Kontroversiel relevans: Tal der virker, er tal der splitter

Hjalte Meilvang[1], ph.d.-stipendiat, Institut for statskundskab, Københavns Universitet

Statistik og indikatorer er centrale redskaber i inddragelsen af viden i politik, forvaltning og offentlig styring, hvor de bidrager med tilsyneladende objektive beskrivelser. Det er imidlertid vanskeligt for en indikator at være på samme tid politisk relevant og ukontroversiel. Tal og statistik bliver *relevante*, når de bruges til noget. Men brug øger risikoen for *kontroverser*. Tal, der virker, virker ofte på en måde, nogen er imod. Denne artikel udfolder denne ide gennem en analyse af de nationale test i den danske folkeskole.

”In God we trust; all others bring data”

“Without data, you're just another person with an opinion.” (Deming 2017)

Disse spidsformuleringer af den amerikanske managementforsker W. Edwards Deming fortæller to ting om typiske opfattelser af kvantitative data. Tal er en foretrukken kommunikationsform i situationer, hvor man ikke stoler blindt på et budskabs afsender, og de udtrykker en form for ’sikker viden’, der er mere værd end individuelle holdninger. Med andre ord ses de som objektive i modsætning til subjektive. Som sådan spiller de en særlig rolle i det moderne massesamfund, hvor deres egenskaber af troværdig og generel viden muliggør koordination og styring (Porter 1995).

De seneste årtiers styringstendenser såsom evidensbasering, evaluering, New Public Management og resultatmålinger gør i høj grad brug af de muligheder for sammenligning mellem enheder over tid, som anvendelse af kvantitativ data tillader (Bhatti et al. 2006; Dahler-Larsen & Kristiansen 2015; Krogstrup 2011). Store informations-

[1] Jeg takker gode kolleager i Institut for Statskundskabs forvaltningsgruppe, dette temanummers redaktørduo Anne Mette Møller og Signe Blaabjerg Christoffersen, den anonyme reviewer samt min ph.d.-vejleder Peter Dahler-Larsen for konstruktive kommentarer og forslag til denne artikel.

mængder kan opsummeres med tilsyneladende entydige tal. Der finder en 'usikkerhedsabsorption' sted, hvor: "…inferences are drawn from a body of evidence, and the inferences instead of the evidence itself, are then communicated." (March & Simon 1958, 165). Dette er nyttigt i en politisk kontekst, hvor kvantitative vurderinger kan udgøre en form for 'forsikring' mod fremtidige overraskelser, når ansvaret deles mellem beslutningstageren og de procedurer, der genererede beslutningens kvantitative grundlag (Lindeberg 2015). Tals entydige og sikre karakter gør dem dermed til en særlig politisk effektiv vidensform. Der er "strength in numbers", som videnskabshistorikeren Theodore Porter udtrykker det (1994, 404).

Det er dog ikke en selvfølgelighed, at tal indtager en sådan styrkeposition. Historien er fyldt med statistiker, der aldrig blev troværdige, og indikatorer, der ikke kom til at tjene deres formål (Asdal 2011). Tals objektivitet tages ikke altid for givet i den offentlige debat, og 'validitet' eller andre metodiske kvaliteter kan ikke i sig selv forklare, hvorfor bestemte studier bliver set som politisk relevante (Contandriopoulos et al. 2010; Weiss & Bucuvalas 1980). Det er ikke sådan, at nogle målinger bare er 'bedre' og derfor ikke bliver kritiseret. I forlængelse af dette argumenterer jeg for, at tals politiske liv almindeligvis udfolder sig i situationer, hvor de hverken bliver taget ukritisk for givet eller afvist fuldstændigt. Noget bliver relevant, når det bruges til noget. Men brug øger samtidig risikoen for kontroverser. Tal, der virker, virker ofte på en måde, nogen er imod. Jeg udforsker derfor i denne artikel en ide om, at tals[2] politik udfolder sig som et dilemma mellem styringsmæssig og politisk relevans; og de kontroverser, denne relevans giver anledning til.

De nationale test i den danske folkeskole fungerer som mit gennemgående eksempel. "Vi lever i et interessant øjeblik i testens historie", skrev evalueringsforskeren Peter Dahler-Larsen (2012, 14) i sommeren 2012 med henvisning til en netop indledt evaluering af testene. Tiderne er ikke blevet mindre interessante siden. De nationale test var i 2016 genstand for intens medieopmærksomhed, gennemgik to høringer i Folketingets undervisningsudvalg – og blev forlangt nedlagt af flere politiske partier. Samtidigt har testene fra begyndelsen haft forskellige formål, der ofte er blevet fremhævet som delvist modstridende. De skal anvendes af lærere som pædagogiske redskaber i undervisningen; de skal fungere som styrings- og informationssystemer på både kommunalt og ministerielt niveau; og de er med folkeskolereformen i 2014 blevet officielle succeskriterier for skolen, idet tre af de såkaldte 'måltal' opgøres via testresultaterne.

Testene er oplagte eksempler på kontroversiel relevans. Deres præcise rolle er et åbent politisk spørgsmål, samtidig med at de løbende debatteres og kritiseres. Tallenes politik er i centrum for min undersøgelse, hvorfor jeg ikke forholder mig til, om testens modstandere eller tilhængere har ret. Jeg forfølger ikke min case ud i klasselokalet eller

[2] Denne artikel handler om 'tal' generelt. Der er en stor litteratur om de særlige karakteristika ved officielle statistikker, ranglister, indikatorer, resultatmålinger etc. I det følgende lægger jeg ikke vægt på deres forskellighed og antager derfor, at mine påstande gør sig gældende på tværs af de forskellige udtryk, et tal kan have. Det ville være interessant at undersøge, om forskellige taltyper opfører sig forskelligt mht. kontroversiel relevans, men dette rækker ude over rammerne af den herværende analyse.

kommunale skoleforvaltninger for at se, om testene faktisk påvirker undervisning og administration, ligesom jeg heller ikke foretager statistiske efterprøvninger af diverse metodiske krav til uddannelsesmæssige test. Fokus er på at følge de argumenter, der fremsættes for og imod testen, hvorfor selve det, at nogen stiller spørgsmålstegn ved testenes reliabilitet eller kritiserer dem for af lede til 'teaching to the test', er interessant for min analyse.

Jeg vil ikke i denne artikel præsentere en veludbygget teori med præcise og færdigudviklede definitioner, men snarere introducere kontroversiel relevans som 'sensitizing concepts' (Bowen 2006), der giver et bud på, hvad man skal fokusere på i analyser af de politiske og metodiske debatter, der former et tal. Først vil jeg præsentere mit grundlæggende syn på tal som politiske og konstruerede samt introducere de nationale test som empirisk eksempel. Derefter vil hovedparten af artiklen udforske begrebsparret *kontrovers* og *relevans*: først hver for sig og siden som en diskussion af de dynamikker, der opstår mellem dem.

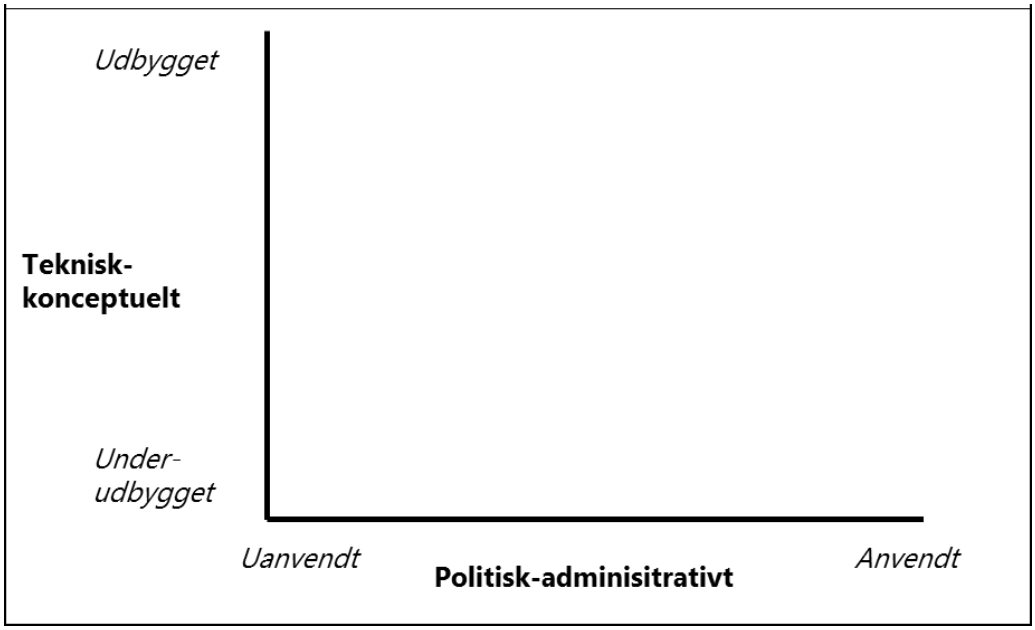
Tal som politisk og metodisk konstruktion

