Notat fra besøg ved Det medisinske Fakultet, UiO

Den 8. – 9. marts 2016 gennemførte Det medisinske Fakultet og jeg et tilsynsbesøg, hvor Ingrid Os, Kristin Wium, Per Grøttum, Krtistian Bjøro og jeg drøftede studiets prøvesystem, digital eksamen, psykometriske målinger og graduerede karakterer. Jeg mødte desuden uddannelsesleder Knut Lundin, Jarle Breivik og Elin Rosvold og fik lejlighed til at drøfte deres udfordringer og overvejelser. Derudover mødtes jeg med modulleder Erik Dissen og drøftede specifikt blokk3-eksamen i modul 1.

Generel konklusion: Jeg er generelt meget imponeret over fakultetets arbejde med at udvikle studiets prøver. Der arbejdes systematisk og visionært med kvalitetssikringen.

Konklusion til fakultetet angående yderligere kvalitetssikring af items i de online multiple choice prøver:
Fakultetet udarbejder løbende glimrende eksamensrapporter med solide psykometriske analyser som modulerne kan støtte sig til. Der kan læres meget af at gennemgå sættet med spørgsmålsstillere, eksamensudvalg, en psykometriker og en pædagogisk konsulent, for rapporten karakteriserer og identificerer de fleste problematiske items. Det kan dreje sig om items, der er alt for lette, eller items, som ikke skelner mellem de dygtigste og de dårligste studerende. Ved gennemgang af disse items kan man konstatere, at der som regel er et 'teknisk' problem – det vil sige, at items bør formuleres bedre.
Jeg vil klart anbefale fakultetet, at man derudover arbejder sig hen mod at følge guidelines fra National Board of Medical Examiners, der aktuelt sidder inde med en internationalt førende ekspertise om multiple choice prøver. NBME fraråder klart, at man bruger False/True formatet. De anbefaler, at man primært bruger formaterne One Best Answer samt Extended Matching. Man kan blandt andet finde deres guidelines her: http://www.nbme.org/pdf/itemwriting_2003/2003iwgwhole.pdf. Endelig vil jeg gøre opmærksom på, at NBME og andre internationale aktører har produceret lavet glimrende og overskuelige guidelines, som alle item skrivere kan støtte sig til, og som fakultetet kunne vælge at oversætte til norsk og distribuere systematisk. Man bør desuden sikre sig at en passende andel af item skriverne og eksamensudvalgene i de enkelte moduler har god indsigt i One best Answer og Extended Matching formatet og kan bidrage til kvalitetssikringen af de enkelte items og sættet. Der er også produceret undervisningsmaterialer og online kurser om item skrivning i de internationale miljøer, fx en 45 minutters online tutorial fra NBME (http://www.nbme.org/IWW/), som man kan overveje at bruge til en sådan opkvalificering. Fakultetet kan naturligvis også vælge at lade sine egne dygtige pædagogiske konsulenter varetage opgaven.

Specifik om blokk3-eksamen i modul 1:

Som et led i tilsynsbesøget mødtes jeg med modulleder Erik Dissen og drøftede blokk3-eksamen i modul 1. Den resterende del af dette notat handler specifikt om de udfordringer, som blokk3 i modul 1 stod med i forbindelse med tilsynet, men nedenstående overvejelser er naturligvis gældende for alle eksaminer med samme problemstilling.

Drøftelsen tog afsæt i en gennemlæsning af dokumenterne

1) Oslo 2014: studieplan for profesjonsstudiet i medisin

2) Eksamenskvalitet og innføring av gradert karakterskala ved profesjonsstudiet i medisin, rapport fra arbeidsgruppen til styringsgruppen for Oslo 2014, februar 2016
3) MED1100 – Medisinstudiet, modul 1
4) Læringsmål og læringsudbytte i modul 1
5) Eksamensbesvarelse Eksamen: MED1100-3 H15 ORD
6) Exam report MED1100-3-H15-ORD, December 15, 2015
7) Eksamen MED1100-3 H15 ORD, Konsekvenser for primærsensuren av de psykometriske analyser av resultaterne
8) Evaluering af modul 1, høsten 2015 (1)

Derudover supplerede Erik Dissen med yderligere informationer og et indblik i de bagvedliggende overvejelser, som han og modulets øvrige ledere og undervisere har gjort sig.

