En oversigt over EVALUERINGSFELTET

med særlig henblik på naturfagsdidaktisk forskning i evaluering

DASERA Forskningsseminar

7. november 2022

Jens Dolin
Institut for Naturfagenes Didaktik

KØBENHAVNS UNIVERSITET

Indhold

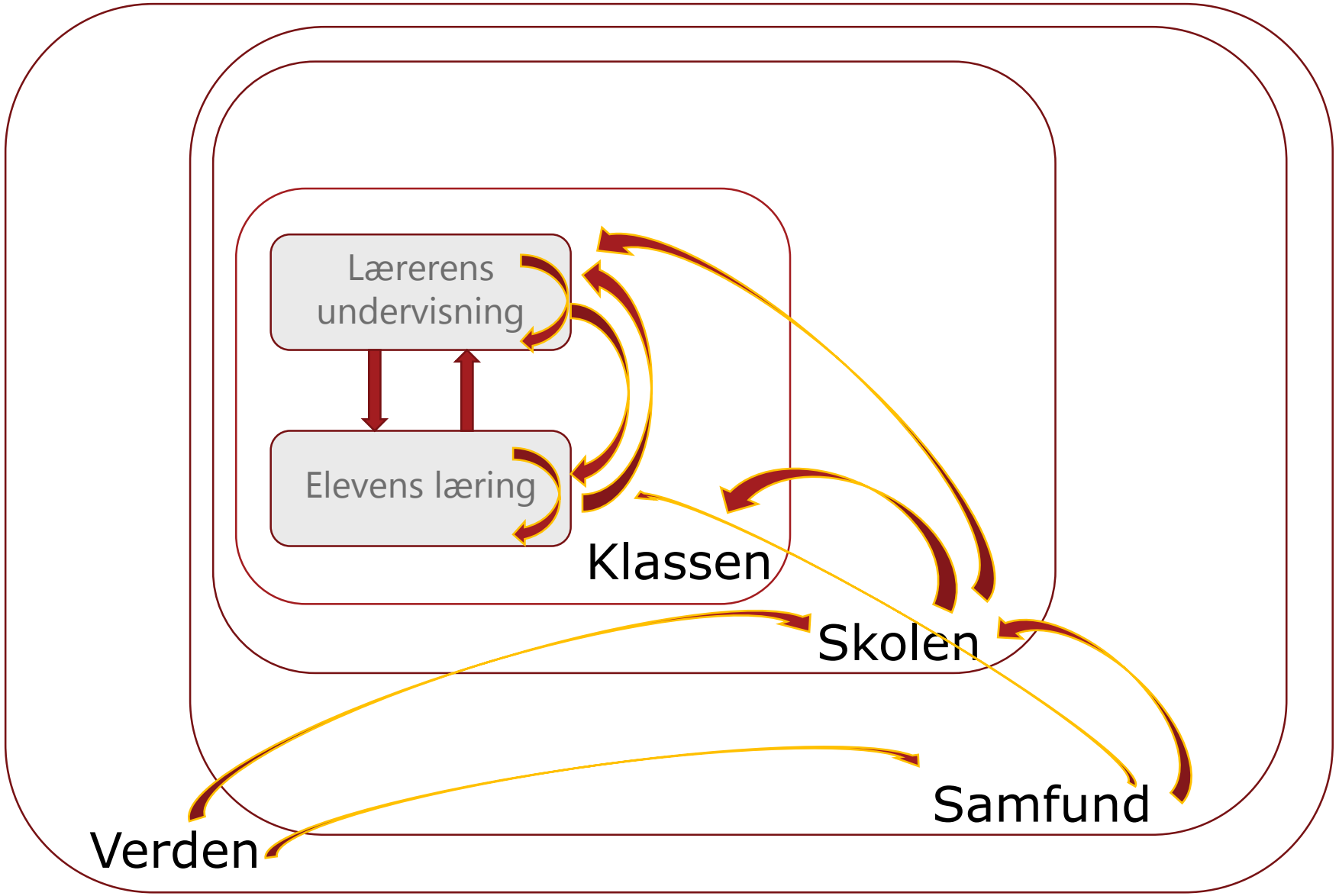
- Evalueringers elementer og kvalitetskrav til evaluering – specielt validitet og reliabilitet
- Karakteristika ved formativ og summativ brug af evalueringer – og deres konsekvenser – og deres mulige sammenhænge
- Evaluering af kompetencer
- Evidens og evalueringers begrænsninger
- Elementer i god evalueringspraksis
- Opbygningen af en evalueringskultur – forskningens rolle

Evalueringers centrale rolle

Et uddannelsessystems evalueringer afspejler de værdier systemet værdsætter og belønner.

Uddannelsessystemets evalueringer har en stærk tilbagesmittende effekt på undervisningen og uddannelsessystemet og bestemmer derfor reelt meget af undervisningens indhold og form.

Evalueringsfeltet



Hvad er evaluering?

Evaluering af en aktivitet handler om at indsamle og vurdere data (i bred forstand) relateret til målene for den aktivitet, der evalueres

Evaluering indebærer *måling* af en præstation og *vurdering* i forhold til nogle kriterier

Ved al måling måler man det måleinstrumentet viser - og ikke nødvendigvis det man er interesseret i! (Spørgsmål om *gyldighed/validitet*)

Al vurdering har et element af subjektivitet i sig. (Spørgsmål om *pålidelighed/reliabilitet*)

Al måling påvirker det, der skal måles

Validitet (gyldighed)

Validitet er et udtryk for, hvorvidt et instrument måler det, som det er tænkt at skulle måle, om resultaterne (tolkningen af testresultaterne) er gyldige for et specifikt formål. Et udtryk for hvilke slutninger man berettiget kan drage på baggrund af testresultater

Validitet kan ikke måles, men skal vurderes.

Begrebsvaliditet (construct validity): Begrebsvaliditeten udtrykker hvorvidt evalueringens målemetode er udviklet på baggrund af den rigtige teori/framework for det, der skal testes. Vurderes fx ved at se hvorvidt testens varians i elevscore korrelerer med variable som man ved er relateret til det givne construct.

Indholdsvaliditet (content validity): Bliver alle relevante aspekter evalueret af metoden?