Statistik og indikatorer kommer et sted fra. De er resultater af en social proces, hvor nogen har besluttet at sætte tal på et fænomen og derefter udviklet eller lånt en fremgangsmåde til at udføre målingen. Utallige studier har undersøgt, hvordan styring og politikudvikling baseret på kvantitative data påvirker og forandrer de målte objekter (fx Bevan & Hood 2006; Espeland & Sauder 2007; Smith 1995; Van Thiel & Leeuw 2002). Det er altså ikke ligegyldigt hvilke tal, der danner grundlaget for politiske debatter, beslutningstagning og administration. Som vi skal se, knytter meget af debatten om nationale test som kvalitetskriterium for undervisning an til uenigheder om, hvordan konkrete testspørgsmål udformes, og hvilken betydning dette får for de overordnede testresultater. Selve målingen er genstand for politik. Fra et skeptisk udgangspunkt kan man derfor ikke bruge tallene til at forstå de målte objekter. Hvis der er politiske valg i en måling, hvordan kan resultatet så beskrive virkeligheden?

Samme kritik kan dog rettes mod sproget generelt. Diskursteori viser, hvordan der er politiske konnotationer i simple beskrivelser, da enhver skildring er påvirket af politiske og samfundsmæssige forhold. Men vi opgiver ikke af den grund at forstå verden gennem sproget. For at håndtere tal, der både er politiske i deres ophav og i deres virkning, anlægger jeg et konstruktivistisk perspektiv, hvor konstruktivisme (i denne sammenhæng) blot betyder, at tal er konstruerede. Den måde, de er udfærdigede på, kan debatteres – og bliver det ofte. De er potentielt kontroversielle, idet der ikke er en enkelt entydigt overlegen måde at måle et fænomen på. Men de er heller ikke grebet ud af den blå luft. Der er begrænsninger for, hvad man kan slippe af sted med. Konstruktivisme handler om den samfundsmæssige konstruktion af vores fælles virkelighed – ikke om

den "samfundsmæssige konstruktion af fiktioner" (Dahler-Larsen 2013, 33f). Verden er nu engang fuld af konstruktioner – og det gælder også de tal, vi beskriver den med.

Undersøgelsen af tals kontroversielle relevans handler derfor nok om politik – men også om videnskab og metode. I en analyse af metoder til opgørelse af bruttonationalproduktet opererer Wouter van Dooren (2009) derfor med to dimensioner i vurderingen af et succesfuldt mål. Målet skal være teknisk-konceptuelt veludbygget, så det kan fungere som en god indikator for den underliggende kompleksitet, men det skal samtidig kunne sætte sig igennem i en politisk-administrativ proces, hvor producenter og brugere af tal forhandler om forskellige måls relative nyttighed.



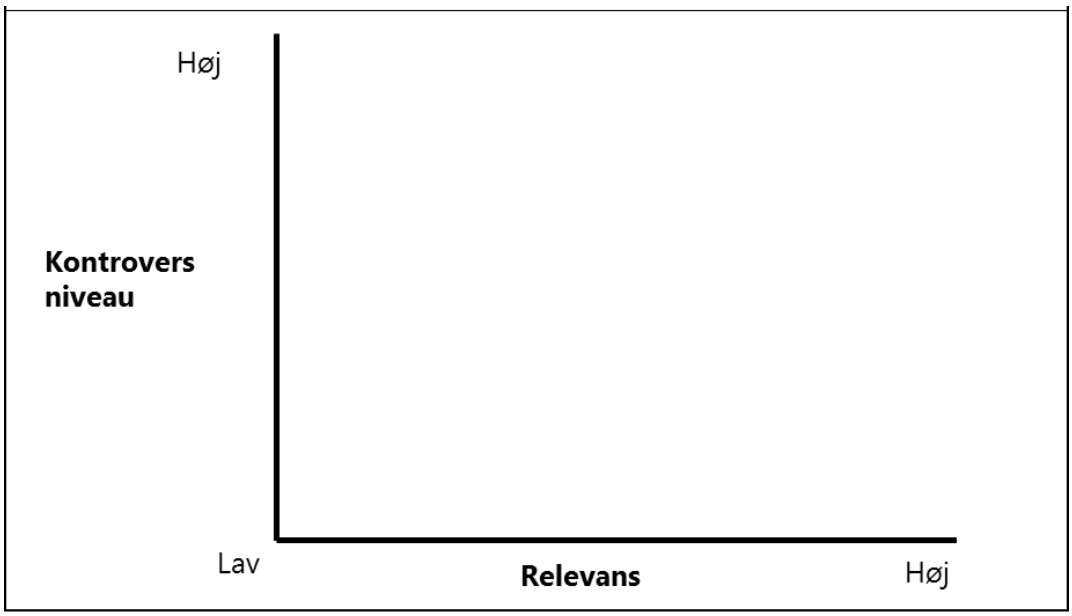
Figur 1. Målingens dimensioner ifølge Van Dooren 2009

Denne fremstillingsform antyder, at der primært er politik i den ene dimension. Først kan man uproblematisk og objektivt vurdere, hvorvidt et mål lever op til diverse metodiske standarder for validitet og reliabilitet og dermed afgøre dets kvalitet som repræsentation af det målte objekt. Dernæst indgår målene – gode såvel som dårlige – i en social proces, hvor deres nyttighed fastsættes af politiske prioriteter. Kvalitet er 'objektivt og givet' – anvendelse er 'politisk og konstrueret'. Betragter man imidlertid tal som sociale fænomener, bliver denne sondring sværere at opretholde, da også den tekniske dimension indeholder politiske og samfundsmæssige spørgsmål. Hvilke metoder opfattes som legitime og hensigtsmæssige? Hvordan afvejes modstridende kvalitetskrav? Hvilke aspekter ved et komplekst objekt skal inddrages i målingen?

En lang tradition med udgangspunkt i særligt Bruno Latours tilgang til videnskabsstudier har undersøgt, hvordan videnskabelige udsagn kommer til at fremstå som sande og objektive, når kontroverser om denne slags spørgsmål dør ud, og udsagnet bliver taget for givet som en uproblematisk beskrivelse (fx Bowker & Star 1999; Latour

1987; 1999). Som tidligere antydet er det ikke nogen selvfølgelighed, at et tal opnår denne position. I et klassisk studie af anvendelse af forskning identificerer Carol Weiss og Michael Bucuvalas (1980) to dimensioner i beslutningstageres vurdering af et givent studie. *Sandhedstest* relaterer sig især til forskningens kvalitet og resultaternes overensstemmelser med det forventede. *Brugbarhedstesten* knytter sig til forskningens handlingsrettethed og beslutningsrelevans og den udfordring, resultaterne udgør for eksisterende policy eller praksis. To af deres konklusioner er særligt relevante for mit argument: *for det første* er forskningens 'kvalitet' især vigtig i kontekster af omtvistede politikspørgsmål og kontroverser; *for det andet* er viden, der udfordrer et områdes status-quo, meget anvendelig. De politiske muligheder, noget giver anledning til, spiller også ind på vurderingen af dets kvalitet.

