NOKUTs utredninger og analyser

Nasjonal deleksamen i anatomi, fysiologi og biokjemi desember 2017

I hvilken grad sensorenes poengsetting kan begrunnes faglig psykometrisk

Hanne Søberg Finbråten og Øystein Guttersrud, desember 2018

NOKUT

NOKUTs arbeid skal bidra til at samfunnet har tillit til kvaliteten i norsk høyere utdanning og fagskoleutdanning, samt godkjent høyere utenlandsk utdanning. Med rapportserien "NOKUTs utredninger og analyser" vil vi bidra til økt kunnskap om forhold innenfor høyere utdanning og fagskoleutdanning som har betydning for kvaliteten i studiene og gi økt kunnskap om forhold knyttet til godkjenning av utenlandsk utdanning i Norge.

Innhold

1 Innledning

Høsten 2014 fikk Nasjonalt organ for kvalitet i utdanningen (NOKUT) i oppdrag fra Kunnskapsdepartementet (KD) å gjennomføre et pilotprosjekt med nasjonale deleksamener, deriblant sykepleierutdanningen (Tokstad & Hamberg, 2017). Nasjonal deleksamen i anatomi, fysiologi og biokjemi (AFB) i bachelorutdanningen i sykepleie er nå en permanent ordning forvaltet av NOKUT.

Hensikten med nasjonale deleksamener er tredelt. Eksamenene gir informasjon om studentenes kunnskapsnivå, og denne informasjonen kan danne grunnlag for kunnskapsbasert lokalt kvalitetsarbeid ved institusjonene. Eksamenene gir også institusjonene mulighet til å sammenlikne seg med hverandre, og dette kan gi informasjon om spesifikke innsatsområder for videreutvikling av utdanningskvaliteten. For det tredje kan nasjonal deleksamen bidra til økt tillit til utdanningene, som at eksamen gir en karakterkalibrerende effekt – at en karakter skal reflektere det samme kunnskaps- og ferdighetsnivået på tvers av institusjonene (Tokstad & Hamberg, 2017).

Dersom resultater fra nasjonale deleksamener skal inngå i kunnskapsgrunnlaget for kvalitetsutvikling, må eksamen gi gyldig og pålitelig informasjon om studentene. Deleksamen gir slik informasjon når poengsettingen kan begrunnes faglig og forsvares psykometrisk. Når sensorveiledningen ikke har tilstrekkelig godt definerte vurderingskriterier, slik at vi ikke vet hva hvert poeng «betyr» faglig, er ikke poengene «faglig forankret». Da er det til liten hjelp om poengene tilsynelatende «virker godt» psykometrisk, for vi kan ikke begrunne vurderingen fra et faglig ståsted. Det er dermed bare når vi kan begrunne poengene faglig at det er interessant å evaluere om poengene kan forsvares psykometrisk. Hvis vi har en ide om at poengene kan begrunnes faglig (ut fra teoretisk vitenskap), men poengene likevel ikke kan forsvares psykometrisk (basert på empirisk evidens), må vi revurdere de faglige vurderingskriteriene og/eller vurderingssystemets «oppbygging». Målet vårt er altså teoretisk funderte vurderingssystemer som kan forsvares empirisk.

1.1 Beskrivelse, tolkning og utdyping av oppdraget

Beskrivelse av oppdraget fra NOKUT:

- *å undersøke intern avhengighet i enkeltoppgaver som «spør etter mer enn én ting»*
- *å finne hvilke faglige vurderingskriterier sensorene synes å ha benyttet for sin poenggivning, når dette ikke er tydelig beskrevet i sensorveiledningen*
- *å bidra til videreutvikling av fremtidige oppgavetyper og sensorveiledninger med tanke på nivåbeskrivelser og kumulative rettemaler.*

Oppdraget er **tolket** til å handle om å gjøre dypere analyser av enkelte oppgaver gitt til deleksamen i AFB i desember 2017, fordi oppdraget tar utgangspunkt i en tidligere kvantitativ analyse av dataene fra denne eksamenen (Guttersrud, 2018). Analysen påpekte blant annet at flere oppgaver ba studentene om å gjøre «mer enn én ting». Guttersrud viste at slike oppgaveformuleringer gjør det vanskelig å tolke resultatene, fordi det er uklart hvilke faglige kunnskaper studenter med «delvis kreditt» da har. Oppdragets pkt. 1 handler om slike oppgaver, og i denne rapporten eksemplifiseres denne typen utfordringer gjennom en dypere analyse av oppgavene 1b, 4c og gitt til deleksamen i AFB i desember 2017.

Når sensorveiledningen ikke tydelig beskriver *hvilke* av vurderingskriteriene som må oppfylles for å få de ulike poengene mellom 0 og full kreditt, f.eks. poengene 1-5 når en enkeltoppgave krediteres opptil 6 poeng, viste analysen til Guttersrud (2018) at oppgaven typisk har såkalte «uordna» poengkategorier. **Uordna poeng** indikerer at vurderingssystemet ikke har fungert optimalt. Uordna poengkategorier signaliserer gjerne at en har valgt en skåringsmodell med «for mange» poengnivåer – flere poengnivåer enn det antallet tydelig atskilte dyktighetsnivåer som oppgaven greier å avdekke blant studentene. Hvis en oppgave ber studentene om å definere f.eks. «puls», vil vi kanskje gi 0 poeng til blanke svar, 1 poeng til de som refererer til «antall slag per minutt» og 2 poeng til de som refererer til «trykkbølge i blodårene». Hvis det ikke opptrer svar med høyere faglig kvalitet enn de som refererer til «trykkbølge», avdekker ikke oppgaven mer enn tre dyktighetsnivåer (0-2 poeng). Hvis vi likevel deler ut f.eks. 0-6 poeng kan ikke poeng 3-6 begrunnes faglig. Grunnen er at vi ikke har vurderingskriterier som svarer til disse «ekstra» poengene. Da vil vi kunne observere det vi kaller «uordna poengkategorier». Uordna kategorier kan også opptre når sensorveiledningen i for liten grad hjelper sensorene med å definere kriteriene eller skillelinjene/tersklene mellom de ulike poengkategoriene. Oppdragets pkt. 2 handler dermed om oppgaver med denne typen utfordringer, og i denne rapporten eksemplifiseres slike utfordringer gjennom dypere analyser av oppgave 4c og 6c gitt

til deleksamen i AFB i desember 2017. Disse to oppgavene reflekterer dermed kvalitetsutfordringene som ligger til grunn for både oppdragets pkt.1 og pkt. 2.

Oppdragets pkt. 3 handler om å gi innspill til videreutviklingen av nasjonal deleksamen i AFB, og da særlig med tanke på utfordringene skissert i oppdragets pkt. 2 angående «faglige vurderingskriterier». Nedenfor er bakgrunnen for oppdraget **utdypet** ved hjelp av eksempler på oppgaver.

Oppgaver som etterspurte to ting, slik som 1b «*Beskriv hva som menes med puls* [deloppgave 1]*, og nevn normalverdier for puls i hvile hos voksne* [deloppgave 2]*»*, er et eksempel på en oppgave som ber studentene om svare på mer enn én ting (jf. oppdragets pkt. 1). Slike oppgaver er problematiske, for vi vet da f.eks. ikke hva det betyr faglig å få 1 poeng av de maksimalt oppnåelige 2 poengene på oppgaven. På den aktuelle oppgaven kan studenter med 1 poeng ha definert puls (f.eks. som «trykkbølge» eller som hverdagsforestillingen «antall slag per tidsenhet»), *eller* de kan ha oppgitt akseptable normalverdier for hvilepuls hos voksne (et intervall eller en enkeltverdi). Videre kan en stille spørsmål ved om «trykkbølge» og «antall «slag» faglig sett fortjener samme kreditt, eller om deloppgave 1 bør krediteres med mer enn 1 poeng (f.eks. full kreditt eller 2 poeng for «trykkbølge», og delvis kreditt eller 1 poeng for «antall slag pr tidsenhet»).

For å hente ut mer informasjon om hva studentene faktisk svarte på 1b, vurderte vi studentenes svar på oppgavens to deloppgaver («beskriv puls» og «nevn normalverdier for puls») på nytt, og vi vurderte besvarelsene på de to deloppgavene som om de to deloppgavene var uavhengige oppgaver. Vi antar dermed at de to deloppgavene ikke har «mer til felles med hverandre» enn de har med de øvrige oppgavene i eksamenssettet, og dette er en antakelse vi må teste (jf. «svaravhengighet» nedenfor). Vi registrerte dermed poengene vi delte ut i to variabler – én for hver deloppgave («beskriv puls» og «nevn normalverdier for puls»). Disse to nye variablene ble deretter rekodet til nye variabler for å evaluere ulike kombinasjoner av vurderingskriterier med tilhørende skåringsmodeller, men dette kommer vi tilbake til. Det var dermed mulig å undersøke den statistiske avhengigheten eller «svaravhengigheten» mellom deloppgavene (jf. oppdragets pkt. 1). Uavhengige oppgaver kan skåres individuelt, mens avhengige oppgaver må skåres som en enhet eller «subtest». Vi oppretter subtester for å håndtere at avhengige oppgaver bryter med prinsippet om «lokal uavhengighet». Det er et empirisk spørsmål om to oppgaver,

utover det underliggende «trekket» eller konstruktet som måles, er statistisk avhengige eller ikke.

På enkelte oppgaver ble studentene kreditert opptil 6 poeng, slik som 4c «*Beskriv hvor de tre hovedtypene av muskulatur finnes i kroppen, og hvordan hver av dem påvirkes av nervesystemet*» (merk at også denne oppgaven ber studentene svare på mer enn én ting: nevne muskeltype, identifisere plassering i kroppen, og beskrive hvordan nervesystemet påvirker muskeltypen). I sensorveiledningen var det ikke tydelig spesifisert hva som var kriteriet for å gi poengene 1-5. Det var dermed opp til sensorene å identifisere de seks faglige «tersklene» mellom de syv poengkategoriene 0-6 (jf. oppdragets pkt. 2). Vi tar det her for gitt at studenter som ikke oppfylte noen av kriteriene skulle få 0 poeng, og at de som oppfylte alle kriteriene skulle få 6 poeng.

Analysen til Guttersrud (2018) viste at oppgave 4c ga *u*ordna poengkategorier, og det kan dermed se ut til at sensorene opererte med flere poengkategorier enn det antallet tydelig atskilte dyktighetsnivåer oppgave 4c greide å avdekke blant studentene. Da analysen til Guttersrud viste at f.eks. gruppa av studenter som fikk 3 poeng på oppgaven ikke var tydelig faglig dyktigere på oppgavesettet enn gruppa som fikk 2 poeng, er det uklart hva sensorene belønnet når de krediterte studenter 3 poeng. Det er med andre ord uklart «hvilke faglige vurderingskriterier sensorene synes å ha benyttet for sin poenggivning» (jf. oppdragets pkt. 2).

På bakgrunn av våre sekundære analyser av dataene fra nasjonal deleksamen i AFB høsten 2017, skisserer vi noen forslag til videreutvikling av eksamen, slik som «kumulative rettemaler» (jf. oppdragets pkt. 3).

2 Fremgangsmåte og analyser

2.1 Utvalg av og antall besvarelser

For å løse oppdraget fra NOKUT må ytterligere informasjon hentes ut fra studentenes svar på eksamensoppgaver, og dette forutsetter ny gjennomgang og ny vurdering av besvarelser. Da det er uhensiktsmessig å vurdere alle de 5062 besvarelsene på nytt, er analysene i denne rapporten basert på et utvalg av besvarelsene. Vi har dessuten begrenset oss til tre oppgaver som vi mener er gode eksempler på de utfordringene oppdraget har til hensikt å belyse: oppgavene 1b, 4c og 6c fra nasjonal deleksamen i AFB desember 2017. Vi hadde tilgang til alle studentenes skriftlige besvarelser på PDF og en SPSS-fil med alle sensorenes poenggiving på hver oppgave, samt samsensuren.

Nedenfor har vi kort beskrevet relevante aspekter ved såkalt «Rasch-analyse». Denne typen analyse bør bygge på data fra minst 250 enheter (Linacre, 1994) – i dette tilfellet eksamensbesvarelser fra 250 studenter. Når svar på oppgaver krediteres mer enn 1 poeng, er det anbefalt å inkludere minst 10 ekstra besvarelser per poengkategori (Linacre, 1994). Da sensorene ga opptil 6 poeng per enkeltoppgave (0-6 poeng eller syv kategorier), bør våre analyser derfor baseres på minst 250 + 70 = 320 besvarelser. Et tilfeldig utvalg av 11 sensorpar (1cd, 1ce, 2ab, 3bc, 4cd, 5de, 7cd, 9bd, 15cd, 16ae, 16de) hadde til sammen sensurert 334 besvarelser (6,6 % av totalt 5062 besvarelser), og dette utvalget av besvarelser danner dermed et tilstrekkelig grunnlag for våre sekundære analyser. Disse 334 besvarelsene ligger til grunn for analysene i denne rapporten. Datapakkene SPSS og RUMM ble brukt til databehandling og Rasch-analyse. Fra et kvalitativt ståsted vil 334 besvarelser kunne være tilstrekkelig til å oppnå «metning» – avdekke de ulike typene av svar som opptrer på en oppgave, men dette har ikke vært det primære målet for våre sekundære analyser av datamaterialet.

