Overvejelser om prøver

Anne Lindebo Holm, læge og klinisk lektor, Enheden for Uddannelsesudvikling, Sundhedsvidenskabeligt Fakultet, Syddansk Universitet.

Anne Lindebo Holm, læge og klinisk lektor, Enheden for Uddannelsesudvikling, Sundhedsvidenskabeligt Fakultet, Syddansk Universitet, .

Anne Lindebo Holms primære arbejdsområder inden for uddannelse er prøver. Hun har blandt andet udformet en ny lægevidenskabelig kandidateksamen efter OSCE princippet (objective structured clinical examination). Hendes nuværende arbejdsområde er kvalitetssikring af prøver og indarbejdning af en samlet prøvestrategi på medicinstudiet. Når tiden tillader det, læser hun Master of Medical Education på University of Dundee.

Prøver og deres resultater er i stigende grad i søgelyset. Både som en del af uddannelsen for den studerende, men også som dokumentation for institutionens arbejde.

Heraf følger, at man må gøre sig nogle overvejelser over hvor sikker man kan være på at bedømmelsen både er pålidelig, og reelt fortæller os noget om det vi måler på.

Denne artikel vil gennemgå nogle tanker man som underviser eller studie bør gøre sig i den forbindelse. Der findes ingen perfekte løsninger eller prøveformer, men basale overvejelser kan medvirke til et øget udbytte af eksamen og det der ligner for såvel studerende som institution.

At afholde prøve eller eksamen er en del af de fleste studier og undervisningsforløb. Prøver på universitetsniveau er sjældent væsentlige emner i den offentlige debat. Det er forfatterens indtryk, at på mange studier lever måden der holdes prøver på, sit stille uforstyrrede liv. Prøverne har ikke i samme omfang som undervisningsmetoderne været fokuspunkt for ændringer.

Udgangspunktet for denne artikel er, at prøver er nødvendige, et middel til læring og en ofte uudnyttet ressource til kvalitetssikring. Vi bruger prøver til at sikre og dokumentere at den studerende har opnået de ønskede mål. Prøverne signalerer hvad uddannelsen værdisætter. For den studerende kan resultatet være bestemmende for hans eller hendes fremtidsmuligheder.

Med dette følger, at man som institution eller underviser må gøre sig nogle tanker om de prøver, man afholder.

Prøver er ressourcekrævende og store dele af en universitetsundervisers tid kan gå med at deltage i opgaveretning og eksamensafholdelse. På medicinstudiet på Syddansk Universitet har vi igennem et stykke tid arbejdet med at gøre dette tidsforbrug så udbytterigt som muligt.

For hvad ved vi om de prøver vi udsætter den studerende for? Ved vi om prøven tester de ønskede kompetencer? Hvor pålideligt er det? Hvor stor er den minimalt accepterbare usikkerhedsmargen af resultatet?

Det vi primært har arbejdet med, er at:
1. Sikre en sammenhæng imellem uddannelsens overordnede mål og de større prøver.
2. Optimere validiteten af de prøver, vi holder.
3. Gøre prøverne mere pålidelige.
4. Bruge afholdte prøver til kvalitetssikring.

De prøveformer vi har arbejdet mest med, er de meget strukturerede prøveformer, som f.eks. stationsprøver og kortsvarsopgaver (multiple choice og lign.), men planlægger at udvide til at omfatte alle de prøver der afholdes på studiet. Desuden har vi udfærdiget en overordnet prøvestrategi som på samme måde som en uddannelses- eller læringsstrategi er styrende for udviklingen.

En af forudsætningerne for at lave gode prøver er, at man har nogle klare mål for uddannelsen eller dens delelementer, som prøven kan laves ud fra. Uklare mål giver uklare prøver.

Vi har også erfaret at kvaliteten af prøven afhænger af om den enkelte underviser føler en forpligtigelse

over for de endelige mål. Vores vilkår er, at mange undervisere bidrager til den samme prøve, og en af de store opgaver er derfor at skabe en fælles forståelse for mål og midler.