Prædiktiv validitet: Evalueringens korrelation med andre mål, som den bør kunne forudsige (fx fremtidige præstationer)

Nødvendigheden af et rammeværk som grundlag for valide evalueringer

Helt generelt er et rammeværk en sammenhængende struktur over begreber og elementer, der tilsammen dækker en forståelse af et område.

Rammeværket udgør herved en tolkning af hvorledes det givne område kan opfattes, hvad der er vigtigt i det og hvilke væsentlige relationer, der er mellem enkeltdelene ⋯

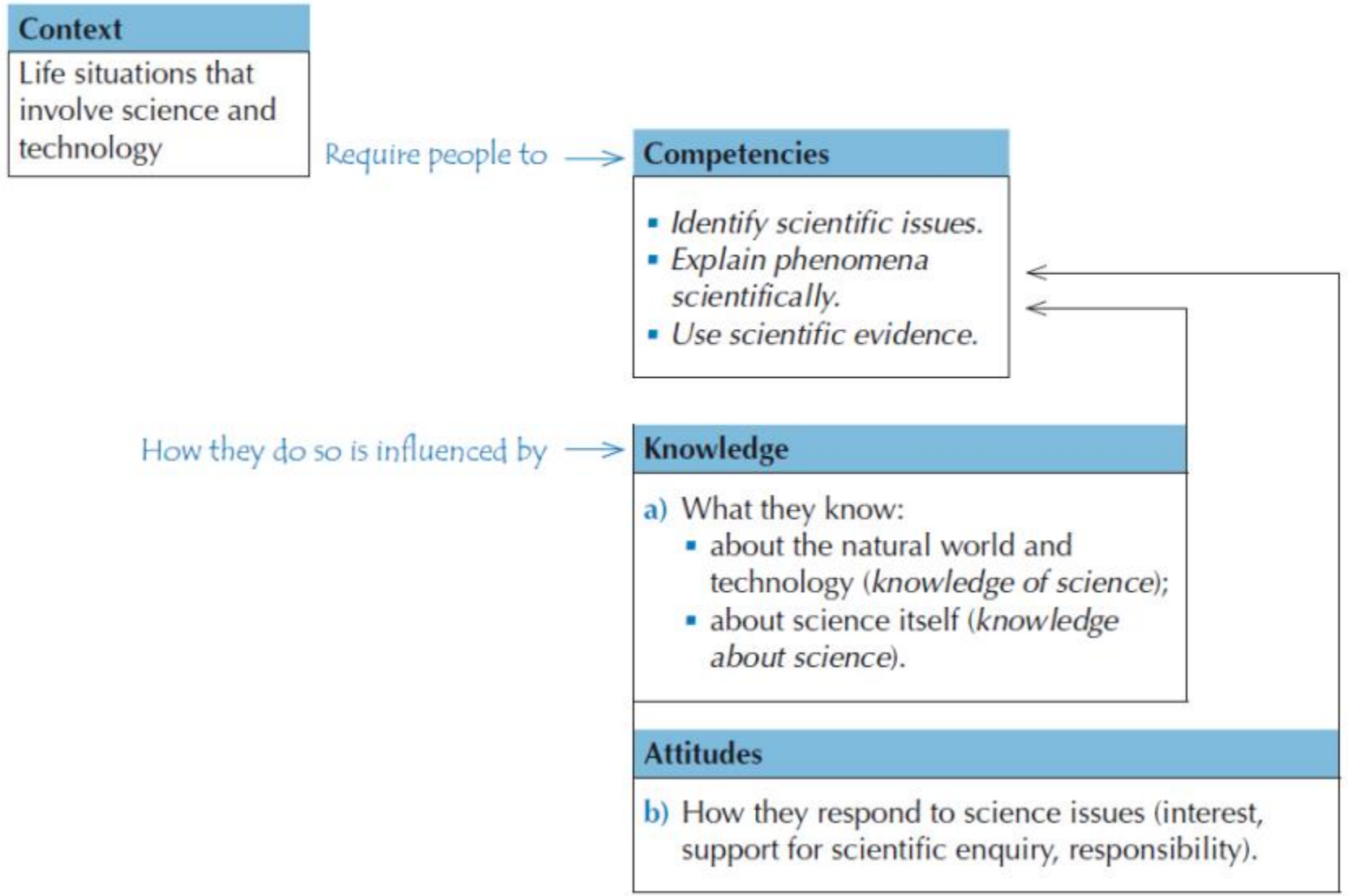


Fig. 2. The PISA 2006 science framework. (OECD 2007, s. 35).

Dolin, J., Ellebæk, J. J., Daugbjerg, P. (2022). *Model for operationelt teoretisk rammeværk*. Naturfagsakademiet – CESE

En Delphi-undersøgelse kan afdække holdninger og synspunkter i et felt

1. Udarbejdelse af forskningsspørgsmål
2. Udvælgelse af et ekspertpanel
3. Udsendelse af 1. runde spørgsmål
4. Indsamle og analysere svar
5. Alle temaer, resumeer og citater distribueres sammen med 2. runde spørgeskema
6. 2. runde svar analyseres
7. De reviderede temabeskrivelser udsendes sammen med 3. runde spørgeskema

Hvad sagde Delphi-undersøgelsen om behov for forskning, udvikling og efteruddannelse i evaluering?

Der er behov for at udvikle en anderledes systematisk og forskningsbaseret evalueringspraksis i naturfagene, en stærkere evalueringskultur.

behov for et større repertoire af evalueringsformater ⋯sikrer at andet end elevernes viden testes.

Formativ evaluering, samt evaluering af naturfaglige kompetencer, omtales som nogle af de største udfordringer for mange lærere.

værktøjer til formativ kompetenceevaluering ⋯ peer-feedback

Også summative aspekter af evaluering bør belyses og kvalificeres viden om karaktergivning, bedømmelseskriterier og den konkrete udformning af prøveformater.

Hvordan håndteres selve prøvesituationen, så den giver eleverne mulighed for at udvise naturfaglige kompetencer, uden at lærerne oplever, at den traditionelle faglige vidensbund forsvinder?

Reliabilitet (pålidelighed)

Reliabilitet referer til *stabiliteten/reproduktionsegenskaben*, dvs. evalueringsmetodens evne til at nå det samme resultat ved gentagne målinger.