2.2 Relevante aspekter ved Rasch-analyse

Rasch-analyse innebærer at en tester dataene opp mot Rasch-modellen (Rasch, 1960), og Rasch-modellen predikerer sannsynlighetene for at en student med en bestemt dyktighet kan oppnå de ulike poengene som deles ut på en oppgave. Nedenfor beskrives kort og kvalitativt de aspektene ved Rasch-analyse som er nødvendige for å kunne lese rapporten.

Dyktigheten til en person er estimert på bakgrunn av personens poengsum på oppgavesettet, mens **vanskegraden** til en oppgave er estimert på bakgrunn av det totale antallet poeng som oppgaven delte ut til personene (f.eks. 50 poeng for en flervalgsoppgave hvor 50 personer svarte riktig og fikk 1 poeng hver). Vi kan altså si at vanskegrad handler om andelen personer som svarte riktig (avkrysningsoppgave), eller som skrev et svar som fortjente kreditt (åpen oppgave). Når dataene fra en oppgave testes opp mot Rasch-modellen, kan vi avgjøre om oppgaven **diskriminerer** eller skiller tilstrekkelig mellom personer med høy og lav dyktighet. I denne rapporten evaluerer vi oppgavenes evne til å diskriminere ved å studere dataenes tilpasning til den grafiske representasjonen av «partial credit»-parametriseringen av Rasch-modellen for endimensjonale data (PCM). En eksamensoppgave er **rettferdig** når den favoriserer – deler ut poeng til – faglig dyktige personer, og det oppnår vi når dataene fra oppgavene har tilstrekkelig tilpasning til den nevnte modellen. Når oppgavene diskriminerer eller skiller tilstrekkelig mellom personer med høy og lav dyktighet, følger det at faglig dyktige personer får høye poengsummer og følgelig gode karakterer.

Når to oppgaver er for like hverandre – tester samme kunnskap eller tilsvarende ferdighet, har oppgavene mer til felles med hverandre enn normaloppgaven i oppgavesettet. Vi sier da at oppgavene er **statistisk avhengige**, og at de dermed bryter med kravet om «lokal uavhengighet» (Marais & Andrich, 2008). Statistisk avhengighet avgjøres ved å studere såkalte «residualkorrelasjoner» (Marais & Andrich, 2008). Denne rapporten har ikke til hensikt å gå i dybden på de ulike aspektene ved en Rasch-analyse, og interesserte lesere henvises derfor til artikkelen.

Basert på faglige vurderingskriterier i sensorveiledningen, vurderer sensorene studentenes svar som f.eks. svake, tilstrekkelige/gode eller svært gode. Svarene kan krediteres f.eks. 0, 1 eller 2 poeng, og dette kaller vi en «**skåringsmodell**». Det følger at skåringsmodellene gir oppgavedata på ordinalnivå. En Rasch-analyse gir informasjon om poengene er såkalt «ordna». Uordna poeng signaliserer at vurderingssystemet ikke har fungert tilfredsstillende godt. Som nevnt, indikerer gjerne uordna poengkategorier at en har valgt en skåringsmodell med «for mange» poengnivåer – flere poengnivåer enn det antallet tydelig atskilte dyktighetsnivåer som oppgaven greier å avdekke blant studentene.

2.3 Diskriminering og klassisk testteori (z-skår)

Eksamenskarakteren er basert på studentens poengsum, og poengsummen er summen av poengene oppnådd på oppgavesettet. Dersom vi *tolker* poengsum som en intervallvariabel, kan vi regne ut studentenes gjennomsnittlige poengsum og estimere standardavviket til fordelingen av poengsummer. Poengsummen til en student kan «standardiseres» (z-skår) ved å regne ut differansen mellom studentens poengsum og den gjennomsnittlige poengsummen, og dividere denne differansen på standardavviket til fordelingen av poengsummer (Ringdal, 2007). Studentens z-skår beskriver dermed studentens prestasjon som antall standardavvik over eller under gjennomsnittet. Ved å sammenlikne den *gjennomsnittlige* z-skåren til studentene som fikk 1 poeng på en bestemt oppgave med den *gjennomsnittlige* z-skåren til de som fikk 2 poeng på oppgaven, kan vi enkelt indikere om poengene total sett synes å diskriminere eller skille mellom studenter med høy og lav dyktighet (der dyktighet er målt ved høy og lav poengsum på oppgavesettet). Denne typen «klassisk» analyse er altså bare basert på poeng og poengsummer – ikke sannsynligheter slik som Rasch-analysen.

3 Analyser

3.1 Oppdragets pkt.1: oppgaver som spør etter mer enn én ting

Deleksamenen i AFB i desember 2017 hadde tre åpne oppgaver hvor det eksplisitt var spurt etter mer enn én ting. Dette gjaldt oppgave 1b «*Beskriv hva som menes med puls, og nevn normalverdier for puls i hvile hos voksne*», 4c «*Beskriv hvor de tre hovedtypene av muskulatur finnes i kroppen, og hvordan hver av dem påvirkes av nervesystemet*», og oppgave 6c «*Nevn hva bukspytt inneholder, og hvilke funksjoner de ulike komponentene i bukspyttet har*».

3.1.1 Oppgave 1b – sensorveiledningen og eksempler på besvarelser som sensorene ga full, delvis og ingen kreditt

På oppgave 1b var det forventet at studenten både beskrev puls og oppga normalverdier for puls. Oppgaven ga opptil to poeng, og i sensorveiledningen stod det følgende:

> *Puls er en trykkbølge som brer seg langs arterien som følge av hjertes kontraksjon. Normalverdier for hvilepuls hos voksne er ca. 50-80 slag/minutt (det bør utvises et visst skjønn når det gjelder svarene på normalverdier for puls).*

Sensorveiledningen beskriver kravet til et fullgodt svar – full kreditt eller 2 poeng, men veiledningen beskriver ikke hva som kreves for delvis kreditt eller 1 poeng. Sensorveiledningen tar heller ikke stilling til om typiske «hverdagsforestillinger», slik som «antall slag per minutt», skal krediteres. Mange studenter ble dermed kreditert (både hele og halve poeng) for beskrivelser som kan kategoriseres som hverdagsforstillinger. Punktlista nedenfor, som inneholder autentiske besvarelser, viser noen av de feil- og hverdagsforestillingene vi avdekket når vi vurderte de 334 utvalgte besvarelsene på nytt.

- *Puls er det samme som antall hjerteslag i minuttet.*
- *Puls er hvor mange ganger hjertet slår per minutt.*
- *Puls er hvor fort hjertet slår.*
- *Puls er hvor hardt hjertet slår.*
- *Puls er det samme som slagvolum.*
- *Puls er blodets bevegelse.*

Det kan i tillegg være verdt å merke seg at en del studenter gir beskrivelser som er mer assosiert med «blodtrykk» enn «puls». Kvalitative undersøkelser kan avdekke om slike studenter blander sammen begrepene, eller om de mener at blodtrykk og puls er det samme. Kanskje kan «puls som *trykk*bølge» gi opphav til feilforestillinger som at puls har noe med blod*trykk* å gjøre.

Basert på erfaring kan vi noen ganger observere sammenheng mellom oppgavers aksjonsverb (f.eks. *beskriv*, *begrunn svaret* eller *trekk konklusjon basert på informasjonen*) og det maksimale antallet poeng en tenker seg at studentene kan oppnå på oppgaven. Dette kan forklare at noen sensorer synes å ha gitt 0,25 poeng for riktig svar på del 2 «*nevn* normalverdier» og mer enn 1 poeng for riktig svar på del 1 «*beskriv* puls». Enkelte studenter har dessuten fått full kreditt til tross for feil normalverdi(er). Ved å sammenholde poengsettingen fra samsensuren med studentenes besvarelser, ser det imidlertid ut til at flesteparten av sensorene delte ut opptil 1 poeng på hver deloppgave.

Punktlista nedenfor viser eksempler på svar på 1b som sensorer ga full kreditt eller 2 poeng. Våre kommentarer til besvarelsene står i fet type i klammeparentes.

- *Puls er en indikator på hjertefrekvensen. Blodet sendes ut gjennom arteriene med kraftige trykkbølger, og kan måles som puls. Normalverdier for puls i hvile hos voksne er 120/80.* [**Studenten beskriver for så vidt puls som en trykkbølge, men oppgir normalverdi for blod*trykk***].
- *Pulsen er de «trykkene» som man kan kjenne i blodåreveggen. Trykket kommer fra hjertets kontraksjon. Normalpuls for voksne er ca 60 (60 «trykk» i minuttet).* **[Studenten beskriver puls upresist, men oppgir enkeltverdi innenfor akseptabelt intervall for puls].**
- *Pulsen er den frekvensen hjertet pumper blod i. Man teller antall «pump» eller hjerteslag i løpet av et minutt: Normal puls i hvile for voksne er mellom 50 og 80 slag per minutt.* **[Studenten beskriver puls upresist, men oppgir et akseptabelt intervall for puls].**
- *Puls er trykkbølgen som sprer seg langs blodårenes vegg når blodet strømmer gjennom blodårene. Normalverdi for puls hos voksne er mellom 50-80, hvor flere ulike faktorer har noe å si, f.eks. fysisk form og ulike sykdommer.* [**Studenten beskriver puls som en trykkbølge, og studenten oppgir akseptabelt intervall for pulsverdier**].

Eksemplene gjenspeiler at det var til dels stor variasjon i kvaliteten på svarene som ble gitt full kreditt. Det er naturlig å anta at ikke alle sensorene ville vært enige i at alle svarene i punktlista ovenfor fortjener full kreditt, og dette kan forklare oppgavens relativt lave sensorreliabilitet (Pedersen, Skeidsvoll, & Tokstad, 2018). Når studenter krediteres for normalverdier for *blodtrykk* (se første eksempel i punktlista ovenfor), er det grunn til å stille spørsmål både ved sensorers arbeid og kvalitetssikringen av samsensuren.

Punktlista nedenfor viser eksempler på svar på 1b som sensorer ga delvis kreditt eller 1 poeng. Våre kommentarer til besvarelsene står i fet type i klammeparentes.

- *Puls og blodtrykk henger sammen. Hjertet har mye å si på hvor fort pulsen slår, samme med blodtrykket. Er du stresset og nervøs, vil hjertet og pulsen slå fortere, er du rolig vil pulsen slå roligere. Du kan måle pulsen blant annet på*

håndleddet. Normalverdier for puls i hvile er 60-75.[**Studenten definerer egentlig ikke puls, men svaret er ikke faglig feil. Studenten krediteres mest sannsynlig for et akseptabelt intervall for puls**].

- *Puls er det samme som hjertefrekvensen, hvor mange ganger hjertet kontraherer (strammer musklene) hvert minutt. 60-80 er normal puls hos voksne i hvile.* **[Beskrivelsen av puls er ufullstendig, men intervallet for puls er akseptabelt. Sensor kan ha gitt noe kreditt for definisjonen, eller bare kreditert intervallet].**
- *Hos voksne ligger normalt puls fra 55-70 i hvile.* **[Studenten har fått 1 poeng for et akseptabel intervall for puls].**
- *Puls er hvor mange ganger hjertet slår per minutt.* **[Studenten beskriver puls upresist, og studenten oppgir ingen normalverdi. Studenten ble kreditert 1 poeng for upresis beskrivelse av puls].**

Eksemplene viser at det ikke er mulig å beskrive hvilken kunnskap studenter med 1 poeng på oppgaven har. Studenter med 1 poeng kan ha definert puls (f.eks. som «trykkbølge» eller som hverdagsforestillingen «antall slag per tidsenhet»), *eller* de kan ha oppgitt akseptable normalverdier (et intervall eller en enkeltverdi). Disse to typene svar reflekterer ulik kunnskap, men de krediteres likevel med «det samme poenget». Dette gjør det vanskelig å bruke oppgaven til å utvikle såkalte «**mestringsbeskrivelser**». Mestringsbeskrivelser kan skissere kunnskaper og ferdigheter til studenter med f.eks. karakter B relativt til studenter med karakter C.

Punktlista nedenfor viser eksempler på svar på 1b som sensorer ikke krediterte (0 poeng). Våre kommentarer til besvarelsene står i fet type i klammeparentes.

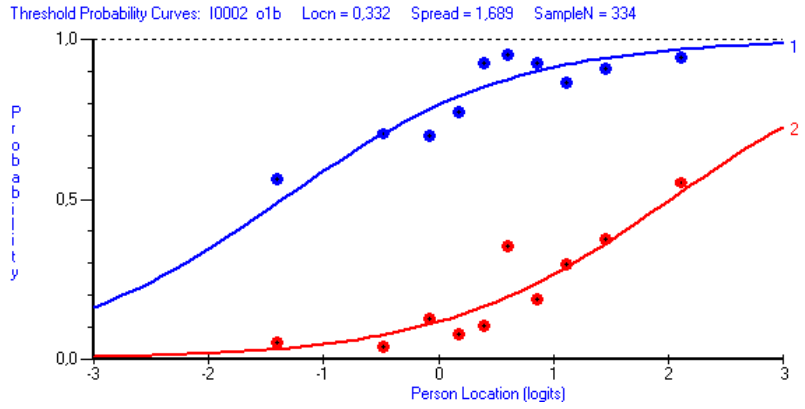
- *Puls er antall slag hjertet gjør i løpet av et minutt. Normalverdi for voksne i hvile er 60-80 slag per minutt.* [**Studenten beskriver puls upresist, men har oppgitt et akseptabelt intervall for puls uten å bli kreditert for dette**].
- *Puls er når hjertet pumper blod ut i arteriene, blir målt i ett minutt og da er antall ganger hjertet pumper blod ut i arteriene på ett minutt. Normal hvilepuls hos voksne ligger mellom 55/60 og 100.* **[Studenten beskriver puls upresist, og intervallet for puls er relativt stort]**.