I det følgende vil jeg forsøge at give en oversigt over hvilke tanker, jeg mener en underviser eller institution kan og bør gøre sig, når man planlægger en prøve.

Basale begreber

At rette eller udforme en prøve er tid der bruges på resultatanalyse efter et længere eksperiment (undervisning). Som med andre analyser er resultatet afhængigt af metoden og dennes anvendelse.

I prøvesammenhæng bruges begreberne validitet; at en prøve måler det den skal måle og reliabilitet; at prøven måler det ønskede på en pålidelig og reproducerbar måde.

Det skal understreges, at såvel validitet som reliabilitet ikke er en direkte funktion af hvilken prøveform der vælges, men derimod af hvordan den valgte prøveform udformes, administreres og bedømmes.

Validitet

Validitet kan deles op i flere underpunkter som er opskrevet i tabel 1.

Tabel 1

De 4 mest almindelige aspekter af en prøves validitet
Face validity
Content validity
Construct validity
Predictive validity

Der findes i litteraturen talrige varianter af denne liste, og der kan argumenteres for, at den skal være længere eller kortere – men jeg har udvalgt 4 aspekter af begrebet validitet, da jeg mener de dækker de vigtigste overvejelser om validitet.

Hvis man vil undersøge om ens prøve er valid, eller man vil designe en valid prøve, kan man starte med at stille sig selv følgende spørgsmål:

1. Hvordan ser prøven ud ved første øjekast. Hvilken effekt vil prøven have på de studerende? – Face validity
2. Dækker jeg indholdet i undervisningen? – Content validity
3. Er denne prøve konstrueret, så den giver mig et billede af om de studerende har opnået de ønskede kompetencer? Construct validity
4. Kan denne prøve give mig og den individuelle studerende information om indsatsområder i fremtiden? – Predictive validity

I eksempel 1 er beskrevet et tænkt undervisningsforløb. I det følgende vil eksemplet være udgangspunkt for de ovenstående spørgsmål.

Eksempel 1

Et kursus for 100 studenter på 10 uger omhandler teorien bag og anvendelsen af 4 forskellige metoder. Udgangskompetence for kurset er:

Den studerende skal efter endt kursus kunne udvælge, applicere og forsvare brugen af …*[metoderne]*… på … *[relevant materiale]*.

Undervisningen består af (pr. uge) 2 forelæsninger, 3 eksaminatorietimer med mulighed for diskussion og 3 øvelsestimer. Der er et foreslået læsemateriale på 200 sider og en øvelsesvejledning.

Den sidste dag i kurset er der en 3 timers skriftlig eksamen, der igennem mange år har bestået i at beskrive en af de 4 metoder i detaljer.

Face validity

Der kan argumenteres for at face validity ikke har noget med validitet at gøre. Det er prøvens ansigt udadtil, men giver ingen mening derudover. Grunden til at det er medtaget er at en dårlig face validity kan sende et signal til såvel studenter som fakultet, der kan være obstruerende for en optimal synergi mellem undervisning og prøve.

De fleste studenter ville acceptere prøven i eksemplet, fordi den har eksisteret i mange år. Der er sikkert cirkulerende standardbesvarelser, og underviserne er fortrolige med prøveformen og kan med stor sikkerhed sige – dette har I brug for til jeres prøve. Der er god enighed omkring niveauet, og det er ikke vanskeligt at rette opgaven.

Prøven kan have den effekt på undervisningen at de studerende vælger de (kostbare) øvelsestimer fra, fordi de i stedet bruger tiden på at læse teori. Til evalueringen vil flere studerende måske sige, at de var spild af tid. De studerende, der dumper, kan sige, at lige den metode der blev stillet spørgsmål i, var svær eller ikke deres stærke side.

