KUNNSKAPSSENTER
FOR UTDANNING

Former for lærervurdering som kan ha positiv innvirkning på skolens kvalitet

En systematisk kunnskapsoversikt

Sølvi Lillejord, Kristin Børte, Erik Ruud, Trond Eiliv Hauge, Therese N. Hopfenbeck, Astrid Tolo, Peder Fisher-Griffiths og Jens-Christian Smeby

Tittel: *Former for lærervurdering som kan ha positiv innvirkning på skolens kvalitet: En systematisk kunnskapsoversikt*

Referanse: Lillejord, S., Børte, K., Ruud, E., Hauge, T. E., Hopfenbeck, T. N., Tolo, A., Fischer-Griffiths, P. & Smeby, J.-C. (2014) *Former for lærervurdering som kan ha positiv innvirkning på skolens kvalitet: En systematisk kunnskapsoversikt*. Oslo: Kunnskapssenter for utdanning, www.kunnskapssenter.no

ISBN: 978-82-12-03332-0

Referanse nr. KSU 1/2014

Publisert: April 2014

Finansiering: Denne rapporten er finansiert gjennom et oppdrag fra Kunnskapsdepartementet.

Forskergruppe: Associate Professor Karen Hammerness, Bard College, New York, USA
Professor emeritus Trond Eiliv Hauge, Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo
Lecturer Therese N. Hopfenbeck, Oxford University Centre of Educational Assessment, UK
Førsteamanuensis Astrid Tolo, Institutt for pedagogikk, Universitetet i Bergen
Professor Jens-Christian Smeby, Høgskolen i Oslo og Akershus

Kontakt: Kunnskapssenter for utdanning, Norges forskningsråd
Besøksadresse: Drammensveien 288, 0283 Oslo
Postadresse: Postboks 564, 1327 Lysaker
Telefon: (+47) 22 03 70 00
Mail: kunnskapssenter@forskningsradet.no
Nett: www.kunnskapssenter.no

KUNNSKAPSSENTER
FOR UTDANNING

Innholdsfortegnelse

Sammendrag

Kunnskapssenter for utdanning presenterer her en systematisk kunnskapsoversikt som har følgende problemstilling (scope):

Hvilke former for lærervurdering kan ha positiv innvirkning på skolens kvalitet

Rapporten har en innledning som danner grunnlag for å forstå forskningen om lærervurdering i en norsk kontekst. Innledningen presenterer arbeidet med kunnskapsoversikten og tar opp særskilte norske forhold, som innføring av et nasjonalt system for kvalitetsvurdering, lærerutdanning i Norge, norske læreres arbeidssituasjon og betingelser i norsk arbeidsliv. Den systematiske kunnskapsoversikten (kapittel 2-6) innledes med metodepresentasjon, hvorpå forskningsresultatene blir presentert og syntetisert.

Lærervurdering i Norge

Det er ingen sterk tradisjon for lærervurdering i Norge. Det som finnes av tiltak, er avhengig av initiativ i enkelte skoler, kommuner eller fylker. Norske lærere har tradisjonelt hatt stor autonomi, og skolens samfunnsoppgave står sentralt i lærerutdanningene. OECD har påpekt at norske lærere sjelden får tilbakemelding på jobben de gjør, og har anbefalt at norske myndigheter integrerer lærervurdering i det nasjonale systemet for kvalitetsvurdering. I arbeidet blir skoleeiere og skoleledere nøkkelaktører, og vurderingen må ha som mål å fremme profesjonsutvikling og skoleutvikling.

Den systematiske kunnskapsoversikten

I arbeidet med den systematiske kunnskapsoversikten er det søkt i fagfellevurderte tidsskrift etter 2009 og i tillegg utført håndsøk. Søkeresultatene er sortert og kategorisert etter bestemte inklusjons- og eksklusjonskriterier. Antall inkluderte referanser som utgjør kjerneartiklene i kunnskapsoversikten er 79. Disse er kartlagt, kvalitetsvurdert og analysert. Funnene er syntetisert på tvers av artiklene ved hjelp av en narrativ syntese. Det er utarbeidet en protokoll som sikrer transparens ved å presentere hvert trinn i den systematiske metoden benyttet i arbeidet med kunnskapsoversikten.

Kunnskapsoversikten viser:

I den inkluderte litteraturen finnes beskrivelser av nasjonale systemer for lærervurdering fra Chile, Kina, Belgia og Portugal. Felles for alle landene er at systemene har doble formål. De skal både bidra til profesjonsutvikling (formativt), og kvalitetssikring og kontroll (summativt). Forskningen viser at det er en tendens til at profesjonsutvikling blir nedprioritert og kontrollfunksjonene overtar. Problemene ser ut til å oppstå når nasjonale intensjoner skal iverksettes lokalt. Om vurderingen skal føre til læring og profesjonsutvikling, er det nødvendig med kunnskapsrike skoleeiere og skoleledere og et godt samarbeid mellom ledelse og lærere.

Forskning om summativ vurdering kommer stort sett fra USA. Her har flere delstater tatt i bruk Value-Added modeller som tar sikte på å måle den enkelte lærers bidrag til elevenes læringsutbytte. Mye av litteraturen som er fanget opp gjennom søkene drøfter kvaliteten på ulike psykometriske modeller og reliabiliteten av de økonomiske modellene som brukes i lærervurderingssystemene. Mange forskere er kritiske til bruken av slike modeller, spesielt når tallene brukes til å ansette, forfremme eller si opp lærere.

Majoriteten av studiene som har sett på prestasjonslønn i vurderingssystemer basert på Value-Added modeller, konkluderer med at de ikke finner positiv effekt av lønnsinsentiver verken på lærernes arbeidsinnsats eller praksis.

I litteraturen som undersøker formativ vurdering, er det identifisert tre tilnærminger: elever som vurderer lærere, egenvurdering og kollegavurdering. Gjennomgående understrekes betydningen av konstruktive og gode arbeids- og læringsmiljø, med ledere som deltar aktivt i tiltak som blir satt i gang for å fremme profesjonslæring. Det finnes lite forskning om elever som vurderer lærere. Forskningen anbefaler enkle metoder som kollega- og egenvurdering, som viser seg å være nyttige for læreres profesjonsutvikling og særlig for nye lærere. God formativ vurdering kan bidra til å heve kvaliteten på skolen, styrke lærerprofesjonen og øke læreryrkets status.

Hva fremmer god lærervurdering

Forskningen viser at medvirkning er nødvendig for at lærervurdering skal lykkes. Bred deltakelse i hele prosessen, fra utforming til evaluering av vurderingssystemer, gir systemene legitimitet. Det må settes av tid og ressurser til arbeidet, og det må foreligge planer for hva som skal gjøres med problemer som avdekkes gjennom vurderingen. Videre er det viktig at vurderingen avgrenses og at det er en felles forståelse av et klart definert vurderingsobjekt. Aktive og engasjerte skoleeiere og skoleledere er nøkkelaktører i en vellykket implementering av lærervurdering. De skoler eller kommuner som klarer å vektlegge læreres profesjonsutvikling i arbeidet med vurderingen, får bedre resultater enn de som lar arbeidet bli for dominert av skjema, kontroll og byråkrati.

Former for lærervurdering som kan ha positiv innvirkning på skolens kvalitet

Resultatene fra den systematiske kunnskapsoversikten viser at følgende fire forutsetninger må være på plass for at lærervurdering skal kunne bidra til god kvalitet i skolen:

- Metodekompetanse
- Medvirkning, ansvar og tillit
- Tydelighet og enkelhet
- Ansvarsplassering, dialog og oppfølging

Kunnskapshull

Kunnskapsoversikten har avdekket at det internasjonalt er forsket lite på praksiser hvor elever vurderer lærere. Videre finnes det lite empirisk forskning om hvilke praksiser lærere anser som virkningsfulle, og hva som kjennetegner god undervisning. Forskning som kombinerer resultatdata og observasjonsdata. ser ut til å gi nye perspektiver og verdifull informasjon, og det trengs flere slike undersøkelser. Forskning på vurderingspraksiser i Norge kan gi verdifull innsikt for utvikling av nasjonale tiltak.

1 Innledning

Lærervurdering er et aktuelt tema i norsk skole- og utdanningsdebatt. Det skyldes blant annet at en OECD-rapport, som analyserte arbeidet med kvalitetsvurdering i Norge (Nusche m.fl. 2011), foreslo at lærervurdering bør inkluderes som en av komponentene i det nasjonale systemet for kvalitetsvurdering. I tillegg viser resultatene fra den internasjonale TALIS-undersøkelsen at tilbakemeldingskulturen i norsk skole er svak sammenlignet med andre land, og at norske lærere i mindre grad enn lærere i andre land får tilbakemeldinger på jobben de gjør. I regjeringsplattformen fra Sundvollen, 7. oktober 2013, s. 56, står det dessuten: «Regjeringen vil la elevene i den videregående opplæringen evaluere undervisningen»[1].

Som et ledd i arbeidet med oppfølgingen av dette har Kunnskapsdepartementet i brev av 10.10.13 bestilt en systematisk kunnskapsoversikt om temaet lærervurdering fra Kunnskapssenter for utdanning (Vedlegg 1). I brevet het det blant annet at begrepet lærervurdering ikke er innarbeidet i norsk utdanningspolitikk, og at det sannsynligvis er store variasjoner i hvor presise og nyttige tilbakemeldinger lærerne får fra elever, foresatte, skoleledelse og skoleeier.

Kunnskapsoversikten skulle omfatte alle former for lærervurdering og inkludere kvalitativ og kvantitativ forskning, artikler om formativ og summativ vurdering, prosesskvalitet og resultatkvalitet. Departementet la til grunn at det jevnlig i prosjektperioden (1. november 2013 – 1. april 2014) skulle avholdes møter mellom departementet, arbeidsgruppen for lærervurdering i GNIST-samarbeid og Kunnskapssenter for utdanning.

