Hva skal vurderes og hvordan?

Utfordringer på muntlig engelsk eksamen i videregående skole

Henrik Bøhn, Høgskolen i Sørøst-Norge

Et avgjørende prinsipp i sluttvurderingen er at den skal være lik og rettferdig. For å oppnå dette er det en forutsetning at de som vurderer har samme oppfatning av hva som skal vurderes og hvordan. Internasjonal forskning på testing av muntlige ferdigheter i andrespråkssammenheng har vist til dels stor variasjon i hvilke kriterier som legges til grunn. I Norge har det vært gjort lite forskning på dette området, men i mitt doktorgradsprosjekt har jeg sett spesifikt på denne problemstillingen. I denne artikkelen presenterer jeg hovedfunnene fra prosjektet og drøfter hvilke tiltak som eventuelt kan iverksettes for å forbedre slike sluttvurderingspraksiser.

Hva skal vurderes og hvordan kan dette gjøres mest mulig likt og rettferdig? Ifølge vurderingsforskriften er det kompetansemålene i læreplanen som danner grunnlaget for vurderingen. For å kunne vurdere i hvilken grad elevene har nådd målene, kan det være hensiktsmessig å benytte en vurderingsmatrise med kjennetegn på måloppnåelse. I vurderingen på skriftlig eksamen er slike matriser utarbeidet av Utdanningsdirektoratet og gjelder på nasjonalt nivå, og sensorene er pålagt å benytte dem (Udir, 2016b). Når det gjelder muntlig eksamen er slike vurderingsmatriser enten veiledende på nasjonalt nivå, som matrisen for standpunktvurdering i engelsk på 10. trinn (Udir, 2016a), eller de er utarbeidet på lokalt nivå, som i videregående skole, hvor det eksisterer et stort antall forskjellige matriser. Det betyr altså at sensorer ikke har noen felles skriftlig standard for vurdering på muntlig eksamen. I internasjonal språktesting er dette svært uvanlig. Matriser er sett på som et særdeles nyttig verktøy for å sikre mest mulig lik og rettferdig vurdering. Men som jeg skal komme tilbake til, vil bruk av slike matriser også kunne ha noen uheldige konsekvenser.

> ***Matriser er sett på som et særdeles nyttig verktøy for å sikre mest mulig lik og rettferdig vurdering***

Doktorgradsprosjekt om norske læreres vurdering av muntlig engelsk. I mitt doktorgradsprosjekt har jeg sett på hvordan lærere i videregående skole vurderer elevers prestasjoner på muntlig eksamen i engelsk fellesfag. Hovedfokuset har vært på hvilke kriterier som legges til grunn. I tillegg har jeg sett på karaktergiving, samt elementer av samsvar mellom lærernes vektlegging og hva som skal vurderes i henhold til læreplanen og tilhørende rundskriv. Prosjektet har bestått av tre delstudier. Totalt 80 lærere deltok i undersøkelsen. Samtlige fikk se en videosekvens av en elev i en autentisk eksamenssituasjon og ble deretter spurt om ulike aspekter ved elevens prestasjon.

Hovedfunn fra undersøkelsen.
Delstudie 1.
I den første delstudien ble 24 lærere intervjuet om kriterier generelt. Resultatene viste at de i stor grad la samme kriterier til grunn i vurderingen. Ikke overraskende var lærerne mest opptatt av språklige trekk ved elevenes kompetanse, som for eksempel uttale, grammatikk og flyt. Aspekter ved innhold ble også trukket fram av et flertall av informantene.

Variasjonen i lærernes responser var særlig knyttet til uttale- og innholdskriteriene. Blant annet var det uenighet om hvorvidt elevene bør vurderes utfra en uttalenorm for morsmålsbrukere, dvs. om elevene må snakke med tilnærmet britisk, amerikansk eller annen morsmålsbrukeruttale for å oppnå toppkarakter. Det var også variasjon i vektingen av innholdskriteriet, i hovedsak definert som elevenes evne til å «besvare oppgaven», samt deres evne til å «reflektere over tema». Funnene viste at lærerne som stort sett underviste på program for studiespesialisering, hadde en tendens til å legge mer vekt på innhold enn lærerne på yrkesfaglige utdanningsprogrammer. For øvrig framstod innholdskriteriet som noe uklart, særlig med hensyn til hvilke temaer lærerne mente var viktige i eksamenssituasjonen.