Reliabiliteten kan udtrykkes som forskellen mellem den 'sande' værdi – som jo er ukendt - og en tilfældig målefejl (systematiske målefejl hører under validitet). Måles typisk vha Cronbach's α.
Reliabilitet kan også udtrykke *korrelationen* mellem sæt af observationsværdier – fx to eller flere individers bedømmelse af et antal objekter. Måles typisk vha Cohen's κ.

Reliabiliteten mindskes ved at:

- Forskellige bedømmere giver forskellig bedømmelse for samme præstation
- Samme elever præsterer forskelligt til forskellige tider
- Den elevernes præstation afhænger af hvilket spørgsmål der trækkes

"The luck of the draw"

Pålidelighedsproblemer ved evaluering

"Et århundredes forskning har konsistent vist at lærere ikke er pålidelige bedømmere af elevers læring, hvis ikke de bruger strategier til at reducere målefejl" (McMillan 2013, p. 110)

Jo mere narrativt orienterede opgaver/problemer, jo lavere pålidelighed.

300 essays rettet af 53 forskellige censorer. 94 % af opgaverne fik 7 forskellige karakterer på en 7-trin skala. (ibid)

Strategier til at reducere vurderingsupålidelighed:

1. Reducere eller eliminere behovet for menneskelig bedømmelse
2. Etablere 'best practice' ved udarbejdelse at opgaverne
3. Opstille retningslinjer for bedømmelse eller skemaer (rubrics)
4. Fremme læreres forståelse af elevers læring (mhp bedre tolkning af elevsvar)
5. Danne praksisfællesskaber til at opbygge en fælles forståelse af hvad man kan forvente af elever.

Censorpålidelighed ved Folkeskolens afgangsprøver

Pålideligheden er beregnet på baggrund af Folkeskolens prøver i dansk og matematik ved sommerprøven 2016. Besvarelserne for 150 elever blev rettet på tre forskellige måder, nemlig af en censor under almindelige rettevilkår, af en kontrolretter og af en ekspertretter.

Dansk:
For to karakterer givet af to forskellige censorer for den samme besvarelse vil der være 40% sandsynlighed for at de to karakterer er ens, og 87% sandsynlighed for at de højst afviger med en karakter.
Matematik:
For to karakterer givet af to forskellige censorer vil der være 72% sandsynlighed for at de to karakterer er ens, og 99% sandsynlighed for at de højst afviger med en karakter.

"... tallene (viser) med stor tydelighed hvor forsigtig man skal være med at lægge prøvekarakterer til grund for elevselektion. Det er simpelthen ikke muligt at give fuldt pålidelige bedømmelser af elevers prøveresultater." (s. 17)

(Dolin, Nielsen og Rangvid, 2018)

Hvor pålideligt skal det blive?

Rapporten om Folkeskolens afgangsprøver anbefaler (bl.a.):

- Censorrettepålideligheden øges ved at censorerne i højere grad end nu opbygger en fælles standard. Dette kan ske ved at etablere og facilitere strukturerede sociale processer blandt censorerne og ved at der etableres mere præcise kriterier.
- Censormøderne i større udstrækning bidrager til en professionalisering af censorerne
- Der strammes op på kriterier for bedømmelser i dansk og sprogfagene (udarbejdet af fagdidaktikere)

((Dolin et al, 2018, s. 7)

Men: Skal det tilstræbes at alle vurderer den samme præstation ens?

Skal det tillades at lærere (som bedømmere) har forskellige holdninger til hvad der er vigtigt og nødvendigt i deres fag?

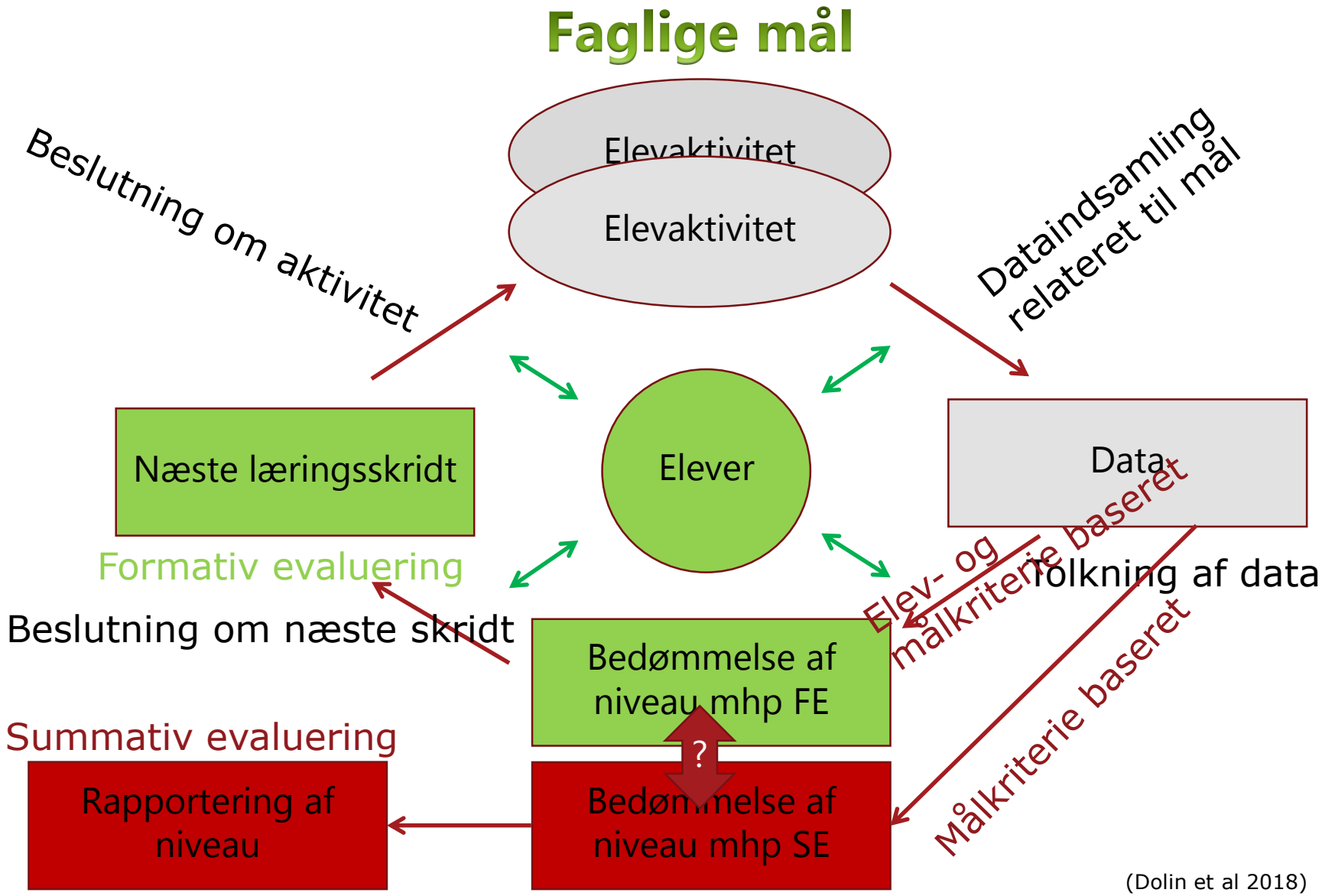
Skal man acceptere – og deklarere – en vis usikkerhed?