Etter å ha gjort prøvesøk i litteraturen, sendte Kunnskapssenter for utdanning en prosjektbeskrivelse til Kunnskapsdepartementet 5. november 2013 (Vedlegg 2). Her formulerte Kunnskapssenteret følgende scope for kunnskapsoversikten:

Hvilke former for lærervurdering kan ha positiv innvirkning på skolens kvalitet?

For å svare på problemstillingen har Kunnskapssenter for utdanning utarbeidet en systematisk kunnskapsoversikt som sammenfatter og syntetiserer resultatene fra forskning om lærervurdering publisert etter 2009. Kunnskapsoversikten begynner i kapittel 2. Innledningsvis beskrives kort hva lærervurdering kan være og hvilke formål den kan tjene.

1 **Politisk plattform for en regjering utgått av Høyre og Fremskrittspartiet, Oslo 2013**
http://www.regjeringen.no/pages/38500565/plattform.pdf

1.1 Lærervurdering som praksisform og forskningstema

Lærervurdering har i hovedsak to formål. For det første er det former for vurdering som brukes til å måle sider ved lærernes kompetanse og prestasjoner. For det andre er det vurderinger og tiltak som skal bidra til lærernes kompetanseutvikling. For å forstå de to formålene med lærervurdering er det viktig å forstå distinksjonen mellom det å *måle* og det å *utvikle*. Vurdering som tar sikte på å *måle resultater* omtales i forskningslitteraturen som *summativ* vurdering (vurdering av det som er oppnådd i form av kompetanse eller prestasjoner). Det vil si at vurderingen tar sikte på å undersøke om lærerne arbeider på måter som bidrar til å innfri krav til de kunnskaper, ferdigheter og kompetanser som elevene skal ha oppnådd. Den skal først og fremst gi myndigheter og allmennheten informasjon om kvaliteten på læreres arbeid, og har en kontrollfunksjon. Lærervurdering som tar sikte på å *utvikle* lærernes kompetanse omtales i forskningslitteraturen som formativ vurdering. Det vil si at vurderingen har til hensikt å legge til rette for lærernes profesjonsutvikling ved å identifisere styrker og svakheter (Isoré 2009). Normalt har imidlertid begge formene for vurdering, altså både summativ og formativ, som mål å *forbedre* prestasjoner og resultater. Man måler for å få informasjon som man kan bruke til å forbedre. De fleste systemer for lærervurdering har derfor som ambisjon å fungere både formativt og summativt.

Forskerne er enige om at lærervurdering som praksis fortsatt er sporadisk og spredt, og at dette kjennetegner de fleste forskningsprosjektene som har undersøkt praksisen. Det ser også ut som om det er vanskelig å forankre vurderingsaktivitetene i organisasjoner og praksis og få til god sammenheng i systemer for lærervurdering. Ideelt sett skal systemer for lærervurdering kombinere og sette sammen ulike metoder og modeller for lærervurdering som til sammen skal si noe om kvaliteten på lærernes arbeid. Derfor er det vanskelig å gi en enhetlig definisjon av lærervurdering. Isoré (2009) viser for eksempel at i de ulike OECD-landene brukes metoder som klasseromsobservasjon, mappevurdering, intervju med lærere, testing av lærere, elevresultater og ulike spørreskjema.

I den internasjonale debatten om lærervurdering er det mulig å identifisere tre tydelige stemmer: 1) forskere som bringer inn den forskningsbaserte kunnskapen, 2) lærere og lærerorganisasjoner som representerer praksisfeltets erfaringsbaserte kunnskap og 3) myndigheter som utformer politikk. Det er ulike motiv, ønsker og formål knyttet til innføring av lærervurdering, både lokalt og nasjonalt, for eksempel om den skal være formativ, summativ eller en kombinasjon av disse.

I 1979 publiserte Benjy Levin en litteraturgjennomgang av forskningen om lærervurdering[2]. Han har identifisert arbeider som kan dokumentere effekt, pålitelighet og gyldighet av de metodene som brukes. Etter en inngående kvalitetsvurdering satt han igjen med 36 artikler som ble inkludert i oversikten, og finner at feltet preges av mange meninger og lite dokumentasjon. Videre er forskningen på de ulike tiltakene for lærervurdering som er satt i gang både spredt og tilfeldig. Den viktigste konklusjonen han trekker av litteraturgjennomgangen er at det er viktig å involvere lærerne i utformingen av systemer for lærervurdering. Videre fremhever han at man ikke bør satse bare på én vurderingsmetode, men kombinere flere metoder og fremgangsmåter.

I 2009 skrev Olivia Little en rapport på oppdrag fra National Education Association[3] som representerer et policy- og praksisperspektiv og er mer erfarings- enn forskningsbasert. I likhet med litteraturgjennomgangen (Levin 1979), viser rapporten at de fleste lærervurderingspraksiser er lokalt utformet og at det er mange forskjellige varianter av dem. Videre slår rapporten fast at dagens praksiser for lærervurdering verken måler

2 Levin, B. (1979): Teacher Evaluation – a Review of Research, *Educational Leadership* 240-245 *http://www.ascd.org/ASCD/pdf/journals/ed_lead/el_197912_levin.pdf (lastet ned 09.03.14)*

3 Little, O. (2009): Teacher Evaluation Systems: The Window for Opportunity and Reform. *National Education Association https://www.google.no/#q=Olivia+Little+Teacher+evaluation+systems (lastet ned 09.03.14)*

lærernes arbeid på en tilfredsstillende måte eller klarer å etablere sammenheng mellom vurderinger og forbedringer av lærernes praksis. Et av argumentene i denne rapporten er at den største utfordringen handler om målinger – hvordan man bruker dem, hva som er det viktigste å måle og hvordan det skal måles.

1.2 Organisering og gjennomføring av arbeidet

Kunnskapssenter for utdanning har organisert arbeidet med den systematiske kunnskapsoversikten som et prosjekt med Kunnskapssenterets direktør som prosjektleder. Det ble opprettet en forskergruppe for å følge prosjektet. Gruppen bestod av følgende forskere: Trond E. Hauge, Therese N. Hopfenbeck, Astrid Tolo, Karen Hammerness og Jens-Christian Smeby.

Først gjennomførte Kunnskapssenter for utdanning elektroniske søk i ulike databaser for å identifisere relevante artikler. I arbeidet med systematiske kunnskapsoversikter brukes søkeord fra problemstillingen (eller scopet). Man går forutsetningsløst ut og søker etter litteratur som anvender søkeordene i tittel og sammendrag – uavhengig av artiklenes forskningsmetode og forskernes teoretiske ståsted. Utgangspunktet er empirisk, og målet er å finne mest mulig forskningslitteratur om et kjernetema som i denne rapporten er *lærervurdering*.

I arbeidet med systematiske kunnskapsoversikter brukes en programvare som er spesielt utviklet for å kunne håndtere (finne, sortere, kategorisere og lagre) store datamengder. Det søkes først og fremst etter artikler i tidsskrift med fagfellevurdering. Kvalitetskrav til artiklene er at de må ha en klart formulert problemstilling, gjøre rede for metodevalg og metodisk fremgangsmåte og at det må være sammenheng mellom problemstilling, funn, drøfting og konklusjon. Etter at de artiklene som skal inngå i rapporten er identifisert, skal forskningen systematiseres og syntetiseres. En systematisk kunnskapsoversikt gir dermed et godt grunnlag for å si noe om mangfoldet i hva som er gjort av tidligere forskning på et felt.

Ambisjonen er å ha identifisert det meste som er å finne om temaet før man avgjør hvilke artikler som skal inkluderes i kunnskapsoversikten. Dette er en noe annen fremgangsmåte enn det som er vanlig i litteraturreviews som forskere normalt gjennomfører. Hensikten med reviews er imidlertid i begge tilfellene å gi et bedre og mer robust kunnskapsgrunnlag på et felt. Her er det snakk om to fremgangsmåter som utfyller og beriker hverandre.

Et kvalitetskriterium i systematisk review- og syntesearbeid er transparens. I arbeidet med systematiske kunnskapsoversikter sikres transparens først og fremst ved at det opprettes en protokoll som spesifikt definerer søkestrategi og hvert påfølgende trinn gjennom hele prosessen. Protokollen inneholder også en beskrivelse av metodevalg, samt refleksjoner omkring styrker og svakheter ved innsamlingsopplegget. Protokollen er tilgjengelig elektronisk på Kunnskapssenterets hjemmeside (www.kunnskapssenter.no)

Artiklene, som ble identifisert gjennom de elektroniske søkene, ble gjennomgått på tittel og sammendrag av Kunnskapssenter for utdanning for å bedømme relevans etter et sett fastlagte kriterier. I etterkant av dette ble det avholdt et arbeidsseminar for forskergruppen ved OUCEA (Oxford University Centre for Educational Assessment) på University of Oxford 3.-4. februar 2014. Leder for GNIST-arbeidsgruppen, professor Eyvind Elstad, UiO, deltok også på arbeidsseminaret. Hele søkeprosessen som Kunnskapssenter for utdanning hadde gjennomført ble gjennomgått, og arbeid med kvalitets- og relevansvurdering av artiklene som foreløpig var inkludert, ble fordelt.

De fleste artiklene som ble identifisert i fagfellevurderte tidsskrift i perioden 2009-2013, og som har lærervurdering som tema, var fra engelskspråklige land, hovedsakelig fra USA. Ettersom historiske og kulturelle forutsetninger samt nasjonale lovverk og læreplaner gir rammer for handlingsrommet til lærere,

ledere og lokale skolemyndigheter, mente forskergruppen at rapporten bør ha en innledning om kvalitetsvurdering i Norge, norsk lærerutdanning og hva som forventes av norske lærere. I perioden 4. til 24. februar har forskergruppen lest og kvalitetsvurdert artikler. Basert på spørsmål som kom opp i løpet av seminardagene i Oxford, er det også gjennomført håndsøk og gjort nye søk i databasene. Dette resulterte i at et tjuetalls artikler ble inkludert for vurdering.

