Dansk Universitetspædagogisk Tidsskrift

Tema Nye perspektiver på evalueringsformer i universitetspædagogik

Årgang 12 nr. 23 / 2017

Titel **Effekt af standardiserede studenterevalueringer på udvikling af undervisning**

Forfatter Frederik Voetmann Christiansen og Simon Sebastian Haag

Sidetal 20-36

Udgivet af Dansk Universitetspædagogisk Netværk, DUN

URL http://dun-net.dk/

Effekt af standardiserede studenter-evalueringer på udvikling af undervisning

Frederik Voetmann Christiansen, lektor, Institut for Farmaci, Københavns Universitet

Simon Sebastian Haag, stud.med., Københavns Universitet

Videnskabelig artikel, fagfællebedømt

Undervisningen evalueres i stigende grad gennem studenterbaserede, standardiserede spørgeskemaer. Evalueringerne tjener flere formål, herunder kvalitetssikring og kvalitetsudvikling. Fokus i denne kvantitative undersøgelse er at finde ud af, om de standardiserede evalueringer forbedrer undervisningens kvalitet. Den grundlæggende hypotese er, at forbedringer burde føre til bedre resultater i evalueringerne over tid. Undersøgelsen er baseret på data fra bacheloruddannelsen i medicin ved Københavns Universitet. Vi beskriver udviklingen i de enkelte kurser i seks på hinanden følgende semestre fra 2011 til 2013 og analyserer udviklingen i enkeltspørgsmålene i spørgeskemaerne. Trods store udsving på tværs af semestrene finder vi ikke evidens for, at de standardiserede spørgeskemaer generelt forbedrer undervisningens kvalitet.

Introduktion

Studenterevalueringer af undervisningen udgør en væsentlig del af undervisningen og kan hjælpe underviserne til løbende at justere indhold og undervisningsaktiviteter til den specifikke gruppe af studerende. Danske Universiteter er forpligtede til at gennemføre studenterevalueringer af undervisning og til at anvende resultaterne af disse systematisk (Akkrediteringsinstitution 2013, s. 13). Studenterevalueringer kan antage mange forskellige former, og der er ingen specifikke krav til, hvordan studenterevalueringerne skal udformes. En meget udbredt evalueringsform er standardiserede spørgeskemaer, der uddeles til de studerende efter endt undervisning, og det er sådanne skemaer, vi vil fokusere på i denne artikel. Standardiserede spørgeskemaer kan være skemaer udarbejdet af underviserne selv, men stadigt flere steder er der tale om mere eller mindre generiske skemaer, der distribueres fra centralt hold, f.eks. gennem fakultetet. Der er næppe tvivl om, at akkrediteringssystemets krav om systematik i evalueringsindsatsen har bidraget til udbredelsen af centraliserede modeller.

I en nylig undersøgelse lavet af Danmarks Evalueringsinstitut peger nogle aktører fra uddannelsesområdet på, at den centraliserede model anses for at have stordriftsfordele, og at den er nem at have med at gøre for underviserne (EVA 2015, s. 42). De centralt initierede evalueringer har dog ifølge EVA-rapporten visse udfordringer. Et væsentligt problem i forbindelse med centralt initierede studenterevalueringer er, at nogle undervisere føler, at evalueringerne ikke er relevante i forhold til den specifik-

ke undervisning (EVA 2015, s. 29). Et andet væsentligt problem kan være, at svarprocenterne på skemaerne ofte er – eller over tid bliver – meget lave. Tidspunktet for udsendelsen og uklarhed omkring opfølgningen på evalueringerne er to forklaringer på de lave svarprocenter, der gives i rapporten.

Forskningslitteraturen om brugen af standardiserede spørgeskemaer til studerende er ikke entydig. På den ene side findes en omfattende litteratur omkring særlige evalueringsinstrumenter, der er validerede og afprøvede. Nogle af de væsentligste er Ramsdens "Course Experience Questionnaire" (CEQ) og Marshs "Students' Evaluations of Educational Quality" (SEEQ), der begge er yderst velbelyste og bredt anvendte (Ramsden, 1991, Marsh, 1982). De to instrumenter adskiller sig fra hinanden ved, at SEEQ fokuserer på den enkelte underviser, mens CEQ fokuserer på undervisningsenheden og anvendes bredere til evaluering af forløb eller hele uddannelser. Fælles for de to er, at instrumenterne antages at knytte an til studerendes læringsudbytte og måle "teaching quality" eller "teaching effectiveness". Dette udbytte antages at kunne vurderes samlet ud fra en række forskellige parametre, der vides at korrelere positivt med studerendes udbytte. I tabel 1 ses de parametre, der ligger til grund for hhv. CEQ og SEEQ.

Tabel 1: Dimensioner, der indgår i hhv. CEQ og SEEQ

CEQ	SEEQ
Good Teaching	Learning
	Instructor Enthusiasm
Clear Goals	Organisation
	Breadth of coverage
Appropriate Assessment	Group Interaction
	Examinations
	Indvidual rapport
Emphasis on Independence	Assignments
Appropriate workload	Workload/Difficulty

Som nævnt er de to instrumenter validerede på flere forskellige måder, bl.a. gennem korrelationer til studerendes eksamensresultater, gennem undervisernes egenvurderinger og de studerendes tilgange til læring.

Studenterevalueringer i form af standardiserede spørgeskemaer kan tænkes at være påvirket af baggrundsfaktorer eller bias, altså at faktorer, der ikke har noget med undervisningens kvalitet at gøre, påvirker resultatet. En lang række ældre og nyere studier undersøger forskellige former for bias, som kan påvirke de studerendes besvarelser – det gælder sådanne faktorer som forudgående interesse, forventet karakter, arbejdsbyrden, holdstørrelsen, underviserens køn, titel, fagområdet, placeringen i studiet, anonymitet m.fl. F.eks. fandt Kwan (1999), at resultaterne påvirkedes af bl.a. holdets størrelse, den akademiske disciplin, kursustypen, og om kurserne var på grunduddannelsen eller på videregående niveau. I modsætning hertil konkluderer Marsh og Bailey (1993), at de fleste kilder til bias ikke påvirker resultatet af studenterevalueringerne i væsentlig grad. Flere nyere studier peger dog på, at bl.a. kønsbias og de studerendes forventning til karakter spiller en væsentlig rolle i forbindelse med studenterevalueringer af undervisere (Stark og Freishtat, 2014; Boring, Ottobin og Stark, 2016).