Simone Ledermann (2012) bygger videre på disse ideer ved at vise, hvordan anvendelse af evalueringskonklusioner afhænger af deres nyhedsværdi og kvalitet: I situationer hvor en høj nyhedsværdi fremmer anvendelse, er der også høje kvalitetskrav til evalueringen. Man skal sige noget nyt for at være nyttig, men det nye er også kontroversielt, hvorfor man skal have sin rygdækning i orden. I 2000'erne blev uddannelsessystemerne i mange lande ramt af et såkaldt 'PISA-chok[3], hvor deres selvforståelse blev udfordret af en ranglisteplacering, der var langt under det forventede. Mange steder gav dette anledning til selvransagelse og reform, hvorved PISA viste sig som uddannelsesmæssigt relevant (Grek 2009). Men PISA er samtidigt ekstremt kontroversiel. Aktører, der er uenige i dens relative vurdering af uddannelsessystemer, antaster testen og sætter spørgsmålstegn ved dens metodiske forudsætninger og politiske virkninger.



Figur 2. Kontroversiel relevans

[3] PISA (Programme for International Student Assessement) udføres hvert tredje år af den internationale organisation OECD (Organisation for Economic Co-operation and Development) som en sammenligning af nationale grundskoleuddannelser.

Jeg erstatter derfor Van Doorens 'teknisk-konceptuelle' akse med spørgsmålet om et tals *kontroversialitet*. I hvor høj grad giver målingen anledning til disputter og slagsmål? Hvor mange og hvilke aktører stiller spørgsmålstegn ved det billede, en bestemt statistik tegner? Hvor antastet er målingens kvalitet? Tilsvarende erstatter jeg den anden akse med tallets *relevans*. Hvilke argumenter og politiske muligheder åbner det op for? Anvendes det til at træffe beslutninger eller fordele resurser? Hvilke praksisser bliver bedømt på baggrunden af målingen?

Hvor i dette skema et bestemt tal befinder sig, er et empirisk spørgsmål. Det er et områdes aktører, der gør noget relevant og kontroversielt. Hvis de sætter spørgsmålstegn ved tallet, kritiserer dets grundlag eller begræder dets betydning, er det kontroversielt. Hvis de anvender det i deres beslutninger, beskriver de målte objekter med udgangspunkt i det, eller simpelthen ikke kan forholde sig til et emne udenom dets målinger, er det relevant. Min fremgangsmåde forudsætter dermed ikke, at man fremsætter normative domme over, hvorvidt et tal *burde* være relevant eller kontroversielt i en given sammenhæng, men blot at man registrerer, om tallets støtter og modstandere forholder sig til det på en måde, man analytisk set vil karakterisere som udtryk for kontroversiel relevans.

Nationale test i den danske folkeskole

Danske folkeskoleelever har siden 2010 i løbet af deres skoletid taget ti test, der afprøver kundskaber og færdigheder i udvalgte fag.[4] Disse test udgør et interessant tilfælde af kontroversiel relevans. I min karakteristik af de debatter og kontroverser, der har ledsaget testene siden før deres vedtagelse, trækker jeg dels på eksisterende akademiske studier, dels på de officielle dokumenter, rapporter og evalueringer, der indgår i testens udvikling og formidling. Endeligt er det generelle narrativ om testens ophav og forløb blevet undersøgt via en stor mængede avis- og magasinartikler fra den relevante periode.

Indførelsen af de nationale test var i høj grad påvirket af udefrakommende dynamikker (Krejsler et al. 2014). En vigtig anledning var en OECD-undersøgelse fra maj 2004, der udråbte 'evalueringskultur' til den enkeltforandring, der kunne medføre størst forbedringer af den danske folkeskole (Ekholm et al. 2004, 129). Det præcise indhold af begrebet evalueringskultur stod imidlertid dengang som nu mindre klart (Dahler-Larsen 2006; Pors 2009). Selvom internationale anbefalinger og rapporter således var centrale, blev de politiske diskussioner ofte ført med reference til danske traditioner for primært at anvende andre evalueringsformer (Andersen 2007; Andersen et al. 2009). En række skolepolitiske aktører udtrykte i deres vision for en dansk evalueringskultur skepsis mht. obligatoriske test (KL et al. 2004), og da den daværende VK-regering først luftede

[4] Der er fire test i dansk/læsning (2., 4., 6. og 8. klassetrin), to i matematik (3. og 6. klassetrin) samt en i hhv. engelsk (7. klassetrin), geografi, biologi og fysik/kemi (8. klassetrin).

ideen, nedlagde socialdemokraterne veto. Med reference til behovet for at 'gøre noget' oven på endnu en skuffende PISA-præstation foreslog regeringen alligevel i december 2004 at indføre nationale test, og over det næste år lykkedes det at komme frem til et kompromis, så en ny folkeskolelov kunne vedtages i foråret 2006 med støtte fra Dansk Folkeparti og Socialdemokratiet.

Vibeke Normann Andersen (2007) har mht. en tidligere lov om offentliggørelse af gennemsnit fra folkeskolens afgangsprøve påpeget, at 'åbenhed og gennemsigtighed' ofte fremstilles som godt i sig selv, hvorimod det er mindre klart, hvordan resultatmålinger konkret skal gavne elevernes læring. De nationale test har løbende været genstand for samme kritik, fx påpegede den daværende radikale leder Marianne Jelved i 2005, at grise ikke bliver "federe af at blive vejet" (DR 2015). Støtter af testene har derfor fra begyndelsen skullet demonstrere relevans. Hvad er det, testresultaterne kan bruges til? Hvordan bliver de et 'pædagogisk værktøj'? At gøre testene relevant for nogen og noget har været centrale elementer i deres politiske udfoldelse.

Relevans – 'at betyde noget for nogen'

Slår man *relevans* op i en ordbog, defineres det som "betydning eller vigtighed i den givne sammenhæng" (Ordnet.dk 2017a). Tal er ikke relevante abstrakt set. De er relevante *for* noget: et bestemt formål, en bestemt aktør, en særlig handling. Et supplerende opslag for *relevant* uddyber: "Det er nødvendigt at aktualisere, modernisere, gøre klassikerne relevante for os her og nu (Ordnet.dk 2017b)." Noget *er* ikke bare relevant, det *gøres* eller *bliver* det. Relevans skal ses som et lokalt, kontekstuelt fænomen, og som resultatet af en proces, hvor succes ikke er garanteret en gang for alle, men løbende skal opretholdes.

Selvom det er et spørgsmåls aktører, der gør et tal relevant ved bl.a. at knytte forbindelser til det, indsætte det i argumenter eller basere handlinger på det, har vi som analytikere alligevel brug for en udfoldelse af begrebet, der tillader identifikation af de empiriske observationer, der tæller som forsøg på relevansgørelse. Det følgende skal ikke ses som en almengyldig typologi, men som et analyseredskab eller analyseramme, der tillader mig at pege på en række forskellige måder, hvorpå et tal kan fremstå relevant.

Relevans har to primære dimensioner: hvordan er noget relevant (anvendelse), og hvor og for hvem er det relevant (forum). Den *første* dimension kan belyses med afsæt i en langvarig diskussion i evalueringslitteraturen af den måde, evalueringsresultater faktisk anvendes til at træffe beslutninger. Udgangspunktet for diskussionen var en undren over, at programmer og initiativer så ud til at blive videreført eller nedlagt uafhængigt af evalueringens anbefalinger. Carol Weiss (1979) forklarer dette med en for 'snæver' forståelse af anvendelse: evalueringer bliver faktisk brugt – bare ikke altid på den forventede måde. Hun foreslår derfor syv 'modeller' for anvendelse, der ofte bliver

reduceret til tre: instrumentel, taktisk/symbolsk og konceptuel (Alkin & King 2016). Evaluering bruges instrumentelt, når resultater direkte lægges til grund for eller informerer handling. Taktisk/symbolsk brug er anvendelsen af evalueringer som legitimering af en beslutning eller som led i et argument. Konceptuel anvendelse dækker over den måde, evalueringsresultater over tid former den evaluerede praksis eller det vurderede fænomen.