Det har vært avholdt tre møter med arbeidsgruppen som er nedsatt av deltakerne i GNIST: 28. november, 17. januar og 17. februar. I møtene har Kunnskapssenter for utdanning presentert søkestrategi og vurderinger som er gjort underveis i arbeidet, samt prinsipper for ekskludering og inkludering av artikler. Ulike spørsmål med relevans for arbeidet med å samle forskningskunnskap om lærervurdering har vært drøftet, blant annet om lærervurdering er et dekkende begrep når det er snakk om arbeid som skal støtte og fremme læreres profesjonslæring. Arbeidsgruppen har fått utkast til rapport til gjennomlesing og kommentar.

1.3 Avgrensninger og utfordringer

Mange som har mye kunnskap om læreryrket, om lærerutdanning, om undervisning, kvalitet i skolen og ledelse av pedagogiske prosesser, vil kanskje savne noen referanser eller bestemte klassiske artikler. En forklaring på litteraturutvalget i den systematiske kunnskapsoversikten som begynner i kapittel 2, er at inkluderte artikler er publisert i fagfellevurderte tidsskrift etter 2009.

Kjernetemaet i denne kunnskapsoversikten er lærervurdering. Slik oppdraget er formulert, skal rapporten også svare på hvilke former for lærervurdering som kan bidra til god kvalitet i skolen. Dette er et bredt tema, og det er vanskelig å finne artikler som svarer direkte på forskningsspørsmålet. Det ble derfor gjort omfattende søk for å fange opp relevante artikler. Utgangspunktet for den systematiske kunnskapsoversikten ble på over 12.000 artikler (se kapittel 2).

For å kunne løse oppdraget innen tidsfristen, er litteratursøkene avgrenset. Det er foretatt noen avveininger med hensyn til hvilke typer artikler som skal inkluderes i rapporten. Følgende hovedkriterier for inklusjon ble satt:

- Studier publisert etter 1. januar 2009
- Studier som omhandler lærervurdering
- Både kvalitative og kvantitative studier
- Både summativ og formativ vurdering
- Studier som omhandler grunnskolen eller videregående skole
- Studier som omhandler kvalitet i skolen

Kunnskapsoversikten omhandler følgelig ikke forskning knyttet til blant annet barnehager og høyere utdanning, elevvurdering (lærere som vurderer elever) eller lærerutdanning.

Den litteraturen som ble fanget opp gjennom søkene er engelskspråklig, med USA som et tyngdepunkt. Kunnskapssenteret har derfor valgt å skrive et innledningskapittel som drøfter problemstillinger knyttet til vurdering i Norge, ser på hva læreres arbeid består i og gir en kort historikk om utviklingen av et nasjonalt

system for kvalitetsvurdering. Noe av litteraturen som refereres i innledningen er fanget opp gjennom de systematiske søkene og lagt i kategorien «kontekst» (se kapittel 2). Den systematiske kunnskapsoversikten begynner med en metodepresentasjon i kapittel 2. Det vil si søkestrategi, inklusjons- og eksklusjonskriterier, innledende organisering og kategorisering av søkeresultatet samt begrunnelse for den narrative syntesemetoden som er benyttet. Selve syntesearbeidet følger i hovedsak fire trinn, der trinn 1 beskrives i kapittel 2. I kapittel 3 presenteres syntesens trinn 2, den forberedende syntesen. Her sammenfattes og aggregeres resultatene fra de inkluderte primærstudiene i nye kategorier. Dette arbeidet gjør det mulig å identifisere mønstre på tvers av artiklene og forberede grunnen for det tredje trinnet av den narrative syntesen, som presenteres i kapittel 4. Her blir funn fra artiklene lest på tvers for å avdekke spenninger og motsetningsforhold i materialet. Målet er å undersøke forhold som kan forklare hvorfor studier konkluderer forskjellig.

Det er en utfordring å oversette engelskspråklig forskningslitteratur om utdanningsspørsmål til norsk. I den systematiske kunnskapsoversikten inngår en rekke ord og begreper som vi mangler gode oversettelser for. I USA og England har vokabularet som brukes for å beskrive utdanningspraksis og vurdering av kvalitet i utdanning de siste tiårene blitt sterkt preget av ord og begreper fra økonomi og marked. På norsk mangler for eksempel en dekkende oversettelse av «high stakes testing» som innebærer at testresultater kan få store konsekvenser i betydningen at skoler kan bli nedlagt eller lærere miste jobben om de ikke klarer å forbedre elevenes læringsutbytte. Heller ikke finnes et godt norsk ord for «accountability» eller det som engelske utdanningspolitikere og forskere legger i «teacher effectiveness». Det er ikke vanlig å snakke om regnskapsplikt eller effektive lærere i Norge. På norsk har effektiv synonymer som *virksom* og *virkningsfull*, og konnotasjoner som *rask* eller *hurtig*. En effektiv person er følgelig en som får tingene raskt unna. For å betegne det samme bruker man på norsk betegnelser som ansvar og lærerdyktighet. I Norge er det heller ingen tradisjon for å snakke om *vellykkede* eller «*suksessfulle*» lærere («successful teachers»). Det vanlige er å si gode lærere. I deler av litteraturen, særlig der «Value-Added»-modeller omtales, har økonomer isolert enkeltfaktorer for å måle effekten hver av dem kan ha på elevenes læringsutbytte. Ved å identifisere og kvantifisere kausale sammenhenger kan man måle effekten av lærerens bidrag til elevenes læringsutbytte. I disse artiklene refereres det derfor faktisk til en lærer*effekt*.

I litteraturen brukes disse tre betegnelsene på lærervurdering: *teacher evaluation, assessment of teachers* og *teacher appraisal*. På norsk brukes vanligvis lærervurdering.

Til slutt diskuterer rapporten resultatene og knytter disse til kvalitet i skolen i lys av den norske konteksten. Forhåpentligvis kan rapporten inspirere til mer forskning på vurderingspraksiser.

1.4 Et tilbakeblikk på arbeidet med kvalitetsvurdering i Norge fra 1988

Hernes-utvalgets rapport, NOU 1988: 28 Med viten og vilje, ga anbefalinger om klarere arbeidsdeling, mer konsentrasjon og samarbeid mellom universiteter og høgskoler, og sa: «Utfordringen for norsk kunnskapspolitikk er at landet ikke får nok kompetanse ut av befolkningens talent. De resultater som nås, er ikke på høyde med de ferdigheter som kan utvikles». Sitatet er et eksempel på det Thomas Popkewitz kaller global circulation of ideas (Lindblad og Popkewitz 2004), for det er sterkt inspirert av rapporten A Nation at Risk, som The National Commission on Excellence in Education la frem i 1983, og slo fast at utdanningens mål er «to develop the talents of all to their fullest»[4].

På slutten av forrige århundre var store deler av OECD-området opptatt av at kunnskap skulle komme til uttrykk i kompetanse, som man både kunne måle og utvikle (Tolo og Lillejord 2009). Antakelsen var at kompetanse så å si utvider seg i bruk. Når en novise får mer erfaring, øker hennes kompetanse. Kompetanse kan man dele med andre, og det er mulig å flytte den fra en kontekst til en annen. Med tid og erfaring, og under de rette betingelser, kan nybegynneren bli en ekspert. Parallelt med utviklingen av en kompetansepolitikk og økt desentralisering, erstattet målstyring regelstyring i Norge[5]. OECD (Nusche m. fl. 2011) påpeker at i et desentralisert system med flere ledelsesnivåer og uklar ansvarsplassering, som det norske, er det utfordrende å etablere et godt system for innhenting av data i bred forstand, både det som kan telles og prosentueres og kvalitativ informasjon. For å kunne dimensjonere datainnhentingen, analysere den informasjonen data gir og legge realistiske planer for hvordan man kan forbedre de resultatene som foreligger, trengs dessuten metodekunnskap og -kompetanse.

Lærervurdering skal inngå som en komponent i det nasjonale kvalitetsvurderingssystemet. Ansvaret for kvalitetsutvikling ligger i kommuner og fylkeskommuner som, i følge OECD, har ulike forutsetninger for å følge opp skoler og lærere.

4 http://datacenter.spps.org/uploads/SOTW_A_Nation_at_Risk_1983.pdf (s. 14)
5 Stortingsmelding nr. 37 (1990-91): Om organisering og styring i utdanningssektoren

Fra skolebasert vurdering til nasjonalt kvalitetsvurderingssystem

I 1987/88 evaluerte en ekspertgruppe fra OECD det norske utdanningssystemet (OECD 1989). I perioden 9.-18. november 1987 hadde gruppen møter med sentrale politikere, fagforeninger og interesseorganisasjoner. De besøkte skoler, høyskoler og universiteter i Oslo, Bergen og Bodø og noterte seg at nordmenn er svært stolte over sine geografiske røtter og mener at alle må få leve og organisere seg slik de ønsker. Likhetsideologien er sterk, og hviler på en blanding av respekt for sentrale myndigheter, parlamentarismen, demokrati og sterkt lokalt selvstyre. Gruppen var overrasket over at det ikke fantes noe system for å følge opp nyansatte lærere og heller ingen prøvetid. De mente å observere en utbredt skepsis til kunnskap fra forskning og forskere blant de som de møtte under sitt besøk. De registrerte dessuten at norske ungdommer ikke var særlig opptatt av utdanning, trolig fordi de fikk like godt betalte jobber uten utdanning som med. Gruppen konkluderte med at det var for mange aktører og for små enheter i det norske systemet, og at det var et opplagt behov for et sentralt system for informasjonsinnhenting og kvalitetsvurdering.