Da en del sensorer krediterte upresise beskrivelser av puls, og noen krediterte akseptable enkeltverdier for puls, kan en tenke seg at de to studentene som fikk 0 poeng kunne fått 2 poeng av et annet sensorpar. I så fall avhenger studenters karakter på eksamen i AFB av hvilken sensor de ble tildelt, og det er urimelig. Sensorveiledningen utdyper at sensorene skal vise skjønn når det gjelder normalverdi, og det er mulig at dette medvirket til sensorenes ulike tilnærminger til oppgaven. Sensorveiledningen burde tydeliggjort om hverdagsforestillingen «antall slag per tidsenhet» skulle krediteres, og om enkeltverdi for normalpuls skulle godkjennes.

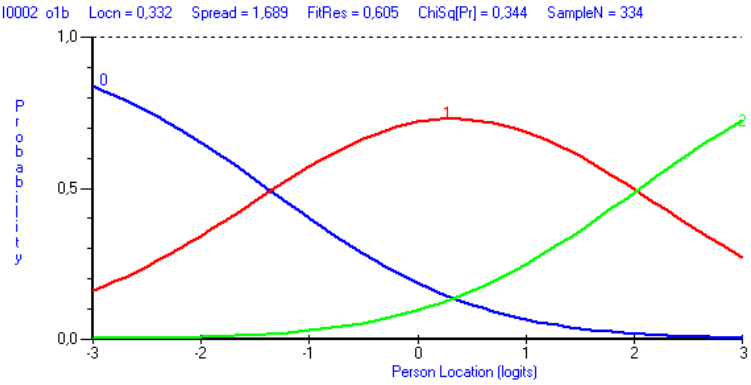
Det synes å være behov for at sensorveiledningene ikke bare beskriver kravet til et fullgodt svar, men at veiledningene også skisserer svar som fortjener delvis kreditt. For å sikre likere vurdering bør sensorene få opplæring i bruk av i sensorveiledningene, og kanskje burde «ekspertsensorer» ta stikkprøver med hensyn på å identifisere avvikende sensorer. En annen måte å løse utfordringen på er at eksamen bare består av avkrysningsoppgaver, slik som f.eks. flervalgsoppgaver. Dette kan igjen utløse andre problemer, som at oppgavene ikke diskriminerer eller skiller tilstrekkelig godt mellom studenter med lav og høy dyktighet (jf. tilpasning til Rasch-modell).

3.1.2 Analyser av data for oppgave 1b basert på original sensur

Figur 1, Figur 2 og Tabell 1 viser analyser av data for oppgave 1b basert på original sensur. Disse analysene er dermed «identiske» med analysene rapportert i Guttersrud (2018), men er her altså basert bare på de utvalgte besvarelsene (n = 334) og ikke hele populasjonen (n = 5062). Analysene er rapportert for å gi de «nye» analysene et sammenlikningsgrunnlag.



Figur 1. Tilpasning til Rasch-modell basert på original sensur av 1b. Røde målepunkter / observerte verdier viser hvordan poeng 2 (vanskegrad 2,0) diskriminerer, og blå målepunkter / observerte verdier viser hvordan poeng 1 (vanskegrad -1,4) diskriminerer.



Figur 2. Sannsynlighetskurver basert på original sensur av 1b. Kurvene viser sannsynligheten for å oppnå 0, 1 og 2 poeng (andreaksen) på 1b som funksjon av dyktighet (førsteaksen).

Tabell 1. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert henholdsvis 0, 1 og 2 poeng på 1b. Analysen er basert på original sensur av 1b.

Poeng	**Gj.sn. z**	**Andel (%)**
0	-0,80	15,3
1	-0,04	64,4
2	0,73	20,4

Tabell 1 viser at gruppen av studenter som fikk 0 poeng på 1b (15 %) har lav gjennomsnittlig dyktighet målt ved z-skår (-0,80), og at gruppen av studenter som fikk 2 poeng (20 %) har høy gjennomsnittlig dyktighet målt ved z-skår (0,73). Vi har delvis kunnskap om hva studenter med 2 poeng på oppgaven «kan», for de skal ha beskrevet puls tilstrekkelig godt og oppgitt akseptable normalverdier for puls. Vi så imidlertid ovenfor at en del svar kreditert 2 poeng ikke holdt tilstrekkelig høy faglig kvalitet. Vi vet lite om hva kandidater med 1 poeng på oppgave 1b «kan». De kan ha gitt god, svak eller ingen beskrivelse av puls, og de kan ha oppgitt et akseptabelt intervall, en akseptabel enkeltverdi eller ikke oppgitt verdi for normalpuls. Det betyr at nasjonal deleksamen ikke gir oss informasjon om hva 64 % eller nær 2/3 av studentene «kan» om puls, og at vi er usikre på kompetansen til de 20 % eller 1/5 av studentene som fikk 2 poeng. Vi velger da å konkludere med at sensorenes poengsetting på 1b er vanskelig å begrunne faglig. Når vi ikke kan begrunne poengene faglig (vurderingssystemet er *ikke* tilstrekkelig godt teoretisk fundert), er det egentlig uinteressant om analysen viser at poengene

kan forsvares psykometrisk. Figur 1 og Figur 2 viser imidlertid at de to poengene på oppgave 1b tilsynelatende fungerte godt psykometrisk sett, men vi vet ikke hva poengene «betyr» faglig.

3.1.3 Analyser av data for oppgave 1b basert på «ny vurdering» av de utvalgte besvarelsene, tydeligere definerte vurderingskriterier og reviderte skåringsmodeller

Når en sammenholder sensorveiledningens vurderingskriterier (jf. «trykkbølge» og intervallet «50-80 slag per minutt» for verdier for normalpuls) og skåringsmodell (0-2 poeng) med studenters besvarelser, utløses et behov for å definere tydeligere vurderingskriterier og kvantitativt evaluere ulike skåringsmodeller (se Tabell 2).

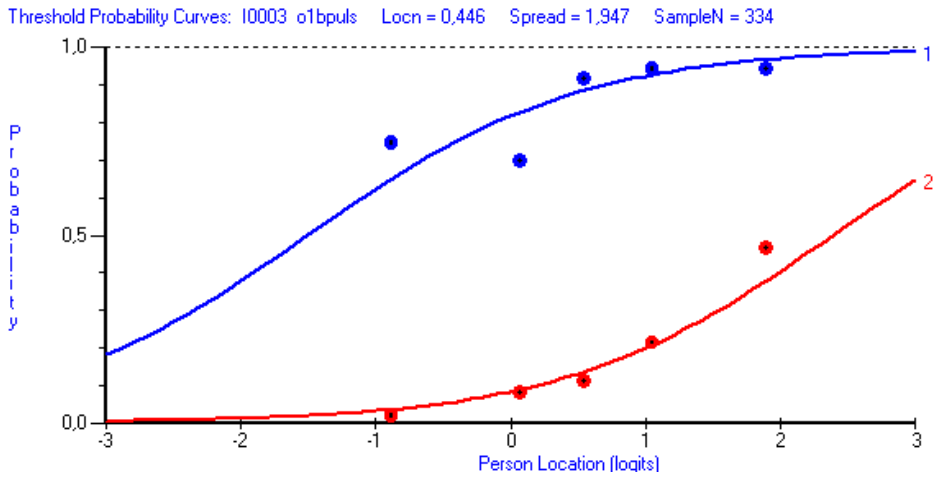
Tabell 2 er ment å illustrere at vi opprettet fem nye variabler (Var1i, Var1ii, Var2i-Var2iii) i SPSS-fila i tillegg til variabelen med sensorenes originale vurdering eller poengsetting av oppgave 1b. Var1i og Var1ii betegner de to nye variablene for deloppgave 1 «beskriv puls», mens Var2i-Var2iii betegner de tre nye variablene for deloppgave 2 «normalverdier for puls». Nedenfor eksemplifiserer vi hvordan opprettelsen av nye variabler i kombinasjon med tydeligere vurderingskriterier gir oss betydelig mer informasjon om studentenes kunnskaper.

Tabell 2. Vurderingskriterier med tilhørende skåring for de fem nye variablene som vi opprettet for oppgave 1b. Variablene ble opprettet i SPSS-fila inneholdende data fra sensorenes vurderinger av alle oppgavene i eksamenssettet.

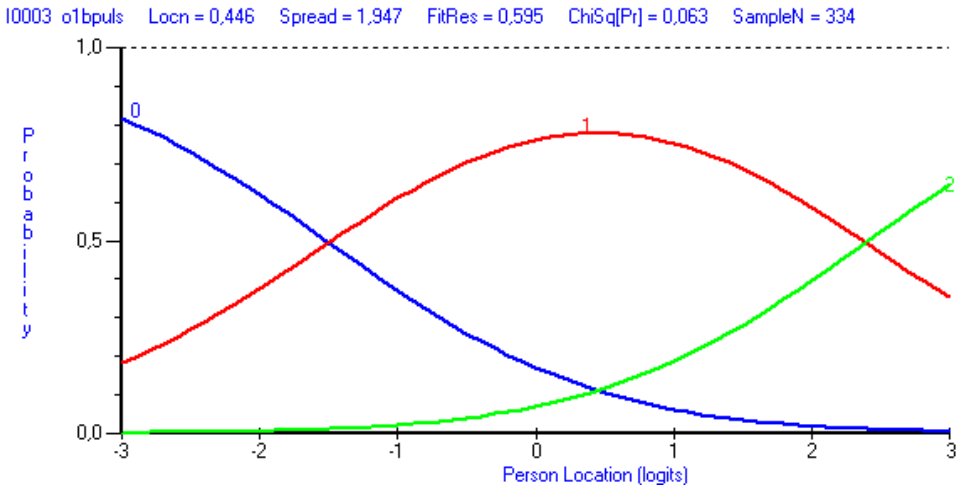
	Nye variabler for oppgave 1b med vurderingskriterier og skåringsmodell				
Skåring	deloppgave 1 «beskriv puls»		deloppgave 2 «normalverdier»		
(poeng)	Var1i	Var1ii	Var2i	Var2ii	Var2iii
2	trykkbølge	-	intervall	-	-
1	hjertefrekvens	trykkbølge	enkeltverdi	intervall	enkeltverdi eller intervall
0	andre svar	hjertefrekvens og andre svar	andre svar	enkeltverdi og andre svar	andre svar

Oppgave 1b deloppgave 1 «beskriv puls»

Ved ny vurdering av svarene på oppgave 1b ble poeng på deloppgave 1 «beskriv puls» registrert i variabelen Var1i. Tabell 2 viser at det ble gitt full kreditt eller 2 poeng til svar som refererte til «puls som trykkbølge». Det ble videre gitt 1 poeng til svar som refererte til «hjertefrekvens», «antall slag pr tidsenhet» eller tilsvarende hverdagsforestillinger. Alle fageksperter vil være enige i at svar kreditert 2 poeng (trykkbølge) er kvalitativt sett bedre enn svar kreditert 1 poeng (antall slag per tidsenhet), og vi kan dermed faglig begrunne poengene våre. Det er dermed interessant å undersøke om denne måten å kreditere oppgaven kan forsvares psykometrisk, og Figur 3 og Figur 4 viser at skåringsmodellen som lå til grunn for Var1i fungerte godt. Studenter som refererte til «trykkbølge» fikk altså 2 poeng på 1b deloppgave 1 «beskriv puls», mens de som svarte «hjertefrekvens» eller tilsvarende fikk 1 poeng.



Figur 3. Tilpasning til Rasch-modell basert på skåringsmodellen som ligger til grunn for variabelen Var1i. Studenter som refererte til «trykkbølge» fikk 2 poeng på deloppgave 1 «beskriv puls», mens studenter som svarte «antall slag i minuttet» eller tilsvarende fikk 1 poeng. Røde målepunkter / observerte verdier viser hvordan poeng 2 (vanskegrad 2,4) diskriminerer, og blå målepunkter / observerte verdier viser hvordan poeng 1 (vanskegrad -1,5) diskriminerer.



Figur 4. Sannsynlighetskurver basert på skåringsmodellen som ligger til grunn for variabelen Var1i. Studenter som refererte til «trykkbølge» fikk 2 poeng på deloppgave 1 «beskriv puls», mens studenter som svarte «antall slag i minuttet» eller tilsvarende fikk 1 poeng. Kurvene viser sannsynligheten for å oppnå 0, 1 og 2 poeng (andreaksen) på 1b deloppgave 1 som funksjon av dyktighet (førsteaksen).

Tabell 3. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert henholdsvis 0, 1 og 2 poeng på 1b deloppgave 1. Analysen er basert på skåringsmodellen som ligger til grunn for Var1i.

