Kvalitetssikring av eksamensoppgaver: eksempler fra sykepleierutdanning

Tradisjonell skoleeksamen

De fleste eksamener i høyere utdanning gjennomføres som det vi kan kalle «tradisjonell skoleeksamen». De nasjonale deleksamenene for grunnskolelærerutdanningen og sykepleierutdanningen er typiske eksempler på tradisjonell skoleeksamen. Med tradisjonell skoleeksamen mener vi en skriftlig eksamen der studentene svarer på **avkrysningsoppgaver**, slik som flervalgsoppgaver, eller selv skriver et svar, som på **åpne oppgaver** (kortsvarsoppgaver). På slike oppgaver krediteres typisk studentene med poeng etter kvaliteten på svaret.

Nedenfor beskriver vi de helt sentrale prinsippene for gyldig og pålitelig måling av kunnskaper og ferdigheter på slike eksamener, men ideene og prinsippene kan overføres til andre summative vurderingsformer.

Faglige begrunnelser (før eksamen)

Til hver eksamen skal det utvikles sensorveiledning med tydelige **vurderingskriterier**. For avkrysningsoppgaver er kriteriene «riktig» eller «galt» avhengig av valgt svaralternativ. På åpne oppgaver finnes det ikke nødvendigvis riktige eller gale svar, men svar som fortjener eller ikke fortjener kreditt. For åpne oppgaver bør derfor vurderingskriteriene beskrive hva som kjennetegner faglig gode svar, middels gode svar, og faglig svake svar. Veiledningen bør også tydeliggjøre forskjellene mellom kategoriene ved å skissere eksempler på faglig gode svar, middels gode svar, og faglig svake svar.

Vurderingskriteriene bør utvikles slik at det er tydelige faglige forskjeller mellom gode og middels gode svar. Dette sikres best gjennom såkalte **kumulative** eller hierarkiske vurderingskriterier. Med kumulative vurderingskriterier mener vi i utgangspunktet et vurderingssystem der gode svar oppfyller kriteriene for middels gode svar, men i *tillegg* oppfyller ytterligere kriterier. Det er mulig å lage velfungerende vurderingssystemer som ikke har en kumulativ struktur, men det er viktig at «gode svar» er tydelig faglig bedre enn «middels gode svar».

Vurderingskriteriene bør ha en tilhørende **skåringsmodell** med heltallige positive poeng og null poeng. På flervalgsoppgaver er det naturlig at riktig svar gir full kreditt (1 poeng), og at feil svar ikke gir kreditt (0 poeng). Denne modellen kan også brukes for åpne oppgaver. Et eksempel på en annen skåringsmodell for åpne oppgaver er at gode svar gir full kreditt (2 poeng), at middels gode svar gir delvis kreditt (1 poeng), og at faglig svake svar ikke gir kreditt (0 poeng). I noen få tilfeller kan vi tydelig skille mellom mer enn tre poengkategorier, og da kan vi bruke følgende skåringsmodell: gode svar gir full kreditt (3 poeng), middels gode svar gir delvis kreditt (2 poeng og 1 poeng avhengig av svarets faglige kvalitet), og faglig svake svar gir ingen kreditt (0 poeng).

Empiriske begrunnelser (etter eksamen)

De fleste vil være enige i at en flervalgsoppgave eller en åpen oppgave er **rettferdig** dersom den belønner faglig dyktige studenter – gir flere poeng til faglig dyktige studenter enn faglig svake studenter. En annen måte å uttrykke dette på er at oppgaven **diskriminerer** eller skiller mellom studenter med høy og lav dyktighet. Dersom oppgavene i et eksamenssett er rettferdige på denne måten og kan forsvares ut fra læringsutbyttebeskrivelsene, har vi sikret at eksamen er et gyldig og pålitelig mål på kunnskaper og ferdigheter.

Etter gjennomføring av eksamen kan vi hente inn og sammenstille sensorenes poengsetting for å evaluere om vurderingskriteriene og de tilhørende skåringsmodellene resulterte i en gyldig og pålitelig måling av kunnskaper og ferdigheter. Den enkleste måten å avgjøre om en oppgave er rettferdig eller «virket godt som eksamensoppgave», er å tolke studentenes poengsum på hele eksamenssettet som et uttrykk for dyktighet. Vi antar da at studenter med høye poengsummer er faglig dyktigere enn studenter med lave poengsummer. I eksemplene nedenfor har vi standardisert studentenes poengsummer (z-skår: et mål på avvik fra gjennomsnittet, hvor gjennomsnittet er uttrykt som 0).