Det er en kompleks opgave at planlægge, gennemføre og evaluere eksaminer og man må derfor i kvalitetsarbejdet udvælge de parametre, der lægges vægt på. I dette notat har jeg særligt kigget på graden af overensstemmelse mellem læringsudbyttebeskrivelsen og den konkrete prøve, prøvens bredde og højde, samt de enkelte items relevans og tekniske kvalitet. Alle disse parametre har stor betydning for prøvens samlede validitet.

<u>Sammenhæng mellem læringsudbyttebeskrivelsen og den konkrete prøve</u>

Blok 3 består af 4 fag (immunologi, anatomi, fysiologi og mikrobiologi) samt nomenklaturlære. Blokken afvikles over 10 uger med 8 pbl-møder, 4 histologikurser, 10 dissektions-, overflade- og PC kurser, samt 9 laboratoriekurser.

Af læringsudbyttebeskrivelsen fremgår det, at blokken primært har vægt på kundskab med 4 x 'kunne beskrive', 35 x 'kunne gøre rede for' og 6 x 'kunne forklare'. Der er ikke beskrevet læringsudbytte på højere taksonomiske niveauer. En skriftlig videnstest med multiple choice som den anvendte er derfor meget velegnet til at teste læringsudbyttebeskrivelsens kundskabsdel.

Udover kundskaber har modulet 15 x færdighedsmål. Mange af disse beskrevne færdigheder har klare træk af kundskab og vil fint kunne testes med en skriftlig prøve, inklusiv en digital multiple choice. Det gælder fx at genkende lymfatiske væv og ræsonnere over sammenhængen mellem opbygning og roller i immunsystemet, at genkende anatomiske strukturer i billedfremstillinger og ræsonnere over hvorvidt disse ser normale ud, eller at ræsonnere over sammenhænge mellem struktur og funktion. Men andre af de beskrevne færdighedsmål har helt klare træk af praktisk færdighed og kan ikke kan testes i en skriftlig prøve, men må testes på anden vis. Det drejer sig for eksempel om praktiske færdigheder som at kunne udføre enkle antistofbaserede immunologiske laboratorieanalyser, at orientere og beskrive makroskopiske vådpræparater, eller at udføre kapillær og venøs blodprøvetagning.

Endelig er der beskrevet seks generelle kompetencemål, som fx at forstå hvordan immunsystemet kan bekæmpe smittestoffer og forårsage sygdomsmanifestationer, kritisk at vurdere udsagn om menneskekroppens normale funktion, eller se menneskets afhængighed af jordens økosystem. De fleste af disse har et klart kundskabspræg og ville kunne testes i en digital skriftlig eksamen. Enkelte, som fx at vise respekt for menneskeliv og døde legemer må vurderes på anden vis.

Jeg anbefaler derfor, at man i blokken diskuterer, hvordan de beskrevne væsentlige praktiske færdigheder, som ikke kan testes i den skriftlig prøve, kunne vurderes opnået på anden vis.

Prøvens bredde og højde

Her har jeg analyseret det eksamenssæt, der blev givet i ved ordinær eksamen H15. Sættet består af 4 opgavesamlinger med i alt 102 spørgsmål. Sættet vidner om, at der omhyggeligt er samplet bredt i fag og emner, så læringsudbyttebeskrivelsens kundskabsmål er dækket i bredden. Virkeligt godt kvalitetsarbejde med breddedækningen.

Langt de fleste items tester faktuel viden og er dermed i overensstemmelse med læringsudbyttebeskrivelsen, som også har et overvejede fokus på om den studerende kan 'beskrive' og 'redegøre for' sin kundskab. Selvom der er et mindre fokus på højere relationelle mål i læringsudbyttebeskrivelsen, så ville jeg dog anbefale, at disse meget specifikt testes i sættet for at skabe en taksonomisk højde over faktuel viden. Jeg har søgt sættet igennem for items, der tester, om den studerende kan 'forklare' eller 'anvende' eller 'vurdere kritisk', men kun fundet meget, meget få af disse items. Dermed opmuntrer sættet desværre de studerende til at fokusere deres læring på faktuel viden, hvilket kan medføre udenadslære frem for dybere forståelse.