Der er flere forskellige formål med at gennemføre evalueringer (se f.eks. EVA 2015, s. 20). Blandt de væsentligste er evalueringernes funktion i forhold til *kvalitetssikring* af undervisningen og evalueringernes bidrag til *kvalitetsudvikling* af undervisningen. I denne artikel vil vi undersøge anvendelsen af studenterevalueringer ved bacheloruddannelsen i medicin ved Københavns universitet. Ved Det Sundhedsvidenskabelige Fakultet betones disse to formål med evalueringen (Det Sundhedsvidenskabelige Fakultet 2016). Ifølge fakultetets retningslinjer er evalueringens formål at "Muliggøre løbende kvalitetsudvikling og kvalitetssikring af undervisningen ved at give evalueringsmæssige input til kursusansvarlige og undervisere". I denne undersøgelse vil vi alene fokusere på standardiserede spørgeskemaers potentiale i forhold til kvalitets*udvikling* af undervisningen, og vi vil ikke forholde os til skemaernes eventuelle kvalitetssikrende funktion.

Hvis evalueringerne bidrager til at forbedre undervisningens kvalitet, er det en rimelig antagelse, at evalueringsresultaterne bør ændre sig i positiv retning over tid. Altså at underviserne bruger evalueringsresultaterne til at foretage ændringer og forbedringer i kurset, og at dette fører til en forbedring af undervisningens kvalitet over tid, som er målbar i evalueringerne.

Dette antog Kember, Leung og Kwan (2002) i en undersøgelse ved det Polytekniske Universitet i Hong Kong. Udviklingen i evalueringsresultaterne over tid blev vurderet ved 25 institutter over en 3-4 årig periode. Undersøgelsen havde et nedslående resultat: Det var ikke muligt at konstatere positive ændringer i evalueringsresultaterne over tid. Dette på trods af, at de i spørgeskemaet indgående dimensioner var meget lig de dimensioner, der indgår i de validerede spørgeskemaer beskrevet ovenfor. Forfatterne påpegede i deres diskussion, at evalueringerne ikke i sig selv kan forbedre undervisningen, og at resultatet formodentlig måtte ses i lyset af utilstrækkelig

opfølgning og manglende incitamentstrukturer i organisationen. Såfremt der ikke følges op på evalueringerne på relevante måder, fører anvendelsen af spørgeskemaer, ifølge denne undersøgelse, ikke til forbedring af undervisningen. Kember, Leung, and Kwan (2002) anfører, at det ville være interessant at undersøge, om deres resultater kan generaliseres til andre universiteter. Undersøgelsen af Kember, Leung og Kwan (2002) har dannet baggrund for vores problemstilling og studiedesign, hvor vi vil undersøge, om brugen af standardiserede spørgeskemaer fører til forbedring af undervisningen over tid. Studiet er for en stor del baseret på et forudgående kandidatspeciale (Haag, 2016).

Metode

Analysen bygger på studerendes besvarelser af spørgeskemaer på bacheloruddannelsen i medicin i en periode på 6 semestre fra foråret 2011 til efteråret 2013 (for ét af de indgående kurser dog kun udviklingen over 5 semestre). I alt indgår 18 forskellige kurser i undersøgelsen med en gennemsnitlig besvarelsesprocent på 49%. Antallet af besvarelser for de enkelte kurser ligger mellem 76 og 214. De fleste af kurserne havde i omegnen af 250 studerende, hvorfor den relativt lave svarprocent ikke burde føre til væsentlig bias i besvarelserne (Nulty, 2008). Da uddannelsen har halvårligt optag, er alle kurserne afholdt i alle semestre, hvilket har givet os mulighed for at undersøge udviklingen i de enkelte kurser over tid. Perioden indskrænker sig til disse 6 semestre, da denne periode var den længste med tilgængelige data og uden væsentlige ændringer i skemaet. Skemaet er efterfølgende blevet ændret væsentligt. I perioden blev alle kurser på medicinstudiet evalueret efter hvert gennemløb på basis af et delvist standardiseret spørgeskema. Hver studerende udfyldte på frivillig basis spørgeskemaet elektronisk efter hvert afsluttet kursus. Der blev stillet mellem 7 og 22 spørgsmål i et spørgeskema afhængig af kursets indhold og form, hvor nogle er specifikke for de enkelte kurser, og andre er generelle spørgsmål. De seks hyppigst stillede generelle spørgsmål, der vurderes med Likert-skala, kan ses i tabel 2.

Tabel 2: De seks generelle spørgsmål, der indgår i de fleste kursers evaluering

Målene	I hvilken udstrækning mener du, at målene for kurset er opfyldt?
Forelæsninger	Hvordan vurderer du dit udbytte af forelæsningerne?
SAU	Hvordan vurderer du dit udbytte af holdundervisningen (SAU-timerne)?
Relevans	Hvordan vurderer du relevansen af kurset for dit videre studie og dit fremtidige arbejde som læge?

Tilfredshed	I hvilket omfang er du generelt tilfreds med kursets indhold?
Opbygning og sekvens	I hvilken grad finder du, at kursets opbygning og sekvens er hensigtsmæssig?

Mindre hyppigt spørges ind til f.eks. udbyttet af demonstrationer og øvelser. Derudover indgår baggrundsinformation omkring spørgsmål vedrørende de studerendes deltagelse, hold mv. Endelig kan spørgeskemaerne suppleres med kommentarer, der skal være af konstruktiv karakter. Disse kommentarer indgår ikke i analysen.

De fleste spørgsmål besvares med afkrydsning på en 7-trins Likert-skala, hvor 1 er "uacceptabelt", 4 er "acceptabelt", og 7 er "optimalt", og vi afgrænser vores undersøgelse til disse spørgsmål. Et enkelt generelt spørgsmål (om det oplevede faglige niveau) går også igen i mange skemaer, men er udeladt fra analysen, da skaleringen på dette spørgsmål var anderledes end for de øvrige kurser.