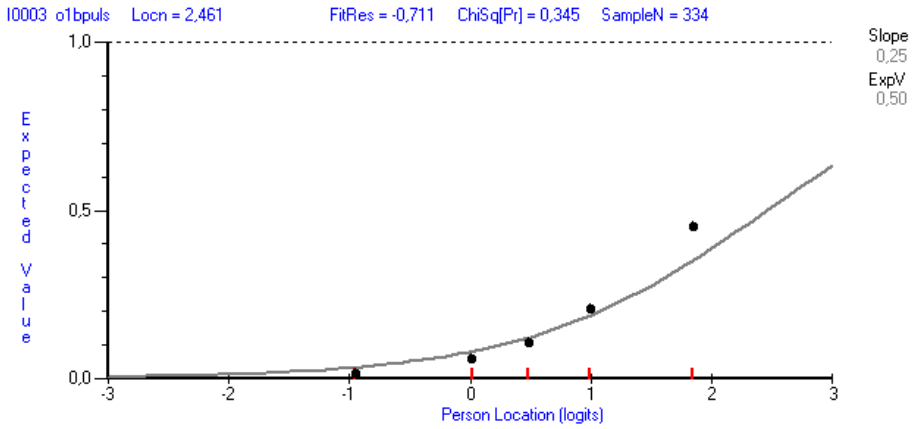
Poeng	Gj.sn. z	Andel (%)
0	-0,71	13,8
1	-0,06	69,8
2	0,85	16,5

Tabell 3 viser at gruppen av studenter som fikk 0 poeng (14 %) har lav gjennomsnittlig dyktighet målt ved z-skår (-0,71), og at gruppen av studenter som fikk 2 poeng (17 %) har høy gjennomsnittlig dyktighet målt ved z-skår (0,85). Resultatene viser at hele 70 % får 1 poeng, og vi vet nå at studenter med 1 poeng på 1b deloppgave 1 anvender hverdagsforestillingen om at puls er «antall slag i minuttet» eller tilsvarende. Disse resultatene avdekker muligens at institusjonene ikke gir undervisning tilpasset studentenes behov. Studentene på sykepleie har vært gjennom 11 år med obligatorisk naturfagundervisning i grunnskolen, men her lærer nok elevene at puls er «hjertefrekvens» i betydningen «antall slag per minutt».

Variabelen Var1i ble rekodet til variabelen Var1ii der svar av typen «puls som trykkbølge» nå ble gitt 1 poeng, mens svar som «hjertefrekvens» og «antall slag pr tidsenhet» ikke lenger ble

kreditert. Vi kan si at vurderingskriteriene ble «strammet inn» eller at «lista for å få poeng ble lagt høyere». Vi kan altså heve og senke poengenes vanskegrad ved å heve og senke vurderingskriteriene.

Figur 5 viser at skåringsmodellen som lå til grunn for Var1ii fungerte godt psykometrisk. Dette var forventet da Figur 5 i praksis viser diskrimineringen til poeng 2 i Figur 3. Vi har rett og slett bare utført en enkel rekoding av poengene gitt på 1b deloppgave 1 «beskriv puls» s.a. 2 -> 1, 1 -> 0 og 0 -> 0. Med denne skåringsmodellen får altså studenter som refererte til «trykkbølge» 1 poeng på deloppgave 1 «beskriv puls», mens de som svarte «antall slag i minuttet» eller tilsvarende får 0 poeng. Denne skåringsmodellen, som bare krediterer 16,5 % av studentene (se Tabell 4), er den skåringsmodellen som er mest lojal mot sensorveiledningens vurderingskriterium for 1b deloppgave 1 (dvs. «puls er en trykkbølge»).



Figur 5. Tilpasning til Rasch-modell basert på skåringsmodellen som ligger til grunn for variabelen Var1ii. Studenter som refererte til «trykkbølge» fikk 1 poeng på deloppgave 1 «beskriv puls», mens studenter som svarte «antall slag i minuttet» eller tilsvarende fikk 0 poeng. Målepunktene / observerte verdier viser hvordan poenget (vanskegrad 2,5) diskriminerer. Denne skåringsmodellen er mest «lojal» mot sensorveiledningens vurderingskriterium for 1b deloppgave 1 «beskriv puls».

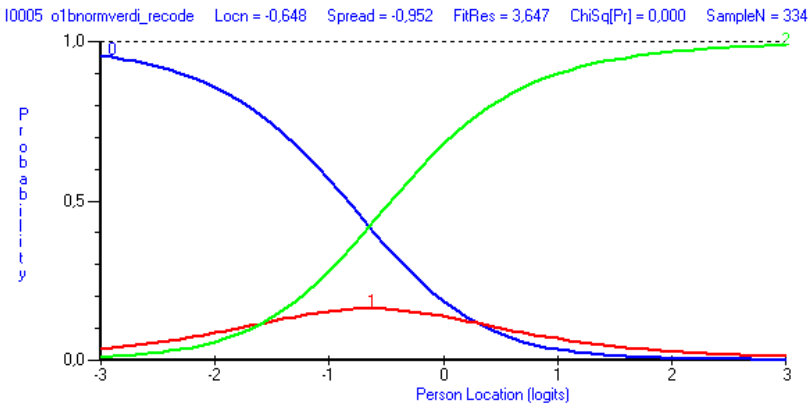
Tabell 4. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert henholdsvis 0 og 1 poeng på 1b deloppgave 1. Analysen er basert på skåringsmodellen som ligger til grunn for Var1ii.

Poeng	**Gj.sn. z**	**Andel (%)**
0	-0,17	83,5
1	0,85	16,5

Oppgave 1b deloppgave 2 «normalverdier for puls»

Ved ny vurdering av besvarelsene på 1b ble poeng på deloppgave 2 «normalverdier for puls» registrert i variabelen Var2i. Tabell 2 viser at det ble gitt full kreditt eller 2 poeng til studenter som oppga et akseptabelt intervall, og at det ble gitt 1 poeng til de som oppga en enkeltverdi innenfor et akseptabelt intervall. Med «akseptabelt intervall» mener vi intervallet 50-80 som angitt i sensorveiledningen. Vurderingskriteriene som ligger til grunn for denne skåringsmodellen er for så vidt et eksempel på «kumulative vurderingskriterier» (jf. oppdragets pkt.3): Alle med 2 poeng (intervall) oppfyller kravet til 1 poeng (enkeltverdi), men de svarer noe mer i tillegg. Studenter med 2 poeng svarer altså kvalitativt sett bedre på 1b deloppgave 2 enn studenter med 1 poeng på deloppgaven, og vi kan dermed faglig begrunne poengene våre for 1b deloppgave 2 «normalverdier for puls». Rasch-analyse viste imidlertid at skåringsmodellen ga uordna poengkategorier (Figur 6).

Tabell 5 viser at bare 10 % fikk 1 poeng, og dette kan forklare at poengkategoriene i Figur 6 er uordna. Figur 6 viser at ingen dyktighetsnivåer langs førsteaksen har 1 poeng som sin mest sannsynlige skår, men dette skyldes altså at få fikk 1 poeng. I slike tilfeller bør vi reskåre oppgaven basert på faglige argumenter.



Figur 6. Sannsynlighetskurver basert på skåringsmodellen som ligger til grunn for variabelen Var2i. Det ble gitt 2 poeng for godkjent intervall og 1 poeng for godkjent enkeltverdi på 1b deloppgave 2 «normalverdier for puls». Kurvene viser sannsynligheten for å oppnå 0, 1 og 2 poeng (andreaksen) på 1b deloppgave 2 som funksjon av dyktighet (førsteaksen).

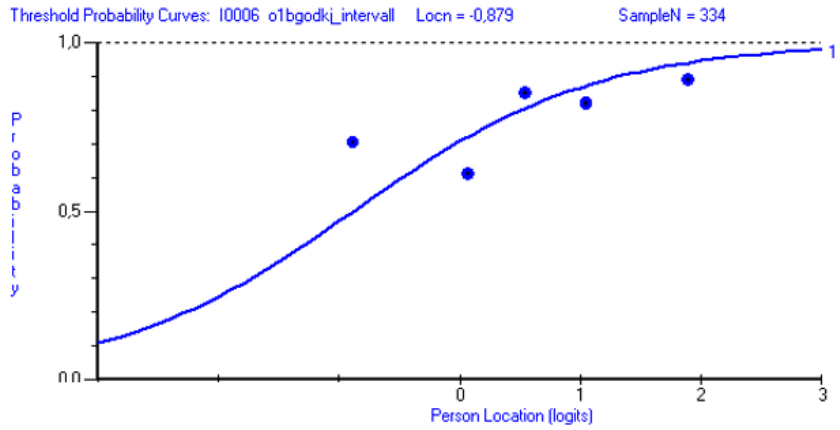
Tabell 5. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert henholdsvis 0, 1 og 2 poeng på 1b deloppgave 2. Analysen er basert på skåringsmodellen som ligger til grunn for Var2i.

Poeng	Gj.sn. z	Andel (%)
0	-0,55	12,3
1	-0,17	10,2
2	0,11	77,5

Et faglig argument for å gi 1 poeng til alle som oppgir enten et akseptabelt intervall for puls eller en enkeltverdi innenfor et akseptabelt intervall, er at studenter som oppgir en akseptabel enkeltverdi viser at de har kunnskap som bør krediteres. Vi valgte imidlertid å undersøke om rekodingen 2 -> 1, 1 -> 1 og 0 -> 0, som krediterer både intervall og enkeltverdi med 1 poeng, fungerer bedre enn rekodingen 2 -> 1, 1 -> 0 og 0 -> 0, som bare krediterer de som oppga et akseptabelt intervall. Disse rekodingene tilsvarer skåringsmodellene som ligger til grunn for henholdsvis Var2iii og Var2ii i Tabell 2.

Figur 8 viser at rekodingen som krediterer både intervall og enkeltverdi fungerer bedre enn rekodingen som bare krediterer et intervall (Figur 7). Dataene viser at 10 % flere får poeng når vi krediterer både intervall og enkeltverdi (Tabell 7), enn om vi bare krediterer de som oppga et akseptabelt intervall (Tabell 6).

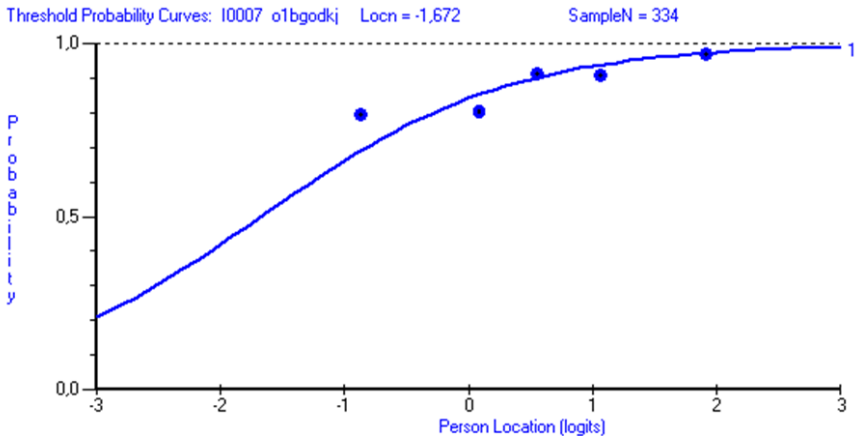
Skåringsmodellen som ligger til grunn for Var2iii medfører tap av informasjon om de som blir kreditert, for vi vet ikke hvem av de med 1 poeng som oppga et intervall og hvem som oppga en enkeltverdi. I praksis kan dette løses ved et tosifret kodesystem der første sifferet angir poeng og det andre sifferet angir type svar: kode 11 = intervall, kode 12 = enkeltverdi, og kode 01 = andre svar. Kode 11 og 12 rekodes deretter til 1 poeng i en ny variabel, mens kode 01 rekodes til 0 poeng.



Figur 7. Tilpasning til Rasch-modell basert på skåringsmodellen som ligger til grunn for variabelen Var2ii. Studenter som oppga et akseptabelt intervall på 1b deloppgave 2 «normalverdier for puls» fikk 1 poeng, mens de som oppga en enkeltverdi fikk 0 poeng. Målepunktene / observerte verdier viser hvordan poenget (vanskegrad -0,9) diskriminerer.

Tabell 6. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert henholdsvis 0 og 1 poeng på 1b deloppgave 2. Analysen er basert på skåringsmodellen som ligger til grunn for Var2ii.

Poeng	Gj.sn. z	Andel (%)
0	-0,36	22,5
1	0,11	77,5



Figur 8. Tilpasning til Rasch-modell basert på skåringsmodellen som ligger til grunn for variabelen Var2iii. Studenter som oppga et akseptabelt intervall eller en akseptabel enkeltverdi på 1b deloppgave 2 «normalverdier for puls» fikk 1 poeng. Målepunktene / observerte verdier viser hvordan poenget (vanskegrad -1,7) diskriminerer.

Tabell 7. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert henholdsvis 0 eller 1 poeng på 1b deloppgave 2. Analysen er basert på skåringsmodellen som ligger til grunn for Var2iii.

Poeng	Gj.sn. z	Andel (%)
0	-0,52	12,3
1	0,07	87,7

Kort oppsummert kan vi si at skåringsmodellene basert på Var1i for 1b deloppgave 1 «beskriv puls» og Var2i for 1b deloppgave 2 «normalverdier for puls» gir oss mest informasjon om studentenes kompetanse. Basert på psykometrisk evidens, vil vi kanskje velge skåringsmodellen basert på Var2iii for 1b deloppgave 2 «normalverdier for puls».