Face validity er en upålidelig faktor, men er vigtigt for et kursus' forløb. Prøven i eksemplet har på papiret en lav face validity. Den ikke har forbindelse med udgangskompetencen og en stor del af undervisningen relaterer ikke til prøven. Da det er en etableret prøve, vil den imidlertid ikke give anledning til de store problemer i form af utilfredshed hos studenterne.

Som underviser skal man være opmærksom på, at nye prøver kan give usikkerhed og negative reaktioner hos de studerende, og det kan være hæmmende for læringen. En god information omkring prøvens udformning, intention, indhold og sværhedsgrad er derfor nødvendig.

En prøve med god face validity er relaterbar til kursets udgangskompetencer og virker relevant for de studerende. Hvis man vil undersøge face validity, skal

man spørge de studerende før og efter prøven er afholdt om deres opfattelse af prøven.

Content validity

Vores prøve i eksemplet tester kun en 1/4 del af det mulige vidensområde, og kun på et teoretisk deskriptivt niveau. Det svarer til situation 3 i figur 1. En stor del af undervisningen er øvelser, og omsætning af teori til praksis. Dette berøres ikke i prøven og giver derfor ikke oplysninger om den studerendes evne til at anvende sin viden.

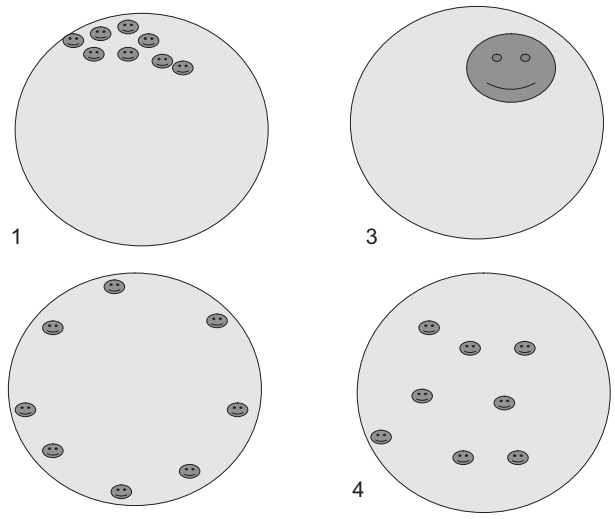
Det er umuligt at teste alt og en selektion er altid nødvendig – men det er ofte bedre at teste bredt, hvis man ønsker at teste viden. Som i situation 4, figur 1. Her er spørgsmålene eller stikprøverne fordelt bredt i emneområdet.

En skematisk metode at etablere content validitet på indholdssiden er blueprinting. Dvs. at man laver en arbejdstegning af sin prøve baseret på vægtningen af indholdet i undervisningen. Den studerende, der har koncentreret sig om det vigtigste, bør også være den, der klarer en prøve godt.

Blueprinting kan også anvendes til at få overblik over, hvilket kognitivt niveau der prøves på.

I det aktuelle eksempel koncentrerer man sig omkring den deskriptive viden, og har ikke højere kognitive kompetencer som analyse og udvælgelse med. Herved bliver en stor del af kurset ikke testet.

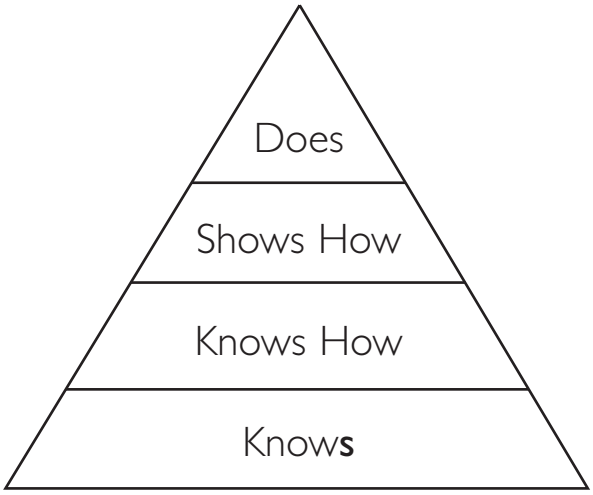
Der findes forskellige måder at vurdere sin prøves kognitive niveau på. Blooms taxonomi (1956) (se figur 1) opdeler viden i 6 stigende niveauer eller kompetencer: Deskriptiv viden, forståelse, applikation, analyse, syntese og evaluering. Til disse hører en lang liste af spørgsmål, man kan stille for at prøve disse niveauer.