Da norske politikere ble konfrontert med at de ikke hadde gode nok data og visste for lite om kvaliteten i utdanningssektoren, begynte diskusjonene om hvordan arbeidet med kvalitetsvurdering skulle gjennomføres. Det ble blant annet satt i gang et prosjekt (EMIL-prosjektet) som skulle utvikle en modell for målstyring og evaluering av den norske skolen (Granheim og Lundgren 1990). Prosjektet førte til en debatt hvor spørsmål om skolens nytteverdi møtte dypt forankrede dannings- og fellesskapsorienterte perspektiver på skolens funksjon i samfunnet. Et statlig system for kvalitetsvurdering stod mot verdien av lokalpolitiske initiativ og betydningen av å gi lærerprofesjonen tillit og handlingsrom (Roald 2010). Et system for kvalitetsvurdering ble nevnt i Stortingsmelding nr. 33 (1991-92): *Kunnskap og kyndighe*t, og Stortingsmelding nr. 47 (1995-96): *Om elevvurdering, skolebasert vurdering og nasjonalt vurderingssystem*, fastsatte som prinsipp at nasjonal vurdering skulle gi grunnlag for utviklingsarbeid og veiledning, sikre kvaliteten på opplæringen og medvirke til mer målrettet ressursbruk, gi informasjon til allmennheten og relevante styringsdata. Skolebasert vurdering ble presentert som metode for systematisk kartlegging, vurdering og rapportering til de instansene som har ansvar for å fastsette skolens rammevilkår. I regjeringens dokumentarkiv finnes en god oversikt over prosjekter i perioden 1992-1997[6].

I 1980-90-årene var det stor interesse for skolebasert vurdering både blant lærere, lærernes organisasjoner og forskere. I 1986 publiserte Tom Tiller en bok basert på et halvt års feltarbeid i en engelsk ungdomsskole.[7] Der så han på hvilke forhold ved skolen som påvirker lærere og lederes læring, og fant at selv om diskursen om skolen preges av samarbeid, fellesskap og integrasjon, kjennetegnes hverdagen ofte av isolasjon, individualisme, oppdeling og segregasjon. Hvis målet er å bruke vurdering for å få til vedvarende organisasjonsutvikling, må skolens organisasjonskultur utfordres. Egenvurdering skulle hjelpe skolene til å holde oppmerksomheten rettet mot det som er sentralt i skolens mandat, nemlig å utvikle barn og unge som «hele» personer – ivareta elevenes affektive, kreative og ferdighetsmessige utvikling i tillegg til den intellektuelle og kognitive. På slutten av 1980-årene tok Rådet for videregående opplæring initiativ til et nasjonalt skolevurderingsprosjekt[8] som la grunnlag for flere publikasjoner og rapporter.[9] Gjennom kontinuerlig egenutvikling (prosesser som tok sikte på å samle og kommunisere informasjon og kunnskap om skolens arbeid) skulle man også kunne styrke allmenhetens tillit til skolene.

---

6 ***http://www.regjeringen.no/en/dokumentarkiv/Regjeringen-Jagland/kuf/Rapporter-og-planer/1997/ Oversikt-over-prosjekter-i-Nasjonalt-vurderingssystem-1992---1997.html?id=422786 (lastet ned 27.02.14)***

7 **Tiller, T. (1986): *Den tenkende skolen, Om organisasjonsutvikling og aksjonslæring på skolens egne premisser.* Oslo: Universitetsforlaget.**

8 **Tiller, T. (1993): *Vurder selv. Skolevurdering i praksis.* Oslo: Universitetsforlaget**

9 **Monsen, L. og Tiller, T. (1991): Effektive skoler. Oslo: AdNotam; Ålvik, T. (1991): *Skolebasert vurdering – en artikkelsamling. Oslo: AdNotam; Hauge, T. E. (1991): Kompetanseutvikling i skolevurdering: erfaringer med intern skolevurdering i grunnskolen.* Sagene lærerhøgskole/Universitetet i Oslo**

Da Stortingsmelding nr. 47 (1995-96) *Om elevvurdering, skolebasert vurdering og nasjonalt vurderingssystem* ble lagt frem, var det mye som tydet på at Norge ville få et omfattende nasjonalt vurderingssystem. I 1997 la Moe-utvalget fram en innstilling som fulgte opp meldingen, men som også pekte på at beredskapen lokalt var dårlig. I Stortingsmelding nr. 28 (1998-99) *Mot rikare mål*, ble de klare anbefalingene fra Moe-utvalget moderert til en nasjonal strategi for å videreutvikle systematisk kvalitetsutvikling i skolen. Skolebasert vurdering ble forskriftsfestet (1997). Kommunene og fylkeskommunene fikk, som skoleeiere, ansvar for å se til at skolene faktisk gjennomførte egenvurdering av aktiviteten sin, og følge opp arbeidet. Det er imidlertid indikasjoner på at mange skoler, kanskje til og med så mange som halvparten, aldri vurderer sin egen aktivitet på systematisk vis[10] (Hopfenbeck og Lillejord 2013).

I 2001 fikk Norge, i likhet med andre land, det som har blitt omtalt som «PISA-sjokk»[11]. Dette satte fart i arbeidet med å få på plass et system for kvalitetsvurdering. I følge Rosenkvist (2010)[12] har følgende land nå nasjonale systemer for summativ vurdering: Australia, den franske delen av Belgia, Danmark, Frankrike, Ungarn, Island, Irland, Japan, Luxembourg, Mexico, Nederland, Norge, Portugal, Slovakia, Sverige og England. I 2000 så PISA særlig på lesing, og tallene viste at norske elever skåret under OECD-gjennomsnittet i sentrale ferdigheter[13]. I 2001 ble Kvalitetsutvalget nedsatt (av statsråd Trond Giske). Utvalgets mandat og sammensetning ble noe justert da Kristin Clemet høsten 2001 overtok som Utdannings- og forskningsminister. Kvalitetsutvalget leverte to rapporter: *Førsteklasses fra første klasse* (2002) og *I første rekke* (2003). Utvalgets viktigste bidrag var innføring av årlige nasjonale prøver og forslag om en *Kvalitetsportal* (Skoleporten) som skulle presentere resultatene fra den årlige kvalitetsmålingen. Nasjonale prøver ble iverksatt våren 2004 og justert under Stoltenberg II (2005-2013). Skolereformen Kunnskapsløftet (K 06) ble innført i 2006 med disse hovedtrekkene: modulbaserte læreplaner med kompetansemål for hele grunnopplæringen og innføring av fem grunnleggende ferdigheter: digitale ferdigheter, muntlige ferdigheter, å kunne lese, å kunne regne og å kunne skrive. Læreplaner med kompetansemål forener kunnskaper og ferdigheter og handler om kunnskap i bruk.

I 2011 var igjen en gruppe fra OECD i Norge – denne gangen for å undersøke det norske systemet for kvalitetsvurdering (Nusche m. fl. 2011). Hensikten med denne gjennomgangen var å se nærmere på hvordan vurderingssystemer kan brukes til å forbedre kvalitet og effektivitet i utdanningssektoren og sikre elevene like muligheter. Det foreligger en synteserapport fra 2013, Synergies for Better Learning, som sammenfatter landrapportene fra 28 land, inkludert Norge (OECD 2013). Her ser OECD på hvordan de ulike komponentene (elevvurdering, lærervurdering, skolevurdering og systemvurdering) i de ulike landenes kvalitetsrammeverk bidrar til å øke elevenes læringsutbytte. Anbefalingen er at Norge bør styrke lærernes kompetanse i å tolke og følge opp resultater fra nasjonale prøver og andre former for elevvurdering, samt styrke lærernes kompetanse i formativ vurdering. Skolelederne, som har ansvar både for lærervurdering og skolevurdering, må bli flinkere til å gi gode tilbakemeldinger, veilede staben og bruke data på en måte som er nyttig for skoleutvikling. Samtidig må skolelederne få god tilbakemelding fra skoleeier, som har ansvar for ekstern vurdering av den enkelte skole og må lære å bruke informasjon fra vurderinger i sine beslutningsprosesser. I mange deler av Norge er det lite realistisk å tro at lokale skoleeiere vil være i stand til å utvikle robuste lokale kvalitetssikringssystemer og bruke disse i oppfølgingen av skolene, blant annet fordi mange små kommuner har få ansatte med kompetanse til å løse disse oppgavene. Det er mer realistisk å se for seg regionale løsninger, for eksempel fylkesvis. I det hele tatt er OECD bekymret over at det i Norge ikke er en omforent enighet mellom de ulike nivåene (skoleeier-skoleleder-lærer) om hva som skal være hensikten med systemet for kvalitetsvurdering.

10 Kristin Clemet i Stortingets spørretime, onsdag den 27. februar 2002 kl. 10.
11 Bergesen, O. K. (2005): *Kampen om Kunnskapsskolen*. Oslo: Universitetsforlaget; Sjøberg, S. (2014): PISA-Syndromet: Hvordan norsk skolepolitikk blir styrt av OECD *Nytt Norsk Tidsskrift* nr. 1 (31) 30–43
12 Rosenkvist, M.A. (2010) «Using Student Test Results for Accountability and Improvement: A literature Review» OECD Education Working Papers, No. 54, OECD. http://dx.doi.org/10.1787/5km4htwzbv30-en
13 OECD (1999) *Measuring Students' Knowledge and Skills: A new Framework for Assessment*. Paris: OECD.

