Multiple choice-spørgsmål i undervisningen

Lotte Dyhrberg O'Neill[1], SDU-Universitetspædagogik

Abstract

Multiple choice (MC) prøver som eksamensformat har eksisteret længe, og deres anvendelse er udbredt internationalt. Som prøveformat komplementerer MC-prøver de øvrige skriftlige prøve-formater, og dets muligheder er værd at kende til for de fleste undervisere. Med udviklingen af moderne undervisningsteknologi spiller MC-spørgsmål sandsynligvis en stigende rolle på de danske universiteter, ikke mindst i forhold til at understøtte læring og eksamination på meget store hold. I denne guide gennemgås otte grundlæggende praksisser for undervisere, som mangler formel viden om dette eksamensformat, og som ønsker at komme godt i gang med at anvende formatet på en pædagogisk forsvarlig måde.

Baggrund

MC-spørgsmål blev udviklet for omtrent hundrede år siden som konsekvens af et behov for et testformat, som var retfærdigt, kunne bruges til relativ bedømmelse, og som var effektivt i forhold til massetestning af et øget optag af studerende og af soldater til første verdenskrig. Formatet blev hurtigt populært i forbindelse med intelligenstestning i 1920'erne, og sidenhen blev det et meget dominerede eksamensformat i Amerikanske uddannelseskontekster (Butler, 2018). Populariteten skyldes:

- at MC-prøver er lette, hurtige og billige at bedømme
- at bedømmelsesprocessen er fri for bedømmerbias
- at formatet tillader en større bredde af indholdsemner i prøven, idet svartiden per spørgsmål er meget kort
- at viden kan testes uafhængigt af skrivefærdigheder (Butler, 2018; Fellenz, 2004).

Disse fordele ved MC-formatet repræsenterer samtidigt væsentlige ulemper ved de fleste andre skriftlige eksamensformater.

Til gengæld er det generelt udfordrende at producere gode MC-spørgsmål uden tekniske fejl (O'Neill, Mortensen, Nørgaard, Øvrehus & Friis, 2019), og det er meget ressourcekrævende at skrive nok spørgsmål til at sikre reproducerbare resultater (Fellenz, 2004). Der skal typisk op mod 100 spørgsmål til for at opnå høje grader af pålidelighed i en eksamenssituation (Paniagua et al., 2016, s. 4). Eksamener med for få spørgsmål har helt generelt en tendens til at resultere i upålidelige scorer, som afspejles i lave reliabilitetskoefficienter. Lav reliabilitet er en trussel for en meningsfyldt fortolkning af eksamensscorerne, og derfor er det også en trussel for eksamensvaliditeten (American Educational Research Association, 2014; Downing, 2002b; Kane, 2006).

[1] Kontakt: ldo@sdu.dk

Dette princip gælder for øvrigt for alle eksamensformater og på tværs af fagområder. MC-formatet har været kritiseret for at medføre testning af lavere ordens tænkefærdigheder (faktuel viden) i for stort omfang (Fellenz, 2004), men den problematik handler mere om forståelsen og anvendelsen af læringsmål end om MC-formatets medfødte egenskaber og muligheder: MC-formatet kan anvendes til at udprøve både faktuel og anvendt viden med (Clay & Root, 2001; Haladyna, 2004; Palmer & Devitt, 2007; Paniagua et al., 2016). MC-prøver er et skriftligt testformat, hvor testtageren vælger et eller flere svar på en liste af svarmuligheder (*selected response format*), som er definerede og formulerede af spørgsmålsstilleren (Downing, 2002a). Det indebærer, at MC-prøver fx ikke er velegnede til at teste læringsmål, hvor studerendes kreativitet, skrivefærdigheder eller mundtlige formidlingsevner er det centrale (Fellenz, 2004). MC-formatet giver ikke testtageren mulighed for at præsentere tankerækken og de argumenter, som ligger bag et afgivet svar. MC-formatet er naturligvis heller ikke velegnet, hvis læringsmålene beskriver motoriske færdigheder, eller hvis de er i kategorien kompetencer, som defineret i den Danske Kvalifikationsramme for de videregående uddannelser (UFM).