Figur 1

Skematisk fremstilling af forskellige prøver. Den store cirkel repræsenterer et emneområde og de små »smileys«, spørgsmål i en prøve. Prøve 1 og 2 vil sjældent være at anbefale. Prøve 3 (en dybere test af et mindre område) kan bruges i kombination med andre. Prøve 4 vil ofte være at foretrække.

Alternativt og mere simpelt kan man anvende Millers trekant (1990) (se figur 2), som er mere handlingsorienteret med 4 kategorier, der bygger oven på hinanden; Ved,Ved hvordan,Viser hvordan, Gør. Hvis man bruger det i et prøveblueprint bliver det til:Viden, anvendelse af viden, kompetence og performance. Millers trekant kan især være anvendelig i handlingsorienterede fag, hvor en viden skal bruges i en konkret sammenhæng, som f.eks. inden for sundhedsvidenskab.



Figur 2

Millers trekant. (fra Miller GE, The assessment of clinical skills/competence/performance. Academic medicine 1990; 65(9);63-67.)

Et blueprint med udgangspunkt i eksempel 1's udgangskompetence kan se ud som i figur 3.

Indhold / Niveau/kompetence	Metode 1	Metode 2	Metode 3	Metode 4
Viden om metode				X
Udvælge metode				
Anvende metode				
Forsvare metode				

Figur 3: Skematisk prøve blueprint over kurset i eksempel 1 med den afholdte prøve som X.

Et dilemma, der kan opstå med content validity, er problemer med at teste viden på et tilstrækkelig højt kognitivt niveau, samtidig med at man dækker kursets vidensindhold på en rationel måde. Den optimale løsning findes ikke – men det kan være nødvendigt at kombinere flere prøveformer. F.eks. en praktisk anvendelsesorienteret og en teoretisk faktaorienteret prøve. I det konkrete tilfælde f.eks. en skriftlig kortsvarsprøve

med spørgsmål bredt i de 4 metoder, og en portefølje baseret på øvelserne gennem kurset med fokus på anvendelse.

To ting skal prøvestilleren holde sig for øje. Hvis en prøve (og undervisningen) ikke betoner det anvendelsesorienterede, kan den studerende senere få svært ved at anvende en faktaviden i relevante situationer. Hvis der er central viden, som man mener den studerende bør have, kan det være rationelt at teste denne viden i »ren form«, men det bør ikke dominere i en sådan grad at anvendelsesaspektet nedprioriteres af den studerende i læringen.

Hvis man vil undersøge sin prøves content validity, bør man holde sit blueprint op imod sit kursus og vurdere om fordelingen, arten og sværhedsgraden af opgaverne er i overensstemmelse med de ønskede udgangskompetencer.

Construct validity

Prøven i eksemplet tester formentlig det den beder de studerende om: At beskrive.

Er prøven konstrueret så den tester de ønskede udgangskompetencer? – Det er mindre sandsynligt.

Deskriptiv viden har en korrelation med anvendt viden, men ofte ikke så stor som antaget. F.eks. er det flere gange vist inden for faget medicin, at f.eks. deskriptive essay-opgaver korrelerer dårligt med hvordan den studerende kan anvende sin viden i praksis, men at multiple choice-opgaver med fokus på løsning af kliniske scenarier har en god korrelation. Dette skyldes efter min overbevisning ikke prøveformen som sådan, men at den studerende tvinges til at foretage et valg, og ikke blot give en forklaring eller beskrivelse.