I modsætning til SEEQ og det af Kember, Leung & Kwan (2002) anvendte skema er det ikke *underviseren,* men *kurset,* der er i fokus i det ved fakultetet anvendte spørgeskema. Dette skyldes formodentlig, at de fleste kurser ved uddannelsen er meget store, og hvert kursus involverer mange forskellige undervisere til f.eks. varetagelse af forelæsningerne, holdundervisningen, øvelser mv.

I Kember, Leung & Kwans undersøgelse var det ikke nødvendigvis de samme kurser, der indgik i et instituts undervisning i to på hinanden følgende semestre. I vores undersøgelse har vi mulighed for at undersøge, om og hvordan evalueringerne af *de samme* kurser ændrer sig i seks på hinanden følgende semestre.

Kember, Leung og Kwan (2002) analyserede ændringer over tid ved forskellige institutter (departments) ved brug af en ”Multivariate Analysis of Variance” (MANOVA) af data, som vi også anvender i den foreliggende undersøgelse. Kember *et al* undersøgte, om der var signifikante forskelle på besvarelserne på institutternes kurser i forskellige år, altså om forskellene i evalueringsresultaterne var tilfældige eller ej. I MANOVA-analysen sammenlignes gennemsnit, idet man tager højde for både spredningen (variansen) indenfor grupperne – eksempelvis at et hold studerende generelt er mere negativt indstillet end det andet – og afhængigheden mellem spørgsmålene, idet det antages, at et positivt (hhv. negativt) svar på det ene spørgsmål fører til et mere positivt (hhv. negativt) svar på det andet.

I vores undersøgelse har vi desuden undersøgt den tværgående udvikling i de hyppigst stillede spørgsmål. Således undersøgte vi, om der er noget, der tyder på, at f.eks. de studerendes vurdering af udbytte af forelæsningerne forandrer sig over tid. Dette er undersøgt med “Analysis of Variance” (ANOVA), som er en simplere statistisk

metode, der blot sammenligner forskelle i gruppernes svar (gennemsnit og varians) for en enkelt variabel.

Output af hhv. MANOVA og ANOVA angives ved F-test-værdien og p-værdien. F-test-værdien er forholdet mellem variansen af gennemsnittet i den samlede prøve og variansen i de enkelte tilfælde. En F-værdi tæt på 1 vil dermed indikere, at nulhypotesen bør fastholdes. Statistisk signifikans er i denne analyse defineret som vanligt ved p-værdi lavere end 0.05.

Værdien p giver kun oplysninger om, hvorvidt en eventuel forandring i gennemsnittene over tid er signifikant, men ikke om ændringen er positiv eller negativ. For at undersøge, om evalueringsresultaterne ændrer sig positivt over tid, er der således også behov for at analysere udviklingen på andre måder, såfremt der er statistisk signifikante ændringer. For MANOVA-analysens vedkommende er dette gjort gennem visualisering af data og kvalitativ analyse. For ANOVA-analysen har vi gennemført Tukey HSD-analyse, hvorved forskelle mellem semestrene kan sammenlignes statistisk.

Resultater

Analyse af de enkelte kurser

MANOVA-analysen af udviklingen over de undersøgte semestres enkelte kurser peger på, at der for 17 ud af 18 kursers vedkommende *er* signifikante forskelle på resultaterne mellem (nogle) semestre. I Tabel 3 er opgjort resultater af MANOVA-analysen med angivelse af output som F-værdi og p-værdi. Kun kursus 15 viser ingen signifikant ændring over tid. At der er store forskelle på de afholdte kurser fra gang til gang siger, som nævnt, ikke noget om, hvorvidt der er en generel positiv udvikling i evalueringsresultaterne over tid.

Tabel 3: Resultater fra MANOVA-analyse af de enkelte kurser angivet som F-værdi og p-værdi. Alle kurser – med undtagelse af kursus 15 – ændrer sig signifikant over tid. Om ændringen er en forbedring eller forværring kan ikke ses ud fra værdierne.

Kursus	F	p
1	5.642	>0.001
2	3.130	>0.001
3	3.068	>0.001
4	4.163	>0.001
5	6.579	>0.001

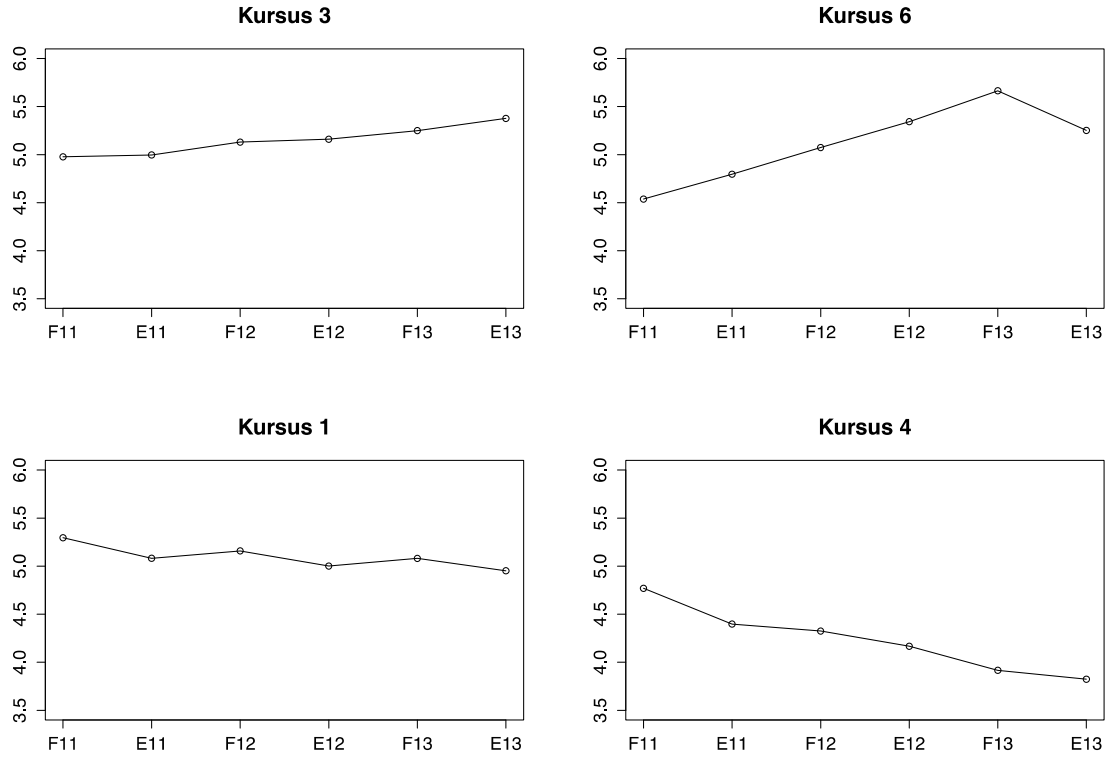
Kursus	F	p
10	1.544	0.004
11	1.621	0.012
12	1.656	0.009
13	2.849	>0.001
14	3.300	>0.001