Multiple choice-formatet kan anvendes i eksamenssituationer som absolut bedømmelse, men også i undervisningssituationer med det primære formål at støtte op om læringen. Forskning har vist, at der findes en generel læringseffekt af testning, som også gælder for MC-spørgsmål og -prøver (Butler, 2018; Butler & Roediger, 2008), men også, at det samtidigt er vigtigt at koble testningen med feedback umiddelbart efter for maksimal læringseffekt (Butler & Roediger, 2008). MC-spørgsmål er også ofte et væsentligt element i et blended eller online-undervisningsdesign, fx som en formativ aktivitet, og mange moderne learning management-systemer giver de studerende mulighed for at teste egen forståelse af tekster og videoforelæsninger som en del af den læring, der foregår i forberedelsestiden. Eftersom både gennemførelsen og bedømmelsen af MC-spørgsmål i dag kan foregå elektronisk via mobiltelefoner og relevant software, som fx kan integreres med underviserens slides, giver MC-spørgsmål også undervisere lynhurtig og effektiv feedback på de studerendes forståelse i storrumssituationer, hvilket kan understøtte muligheden for at tilpasse undervisningen, mens den pågår.

I det følgende gennemgås otte grundlæggende råd til undervisere, som vil anvende multiple choice-spørgsmål i undervisnings- og/eller eksamenssammenhæng på en professionel og pædagogisk forsvarlig måde.

Praksispunkter

1. Afgør formålet med testen
2. Vælg det rigtige format
3. Udvælg indhold med omhu
4. Inkluder stimulusmateriale
5. Stil fokuserede og klare spørgsmål
6. Undgå konstruktionsfejl
7. Undersøg, hvordan dine spørgsmål fungerer
8. Brug resultaterne til kvalitetssikring

Tip 1: Afgør formålet med testen

Det er generelt en god ide at starte med at identificere, hvad hovedformålet med testen er, fordi det kan påvirke den måde, testspørgsmålene skal udvikles og administreres på, og den måde testens resultater bruges på. Uanset hvad testformålet end måtte være, bør man dog altid stille fokuserede og klare MC-spørgsmål (tip 5), ligesom man altid bør forsøge at undgå tekniske fejl (tip 6). Hvis hovedformålet med testspørgsmålene primært er at støtte op om de studerendes læring i undervisningen, er det særligt vigtigt, at de studerende får umiddelbar adgang til formativ feedback på både korrekte og forkerte svarmuligheder. Moderne teknologi gør det heldigvis nemt at automatisere og individualisere den form for feedback. Hvis formålet med MC-spørgsmålene er at støtte op om de studerendes motivation for faget i løbet af semestret, kan det være særligt vigtigt at have ekstra meget fokus på at vælge testindhold med autentisk stimulusmateriale (tip 4). Hvis formålet med testen er at kommunikere om eller træne til eksamen, skal indholdet i testspørgsmålene selvfølgelig udvælges med særlig omhu, så de giver et retvisende billede af, hvad man kan forvente af eksamen som studerende (tip 3). Hvis det primære formål med at anvende MC-spørgsmål er at evaluere undervisningen, er det vigtigt at evaluere testresultaterne og item-statistikken (tip 7 og 8). Hvis formålet med testen derimod er at bedømme præstationer i testsituationer, hvor der er meget på spil for de studerende (fx optagelsesprøver, eksamenssituationer o.l.), er alle 8 tips i denne guide særligt vigtige (Paniagua et al., 2016).

Tip 2: Vælg det rigtige format

Hvis det i testkonteksten er udpræget vanskeligt at konstruere plausible distraktorer (forkerte svarmuligheder, som er tilstrækkeligt udfordrende for testtagerne), kan det meget vel være en bedre løsning at vælge et skriftligt testformat, hvor testtagerne selv formulerer svaret som fx 'fill-in-the-blanc'-formater, 'very short answer'- eller 'short answer'-spørgsmålsformater (Clay & Root, 2001; Downing, 2002a; Sam et al., 2018).

Hvis MC-spørgsmål er fundet velegnede i konteksten, skal der også foretages et valg om den type af MC-spørgsmål, som med fordel kan anvendes. Overordnet kan MC-typerne inddeles i 2 familier: 'True-False'-familien (Figur 1) eller 'One-Best-Answer'-familien (Figur 2). Det bliver alt for vidtgående at beskrive samtlige undertyper i denne guide, så spørgsmålsstillere opfordres til at orientere sig nærmere om de mange afarter i hver af de to familier, deres respektive fordele og ulemper og den tilhørende evidens i relevante lærebøger og manualer (Haladyna, 2004; Paniagua et al., 2016). Dog er det værd at nævne her, at den store nationale testinstitution National Board of Medical Examinars i USA forsøger at undgå at anvende 'True-False'-formatet, fordi de har fundet, at det i praksis er vanskeligere at konstruere 'True-False'-spørgsmål af høj kvalitet, men også fordi formatet synes at have en tendens til at inducere testning af faktuel viden i overdrevent omfang (Haladyna, 2004; Paniagua et al., 2016).

Figur 1: Et 'True-False'-multiple choice-eksempel

Hvad adskiller havmus fra resten af bruskfiskene?