	Instrumentel	**Taktisk/symbolsk**	**Konceptuel**
Offentlig	Tallet refereres i den offentlige debat: *En rangliste baseret på testresultater omtales. Landsgennemsnittet diskuteres.*	Tallet bruges til at konstruere bestemte argumenter: *"Disse skoler har dårlige resultater, derfor skal vi gøre x, y, z"*	Diskursiv forståelse af det målte fænomen er påvirket af tallet: *God undervisning italesættes som synonymt med høje testresultater.*
Administrativ	Bruges til at træffe beslutninger: *Inspektioner/tilsyn på skoler, hvor resultaterne er for lave. Ressourcefordeling baseret på resultater.*	Retfærdiggør administrative beslutninger: *En skolelukning motiveres med dårlige testresultater*	Forståelsen af fænomenet i administrativ praksis er påvirket af tallet. *Høje testresultater er generelt administrativt succeskriterium i en kommune.*
Faglige	Tallet motiverer/informerer praksis: *Testene er et pædagogisk redskab til at identificere fokusområder og sætte prioriteter i undervisningen.*	Tallet bruges som retfærdiggørelse af faglige vurderinger: *Beslutninger om særlige indsatser overfor en elev motiveres med testresultater.*	Forståelse af fænomenet i faglig praksis er påvirket af tallet: *Undervisningen fokuserer på testens opgave- og spørgsmålstyper.*

Tabel 1. Dimensioner i relevans

Et studie af 'fora' for ansvarlighed (accountability) (Willems & van Dooren 2012) giver inspiration til at udfolde den *anden* dimension som 'offentlige, administrative og faglige fora for relevans'. Det *offentlige* forum dækker over den offentlige debat om emnet. Relevans betyder her, at debatdeltagere implicit eller eksplicit refererer til tallet. I det *administrative* forum bliver et tal relevant, hvis det anvendes i forvaltningsmæssigt arbejde af juridisk eller bureaukratisk karakter. Endeligt udgør det daglige arbejde i organisationens yderste lag et *fagligt* forum, hvor et tal er relevant, hvis det har betydning for eller anvendes af praktikere eller frontlinjemedarbejdere. Tabel 1 (se forrige side) sammenkobler de to dimensioner som illustration af mulige relevansformer.

Som det fremgår, er der mange måder, de nationale test kunne blive relevante for dansk uddannelses- og skolepolitik. Men ikke alle er blevet til virkelighed. Tabel 1 er derfor ikke en oversigt over, hvad der blev realiseret, men en påmindelse om de forskellige udtryk, et tals relevans kan tage. Det er så et empirisk spørgsmål, hvilken relevans et tals støtter faktisk arbejder for.

For de nationale test var det tidligt et centralt argument, at en styrkelse af det faglige niveau især ville komme de svagere elever til gavn. Undervisningsministeriet lægger i sin kommunikation vægt på, at testresultaterne er en nem og systematisk måde, lærere kan danne sig overblik over elevernes niveau (fx Ravn 2008), hvilket især skal bidrage til at rette opmærksomheden mod de elever og skoler, der har det svært. Dette er et slagkraftigt argument, og daværende statsminister Anders Fogh Rasmussen kunne retorisk fremstille sig som chokeret over: "at Socialdemokraterne i ramme alvor vil modsætte sig initiativer [der] er til størst gavn for de elever, der kommer med den svageste baggrund" (Carlsen 2004). Da testene bliver behandlet i Folketinget er det derfor ifølge Socialdemokratiets ordfører afgørende for parties opbakning, at testene kan fungere som et pædagogisk redskab (Folketinget 2006a). Desuden er denne relevansform et forsonende træk for ellers kritiske aktører som Danmarks Lærerforening (DLF) og Radikale Venstre.

At fremme den pædagogiske relevans er blevet en bunden opgave i testens implementering. Undervisningsministeriet lancerer derfor en pjece med eksempler på en lærerfaglighed, der baserer elevvurderinger på testresultater (Undervisningsministeriet 2011). Man forsøger så at sige at 'skabe sine egne brugere' (Young 2006) ved at påvirke de praktikere, der skal opfatte testene som fagligt relevante. Sådanne relevansgørelser er dog langt fra garanteret succes. Hvad enten de er politiske udmeldinger eller mere forvaltningsmæssigt udfoldende implementeringsstrategier, er de kun forsøg, der er afhængige af støtte – og sårbare overfor modstand. En relevansgørelses videre skæbne afhænger dermed bl.a. af de kontroverser, den giver anledning til.

Kontroverser: 'at virke på en måde, man kan være imod'

Kontroverser er episoder, hvor en målings tekniske og politiske valg bliver synlige. De genåbner et velfungerende og naturaliseret tal, og gør det derved muligt at studere de processer, der får en bestemt beskrivelse af verden til at fremstå naturlig (Latour 1987). At fokusere på kontroverser er dog i mit argument mere end et metodisk princip, idet jeg undersøger idéen om, at tal i offentlig forvaltning og politik kan fungere, selvom de er genstand for åbne kontroverser.

Kontroverser udfolder sig ikke nødvendigvis offentligt. Dataindsamling kan være genstand for træghed og diffus modstand, hvis de faggrupper, der skal arbejde med en statistik, ikke støtter op om den (Pors 2016). I denne artikels undersøgelse af de nationale tests begrænser jeg mig dog overvejende til offentlige kontroverser. En første læsning af debatten om testenes styrker og svagheder kan tilskrive aktørerne interessebaserede positioner: DLF er imod, idet lærere ikke vil kontrolleres, og Undervisningsministeriet embedsmænd er for, da de vil have 'styr' på en uregerlig folkeskole. Talmodstandere opdyrker kontroverser, som når det (delvis DLF- kontrollerede) fagblad

Folkeskolen[5] opfordrer lærere til at skrive ind med eksempler på elevers (dårlige) oplevelse af testsituationen (Ravn 2010).

Selvom motiverne til at stille bestemte spørgsmål kan fremstå som politiske, vil ammunitionen i kontroverser alligevel ofte være metodiske og 'videnskabelige' udsagn om validitet eller reliabilitet. For at få betydning, skal kritikpunkter dog 'konstrueres' som signifikante problemer. Selv den mest metodisk velfunderede kritik bliver irrelevant, hvis den ikke vækker genklang, og et tals støtter derfor omkostningsfrit kan ignorere den. Noget skal opfattes som problematisk, før det faktisk bliver en udfordring (Blyth 2013). Problemer eller styrker ved et tal er dog ikke noget, den enkelte kan finde på. Som tidligere fremhævet er sociale konstruktioner ikke grebet ud af den blå luft, men afhængigt af intersubjektive kriterier for gode målinger. Med Weiss og Bucuvalas' udtryk er der en videnskabelig kanon, man kun delvist kan se bort fra (1980, 307). Men da denne kanon netop er konstrueret, er der plads til uenigheder om, hvilke metoder der tæller, og hvilke konsekvenser der er acceptable.

I princippet er der ikke nogen forhåndsbegrænsninger på, hvad der kan blive kontroversielt. Alt, hvad der i praksis diskuteres som kritik af en måling, kan være centrum for en kontrovers. For de nationale test falder krikken overvejende i fire grupper. For det første kan *kvantificeringen som sådan* erklæres problematisk. Selve det at beskrive noget med tal fremstilles som en urimelig reduktion af komplekse og unikke emner, der ikke bør behandles på numerisk form (Espeland & Stevens 1998). Undervisning og læring bliver af skolefolk ofte beskrevet som brede fænomener, hvis formål ikke kan fanges af en enkelt test. Dorte Lange fra DLF kritiserer fx i 2008, at testen ikke kan måle, om eleverne lærer "at skrive digte, der får hårene til at rejse sig" (Ravn 2008).

Matematiklærerforeningen påpeger tilsvarende i et høringssvar, at testene ikke opfanger store dele af fagets formål (Folketinget 2006a), hvilket er i tråd med en generel kritik om, at opgavetyperne umuliggør målinger af de mere kreative og problemløsende kompetencer. Testens tilhængere svarer i den forbindelse ofte, at det heller aldrig har været meningen, at testen skal dække alt (fx Nørby 2016). Det er ikke i sig selv et problem, at der er elementer af elevernes læring, der ikke måles – så længe undervisningen ikke tilrettelægges udelukkende med testen in mente. Læring bør ikke reduceres til testscorer, men dette udelukker ikke måling af visse aspekter af læring.

