Tidsskriftet
DEN NORSKE LEGEFORENING

Eksamenslære for dummies

KRONIKK

PETTER GJERSVIK

E-post: petter.gjersvik@medisin.uio.no
Petter Gjersvik er professor ved Institutt for klinisk medisin ved Universitetet i Oslo, hvor han er undervisningsleder i hud- og veneriske sykdommer og leder av en eksamenskommisjon ved profesjonsstudiet i medisin. Han er også medisinsk redaktør i Tidsskriftet.
Forfatteren har fylt ut ICMJE-skjemaet og oppgir ingen interessekonflikter.

En eksamen i medisinstudiet skal teste kunnskaper og ferdigheter som studentene får bruk for som leger – ikke studentenes evne til å bestå eksamen. Hva er den beste eksamensformen? Hva tester man egentlig ved eksamen?

Vurdering og evaluering av studenters kunnskaper og ferdigheter – på engelsk kalt *assessment* – er en egen disiplin innen fagområdet medisinsk utdanning (1, 2). Kall det gjerne *eksamenslære*, selv om vurderinger også kan skje på andre måter enn gjennom en eksamen. Det finnes lærebøker, kurs og vitenskapelige konferanser som bare dreier seg om evalueringsmetoder innen utdanning. Det er utviklet prosedyrer og regler for hvordan eksamensoppgaver bør lages og hvordan svar skal vurderes og skåres, og kvantitative forskningsmetoder for å vurdere hvor godt eksamensoppgaver fungerer og hvordan karaktersettingen bør være slik at den oppleves som rettferdig og pålitelig.

En meningsfull diskusjon om eksamen er avhengig av en god dialog mellom faglærere, eksamenseksperter og studenter. For faglærere, eksaminatorer og sensorer kan møtet med eksamensteori være vanskelig, fordi det er etablert en terminologi og logikk som kan virke uklar, uheldig og forvirrende, særlig når ordene som brukes, er engelske.

Her følger en gjennomgang av sentrale eksamenstermer, de fleste først og fremst knyttet til skriftlige eksamensformer, med en ikke-eksperts forsøk på å forklare dem for nybegynnere. Styrker og svakheter ved ulike eksamensformer og oppgavetyper omtales kort, for termene forstås best i kontekst. For mer presise og utfyllende forklaringer henvises leseren til faglitteratur (2). Temaet bør interessere både studenter og leger, særlig faglærere ved våre medisinske fakulteter, for vurderingsmetoder påvirker både studenters og legers læringsatferd.

Det grunnleggende

Formative tester betegner tester som bl.a. brukes for å avklare hvor mye studentene kan på forhånd, slik at underviser tilpasser undervisningen til studentenes forutsetninger (1, 2). Slike tester kan bidra til økt motivasjon for læring og bedre studievaner (3), men de har ellers ingen konsekvenser for den enkelte. *Summative tester* betegner tester som har til hensikt å kontrollere hva studentene har lært, og innebærer at det settes en karakter, enten bestått / ikke bestått eller etter en gradert karakterskala. Kort sagt en eksamen.

Validitet og *reliabilitet* er velkjente termer for alle som driver eller bruker forskning. Høy validitet innebærer at testen tester det den er ment å teste, i denne sammenhengen studentenes kunnskaper og ferdigheter innen representative deler av faget. Med andre ord: Er eksamensformen og eksamensoppgavene dekkende og passende? Høy validitet er det aller viktigste. Høy reliabilitet innebærer at testens utfall, dvs. karakteren, er pålitelig og reproduserbar. Karakteren må oppfattes som rettferdig av studentene.

Flervalgsoppgaver = avkrysningsoppgaver

Testing av medisinstudenters kunnskaper gjøres nå ofte i form av en digital eksamen, dvs. med bruk av PC. Basert på erfaringer fra bl.a. USA brukes i økende grad såkalte *multiple choice questions*, ofte forkortet *MCQ* (1, 2). På norsk kan slike oppgaver kalles *flervalgsoppgaver* (4), men *avkrysningsoppgaver* er også dekkende. Slike oppgaver innebærer at det oppgis flere svaralternativer, vanligvis tre–fem, der kun ett er det «mest riktige svar» (*single best answer*). De andre svaralternativene kalles *distraktorer*. Disse svaralternativene må kunne oppfattes som plausible, må ikke skille seg ut, og må ikke være åpenbart gale, men altså være mindre riktige enn «mest riktige svar». *Flerresponsoppgaver* (*multiple response questions*) er en variant av flervalgsoppgaver der kandidatene skal velge to–tre riktige av fem–åtte oppgitte svaralternativer. Det finnes også andre, mindre brukte oppgavetyper.

Den viktigste fordelen med slike avkrysningsoppgaver er at de kan besvares på kort tid. En eksamen med kun slike oppgaver kan derfor inneholde flere oppgaver enn ellers og dermed dekke større deler av faget. Dessuten gjøres skåringen automatisk, ettersom bare ett svaralternativ er «mest riktig» (på flerresponsoppgaver flere). Ved å starte oppgaven med en beskrivelse av en klinisk situasjon vil oppgaven kunne illudere en klinisk beslutningsprosess (5). Oppgavesett med gode flervalgsoppgaver er vist å skille ganske godt mellom sterke, middels sterke og svake studenter (2).