Gruppediskussion

Diskutér validitets og pålidelighedsproblemer i forhold til den/de evalueringer, som gruppen arbejder med.

Kan I skitsere et første framework for evalueringen?
Kan I påpege de væsentligste kilder til upålidelighed?

Formativ og summativ evaluering af elevlæring



(Dolin et al 2018)

Definitioner

Formativ (brug af) evaluering har til formål at fremme læring gennem feedback – så lærerne kan forbedre deres undervisning og eleverne deres læring, *dvs. evaluering for læring.*

Summativ (brug af) evaluering har til formål at teste individuelt niveau af læring/performance – for at kunne følge en udvikling eller sammenligne med givne standarder, *dvs evaluering af læring.*

Formativ og summativ refererer til hvorledes resultaterne af evalueringen bruges. Alle evalueringsmetoder kan bruges såvel formativt som summativt.

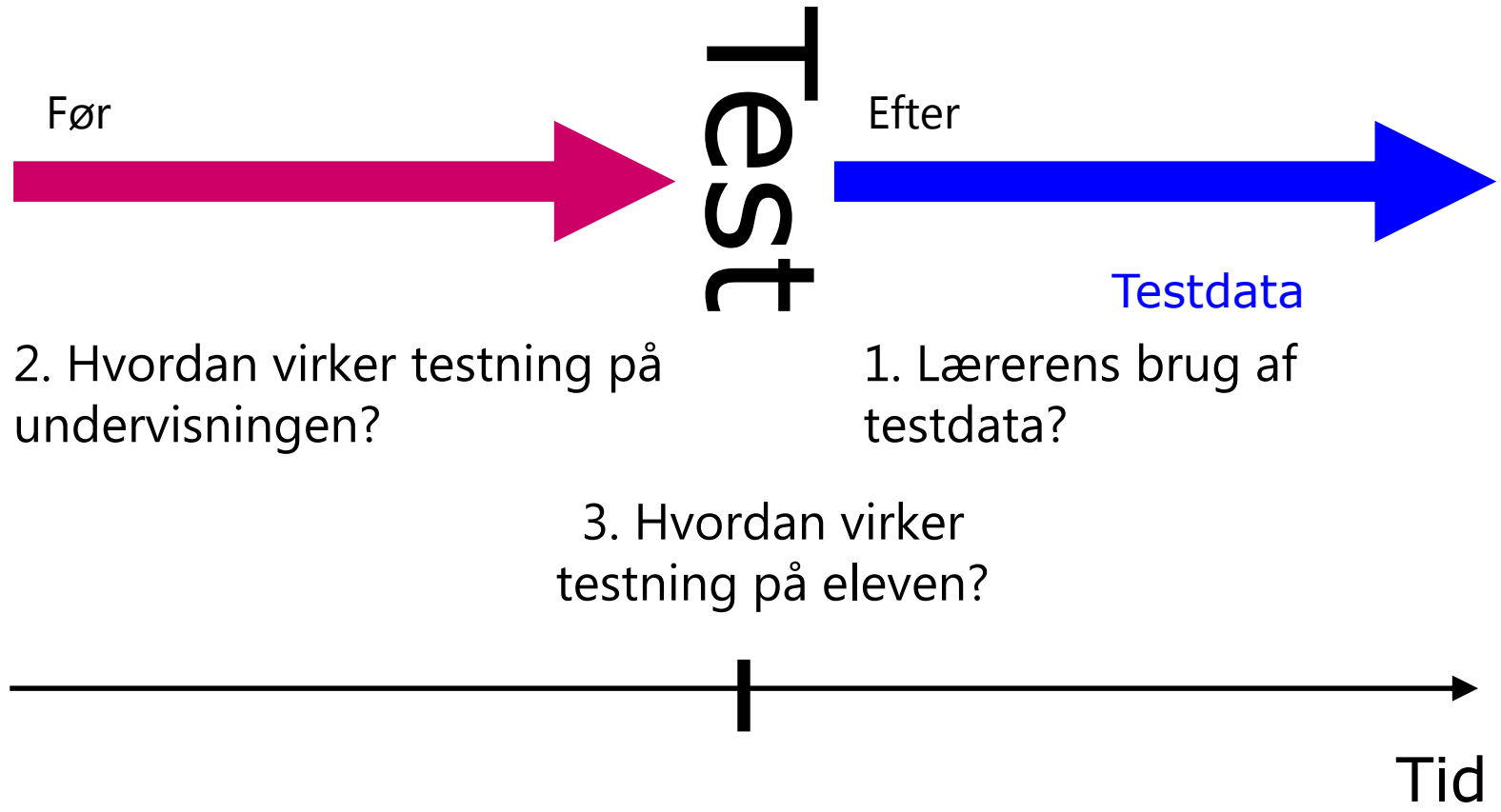
Summativ evaluering på et niveau kan fungere som formativ evaluering på et andet niveau. Fx prøveresultater som lærer- og skoleevaluering, PISA-resultater som udtryk for folkeskolens effektivitet, karakterer som udtryk for løfteevne, undervisningsevalueringer som udtryk for uddannelseskvalitet.