En anden, mere udbredt, type af kritik går på *testens udførelse*. Det er i princippet acceptabelt at sætte tal på læring, men den konkrete måling givet et forfejlet eller upålideligt billede. En række af testopgaverne er blevet kaldt meningsløse – med et læsespørgsmål til fjerde klasse om betydningen af 'inkommensurabel' som et berygtet eksempel (Andersen 2016). Som svar forsøger Undervisningsministeriet at øge tilliden til opgaverne ved bl.a. at invitere fagbladet Folkesolen til at følge udviklingen af testop-

[5] *Folkeskolen* er som kilde en speciel størrelse, idet bladet er en "helt central arena" for skoledebat (Hermann 2007, 22), og den publikation der mest vedholdende, detaljeret og kompetent beskæftiger sig med de nationale test – men samtidig bærer den redaktionelle linje (til tider) præg af bladets rolle som DLF's medlemsblad.

gaver og ved at være generelt åbne om processen (Folkeskolen, 4/6/14; Undervisningsministeriet 2016).

Testene har løbende været udsat for debatter om deres pålidelighed. Disse problemer brød ud i lys lue i efteråret 2015, hvor en skole berettede om betydelige udsving i elevernes resultater, da man med kort interval lod dem gentage samme test. I medierne blev der fra flere sider stillet spørgsmålstegn ved, om resultaterne overhovedet afspejler elevernes faglige kunnen. For at imødegå kritikken satte undervisningsministeren ministeriet til at undersøge yderligere. I et offentligt samråd i Folketingets undervisningsudvalg karakteriserede ministeren udsvingene som acceptable (Nørby 2016), men professor Jeppe Bundsgaard var ikke overbevist og fortsatte offentligt med at stille spørgsmålstegn ved testens pålidelighed (Ravn 2016).

Hvor de foregående kontroverser går på selve tallet og dets tekniske infrastruktur, kan også en *bestemt relevansform* være genstand for kritik. Offentliggørelse af testresultater giver mange anvendelsesmuligheder. Testresultater kan hjælpe forældre til at træffe informerede beslutninger under 'frit-valg' politiker, og testscorer kan fungere som referencer i offentlige diskussioner om uddannelseskvalitet. Det var imidlertid tidligt et emne, at testene ikke måtte medføre rangordning af skoler, hvorfor offentliggørelse af resultater efter socialdemokraitsk ønske bliver udelukket i den oprindelige aftale om testene (Gustafsson 2012). Det gav derfor anledning til heftig kritik fra forligskredsen, da regeringen i efteråret 2005 forsigtigt antydede planer om delvis offentliggørelse (Folkeskolen 2005).

At tale for offentliggørelse er derfor en sikker måde at ryste både det politiske forlig bag testen og store dele af den danske uddannelsesverden. Det er imidlertid præcist, hvad daværende statsminister Lars Løkke Rasmussen gjorde i begyndelsen af 2010, da han lagde op til et opgør mod offentliggørelsesforbuddet, da man ikke "i et åbent, demokratisk samfund" bør hemmeligholde resultater (Olsen 2010). Forargelsen var massiv. Danmarks Lærerforening frygtede konkurrence mellem skoler, at forældre ville få et forvrænget billede af skolen, og at undervisningen indsnævres til de testede emner. Skolelederforeningen og forældreorganisationen Skole og Samfund var tilsvarende kritiske og en række eksperter skød idéen ned. Sven Erik Nordenbo, der netop havde gennemførte en undersøgelse af testenes pædagogiske potentiale, erklærede sig decideret "forbitret" (Børsting & Fuglsang 2010).

Endeligt kan selve de *negative konsekvenser* af en bestemt relevansform vække kritik. Der er selvfølgelig et vis overlap med det foregående, idet modstanden mod en relevans ofte skyldes frygten for dens konsekvenser. Offentliggørelses-kontroversen kunne fx være opstillet som en kritik af de skadelige virkninger af konkurrence i skolerne, og således have sat konsekvenser fremfor relevansformen i centrum. En faktisk forekommende konsekvenskritik går på den 'teaching to the test', som testene siges at give anledning til. I efteråret 2016 udgav skoleforsker Jeppe Bundsgaard i forbindelse med en eksperthøring i Folketinget en rapport om testens pædagogiske anvendelser. Han påpegede bl.a., at mange lærere er begyndt at undervise direkte i testens opgavetyper, hvilket er netop den 'teaching to the test', alle er enige om skal undgås (Bundsgaard

& Puck 2016). Dette fik Alternativets Carolina Maier til at 'dumpe' testen og foreslå den afskaffet (Folketinget 2017; Maier 2016).

Kontroversiel relevans: slagsmål om en test

Indtil nu har jeg overvejende behandlet kontroversialitet og relevans hver for sig. Da jeg betragter deres indbyrdes dynamikker som empiriske spørgsmål, der ikke udspiller sig ens i alle tilfælde, vil jeg i stedet for yderligere konceptuelle afklaringer foretage en række nedslag i de nationale tests politik for derigennem at kaste lys over samspillet mellem relevansformer og kontroverser.

Den kontroversielle offentliggørelse

Som vi har set, blev offentliggørelse af testresultater udelukket som del af det oprindelige forlig. Da daværende statsminister Løkke Rasmussen genåbnede debatten i 2010, førte det til kraftige reaktioner, som han forsøgte at imødegå ved at præcisere hensigten: "Regeringen går ind for åbenhed og gennemsigtighed, men vi går ikke ind for, at skolerne skal i en gabestok." (Olsen 2010). Over det næste år forsatte diskussionen, indtil tingene i foråret 2011 spidsede til, hvis en ændring af folkeskoleloven skulle vedtages inden det kommende valg. Regeringen bøjede af og erklærede sig som principielle tilhængere af offentliggørelse, men anerkendte at forslaget var for kontroversielt. Offentliggørelse af testresultaterne har siden da ligget dødt som politisk emne.

Ved at udelukke offentliggørelse af resultater på skole- eller kommuneniveau umuliggøres oplagte relevansformer. Offentlige testresultater kunne have været en måde at identificere eksempler på uddannelseskvalitet og derved placeret testen centralt i diskussionen af, hvad der konstituerer gode skoler og god undervisning. I den nuværende udformning, hvor resultaterne kun offentliggøres på landsplan, kan debatten udelukkende handle om den nationale udvikling over tid, hvilket vækker mindre opmærksomhed end den rangering af kommuner eller skoler, man allerede i dag kan foretage med udgangspunkt i karaktergennemsnit fra folkeskolens afgangsprøver. Der er derfor færre grunde til at henvise til de nationale test i uddannelsespolitiske debatter. En potentiel relevans bliver altså valgt fra pga. dens kontroversialitet, hvilket i figur 2 ville være en bevægelse nedad og mod venstre.

Et genstridigt pædagogisk værktøj

Pædagogisk anvendelse var oprindeligt det primære argument for indførelsen af nationale test. Evalueringskulturen skulle styrkes via tilvejebringelse af 'objektiv' information om elevernes resultater, hvilket kunne føre til refleksion og kvalitetsforbedringer på de enkelte skoler. Dette var fra starten udfordret i praksis, da de fleste lærere stoler mere på deres erfaringsmæssige hverdagsviden. En Rambøll-evaluering udkom fx i 2013 med konklusionen, at testene overvejende anvendes til at opsummere resultater og derfor ikke tjener deres primære formål som pædagogiske redskaber (Rambøll 2013). Desuden sættes der, som vi har set, løbende spørgsmålstegn ved testenes pålidelighed som beskrivelse af elevernes faglige niveau.

Som faglig relevans kan pædagogisk brug kun vanskeligt tvinges igennem, da den fordrer opbakning fra de praktikere, der skal anvende den. Når lærerens egen vurdering af en elev og dennes testscore afviger, bliver det er spørgsmål om de to vidensformers relative legitimitet (Kousholt 2015). Testviden kan ifølge Kousholt tilskrives høj legitimitet fra deres officielle og obligatoriske status, men testene er stadig udfordrede af kontroverser om upålidelighed og meningsløse testopgaver, hvorfor læreren kan tænkes at lægge mindre vægt på dem. Selvom mange test-tilhængere og ministerielle embedsmænd rutinemæssigt omtaler testen som pædagogisk relevant, er det en alvorlig udfordring, at mange lærere er skeptiske overfor dens nyttighed. Socialdemokraten Pernille Rosenkrantz-Theil udtalte i 2016 endda i relation til dette, at hun heller aldrig havde gjort sig "nogen illusioner om, at de nationale test skulle være et pædagogisk redskab" (Riise 2016), hvilket tidligere ville have været en uhørt kommentar. Vi ser, at en ellers central relevansform over tid er blevet undermineret af de kontroverser, den giver anledning til. Hvis testene skal undgå at synke ned ad relevansaksen i figur 2, skal de pædagogiske perspektiver altså erstattes med noget andet.