I et eget kapittel i landrapporten til Norge diskuterer OECD lærervurderingens plass i det norske kvalitetsvurderingssystemet (Nusche m. fl., 2011 73-90). De sier at mens det er klare nasjonale forventninger om at lærerne skal få tilbakemeldinger på den jobben de gjør, er ikke dette lovregulert. Mangelen på karriereveier og anerkjennelse av godt utført arbeid motiverer ikke lærere til å forbedre innsatsen sin. Heller ikke er det utviklet nasjonale prestasjonskriterier eller referansestandarder. Den enkelte skoleeier står fritt til å utvikle sitt eget system for lærervurdering, men få av dem har systematisert dette arbeidet. Den vanligste formen for tilbakemelding til norske lærere er den årlige medarbeidersamtalen. Dermed er ikke norske lærere sikret tilbakemelding fra arbeidsgiver på den profesjonelle delen av arbeidet sitt. Det er ingen rutiner på plass for å sikre at alle lærere får sin praksis observert eller tilbakemelding på om de har hatt en god profesjonsutvikling. OECD anbefaler derfor at Norge gjør lærervurdering til en del av kvalitetsvurderingssystemet og understreker at vurdering av læreres arbeid har lite for seg med mindre man kobler vurderingen til profesjonslæring.

OECD er nå opptatt av implementeringsutfordringer i komplekse systemer[14] og har satt i gang prosjektet Governing Complex Education Systems. Norge bidrar i denne studien med en evaluering av satsingen *Vurdering for læring* (Hopfenbeck m. fl. 2013). Den norske rapporten viser at det fortsatt er store forskjeller mellom norske kommuner når det gjelder hva slags vurderingskompetanse de har utviklet og hvordan de arbeider med implementering. Kommuner som lykkes med å innføre ny vurderingspraksis kjennetegnes av en høy grad av tillit og dialog mellom kommunal ledelse, skoleledere og lærere. Det er også variasjoner i hvilken vurderingskompetanse lærere har, og hvordan de forstår den nye vurderingsparagrafen. Dette bekreftes i en rapport fra prosjektet Forskning på individuell vurdering i skolen,[15] hvor analyser av data fra åtte skoler viser stor variasjon mellom skolene med hensyn til hvor langt de har kommet i vurderingsarbeidet. Lærernes vurderingspraksis kan variere betydelig på en og samme skole, og en av de største utfordringene for skolene er å samle seg om en felles forståelse av hva som er god vurderingspraksis.

14 Fazekas, M. and Burns, T. (2012): Exploring the Complex Interaction Between Governance and Knowledge in Education. OECD, Paris, Education Working Papers, No. 67.
15 Sandvik, L. V. og Buland, T. (2013) *Vurdering i skolen.* Operasjonaliseringer og praksiser. Delrapport 2. NTNU

1.5 Generelle prinsipper for vurdering

Gjennom den nye forskriften til opplæringsloven §§ 3-2 og 3-11 har elever fått rett til medvirkning i vurderingsarbeidet[16]. Forskriften sier at elevene lærer best når de

1. forstår hva de skal lære og hva som er forventet av dem,
2. får tilbakemeldinger som forteller dem om kvaliteten på arbeidet eller prestasjonen,
3. får råd om hvordan de kan forbedre seg, og
4. er involvert i eget læringsarbeid ved blant annet å vurdere eget arbeid og utvikling.

Disse fire prinsippene for vurdering har dannet grunnlag for en omfattende satsing på vurdering i Norge gjennom programmet *Vurdering for læring* (2010–2014)[17], som i stor grad bygger på arbeidet til den britiske forskergruppen Assessment Reform Group og spesielt Black og Wiliam (1998; 2009), Stobart (2008) og Hayward og Spencer (2010). Prinsipper for vurdering som er utviklet av Assessment Reform Group (2006) kan betraktes som allmenngyldige på den måten at de like gjerne kan styre arbeidet med god vurderingspraksis i en skole som i et klasserom, kommune eller bedrift – og de er like gyldige for god lærervurdering som de er for god elevvurdering (og for vurdering på organisasjonsnivå). Det er dessuten konsensus blant forskere om disse prinsippene.

- De ressursene som trengs for å gjennomføre vurderingen, må stilles til rådighet (ekspertise, økonomi, tid) og innsatsen må balanseres mot det man forventer å få ut av aktiviteten
- Vurderingen må planlegges og gjennomføres på en slik måte at den eller de som blir vurdert opplever resultatet av vurderingen som gyldig og troverdig
- Vurderingen må avgrenses til og konsentrere seg om bestemte sider ved arbeidet, men likevel ta hensyn til at det som vurderes inngår i en større sammenheng
- Vurdering bør gjennomføres på måter som ikke bare måler prestasjoner, men som også generelt bidrar positivt til arbeidet som utføres i skolen og styrker læringsmiljøet
- Vurdering har generelt stor innvirkning på praksis i et felt, så vurderingen må planlegges og gjennomføres på en slik måte at uønskete virkninger av vurderingen minimeres

Vurdering er en praksis som hele tiden kan bli bedre. Erfaring fra arbeid med vurdering viser at informasjon som samles inn for ett formål ikke nødvendigvis kan brukes til flere formål. Videre må ikke vurderingssystemer bare basere seg på kvantitative data (summativ), men hente informasjon fra alle tilgjengelige kilder i skolen og i skolens omgivelser. Man kan ikke vurdere alt på en gang, og hele prosessen rundt vurderingen må være transparent. Ingen skal oppleve vurderingen som urimelig eller nedverdigende. Vurderingene som blir gjort må dessuten dokumenteres, slik at alle involverte kan kjenne seg igjen og stole på resultatet.

Flere tiår med forskning har konkludert med noen kjøreregler som gjelder for god vurderingspraksis. For det første må man bli enige om hva som er gjenstanden for vurderingen – hva skal vurderes? Det andre er *hvordan* det skal vurderes og målet med vurderingen. I dag er det bred enighet om at vurderingen ideelt sett skal ha forbedring som mål. Når forbedring er målet med vurderingen, får det konsekvenser for *hvilke* metoder som brukes og hvordan kunnskapen som genereres gjennom evalueringen brukes i forbedringsarbeidet. Når disse enkle kjørereglene ikke blir fulgt, kan resultatet bli uheldige og til og med ødeleggende praksiser (Stobart 2008).

16 http://lovdata.no/dokument/SF/forskrift/2006-06-23-724
17 http://www.udir.no/Vurdering-for-laring/

Vurdering er alltid vurdering av noe

Vurdering er alltid vurdering av noe i forhold til noe annet. Det vil si at selve vurderingsobjektet, altså hva det er som skal vurderes, må være klart definert. Når det finnes en klar standard å vurdere i forhold til, kan det være enkelt å anslå om det som vurderes møter standarden eller ikke. Skal man lage skruer av en viss størrelse og fasong, er det viktig at hver skrue er nøyaktig slik den skal være og passer til det formålet den skal tjene. Når man skal vurdere et komplekst arbeid, mangler imidlertid ofte både standarder og klare kriterier. For eksempel mangler klare og omforente indikatorer for hva som kjennetegner god undervisning. Forarbeidet, altså den analysen som blir gjort før man setter i gang med vurderingen, målet og planer for oppfølging av vurderingen, er svært viktig (Lillejord og Hopfenbeck 2013; Hopfenbeck og Lillejord 2013). I all vurdering må den som skal vurdere på forhånd tenke gjennom hva man skal gjøre med eventuelle problemer som avdekkes eller oppstår i prosessen, legge en plan og sette av nødvendige ressurser – både personal og økonomi.

Både håndverkere og profesjonsutøvere har standarder for sitt arbeid – men samtidig er det noe som gjør enkelte håndverkere og profesjonsutøvere bedre enn andre – uten at det er umiddelbart enkelt å si nøyaktig hva dette er. Delvis skyldes det at det kan være mulig å lage standarder for deler av jobben, mens det er vanskeligere å standardisere *hvordan* yrkesutøvelsen skal skje. Her er det forskjell på enkle, tekniske arbeidsoppgaver og avansert profesjonsutøvelse. Hvis det er en forventning om at lærerne skal bli innovative og bidra til å skape nye praksisformer i skolen, bør det ikke lages standarder for alle deler av læreryrket. Her gjelder det om å finne en balanse mellom hvor standarder kan være tjenlige og når de kan virke sementerende.

Vurdering handler om å gradere kvalitet

I arbeid med vurdering, tilskrives noe verdi, det fratas verdi eller verdien av det nyanseres. Vurdering handler derfor også om gradering av kvalitet. I de fleste sammenhenger er imidlertid ikke kvalitet et entydig begrep. Det som representerer god kvalitet for én person kan være middels eller til og med dårlig kvalitet for en annen. I hverdagslivet er det vanlig å snakke om mer eller mindre vakker arkitektur, god og dårlig service, dyktige eller kjedelige forelesere. Her dreier det seg om produkter, tjenester og opplevelser. Grunnen til at mennesker er uenige om hva som er godt og dårlig er at vurdering av kvalitet baserer seg på menneskelig skjønn. All vurdering innebærer et visst element av skjønnsutøvelse, men de som vurderer kan på forhånd bli enige om hva man skal se på, hva man skal se etter og hva man skal gjøre med *resultatet* av vurderingen. Alle vurderingsprosesser innebærer overraskelsesmomenter – ting er ikke alltid slik man tror. På sitt beste er derfor vurdering en læreprosess både for den som vurderer og for den eller de som blir vurdert.

Kvalitet kan fastsettes objektivt gjennom entydige standarder. Som regel har imidlertid alle kvalitetsvurderinger innslag av subjektivitet. Det er forskjell på å vurdere kvaliteten på produkter og prosesser, og vurdering av prestasjoner er normalt prosessvurdering. Derfor er det flere dommere i idretter som hopp og kunstløp. Selv om dommerne følger kriterier, vurderer de ofte forskjellig og gir ulike karakterer fordi prestasjonsvurdering inneholder elementer av skjønn. Også ved skjønnsutøvelse kan personer imidlertid ofte være enige om hva som holder bedre kvalitet enn noe annet. Det er imidlertid viktig å ha klart for seg disse to sidene ved vurderingsarbeidet. Noe kan standardiseres og vurderes ved hjelp av kriterier og indikatorer. Noe er det vanskeligere å standardisere, men man kan likevel snakke seg frem til enighet om hvilken kvalitet eller verdi det har.