☒ A. Gællelåg
☒ B. Manglende skæl
☐ C. Dorsalt nerverør
☐ D. Gælletarm
☒ E. Overkæben er sammenvokset med neurokraniet

Note. Dette 'True-False'-spørgsmål består af et 'lead-in' (selve det spørgsmål, der skal besvares) og svarmulighederne. Testtageren skal angive alle korrekte svarmuligheder på listen. 'True-false'-formatet fungerer kun, hvis alle svarmuligheder på listen er enten 100 % rigtige eller 100 % forkerte. I nogle kontekster kan det være vanskeligt at konstruere 'distraktorer' (forkerte svarmuligheder), som er tilstrækkeligt plausible til at være en passende udfordring for testtagerne. Det kan påvirke testens muligheder for at skelne tilstrækkeligt mellem eksaminanders præstationer (reliabiliteten) negativt.

Figur 2: Et 'One-Best-Answer'-multiple choice-eksempel

Tove (28 år) klager over smerter i lænden med udstråling til bagsiden af højre lår. Smerterne opstod, da hun skulle løfte sin 1-årige datter ud af bilen. At bøje sig bagover forværrer smerterne, mens det lindrer at ligge på ryggen med benene trukket op mod hagen. Strakt benløft-test på højre side giver forværring af smerterne i lænden ved ca. 70 grader. Neurologisk undersøgelse: Intet abnormt.

Hvilken diagnose er mest sandsynlig?

☒ A. Facetleds syndrom
☐ B. SI-leds syndrom
☐ C. Triggerpunkt i M. Piriformis
☐ D. Prolaps med rodaffeksion

Note. Dette 'One-Best-Answer'-spørgsmål består af en vignet (en skriftlig case), et 'lead-in' (spørgsmål, som skal besvares) og svarmulighederne. Testtageren skal angive det ene svar på listen, som vurderes at være bedst på baggrund af konstellationen af oplysninger i situationen, som er beskrevet i vignetten. 'One-Best-Answer'-formatet er bedst i faglige kontekster, hvor svarmulighederne ligger på et kontinuum af rigtighed eller sandsynlighed. Der skal være *et* af svarene, som reelt er bedre, mere rigtigt eller mere sandsynligt end de øvrige, og distraktorerne skal samtidigt være tilstrækkeligt plausible til at være en passende udfordring for målgruppen af testtagere, før at formatet fungerer optimalt.

Tip 3: Udvælg spørgsmålenes indhold med omhu

Uanset om der produceres MC-spørgsmål til læringsformål eller til eksamensformål, er det nødvendigt at overveje ikke blot hvilket indholdsemne, men også hvilken type af viden, som skal undersøges. Det taksonomiske niveau af den viden, som skal testes (Figur 3), vil både påvirke formuleringen af 'lead-in'-delen af MC-spørgsmålet og have betydning for valget (eller fravalget) af stimulusmateriale.

Figur 3: Blooms reviderede kognitive taksonomi

Note. Taksonomien er beskrevet i flere detaljer af Krathwohl (2002). MC-spørgsmål er ikke velegnede til at teste læringsmål, hvor studerendes kreativitet, skrivefærdigheder eller mundtlige formidlingsevner er det centrale. MC-formatet giver heller ikke testtageren mulighed for at præsentere tankerækken og de argumenter, som ligger bag et afgivet svar. Men et velkonstrueret MC-spørgsmål kan godt teste studerendes evner til at evaluere (bedømme, vægte information, vælge en løsning).

Det anbefales generelt, at der er en høj grad af kongruens (Constructive Alignment) mellem læringsmålene, læringsaktiviteterne og eksamen på videregående uddannelser (Biggs & Tang, 2007). Bruges MC-spørgsmål derfor til at understøtte læringen i undervisningen, er det vigtigt, at de repræsenterer centrale læringsmål og læringsaktiviteter, som fylder i forberedelsen og undervisningen. Bruges MC-spørgsmål i eksamenssammenhæng, er det særligt vigtigt for validiteten af eksamen, at indholdet i testen er tilstrækkeligt dækkende i bredden, dvs. tilstrækkeligt dækkende i forholdet til alle fagets emner (American Educational Research Association, 2014; Downing, 2002b, 2003; Downing & Haladyna, 2004; Kane, 2006). Det kan derfor være en god ide at skabe sig et overblik over forholdet mellem læringsmål, indholdsemner og læringsaktiviteter i konteksten, før man bestemmer indholdsemne og taksonomisk niveau for de spørgsmål, som skal produceres. Hvis prøven fx har 80 eller 100 spørgsmål, som måske skal produceres af flere spørgsmålsstillere, kræver det en vis styring at sikre tilstrækkelig kongruent indholdsdækning, som er helt centralt for prøvevaliditeten (Downing, 2002b).