En mere styrbar relevans

Med folkeskolereformen i 2014 begynder der at ske noget med de nationale tests styringsmæssige anvendelse. Reformen opstiller fire mål for elevernes faglige udvikling, hvoraf de tre baserer sig på resultater fra de nationale test i dansk og matematik. En succesfuld implementering af folkeskolereformen bliver dermed delvist gjort til et spørgsmål om højere testscore. Fra efteråret 2015 udgiver undervisningsministeriet en årlig statusredegørelse, der kortlægger fremskridt mod de nationale mål. Hvor man i den tidligere nationale præstationsprofil blot kunne følge udviklingen i testscore, knyttes resultaterne nu direkte til den politiske målsætning om, at alle elever skal blive 'så dygtige, de kan'. Testen opnår altså større potentiel relevans som reference i den uddannelsespolitiske debat, hvor fravær af fremskridt kan udnyttes til at kritisere folkeskolereformen og forligskredsens skolepolitik.

Derudover indgår testene nu direkte i det ministerielle kvalitetstilsyn, hvor testscorer på de tre nationale måltal indgår i de indikatorer, der danner grundlag for identifikation af skoler og kommuner med utilstrækkelige resultater. På grundlag af en helhedsvurdering af skolens resultater beslutter ministeriet efter denne screening, om man vil igangsætte yderligere tiltag. Testene bliver således en faktor, kommunale skolemyndigheder er nødt til at forholde sig til, idet længerevarende fravær af fremskridt mod de nationale måltal kan medføre påtale fra Undervisningsministeriet. Da kommunerne samtidigt er forpligtiget til at behandle de nationale måltal i årlig kvalitetsrapporter, opnår testene administrativ relevans som basis for ministerielle og kommunale beslutninger.[6]

Dette er ikke ukontroversielt. Modstandere af testen har gentagene gange påpeget, at det kontrolfokus, der ligger bag de kommunale kvalitetsrapporter og det ministerielle tilsyn, risikerer at introducere sammenligning af bagvejen, idet det er offentligt tilgængeligt, hvilke kommuner der ikke lever op til målsætningerne (Folketinget 2006b; Hellisen 2014). Forskellen på testen som led i en styringsrelation og som pædagogisk værktøj er således ikke, at kun sidstnævnte er kontroversiel, men snarere at administrativ relevans i højere grad kan fastsættes fra centralt hold,[7] idet den ikke i samme grad er afhængig af praktikernes opbakning. Ikke alle relevansformer er altså lige sårbare overfor alle kontroverser, hvilket komplicerer figur 2's billede af en række positioner, der alle er udstyret med et bestemt niveau af kontroversialitet og relevans. En given kontrovers' skadelighed afhænger af, hvilken relevans man fokuserer på.

Hønen eller ægget

Taget hver for sig kan forløbet omkring offentliggørelse og pædagogisk anvendelse antyde, at tidsfølgen altid går fra relevans til kontrovers, således at relevansgørelsens succes afhænger af, om de kontroverser, den giver anledning til, kan håndteres eller ej. Selvom denne tolkning har meget på sig, er det ikke den eneste måde, de to faktorer kan spille sammen på. I visse tilfælde kan en ny relevansform også være svar på en kontrovers. Socialdemokraten Pernille Rosenkrantz-Theil lægger, som vi så tidligere, ikke

[6] Undersøgelser, der går tættere på den kommunale virkelighed kommer til samme konklusioner. I en rapport fra folkeskolens følgeforskningsprogram undersøger Bente Bjørnholt og Karl Fritjof Krassel (2016) via både kvantitative og kvalitative midler, hvordan målstyring, baseret på bl.a. nationale test, har fået en større rolle i kommunalt skolearbejde efter reformen - om end styringen i tråde med dansk skoletradition overvejende er 'blød' og dialogbaseret.

[7] Dette afhænger af, at den centrale myndighed har redskaberne til at sikre implementering af styringsrelationen, hvilket selvfølgelig er et stort 'hvis' i mange tilfælde. Fænomener som 'gaming', (Bevan & Hood 2006; Smith 1995; Van Thiel & Leeuw 2002) eller symbolsk efterlevelse (Meyer & Rowan 1977) er måder at omgå central styring. Tallenes politik udfolder sig indenfor organisatoriske og institutionelle rammer, hvorfor mere dybdegående analyser af et bestemt tal sikkert kan beriges ved at inddrage andre teoretiske perspektiver. På dette sted har jeg dog været optaget af at udforske det bidrag, er fokus på kontroversiel relevans giver.

vægt på testen som pædagogisk redskab, men ser den i stedet som en kontrolmekanisme, der skal hjælpe med at undgå en uddannelsesmæssig "Tønder-sag hver tredje måned" (Riise 2016). Denne kommentar faldt under en debat af de mange kritikpunkter, der er rejst ved dens pædagogiske anvendelighed. Samme pointe har daværende undervisningsminister Ellen Trane Nørby, der fremhæver, at vi ikke har noget at sætte i stedet for testene, og derfor ikke kan undvære dem (Ravn 2016). Testene forsvares her som den (mindst ringe) måde at sikre kvalitet i skolen, og derfor skal vi forsætte med at bruge dem på trods af deres mangler. At argumentere for en bestemt relevansform fungerer her som et svar til kritikerne og dermed som en måde at håndtere kontroversen. Hvis testene blot skal sikre ensartet kvalitet, er det jo ikke er problem, at lærerne ikke finder dem relevante.

Konklusion

Jeg har i denne artikel udfoldet en ide om, at begrebsparret 'kontroversiel relevans' kan fungere som indgangsvinkel til at analysere tals politik. Tal ses ofte som en særlig objektiv vidensform, hvilket giver dem politisk nyttige konnotationer af sikkerhed og troværdighed. Men i praksis er mange tal løbende genstand for kontroverser, hvor deres kvalitet som beskrivelse af de målte objekter udfordres. Kontroversiel relevans tilbyder en analyseramme, der tager hensyn til denne dobbelthed ved at rette opmærksomheden mod de påstande om relevans, et tals støtter rejser, og de kontroverser, det bliver genstand for.

De nationale test har gennemgået en udvikling, hvor de først blev foreslået med reference til de generelle – men uspecificerede – fordele ved 'åbenhed og gennemsigtighed'. Da offentliggørelse af testresultaterne var for kontroversielt, blev testenes status som 'pædagogisk redskab' den dominerende måde at italesætte relevans. Denne pædagogiske relevans har imidlertid altid været udfordret af en række kontroverser, og anvendelse af testresultaterne til styring og kvalitetssikring er blevet en mere fremherskende relevansgørelse, der ofte forsvares med argumentet om, at selvom testene ikke er perfekte, er de det bedste, vi har. Denne analyse af de nationale test peger på en række dynamikker i forholdet mellem kontroverser og relevans og giver dermed indikationer af, hvad der foregår mellem figur 2's akser.

For det første måtte oplagte relevansformer knyttet til testenes offentliggørelse fravælges pga. deres kontroversialitet. Som forventet ser der i praksis ud til at være et vist 'trade-off' mellem de to begreber, således at håndteringen af en kontrovers kan udelukke bestemte relevansformer. *Dernæst* afslørede forløbene mht. testene som pædagogisk værktøj og styringsredskab, at der er forskel på, i hvor høj grad en given relevansform kan fastsættes institutionelt: den pædagogiske relevans blev aldrig for alvor accepteret, hvilket muligvis har været medvirkende til et skift i retning af testene som styringsredskab. Dette antyder, at hvad der tæller som en problematisk kontrovers er af-

hængigt af, hvilken relevans man fokuserer på. Det er invaliderende for testen som pædagogisk redskab, hvis individuelle elever vurderes upræcist – men som overordnet screeningsværktøj kan testen stadig fungere, idet resultaterne blot skal fremstå stabile på skole- eller kommuneniveau. *For det tredje* er forholdet ikke kun er envejs, da vigtigheden af en bestemt relevans også kan bruges som middel til at håndtere kontroverser, som vi så mht. den nye dagsorden om testene som kvalitetssikrings- og kontrolredskab. En analyse bør dermed ikke kun skride frem som en identifikation af, hvor kontroversielle et tal fremstår i forskellige situationer, men også tage højde for, at italesættelse af en ny relevansform kan hjælpe med at afmontere en eksisterende kontrovers.