6	6.304	>0.001
7	3.062	>0.001
8	6.857	>0.001
9	3.330	>0.001

15	1.359	0.093
16	4.184	>0.001
17	3.161	>0.001
18	2.065	>0.001

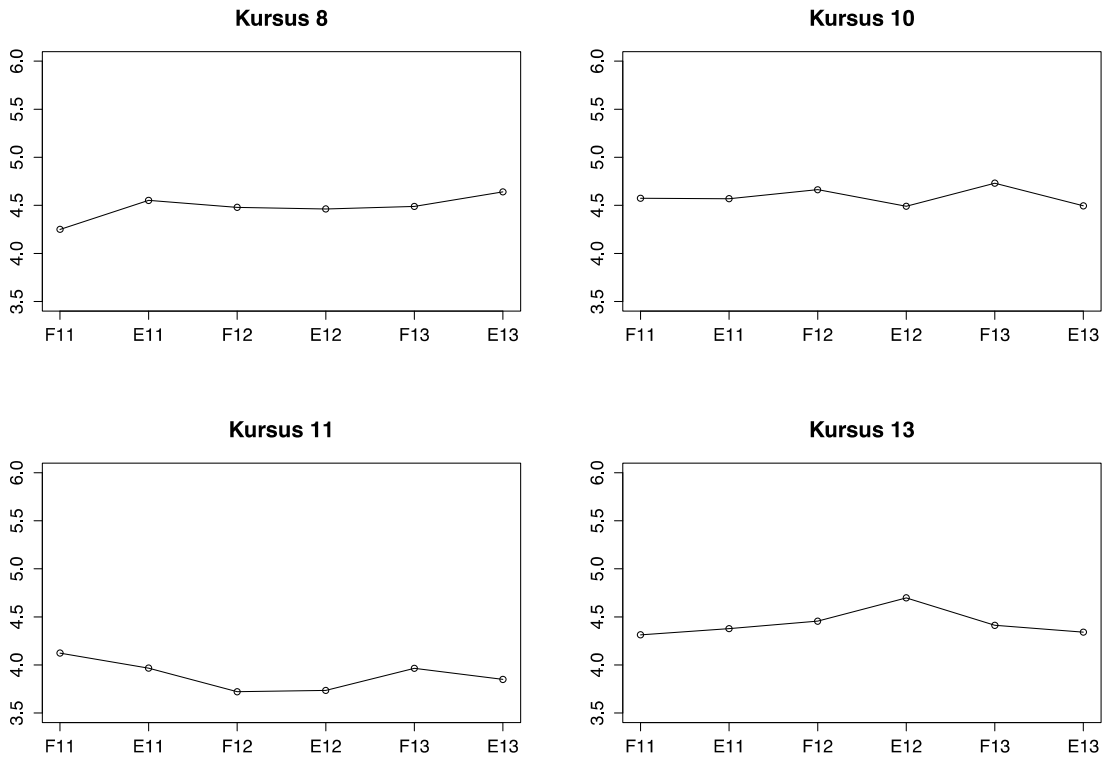
For at undersøge den konkrete udvikling i de enkelte kurser har vi lavet visualiseringer af de samlede gennemsnit for de enkelte kurser for at bestemme den tidslige udvikling i resultaterne. Plottene viser, at kun to ud af 18 kurser har en tydelig positiv udvikling over tid, mens to andre kurser har en negativ udvikling over tid (jf. figur 1).

Figur 1: Kurser med hhv. positive og negative udviklinger over tid baseret på kvalitativ analyse. X-aksen viser tidspunktet (F11 = forårssemestret 2011, E12 = efterårssemestret 2012 osv.). Y-aksen viser den gennemsnitlige bedømmelse på 7-trins Likert-skalaen begrænset fra 3.5-6 for bedre visualisering.



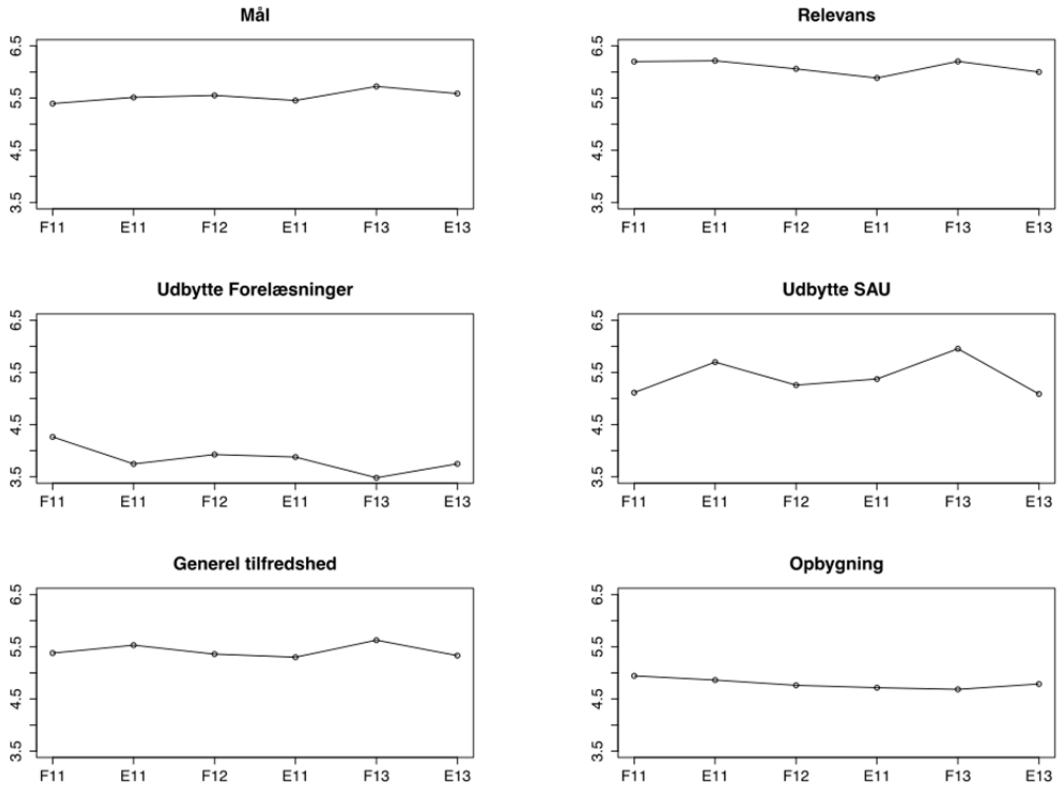
På baggrund af den kvalitative analyse er der for de øvrige 14 kursers vedkommende ikke baggrund for at konkludere, at der er en positiv udvikling over tid. For disse kurser er der tale om enten meget konstante niveauer over tid eller udsving fra et semester til et andet uden nogen klar trend. Eksempler på blandede og konstante udviklinger kan ses i figur 2.