> ***Tre av lærerne ga karakteren 2, femten stykker ga karakteren 3, mens seks lærere ga karakteren 4 […]. Variasjonen i karaktersettingen var først og fremst knyttet til lærernes programtilhørighet***

Videre ble informantene bedt om å sette karakter på prestasjonen vist i videosekvensen. Tre av lærerne ga karakteren 2, femten stykker ga karakteren 3, mens seks lærere ga karakteren 4. Dette må kunne sies å være akseptabelt, siden de ikke hadde noen felles vurderingsmatrise. Variasjonen i karaktersettingen var først og fremst knyttet til lærernes programtilhørighet. Lærerne som i hovedsak underviste på studiespesialisering var gjennomgående strengere enn lærerne som underviste på yrkesfag.

Et siste funn det er verdt å nevne, er at felles forståelse av vurderingskriterier ikke automatisk betyr at lærerne vurderer prestasjoner

likt. Det er også viktig at de er enige i hvordan en prestasjon skal nivåplasseres på karakterskalaen. For eksempel var lærerne enige om at «flyt» var et viktig vurderingskriterium. I karaktersettingen vurderte imidlertid noen elevens flyt som god, andre som dårlig.

Delstudie 2

I den andre delstudien ble intervjuer og spørreskjema brukt for å undersøke til sammen 70 læreres tilnærming til aspekter ved uttale, her forstått som produksjon av enkeltlyder, intonasjon, trykk, rytme osv. Uttale er et relevant forskningsområde i vurderingssammenheng ettersom internasjonal forskning indikerer at det lenge har vært neglisjert felt (se f.eks. Isaacs, 2014). Dette er interessant i en norsk kontekst, ettersom uttale ikke engang ble nevnt i læreplanen for fellesfag engelsk i videregående skole før den ble revidert i 2013. Etter revisjonen sier læreplanen kun at «eleven skal kunne […] bruke mønstre for uttale, intonasjon, ordbøying og varierte setningstyper i kommunikasjon». Her kan man spørre seg hva mønstre for uttale og intonasjon faktisk innebærer.

For øvrig hevder den amerikanske lingvisten Richard Levis at undervisningen med hensyn til uttale i engelskopplæringen har vært styrt av to motstridende prinsipper: Det ene har hatt som mål at eleven skal kunne snakke som en morsmålsbruker (såkalt «nativeness»); det andre har kun hatt som mål at eleven skal kunne gjøre seg forstått (såkalt «intelligibility») (Levis, 2005). Internasjonale forskere på feltet er i dag overveiende enige om at det er forståelse som bør være det overordnede målet (se f.eks. Munro & Derwing, 2015), og det har derfor i den senere tid blitt gjennomført studier som har sett på hvilke aspekter ved uttale som er mer eller mindre viktige for forståelse. Denne forskningen er imidlertid ikke spesielt omfattende og generaliserbar. Allikevel synes forskerne å være enige om at følgende uttaleaspekter er viktige for forståelse:

1. Enkeltlyder: Særlig konsonantlyder, men også vokallyd-kontraster (f.eks. skillet mellom kort og lang vokal i *pitch* og *peach*)
2. Trykk i enkeltord (f.eks. *SEcond* og ikke *seCOND*)
3. Trykk i setninger (f.eks. *I work more than YOU do*, ikke *I work more than you DO*).

Når det gjelder andre aspekter ved uttale, er forskningen mindre entydig. En god del fonetikere hevder for eksempel at intonasjon er viktig for kommunikasjon, men en del studier har vist at dette aspektet har lite å si for at budskapet skal bli forstått (se f.eks. Jenkins, 2009).

På bakgrunn av denne forskningen valgte jeg å spørre lærerne om hvorvidt de oppfattet disse fire uttaleaspektene som viktige kriterier for å vurdere elevers uttale. I tillegg ble de spurt om hvorvidt tilnærmet morsmålsbrukeruttale er nødvendig for at elevene skal oppnå høyeste karakter. Funnene viste at lærerne i moderat til stor grad var opptatt av de første tre punktene nevnt ovenfor. Intonasjon, derimot, var de gjennomgående mindre opptatt av eller usikre på. Hva gjelder morsmålsbrukeruttale, var lærerne tydelig uenige. Seks av de 46 respondentene som svarte på spørreskjemaet, mente at elevene absolutt ikke burde vurderes i forhold til en morsmålsbrukernorm, mens sju av 46 mente at de burde bli det.