Forskningsspørgsmål:

Hvordan indvirker indførelsen af testning læreres didaktiske beslutninger og elevers læringsadfærd?

(Nordenbo et al, 2009)

Forskningsset-up for Pædagogisk brug af test



5986 referencer → 118 dokumenter → 61 genbeskrivelser
→ narrative synteser baseret på 43 forskningsprojekter

Hvordan virker testning på undervisningen og eleverne?

Undervisningen

- indsnævret eller fordrejet curriculum, idet faglige tankegange forsimples, faktaviden og mekaniske færdigheder betones på bekostning af kreative og æstetiske perspektiver
- undervisningstid allokeres til det/de fag og fagområder, der skal testes i, på bekostning af de fag og fagområder, der ikke testes i, og
- undervisningen kan forfalde til træning til testen og udenadslæren

Eleverne

- elevernes testresultater stiger ved indførelse af test, men først efter nogle år
- når en test annonceres, kan det udløse følelsesreaktioner som nervøsitet og angst
- eleven forbereder sig ved at lære udenad og memorere sætninger
- for bedre præsterende elever stiger motivation, mens svagere præsterende taber modet
- det testresultat, som eleven får ved testen, kan virke ind på fremtidig motivation og selvværd.

(Nordenbo et al, 2009)

Kulturelle skævheder

Undersøgelsesland	Antal studier
USA	32
Kina/Hong Kong/Taiwan	6
U.K.	5
Israel	4
Australien	3
Canada	2
Sverige	2
Danmark	1
Svejts	1
Trinidad Tobago	1
Sri Lanka	1
Japan	1
Uspecificerede eller mange lande	4

Præstationsorientering vs mestringsorientering

”En **mestringsorienteret** person fokuserer på den aktuelle opgave og relaterer sig især til at udvikle kompetence og opnå forståelse og indsigt.
En **præstationsorienteret** person fokuserer på selvet og relaterer sig især til hvorledes evner bliver bedømt og hvorledes man selv præsterer, især i forhold til andre.”
(Midgley et al 2001, p.77 – egen oversættelse).

Der er stærk evidens for at præstationsorientering underminerer den *indre motivation*.
Der er indikationer på at *ydre motivation* leder til ‘overflade’- snarere end ‘dyb’ læring
(Harlen 2012, p.174, egen oversættelse)

Elever med en præstationsorientering har sværere ved at få udbytte af formative evalueringer, mindre glade for gruppeprocesser, mindre tilbøjelige til at kaste sig ud i nye udfordringer.

”Karaktergivning og eksamen er omstændigheder, som tilskynder til en præstationskultur hos mange elever – simpelthen fordi de lægger op til sammenligning og konkurrence”
(Krogh&Andersen, p.174)

John Hattie - synlig læring

Growth Effect Size på 0,4 svarer til 1 års fremgang (ved almindelig skolegang).
Vi søger effekt på over 0,4.

- Lade elever gå et år om (-0,18)
- Erfaring fra andet arbejde end undervisning (0,09)
- ”Teacher subject matter knowledge” = lærerens faglige viden (0,09)
- Efteruddannelse af lærere (0,12)
- Elevernes læringsstil (0,18)
- Reduktion af klassestørrelse (0,21)
- Elev forventninger (1,44)
- Lærer troværdighed (1,07)
- Formativ evaluering (0,90)
- Klassediskussion (0,82)
- Feedback + lærer klarhed (0,75)
- Lærer-elev relationer (0,72)

(data udleveret ved konference med John Hattie og Martin Renton, Ringsted, 20.11.14
Delvist baseret på Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), 81–112. doi:10.3102/003465430298487)

Fire formative evalueringsaktiviteter og fem strategier

Ifølge Black and Wiliam (2012, s.208), er der fire hovedtyper af formative aktiviteter:

- Klassedialog (incl. diskussion)
- Rette med kun kommentarer
- Selv- og kammeratevaluering
- Formativ brug af (summative) tests

Ved at kombinere de tre feedback-trin med de tre agenter i formative processer opstiller de fem hovedstrategier for formativ evaluering (s. 209):

	Hvor skal jeg hen?	Hvor er jeg?	Hvordan kommer jeg derhen?
Lærer	1. Klargøre mål og kriterier	2. Facilitere diskussioner og aktiviteter/opgaver som klargør elevforståelse	3. Sørg for feedback som bevæger den lærende mod målene
Kammerat	Forstå og dele formål og kriterier	4. Aktivér hinanden som ressourcer for hinandens læring	
Elev	Forstå mål og kriterier	5. Gør eleverne til ejere af egne læreprocesser	

At finde balancen mellem den formative og den summative brug af evalueringer

Hvorledes kan man reducere de negative effekter af summative evalueringer og fremme formativ brug af evalueringer?

Feedback for at øge læringen

Fremmer mestringsorientering
Styrker selvværd

Genererer styringsdata

Fremmer præstationsorientering
Øger stress

Formativ og summativ brug af evalueringer kan ses som et kontinuum

	Formativ	<->	Summativ	
	Informal form.	Formal form.	Informal sum	Formal sum
Primært fokus	Hvad er næste læringstrin		Hvad er der opnået til dato	
Eksempel	Klassedialog	Skriftlig fb uden karakterer	Mindre tests Opgaver m karakterer	Eksamen og prøver
Formål	Informere næste læringstrin	Informere næste læringstrin + uvplan	Monitorere elevniveauer ift uvplan	At dokumentere individuelt elevniveau
Bedømmes af	Elever og lærer	Lærer og elever	Lærer	Lærer og censor

Hvor skarp er denne grænse?

