for studenter er sluttvurdering i seg selv utilstrekkelig: «*A corollary of this is that research into the reliability of analytic grading needs to go beyond the results it produces (the grades) if formative effectiveness is to be understood and promoted*» (Sadler, 2009b, p. 168).

Mens implisitt vurdering teoretisk sett kan dra veksler på et stort og raffinert sett av kriterier, er situasjonen annerledes ved analytisk vurdering. Praktiske hensyn tilsier at utvalg må til, noe som betyr at kriterier enten blir sammenslått eller utelatt. Hvilken løsning som blir valgt, rører ved vurderingens gyldighet (validitet), mens flertydige kriterier påvirker pålitelighet (reliabilitet). Vurdering krever også referanse til standarder (nivå) som enten kan beskrives, eksemplifiseres eller fastsettes relativt ut fra gruppens prestasjon. Det nye evalueringsrammeverket for sang, som vist i tabell 2, utgjør prinsipielt sett grunnlaget for det som i forskningslitteraturen blir omtalt som en ”rubrikk”, eller på engelsk ”rubric” (Stevens & Levi, 2004). Dette er et redskap som blir benyttet i stort omfang i mange land. Hensikten er å bidra til en raskere og mer rettferdig vurdering, samtidig som rubrikker kan brukes ved tilbakemelding til studentene. I vårt tilfelle er antall kriterier for stort til at verbale beskrivelser kan omtales i forhold til ulike nivåer. En forenkling, som vist i tabell 3, muliggjør verbale beskrivelser for hver kriteriekategori, men mange nyanser går i så fall tapt i forhold til det opprinnelige kriterisettet i tabell 2.

Tabell 3) Forenklet tabell for vurdering av sangprestasjoner (basert på tabell 2)

Karakterer	A = 5	B = 4	C = 3	D = 2	E = 1	F = 0
Teknikk						
Språk/tekst						
Musikalitet						
Formidling						
Repertoarvalg (for eksamenskonsert)						

Både tabell 2 og tabell 3 angir standard i form av en score fra 5–0 med direkte konvertering til bokstavkarakterer fra A–F (stryk). Dette reiser vesentlige spørsmål i forhold til potensiell bruk av disse tabellene. Med relevans til denne undersøkelsen hevder Bonderup Dohn følgende: «*De vurderer ikke først, hvilket kvalitativt nivå præstationen ligger på, hvorefter de tilskriver en kvantitativ karakter; i stedet tænker de nivåerne gennem skalaen*» (Bonderup Dohn, 2006). Nå har vi ikke data som forteller hvordan evalueringsarbeidet faktisk foregår i hvert enkelt tilfelle, men prinsipielt er karakterene et rapporteringsredskap som er lite egnet til å fange detaljene i en helhetlig prestasjon: «*Redskapet er ganske enkelt ikke nuanceret nok til at have denne rolle som perceptuelt utgangspunkt*» (Bonderup Dohn, 2006, s. 43). Nytt forsknings- og utviklingsarbeid vil vise om konvertering til karakter først bør skje på grunnlag av samlet score. Arbeid gjenstår også for å gjøre begrepet ”standard” bedre konseptuelt forståelig og mer operativt tilgjengelig. I denne rapporten er ikke denne utfordringen utdypet nærmere, men den fortjener grundig analyse i framtidig forskning, slik Sadler antyder i følgende sitat: «*Where standards are acknowledged as the key issue, common ways of specifying them are fundamentally inadequate*» (Sadler, 2003).

Utfordringer med å vurdere prestasjonsnivå er naturligvis ganske forskjellig i ulike disipliner. Noen disipliner er mer eksakte i sin struktur og innhold, noe som bidrar til at prestasjoner lettere lar seg måle og kvantifisere enn tilfellet er i kunstneriske og utøvende disipliner. Den prinsipielle tenkningen rundt evalueringsteoretiske spørsmål er derfor dels de samme, dels noe forskjellig. Den største ufordringen i utøvende disipliner består i at standarder ikke er direkte målbare. Vurdering og karaktersetting forutsetter derfor et tolkende mellomledd, eller profesjonelt skjønn. Dette skjønnet skal utøves med vekt på prestasjonen, men i vårt tilfelle er også ”utstråling” og ”scenisk fremstilling og iscenesettelse” oppgitt som kriterier. Fristelsen til å la seg påvirke av ikke-relevante forhold som utseende og påkledning er nærliggende. Tilsvarende utfordring er også observert under sensurmøter, der ikke-faglige hensyn iblant trekkes inn som argument, for eksempel sykdom og fravær i studiet, omtalt i litteraturen som ”infidelity” (Sadler, 2009a).