Avhengighet mellom 1b deloppgave 1 «beskriv puls» og 1b deloppgave 2 «normalverdier for puls»

Deloppgavene 1 og 2 viser ikke betydelig statistisk avhengighet (residualkorrelasjon < 0,2). I praksis betyr dette at 1b deloppgave 2 ikke oppfattes som «lettere» for de som ble kreditert på 1b deloppgave 1, enn for de som *ikke* ble kreditert på 1b deloppgave 1. Dette kan forklares ved at de fleste studentene i løpet av AFB-kurset registrerer pulsen på hverandre, og ved at de kan telle pulsen på seg selv under eksamen. Det å kunne oppgi en akseptabel enkeltverdi for puls er dermed uavhengig av om studenten vet at puls er en «trykkbølge». Oppgave 1 deloppgave 1 og deloppgave 2 kunne dermed blitt sensurert uavhengig av hverandre, og deleksamenen kunne dermed gitt oss noe mer informasjon om studentenes kunnskaper.

3.1.4 Oppsummering oppgave 1b

Sensorenes poenggiving på 1b er vanskelig å begrunne faglig, for vi vet ikke hva studenter med 1 poeng på oppgaven «kan». Det er dessuten ikke konsensus blant sensorene om svar av typen «puls er antall slag i minuttet» fortjener kreditt på den første delen av oppgaven «beskrive puls». Det følger heller ikke av sensorveiledningen om sensorene skal kreditere akseptable enkeltverdier på oppgavens siste del om «normalverdier for puls». Våre analyser tyder på at en kunne gitt opptil 2 poeng på oppgavens første del og opptil 1 poeng på oppgavens siste del, og at de to deloppgavene kunne vært kreditert uavhengig av hverandre. Denne måten å vurdere

oppgaven på gir oss langt mer informasjon om studentenes kunnskaper enn det den originale sensuren gjorde.

Vi fant at svar av typen «puls er trykkbølge» er assosiert med høy kompetanse i fysiologi, mens svar av typen «antall slag i minuttet» er assosiert med lav kompetanse i fysiologi. Det å kunne oppgi et akseptabelt intervall for normalpuls er også assosiert med lav kompetanse i fysiologi. Vi kan trekke disse slutningene fordi poengene som representerer disse typene svar har henholdsvis høy og lav vanskegrad.

Vi anbefaler at oppgaver spør om én ting, og at sensorveiledningene tydeligere beskriver kjennetegn eller karakteristikker på svar som fortjener delvis kreditt og svar som fortjener full kreditt. Per i dag beskriver sensorveiledningene bare hva som forventes av et fullgodt svar.

3.2 Oppdragets pkt.2: identifisere faglige vurderingskriterier – oppgaver med uordna poengkategorier

Uordna kategorier signaliserer, som nevnt, at vurderingen av besvarelser ikke har fungert optimalt. Uordna kategorier opptrer typisk når sensorene anvender flere poengkategorier enn det antallet dyktighetsnivåer oppgaven greier å avdekke blant studentene. Uordna kategorier kan også opptre når sensorveiledningen i for liten grad hjelper sensorene med å definere kriteriene for de ulike poengkategoriene (jf. oppdragets pkt. 2). Analysene til Guttersrud (2018) viste at totalt 11 oppgaver hadde uordna poengkategorier (1a, 1d, 2b, 2d, 3a, 3b, 4c, 5b, 5c og 6c). Vi går her i dybden på oppgave 4c og 6c fra deleksamen i AFB i desember 2017, siden disse oppgavene også spurte om «mer enn én ting».

3.2.1 Oppgave 4 c – sensorveiledningen og eksempler på kreditering

Oppgave 4c «*Beskriv hvor de tre hovedtypene av muskulatur finnes i kroppen, og hvordan hver av dem påvirkes av nervesystemet*» er et eksempel på en eksamensoppgave som har flere svakheter, men oppgaven er god på den måten at den kopler sammen kroppens ulike systemer. Oppgaven forventer at studentene navngir tre hovedtyper av muskler, beskriver hvor i kroppen hver av disse tre hovedtypene er lokalisert, og hvordan nervesystemet påvirker hver av dem. Slik sett ber oppgaven studentene svare på ni forskjellige ting, men sensorveiledningen er utformet slik at navnene på muskeltypene er en forutsetning for å få poeng, men ikke

tilstrekkelig for å få poeng. Oppgaven skal da tolkes til å etterspørre seks forskjellige ting, og oppgaven ga opptil 6 poeng. Sensorveiledningen ga følgende informasjon:

Skjelettmuskulatur:

- *Finnes i bevegelsesapparatet.*
- *Utfører viljestyrte muskelsammentrekninger og påvirkes av det somatisk-motoriske nervesystemet. (2 poeng)*

Glatt muskulatur:

- *Finnes hovedsakelig i indre organer og blodårer.*
- *Utfører ikke-viljestyrte muskelsammentrekninger og påvirkes av det autonome nervesystemet. (2 poeng)*

Hjertemuskulatur:

- *Finnes bare i hjertet.*
- *Utfører ikke-viljestyrte muskelsammentrekninger og påvirkes av det autonome nervesystemet. (2 poeng)*

På oppgave 4c er det vanskelig å begrunne poengene faglig. Årsaken er at vurderingskriteriene for hvert av de seks poengene som skal deles ut ikke er gjensidig utelukkende eller «tydelig beskrevet i sensorveiledningen» (jf. oppdragets pkt. 2). Det er med andre ord vanskelig å vite hva studenter med 1-5 poeng på oppgaven «kan».

Studenter med 1 poeng på oppgaven kan f.eks. ha svart at «skjelettmuskulatur» er viljestyrt (uten å nevne at skjelettmuskulaturen finnes i bevegelsesapparatet). Andre studenter med 1 poeng på oppgaven kan f.eks. ha svart at «hjertemuskulatur» er en type muskel (kvalifiserer ikke for kreditt), og at denne muskeltypen bare finnes i hjertet (skal krediteres). Studenter med 2 poeng kan f.eks. ha svart at «glatt muskulatur» er viljestyrt og finnes i indre organer. Det er med andre ord vanskelig å begrunne faglig at studenter med ulik skår på oppgaven har tydelig

forskjellig kunnskap om emnet. Basert på antall poeng en student har fått på oppgaven, vet vi lite om *hvilke* muskeltyper studenten kan referere til, om studenten vet *hvor* i kroppen vi finner ulike muskeltyper, og om studenten vet noe om *hvordan* nervesystemet påvirker muskler.

Nedenfor er det et eksempel på svar på 4c som sensorer ga 2 poeng. Vår kommentar til besvarelsen står i fet type i klammeparentes.

- *Vi har glatt muskulatur, tverrstripa muskulatur. Disse to typene finner du i for eksempel hjertet. Muskler påvirkes av nervesystemet ved at nevrotransmittere skilles ut fra vesikler i presynaptisk membran. Eksempel på nevrotransmittere i en nevromuskulær synapse er Acetylcolin, dopamin, adrenalin, norandrenalin, GABA og glutamat.***[Studenten nevner to muskeltyper, men beskriver ikke hvordan *hver* av dem påvirkes av nervesystemet. Studenten mestrer dermed ikke sammenkoplingen av muskelsystemet og nervesystemet. Studenten ramser opp en rekke transmittere uten å skille mellom hvilke som er typiske for det somatisk-motoriske nervesystemet og hvilke som er typiske for det autonome nervesystemet].**

Nedenfor er det et eksempel på svar på 4c som sensorer ga 1 poeng. Vår kommentar til besvarelsen står i fet type i klammeparentes.

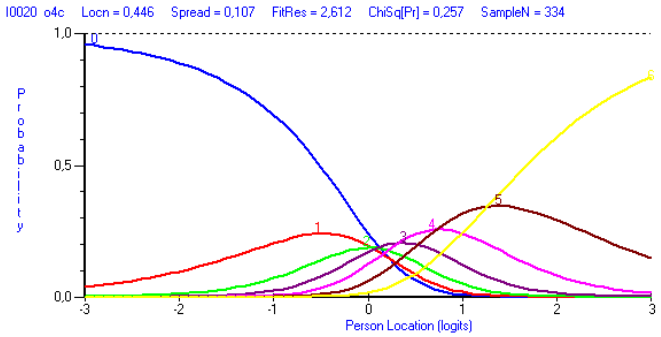
- *Glatt muskulatur finner vi i indre organer, den er ikke viljestyrt, den styres av det autonome nervesystemet. Tverrstripet muskulatur finner vi i skjelettmuskulaturen, den er viljestyrt og styres av det motoriske nervesystemet* **[Studenten nevner to muskeltyper og gjengir korrekt hvor muskeltypene finnes, og hvordan de påvirkes av nervesystemet. Svaret kunne vært mer utfyllende, men vi ville likevel forventet at studenten fikk mer enn 1 poeng].**

Ut fra disse eksemplene er det vanskelig å begrunne faglig at studenten som ble kreditert 2 poeng er dyktigere enn studenten som ble kreditert 1 poeng. Eksemplene illustrer også at sensorveiledningen må være tydeligere på hva som skal kreves for at studentene skal få hvert av poengene 1-5.

3.2.2 Analyser av data for oppgave 4c basert på original sensur

Dataene i Tabell 8 styrker hypotesen om at det er vanskelig å begrunne faglig at studenter med ulik skår på oppgaven har tydelig forskjellig kunnskap om emnet. Tabell 8 viser at det ikke er betydelig forskjell i dyktighet (målt ved z-skår) mellom studenter som fikk 0 poeng og studenter som fikk 1 poeng på 4c, og med «betydelig» mener vi her forskjell i z-skår større enn 0,5. Det er heller ikke betydelig forskjell i dyktighet mellom studenter som fikk 2 poeng og studenter som fikk 3 poeng på 4c, og tilsvarende gjelder for studenter med 5 poeng og studenter med 6 poeng. Det er da naturlig å tenke seg at poenggivingen ikke kan forsvares psykometrisk, og Figur 9 viser nettopp at oppgave 4c har uordna poengkategorier. Når poengene verken kan begrunnes faglig eller forsvares psykometrisk, har vi ikke et godt verktøy for å vurdere studentenes kunnskaper. På 4c ble sensorene dermed gitt et umulig oppdrag, for de skulle skille mellom flere dyktighetsnivåer enn det oppgaveformuleringen og vurderingskriteriene greier å avdekke. Gitt at eksamener sjelden eller aldri produserer tilstrekkelig mengde informasjon/ høy nok reliabilitet til pålitelig å gruppere studenter i seks grupper / dele ut **seks** karakterer A-F, framstår det urimelig at en skal kunne skille pålitelig mellom **sju** nivåer av dyktighet (0-6 poeng) basert på bare én oppgave.

Slik oppgave 4c er utviklet, og den tilhørende sensorveiledningen med vurderingskriterier er konstruert, vet vi bare hva studenter med 0 og 6 poeng på oppgaven «kan». Ifølge analysen til Guttersrud (2018) fikk 20 % av studentene 0 poeng og 15 % fikk 6 poeng (n = 5062). Det betyr at nasjonal deleksamen i AFB ikke gir oss informasjon om hva 65 % eller 2/3 av studentene kan og ikke kan om det fagstoffet oppgave 4c etterspør. Dersom 4c hadde vært erstattet av f.eks. 4c1 ««*Beskriv hvor i kroppen vi finner glatt muskulatur*» (1 poeng) og f.eks. 4c2 *«Beskriv hvordan nervesystemet påvirker skjelettmuskulaturen»* (1 poeng), hadde sensorene hatt en enklere jobb, og sensorenes poengsetting kunne gitt oss eksplisitt informasjon om studentenes kunnskaper.

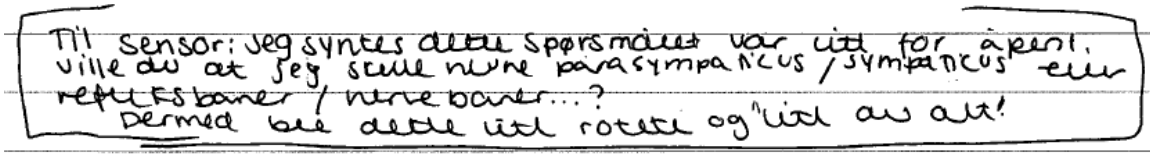


Figur 9. Sannsynlighetskurver basert på original sensur av 4c. Kurvene viser sannsynligheten for å oppnå 0-6 poeng (andreaksen) på 4c som funksjon av dyktighet (førsteaksen).

Tabell 8. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert med poeng 0-6 på oppgave 4c. Analysen er basert på original sensur av 4c.

Poeng	Gj.sn. z	Andel (%)
0	-1,05	18,0
1	-0,89	9,9
2	-0,38	10,5
3	-0,19	11,7
4	0,27	14,4
5	0,68	17,7
6	1,00	18,0

Følgende sitat viser at ikke bare sensorene fikk en utfordring på 4c, for studenter opplevde også at det var vanskelig å vite hvilke kriterier de ville bli vurdert etter:



Studentsitatet kan forklare hvorfor det var store forskjeller i hvor detaljert studentene besvarte delspørsmålet om hvordan nervesystemet påvirker muskeltypen. Noen studenter har bare nevnt om muskeltypen påvirkes av det autonome eller det somatisk-motoriske nervesystemet, mens andre studenter gjør rede for innerveringen og beskriver sympatisk og parasympatisk stimulering. De studentene som trekker inn dette viser en betydelig større forståelse av emnet, men sensorveiledningen har ingen «kumulativ struktur» slik at den fanger opp og krediterer denne typen svar. Vi fant dessuten at en del studenter feilaktig svarer at skjelettmuskulatur og tverrstripet er to ulike typer muskulatur.