Denne artikels konceptuelle og empiriske udforskning har blot været et første skridt ind i et terræn, der skal udfyldes med yderligere indhold. Sammenspillet mellem kontroverser og relevans er tydeligvis komplekst, og empirien udfordrer adskillige steder figur 2's simple afbildning af deres forhold. Fx er det vanskeligt at vurdere, om testene generelt bliver mere eller mindre kontroversielle, når vi på samme tid observerer øget kritik af deres pædagogiske egenskaber og en gryende accept af deres roller som styringsredskab. Mit formål i denne artikel har da heller ikke været at opstille en færdig teori, men at præsentere en analytisk tilgang der retter opmærksomheden mod den måde, interessemæssige og metodiske/videnskabelige forhold spiller sammen i debatter om tal. Tal er altid resultat af bestemte metodiske valg, og hvis tallene siger noget om politisk relevante emner, er disse valg også politiske – og dermed ofte kontroversielle. Med Judith Innes' ord: "the only way a statistician can keep out of politics is to collect irrelevant data" (1990, 75).

Litteratur

Alkin, M. C. and King, J. A. (2016). The Historical Development of Evaluation Use. *American Journal of Evaluation*, vol. 37(4), pp. 658-579.

Andersen, A. S. (2016). Lærer undrede sig: "Inkommensurabel" fjernes fra danskprøve. [online], DR, Available at: http://www.dr.dk/nyheder/regionale/midtvest/laerer-undrede-sig-inkommensurabel-fjernes-fra-danskproeve [Accessed 20 Oct. 2016].

Andersen, V. N. (2007). Transparency and Openness: A Reform or Education Policy?. *Scandinavian Political Studies*, vol. 30(1), pp. 38-60.

Andersen, V. N., Dahler-Larsen, P. and Pedersen, C. S. (2009). Quality assurance and evaluation in Denmark. *Journal of Education Policy*, vol. 24(2), pp. 135-147.

Asdal, K. (2011). The office: The weakness of numbers and the production of non-authority. *Accounting, Organizations and Society*, vol. 36(1), pp. 1-9.

Bevan, G. and Hood, C. (2006). Whats measured is what matters: Targets and Gaming in the English Public Health Systems. *Public Administration*, vol. 84(3), pp. 517-538.

Bhatti, Y., Hansen, H. F. and Rieper, O. (2006). *Evidensbevægelsens udvikling, organisering og arbejdsform: En kortlægningsrapport* [online] København: AKF-Forlaget. Available at: http://curis.ku.dk/ws/files/15320254/evidens.pdf [Accessed 23 May 2017].

Bjørnholt, B. & Krassel, K. F. (2016). *Midtvejs i folkeskolereformen – en midtvejsmåling af den kommunale styring i forbindelse med folkeskolereformen* [online] København: KORA. Available at: https://www.kora.dk/udgivelser/udgivelse/i13592/ [Accessed 23 May 2017].

Blyth, M. (2013). Paradigms and Paradox: The Politics of Economic Ideas in Two Moments of Crisis. *Governance*, vol. 26(2), pp. 197-215.

Bowen, G. A. (2006). Grounded Theory and Sensitizing Concepts. *International Journal of Qualitative Methods*, vol. 5(3), pp. 12-23.

Bowker, G. C. and Star, S. L. (1999). *Sorting Things Out: Classification and Its Consequences*. Cambridge: MIT Press.

Bundsgaard, J. and Puck, M. R. (2016). *Nationale test : Danske lærere og skolelederes brug, holdninger og viden* [online]. Aarhus: DPU, Aarhus Universitet og Center for Anvendt Skoleforskning, University College Lillebælt. Available at: http://pure.au.dk//portal/files/102516329/Ebog_Nationale_test_FINAL_september_2016.pdf [Accessed 23 May 2017].

Børstin, M. And Fuglsang, J. (2010). Haarder og Løkke er i åben strid om skoletest. *Politiken*, [online]. Available at: http://politiken.dk/indland/politik/ECE889088/haarder-og-loekke-er-i-aaben-strid-om-skoletest/ [Accessed 20 Oct. 2016].

Carlsen, E. M. (2004). Ledende artikel: Derfor vinder Fogh. *BT*, section 1, p. 2.

Contandriopoulos, D., Lemire, M., Denis, J.-L. and Tremblay, É. (2010). Knowledge Exchange Processes in Organizations and Policy Arenas: A Narrative Systematic Review of the Literature. *Milbank Quarterly*, vol. 88(4), pp. 444-483.

Dahler-Larsen, P. (2006). *Evalueringskultur: et begreb bliver til*. Odense: Syddansk Universitetsforlag.

Dahler-Larsen, P. (2012). Når test skal testes. *Unge Pædagoger*, 2012(3), pp. 14-22.

Dahler-Larsen, P. (2013), *Evaluering af projekter: og andre ting, som ikke er ting*. Odense: Syddansk Universitetsforlag.

Dahler-Larsen, P. and Kristiansen, M. B. (2015). Tema: Hvad kom der ud af evalueringsbølgen?. *Økonomi & Politik*, vol. 88(1)-

Deming, E. (2017). [online]. *Goodreads.com.* Available at: https://www.goodreads.com/author/quotes/310261.W_Edwards_Deming [Accessed 23 May. 2017].

DR, (2005). *Den første partileder-debat*. [online]. Available at: http://www.dr.dk/nyheder/htm/baggrund/tema2005/fvvalg/381.htm. [Accessed 23 Jan. 2017].

Ekholm, M., Mortimore, P., David-Evans, M., Laukkanen, R. and Valijarvi, J. (2004). *OECD-rapport om grundskolen i Danmark – 2004* (Uddannelsesstyrelsens temahæfteserie, 5). København: Undervisningsministeriet.

Espeland, W. N. and Sauder, M. (2007). Rankings and reactivity: How public measures recreate social worlds. *American Journal of Sociology*, vol. 113(1), pp. 1-40.

Espeland, W. N. and Stevens, M. L. (1998). Commensuration as a Social Process. *Annual Review of Sociology*, vol. 24, pp. 313-343.

Folkeskolen, (2005). *Koks i forligskredsen*. [online]. Available at: https://www.folkeskolen.dk/39736/koks-i-forligskredsen. [Accessed 26 Jan. 2017].

Folketinget. (2006a). *Førstebehandling af L 101 Forslag til lov om ændring af lov om folkeskolen (Styrket evaluering og anvendelse af nationale test som pædagogisk redskab samt obligatoriske prøver m.v).* Folketinget.dk. Availbale at: http://www.ft.dk/samling/20051/lovforslag/l101/beh138/forhandling.htm?startItem=-1#alleindlaeg [Accessed 25 Jan. 2017].

Folketinget. (2006b). *Førstebehandling af L 170 Forslag til lov om ændring af lov om folkeskolen. (Præcisering af folkeskolens formål, ekstra timer i dansk og historie, elevplaner, offentliggørelse af landsresultater af test, præcisering af det kommunale ansvar samt etablering af nyt råd for evaluering og kvalitetsudvikling af folkeskolen)*. Folketinget.dk. Available at: http://www.ft.dk/samling/20051/lovforslag/L170/BEH1-66/forhandling.htm#dok [Accessed14 Feb. 2017].

Folketinget. (2017). *Førstebehandling af B 55 Forslag til folketingsbeslutning om at gøre de nationale test til frivillige redskaber i folkeskolen.* Folketinget.dk. Available at: http://www.ft.dk/samling/20161/beslutningsforslag/B55/BEH1-58/forhandling.htm#dok [Accessed 14 Feb. 2917].

Grek, S. (2009). Governing by numbers: the PISA Effect in Europe. *Journal of Education Policy*, vol. 24(1), pp. 23-37.

Gustafsson, L. R. (2012). *What Did You Learn in School Today?: How Ideas Mattered for Policy Changes in Danish and Swedish Schools 1990-2011*. Aarhus: Aarhus Universitet.