I dagligtale kan man lett få inntrykk av at ”kriteriebasert” vurdering er et rimelig entydig begrep, noe som slett ikke er tilfelle. Sadler illustrerer dette ved følgende fire modeller, som alle i en viss forstand er ”kriteriebaserte” (Sadler, 2005): ”Achievement of course objectives” (1); ”Overall achievement as measured by total scores” (2); ”Grades reflecting patterns of achievement” (3) og ”Specified qualitative criteria and attributes” (4). Den første modellen framstår intuitivt som logisk og forståelig, mens de øvrige punktene fortjener en kort forklaring: Som den andre modellen antyder, skjer vurdering og karaktersetting ofte med utgangspunkt i poengscorer uttrykt på en skala fra 0–100. Dette kan imidlertid ikke anses som kriteriebasert vurdering i streng betydning, fordi scoren ikke forteller noe absolutt om prestasjonen. Den tredje modellen framstår som et alternativ i de tilfeller der det settes

visse krav til flere eller alle deler av prestasjonen. En totalscore kamuflerer ofte både sterke og svake sider ved en prestasjon, slik at en kandidat kan bestå selv om enkelte deler av besvarelsen står til stryk. Den siste modellen kan eksemplifiseres med verbale nivåbeskrivelser uttrykt som karakterer fra A til F, med E som dårligste ståkarakter.

På nasjonalt nivå er vurderingsarbeidet blitt komplisert ved at karakterbeskrivelsene har prinsipielt ulik evalueringsteoretisk forankring. Ved NMH ble karakterbeskrivelsene i tillegg ansett som lite egnet som grunnlag for evaluering innen utøvende disipliner, noe som motiverte vårt arbeid med eksplisitte kriterier tilpasset behovene i sang som instrument. Selv om utvikling av kriterier i seg selv har vært en lærerik prosess, erfarte vi at analytisk vurdering i streng betydning er en svært krevende øvelse for faglærere, og neppe enklere for studenter, slik følgende sitat viser: «*Én ting er å skjønne ordene, å skjønne det med hodet sitt, noe annet er jo å gjøre det. Og det er forholdsvis lang vei fra at du kan forstå det, eller tror du forstår det, til du fikser det*» (Faglærer, NMH, 22.09.09). Den gode nyheten består i at kriterier og standarder kan vises via pedagogisk velvalgte eksempler, som demonstrasjon av spekteret av ulike prestasjoner:

> *«When multiple criteria are used in appraisals of quality, which is a common situation in higher education, a single case cannot constitute a standard, although it may exemplify it. Other examples relating to the same standard ordinarily may be expected to differ from one another» (Sadler, 2005, p. 192).*

Eksplisitte kriterier erstatter ikke det Bonderup Dohn omtaler som "viden i praksis", men kan trolig bidra til bedre fokusering ved vurdering. I utøvende disipliner kan bruken av eksempler støtte opp om og tydeliggjøre kriterier. Slik kan kompleks vurdering synliggjøres og prestasjoner gjenkjennes som "gestalt", eller helhet, av studentene. Denne tilnærmingen aktualiserer også et stort læringsfremmende potensial som ligger i selvvurdering som mulighet. Dersom kandidaten får bedre innsikt i hva kriterier og standarder betyr, kan det i sin tur danne grunnlag for selvforbedring som strategi (Gynnild, Holstad & Myrhaug, 2008; Zimmerman & Schunk, 2001).

I skrivende stund er de generelle karakterbeskrivelsene fra Universitets- og høgskolerådets tilleggsrapport (Glasser, 2009) ute på høring. I høringsnotatet sies imidlertid intet om hvilke prinsipper som ønskes lagt til grunn for tilbakemeldingen. I den grad diskusjonen kun dreier seg om formuleringer, uten forankring i en prinsipiell, evalueringsteoretisk referanse, er det grunn til å tro at institusjonenes forståelse av hva de selv gjør eller bør gjøre forblir uklar. Spørsmålet om *kriterier* handler grunnleggende sett om hva som skal vurderes, mens diskusjonen om *standard* har spesiell relevans i forhold til karakterfastsetting. Kvalifikasjonsrammeverket forutsetter at tilsiktet læringsresultat formuleres slik at det er mulig å konstatere om målet er nådd. Her legges grunnlaget for en kriteriebasert vurdering i en annen betydning enn det som til nå har rådet grunnen her til lands. Spørsmålet blir reist om hva studentene egentlig skal kunne, hvilke kunnskaper og ferdigheter de skal utvikle. Dette krever at karakterbeskrivelsene blir tydeligere med hensyn til evalueringsteoretisk referanse. Det vesentligste spørsmålet blir om studentene når de faglige krav som er satt, mens spørsmål om karakterfordelingens profil har mindre interesse.