Figur 2: Eksempler på kurser uden en klar trend baseret på kvalitativ analyse. Se signaturforklaringen i Figur 1.



Visualisering af de enkelte spørgsmål inden for hvert kursus bekræfter det billede, der er givet ovenfor. De fire kurser med hhv. stigende og faldende tendens over tid viser også en hhv. stigende og faldende trend i (de fleste af) de spørgsmål, der indgår. For de øvrige 14 kursers vedkommende er der konstante eller meget blandede besvarelser. Figur 3 viser et eksempel på et kursus uden en entydig udvikling i enkeltspørgsmålene over tid. Eksemplet viser 6 af de 8 spørgsmål, der indgår i analysen for dette kursus.

Figur 3: Eksempel på udvikling i underspørgsmål for et kursus uden klar trend. Se signaturforklaringen i hhv. Figur 1 og Tabel 2.



Tværgående analyse af spørgsmålene

Som nævnt er der seks spørgsmål, der går igen i langt de fleste af spørgeskemaerne. Er der nogle af disse spørgsmål, der viser en positiv udvikling over tid, når man kigger på tværs af kurserne? Er der, for eksempel, en positiv udvikling i studerendes vurdering af udbytte af forelæsningerne på studiet over tid? For at undersøge dette har vi gennemført ANOVA-analyser af besvarelserne for de enkelte spørgsmål på tværs af alle kurserne. For visse af kurserne indgik det samme spørgsmål i flere varianter, f.eks. spørges i nogle kurser til udbyttet af forelæsninger i forhold til forskellige dele af kurset, så dette spørgsmål blev stillet op til 3 gange i et enkelt kursus. I analysen er alle besvarelser på det samme spørgsmål medtaget, hvorved antallet af spørgsmål kan være større (eller lavere) end antallet af kurser.

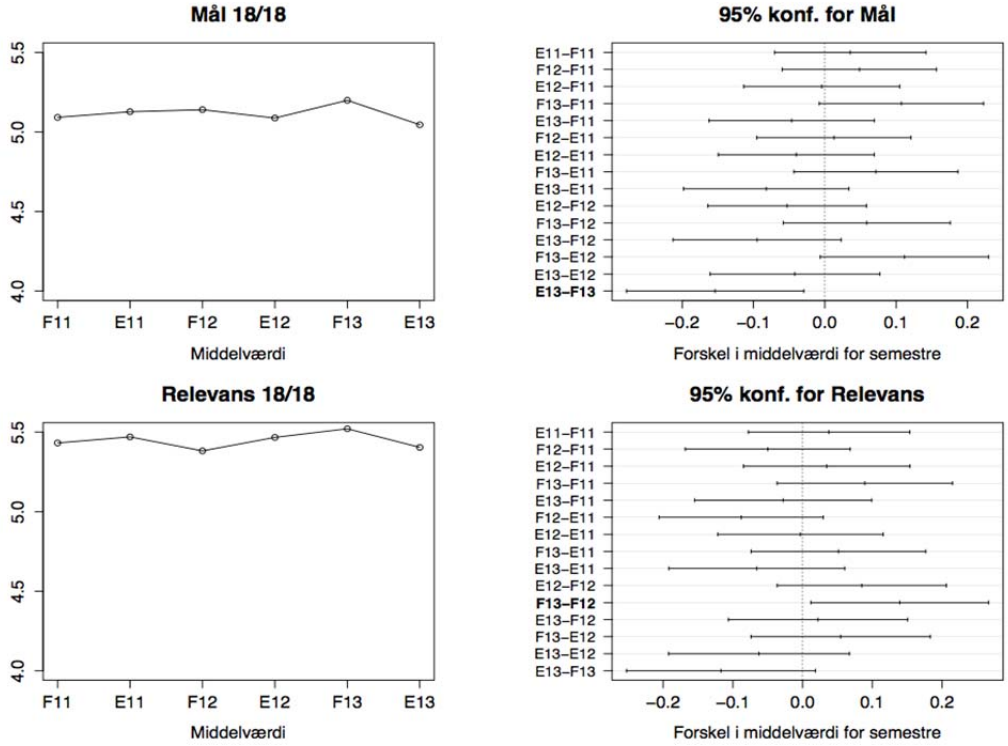
Tabel 4: ANOVA-analyse af enkeltspørgsmål på tværs af kurserne angivet som F-værdi og p-værdi. I nogle kurser spørges ind til samme emne med flere spørgsmål, hvorfor der kan være større antal spørgsmål end kurser. Spørgsmålene er nærmere beskrevet i Tabel 2.