3.2.3 *Analyser av data for oppgave 4c basert på «ny vurdering» av de utvalgte besvarelsene* og tydeligere definerte vurderingskriterier

Som for oppgave 1b, var det for oppgave 4c behov for å definere tydeligere vurderingskriterier, og vurdere besvarelser på nytt på bakgrunn av disse vurderingskriteriene. Sensorveiledningen indikerer at studentene kan få 2 poeng dersom de beskriver hvor muskeltypen finnes og hvordan

den påvirkes av nervesystemet (den første delen av sensorveiledningens tre deler), men sensorveiledningen beskriver ikke hvordan sensorene skal kreditere når studenten bare oppgir hvor muskeltypen finnes. Vår gjennomgang av eksamensbesvarelser tyder på at sensorene krediterer det å beskrive *hvor* muskeltypen finnes, og *hvordan* nervesystemet påvirker muskeltypen, med 1 poeng hver. Dette er imidlertid bare én av flere måter å tolke sensorveiledningen på.

For hver muskeltype (skjelettmuskulatur, glatt muskulatur og hjertemuskulatur) opprettet vi én variabel for å kunne kreditere studenter som svarte *hvor* muskeltypen finnes, og én variabel for å kunne kreditere studenter som refererte til *hvordan* nervesystemet påvirker muskeltypen. Alle de 334 utvalgte besvarelsene ble kreditert 0 eller 1 poeng på hver av disse seks nye variablene. Hensikten med å opprette seks nye variabler var å kunne studere svaravhengighet mellom det å vite hvor en muskeltype finnes og hvordan nervesystemet påvirker den, samt å undersøke svaravhengighet på tvers av muskeltypene. I tillegg kunne vi enkelt beregne andelen studenter som kun svarte *hvor* muskeltypen finnes, som kun svarte *hvordan* nervesystemet påvirker muskeltypen, og andelen som refererte til begge deler (Tabell 9).

Tabell 9 viser at 5,7 % + 45,2 % ≈ 51 % vet hvor vi finner skjelettmuskulatur. Tilsvarende tall for hjertemuskulatur og glatt muskulatur er henholdsvis 58 % og 64 %. Resultatene viser at mellom halvparten og 2/3 av studentene kan gjengi navnet på de tre muskeltypene og beskrive hvor i kroppen vi finner dem. Overraskende nok er det færre som nevner skjelettmuskulatur enn glatt muskulatur og hjertemuskulatur. «Bare» 45 % oppgir både hvor skjelettmuskulatur finnes, og hvordan nervesystemet påvirker skjelettmuskulatur.

Tabell 9 rapporterer også gjennomsnittlig z-skår for ulike grupper av studenter. Dataene viser *ikke* entydig og betydelig forskjell i dyktighet mellom gruppen av studenter som bare oppga hvor en muskeltype finnes (ca. 6-7 %), og gruppa som bare refererte til hvordan nervesystemet påvirker den samme muskeltypen (varierer mellom ca. 4 % og 19 % avhengig av muskeltype): Forskjellene i gjennomsnittlig z-skår er lik eller mindre enn 0,5 for «kun finnes» og «kun påvirkes» i Tabell 9. Vi kan dermed forsvare at vi ga 1 poeng til de som bare oppga hvor muskeltypen finnes, og 1 poeng til de som bare refererte til hvordan nervesystemet påvirker muskeltypen. Vi ga altså her det «samme poenget» til to ulike typer kunnskap, men de to typene kunnskap synes ikke å representere betydelig forskjellig dyktighet. De studentene som refererte til både hvor vi finner muskeltypen og hvordan nervesystemet påvirker muskeltypen, har i

gjennomsnitt høy dyktighet (gjennomsnittlig z-skår varierer mellom ca. 0,7 og 0,8). Slike svar kan da krediteres 2 poeng. Siden vi opprettet mange variabler, har vi **ikke verifisert ulike mulige skåringsmodeller**.

Vi forventer ikke at sensorer rapporterer på seks variabler på enkeltoppgaver. Vårt beste råd er derfor å unngå å formulere oppgaver som i praksis spør om seks ulike ting.

Tabell 9. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som har oppgitt enten hvor muskeltypen finnes, hvordan den påvirkes av nervesystemet, eller begge deler (oppgave 4c).

Kode	**Skjelettmuskulatur**		**Glatt muskulatur**		**Hjertemuskulatur**	
	Gj.sn. z	**Andel (%)**	**Gj.sn. z**	**Andel (%)**	**Gj.sn. z**	**Andel (%)**
Feil eller blankt	-1,18	30,5	-1,28	24,9	-0,95	38,6
Kun finnes	-0,37	5,7	-0,55	7,2	-0,45	5,7
Kun påvirkes	0,13	18,6	-0,16	11,1	-0,08	3,9
Finnes og påvirkes	0.79	45,2	0,66	56,9	0,76	51,8

3.2.4 Oppsummering oppgave 4c

Oppgave 4c er et eksempel på en oppgave som etterspør så mange aspekter at det er umulig for sensorene å poengsette den på en måte som gir oss entydig informasjon om hva studentene kan og ikke kan. Det er dermed vanskelig å faglig begrunne de ulike poengene, og oppgave 4c gir dessuten uordna poengkategorier. Dersom 4c hadde vært erstattet av f.eks. 4c1 «*Beskriv hvor i kroppen vi finner glatt muskulatur*» (1 poeng) og f.eks. 4c2 *«Beskriv hvordan nervesystemet påvirker skjelettmuskulaturen»* (1 poeng), hadde sensorene hatt en langt enklere jobb. Dessuten kunne poengsettingen gitt oss mer eksplisitt informasjon om studentenes kunnskaper enn det 4c gjør.

Da vi observerte svaravhengighet mellom ulike muskeltyper, kunne en ha utviklet en oppgave som handlet om bare én av muskeltypene. Oppgaven «*Beskriv hvordan glatt muskulatur påvirkes av nervesystemet*», kunne hatt følgende sensorveiledning og skåringsmodell (jf. oppdragets pkt. 3 om kumulative vurderingskriterier:

0 poeng = feil eller manglende svar

1 poeng = oppgir at glatt muskulatur ikke er viljestyrt

2 poeng = oppgir at glatt muskulatur styres av det autonome nervesystemet

3 poeng = refererer til hva som skjer i muskulaturen ved sympatisk versus parasympatisk stimulering

Et annet alternativ kunne være å utvikle oppgave 4c til følgende flervalgsoppgave, hvor en enkelt kan erstatte «glatt muskulatur» med «skjelettmuskulatur» eller «hjertemuskulatur»:

Hvilken påstand er best?

A) Glatt muskulatur finnes i bevegelsesapparatet og i fordøyelsessystemet, og den reguleres av det motoriske nervesystemet.

B) Glatt muskulatur finnes i indre organer og i blodårer, og den reguleres av det autonome nervesystemet.

C) Glatt muskulatur finnes i indre organer og i fordøyelsessystemet, og den reguleres av det motoriske nervesystemet.

D) Glatt muskulatur finnes i bevegelsesapparatet, og den reguleres av det autonome nervesystemet.

3.2.5 Oppgave 6c – sensorveiledningen og eksempler på kreditering

Oppgave 6c «*Nevn hva bukspytt inneholder, og hvilke funksjoner de ulike komponentene i bukspyttet har*» ber igjen studentene om å gjøre to ting. Ifølge sensorveiledningen gir oppgave 6c opptil 4 poeng:

- *amylase - spalter karbohydrater (1 poeng)*
- *proteaser (blant annet trypsin) - spalter proteiner (1 poeng)*
- *lipase - spalter fett (1 poeng)*
- *hydrogenkarbonat (HCO_3^-)/bikarbonat - nøytraliserer saltsyre fra ventrikkelen (1 poeng)*

Det forventes ikke at nukleaser er med i svaret.

Sensorveiledningen spesifiserer altså at sensorene skal gi ett poeng per riktig komponent med tilhørende funksjon. Det er dermed sannsynlig at sensorer krediterer halve poeng for komponent uten tilhørende funksjon. Sensorveiledningen spesifiserer ikke om «base» skal krediteres på lik linje med hydrogenkarbonat og bikarbonat.

Sensorveiledningen har ikke kumulative vurderingskriterier, og det er umulig å vite hva studenter med 1-3 poeng på oppgaven kan og ikke kan. En student kan f.eks. ha blitt kreditert 2 poeng for å oppgi at amylase spalter karbohydrater og at proteaser spalter proteiner, mens en annen student kan ha fått 2 poeng for å nevne alle fire komponentene uten å referere til noen av komponentenes funksjoner.

Punktlista nedenfor viser eksempler på sensorers kreditering av svar på 6c. Våre kommentarer til besvarelsene står i fet type i klammeparentes. Svar som ble gitt full kreditt (4 poeng):

- *Bukspytt inneholder enzymene amylase, lipase, protease og base, og hjelper tynntarmen bryte maten ned i sine minste bestanddeler, og tilføre spytt slik at det lettere kan absorberes nedover tynntarmen. Bukspytt skilles ut i øverste del av tynntarm (duodenum). Amylase bryter ned karbohydrater, lipase bryter ned fett, protease bryter ned proteiner. Base vedlikeholder pH.* **[Studenten svarer korrekt om de tre fordøyelsesenzymene, men oppgir «base» i stedet for hydrogenkarbonat (HCO_3-)/bikarbonat. At base vedlikeholder pH er også upresist, siden funksjonen er å nøytralisere det sure mageinnholdet fra ventrikkelen].**
- *Bukspytt er an blanding av hydrogenkarbonat (HCO_3-) fra kanalcellene og fordøyelsesenzymer fra acinarcellene. Deres funksjoner er: i) HCO_3- nøytraliserer den sure magesyra i tynntarmen slik at tarmslimhinnen ikke ødelegges. Det danner også riktig arbeidsmiljø for α-amylasen, ii) α-amylasen bryter ned karbohydrater til små disakkarider, iii) lipase bryter ned fett til fettsyrer og glyserol, iv) DNAser/RNAser bryter ned nukleinsyrer, v) trypsinogen, chymotrypsinogen, proelaktase og prokarboksypeptidase aktiveres til trypsin, chymotrypsin, elastase og karboksypeptidase, og bryter ned proteiner til små peptidfragmenter.* **[Studenten refererer til alle komponentene og deres funksjoner. Studenten kunne nevnt nedbrytning av disakkarider til monosakkarider, og nedbrytning av**

peptidfragmenter til aminosyrer. Studenten bruker flere fagtermer inkludert hydrogenkarbonat (i stedet for base). Dette svaret reflekterer mest sannsynlig høyere kompetanse enn det forrige svaret som også fikk 4 poeng].

Svar som ble gitt delvis kreditt (3 poeng [3,5]):

- *Amylase spalter polysakkaridene fra ventrikkelen til disakkarider. Lipase spalter triglycerider til monoglycerider og frie fettsyrer. Protease spalter polypeptider fra ventrikkelen til tri- og di-peptider. Base nøytraliserer den kraftige syra i magesaften når innholdet når duodenum.* **[Besvarelsen er av tilsvarende kvalitet som den første besvarelsen ovenfor (4 poeng). Vi antar at studenten har blitt trukket for å skrive base i stedet for hydrogenkarbonat].**
- *Lipase, amylase og protease nedbryter karbohydrater, protein, fett til mindre bindinger som kan enten absorberes i blodbanen eller i lymfeårer.* **[Studenten nevner fordøyelsesenzymene, men oppgir ikke hvilken komponent som har hvilken funksjon. Studenten beskriver ikke base/ hydrogenkarbonat].**

Svar som ble gitt delvis kreditt (2 poeng):

- *Protease bryter ned protein, lipase bryter ned karbohydrat, emylase bryter ned fett, base spalter enzymer.* **[Studenten oppgir fire komponenter, men skriver emylase i stedet for amylase og base for hydrogenkarbonat. Dessuten har studenten oppgitt feil funksjon for lipase, «emylase» og base. Uklart hvilken kompetanse studenten har, og hva studenten ble kreditert for].**
- *Bukspyttets innhold: enzymer som hjelper til nedbrytningen av næringsstoffer. F.eks. amylase, lipase. Skiller ut enzymet amylase som bryter ned karbohydrater til mindre disakkarider. Skiller ut lipase som bryter ned fett til mindre fettsyrer. Skiller ut transportprotein som transporterer næringsstoffene til epitelcellene i tarmen. Angiotensinogen som regulerer vann-og salt/elektrolyttopptaket i nyrene.* **[Studenten har oppgitt to av fire komponenter med tilhørende funksjoner, og er mest sannsynlig kreditert med 1 poeng for hver komponent].**
- *Bukspytt inneholder: amylase bryter ned karbohydrater, lipase bryter ned fett, protease bryter ned proteiner, base opprettholder riktig pH-verdi, motvirker syren.*

[Besvarelsen er tilnærmet identisk med sensorveiledningen, bortsett fra at studenten oppgir base i stedet for hydrogenkarbonat. Studenten er likevel kreditert bare 2 poeng].