Man kunne i modulet derfor drøfte, om det er hensigtsmæssigt at kundskabskravene er sat taksonomisk lavt med langt de fleste kundskabsmål på SOLO-niveau 'multistrukturel' og kun seks højere kundskabsmål på SOLO-niveauet 'relationel'. Ved at løfte niveauet i læringsudbyttebeskrivelse og prøven bare en anelse kunne man sikre en større sammenhæng til de valgte undervisningsformer (PBL, TBL og øvelser), som jo netop IKKE er rettet mod udenadslære, men dyb forståelse og anvendt viden. I undervisnings-evalueringerne har de studerende udtrykt frustrationer over netop prøvens indhold og skriver fx:

- Jeg syns ikke spørsmålene på digital eksamen i blokk 3 gjenspeilet helt det vi har fokusert på i Pbl og Tbl. Viktig at det er samsvar mellom disse syns jeg.
- Blokk 3 har opplevdes litt rotete. Veldig gode forelesninger i immunologi og gode kurs generelt i alle fag. Vanskelig å få oversikt over pensum i mikrobiologi (noe som kom tydelig fram på blokk 3-eksamen, da spørsmålene var mye mer detaljerte enn forventet).
- Blokk 3 inneholdt mange spennende temaer, med stor motivasjon fra min side. Det var derimot få studenter som forstod omfanget av mikrobiologi da vi oppfattet det mer som "små-kurs" som vi egentlig hadde stort tidspress på å fullføre fra gang til gang. Til tross for forelesninger tidlig på kursene ble ikke disse oppfattet eller bemerket som såpass viktige at de temaene skulle få bortimot 50 % av eksamen i blokk 3. Sett opp mot grundighet rundt blodkursene, mtp forberedelser med flere forelesninger, gjennomgang og tidsbruk var det overraskende at mikrobiologi fikk så mye "plass".
- Blokk 3 var litt mer uoversiktelig, og var preget av at vi hadde dårlig tid. Kun fokus på pugging, og lite fokus/mulighet/tid til å få en ordentlig forståelse. Her var det også gode forelesere, men eksamen svarte ikke til det som ble forespeilet - verken av forelesere eller læringsmål. Når det er sagt, syns jeg det var positivt med digital eksamen - denne eksamensformen syns jeg fungerte veldig bra
- Ikke ha kliniske forelesninger(alltså forelesninger som ikke er relevante for eksamen) like før eksamen
- Fortell elevene fra starten av hvilke mikrober de skal kunne, og ha gjerne mer teori og litt mindre praktisk.

Det er klart af kommentarerne fra de studerende, at nogle af dem ikke oplever en sammenhæng mellem læringsudbyttebeskrivelse, undervisningsformerne og prøven. Man kunne derfor vælge at justere prøvens højde, så den også tester problemløsning, opgaveløsning, og praktisk klinisk tankegang. Alle disse kan testes med online multiple choice formatet, hvis man i langt højere grad skriver items, der ikke kun tester faktuel viden.

Jeg anbefaler derfor, at man i blokken arbejder fokuseret med at få flere items på højere taksonomisk niveau ind i sættet, så man specifikt tester den studerendes evne til at forklare, anvende kundskab samt vurdere udsagn eller evidens.

De enkelte items relevans og tekniske kvalitet

Det er vanskeligt at skrive gode items, ikke mindst gode items, der tester de højere taksonomiske niveauer. Det er dog afgørende vigtigt for prøvens samlede kvalitet, at de fleste items har en høj individuel kvalitet og jeg vil derfor dvæle ved dette spørgsmål. Jeg ønsker ikke at kritisere enkelte undervisere, men da jeg gerne vil give eksempler på mindre gode items, så har jeg alligevel valgt fire tilfældige nedslag i sættet, som jeg diskuterer i det efterfølgende. Nedenstående er derfor alene valgt som generelle eksempler på almindelige problemstillinger:

Opgave 1, del 3, spørgsmål 1 (18 725):
Der er skrevet en glimrende klinisk case. Der er også stillet et godt spørgsmål: 'Hva er den mest sannnsynlige årsaken til at Arne har vann i lungerne?' Men der er desværre valgt et svarformat, hvor den studerende skal give TO svar. Dette komplicerer et item rigtig meget, uden at man får testet kompleks viden. Man tester alene, om den studerende kan gennemskue et kompliceret svar format. Det kan studerende i almindelighed ikke, hvilket man også kan se i den psykometriske analyse af dette item. Den viser, at dette item er moderat vanskeligt (52 % svarer rigtigt), men item skelner ikke tilstrækkeligt mellem de dygtigste og de dårligste, da de dygtige studerende svarer ligeså spredt og tilfældigt, som de øvrige studerende. Jeg vil klart anbefale, at man til databasen skriver dette item om, så der kun er et bedste svar på spørgsmålet.