Hellisen, H. (2014). *Nu skal resultatstyringen være lov*. [online], Folkeskolen. Available at: http://www.ft.dk/samling/20051/lovforslag/L170/BEH166/forhandling.htm#dok. [Accesed 14 Feb. 2017].

Hermann, S. 2007. *Magt og Oplysning - Folkeskolen 1950-2006*. København: Unge Pædagoger.

Innes, J. E. (1990). *Knowledge and Public Policy: The Search for Meaningful Indicators*, 2. ed. New Brunswick, NJ: Transaction Publishers.

Karen, R. (2014). *Folkeskolen tester opgaveudvikling: Lang vej fra tanke til test*. [online], Folkeskolen. Available at:

http://www.folkeskolen.dk/545682/folkeskolen-tester-opgaveudvikling-lang-vej-fra-tanke-til-test. [Accessed 13 Feb. 2017].

Karen, R. (2016). *Skoleprofessor: Problemer med nationale test er større, end jeg troede*. [online], Folkeskolen. Available at: http://www.folkeskolen.dk/583736/skoleprofessor-problemer-med-nationale-test-er-stoerre-end-jeg-troede [Accessed 20 Oct. 2016].

KL, DLF, DS, S&S, and BKF. (2004). *Folkeskolens svar på OECD's anbefalinger - Tilbagemelding til undervisningsministeren fra KL, Danmarks Lærerforening, Lederforeningen, Danmarks Skolelederforening, Skole og Samfund, Børne- og Kulturchefforeningen* [online]. Available at: http://www.dlf.org/media/44138/folkeskolenssvarpaaoecd.pdf [Accessed 23 May 2017].

Kousholt, K. (2015). Vidensformers legitimitet i skolepraksis. *Nordic Studies in Education*, vol. 35(3/4), pp. 168-183.

Krejsler, J. B., Olsson, U. and Petersson, K. (2014). The transnational grip on Scandinavian education reforms - The open method of coordination challenging national policy-making. *Nordic Studies in Education*, vol. 34(3), pp. 172-186.

Krogstrup, H. K. (2011). *Kampen om evidens: Resultatmåling, Effektevaluering og Evidens*. København: Hans Reitzel.

Latour, B. (1987). *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge: Harvard University Press.

Latour, B. (1999). *Pandora's Hope: Essays on the Reality of Science Studies*. Cambridge: Harvard University Press.

Ledermann, S. (2012). Exploring the Necessary Conditions for Evaluation Use in Program Change. *American Journal of Evaluation*, vol. 33(2), pp. 159-178.

Lindeberg, T. (2015). Evaluering som politisk forsikring. *Økonomi & Politik*, vol. 88(1), pp. 23.

Maier, C. M. 17-9-2016, *De nationale tests dumper i test* [blog]. Folkeskolen.dk. Available at: https://www.folkeskolen.dk/593504/de-nationale-tests-dumper-i-test [Accessed 26 Jan. 2017].

March, J. G. and Simon, H. A. (1958). *Organizations*. New York: Wiley.

Meyer, J. W. and Rowan, B. (1977). Institutionalized Organizations: Formal Structure as Myth and Ceremony. *American Journal of Sociology*, vol. 83(2), pp. 340-363.

Nørby, E. T. (2016). *Talepapir - Åbent samråd i BUU om de nationale test* [online]. Folketinget.dk. Available at: http://www.ft.dk/samling/20151/almdel/buu/bilag/122/1607817.pdf [Accessed 19 Nov. 2016].

Olsen, J. V. (2010). *Lars Løkke: Det er slut med topstyring af folkeskolen*. [online], Folkeskolen. Available at: https://www.folkeskolen.dk/62803/lars-loekke-det-er-slut-med-topstyring-af-folkeskolen. [Accessed 20 Oct. 2016].

Olsen, J. V. (2010). *Ugen, hvor statsministeren satte skolen på den anden ende*. [online], Folkeskolen. Available at: http://www.folkeskolen.dk/61169/ugen-hvor-statsministeren-satte-skolen-paa-den-anden-ende [Accessed 20 Oct. 2016].

Ordnet.dk (2017a). 'Relevans'. [online], *Ordnet.dk*. Available at: http://ordnet.dk/ddo/ordbog?query=relevans [Accessed 23 May 2017].

Ordnet.dk (2017b). 'Relevant'. [online], *Ordnet.dk*. Available at: http://ordnet.dk/ddo/ordbog?query=relevant [Accessed 23 May 2017].

Pors, J. G. (2009). *Evaluering indefra: politisk ledelse af folkeskolens evalueringskultur*. Frederiksberg: Samfundslitteratur.

Pors, J. G. (2016): The ghostly workings of Danish accountability policies. *Journal of Education Policy*, vol. 31(4), pp. 466-481.

Porter, T. (1994). Making Things Quantitative. *Science in Context*, vol. 7(3), pp. 389-407.

Porter, T. (1995). *Trust in Numbers: The Pursuit of Objectivity in Science and Public Life*. Princeton: Princeton University Press.

Rambøll. (2013). *Evaluering af de nationale test i folkeskolen* [online]. København: Rambøll Management. Availabel at: http://www.ramboll.dk/medier/rdk/ramboll-evaluerer-nationale-test [Accessed 23 May 2017].

Ravn, K. (2008). *Test med mange formål*. [online], Folkeskolen. Available at: https://www.folkeskolen.dk/51098/test-med-mange-formaal. [Accessed 25 Jan. 2017].

Ravn, K. (2010). *De nationale test – hvordan er de?*. [online], Folkeskolen. Available at: https://www.folkeskolen.dk/61344/de-nationale-test--hvordan-er-de. [Accessed 25 Jan. 2017].

Ravn, Karen. (2016). *Nu ved politikerne, at de nationale test ikke måler præcist*. [online], Folkeskolen. Available at: http://www.folkeskolen.dk/593625/nu-ved-politikerne-at-de-nationale-test-ikke-maaler-praecist. [Accessed 21 Oct. 2016].

Riise, A. B. (2016). *S-profil: Giv gode kommuner fem års pause fra nationale test*. [online], Folkeskolen. Available at: http://www.folkeskolen.dk/592251/s-profil-giv-gode-kommuner-fem-aars-pause-fra-nationale-test [Accessed 20 Oct. 2016].

Smith, P. (1995). On the unintended consequences of publishing performance data in the public sector. *International Journal of Public Administration*, vol. 18(2-3), pp. 277-310.

Undervisningsministeriet. (2011). *Brug testresutlaterne - Inspiration til pædagogisk brug af restultater fra de nationale test*. København: Styrelsen for Evaluering og Kvalitetsudvikling af Folkeskolen (Skolestyrelsen) .

Undervisningsministeriet. (2016). *Opgaveproduktion og kvalitetssikring af opgaver til de nationale test* [online]. København: Ministeriet for Børn, Undervisning og ligestilling (Styrelsen for Undervisning og Kvalitet). Available at: http://www.uvm.dk/Aktuelt/~/UVM-DK/Content/News/Udd/Folke/2016/Sep/160912-Faktanotater-om-

opgaveproduktion-og-statistisk-sikkerhed-i-de-nationale-test [Accessed 23 May 2017].

Van Dooren, W. (2009). A Politico-administrative Agenda for Progress in Social Measurement: Reforming the Calculation of Government's Contribution to GDP. *Journal of Comparative Policy Analysis: Research and Practice*, vol. 11(3), pp. 309-326.

Van Thiel, S. and Leeuw, F. L. (2002). The Performance Paradox in the Public Sector. *Public Performance & Management Review*, vol. 25(3), pp. 267-281.

Weiss, C. H. (1979). The Many Meanings of Research Utilization. *Public Administration Review*, vol. 39(5), pp. 426-431.

Weiss, C. H. and Bucuvalas, M. J. (1980). Truth Tests and Utility Tests: Decision-Makers' Frames of Reference for Social Science Research. *American Sociological Review*, vol. 45(2), pp. 302-313.

Willems, T. and van Dooren, W. (2012). Coming to Terms with Accountability. *Public Management Review*, vol. 14(7), pp. 1011-1036.

Young, J. J. (2006). Making up users. *Accounting, Organizations and Society*, vol. 31(6), pp. 579-600.