En måde at skabe sig det nødvendige overblik på kan være at lave et såkaldt *blueprint* for det fag, som testen skal dække (Tabel 1).

Tabel 1: Et eksempel på et blueprint for en test med 20 spørgsmål

Indholdsemne	**Kontakttimer**	**Vægtning**	**Antal spørgsmål**	**Taksonomisk niveau**		
				Huske/ forstå	Anvende/ analysere	Evaluere
Arbejderbevægelsen	2	0,20	4	1	2	1
Grundloven	1	0,10	2	0	1	1
Englandskrigene	4	0,40	8	2	5	1
Imperialismen	3	0,30	6	1	2	3
Total	10	1	20	4	10	6

Note. Tidsforbruget og de gennemførte undervisningsaktiviteters taksonomiske niveau danner grundlag for sammensætningen af prøvespørgsmålene. Eksemplet her er inspireret af Eweda et al. (2020), og antallet af spørgsmål er lavt for at lette formidlingen.

Blueprintet bruges derefter til at forudbestemme indhold og taksonomisk niveau for hvert spørgsmål, som skal indgå i testen (Ahmad & Hamed, 2014; Eweda, Bukhary & Hamed, 2020; Roberts, Newble, Jolly, Reed & Hampton, 2006). På den måde kan man understøtte og evt. dokumentere om nødvendigt, at kompositionen af spørgsmål og testformater i testen afspejler eller dækker faget på en fornuftig måde (Dent, Harden & Hunt, 2017; Downing & Yudkowsky, 2009; NBME, 2019).

Ofte tillader ressourcer og rammer ikke, at der udprøves i samtlige mulige indholdsemner i en eksamen – specielt ikke i større og meget indholdstunge fag. De humane ressourcer, som allokeres til udvikling af testindhold, er typisk særligt begrænsende for graden af indholdsdækning, som er mulig i MC-prøver. Når der derfor skal prioriteres i testindholdet, anbefales det, at man bruger den tilgængelige testtid klogt. Det indebærer, at man som testudvikler prioriterer emner og indhold, som udgør vigtige begreber eller koncepter ('kernepensum'), almindeligt forekommende problemstillinger, katastrofale eller fatale problemstillinger etc. *før* andet mere esoterisk eller perifert indhold, og at man undgår at bruge den dyrebare testtid på at stille trickagtige spørgsmål (Haladyna, 2018; Paniagua et al., 2016). Valg af MC-spørgsmålenes indhold kan betragtes som en *optimerings-opgave*, hvor opgaven består i at dække flest mulige vigtige emner og læringsmål i den tilgængelige testtid eller med de tilgængelige menneskelige ressourcer (Downing, 2002b, 2003; Downing & Haladyna, 2004).

Tip 4: Inkluder stimulusmateriale

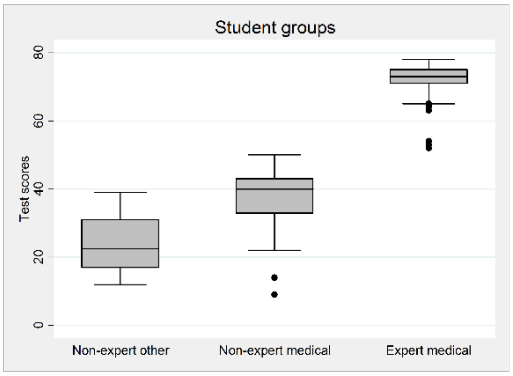
Hvis man påtænker at eksaminere eller teste sine studerendes tænkefærdigheder på *højere taksonomiske niveauer* (Krathwohl, 2002), dvs. om de fx kan anvende viden, analysere eller evaluere information i en given situation på baggrund af lærte principper, begreber eller teorier (jf. Figur 3), så skal MC-spørgsmålet indeholde en eller anden form for stimulusmateriale, som muliggør og afkræver den ønskede form for tankevirksomhed. Der skal være noget (en situation, et problem, et artefakt etc.) man kan anvende sin faktuelle eller teoretiske viden på (Clay & Root, 2001; Paniagua et al., 2016; Scully, 2017).

Brug af stimulusmateriale kan ikke blot hjælpe med at understøtte, at højere taksonomiske niveauer testes. Det kan også tydeliggøre og eksemplificere *relevans* og *anvendelighed* af læringsmål, indhold og læringsaktiviteter i kurset, som igen kan understøtte de studerendes *motivation* (Harden & Laidlaw, 2013). Hvis stimulusmateriale og tilhørende spørgsmål har oprindelse i reelle, aktuelle problemstillinger fra professionel praksis, kan det tilføre en højere grad af *autenticitet* til lærings- eller eksamenssituationer.