Opgave 1, del 3, spørgsmål 2 (18 730):
Samme kliniske case er også basis for dette efterfølgende item. Desværre har casen ingen betydning eller sammenhæng med spørgsmålet, der lyder: 'Hvilken af følgende er hovedtegn på betennelse?' Hvis man dækker for casen kan spørgsmålet alligevel besvares (anbefales ikke). Spørgsmålet er desuden stillet, så man ikke kan formulere et svar på det uden at se på svarmulighederne (anbefales heller ikke). Det er derfor et rent faktuelt spørgsmål uden ansatser til anvendt klinisk viden og der testes kun helt simpel lærebogs-gengivelse. Spørgsmålet er dermed for let, hvilket man også kan se af den psykometriske analyse af dette item. Den viser, at 100 % af de studerende kunne svare rigtigt og spørgsmålet fungerer derfor ikke i den karaktergivende videnstest. Jeg vil klart anbefale, at man til databasen helt omskriver spørgsmål og svar, så de direkte relaterer sig til den kliniske case og tester anvendt viden.

Opgave 1, del 3, spørgsmål 11 (18 751) og 13 (18 753):
Også disse to spørgsmål hører under ovenstående kliniske case. Igen er der desværre ingen sammenhæng mellem spørgsmålene og den kliniske case overhovedet. Den kliniske case fremtræder dermed direkte vildledende for de dygtigste studerende (frarådes). Spørgsmål 11 lyder 'Hvilke utsagn er riktig?' Og spørgsmål 13 lyder 'Hvilken av påstandene under er ikke riktig om lymfeåresystemet?' Spørgsmålene er af typen Sandt – Falsk (frarådes). Svarmulighederne består af løsrevne sætninger, som man ikke kan kategorisere eller se sammenhæng i (anbefales ikke). Dermed bliver de studerendes svar ofte til overfladisk gætteri. Dette fremgår også af den psykometriske analyse, da kun 33 % af de studerende svarer korrekt på spørgsmål 11. Lidt flere svarer rigtigt på spørgsmål 13, men begge items er præget af, at de studerendes

svarmønster er meget spredt. Både de dygtige og de dårligste studerende svarer lige tilfældigt. Jeg vil klart fraråde, at man bruger Sandt-Falsk spørgsmål.

Ovenstående er eksempler på det, som man kalder 'tekniske fejl' i et item. Der dækker over uheldige formuleringer og kombinationer, som forskning og analyser gennem tiden har vist giver et dårligt test-resultat (ringe validitet). Tekniske fejl kan forholdsvist let rettes, hvis item skriverne får en klar guideline, og et antal af dem får udvidet indsit i test formaterne One Best Answer samt Extended Matching og dermed kan fungere som ressourcepersoner for modulet.

Samlede anbefalinger til blok 3

1. at man i blokken diskuterer, hvordan de væsentlige praktiske færdigheder beskrevet i blokkens læringsudbyttebeskrivelse kunne testes eller vurderes opnået.
2. at man i blokken arbejder fokuseret med at få flere items på højere taksonomisk niveau, hvor man tester den studerendes evne til at forklare, anvende sin kundskab og vurdere viden eller evidens.
3. at man primært skriver items i formaterne One Best Answer samt Extended Matching.
4. at man undlader at bruge Sandt-Falsk formatet.
5. at man med fakultetet drøfter, hvordan alle item skrivere kunne få adgang til en overskuelig simpel guideline til 'gode items'.

Til slut vil jeg gerne understrege, at de ovenstående problemstillinger er helt almindelige for alle kurser, der bruger multiple choice test og kritikken er derfor ikke rettet mod hverken blokk 3, uddannelsen eller fakultetets arbejde med prøvekvaliteten. Ovenstående er en understøttelse af fakultetets igangværende solide kvalitetssikring af prøvernes validitet og kvalitet. Online multiple choice prøver er glimrende til at teste kundskab bredt og højt med meget stor pålidelighed og det anbefales, at man - trods udfordringerne - fastholder prøveformen og forsætter det langsigtede træk med at øge kvaliteten.

Med venlig hilsen
Anne Mette Mørcke
Tilsynssensor
Chefkonsulent ved Copenhagen Academy for Medical Education and Simulation