Studenter lærer seg å identifisere riktig svar ut fra hvordan svaralternativene er formulert – de blir testsmarte

Gode avkrysningsoppgaver er imidlertid vanskelige å lage, og ikke alle temaer egner seg like godt. Kritikk mot denne type oppgaver går også ut på at de ikke gjenspeiler klinisk virkelighet på en særlig god måte, at de i begrenset grad tester kandidatenes evne til refleksjon og kunnskapsanvendelse, og at studentenes studieatferd påvirkes negativt (2, 6). Dessuten er muligheten stor for å få riktig svar ved ren gjetting: 25 % ved fire svaralternativer og henholdsvis 33 % og 50 % hvis kandidaten klarer å identifisere ett eller to svaralternativer som gale. Noen kandidater vil kjenne igjen riktig svar når de leser de oppgitte svaralternativene (det som på engelsk kalles *cueing*, på norsk *gjenkjenning* av *stikkord*). Offentliggjøring av tidligere eksamensoppgaver og erfaring fra tidligere eksamener bidrar til at studenter lærer seg å identifisere riktig svar ut fra hvordan svaralternativene er formulert – de blir testsmarte (fra engelsk *test wise*, *street smart*). Disse ulempene med flervalgsoppgaver blir ofte undervurdert og underkommunisert.

Fritekstoppgaver = kortsvaroppgaver

Oppgaver ved digital eksamen kan også innebære at kandidatene skal svare med en kort tekst. Slike oppgaver bør kalles *fritekstoppgaver* eller *kortsvaroppgaver* (4).

Mange kaller fritekstoppgaver for *essayoppgaver* (eller *miniessayoppgaver*) (7), men dette er i beste fall misvisende. Essay er en litterær sakprosasjanger med lange tekster publisert først og fremst i tidsskrifter og bøker, nærmest som en liten avhandling (8). Uttrykket gir assosiasjoner til en gammeldags eksamensform som for lengst er avskaffet, der kandidatene ble bedt om å skrive en lengre utredning om et oppgitt emne. Å skrive et essay er altså det stikk motsatte av hva man ønsker at studentene skal gjøre ved en digital eksamen, nemlig å svare på en oppgave med en kort, presis og konsis tekst, gjerne stikkordspreget eller med kun én–to setninger. Å bruke ordet *essay* i denne sammenhengen er snarere en oppfordring til å skrive langt, noe mange studenter dessverre gjør, særlig når de ikke helt vet hva de skal

svare.

Fordelen med fritekstoppgaver er at kandidatene må svare uten hjelp av oppgitte alternativer, på samme måten som leger må agere i klinisk virksomhet (6). Slike oppgaver vil ofte gi et sannere og mer autentisk bilde av kandidatens kompetanse (9, 10). Men også fritekstoppgaver og skåringsveiledninger kan være vanskelige å lage. Skåring av besvarelsene er tidkrevende og kan variere avhengig av skårers bakgrunn og forutsetninger. Konsistent skåringspraksis kan likevel fremmes med gode skåringsveiledninger, forhåndstrening og konsensusskåring, dvs. at skårerne justerer sine skårer ved store avvik. Antallet fritekstoppgaver kan ikke være for høyt, fordi det kan ta noe lengre tid å besvare dem. I England er det utviklet et dataprogram med *very short answer questions*, der svarene kan skåres ved hjelp av et dataprogram (9), slik jo skåringen av svarene på flervalgsoppgaver gjøres.

Psykometri

Det er utviklet en rekke kvantitative forskningsmetoder for å vurdere hvor godt eksamensoppgaver fungerer (2). Er en oppgave lett (nesten alle svarer riktig) eller vanskelig (nesten ingen svarer riktig)? Hvordan er fordelingen av oppgavene etter deres vanskelighetsgrad (eng. *item difficulty*) i et oppgavesett? Hvor godt skiller en oppgave sterke, middels sterke og svake kandidater fra hverandre? Hvis omtrent like mange studenter har valgt hvert av svaralternativene, tyder dette på at de i stor grad har gjettet. Slike metoder gjør at eksamenskommisjonen kan identifisere eksamensoppgaver som ikke har fungert tilfredsstillende, og vurdere å ta dem ut av sensurgrunnlaget (7).

Tilsvarende er det utviklet metoder som vurderer avvik og presisjon ved skåringer av besvarelser på fritekstoppgaver (1, 2). Avviker skåringene fra én skårer for mye fra skåringene fra andre skårere? Slike metoder om *skåringssamsvar* kan identifisere skårere som er «for snille» eller «for strenge» på én eller flere oppgaver, og eventuelt justere skåringer der avviket er for stort, dvs. *skåringsjustering* (eng. *rater alignment*).

Disse metodene for kvalitetssikring av eksamensoppgaver og skåringspraksis kalles *psykometri*. I denne sammenhengen betyr altså psykometri ikke måling av personers psykologiske egenskaper, slik man gjerne skulle tro, men måling av hvordan eksamensoppgaver hver for seg og samlet fungerer, og i hvilken grad vurderingen av kandidatenes besvarelser har skjedd på en konsistent og pålitelig måte.