Svar som ble gitt delvis kreditt (1 poeng):

- *Pancreas (bukspyttkjertelen) produserer lipase og amylase, rett miljø tarm enzym, ta opp næring, galle, spalting av fett. Bukspytt spalting av næringsstoff.* **[Vi antar at studenten har blitt kreditert for å nevne lipase og amylase. Studenten nevner ikke de øvrige komponentene, og studenten klargjør heller ikke komponentens funksjoner].**
- *Bukspytt inneholder enzymer og hormoner. Enzymer spalter næringsstoffer (karbohydrat, fett, protein).* **[Studenten nevner korrekt at bukspytt inneholder enzymer, men ikke hvilke. Vi antar at studenten har blitt kreditert for å nevne at enzymene spalter næringsstoffene karbohydrater, fett og proteiner].**
- *Bukspytt inneholder nøytraliserende stoffer som hjelper i fordøyelsen. I tolvfingertarmen blir matens innhold nøytralisert så den kan gå videre til tynntarmen. Grunnen er magesekkens surt miljø/pH.* **[Studenten oppgir ingen av komponentene, og studenten nevner bare at bukspyttet har en «nøytraliserende» effekt. Studenten oppgir ikke at bukspyttet inneholder base/hydrogenkarbonat. Det er tvilsomt om studenten fortjener 1 poeng].**

Svar som *ikke* ble kreditert (0 poeng):

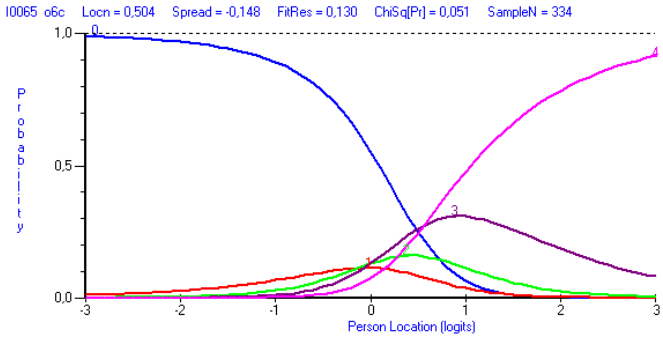
- *Pancreas produserer og skiller ut bukspytt. Bukspytt inneholder galle, saltsyre, Na+. Galle har funksjon ved at det skilles ut for å bryte ned fet mat/fett. Saltsyre har funksjon ved at det har lav Ph og skal bryte ned maten. Na+ (Natrium) får vi generelt lite av i kosten, så det skilles ut for å ha riktig balanse av det.* **[Studenten nevner verken navn på aller funksjonen til de ulike komponentene. Studenten skriver pH på feil måte, og det som står om salt er feil. Det er riktig at pancreas skiller ut bukspytt].**
- *Spyttets funksjoner: bøte maten/det som skal gjennom fordøyelsessystemet og kjemisk nedbrytning av det som skal gjennom fordøyelsessystemet. Spyttkjertelens funksjoner:*

skille ut enzymer som bidrar til kjemisk nedbrytning. **[Studenten skriver om spytt – ikke bukspytt slik oppgaven etterspør].**

Poengsettingen viser at sensorene også på oppgave 6c synes å kreditere noe ulikt. Noen sensorer krediterer «base», mens andre ikke gjør det. Da aksjonsverbet i oppgave 6c var «nevn», kan en ikke forvente beskrivelser av fordøyelsesenzymenes funksjon utover det å nevne komponentenes funksjoner. Den nederste av de to besvarelsene som ble kreditert 3 poeng, oppga imidlertid ikke komponentenes funksjoner.

3.2.6 Analyser av data for oppgave 6c basert på original sensur

Figur 10 viser at oppgave 6c har uordna kategorier, og det er dermed igjen evidens for at sensorene har delt ut flere poeng enn det antallet dyktighetsnivåer som oppgaven greier å avdekke hos studentene. Tabell 10 viser at det ikke er betydelige forskjeller i dyktighet mellom gruppene som fikk 1-3 poeng, fordi forskjellene i gjennomsnittlig z-skår er mindre enn 0,5. Fra et psykometrisk ståsted synes det derfor fornuftig å skåre oppgaven 0-1 poeng istedenfor 0-4 poeng, og én mulighet er å reskåre 0 -> 0, 1 -> 0, 2 -> 0, 3 -> 1 og 4 -> 1. Denne reskåringen kan ikke begrunnes faglig, for sensorene opererte her ikke med felles og faglig funderte vurderingskriterier. Fra et faglig ståsted kan derfor andre reskåringer være «riktigere».



Figur 10. Sannsynlighetskurver basert på original sensur av 6c. Kurvene viser sannsynligheten for å oppnå 0-4 poeng (andreaksen) på 6c som funksjon av dyktighet (førsteaksen).

Tabell 10. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert 0-4 poeng på oppgave 6c.

Poeng	Gj.sn. z	Andel (%)
0	-0,97	32,3

1	-0,06	6,3
2	0,04	10,5
3	0,35	19,8
4	0,79	31,1

3.2.7 Analyser av data for oppgave 6c basert på «ny vurdering» av de utvalgte besvarelsene, tydeligere definerte vurderingskriterier og reviderte skåringsmodeller

Vi opprettet fire nye variabler for oppgave 6c – én variabel for hver av komponentene amylase, protease, lipase og hydrogenkarbonat. Ved gjennomlesing av de 334 utvalgte besvarelsene, brukte vi følgende kodesystem for variablene for amylase, protease og lipase:

01 = feil eller ubesvar

11 = oppgir bare navn på komponenten

12 = oppgir bare funksjonen til komponenten

21 = oppgir *både* navn og funksjonen til komponenten

For variabelen for hydrogenkarbonat brukte vi følgende koder:

01 = feil eller ubesvart

11 = oppgir bare «base»

12 = oppgir bare «hydrogenkarbonat» eller «bikarbonat»

13 = referer bare til funksjonen til hydrogenkarbonat (f.eks. nøytraliserer syre)

21 = oppgir både «base» og funksjon

22 = oppgir både hydrogenkarbonat eller bikarbonat og funksjon

For hver av de fire nye variablene opprettet vi ytterligere tre nye variabler ved å reskåre fra koder til poeng (se Figur 11). Hensikten med å opprette ytterligere tre nye variabler for hver

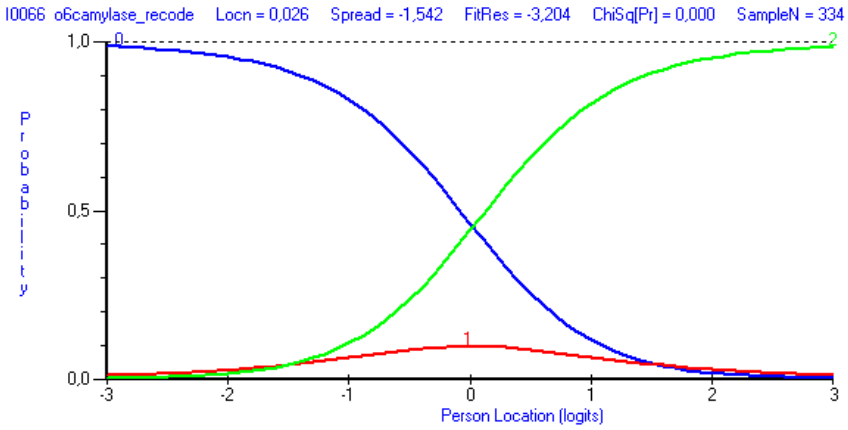
komponent, var å kunne studere svaravhengighet mellom navn og funksjon for én og samme komponent samt mellom ulike komponenter.

o6c_amylase_koder	o6c_amyNavn	o6c_amyFunk	o6c_amyPoeng
11	1	0	1
11	1	0	1
11	1	0	1
21	1	1	2
21	1	1	2
21	1	1	2

Figur 11. Utsnittet av SPSS-fila viser den nyopprettede variabelen «amylase» der studentenes besvarelser er kodet. Utsnittet viser også ytterligere tre nye variabler som ble opprettet for å undersøke svaravhengighet mellom type svar og mellom ulike typer komponenter (disse variablene er skårte variabler basert på rekoding av variabelen «amylase»).

Vi har altså tolket oppgave 6c som fire deloppgaver, og vi har poengsatt disse fire deloppgavene uavhengig av hverandre. Da kan vi evaluere poengsettingen av hver deloppgave og etterpå undersøke den statistiske avhengigheten mellom dem.

Vi prøvde ut en skåringsmodell der vi ga 1 poeng for å nevne navnet på komponenten, og 2 poeng for å nevne navnet og funksjonen. Figur 12 viser sannsynlighetskurvene for deloppgaven om «amylase», og figuren viser at deloppgaven for amylase hadde uordna poengkategorier. Det betyr at vurderingen ikke er optimal, og at vi kanskje burde gitt ett poeng for både navn *og* funksjon. Analyser viste at det samme gjaldt for de tre andre komponentene.



Figur 12. Sannsynlighetskurver for deloppgaven om «amylase» på oppgave 6c (én av fire deloppgaver). Kurvene viser sannsynligheten for å oppnå 0-2 poeng (andreaksen) på denne deloppgaven som funksjon av dyktighet (førsteaksen). Merk at de tre andre deloppgavene for de tre andre komponentene hadde tilsvarende sannsynlighetskurver.

Tabell 11. Gjennomsnittlig z-skår for gruppene av studenter (n = 334) som ble kreditert med poeng 0-2 på deloppgaven om amylase på oppgave 6c.

Kode	**Gj.sn. z**	**Andel (%)**
Feil eller ubesvart (kode 01)	-0,97	34
Kun navn (kode 11)	-0,07	5
Navn og funksjon (kode 22)	0,55	61

Figur 12 viser at vurderingen av deloppgaven om amylase på oppgave 6c ikke fungerer optimalt, fordi poengene ikke er «ordnet». Tabell 11 viser at gruppa av studenter som ble kreditert 1 poeng (dvs. kode 11) på deloppgaven om amylase på oppgave 6c, har relativt lav gjennomsnittlig dyktighet (z = -0,07). Da studentene som refererer til både amylase og funksjonen til amylase (kode 22) i gjennomsnitt er faglig dyktigere (z = 0,55), virker det rimelig å gi 0 poeng istedenfor 1 poeng til studentene med kode 11 og 12, og gi 1 poeng til studentene med kode 22. Denne poengsettingen kan begrunnes faglig og den kan forsvares psykometrisk.

Vi observerte svaravhengighet mellom det å vite navnene på de fire ulike komponentene (residualkorrelasjoner på mellom 0,45 og 0,81). Det er dermed «unødvendig» å etterspørre mer enn én av komponentene, for dette gir lite ny informasjon om studentene – det er altså slik at studenter som oppgir én av de fire komponentene gjerne oppgir også andre komponenter. Vi observerte også svaravhengighet mellom det å navngi en komponent og nevne funksjonen dens, og dessuten mellom navn og funksjon mellom de ulike komponentene. Sistnevnte gjaldt også for det å oppgi navn og funksjon for base. Imidlertid observerte vi ikke svaravhengighet mellom det å oppgi navn og funksjon for fordøyelsesenzymene, og det å oppgi navn og funksjon for hydrogenkarbonat eller bikarbonat. Det mest effektive hadde kanskje vært å bedt studentene oppgi funksjonen til én navngitt komponent, og heller brukt eksamenstiden til å teste andre typer kunnskaper og ferdigheter.

3.3 Oppdragets pkt.3: Sensorveiledninger med hierarkiske og kumulative kriterier

Figur 13 viser et eksempel på en flervalgsoppgave fra nasjonal deleksamen i AFB. Dersom vi gir 1 poeng for riktig svar, er det entydig hvilken kunnskap dette poenget representerer. Alle som fikk poenget «kan anvende anatomiske retningsbeskrivelser, som at kneleddet ligger proksimalt for ankelleddet».

7.1 Hvilket utsagn er riktig?

a) Ryggsøylen ligger lateralt for ribbena
b) Albuen ligger distalt for håndleddet
c) Kragebeinet ligger medialt for brystbeinet
d) Kneleddet ligger proksimalt for ankelleddet

Figur 13. Flervalgsoppgave 1 fra nasjonal deleksamen i AFB desember 2017. Svaralternativ D er riktig.

Sensorveiledningene til åpne oppgaver på eksamen i AFB desember 2017 beskriver bare kriteriene for hva en forventer av et fullgodt svar – svar som gir full kreditt. Sensorveiledningen i figur 14 har to svakheter ved at den i) ikke beskriver hva som kreves for «delvis kreditt» (1 poeng), og at den ii) ikke skisserer hvor «grensene mellom» poengene 0-2 går. Grensene kunne vært tydeliggjort ved å gi eksempler på svar en tenker seg at fortjener henholdsvis full, delvis og ingen kreditt. Vi kan anta at kandidater med 2 poeng på oppgave 2c «kan beskrive oksygenmetning som et mål på andelen jernatomer i hemoglobin som har bundet til seg oksygengass» (se sensorveiledningen i Figur 14). Det er imidlertid ikke entydig hvilken kunnskap kandidater med 1 poeng på oppgaven har. Det fremgår ikke hvor grensen går mellom 2 og 1 poeng eller mellom 1 og 0 poeng, og det er følgelig ikke mulig for sensorene å komme til konsensus om hvilke krav som skal stilles til svar som fortjener delvis kreditt (1 poeng).

Oppgave 2

c) Beskriv hva som menes med oksygenmetning. (2 poeng)

Sensorveiledning:
Oksygenmetning er et mål for hvor stor prosentandel av jernatomene i hemoglobin som har bundet til seg oksygen. (2 poeng)

Figur 14. Oppgave 2c (åpen oppgave) fra nasjonal deleksamen i AFB desember 2017.