Selv om denne rapporten fokuserer spesielt på vurdering av kunstneriske prestasjoner (sang), er det naturlig å parallellisere problemstillingene også til humanistiske og samfunnsvitenskapelige disipliner. Universitetsstudier på disse fagområdene forutsetter oftest eksamensprestasjoner i form av tekst (essay) eller tale (muntlig prøve). Ved vurdering – som ideelt sett forventes å gi et pålitelig og gyldig mål på studentens læring – står utfordringene i kø når uttrykte målkrav skal undersøkes systematisk i

konkrete vurderingsordninger. Læringsmålene er i mange tilfeller enten fraværende, utydelige, utilstrekkelige eller simpelthen lite egnet for eksaminasjon. Vurderingen forutsetter ofte et tilsvarende diffust skjønn som neppe kan forklares eller forstås fullt ut av studentene, og som i enkelte tilfeller bidrar til mange ubegrunnete klager på sensurvedtak.

2.3 Konklusjon

Studier av karakterfordelinger ved nasjonale institusjoner for utøvende musikkutdanning har dokumentert en "skeiv" fordelingsprofil til fordel for karakterene A, B og C. Dette harmonerer dårlig med normbaserte karakterfordelinger, som foreslått gjennom European Credit Transfer System (ECTS). Analysen viser at bruken av generelle karakterbeskrivelser ikke ble opplevd som hensiktsmessig i utøvende musikkutdanning. Karakterbeskrivelsene er flertydige, og bygger i praksis på to ulike evalueringsteoretiske prinsipper som vanskelig lar seg forene. Generelle, normbaserte retningslinjer for karakterfordeling står i motstrid til ønsket om kriteriebasert vurdering, noe som også inngår i Universitets- og høgskolerådets veiledninger til institusjonene.

Et forsknings- og utviklingsarbeid i forlengelsen av interne høringer og vedtak ved NMH har utvidet sangmiljøets innsikt i hva kriteriebasert vurdering *kan* bety – på godt og vondt. Å utvikle et felles rammeverk for analytisk vurdering var en krevende øvelse, men faglærerne mener det nye redskapet bidrar til en mer *konsistent* vurdering. I tillegg fungerer det som en felles referanse for undervisning og veiledning, spesielt på bachelornivå. Enhver analytisk vurdering er likevel isolert sett reduksjonistisk ved at den aldri evner å fange den kompleksitet og de subtile nyanser som alltid eksisterer i kunstneriske prestasjoner. For studentene kan analytisk vurdering være et steg på veien til å utvikle en selvstendig, implisitt vurdering som i sin tur danner grunnlag for selvdrevet læring og livslang utvikling i et utøvende yrke. Som læringstiltak vil trolig helheten i kunstneriske prestasjoner med hell kunne uttrykkes og erfares gjennom pedagogisk velvalgte eksempler som viser forskjellene mellom gode og mindre gode prestasjoner i praksis.

Undersøkelsen viser at et flertydig begrep som "kriteriebasert vurdering" verken er forklart eller eksemplifisert i de nasjonale retningslinjer for vurdering og karaktersetting som institusjonene er forventet å etterleve. Undersøkelsen viser også at de nasjonale karakterbeskrivelsene her til lands dels bygger på et *normbasert*, dels på et *kriteriebasert* grunnlag. Den tolkningen av kriteriebasert vurdering som er lagt til grunn for dette utviklingsarbeidet, gir i hvert fall i teorien et grunnlag for kommunikasjon mellom undervisere/bedømmere og studenter, og dermed for bedre læring. *Summativ* vurdering av kunstneriske prestasjoner er en kompleks utfordring som aldri lar seg sette på formel. Derfor er det også grunn til å minne om begrensninger som er beheftet ved vurderingsordninger som er analytiske i sin tilnærming, ikke minst i kunstneriske og utøvende disipliner der helhet og sammenheng i prestasjonen alltid framstår som et vesentlig element. Til sist faller det naturlig å spørre om "analytisk vurdering" i streng betydning er mulig, eller om dette er en begrepsmessig konstruksjon som savner rot i virkeligheten? Evaluering utføres av mennesker i konkrete situasjoner, og subjektivitet er en forutsetning for all vurdering.