Standardsetting

Sensur er fastsettelse av endelig karakter, enten i form av bestått / ikke bestått eller etter en gradert karakterskala, f.eks. A–F, der A er beste karakter og F ikke bestått. Det viktigste, og ofte vanskeligste, er å fastsette grensene for bestått / ikke bestått. Slike prosesser kalles *standardsetting* (2).

Ulike eksamensformer og oppgavetyper har sine styrker og svakheter, og meningene om dem er mange og ofte motstridende

Ideelt burde vanskelighetsgraden av oppgavesett holdes stabil fra år til år, men dette er i praksis vanskelig å få til (11). *Relativ standardsetting* tar utgangspunkt i alle studentenes prestasjoner og en vurdering av oppgavenes vanskelighetsgrad. *Absolutt standardsetting* innebærer at grensen for bestått er bestemt på forhånd. Det er utviklet en rekke matematiske modeller for hvordan grensen for bestått kan fastsettes, men disse er kompliserte og ressurskrevende (2, 11).

I praksis vil en eksamenskommisjon i fastsetting av beståttgrensen oftest basere seg på en pragmatisk tilnærming der man benytter faglig skjønn. Ved graderte karakterer kan de øvrige karakternivåene fastsettes med utgangspunkt i beståttgrensen og en tilsvarende vurdering av grensen for beste karakter, A.

Mangfold og helhet

En eksamen skal teste kunnskaper og ferdigheter som studentene vil ha bruk for som leger – ikke studentenes evne til å bestå eksamen. Ulike eksamensformer og oppgavetyper har sine styrker og svakheter, og meningene om dem er mange og ofte motstridende (2, 12). Snarere enn å velge den ene oppgaveformen fremfor den andre, bør en digital eksamen inneholde *både* avkrysningsoppgaver og fritekstoppgaver. De praktiske utfordringene dette innebærer, er håndterbare og overkommelige. I tillegg må man ha kliniske og muntlige eksamener som bedre tester studentenes resonneringsevne og kliniske ferdigheter enn det en digital eksamen gjør. Slike eksamener kan i stor grad gjøres standardiserte for å sikre høy validitet og konsistent skåringspraksis.

Hensikten med eksamen i medisinstudiet er at samfunn, helsevesen og pasienter skal være sikre på at universitetene utdanner leger som duger. Dessuten virker testing og eksamen motiverende for læring. Så enten man liker det eller ikke: Eksamen er viktig.

LITTERATUR:

1. Epstein RM. Assessment in medical education. N Engl J Med 2007; 356: 387–96. [PubMed][CrossRef]

2. Schuwirth LWT, Ash J. Principles of assessment. I: Walsh K, red. Oxford Textbook of Medical Education. Oxford: Oxford University Press, 2013.

3. Larsen DP, Butler AC. Test-enhanced learning. I: Walsh K, red. Oxford Textbook of Medical Education. Oxford: Oxford University Press, 2013.

4. NTNU. Eksamensoppgaver – medisin – MH. https://innsida.ntnu.no/wiki/-/wiki/Norsk/Eksamensoppgaver+-+Medisin+-+MH Lest 15.2.2020.

5. Schuwirth LWT, Verheggen MM, van der Vleuten CP et al. Do short cases elicit different thinking processes than factual knowledge questions do? Med Educ 2001; 35: 348–56. [PubMed][CrossRef]

6. Gjersvik P. Eksamen og legers virkelighet. Tidsskr Nor Legeforen 2018; 138. doi: 10.4045/tidsskr.18.0472. [CrossRef]

7. Universitetet i Oslo. Medisin (profesjon). Oppbygging og gjennomføring. https://www.uio.no/studier/program/medisin/oppbygging/ Lest 15.2.2020.

8. Skei HH. Essay. I: Store norske leksikon. https://snl.no/.search?queryessay. Lest 18.2.2020.

9. Sam AH, Field SM, Collares CF et al. Very-short-answer questions: reliability, discrimination and acceptability. Med Educ 2018; 52: 447–55. [PubMed][CrossRef]

10. Sam AH, Westacott R, Gurnell M et al. Comparing single-best-answer and very-short-answer questions for the assessment of applied medical knowledge in 20 UK medical schools: Cross-sectional study. BMJ Open 2019; 9: e032550. [PubMed][CrossRef]

11. Colberg AB, Vatn D, Standal R et al. Hvordan kan strykprosenten ved eksamen stabiliseres? Tidsskr Nor Legeforen 2017; 137. doi: 10.4045/tidsskr.17.0025. [CrossRef]

12. Hift RJ. Should essays and other "open-ended"-type questions retain a place in written summative assessment in clinical medicine? BMC Med Educ 2014; 14: 249. [PubMed][CrossRef]

Publisert: 11. juni 2020. Tidsskr Nor Legeforen. DOI: 10.4045/tidsskr.20.0142
Mottatt 18.2.2020, godkjent 16.4.2020.