Stimulusmateriale i et MC-item kan bestå af fx: en kort tekst (en kasuistik, en artikel, et digt etc.), et diagram, en tabel, en graf, en illustration, kunst, noder, fotos eller andet visuelt materiale eller et fysisk artefakt (Figur 4). Hvis testen skal foregå elektronisk, kan stimulusmaterialet også være en videofil eller en lydfil, hvor det er mere relevant.

Figur 4: Eksempler på stimulusmateriale

A 

B 

C 

Note. Fotomateriale, figurer eller artefakter kan danne baggrund for fortolkninger (A), evalueringer (B) og valg/beslutninger, som skal træffes (C), og dermed understøtte MC-spørgsmål, som søger at udprøve højere taksonomiske niveauer (jf. Figur 1).

Tip 5: Stil fokuserede og klare spørgsmål

Kvaliteten af de studerendes besvarelser i skriftlige eksamener afhænger ikke kun af de studerendes viden, men er også påvirket af *spørgsmålenes kvalitet*. Dette gælder naturligvis også for multiple choice-formatet (Clay & Root, 2001; Haladyna & Rodriguez, 2013). Det er en tommelfingerregel, at selve det spørgsmål som skal besvares ('lead-in') udgør en kort, klar, hel og fokuseret sætning, som slutter med et spørgsmålstegn. 'Lead-in' i figur 1 ('Hvad adskiller havmus fra resten af bruskfiskene?') og i figur 2 ('Hvilken diagnose er mest sandsynlig?') er eksempler, som overholder denne regel. 'Lead-in' skal være så fokuseret eller specifikt, at man som testtager i princippet skal kunne afgive det korrekte svar, selv hvis man ikke kunne se svarmulighederne ('Cover-the-

options'-reglen). Samtidigt bør alle svarmulighederne tilhøre det samme domæne, som 'lead-in' fokuseres på. Dvs. hvis 'lead in' fokuserer på fx diagnoser (som i figur 2), så bør alle svarmuligheder være diagnoser. Eksempler på et 'lead-in' som *ikke* opfylder dette krav er fx 'Hvilket af følgende udsagn er mest korrekt?' eller 'International klassifikation af funktionsevne er'. Hvis man alligevel vælger at bryde med princippet om at fokusere 'lead-in', så skal samtlige svarmuligheder på listen kunne kategoriseres som værende enten helt rigtige eller helt forkerte (Case & Swanson, 2002; Downing & Yudkowsky, 2009; Paniagua et al., 2016).

Hvis det er svært at komme i gang med at skrive et MC-spørgsmål, kan det anbefales at man lader sig inspirere af de talrige generiske 'lead-ins' ('shells'), som findes beskrevet i MC-litteraturen (Haladyna, 2018; Paniagua et al., 2016).

Tip 6: Undgå konstruktionsfejl

Der findes en perlerække af konstruktionsfejl, som man som spørgsmålsstiller bør kende til og forsøge at undgå, fordi de har vist sig at 'forurene' det begreb, man påtænker at teste (den faglige viden). Konstruktionsfejlene resulterer typisk i, at man i et eller andet omfang kommer til at teste *faglig irrelevant sværhed* eller testtagerens *testsnuhed* frem for ren faglig viden. Tabel 2 viser eksempler på konstruktionsfejl, som er identificeret i items hvor stimulusmaterialet udgøres af en skriftlig vignet.

Tabel 2: Eksempler på almindelige konstruktionsfejl i multiple choice-spørgsmål

Irrelevant sværhed	**Testsnuhed**
Svarmulighederne er lange, komplicerede eller dobbelte	Grammatiske hints – en/flere distraktorer passer ikke grammatisk til 'lead-in' og kan derfor hurtigt udelukkes.
Numeriske svarmuligheder angives forvirrende (fx i ikke-numerisk rækkefølge eller i et uensartet format).	Logiske hints – en undergruppe af svarmulighederne dækker tilsammen hele universet af mulige svar, sådan at de øvrige distraktorer automatisk kan udelukkes som reelle svarmuligheder
Vage ord (sjældent, vanligvis etc.) i svarmulighederne	Absolutte ord (altid, aldrig etc.) i svarmulighederne
Ikke-parallelt sprog i svarmulighederne	Det korrekte svar er længere, mere specifikt eller mere komplet end distraktorerne
Svarmuligheder i ulogisk orden	Ordgentagelser – et ord eller en vending i stammen kan genfindes i det korrekte svar
'Ingen af ovenstående' eller 'alle ovenstående' som svarmulighed	Konvergensstrategien kan anvendes – det korrekte svar er det, som har flest elementer til fælles med distraktorerne

Stammen[a] er unødigt kompliceret	
Svaret på et item afhænger af svaret på et af de andre items	
Negeringer i 'lead-in' (fx 'ikke' eller 'undtagen')	
Overlappende svarmuligheder	

Note. Tabellen er inspireret af Paniagua et al. (2016).
[a]Stammen udgøres af stimulusmaterialet (fx en skriftlig vignet) og selve det spørgsmål ('lead-in'), som skal besvares.