Delstudie 3

I den tredje delstudien ble 10 nye lærere intervjuet om hvordan de tolket innholdskriteriet. Funnene viste at lærerne i stor grad forstod innhold som et todimensjonalt vurderingskriterium i tråd med Blooms reviderte taksonomi for læringsmål (Anderson & Kratwohl, 2001). Det betyr at innhold kan deles inn i en kunnskapsdimensjon (hva), og en kognitiv prosessdimensjon (hvordan). Dette er for øvrig i overensstemmelse med hvordan kompetansemålene i læreplanen er oppbygd. Følgende eksempel illustrerer dette: «Målet for opplæringen er at eleven skal kunne […] drøfte framveksten av engelsk som et verdensspråk». Her utgjør substantivfrasen «framveksten av engelsk som et verdensspråk» kunnskapsdimensjonen, mens verbfrasen «drøfte» utgjør den kognitive prosessdimensjonen. Det som gjør denne taksonomien særlig relevant som modell for vurdering av innhold, er at de to dimensjonene er hierarkisk ordnet, fra det enkle til det avanserte. Dette gjelder særlig den kognitive prosessdimensjonen. En slik inndeling egner seg godt for vurdering av elevenes prestasjoner på forskjellige nivåer. I kombinasjon kan de to innholdsdimensjonene visualiseres med følgende matrise (nevnte kompetansemål er satt inn for å illustrere):

Den kognitive prosessdimesjonen

Kunnskapsdimensjonen	Huske	Forstå	Anvende	Analysere	Evaluere	Skape
Fakta-kunnskap				Drøfte framveksten av engelsk som et verdensspråk		
Begreps-kunnskap						
Prosedyre-kunnskap						
Metakognitiv kunnskap						

Blooms reviderte taksonomi for læringsmål

Det som er spesielt interessant i denne sammenhengen er at den kognitive prosessdimensjonen synes å være viktigere for lærerne enn kunnskapsdimensjonen. Med andre ord er det ikke så viktig hva elevene bringer inn i diskusjonen, men hvordan de gjør det – at de er i stand til å analysere, reflektere og evaluere. Dette er trolig en konsekvens av at læreplanen er svært omfattende, og at den i liten grad krever kjennskap til detaljerte faktakunnskaper. Dermed ser det ut til at lærerne blir mindre opptatt av elevenes kjennskap til spesifikke temaer og mer fokusert på elevenes evne til å analysere og reflektere over hva nå enn temaene for eksamen måtte være. Som en lærer formulerte det: «Jeg tenker at vi av og til måler litt sånn generell modning […] mer en type generell intelligens eller generell kunnskap som kanskje ikke alltid er så knyttet til engelskfaget».

Her må det også nevnes at på spørsmål om hva lærerne faktisk oppfattet som relevante temaer for vurderingen (kunnskapsdimensjonen), var det ikke alle kompetansemålene de fant like viktige. De var enige om at temaene som nevnes under hovedområdet *Kultur, samfunn og litteratur* var sentrale, men de var mer usikre når det gjaldt for eksempel målet som omhandler læringsstrategier, dvs. «vurdere og bruke […] læringsstrategier for å videreutvikle egne ferdigheter i engelsk». Som en lærer sa det: «Nei, det er en metakunnskap og skal ikke testes». Det at slike temaer blir forbigått av noen lærere, men fremhevet av andre, er naturlig nok uheldig. [1]

Oppsummering av hovedfunn

Det er viktig å understreke at antall lærerinformanter ikke utgjør noe representativt utvalg. Resultatene kan derfor ikke uten videre generaliseres til sluttvurdering i muntlig engelsk allment i Norge. Allikevel peker funnene på en del interessante forhold som det er verdt å diskutere nærmere. Først er det imidlertid viktig å presisere at lærerne i studien i stor grad hadde et sammenfallende syn på hvordan muntlige prestasjoner skal vurderes på eksamen. De var i hovedsak enige om hvilke kriterier som skal gjelde, og deres læreplanfortolkning var i rimelig grad i samsvar med det kompetansemålene angir. Det borger godt for kvaliteten til denne typen sluttvurdering. Men som funnene ellers indikerer, er det også problemområder som bør undersøkes nærmere. Dette gjelder først og fremst fire forhold:

- vurdering av uttale
- vurdering av innhold
- nivåfastsetting av elevenes prestasjoner
- tendensen til at programtilhørighet påvirker karakterfastsettingen

Hvordan kan engelsk muntlig eksamen gjøres mer lik og rettferdig? Når det gjelder variasjonene i lærernes tilnærming til uttale, særlig med tanke på morsmålsbrukeruttale og intonasjon, er det fullt forståelig at disse bidrar til forvirring. Læreplanen er fortsatt vag når det gjelder uttale, og pedagogisk praksis synes å ha vært usystematisk på dette punktet. Prinsippene om «nativeness» og «intelligibility» lever side om side med tilhengere på hver sin kant. Dessuten er forskningen som nevnt ikke entydig når det gjelder hvilke fonologiske trekk som er viktige for forståelse. For å høyne kvaliteten på denne typen vurderingspraksis bør derfor utdanningsmyndighetene bidra til å tydeliggjøre kompetansemålet som gjelder uttale, slik at det gir mindre grunnlag for individuelle fortolkninger. Her må imidlertid også lingvistene på banen for å klargjøre forskjellen på «nativeness» og «intelligibility» og bidra med mer forskning på hvilke fonologiske elementer som er avgjørende i kommunikasjon. Sist, men ikke minst, må det fortsatt legges til rette for gode tolkningsfellesskap for å oppnå en felles forståelse av vurdering av uttale blant lærerne.

Når det gjelder vurdering av innhold, tyder funnene på at kunnskapsdimensjonen ikke er klart nok definert for en del sensorer. Det at enkelte avviser at noen kompetansemål skal kunne bedømmes er et problem for vurderingens gyldighet. I tillegg er variasjonene som gjelder vektingen av innholdskriteriet en utfordring. Når enkelte lærere legger stor vekt på elevenes evne til å reflektere og svare på oppgaven, mens andre toner dette ned, er det problematisk. En mulig løsning på begge disse problemene kan være å innføre nasjonale vurderingsmatriser.

Når det gjelder nivåfastsettingen av elevenes prestasjoner i forhold til kriteriene, er dette et område som er velkjent i vurderingslitteraturen (se f.eks. Eckes, 2009). I min studie var for eksempel lærerne uenige om hvorvidt eleven hadde god eller dårlig flyt. Slik variasjon i bedømmelsen vil imidlertid kunne kompenseres gjennom sensorskolering og utvikling av tolkningsfellesskap, der for eksempel videoklipp av elever i autentiske eller ikke-autentiske eksamenssituasjoner brukes for å eksemplifisere prestasjoner på de forskjellige nivåene. En slik praksis vil tilsvare bruk av eksempeltekster til skriftlig eksamen.[2]

Til sist må funnet som gjelder variasjon knyttet til programtilhørighet nevnes. Tendensen til at lærerne i program for studiespesialisering var generelt strengere enn lærerne i de yrkesfaglige utdanningsprogrammene er problematisk med hensyn til rettferdige karakterutfall. Det kan sikkert være mange grunner til at det blir slik. En sannsynlig årsak er at sensorer lar seg påvirke av det generelle prestasjonsnivået til elevene de vanligvis underviser. Det vil i praksis si at sensorene sammenligner prestasjoner med det de er vant til. En mulig løsning på dette problemet vil være sensorskolering med videoeksempler for å utvikle en felles tolkningsforståelse av hvilke prestasjoner som skal gis hvilke karakterer.

Avsluttende betraktninger. En sluttvurderingspraksis der sensorer ikke har felles vurderingsmatriser og der sensorskolering er til dels fraværende, kan virke underlig i et internasjonalt testperspektiv. Dette inntrykket forsterkes ved at det for skriftlig eksamen eksisterer tydelige nasjonale føringer. Det er rimelig stor konsensus i testlitteraturen om at felles vurderingsmatriser og systematisk sensorskolering i de aller fleste tilfeller vil høyne kvaliteten på vurderingen (se f.eks. Fulcher, 2012; Taylor & Galaczi, 2011).