Spørgsmål	**Antal kurser**	**Antal spørgsmål**	**F**	**p**
Målene	18	18	3.082	0.0090
Forelæsninger	15	23	6.252	0.0000
SAU	16	23	3.216	0.0070
Relevans	18	18	2.523	0.0273
Tilfredshed	17	17	1.975	0.0790
Opbygning og sekvens	18	18	4.946	0.0000

ANOVA-analyse af de enkelte spørgsmål viser signifikante forskelle ($p<0.05$) mellem besvarelserne fra forskellige semestre for fem ud af seks af de gennemgående spørgsmål (jf. Tabel 4). Eneste undtagelse er spørgsmålet om tilfredshed ($p=0.08$).

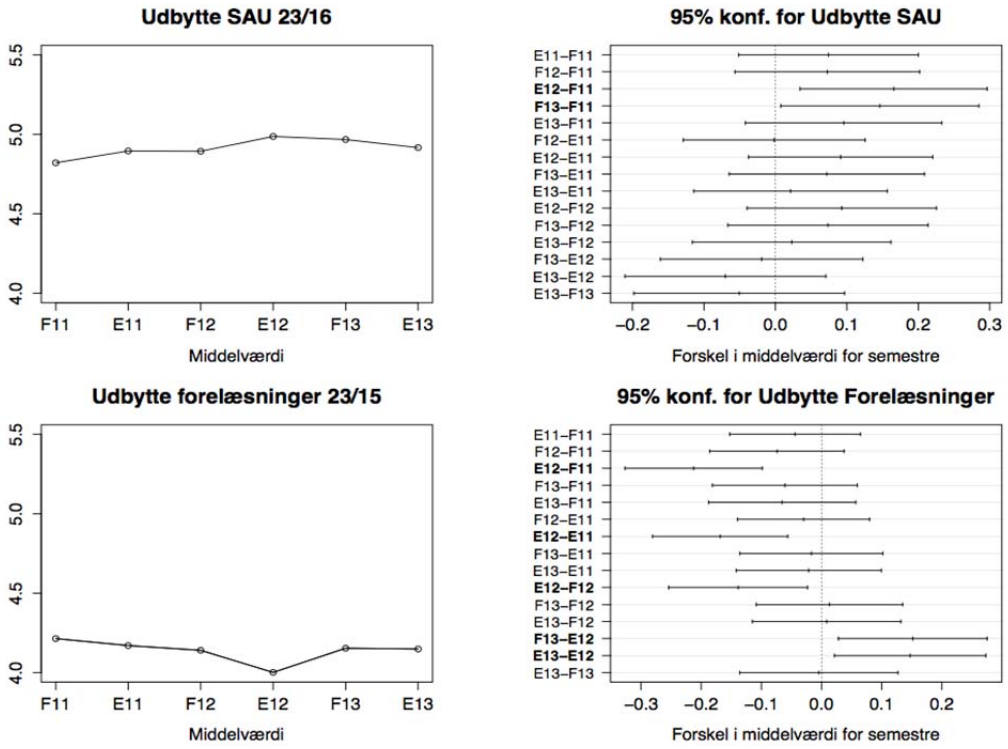
For alle spørgsmål med signifikante ændringer er der gennemført parvis Tukey HSD-analyse, hvorved forskellene i middelværdi mellem de forskellige semestre kan visualiseres. I Tukey HSD-analysen tages højde for, at antallet af sammenligninger er højt med evt. falsk positive fund til. Tukey HSD-testen korrigerer for dette. Figurerne 4-6 viser plots af gennemsnit for besvarelserne i de enkelte semestre og resultaterne af Tukey HSD-analysen med markerede konfidensintervaller (95%). Semesterpar, hvor forskellen i middelværdi (med 95% konfidensinterval) er større eller mindre end nul, er markeret med fed i figurerne til venstre.

Figur 4: Udvikling i spørgsmål "Mål" og "Relevans" og tilhørende parvis Tukey HSD-analyse. Når konfidensintervallerne ikke krydser 0, er der en signifikant forskel på de to semestre – disse er markeret med fed.



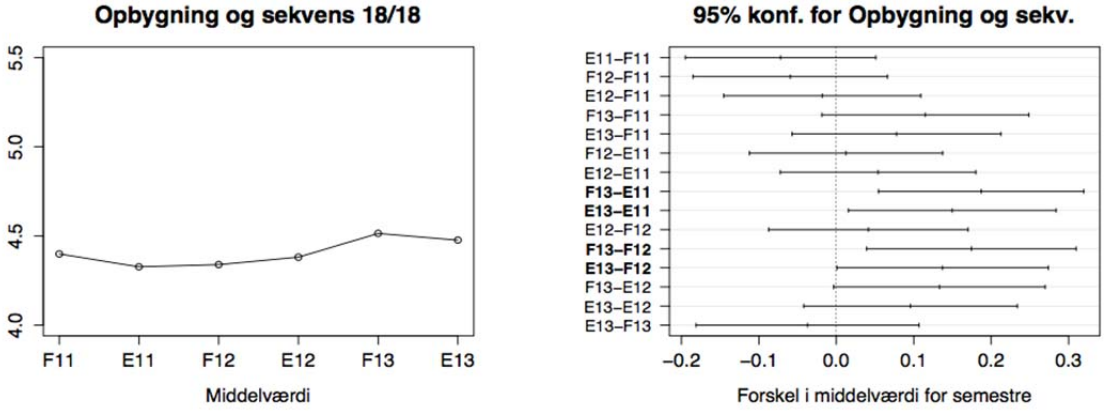
For spørgsmålene "Mål", "Relevans", "Forelæsninger" og "SAU" (jf. Tabel 2) gælder, at resultaterne kan forklares med, at et enkelt semester ligger væsentligt lavere eller højere end de øvrige. Som eksempel kan tages spørgsmålet "Forelæsninger" (jf. figur 5). Her ligger værdien for efteråret 2012 væsentligt lavere end for de øvrige semestre, og dette forhold alene forklarer, at så mange parvise forskelle i middelværdier er forskellige fra nul.

Figur 5: Udvikling i spørgsmål "SAU" og "Forelæsninger" og tilhørende parvis Tukey HSD-analyse



For spørgsmålet "Opbygning og Sekvens" (figur 6) kan det ses, at de to sidste semestre (F13 og E13) ligger markant højere end to tidligere semestre (E11 og F12). For dette ene spørgsmål (ud af 6) finder vi det derfor rimeligt at sige, at der har været en positiv udvikling over tid. Vi finder, at der med en enkelt undtagelse ikke er en positiv udvikling over tid i besvarelsen af spørgsmålene.