(Dolin, Black et al, 2018)

At kombinere formativ og summativ brug af evalueringer

EVA (2016) anbefaler en klar adskillelse mellem øve- og prøverum

Det er nødvendigt at den didaktiske kontrakt er klar

At *anvende formative evalueringer summativt* vil kræve en målorienteret strukturering af (dele af) undervisningen (dvs. formulering af præcise krav og kriterier), således at de kan foretages stringent og pålideligt. Det kræver en stor arbejdsindsats, som måske vil ændre undervisningsrummet – på godt og ondt

Brug af *summative evalueringer til formative formål* forudsætter at

- formål og kriterier for den summative evaluering stemmer overens med formål og kriterier i den givne undervisning
- at den summative evaluering fokuserer på at teste hvorledes den studerende lærer
- at den summative evaluering kan virke motiverende
- at den summative evaluering tester alle dimensioner i det, der skal læres

Diskussionsspørgsmål

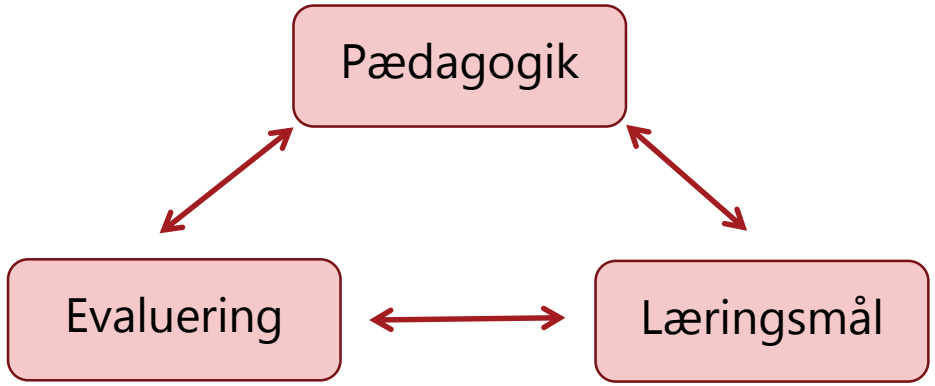
Hvilke data er I interesserede i at få i jeres project, og hvorledes vil I bruge dem summativt (hvis I vil det) og hvorledes formativt?

Kan I kombinere de to anvendelser – dvs. bruge de fundne data både formativt og summativt – og hvilke krav stiller det til evalueringsprocessen og hvilke problemer skal der tages højde for?

Evaluering, pædagogik og læringsmål

Undervisningstilgang

Transmission - tolkning



Evalueringsparadigme

Postpositivistisk - sociokulturelt

Målkategorier

Information/viden - kompetencer

Der skal være alignment mellem mål, pædagogik og evaluering. Nye læringsmål og nye undervisningsformer kræver nye evalueringsformer.

To forskellige testparadigmer

Forskning viser, at ligesom elevers læring afhænger af situationen, hvori de lærer, så afhænger deres præstation af situationen de sættes i.

Post-positivistisk

Elevens evner anses for konstante, på tværs af forskellíge evalueringssituationer.

Tests kan derfor være

- ikke-interaktive
- individuelle
- statiske
- produkt-orienterede
- begrænset i brug af artefakter
- lokaliseret i specielle settings

Socio-kulturel

Elevens evner anses for afhængige af/koblet til evalueringssituationen.

Tests skal derfor være

- interaktive
- samarbejdsorienterede
- dynamiske
- proces-orienterede
- inddragende i brug af artefakter
- realiseret i autentiske situationer

130 elever, der var testet i PISA 2006 blev i 'Validering af PISA'-projektet evalueret i en mere almindelig skolesituation i en række PISA-opgaver, de havde haft i PISA-testen.

Forskellige test paradigmer – forskellige elevscore

Ændring af test-format forøger elevers gennemsnitlige performance ifølge PISA-kriterierne fra et gennemsnit på 0.54 til 0.68-0.71

–en forøgelse på over >26%!

Svage elever klarede sig relativt bedre end stærke elever.

En (traditional) psykometriker vil sige: Den socio-kulturelt orienterede test er lettere for eleven (de må snakke med hinanden og de har hjælpemidler!).

En socio-kulturelt orienteret forsker vil sige: Viden er indlejret i konteksten og afhængig af omstændighederne – den er situeret.

Evaluering af kompetencer

Kompetencerne bedømmes ved at demonstrere handlen i en række konkrete kontekster sammenholdt med et sæt åbne og anerkendte kriterier. Dvs

1. Processer må evalueres i en proces, dvs. der må etableres en *relevant situation* som muliggør realisering af de ønskede kompetencer
2. Der må opstilles *kriterier* for god praksis (som er målbare), baseret på et rammeværk. Fx i form af en operationaliserbar template.
3. Der må opstilles en (ideel) *progression* af den kompetente handlen inden for de enkelte kriterier.

Slutevalueringen på modelkompetence i Autentisk Fysik-projeket

Situation: 3 timers praktisk arbejde i grupper i laboratoriet
Grupperne trækker ukendt problem, som skal modelleres

I forbindelse med den praktiske del vurderedes elevernes evne til at

udvælge problemets variable
udarbejde en plan for udførelse af forsøg (variabelkontrol)
gennemføre (mindst) én måleserie
give en grafisk fremstilling
udføre regressionsanalyse
opstille en model
samarbejde og løse problemer
begå sig i laboratoriet og benytte måleudstyret fornuftigt

Lærer og censor går rundt blandt grupperne og iagttager og spørger og noterer i et skema.
Bedømmelsespålidelighed ca. 85%

Generelle retningslinjer for god evalueringspraksis

Criteria for Evaluating Systems for Student Assessment
"…assessment for any purpose should provide information meeting the criteria of validity, reliability, desired impact, and good use of resources." (Harlen, 2007)

The Assessment Reform Group
Assessment for Learning is the process of seeking and interpreting evidence for use by learners and their teachers to decide where the learners are in their learning, where they need to go and how best to get there.
10 principles for AfL (ARG): AfL

- is part of effective planning
- focuses on how students learn
- is central to classroom practice
- is a key professional skill
- is sensitive and constructive
- fosters motivation
- promotes understanding of goals and criteria
- helps learners know how to improve
- develops the capacity for self-assessment
- recognises all educational achievement

Det amerikanske BEAR Assessment System (https://bearcenter.berkeley.edu/page/about-bear):
1. Test og evaluering skal bygge på et udviklingsperspektiv for læring.
2. Det der undervises i og det der testes og evalueres, skal være i overenstemmelse.
3. Lærere skal kunne håndtere og bruge data.
4. Test og evaluering i skolen skal leve op til sunde standarder for validitet (gyldighed) og reliabilitet (pålidelighed) (Wilson, 2009).

Forskningsmetoder og Evidens

Evidens vil sige at der er signifikante og målbare korrelationer mellem årsag og virkning.