Vurdering har alltid et formål

Selv om formålet med vurderingen ikke alltid er like klart formulert, er det alltid et formål med en vurdering. Man ønsker å oppnå noe. I verste fall blir en vurdering satt i gang uten tanke for hva man ønsker å forandre fra eller til. Uten en tilstandsanalyse (ståstedsanalyse) og en plan for hva som kan eller bør forandres i hvilken retning og i hvilket tempo, kan man miste kontrollen underveis i vurderingsarbeidet. Det er lett å undervurdere hvor vanskelig det er både å samle og analysere data. Mange samler inn mer enn de trenger, og det finnes mange eksempler på dårlige vurderingspraksiser både i skole og arbeidsliv. Aviser rapporterer om tilfeller av vurdering på arbeidsplassen som de ansatte opplever som lite produktive, til og med direkte ydmykende, og vi hører om barn og voksne som har måttet slite med dårlig selvbilde etter å ha blitt utsatt for tankeløs vurdering. Dårlig vurderingspraksis kan føre til dårlig læringsmiljø i skoler og kommuner. Før man etablerer et system for vurdering, er det derfor viktig på forhånd å bestemme seg for hva man vil oppnå. Systemet som etableres, må speile målet for vurderingen. Hvis målet er utvikling av lærerprofesjonalitet og god skoleutvikling, må systemet utformes slik at det blir mulig å nå dette målet.

Vurdering skal prinsipielt føre til forbedring

At vurdering skal føre til forbedring høres kanskje selvsagt ut for mange. Man vurderer ikke bare for å kontrollere, men for å få et bilde av hvordan det står til på et område. Hensikten er at man skal gjøre noe med eventuelle problemer som avdekkes. Derfor er det viktig å legge en plan for hva som skal skje etter at vurderingen er gjennomført. En vurderingsprosess genererer både resultatdata og prosessinformasjon. Resultatdata kommer i form av tall som kan prosentueres og fremstilles i tabeller og grafer. Prosessdata er det vanskeligere å håndtere. Det er informasjon som kommer underveis i prosessen, og det kan være vanskelig å få synliggjort den kunnskapen som genereres i prosessen uten en plan for hvordan den skal dokumenteres underveis.

Forskningslitteraturen skiller mellom summativ og formativ vurdering. Summativ vurdering handler om å måle resultatet av et arbeid, mens formativ vurdering er ulike tiltak som skal forbedre resultatet. Både summativ og formativ vurdering har imidlertid som mål å *forbedre* prestasjoner og resultater. Forskningen som er gjennomført de siste tiårene er imidlertid entydig på at de beste resultatene får man der ledere og lærere arbeider systematisk med formativ vurdering, det vil vi bruker summative resultater formativt (Black og Wiliam 1998).

For å kunne bruke informasjon som fremkommer gjennom vurdering til å forbedre, må man ha både kunnskap om det som skal vurderes, metodekompetanse og kompetanse i å lede mennesker i endringsprosesser. Å bruke informasjon som fremkommer gjennom vurdering til å forbedre praksis, er et analytisk og relasjonelt arbeid. Aller viktigst er det å etablere enighet mellom alle involverte om hva som er problemet og hvordan det skal løses. De som er tettest på problemene, må involveres i prosessen som handler om å bli enige om hva man skal gjøre. Det er en lederoppgave å sørge for forankring.

Vurderingskunnskapen skal brukes

Når man har vurdert, må den som har blitt vurdert rask få konkret og tydelig tilbakemelding. Tilbakemeldingen må være slik at den det gjelder forstår hva som må gjøres. Den som gir tilbakemelding må undersøke om den er riktig fortått og forsikre seg om at den som har fått tilbakemeldingen er i stand til å endre praksis. Vurderingskunnskapen må videre komme hele organisasjonen til gode ved at ledelsen bruker

det som fremkommer gjennom vurderingen til å planlegge skolens utviklingsarbeid. Det forventes at lærere skal vurdere elever med klokskap og profesjonalitet. Det kan være en god rettesnor også for arbeidet med skole- og lærervurdering.

Det er mye å lære fra andre land som har innført systemer for lærervurdering (se kapittel 3.1, som presenterer forskning fra Chile, Kina, Belgia og Portugal). Samtidig er det viktig å anerkjenne særegne (historiske og kulturelle) trekk i det norske samfunnet som bidrar til vårt utdanningssystem og de evaluerings- og vurderingstradisjonene som rår her.

1.6 Læreres arbeid

I alle spørsmål om vurdering, både skole-, elev- og lærervurdering, er en viktig del av forarbeidet å spørre seg selv *hva skal vurderes?* Det er ikke mulig å vurdere «alt», og hvis man begynner for ambisiøst, kan det bli store datamengder og mye å holde rede på i analyseprosessen. Et vanlig problem er å begynne datainnsamlingen for tidlig og ofte før det er avklart hva man særlig trenger svar på. Når det gjelder lærervurdering, er et sentralt spørsmål hva lærere utdannes til. Hvilke lærerkompetanser fremheves i lærerutdanningen – hva sier forskriftene at lærere skal kunne og kunne gjøre? Det er viktig å skaffe seg en oversikt over hva lærere faktisk gjør og hva det forventes at de skal gjøre.

I følge Imsen (2009) definerte normalplanene av 1939 arbeidsoppgavene for skoleinspektør, skolesjef, skolestyrer og lærer gjennom forskrifter. Når læreren hadde autoritet i dette systemet, skyldtes det at læreren hadde mer utdanning enn folk flest, samt en sterk tro på at kunnskap ville bringe samfunnsutviklingen videre. Dette kunnskapshegemoniet ble gradvis redusert i løpet av 1970- og 80-tallet, og i 1997 kom nye læreplaner som definerte både innhold og arbeidsformer og utfordret prinsippet om lærernes metodefrihet. På 2000-tallet har lærerne fått mange nye oppgaver. De skal ikke bare undervise i fag og veilede som før, de skal i tillegg arbeide i team, involvere elevene i utforming av læreplanmål, tolke kvantitative og kvalitative data, utvikle nye praksis- og evalueringsformer, vurdere forskningsresultater og delta i skolens organisasjonsutvikling.

OECD bemerker at norske lærere har stor autonomi, generelt høy tillit i befolkningen (Nusche m. fl. 2011 s. 77), og konkluderer med at det ser ut til å være enighet i Norge om at lærere skal møtes med tillit, ikke med kontroll. Norske lærere står normalt fritt til å avgjøre innhold i undervisningen samt å velge hvilke undervisningsmetoder og læremidler de skal bruke. I følge OECD er en konsekvens av at lærere møtes med tillit og nyter stor autonomi i sin yrkesutøvelse at de setter pris på at skoleledere tar seg tid til å gi dem tilbakemeldinger på arbeidet de gjør. Selv om det er nasjonale krav om at skoleledelsen årlig skal gi lærere faglige tilbakemeldinger, sier så mange som 26,2 % av de som svarer i TALIS at de aldri får slik tilbakemelding. Norske skoleledere bruker mer tid på administrasjon enn på pedagogiske oppgaver som veiledning og tilbakemeldinger. Det som finnes av lærervurdering i Norge er overlatt til enkelte skolers eller kommuners initiativ og avhenger av hvordan ledere oppfatter det å være skoleleder eller av skolens tradisjon for skolevurdering (op. cit. s. 81).

Rammeplanene for lærerutdanning gir en viss pekepinn på hva myndighetene mener at læreres arbeid går ut på. Kravene er omfattende og til dels ulikt definert for de ulike utdanningene (1-7, 5-10 og 8-13). Forventningen om at lærere skal forberedes til å delta i skolens og profesjonens utvikling står sterkt, og mye av det en nyutdannet lærer skal bidra med går langt ut over det som skjer i klasserommet. Felles for de tre lærerutdanningene er at alle lærere skal beherske et bredt repertoar av undervisningsmetoder, kunne

kommunisere og samarbeide med elever, kollegaer og foreldre, samt delta i utvikling av skoleorganisasjonen og bidra til profesjonsutvikling. Det er krav om at den som skal bli lærer i ungdomsskole og videregående skole (lektorutdanningen) skal ha inngående kunnskap om vitenskapelige problemstillinger, mens den som skal arbeide på trinn 1-7 skal ha kunnskap om grunnleggende ferdigheter. Det er ulike forventninger til lærere på ulike klassetrinn.

I lærervurdering er det viktig å stille spørsmål som: Hvilke sider ved lærernes arbeid ønsker man å få mer informasjon om? Hvis svaret på dette spørsmålet er *undervisningen*, er det nødvendig å undersøke hva som kjennetegner undervisningssituasjonen for lærerne i den skolen eller kommunen hvor vurderingen skal gjennomføres. Det kan handle om hvordan det hovedsakelig undervises – i team, to-lærerordninger eller med assistenter. Det er også nødvendig å undersøke hvilke typer av undervisning som faktisk foregår i skolen. Her er det neppe én praksis som gjelder, men mange ulike. En lærers undervisning er ikke bare det som skjer i klasserommet eller i gruppen, men både planlegging av undervisningen, gjennomføring og vurdering – altså tre faser, som hver for seg er komplekse og kompliserte aktiviteter som forutsetter og griper inn i hverandre (Munthe 2013; Klette 2013; Hopfenbeck og Lillejord 2013). Å si at undervisning er komplekse prosesser er ikke det samme som å si at de (dermed) ikke kan vurderes. All vurdering må imidlertid følges av spørsmål om man faktisk vurderer det man har til hensikt å vurdere.