Jo mere kompleks anvendelsen af viden er, jo vanskeligere er det at forudsige om den student, der kan beskrive en metode i detaljer, også er den, der er bedst til at udvælge og anvende metoden.

Hvis man vil validere sin prøve mhp. construct validity, dvs. om prøven der er designet, tester de underliggende strukturer, kan man bruge de samme metoder som når man validerer et spørgeskema. F.eks. udtalelser om relevans af opgaven fra et panel af eksperter inden for fagområdet, afprøvning på relaterede grupper af studenter eller sammenligning med andre prøver der tester det samme.

Predictive validitet

Kan den afholdte prøve i eksemplet give oplysninger om, hvordan den enkelte studerende vil klare sig fremover – f.eks. i et senere kursus, i et erhverv eller som forsker? Forudsætninger er selvfølgelig at kurset og prøven har en relevans. Mener aftager, at det der fokuseres på, er relevant?

Prædiktiv validitet kan være svært at vurdere, men kan være relevant at få belyst – specielt i forbindelse med større projekter (embedsprøver og specialer). Hvis man vil undersøge prædiktiv validitet, skal man sammenholde resultatet med prøver eller udtalelser længere fremme i studieforløbet eller på aftagerside.

Reliabilitet

Reliabilitet (pålidelighed) bør altid overvejes, når man vælger sin prøveform og hvordan man vil administrere den. Jo mere, der står på spil for den studerende til prøven, jo mere pålidelig må prøven være. En rimelig fejlmargen kan accepteres ved en primært uddannende (formativ) prøve, men den bør være beskeden ved en afsluttende prøve.

Teknisk set er pålidelighed afhængig af 3 ting. Hvor meget der prøves i, hvor mange bedømmere der er, og hvor lang tid, der bruges på det.

Helt basalt kan man sige, at jo flere observationer man har – jo mere pålidelig bliver prøven. En prøve med 20 spørgsmål vil være mere pålidelig end en med 10 af samme slags. En traditionel mundtlig eksamen skal f.eks. anslagsvis bruge 4 timer på at give samme pålidelighed, som en 2 timers kortsvars skriftlig opgave.

Deraf følger at prøveformer, hvor man kan svare på mange spørgsmål pr. tidsenhed er meget pålidelige i forhold til den tid der bruges på det (en af grundene til at multiple choice er så anvendeligt til »populationstestning«). Det er dog vigtigt at huske, at pålidelighed ikke er det samme som validitet. 50 bedømmelser af prøven i eksemplet gør ikke prøven pålidelig til at bedømme andet end at studenten kan beskrive en laboratorieøvelse.

Optimalt bedømmes alle spørgsmål ens for alle. Hvis man f.eks. har en prøve med 4 elementer, giver det størst pålidelighed, hvis hvert element bliver bedømt af den samme hos alle studenter, og der i stedet er 4 bedømmere, og deres udsagn kombineres til sidst. Ofte er det modsatte tilfældet: At en students hele præsentation bliver bedømt af en bedømmer,og medstudentens præsentation af en helt anden bedømmer, der måske har andre standarder eller en anden interaktion med studentens arbejde.

Hvis man har spørgsmålsspecifikke bedømmere, mindskes »halo-effekten«, dvs. man undgår at et godt eller dårligt besvaret spørgsmål 1 påvirker bedømmelsen af spørgsmål 2.

Reliabilitet er ofte en handel man slår af med ressourcer. Jo flere studenter, jo sværere kan det være at give de studerende en pålidelig bedømmelse af f.eks. 20 siders essayopgaver ud fra et spørgsmål. Det ligger i sagens natur, at den samme bedømmer ikke kan bedømme dem alle.