Der er tidligere publiceret mange fine eksempler på de tekniske fejl beskrevet i tabel 1, som jeg kan opfordre interesserede læsere til at studere nærmere (Haladyna, 2004, 2018; Paniagua et al., 2016).

Tip 7: Undersøg, hvordan dine spørgsmål fungerer

Det er en god ide at forsøge at gøre lidt ekstra ud af at kvalitetssikre spørgsmålene, hvis de skal anvendes i en eksamenssituation, hvor der er noget på spil for de studerende. *Før eksamen* kan det anbefales, at man lader sine kolleger, fagfæller eller evt. censor gennemføre testen uden at kende til den påtænkte svarnøgle, sådan at man kan identificere fejl eller uenigheder, før de studerende udsættes for sættet. Det er let at komme til at lave fejl i en svarnøgle i et større eksamenssæt, og der kan også være faglige uenigheder om det korrekte svar inden for et fagfelt. *Efter eksamen* (og før karaktergivning) kan man anvende statistik baseret på eksamens-besvarelserne til at identificere spørgsmål, som ikke fungerede som ønsket, og som eventuelt kunne være fejlbehæftede.

Tabel 3 illustrerer eksempler på statistik, som kan hjælpe den eksamensansvarlige med at identificere dårligt fungerende spørgsmål. Hvis man som testansvarlig ikke automatisk har adgang til lignende statistik, findes der lettilgængelige guides til, hvordan den udregnes manuelt (Varma, 2006). I eksemplet i tabel 3 svarede kun 30 % af eksaminanderne spørgsmål 28 korrekt, hvilket indikerer, at spørgsmålet var svært. Diskriminationsindekset (DI) var tæt på 0, hvilket indikerede, at eksaminandernes besvarelse af spørgsmål 28 harmonerede dårligt med, hvordan de klarede sig i de øvrige spørgsmål i eksamenssættet. Svarmønstret indikerede, at kun 25 % af de bedste eksaminander (Top) svarede korrekt, mens 41 % af de ellers dårligst præsterende eksaminander besvarede spørgsmålet korrekt. Svarene synes at være meget jævnt fordelt over de tre svarmuligheder for alle eksaminander uanset dygtighed. Statistikken for dette item fik de testansvarlige til at tjekke for konstruktionsfejl, og der blev fundet adskillige fejl, bl.a. et 'lead-in', som ikke var fokuseret, og tilhørende svarmuligheder i forskellige domæner. Statistikken illustreret i tabel 3 kan også hjælpe med at identificere distraktorer, som ikke var tilstrækkeligt plausible (dvs. blev valgt af <5 % af respondenterne), hvilket påvirker spørgsmålets diskriminationsevne og reliabilitet negativt. Det er ofte en relativt stor udfordring at skrive plausible distraktorer for spørgsmålsstillere. Der synes heldigvis at være relativ bred og forskningsbaseret konsensus om, at det for det meste er tilstrækkeligt at have tre svarmuligheder per spørgsmål, hvis ellers sættet indeholder et tilstrækkeligt stort antal spørgsmål (Haladyna & Downing, 1993; Kilgour & Tayyaba, 2016; Rodriguez, 2005; Royal & Stockdale, 2017). Haladyna (2004, s. 113) forsvarede dette råd med ordene:

'One Criticism of using fewer instead of more options for an item is that guessing plays a greater role in determining a student's score. The use of fewer distractors will increase the chances of a student guessing the right answer. However, the probability that a test taker will increase his or her score significantly over a 20, 50 or 100 item test by pure guessing is infinitesimal.' Spørgsmål med DI på over 0,25 og sværhedsgrader på 45-75 % kategoriseres typisk som gode i almindelige eksamenskontekster (Downing & Yudkowski, 2009). Generelt bør der altså i et godt eksamenssæt være mange items af middel sværhed og ganske få, som er hhv. meget lette og meget svære. Disse kvalitetskrav kan være ganske vanskelige at leve op til i en dansk kontekst pga. gældende lovgivning om studerendes ret til at klage over eksamen, som har medført en praksis baseret på juridiske fortolkninger, hvor eksamenssæt skal udleveres til studerendes granskning på forlangende. Denne praksis underminerer opbygningen af sikre MC-banker, hvor afprøvede spørgsmål af høj kvalitet og kendt sværhedsgrad kan akkumuleres og genbruges. Denne situation udfordrer ikke blot muligheden for at optimere prøvevaliditeten, men også bæredygtigheden og acceptabiliteten af store MC-eksamener i danske universitetskontekster, fordi det i længden er meget ressourcetungt at skulle konstruere et helt nye eksamenssæt til hver eksamen i et forsøg på at opretholde testsikkerheden og sikre validiteten.