> ***Det er rimelig stor konsensus i test-litteraturen om at felles vurderingsmatriser og systematisk sensorskolering i de aller fleste tilfeller vil høyne kvaliteten på vurderingen***

Her må det imidlertid understrekes at tradisjonell språktesting i noen grad skiller seg fra vurdering i språkundervisningen fordi testing ofte ikke involverer noen undervisningskomponent. I språkopplæring derimot er vurdering, undervisning og læring tett sammenvevd. Her hjemme må lærere forholde seg til en svært omfattende læreplan som få vurderingsmatriser vil kunne reflektere fullt ut. Nasjonalt gitte vurderingsmatriser vil da ha den svakheten at de vil kunne overforenkle læreplanens mål. Som andre har påpekt, vil et for sterkt fokus på kriterier og kjennetegn kunne føre til at de «skaper en ny struktur i faget» (Throndsen m.fl., 2009: 109). En slik utvikling er uheldig fordi den kan føre til at viktige aspekter ved fagets egenart overses. I lys av disse innvendingene synes Throndsen m.fl. å ha et godt poeng når de anbefaler at kjennetegn på måloppnåelse innføres på nasjonalt nivå, men at de gjøres veiledende, slik tilfellet er for muntlig standpunktvurdering i engelsk på 10. trinn.

Til slutt vil jeg igjen få trekke fram poenget med bruk av videofilmede elevprestasjoner i sensorskoleringen. For at slik skolering skal kunne styrke tolkningsfellesskapet er det et poeng at den gjøres til et nasjonalt anliggende. Det vil innebære at alle har tilgang til de samme eksempelvideoene. Naturlig nok vil dette gi noen praktiske utfordringer med hensyn til blant annet personvern. Hvordan slike utfordringer konkret skal løses har ikke vært vurdert i denne studien, men som denne gjennomgangen har vist, er det flere gode grunner til at et slikt tiltak bør vurderes.

Referanser

Anderson, L. W., & Kratwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives.* New York: Longman.

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. I: A. Brown & K. Hill (Red.), *Tasks and criteria in performance assessment: Proceedings of the 28th Language Testing Research Colloquium.* Frankfurt: Peter Lang, Vol. 13.

Fulcher, G. (2012). Scoring performance tests. I: G. Fulcher & F. Davidson (Red.), *The Routledge Handbook of Language Testing* (s. 378-392). Oxford: Routledge.

Isaacs, T. (2014). Assessing pronunciation. I: A. J. Kunnan (Red.), *The companion to language assessment,* Vol. 1, s. 140-155.

Jenkins, J. (2009). (Un)pleasant? (In)correct? (Un) intelligible? ELF speakers' perceptions of their accents. I: A. Mauranen & E. Ranta (Red.), *English as a Lingua Franca: Studies and findings* (s. 10-36). Newcastle: Cambridge Scholars Publishing.

Levis, R. (2005). Changing contexts and shifting paradigms in pronunciation teaching. *TESOL Quarterly*, Nr. 39(3), s. 369-377.

Munro, M. J., & Derwing, T. (2015). Intelligibility in research and practice: Teaching priorities. I: M. Reed & J. M. Levis (Red.), *The handbook of English pronunciation* (s. 377-398). Chichester: Wiley-Blackwell.

Taylor, L., & Galaczi, E. (2011). Scoring Validity. I: L. Taylor (Red.), *Examining speaking: Research and practice in assessing second language speaking*, Vol. 30, s. 171-233.

Throndsen, I., Hopfenbeck, T. N., Lie, S., & Dale, E. L. (2009). *Bedre vurdering for læring: Rapport fra "Evaluering av modeller for kjennetegn på måloppnåelse i fag.*

Kunnskapsdepartementet (KD). (2013 [2006]). *Læreplan i engelsk. I: Læreplanverket for Kunnskapsløftet (LK06)*. Oslo: Utdanningsdirektoratet.

Utdanningsdirektoratet (Udir). (2016a). Engelsk: Kjenneteikn på måloppnåing. Oslo: Utdanningsdirketoratet.

Utdanningsdirektoratet (Udir). (2016b). Vurderingsveiledning: Om vurdering av eksamensbesvarelser 2016. Oslo: Utdanningsdirektoratet.

Fotnoter

[1]Se for eksempel vurderingsmatrisen for Sør-Trøndelag fylkeskommune, der læringsstrategier er inkludert: https://www.stfk.no/Tjenester/opplaring/privatist/Vurderingskriterier/

[2]Eksamensbesvarelser med begrunnelse for karakteren - videregående: http://www.udir.no/eksamen-og-prover/eksamen/Eksamensbesvarelser-med-begrunnelse-for-karakteren/