De to flervalgsoppgavene nedenfor (oppgave 5.12 og 5.18) er hentet fra nasjonal deleksamen for sykepleiere i desember 2016. Analysene viser at 64 % svarte riktig på 5.12, og at disse studentene i gjennomsnitt fikk 57 poeng (z = 0.36) av totalt 100 poeng på hele eksamenssettet. 36 % svarte galt på 5.12, og disse studentene i gjennomsnitt fikk 35,5 poeng (z = -0,64) av totalt 100 poeng på hele eksamenssettet. Avstanden i z-skår mellom de som svarte riktig og de som svarte galt viser at oppgaven totalt sett diskriminerer eller skiller godt mellom studenter med høy og lav dyktighet, og at dette er en svært god eksamensoppgave. Vi ser på tilsvarende måte at oppgave 5.18 skiller svakere mellom studenter med høy og lav dyktighet, og at oppgaven dermed er en litt urettferdig eksamensoppgave. Oppgave 5.12 har dermed bedre egenskaper som eksamensoppgave enn oppgave 5.18:

5.12 Hvilket utsagn om nyrene er riktig?
A. Filtrasjon av blodet foregår i distale tubuli
B. Råurin/preurin inneholder erytrocytter og store proteiner
C. Reabsorpsjon foregår i tubuli og samlerør
D. Sekresjon er transport av urin fra samlerør til nyrebekken
Riktig svar: C

Poeng	Gj.sn. poeng	Andel (%)
0	35,5	36
1	57,0	64
Total	49,3	100

Poeng	Gj.sn. z	Andel (%)
0	-0,64	36
1	0,36	64
Total	0,00	100

5.18 Hvilket utsagn om temperaturregulering ved feber er riktig?
A. Ved stigende feber er blodstrømmen i huden nedsatt
B. Ved stigende feber øker blodstrømmen i huden
C. Den forhåndsinnstilte temperaturen i hypotalamus senkes når feberen stiger
D. Det er typisk å få skjelvinger når feberen synker
Riktig svar: A

Poeng	Gj.sn. poeng	Andel (%)
0	45,4	61
1	55,4	39
Total	49,3	100

Poeng	Gj.sn. z	Andel (%)
0	-0,18	61
1	0,28	39
Total	0,00	100

Den åpne oppgaven om «puls» nedenfor er hentet fra nasjonal deleksamen for sykepleiere i desember 2017. Denne oppgaven ba uheldigvis studentene om to ting (deloppgave 1 «beskriv puls» og deloppgave 2 «nevn normalverdier for puls»). I sensuren så vi at den lite presise sensorveiledningen førte til at sensorene vurderte besvarelsene ut fra mer eller mindre individuelle kriterier, slik at svar på tilnærmet samme faglige nivå ble kreditert vidt forskjellig. Det var heller ikke mulig å se ut fra sensuren hva studentene var kreditert for:

> **1 b) Beskriv hva som menes med puls, og nevn normalverdier for puls i hvile hos voksne. (2 poeng)**
>
> Sensorveiledning: Puls er en trykkbølge som brer seg langs arterien som følge av hjertets kontraksjon. Normalverdi for hvilepuls hos voksne er ca. 50 – 80 slag/minutt (det bør utvises et visst skjønn når det gjelder svarene på normalverdier for puls).

Vi har analysert oppgaven som to uavhengige oppgaver. På deloppgave 1 er det gitt full kreditt (2 poeng) til gode svar som refererer til «trykkbølge», delvis kreditt (1 poeng) til middels gode svar som refererer til «antall slag per tidsenhet», og ingen kreditt (0 poeng) til andre typer svar. Det er da tydelig at gode svar holder høyere faglig kvalitet enn middels gode svar, og vi ser at oppgaven diskriminerer eller skiller godt mellom studenter med ulik dyktighet (Tabell X): Gruppa av studenter (16 %) som refererte til «trykkbølge» ble kreditert 2 poeng, og de skårte i gjennomsnitt bedre på hele eksamenssettet (z = 0,85) enn gruppa av studenter (70 %) som refererte til «antall slag per tidsenhet» (z = -0,06). En liten gruppe (14 %) av relativt sett faglig svake studenter (z = -0.70) fikk ikke

poeng på deloppgave 1. På deloppgave 2 ble alle som oppga et akseptabelt intervall eller en enkeltverdi innenfor intervallet kreditert 1 poeng (Tabell Y): Gruppa av studenter (88 %) som ble kreditert 1 poeng skårte i gjennomsnitt bedre på hele eksamenssettet (z = 0,07) enn gruppa av studenter (12 %) som ikke ble kreditert (z = -0,52). Ved bruk av de nevnte vurderingskriteriene og de tilhørende skåringsmodellene, deler vi altså ut poeng som kan begrunnes faglig og forsvares empirisk. Denne måten å rette oppgaven på resulterer dermed i en rettferdig vurdering.

Tabell X. Beskriv puls. Gjennomsnittlig z-skår for gruppene av studenter som ble kreditert henholdsvis 0, 1 eller 2 poeng på deloppgave 1 om puls.

Poeng	Gj.sn. z	Andel (%)
0	-0,71	14
1	-0,06	70
2	0,85	16
Total	0,00	100

Tabell Y. Normalverdier for puls. Gjennomsnittlig z-skår for gruppene av studenter som ble kreditert henholdsvis 0 eller 1 poeng på deloppgave 2 om puls.

Poeng	Gj.sn. z	Andel (%)
0	-0,52	12
1	0,07	88
Total	0,00	100