Ved klassiske evidensstudier holdes alle variable pånær én (som varieres) konstant, og man måler effekten af variationen af denne variabel. Man vil så sige at alt andet lige skyldes effekten denne variabel. Men alt andet er aldrig lige når man taler undervisning. Det er svært at kontrollere konteksten.

Man kompenserer for forskellighederne ved at måle på store populationer – men så skal man måle på noget relativt simpelt.

Derfor skelnes mellem global evidens og lokal evidens.

Evidenshierarkiet

1.Systematiske reviews og metaanalyser
2.Random Clinical Trials (RCT)
3.Kohortestudier
4.Case-kontrol studier
5.Tværsektionelle interviews
6.Caserapporter

Styrken af evidens bliver bestemt ud fra:

Hvilken type forskning, der er lavet.
Hvor stor og grundig undersøgelsen er.
Hvor stor den målte effekt er.
Hvor mange studier, der peger i samme retning.
Hvor mange studier, der peger i en anden retning.
(https://videnskab.dk/krop-sundhed/hvad-er-videnskabelig-evidens)

Myter om evidens og mulige konsekvenser af evidensbasering

Fem myter om evidens

1. Evidens baseres kun på kvantitative analyser af RCT-studier
2. Evidens handler kun om effekt
3. Evidens er lig endelig og sikker viden
4. Evidens er en kogebogsopskrift
5. Evidens er opfundet af politikere for at styre praksis

(https://dpu.au.dk/forskning/danskclearinghouseforuddannelsesforskning/omclearinghouse/fem-myter-om-evidens/)

Mulige – uheldige – konsekvenser af (traditional) evidensbasering af uddannelser og undervisning

- Det målelige som kvalitetsparametre – et forsimplet kvalitetsbegreb
- Negligering af lokale forhold (dvs. global harmonisering)
- Nedbrydning af eksisterende (lærer)viden – afprofessionalisering
- Stereotyp undervisningskultur (Best Practice finds ikke)

Evidens vs eksemplaritet

Inden for læreruddannelserne arbejdes med at forbinde begrebet *eksemplaritet* med begrebet *dømmekraft*:
"Der nok kan være evidens for noget, men at overførbarheden til praksis er aldrig er 1:1. Her må vi i den konkrete situation forlade os på vor egen lærerfaglige autonomi og vort pædagogiske skøn."
(Korsgaard og Clausen, 2022)

Korsgaard og Clausen trækker på Martin Wagenscheins eksemplarisk undervisning og Elgins instantiering og ekspressivitet.
"En central pointe er hér, at eksemplaritet ikke bare sker, når vi didaktiserer vores undervisning med *implicit* og *eksplicit modellering*, men at det *altid allerede* er på spil."
Jfr PCK-begrebet, der er båret af erfaringers eksempler.

Opbygning af en evalueringskultur

Eksplicitte begrundelser og procedurer

Er karakteriseret ved det implicitte

"En evaluering udføres med sigte på et formål (en eller anden form for anvendelse), hvorimod kulturer udmærker sig ved at holde noget helligt, som ikke lader sig reducere til noget instrumentelt."

(Dahler-Larsen, 2006, s. 36)

En skoles evalueringskultur ser evaluering som et didaktisk anliggende for hele skolen – med sigte på at øge elevernes læring og vurdere deres læring på et professionelt grundlag

Et uddannelsessystems evalueringskultur ser evaluering som et uddannelsespolitisk anliggende for alle aktører – med henblik på at sikre at rammebetingelserne har sigte på at øge elevernes læring frem for accountability og kontrol.

At opbygge en evalueringskultur er ikke kun et spørgsmål om at bruge de 'rigtige' evalueringsværktøjer.
Det handler også om at forholde sig til de mange virkninger, værktøjerne vil have.

Første reaktion:
Yes! Endelig

Læste aftalen (23s)

Anden reaktion:
Hvorfor …?
Hvorfor ikke …?

Tredje reaktion:
Hvorfor lyttede de ikke (mere) til eksperterne?

Folkeskoleforligskredsen har den 29. oktober 2021 indgået en aftale om det fremtidige evaluerings- og bedømmelsessystem i folkeskolen.

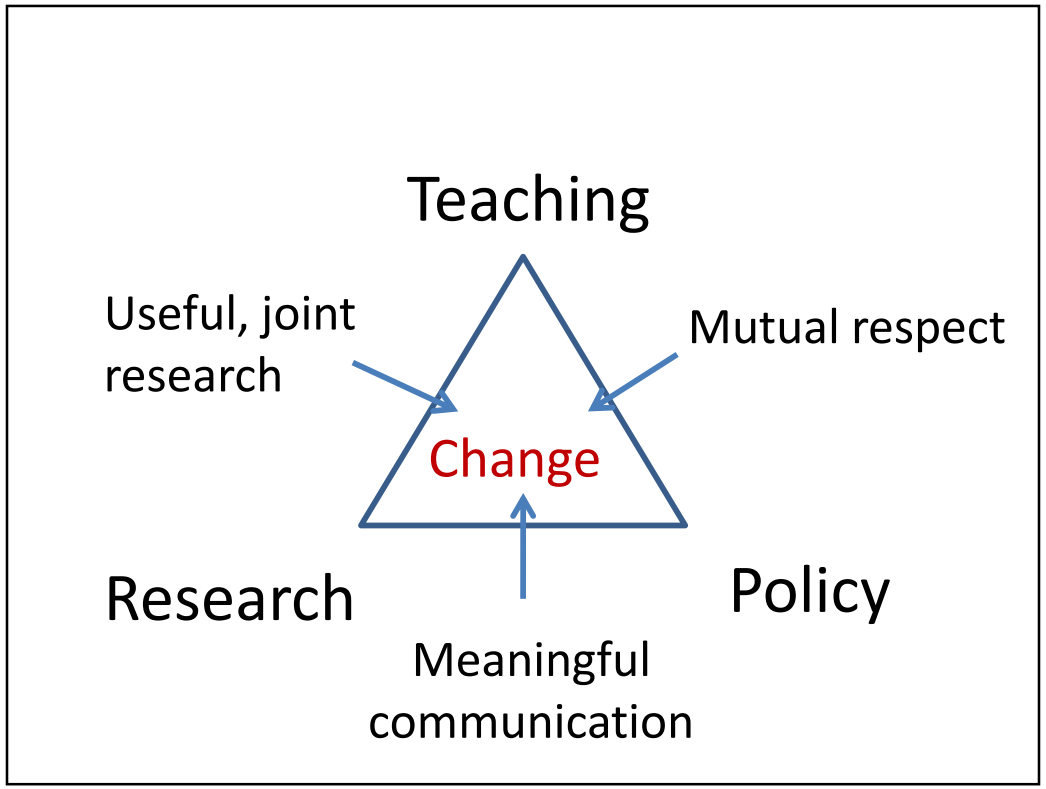
(https://www.uvm.dk/aktuelt/nyheder/uvm/2021/okt/211029-bred-aftale-om-fremtidigt-evaluerings-og-bedoemmelsessystem)

Så: Jeg kontaktede de uddannelsespolitiske ordførere for tre af partierne
- og fik et interview med dem.