Tabel 3: Et eksempel på item-statistik, som kan bruges til kvalitetssikring efter eksamen

Item nr.	Sværhedsgrad (%)	DI	Svar	Alle (%)	Top (%)	Bund (%)
28	30	0,00	A*	30	25	41
			B	41	41	38
			C	28	34	22

Note. DI=diskriminationsindekset angiver korrelationen mellem besvarelsen af item 28 og så sumscoren på de øvrige items i testen på tværs af eksaminanderne, Alle=alle eksaminander, Top=den fjerdedel af eksaminanderne, som klarede sig bedst til prøven, Bund=den fjerdedel af eksaminanderne, som klarede sig dårligst til prøven.

* angiver det korrekte svar

Tip 8: Brug resultaterne til kvalitetssikring

Et spørgsmål som item 28 i tabel 3 bør trækkes ud af eksamenssættet, før scorerne omregnes til karakterer, da det er skadeligt for eksamenssættets samlede validitet og derfor også uretfærdigt overfor de studerende. Generelt er det en god ide at begynde med at screene eksamenssættet for items med DI tæt på 0 eller endda negative værdier. Dernæst er det en god ide at tjekke, om svarnøglen for disse items faktisk var korrekt. Hvis der ingen fejl er i svarnøglen, bør man tjekke, om der er konstruktionsfejl i disse spørgsmål. Det kan anbefales, at man er flere om at vurdere spørgsmål for fejl, da det er en udfordrende opgave, som kræver både faglig ekspertise og konsensus (O'Neill et al., 2019). Hvis man ikke finder nogen konstruktionsfejl, skal man ikke fjerne spørgsmålet fra eksamenssættet før karaktergivning, hvis man ellers har valgt spørgsmålets indhold med tilstrækkelig omhu i forhold til læringsmålene og læringsaktiviteterne (jf. tip 3 herover). I stedet bør man overveje, hvorfor de studerende så har svaret forkert. Kan der have været forvirrende eller modsatrettede

oplysninger i undervisningsforløbet (fx mellem underviser og materialer), har undervisningen været utilstrækkelig etc. er oplagte spørgsmål, man bør overveje. Statistikken kan således både bruges til kvalitetssikring af eksamen og til udvikling af undervisning. Statistikken eksemplificeret i tabel 2 kan også være nyttig i forbindelse med eksamensklager, fordi den repræsenterer evidens for kvaliteten af enkelte spørgsmål/ hele eksamen, som baserer sig på hele populationens faktiske præstationer, hvilket bør vægte tungt i forhold til at støtte eller afvise enkeltindividers kritik.

Konklusion

Der findes naturligvis store mængder af viden om MC-spørgsmål og -prøver i diverse udmærkede lærebøger, manualer og videnskabelige artikler m.m., som også er relevant. Jeg har fx bevidst fravalgt at give tips i bydeform om et kontroversielt emne som *negative marking* (korrektion for gætteri), om hvilket der ikke er konsensus i det akademiske samfund (Burton, 2005; O'Neill, 2017), men som af og til optager undervisere. De otte tips beskrevet herover er tænkt som udvalgte, indledende og grundlæggende viden om MC som prøveformat og som konkrete forslag til praksis, som der er bredere konsensus om.

Referencer

Ahmad, R. G., & Hamed, O. A. (2014). Impact of adopting a newly developed blueprinting method and relating it to item analysis on students' performance. *Medical Teacher, 36*(sup1), S55-S61.

American Educational Research Association, A. P. A., National Council of Measurement in Education. (2014). *Standards for educational and psychological testing*. Whashington (DC): American Educational Research Association.

Biggs, J., & Tang, C. (2007). *Teaching for quality learning at unversity* (3rd ed.). Maidenhead, England: Society for Research into Higher Education & Open University Press.

Burton, R. F. (2005). Multiple-choice and true/false tests: myths and misapprehensions. *Assessment & Evaluation in Higher Education, 30*(1), 65-72.

Butler, A. C. (2018). Multiple-choice testing in education: Are the best practices for assessment also good for learning? *Journal of Applied Research in Memory and Cognition, 7*(3), 323-331.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616.