Nettopp derfor oppstår mange spørsmål om hvordan profesjonell evalueringskompetanse kan utvikles og styrkes, slik at vurdering ikke framstår som tilfeldig, urimelig eller urettferdig. Slik kompetanse kan aldri bli "intuitiv" uten språklig mediert viten. Samtidig kan den neppe bli profesjonell uten faglig sosialisering og et arsenal av taus kunnskap. For bedømmere med liten eller ingen faglig ballast eller erfaring vil verbalt uttrykte kriterier i seg selv være utilstrekkelig. Vurdering av kunstneriske prestasjoner bygger alltid på et element av gjenkjennelse i forhold til "situationens holistiske gestalt"

(Bonderup Dohn, 2006) som grunnlag for sammenlikning og karakterfastsetting. Læringsmessig sett kan det være mye å hente der refleksjon over kriterier og standarder i forbindelse med vurdering løper parallelt med faglig sosialisering og erfaring.

Erkjentlighet:
Rapporten er et resultat av et utviklingsarbeid i samarbeid mellom forfatteren og faglærere i det sangfaglige miljøet ved Norges musikkhøgskole. En takk rettes derfor til følgende kolleger som har bidratt med kommentarer og innspill: Folke Bengtsson, Svein Bjørkøy, Håkan Hagegård, Mona Julsrud, Barbro Marklund-Pêtersone og Kirsten Taranger. En spesiell takk til professor Svein Bjørkøy, som har ledet interne drøftinger om valg og utforming av vurderingskriterier. Han har også skrevet det reviderte kriteriesettet og "ordboka". Forfatteren gjennomførte prosjektet i bistilling som seniorrådgiver ved NOKUTs Utrednings- og analyseavdeling i 2008-2010.

Referanser

Biggs, J., & Tang, C. (2007). *Teaching for Quality Learning at University. What the Student Does* (3rd ed.). Berkshire & New York: Society for Research into Higher Education & Open University Press.

Bonderup Dohn, N. (2006). Karaktergivning - intuitiv ekspertise eller 'viden i praksis'. *Dansk Universitetspedagogisk Tidsskrift, 1*, 38-46.

Commission, E. (2009). ECTS Users' Guide Available from http://ec.europa.eu/education/lifelong-learning-policy/doc/ects/guide_en.pdf

Dreyfus, H., & Dreyfus, S. (1991). *Intuitiv ekspertise: den bristede drøm om tænkende maskiner*. [København]: Munksgaard.

Dreyfus, H., & Dreyfus, S. (1999). Mesterlære og ekstpertens læring. In K. Nielsen & S. Kvale (Eds.), *Mesterlære-læring som social praksis*. København: Reitzels Forlag A/S.

Flyvbjerg, B. (2010). Fem misforståelser om case-studiet. In L. Tanggaard & S. Brinkmann (Eds.), *Kvalitative metoder*. København: Hans Reitzels Forlag.

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., Holstad, A., & Myrhaug, D. (2008). Identifying and Promoting Self-Regulated Learning in Higher Education: Roles and Responsibilities of Student Tutors. *Mentoring & Tutoring: Partnership in Learning, 16*(2), 147-161.

Polanyi, M. (1958). *Personal knowledge*. London: Routledge and Kegan Paul.

Price, M. (2005). Assessment standards: the role of communities of practice and the scholarship of assessment. *Assessment & Evaluation in Higher Education, 30*(3), 215 - 230.

Sadler, D. R. (1989). Formative Assessment and the Design of Instructional Systems. *Instructional Science, 18*(2), 119-144.

Sadler, D. R. (2003). *How criteria-based grading misses the point?* Paper presented at the ETL Conference.

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. (2009a). Fidelity as a precondition for integrity in grading academic achievement. *Assessment & Evaluation in Higher Education*.

Sadler, D. R. (2009b). Indeterminacy in the use of preset criteria for assessment and grading. *Assessment & Evaluation in Higher Education, 34*(2), 159 - 179.

Sadler, D. R. (2010). Assessment in higher education. In M. Baker, B. & Peterson, P. (Ed.), *International encyclopedia of education*. Oxford: Elsevier.

St.meld. nr. 27 (2000-2001) Gjør din plikt - Krev din rett (2001). Retrieved from http://www.regjeringen.no/Rpub/STM/20002001/027/PDFA/STM200020010027000DDDPDFA.pdf.

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.

Wrigley, W. J. (2005). *Improving music performance assessment.* Griffith University, Brisbane.

Yorke, M., Bridges, P., & Woolf, H. (2000). Mark Distributions and Marking Practices in UK Higher Education: Some Challenging Issues. *Active Learning in Higher Education, 1*(1), 7-27.

Zimmerman, B. J., & Schunk, D. H. (2001). *Self-regulated learning and academic achievement: theoretical perspectives* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum.

Zuber-Skerritt, O. (2002). The concept of action learning. *The Learning Organization, 9*(3), 114-124.