I Norge er relasjonene mellom arbeidsgivere og arbeidstakere normalt preget av en høy grad av tillit og stor grad av arbeidstakermedvirkning. I likhet med alle norske arbeidstakere er lærere underlagt arbeidsmiljøloven[18]. Lovens overordnede formål (§ 1-1) er å sikre et inkluderende arbeidsmiljø som gir grunnlag for en helsefremmende og meningsfull arbeidssituasjon med full trygghet mot fysiske og psykiske skadevirkninger. Det presiseres at for eksempel bruk av prestasjonslønn skal være slik at arbeidstakere ikke utsettes for uheldige fysiske eller psykiske belastninger (§ 4-1) og at arbeidet legges til rette slik at arbeidstakeres integritet og verdighet ivaretas (§ 4-3). Ellers er loven klar på at arbeidstaker skal sikres selvbestemmelse, innflytelse og faglig ansvar, samt muligheter for personlig og faglig utvikling gjennom sitt arbeid (§ 4-2). Som gruppe er lærerne også regulert av Opplæringsloven[19] som fastsetter at det er skoleeiers ansvar å sikre nødvendig kompetanseutvikling for sine ansatte i grunnopplæringen (§ 10-8). Lovverket samstemmer med kravene som blir stilt gjennom forskriftene om lærerens deltakelse i organisasjons- og profesjonsutvikling.

I alle yrker får noen arbeidsoppgaver mer oppmerksomhet enn andre. Når det gjelder lærere, er det tiden sammen med elevene i basisgrupper og klasser som vanligvis oppfattes som den viktigste delen av jobben. Men hvis man reduserer lærervurdering til kun det som skjer i klasserommet, mister man av syne alt arbeidet som ligger *rundt* undervisningen. Det kan handle om individuell og kollektiv elevkontakt, det som skjer i friminuttene, før og etter skoletid, samtaler med kollegaer om elever, samtaler med foreldre, skoleledelsen, PPT, helsesøster eller annet støtteapparat. Dette er relasjonsarbeid som gir rammer for undervisningen og bidrar til læringsmiljøet i en klasse. Om man bare vurderer det den enkelte læreren presterer i en time eller en undervisningsøkt, risikerer man å overse kollegiets eller skoleledelsens bidrag til den enkeltes lærers prestasjon, altså skolefaktorens betydning for den enkelte. Det som et enkelt individ presterer er ofte et resultat av den samlede innsatsen i organisasjonen vedkommende arbeider i og er en del av. Ledelse, kollegialitet, skolekultur og materielle ressurser har betydning for den enkelte lærers muligheter for utfoldelse.

18 http://lovdata.no/dokument/NL/lov/2005-06-17-62 (lastet ned 24.02.14)
19 http://lovdata.no/dokument/NL/lov/1998-07-17-61 (lastet ned 24.02.14)

1.6.1 Hva bruker lærere arbeidstiden sin på?

I desember 2009 leverte et utvalg, ledet av Fylkesmann Kirsti Kolle Grøndahl, rapporten Tidsbruk i skolen (Tidsbruksutvalget 2009). Der kom det frem at lærere ønsker å bruke mer tid på undervisningsrelaterte arbeidsoppgaver, faglig oppfølging av elevene, faglige møter og kompetanseutvikling. De opplever at de bruker for mye tid på konfliktløsning, holde ro og orden, møter av ikke-faglig art, praktiske oppgaver, dokumentasjon rundt enkeltelever, oppfølging og kontakt med enkeltelever, lokalt læreplanarbeid, rapportering til skoleledelse og skoleeier samt kontakt med foreldre/foresatte og enkeltelever utenom undervisningen. Tidsbruksutvalget mente at fordi lærerens kompetanse og samspillet med elevene har størst betydning for elevenes motivasjon og læringsutbytte, må også forholdene legges til rette slik at lærerne kan bruke mest mulig av tiden på skolens kjerneoppgaver: Undervisning, vurdering og planlegging av undervisningen. De konkluderte med at det er behov for økt voksentetthet i skolen.

Blant de områdene utvalget mente det måtte satses på fremover, stod *god ledelse* øverst på listen, både skoleledelse og klasseledelse. En side ved lederskapet er at skoleeier må sørge for at kommunen har nødvendig skolefaglig kompetanse, nyansatte lærere må få oppfølging, andre yrkesgrupper må inn i skolen for å sikre lærerne tid til kjerneoppgavene, regelverket må bli tydeligere og dokumentasjonskravene må reduseres.

I en nyere artikkel som undersøker hva lærere bruker arbeidstiden til, tar Philipp og Kunter (2013) utgangspunkt i forskning som viser at lærere er i en risikogruppe for utbrenthet, noe som både har helsemessige implikasjoner og konsekvenser for jobbinnsatsen. De mener at det derfor er viktig å finne ut hva lærere bruker tiden sin på og hva de selv opplever som følelsesmessig utmattende. Når lærere blir spurt, oppgir de tidspress og arbeidsmengde som de største problemene i yrket. Dette samsvarer med arbeidstakerundersøkelser som viser at ansatte i ulike yrker rapporterer at det er viktig for jobbtrivsel å føle at man har kontroll over arbeidsdagen sin.

En oversikt fra OECD (2011) viser at undervisningsoppgaver utgjør ca. halvparten av læreres arbeidstid. Denne delen av lærernes arbeid er i mange OECD-land avtalefestet sentralt, mens det som ligger utenfor (for eksempel møter, ekskursjoner, prosjektarbeid, administrative oppgaver, oppfølging av enkeltelever, kontakt med PPT, foreldresamtaler, kollegasamarbeid, tolking av testresultater, kurs, etter- og videreutdanning) reguleres lokalt. Philipp og Kunter (2013, s. 5) har undersøkt et utvalg på 1939 tyske lærere fra 198 videregående skoler[20]. De fant at lærerne, i tillegg til den tiden som gikk med til å gjennomføre selve undervisningen, brukte nesten 32 timer per uke på ca. 11 undervisningsrelaterte arbeidsoppgaver. Den aktiviteten som tok mest tid var individuell timeforberedelse. Deretter kom retting av tester, tilbakemeldinger på elevenes hjemmearbeid og administrative oppgaver. Tilsvarende tall finner Gunter m. fl. (2005) i England og Bruno m. fl. (2012) i USA. De øvrige arbeidsoppgavene var dokumentasjon av elevenes prestasjonsnivå, elevkontakt, prosjektorganisering, veiledning, ekskursjoner, foreldremøter, faglig oppdatering, administrative oppgaver og deltakelse i skolens arrangementer. Interessant nok viste tallene at mer erfarne lærere brukte mer tid på administrative oppgaver og mer tid på å rette hjemmearbeid enn de yngre, og det var heller ingen stor aldersvariasjon i tallene som gjaldt hvilke arbeidsoppgaver lærerne rapporterte at de brukte tid på. Ulike undersøkelser viser at lærere i England, USA, New Zealand og Tyskland rapporterer at de arbeider ca. 50 timer per uke (Philipp og Kunter 2013, s. 2).

20 69,9 % av respondentene arbeider fulltid; aldersspennet er 25-65 år; de har fra 1-44 års erfaring og 51,3 % av utvalget er kvinner

Når spørsmålet er hvordan lærervurdering skal struktureres, er det nødvendig å tenke grundig gjennom *hva* som skal vurderes. Læreres arbeid er komplekst og omfattende, og det er vanskelig å si at én aktivitet, for eksempel den delen av undervisningen som kan observeres i klasserommet, ikke henger sammen med andre deler av lærerens jobb. Samtidig er det mulig å argumentere for at det arbeidet som skjer før og etter selve undervisningen bidrar til og understøtter selve undervisningen. Godt for- og etterarbeid er en forutsetning for god undervisning. Det er i tillegg viktig å ta hensyn til sentrale utviklingstrekk, slik at vurderingssystemet tar høyde for nye utdanningspraksiser. I morgendagens skole vil det utvilsomt være andre utfordringer enn i gårsdagens, og et system for kvalitetsvurdering må ta hensyn til at samfunnet forandrer seg. Den teknologiske utviklingen åpner for eksempel muligheter vi ikke kan forutse i dag. Samtidig som noe vil være stabilt, vil det være andre forventninger til lærerkompetanse i 2020 enn i dag.

Stortingsmelding nr. 30 (2003-2004), *Kultur for læring*, beskrives kompetanse som «hva man gjør og får til i møte med utfordringer» (s. 31). Definisjonen tar utgangspunkt i OECD-prosjektet DeSeCo[21], som beskriver kompetanse som evnen til å mestre komplekse utfordringer både i utdanning og yrkesliv. Kompetanse er ikke noe man lærer en gang for alle, men en sum av vedvarende kognitive og sosiale prosesser (Tolo 2011). Lærere forbereder seg på å undervise slik at elevene utvikler nøkkelkompetanser (key competencies) eller ferdigheter for det 21. århundre (21 Century skills) som får betydning i fremtidens skole.

- *Måter å tenke på*: Kreativitet, kritisk tenkning, problemløsning, kunne foreta valg i læringsprosessen
- *Måter å arbeide på*: Kommunikasjon og samarbeid
- *Arbeidsredskaper*: IKT og digitale ferdigheter (literacy)
- *Livsferdigheter*: Dannelse, kompetanse for livslang læring, personlig ansvar og sosialt ansvar

Her videreføres UNESCOs ambisjon om at skolen skal fremme internasjonal forståelse, toleranse, respekt for menneskerettigheter, demokrati, kulturelt mangfold og identitet. Derfor må elevene lære å forstå, lære å gjøre, lære å leve sammen og lære å være (Fevolden og Lillejord 2005).