Figur 15 viser sensorveiledningen til oppgave 2b. Oppgave 2b ber kandidatene «gjøre rede for», og det forventes dermed at kandidatene skal kunne vise utdypende forståelse av og kan begrunne et tema eller fenomen. Dette kan f.eks. være sammenhengen mellom biokjemisk/fysiologisk prosess og anatomisk oppbygning. Sensorveiledningen bestod av fem setninger/aspekter som indikerte hva en forventet av et fullgodt svar. Igjen skisseres bare kriteriene for et fullgodt svar som gir full kreditt (6 poeng). Det er ikke entydig hva poengene 1-5 betyr faglig, og det er ikke mulig å avgjøre hvor grensene mellom poengene går. Det var dermed opp til sensorene å avgjøre hvilke og hvor mange av aspektene som skulle kreves for hver av de fem ulike kategoriene av delvis kreditt eller 1-5 poeng (jf. oppdragets pkt. 2). Analysen til Guttersrud (2018) viste at oppgaven ga «*u*ordna poengkategorier». Analysen til Guttersrud (2018) viste også at det var *u*betydelige forskjeller i gjennomsnittlig dyktighet mellom gruppene av studenter som ble kreditert de ulike poengverdiene. Da f.eks. gruppa av studenter som fikk 3 poeng på oppgave 2b ikke var tydelig faglig dyktigere på eksamenssettet enn gruppa som fikk 2 poeng på oppgaven, er det uklart hva sensuren egentlig belønnet på denne oppgaven. Det var da uklart «hvilke faglige vurderingskriterier sensorene synes å ha benyttet for sin poenggivning».

Oppgave 2

b) Gjør rede for hvordan gassutvekslingen foregår mellom alveoler og lungekapillærer. (6 poeng)

Sensorveiledning:
Forskjeller i partialtrykk / konsentrasjon av O_2 og CO_2 i alveolluften og i lungekapillærene er en forutsetning for gassutvekslingen.
I alveolluften er konsentrasjonen av O_2 høyere enn i blodet som kommer til lungene, mens konsentrasjonen av CO_2 er høyere i blodet enn i luften i alveolene.
O_2 diffunderer fra alveolluften til kapillærene, mens CO_2 diffunderer fra blodet til alveolluften, inntil likevekt av begge gassene er nådd.
Veggen mellom alveolluften og blodet i kapillærene er tynn. Den korte diffusjonsavstanden fra alveoler til lungekapillærer er en forutsetning for tilstrekkelig diffusjon. (6 poeng)

Transport av O_2-molekyler og CO_2-molekyler i blodet ligger utenfor oppgaven.

Figur 15. Oppgave 2b (åpen oppgave) fra nasjonal deleksamen i AFB desember 2017.

Utfordringene ovenfor kan reduseres ved å avgrense oppgavene og definere sensorveiledninger der i) kriteriene er **hierarkiske og kumulative** i betydningen at full kreditt refererer til kriteriet

for delvis kreditt med ytterligere kriterier, eller ii) kriteriene er **hierarkiske** i betydningen at full kreditt refererer til kvalitativt sett utvilsomt faglig bedre svar enn kriteriet for delvis kreditt.

Nedenfor har vi skissert en **hierarkisk og kumulativ sensorveiledning** for oppgave 2b i Figur 15. Merk at svar som krediteres f.eks. 3 poeng inneholder det samme som 2 poeng, men med tilleggskriterier:

> Full kreditt (3 poeng): svar som refererer til at i) gassutvekslingen foregår ved diffusjon, ii) at forutsetningen for diffusjon er forskjeller i partialtrykk/konsentrasjon av O2 og CO2 i alveoler og kapillærnett, og som i tillegg iii) klargjør forutsetninger for diffusjon ved bruk av begrepene partialtrykk og diffusjonsavstand.
>
> Delvis kreditt (2 poeng): svar som bare refererer til at i) gassutvekslingen foregår ved diffusjon og ii) at forutsetningen for diffusjon er forskjeller i partialtrykk/konsentrasjon av O2 og CO2 i alveoler og kapillærnett.
>
> Delvis kreditt (1 poeng): svar som bare refererer til at i) gassutvekslingen foregår ved diffusjon.
>
> Ingen kreditt (0 poeng): andre typer svar.

Et annet eksempel på en **hierarkisk og kumulativ sensorveiledning** er en sensorveiledning der en får 1 poeng for å svare riktig på et gitt spørsmål, og hvor en får 2 poeng om en i *tillegg* forklarer hvorfor dette svaret er riktigst. Her kan en f.eks. tenke seg en flervalgsoppgave med tre svaralternativer kombinert med en begrunnelse for hvorfor ett av svaralternativene er best.

Et eksempel på en **hierarkisk *ikke*-kumulativ sensorveiledning** der kriteriet for full kreditt refererer til kvalitativt sett faglig bedre svar enn kriteriet for delvis kreditt, er gitt nedenfor. Den tidligere omtalte oppgave 1b «*Beskriv hva som menes med puls* [deloppgave 1], *og nevn normalverdier for puls i hvile hos voksne* [deloppgave 2]», kan avgrenses til deloppgave 1 «*Beskriv hva som menes med puls*» og krediteres på følgende måter (der eksempler er gitt for å utdype og definere grenser mellom poengene):

> Full kreditt (2 poeng): svar som refererer til «trykkbølge som sprer seg langs arteriene».
>
> - *Puls er en trykkbølge i blodårer*

Delvis kreditt (1 poeng): svar som refererer til «antall slag pr tidsenhet» eller tilsvarende.

- *Antall hjerteslag i minuttet*
- *Hjertefrekvens*

Ingen kreditt (0 poeng): andre typer svar eller ubesvart.

- Puls er det samme som blodtrykk

En alternativ sensorveiledning med **tosifrete** k**o**der der det første sifferet referer til poeng og det andre til type svar:

Full kreditt (2 poeng)

Kode 21: svar som refererer til «trykkbølge som sprer seg langs arteriene».

- *Puls er en trykkbølge i blodårer*

Delvis kreditt (1 poeng)

Kode 11: svar som refererer til «antall slag pr tidsenhet» eller tilsvarende.

- *Antall hjerteslag i minuttet*

Kode 12: svar som refererer til «frekvens» eller tilsvarende.

- *Hjertefrekvens*

Ingen kreditt (0 poeng)

Kode 01: andre typer svar

- *Puls er det samme som blodtrykk*

Kode 99: ubesvart (blanke svar).

Bruk av flere kategorier innenfor hvert poeng kan virke omfattende, men så lenge en begrenser antall koder kan systemet effektivisere rettingen. Én fordel med kodesystemet er at en tar vare

på informasjon som deretter kan brukes formativt inn i kvalitetsutvikling av undervisningen. En annen fordel er at en kan omdefinere krediteringen: Dersom en finner ut at f.eks. svar av typen «hjertefrekvens» likevel ikke bør krediteres, kan en omdefinere kode 12 til kode 02. Vi kan altså enkelt endre poengsettingen i ettertid uten å måtte rette besvarelser på nytt.

4 Oppsummering og anbefalinger for fremtidige eksamener

I denne rapporten har vi særlig diskutert problemstillinger knyttet til hvilke vurderingskriterier som ligger til grunn for kreditering av svar på åpne oppgaver (f.eks. kreditering av svar som refererer til «tetthetsbølge» og til «slag pr minutt» for oppgave 1b om puls), vi har reist problemstillinger knyttet til oppgaver som «spør etter mer enn én ting» (f.eks. 1b «beskriv puls» og «nevn normalverdier for puls»), og vi har undersøkt statistisk avhengighet mellom ulike deler av oppgaver (f.eks. svaravhengighet mellom ulike muskeltyper på oppgave 4c). Vi anbefaler at

- hver oppgave spør om én ting (jf. «beskriv puls» versus «beskriv puls og oppgi normalverdier for puls»)
- oppgaver som spør om én ting har et begrenset omfang (f.eks. spør hvordan nervesystemet påvirker glatt muskulatur – ikke hvordan nervesystemet påvirker både glatt muskulatur, hjertemuskulatur og skjelettmuskulatur)
- aksjonsverbene (f.eks. beskriv og forklar) er tydeligere «definert» for kandidatene, og at en fortsatt sikrer at aksjonsverbet i hver oppgave reflekterer de faktiske vurderingskriteriene (f.eks. at en enkeltoppgave ber kandidatene om å «forklare» dersom sensorveiledningen stiller krav om «forklaring»)
- en bruker et begrenset antall ulike aksjonsverb i læringsutbyttebeskrivelsene og de tilhørende eksamensoppgavene, og at en sikrer at aksjonsverbene i eksamensoppgavene tillegges samme tolkning som i læringsutbyttebeskrivelsene (for å ta høyde for at det i fremtiden er to ulike faggrupper som utvikler eksamen og læringsutbyttebeskrivelsene)
- aksjonsverbene brukes bevisst til å gi ønsket fordeling innenfor de kognitive kategoriene eksamen skal teste (andelen oppgaven innenfor eksempelvis «gjengi kunnskap», «anvende kunnskap» og «analysere/vurdere»)

- sensorveiledningene er hierarkiske i betydningen at de også beskriver kriteriene for delvis kreditt – ikke bare kriteriene som gjelder for svar som gir full kreditt (jf. «tetthetsbølge» og til «slag pr minutt»)
- sensorveiledningene tydeliggjør grensene mellom de ulike poengene på en oppgave (på oppgave 1b om puls burde sensorveiledningen f.eks. klargjort om svar av typen «hjertefrekvens» skal gi delvis eller ingen kreditt).
- det arrangeres et forsensurmøte der sensorene, etter å ha rettet noen få besvarelser hver, møtes og diskuterer problemstillingene som har oppstått.
- sensorene får opplæring i bruk av sensorveiledningene på f.eks. forsensurmøtet.
- en i ettertid sammenlikner karakterfordelingene til sensorparene for å kunne identifisere om noen legger kravene for høyt (er for strenge) eller for lavt, og at en gir tilbakemelding til aktuelle sensorer om dette.

I rapporten har vi gitt innspill til videreutvikling av oppgaver og sensorveiledninger med hierarkiske og tydelige vurderingskriterier, og hvordan sensorveiledninger med hierarkiske og tydelige vurderingskriterier også kan være kumulative. Sensorveiledninger med hierarkiske vurderingskriterier er en forutsetning for å kunne utvikle «mestringsbeskrivelser» eller «karakterbeskrivelser» – hva kandidater med ulike karakterer «kan» og «ikke kan». Grunnen er at slike sensorveiledninger gir oss informasjon om hva hvert poeng betyr faglig. Sensorveiledninger med hierarkiske vurderingskriterier kan dessuten bidra til å øke sensorreliabiliteten.

For å ytterligere tydeliggjøre hva hvert poeng som deles ut betyr faglig, bør en i fremtidige oppgaver unngå å spørre om mer enn én ting. En bør også reflektere over hvor mange dyktighetsnivåer en gitt oppgave kan avdekke, og begrense antall poengkategorier i henhold til dette. Antall poengkategorier bør ikke styres av oppgavens aksjonsverb (f.eks. beskriv, gi eksempel på, gjør rede for, forklar) eller hvor vanskelig en antar at oppgaven er. Oppgaver som ber kandidatene «gjør rede for» tolkes gjerne til å være vanskeligere enn en oppgave som ber om en «beskrivelse», og en opererer dermed typisk med flere poengkategorier i det første tilfellet. Dette kan selvfølgelig være fornuftig i noen tilfeller, men da under forutsetning av at «gjør rede for» faktisk avdekker flere dyktighetsnivåer i svarene enn «beskriv».

Vi finner få oppgaver der kandidatene må «forklare», og vi finner ytterst få oppgaver som ber kandidatene om å «analysere» eller «vurdere» (ofte kalt «høyere ordens kognitive ferdigheter»). Da Universitets- og Høgskolerådets (2011) karakterbeskrivelser legger til grunn at kandidatene også skal testes innenfor slike kognitive ferdigheter, bør en avgjøre om oppgavene skal teste mer enn bare det å «gjengi kunnskap». Læringsutbyttebeskrivelsene for emnet AFB skisserer kunnskaper og ferdigheter som er av betydning for faglig forsvarlig utøvelse av sykepleie. På bakgrunn av dette kunne en inkludere oppgaver der studentene må anvende kunnskaper for å f.eks. «forklare» ulike pasientobservasjoner.

5 Referanser

Guttersrud, Ø. (2018). Gjennomgang av eksamensoppgaver – nasjonal deleksamen i AFB, høst 2017. Innlegg presentert på evalueringsmøte i regi av NOKUT. Gardermoen, mars 2018.

Linacre, J. M. (1994). Constructing measurement with a many-facet Rasch model. In M. Wilson (Ed.), *Objective Measurement: Theory in Practice* (Vol. II). Newark NJ: Ablex.

Marais, I., & Andrich, D. (2008). Formalizing dimension and response violations of local independence in the unidimensional Rasch model. *Journal of applied measurement, 9*(3), 200-215.

Pedersen, L. F., Skeidsvoll, K. J., & Tokstad, K. (2018). Nasjonal deleksamen i anatomi, fysiologi og biokjemi i sykepleierutdanningen - høsten 2017: NOKUT.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests* (Expanded ed.). Chicago: University of Chicago Press.

Ringdal, K. (2007). *Enhet og mangfold : samfunnsvitenskapelig forskning og kvantitativ metode* (2. utg. ed.). Bergen: Fagbokforl.

Tokstad, K., & Hamberg, S. (2017). Nasjonal deleksamen – et pilotprosjekt og en mulighetsstudie: NOKUT.

Universitets- og Høgskolerådet (2011). Karaktersystemet – generelle, kvalitative beskrivelser. Hentet fra https://www.uhr.no/_f/p1/i4bfb251a-5e7c-4e34-916b-85478c61a800/karaktersystemet_generelle_kvalitative_beskrivelser.pdf