Skolerne og lærerne kan ikke gøre det alene

Der skal opbygges dialogfora hvor parterne kan udvikle meningsfulde evalueringsmodeller

"In research we are able to treat policy, practice and assessment as discrete subject fields for study. If research is to contribute to changes in practice, all three have to come together in some orchestrated fashion."
(Aikenhead, citeret efter Fensham (2009). The Link Between Policy and Practice in Science Education: The Role of Research. *Science education*, p. 1078).



Forskere må forfølge forskningsspørgsmål, der er relevante for lærere og policy-makers.

Lærere må ændre undervisningspraksis i lyset af forskningsresultater og policy krav.

Policy-makers må ændre. uddannelsespolitik i overensstemmelse med forskning og hensyn til læreres muligheder for at udføre deres arbejde.

Opmærksomhedspunkter

Ikke alt vigtigt kan måles

Mange kvaliteter er (næsten) umålelige, fx stemningen i en klasse, elevernes dannelsesprocesser, glæden ved engagement, …

Opstillede mål målretter

De fleste aktiviteter og udvikling og prioritering vil rette sig mod de opstillede indikatorer – på bekostning af noget vi også synes er vigtigt og vil bevare.

Man får, hvad man evaluerer
Det man ikke evaluerer, får man ikke

Afsluttende

Evalueringer er uddannelsessystemets og skolens mest indflydelsesrige værktøj, som påvirker elever, lærere og ledelse – de skal anvendes med kritisk omhu.

Summativ brug tilskynder eleverne til en præstationsorientering, lærerne til en forsimplet undervisning og ledelsen til at agere på et (måske for) enkelt grundlag, hvorfor der overordnet set skal mere focus på formativ brug af evalueringer og mindre på summative – for at fremme elevernes læring og trivsel

Både kvalitative og kvantitative evalueringsmetoder kan have stor værdi med hver deres udsigelseskraft og anvendelighed.

Udvikling af læringsfremmende og konsistente måder at evaluere på er ikke kun et spørgsmål om evalueringsværktøjer, men i lige så høj grad et spørgsmål om at forholde sig til de mange virkninger, værktøjerne vil have - om opbygningen af en evalueringskultur.

Opbygningen af en evalueringskultur vedrører hele uddannelsessektoren og kræver et aktivt samspil mellem lærere, forskere og politikere.

Referencer

Assessment Reform Group ( ). *Assessment for Learning – beyond the black box*. http://www.assessment-reform-group.org.uk/

Dolin, J., Black, P., Harlen, W. & Tiberghien, A. (2018). Exploring relations between formative and summative assessment. In Dolin, J. & Evans, R. (eds.). *Transforming Assessment. Through an Interplay Between Practice, Research and Policy*. Springer International Publishing. (s. 53-80).

Dolin, J., Bruun, J., Constantinou, C. P., Dillon, J., Jorde, D. (2018), and Peter LabuddePolicy Aspects: How to Change Practice and in What Direction. In Dolin, J. & Evans, R. (eds.). *Transforming Assessment. Through an Interplay Between Practice, Research and Policy*. Springer International Publishing. (s. 249-278).

Dolin, J., K. Nielsen & B.S. Rangvid (2018). *Rapport fra Følgegruppen for én bedømmer ved folkeskolens prøver*. København: Undervisningsministeriet.

Dolin, J., Ellebæk, J. J., Daugbjerg, P. (2022). Model for operationelt teoretisk rammeværk. Naturfagsakademiet – CESE.

Harlen, W. (2007). Criteria for Evaluating Systems for Student Assessment. *Studies in Educational Evaluation 33* , 15–28

Harlen, W. (2012). The Role of Assessment in Developing Motivation for Learning. In: Gardner, J. (ed.). *Assessment and Learning*. London: SAGE.

Hattie, J. (2009). Visible learning : a synthesis of over 800 meta-analyses relating to achievement. London; New York: Routledge.

Korsgaard, M. T. & Clausen, C. H. (2022). Om eksemplaritet og lærer(ud)dannelse. Fra mimesis til autonomi. *Nordic Studies in Education, 42(3)*, 289–305. https://doi.org/10.23865/nse.v42.3675.

McMillan (2013). I: McMillan (ed.). *SAGE Handbook of Research on Classroom Assessment*. Los Angeles: SAGE

Midgley, C., Kaplan, A. & Middleton, M. (2001). Performance-Approach Goals: Good For What, For Whom, Under What Circumstances, and At What Cost? *Journal of Educational Psychology*, 93(1), s. 77-86.

Nordenbo, S. E., Allerup, P., Andersen, H.L., Dolin, J., Korp, H., Larsen, M.S., Olsen, R.V., Svendsen, M.M., Tiftikçi, N., Wendt, R.E., Østergaard, S. (2009). *Pædagogisk brug af test – Et systematisk review*. København: Danmarks Pædagogiske Universitetsforlag og Dansk Clearinghouse for Uddannelsesforskning

OECD (2007). *PISA 2006. Science Competencies for Tomorrow's World. Volume 1 – Analysis*. Paris: OECD. https://www.oecd-ilibrary.org/education/pisa-2006_9789264040014-en (Access date: March 16, 2022).

Sara Tougaard, Jan Sølberg og Bella Marckmann (2019. Evalueringstilgange i naturfag i grundskolenNaturfagenes evaluerings- og udviklingscenter - neuc.dk