Case, S., & Swanson, D. B. (2002). *Constructing written test questions for the basic and clinical sciences* (3rd ed.). Philadelphia (PA): National Board of Medical Examiners.

Clay, B., & Root, E. (2001). *Is this a trick question?: A short guide to writing effective test questions*. Kansas Curriculum Center.

Dent, J., Harden, R. M., & Hunt, D. (2017). *A practical guide for medical teachers*: Elsevier Health Sciences.

Downing, S. M. (2002a). Assessment of knowledge with written test forms. In *International handbook of research in medical education* (pp. 647-672): Springer.

Downing, S. M. (2002b). Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. *Advances in Health Sciences Education, 7*(3), 235-241.

Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. *Medical education, 37*(9), 830-837. doi:10.1046/j.1365-2923.2003.01594.x

Downing, S. M., & Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Medical Education, 38*(3), 327-333.

Downing, S. M., & Yudkowsky, R. (2009). *Assessment in health professions education*. New York, NY: Routledge.

Eweda, G., Bukhary, Z. A., & Hamed, O. (2020). Quality assurance of test blueprinting. *Journal of Professional Nursing, 36*(3), 166-170.

Fellenz, M. R. (2004). Using assessment to support higher level learning: the multiple choice item development assignment. *Assessment & Evaluation in Higher Education, 29*(6), 703-719.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items* (3rd ed.): Routledge.

Haladyna, T. M. (2018). Developing Test Items for Course Examinations. IDEA Paper# 70. *IDEA Center, Inc.* Retrieved from https://files.eric.ed.gov/fulltext/ED588351.pdf

Haladyna, T. M., & Downing, S. M. (1993). How many options is enough for a multiple-choice test item? *Educational and Psychological Measurement, 53*(4), 999-1010.

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*: Routledge.

Harden, R. M., & Laidlaw, J. M. (2013). Be FAIR to students: four principles that lead to more effective learning. *Medical Teacher, 35*(1), 27-31.

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement* (pp. 17-64). Westport, CT: ACE/Praeger.

Kilgour, J. M., & Tayyaba, S. (2016). An investigation into the optimal number of distractors in single-best answer exams. *Advances in Health Sciences Education, 21*(3), 571-585.

Krathwohl, D. R. (2002). A revision of Bloom's taxonomy: An overview. *Theory into practice, 41*(4), 212-218.

NBME. (2019). *Test Blueprinting II: Creating a Test Blueprint*. Retrieved from https://www.nbme.org/PDF/Publications/Test-Blueprinting-Lesson-2.pdf:
O'Neill, L. D. (2017). Aben der nægtede at dø. Multiple choice-prøver og korrektion for gætteri. *MONA-Matematik-og Naturfagsdidaktik*(1).

O'Neill, L. D., Mortensen, S. M. R., Nørgaard, C., Øvrehus, A. L. H., & Friis, U. G. (2019). Screening for technical flaws in multiple-choice items. A generalizability study. *Dansk Universitetspaedagogisk Tidsskrift, 14*(26), 51-65.

Palmer, E. J., & Devitt, P. G. (2007). Assessment of higher order cognitive skills in undergraduate education: modified essay or multiple choice questions? Research paper. *BMC Medical Education, 7*(1), 49.

Paniagua, M., Swygert, K., Haist, S., Merrill, J., Hussie, K., Deruchie, K., . . . Tyson, J. (2016). Constructing written test questions for the basic and clinical sciences. *Philadelphia, PA: National Board of Medical Examiners*.

Roberts, C., Newble, D., Jolly, B., Reed, M., & Hampton, K. (2006). Assuring the quality of high-stakes undergraduate assessments of clinical competence. *Medical Teacher, 28*(6), 535-543.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24*(2), 3-13.

Royal, K. D., & Stockdale, M. R. (2017). The impact of 3-option responses to multiple-choice questions on guessing strategies and cut score determinations. *Journal of Advances in Medical Education & Professionalism, 5*(2), 84-89.

Sam, A. H., Field, S. M., Collares, C. F., Vleuten, C. P., Wass, V. J., Melville, C., . . . Meeran, K. (2018). Very-short-answer questions: reliability, discrimination and acceptability. *Medical education*.

Scully, D. (2017). Constructing multiple-choice items to measure higher-order thinking. *Practical Assessment, Research, and Evaluation, 22*(1), 4.

Varma, S. (2006). Preliminary item statistics using point-biserial correlation and p-values. *Educational Data Systems Inc.: Morgan Hill CA*.https://eddata.com/wp-content/uploads/2015/11/EDS_Point_Biserial.pdf