Figur 6: Udvikling i spørgsmål "Opbygning og sekvens" og tilhørende parvis Tukey HSD-analyse



Diskussion

Resultatet af denne undersøgelse af studenterevalueringer af undervisningen ved medicinuddannelsen i København svarer til det, Kember, Leung & Kwan (2002) fandt ved det Polytekniske Universitet i Hong Kong. Der er – trods store udsving mellem semestrene – ikke noget, der tyder på, at der er generelle stigninger i resultaterne over tid. Spørgsmålet er, om dette fund er ensbetydende med, at kvaliteten af undervisningen ikke bliver bedre. Dette spørgsmål må ses i lyset af spørgeskemaets validitet: Afspejler spørgsmålene reelt forhold, der er knyttet til de studerendes læring?

Det er ikke klart, på hvilken baggrund Det Sundhedsvidenskabelige Fakultet har udviklet de spørgsmål, de studerende skal tage stilling til i evalueringerne, og der kan være grund til at overveje validiteten af visse af spørgsmålene – hvilket fakultetet også har gjort efterfølgende. I forhold til spørgsmålene vedrørende udbyttet af forelæsninger og SAU-undervisning er det f.eks. ikke klart, hvad "udbytte" vil sige. Betyder det, at forelæsningen har hjulpet den studerende med at forstå stoffet? Eller at den studerende blev motiveret til at læse mere om emnet derhjemme? Og hvad skal underviserne lægge i, at der evt. gives en lav score i disse spørgsmål? Hvilken type opfølgning skal en lav score give anledning til?

Spørgsmålet vedrørende "Relevans" er det næstbedst evaluerede i hele spørgeskemaet med et gennemsnit på 5.4 point på tværs af alle kurser. Kan de studerende reelt tage stilling til relevansen af undervisningen i forhold til deres fremtidige virke som læge?

Der kan således være grund til at betvivle det anvendte spørgeskemas validitet, og der kan være inspiration at hente i nogle af de validerede instrumenter, vi har beskrevet, f.eks. SEEQ. Det kunne være spørgsmål om underviserens entusiasme for at undervise kurset, om stimulation af den studerendes interesse i faget, om interaktionen mellem de studerende og underviseren, om undervisningens organisation og klarhed. Alle er muligheder, der i litteraturen beskrives som indikatorer for effektiv undervisning (Marsh, 1982; Rowley, 2003). Brug af mere specifikke spørgsmål kunne hjælpe til at identificere problemer og give underviserne handlemuligheder.

En vanskelighed ved at anvende spørgsmål fra de validerede skemaer er, at spørgeskemaer som SEEQ fokuserer på *underviseren* i stedet for *kurset* generelt. Da kurserne på medicinstudiet ofte undervises af mange forskellige undervisere, er det ikke ligetil at overtage spørgsmålene. Det vil derfor være nødvendigt at tilpasse spørgsmålene på relevante måder.

Selv med en sådan ændring af spørgeskemaets udformning er der grund til at stille spørgsmål ved, om standardiserede spørgeskemaer kan bidrage til kvalitetsudvikling af undervisningen. Spørgeskemaet anvendt af Kember et al (2002) spurgte netop til

disse forhold, og resultatet af deres undersøgelse var det samme: at der ikke kunne konstateres en positiv udvikling i resultaterne over tid. Kember et al. peger på, at især manglende opfølgning på evalueringerne og manglende incitamentstrukturer til belønning af god og fremragende undervisning er en del af forklaringen på, at der ikke kan konstateres positive udviklinger over tid. Det Sundhedsvidenskabelige Fakultet har efterfølgende indført procedurer, der skal bidrage til en stærkere opfølgning på evalueringerne, f.eks. skriftlige tilbagemeldinger fra kursuslederne på resultaterne af evalueringerne, men om dette er en tilstrækkelig opfølgning vides ikke. Der er næppe tvivl om, at positive evalueringsresultater generelt er forbundet med prestige og anerkendelse fra kolleger, men mange universitetsansatte oplever, at høj kvalitet i undervisningen ikke fører til karrieremæssige eller lønmæssige fordele. Ifølge Kvalitetsudvalgets rapport "Høje Mål" er det kun et lille mindretal af de universitetsansatte, der på tværs af sektoren oplever, at kvaliteten af deres undervisning har meget eller en del betydning for løn (11%), forfremmelse (ca. 13%) eller anden ledelsesmæssig anderkendelse (ca. 26%) (Kvalitetsudvalget, 2014, s. 85 – se også rapportens bilag 3, tabellerne 288-292).

Eventuel manglende opfølgning kan også skyldes manglende pædagogiske forudsætninger hos underviserne. Vi er ikke i tvivl om, at flertallet af underviserne gerne vil levere høj standard i deres undervisning, men det er muligt, at nogle ikke kan se, hvordan de kan gøre det bedre. Overfyldte powerpoint slides og monotone monologer er stadig en del af virkeligheden i undervisningen. På fakultetet indgår pædagogisk kompetenceudvikling i standarden for uddannelseskvalitet (SUND's standarder for uddannelseskvalitet), f.eks. skal nye undervisere deltage i universitetspædagogiske kurser, og der gives på fakultetet en række andre frivillige tilbud om deltagelse i workshops og kurser. Løbende pædagogisk kompetenceudvikling har dog i vid udstrækning været overladt til den enkelte. Fra marts 2016 stilles dog krav til de fastansatte om løbende pædagogisk efteruddannelse, som skal drøftes i forbindelse med MU-samtaler.