Metoder, der kan anvendes her, er dels gennem-

diskuterede rettenøgler hvor bedømmerne sammen udformer et sæt »regler« eller en rettenøgle, dels at bruge flere bedømmere, jo tættere en student er på »dumpegrænsen«. På lægeuddannelsen ved Syddansk Universitet har vi arbejdet med det første koncept, og vores erfaring er, at det kræver et godt samarbejde mellem bedømmerne før opgaven stilles, men hvis det er til stede, opnås en god interbedømmeroverensstemmelse, og det effektiviserer retningen. Det andet koncept bruger vi også, men mere uformelt – men med indførelsen af komplekse prøveformer som f.eks. porteføljeeksamen, vil der være behov for en mere formaliseret tilgang

Jo mere kompleks en sag er, jo flere bedømmere kræver det. Et multiple choice spørgsmål kræver kun én bedømmer – et speciale måske op til fem, hvis der skal opnås en god pålidelighed. Hvor mange bedømmere der skal til, afhænger dels af deres indbyrdes forskellighed (hawks and doves), og hvordan de interagerer med hinanden.

Reliabilitet måles oftest som en Cronbachs alpha eller ANOVA. Dvs. et tal mellem 0 og 1. For en prøve der har stor betydning for den studerende, skal reliabilitetskoefficienten være over 0,8-0,9. Det kan groft oversættes med, at man har en fejlmargen på 10-20 %.

Hvis man ønsker at beregne reliabilitet på sin prøve, anbefales det at kontakte en statistiker.

Konklusion

Med overgangen til studier og kurser defineret ved udgangskompetencer; »outcome based education« er focus på »outcomebased assessment« forestående. Heri ligger at vi skal holde vores prøver op imod vores udgangskompetencer. Hvis vi skal have prøverne til at arbejde for læringen og bruge dem som mål for vores succes internt og ekstern, er det nødvendigt at indkorporere overvejelser om validitet og reliabilitet i planlægningen, og ikke mindst evalueringen af de prøver, vi afholder.

At opnå valide og pålidelige prøver er ikke nødvendigvis afhængig af at opfinde nye prøveformer, men snarere at anvende og kombinere prøveformer til at opnå et så nuanceret billede som muligt. Prøveformen i sig selv er ikke garant for hverken validitet eller reliabilitet.

Som tiden udvikler sig tegner der sig et billede af, at betydningen af prøveresultater kan få en (endnu større) betydning for den studerendes fremtidsmuligheder. Formentlig vil vi i fremtiden møde større krav fra såvel studerende som aftagere til de prøver, vi stiller som dokumentation og dom. Kan vi da forsvare at afholde en prøve, som f.eks. kun med 50 % sandsynlighed præcist kan belyse, om den studerende mestrer det testede (dette resultat vil være standard ved 1 times mundtlig eksamen)? Kan vi argumentere for at den prøve, vi holder, virkelig tester den kompetence, vi lover studenter og samfundet at de har, når de forlader vores institution?

Disse spørgsmål har selvfølgelig ikke noget enkelt svar. Prøver er tradition, og når de tages op til diskussion, afføder det mange nye spørgsmål om, hvad det at være universitetsstuderende er, og hvad målet med et studie bør være. Denne diskussion er ikke mindre interessant, og tæt forbundet med den måde vi vælger at prøve vores studerende på.

Litteratur

Bloom, B.S. (Ed.) (1956) *Taxonomy of educational objectives: The classification of educational goals: Handbook I, cognitive domain*. New York; Toronto: Longmans, Green.

Miller GE, The assessment of clinical skills/competence/performance. *Academic medicine* (1990); *65(9)*; 63-67

Schuwirth,L.W. & van der Vleuten,C.P. (2004a). Changing education, changing assessment, changing research? *Med Educ* 38, 805-812.

Schuwirth,L.W. & van der Vleuten,C.P. (2004c). Different written assessment methods: what can be said about their strengths and weaknesses? *Med Educ* 38, 974-979.

Noter

De to ovenstående artikler af Schuwirth kan anbefales til et overblik over området om reliabilitet og validitet set i relation til forskellige skriftlige prøveformer.

Ønskes der en fuldstændig litteraturliste, kan dette fås ved henvendelse til forfatteren.