Med baggrund i denne analyse vil vi sætte spørgsmålstegn ved, om de standardiserede studenterspørgeskemaer bidrager positivt til udvikling af kvaliteten af undervisningen. Vi vil dog ikke udelukke, at studenterevalueringerne kan tænkes at have en kvalitetssikrende funktion – at systemet kan bidrage til, at "alarmklokkerne ringer", hvis evalueringsresultaterne pludselig falder markant eller ligger meget lavt over en længere periode. Selv hvis evalueringerne har en sådan funktion, finder vi, at det bør overvejes, om der ikke findes relevante alternativer i form af evalueringsformater og opfølgningsprocedurer, der kan udfylde såvel kvalitetssikrings- som kvalitetsudviklingsformålene. For blot at give ét eksempel på en alternativ måde at evaluere undervisningen på kan vi henvise til en model, der også kan anvendes af uddannelserne på Det Sundhedsvidenskabelige Fakultet – en model, der benævnes *Dialogmodellen.* Denne model "er centreret omkring en systematiseret dialog mellem

studerende, kursusansvarlige og institutter i undervisningsregi eller på særskilte dialogmøder. Kursusansvarlige og studerende foretager i fællesskab en skriftlig opfølgning på den mundtlige dialog" (Det Sundhedsvidenskabelige Fakultet, 2016). Modellen anvendes på en række af uddannelserne på fakultetet, herunder på odontologi, farmaci, folkesundhedsvidenskab m.fl. Modellen er baseret på, at der afholdes dialogmøder med undervisere og holdrepræsentanter med udgangspunkt i et fælles skriftligt produkt udarbejdet af de studerende. Antal af uddannelser ved fakultetet, der anvender modellen, har været stigende, men om denne model fører til kvalitetsudvikling af undervisningen over tid er ikke undersøgt og kan næppe undersøges kvantitativt.

Konklusion

Vi har undersøgt den tidslige udvikling i evalueringsresultaterne fra bacheloruddannelsen i medicin over 6 semestre fra foråret 2011 til efteråret 2013. Mens der i mange tilfælde er store forskelle på evalueringsresultaterne fra de forskellige semestre, finder vi ikke, at der generelt er en positiv udvikling i evalueringsresultaterne over tid. Vi finder heller ikke en generel positiv udvikling over tid, når enkeltspørgsmålene analyseres på tværs af de undersøgte kurser.

At der ikke generelt kan konstateres en positiv udvikling i evalueringsresultaterne over tid kan skyldes, at det anvendte skemas validitet er lav, da det kun til en vis udstrækning afspejler dimensioner, der vides at korrelere med studerendes læringsudbytte. Undersøgelsen af Kember, Leung og Kwan (2002) anvendte dog et skema, der må formodes at have højere validitet, og de nåede et tilsvarende nedslående resultat.

Samlet set peger vores undersøgelse af brugen af standardiserede spørgeskemabaserede studenterevalueringer på, at der er grund til at genoverveje, om standardiserede spørgeskemaer er den rette måde at udvikle undervisningen over tid på – særligt da underviserne ofte kun afsætter begrænset tid til evaluering af undervisningen i det hele taget. Der er mange andre måder at evaluere undervisningen på.

Det skal understreges, at undersøgelsen ikke har forholdt sig til, om de standardiserede spørgeskemaer kan have en kvalitets*sikrende* funktion, men alene peger på, at skemaerne i det konkrete tilfælde ikke har en konstaterbar kvalitets*udviklende* funktion.

Frederik Voetmann Christiansen *er lektor i naturvidenskabsdidaktik ved Institut for Farmaci ved Det Sundhedsvidenskabelige Fakultet, Københavns Universitet. Hans forskning og undervisning er primært inden for universitetspædagogik og uddannelsesforskning, natur- og sundhedsvidenskabelig didaktik, samt natur- og sundhedsvidenskabelig videnskabsteori.*

Simon Sebastian Haag *er medicinstuderende på Københavns Universitet og afslutter i 2017. Som illustrator for bl.a. det tyske forlag Thieme Stuttgart og websiden meditricks.de har han arbejdet med audio-visuelle, web-baserede læremetoder til medicinstuderende.*

Litteratur

Akkrediteringsinstitution, Danmarks. 2013. *Vejledning Om Institutionsakkreditering*. Danmarks Akkrediteringsinstitution.

Boring, A., K. Ottobin, and P. B. Stark. 2016. "Student Evaluations of Teaching (Mostly) Do Not Measure Teaching Effectiveness." *ScienceOpen Research*, Januar, 1–11.

Det Sundhedsvidenskabelige Fakultet. 2016. *Evaluering Af Undervisning På SUND*. København: Det Sundhedsvidenskabelige Fakultet.

EVA. 2015. *Undervisningsevaluering På de Videregående Uddannelser*. København: Rosendahls.

Haag, 2016. *Effekten af studenterbaserede, standardiserede spørgeskemaer på undervisningens kvalitet på Bacheloruddannelsen i Medicin*. Kandidatspeciale. København: Det Sundhedsvidenskabelige Fakultet, Københavns Universitet.

Kember, D., D. Y. P. Leung, and K. P. Kwan. 2002. "Does the Use of Student Feedback Questionnaires Improve the Overall Quality of Teaching?" *Assessment and Evaluation in Higher Education* 27 (5): 411–25.

Kwan, K.P. 1999. "How Fair Are Student Ratings in Assessing the Teaching Performance of University Teachers?" *Assessment and Evaluation in Higher Education* 24 (2): 181–95.

Kvalitetsudvalget, 2014. Høje Mål – Fremragende undervisning i de videregående uddannelser. Udvalget for kvalitet og relevans i de videregående uddannelser. Uddannelses- og forskningsministeriet. URL: http://www.ufm.dk/kvalitetsudvalget (Tilgået 15/11 2016).

Marsh, H. W. 1982. "SEEQ: A Reliable Valid, and Useful Instrument for Collecting Students' Evaluations of University Teaching." *British Journal of Educational Psychology* 52 (1): 77–95.

Marsh, H.W., and M. Bailey. 1993. “Multidimensional Students Evaluations of Teaching Effectiveness: A Profile Analysis.” *The Journal of Higher Education* 64 (1): 1–18.

Nulty, D. D. “The adequacy of response rates to online and paper surveys: what can be done?” *Assessment & Evaluation in Higher Education* 33(3): 301-314.

Ramsden, Paul. 1991. “A Performance Indicator of Teaching Quality in Higher Education: The Course Experience Questionnaire.” *Studies in Higher Education* 16 (2).

Rowley, J. 2003. Designing student feedback questionnaires. *Quality Assurance in Education* 11(3), pp. 142-149.

Stark, P. B., and Rechard Freishtat. 2014. “An Evaluation of Course Evaluations.” *ScienceOpen Research*.