21 DeSeCo (Definition and Selection of Competencies), se for eksempel Salganik, L. H. og D. S. Rychen (2003).

1.6.2 Kjennetegn ved god undervisning

Tema for denne kunnskapsoversikten er lærervurdering. I den forskningen som er identifisert om lærervurdering etter 2009, er det stor oppmerksomhet om lærernes undervisning, det vil si at læreres undervisning er et sentralt vurderingsobjekt. Det finnes svært mye forskning om undervisning; mye handler om hva som kjennetegner god undervisning og hvordan man skal fremme god undervisning. Det kunne ha vært laget en egen kunnskapsoversikt av denne litteraturen. I den systematiske kunnskapsoversikten er det imidlertid inkludert forskning hvor undervisning er et vurderingsobjekt.

Det er ingen sterk tradisjon i utdanningssektoren eller blant utdanningsforskere i retning av å utvikle standarder for undervisning. Man har vært mer opptatt av *prinsipper* for god undervisning. Flere år med klasseromsforskning har imidlertid ført til at stadig flere forskere interesserer seg for spørsmålet om hva som kjennetegner gode lærere, altså hva dyktige lærere *gjør* (Klette 2013). I mangel på entydige og omforente standarder for eller indikatorer på hva som kjennetegner godt lærerarbeid, fant tre forskere ut at de ville velge ut noen lærere som hadde fått nasjonale priser for sin undervisning (Grant m. fl. 2013). De ville identifisere fellestrekk ved disse lærernes praksis som gjorde dem til så dyktige undervisere at de vant priser for arbeidet sitt. De undersøkte i to forskjellige land (Kina og USA), og fant frem til 16 kinesiske og 16 amerikanske prisbelønte lærere[22].

For å kunne gjøre en fornuftig sammenligning av lærere som praktiserer i to så forskjellige kulturer som Kina og USA, så forskerne både på lærernes praksis, altså det man kunne observere at de gjorde, og intervjuet dem for å finne ut hvordan lærerne tenkte om utdanningens samfunnsmessige funksjon og sin profesjonsutøvelse. I studien var det et mål å finne ut hvilke undervisningsaktiviteter de kinesiske og amerikanske lærerne brukte, i hvor stor grad undervisningen var lærerdominert, hva som kjennetegnet elevaktiviteten i klassene, om læreren tok hensyn til de ulike kognitive nivåene som er beskrevet i Blooms taksonomi, hvilke strategier for klasseledelse lærerne valgte og hvordan de reflekterte over sin praksis, særlig med tanke på pedagogiske prosesser som vurdering for læring og differensiering (tilpasset opplæring).

Det viste seg at det var store likhetstrekk mellom de kinesiske og amerikanske lærernes klasseromspraksis (Grant m. fl., 2013 s. 262). Begge lærergruppene tok i bruk et bredt spekter av undervisningsaktiviteter, som også tok hensyn til taksonomi og spredte seg over flere kognitive nivåer. De brukte aldri lang tid på å presentere nytt lærestoff, men kunne ha flere kortere presentasjoner i løpet av en time.

Selv om de fleste læringsaktivitetene var styrt av læreren, var undervisningen preget av høyt elevengasjementet og stor elevaktivitet. Både de kinesiske og de amerikanske lærerne var svært lydhøre overfor elevene, og endret raskt undervisningen sin når elevenes tilbakemeldinger tilsa at det var nødvendig. Alle lærerne brukte også en rekke vurderingsstrategier som ga dem løpende informasjon om hvor elevene var i sin læreprosess. De bidro til at læringsmiljøet ikke bare var trygt, men også intellektuelt stimulerende og preget av godt humør (*fun*). I de intervjuene forskerne gjennomførte, viste det seg imidlertid at det var kulturelle forskjeller mellom de kinesiske og amerikanske lærerne i spørsmål om hvordan de planla undervisningen, hvordan de faktisk underviste, hvordan de arbeidet med klasseledelse og hvordan de tilpasset undervisningen.

Denne artikkelen viser at det – på tross av at det er interessante likhetstrekk med hensyn til hva som betraktes som god klasseromspraksis i ulike kulturer – er vanskelig å løsrive det som foregår i timene fra den kulturelle og sosiale konteksten undervisningen foregår i. Lærere har noen motiver og *grunner* for å arbeide slik de gjør. Det hjelper dem å få satt ord på dette grunnlaget, det de oppfatter setter arbeidet de gjør inn i en større sammenheng. Forskerne konkluderer med at USA og Kina er inne i nesten motsatte utviklingstrender, og kanskje har en del å lære av hverandre. Mens kinesiske lærere oppmuntres til mer elevsentrert og

22 Utvalget på 32 bestod av 23 kvinner og 9 menn. Sju lærere hadde mellom 5-10 års undervisningserfaring, mens 25 hadde mer enn 10 års erfaring. 18 lærere var i grunnskolen og 14 i videregående.

undersøkende arbeid, går USA i retning av økt standardisering og sentralisering. Mens kinesiske lærere arbeider mer kollektivt, blir det stadig sterkere individorientering i USA.

Oppsummert ser det ut som om lærervurdering forutsetter at man avklarer hva som skal være vurderingsobjektet. Det er viktig å se på hva lærerne er utdannet til, hva jobben deres består i, hva de bruker arbeidstiden sin til og hvilke deler av lærerens jobb som henger så uløselig sammen at det ikke er nyttig å isolere og vurdere bare ett av elementene.

1.7 Lærervurdering i OECD-området

Det finnes mange land i OECD-området som har innført systemer for lærervurdering. Noen har lang erfaring med slikt arbeid, andre har relativt nylig initiert lærervurdering (for eksempel Chile, som innførte et system i 2005). De fleste forskningsarbeidene som er inkludert i den systematiske kunnskapsoversikten om lærervurdering er fra USA. Der har accountability-politikken fra årtusenskiftet ført til mange tiltak for standardisering slik at man kan stille skoler og lærere til ansvar for elevenes læringsresultater. I de ulike statene har administrasjonen innført mer og mindre helhetlige vurderingssystemer, og noen av disse har blitt fulgt av forskere – over kortere og lengre tid. I forbindelse med reformen No Child Left Behind (2002) ble det innført årlige fremdriftsrapporter (Adequate Yearly Progress – AYP) som avdekket forskjeller i elevenes målbare læringsresultat. I 2009 ble initiativet Race to the Top lansert av det amerikanske utdanningsdepartementet (U.S. Department of Education). En viktig del av denne satsingen var at de lokale skoledistriktene skulle utvikle datasystemer som hadde til hensikt å gi informasjon om hvorvidt elevenes læringsutbytte ble forbedret.

En rapport fra OECD (Isoré 2009) ser nærmere på lærervurdering i grunn- og videregående skole. Det finnes et stort og variert antall lærervurderingssystemer i OECD. Systemene har ulike tilnærminger og forutsetter avveiinger mellom mål og ambisjoner på den ene siden og tilgjengelige ressurser på den andre. Rapporten finner at de beste systemene for lærervurdering er de som har en helhetlig tilnærming, tar i bruk varierte metoder, bygger på informasjon fra mange ulike kilder og inkluderer lærerne gjennom hele prosessen, fra utforming av systemet til innføring og løpende evaluering. Hensikten med vurderingssystemene, forventningene til hva lærere skal kunne og konsekvensene av lærervurdering må være tydelig for alle involverte parter.

Nøkkelaktører og metoder i systemer for lærervurdering

Rapporten identifiserer aktører som har ulike roller i systemer for lærervurdering. Mens sentrale myndigheter setter nasjonale standarder, får lokale myndigheter ansvar for den praktiske implementeringen av tiltakene. Skoleledere arbeider med den praktiske innføringen av systemene, og utdanningsforskere kan bidra til å evaluere og utvikle dem. Lærerorganisasjoner ivaretar lærernes interesser og kan bidra til å skape enighet om og aksept for vurderingssystemer, evaluere og utvikle dem. Elever og foreldre bidrar med vurderinger av læreres arbeid. Metodene som brukes i vurderingssystemer påvirker utfall, typen av data og kvalitet på data. Derfor må valg av metode ses i sammenheng med vurderingsstrategi. Kvantitative mål og metoder egner seg til summative vurderinger, mens kvalitative og fortolkende tilnærminger egner seg bedre til formativ vurdering. Den vanligste vurderingsformen ser ut til å være klasseromsobservasjon. Andre mye brukte former er mappevurdering – som dokumenterer bredden av lærernes arbeid, intervju eller samtaler

med lærere, elevresultater (som for eksempel brukes i Value-Added-tilnærminger). I noen OECD-land, for eksempel Mexico og Chile, testes lærernes kunnskaper i fag, didaktikk og pedagogikk. Flere land bruker også spørreskjema som fylles ut av rektor, foreldre, kollegaer og/eller elever.

Ved internvurdering er det skoleledelsen eller representanter for ledelsen som vurderer (i 60 % av OECD-landene er det rektor som vurderer). Eksternvurdering gjennomføres ofte av andre lærere med lang erfaring som har fått opplæring i å vurdere læreres arbeid. Egenvurdering kan være nyttig for profesjonsutvikling, egner seg best til formativ vurdering og inngår vanligvis i mapper. Elevvurdering brukes i begrenset omfang, kun i Mexico, Spania og Sverige. Det finnes svært lite forskning på hvordan elever kan vurdere og hva kvaliteten på slik vurdering er. Godt utformede skjemaer med tydelige spørsmål kan gi verdifull informasjon, men må brukes sammen med andre metoder.

OECD (Nusche m. fl. 2011) mener at det er vanskelig å se for seg at Norge, med sitt desentraliserte skolesystem, skal klare å utvikle et godt system for oppfølging av lærere eller produktive former for skolevurdering uten å satse på skolelederne. De mener imidlertid at det trengs en betydelig endring i kulturen for skoleledelse, blant annet må norske skoleledere oppmuntres til å bruke det handlingsrommet de har, bli læringsledere og læres opp til å forstå hvordan god undervisning kan bli bedre undervisning gjennom vedvarende arbeid med å forbedre praksis (s. 39-40). Også norske kommuner må styrke sin kunnskap om hvordan data som fremkommer gjennom prøver av ulike slag kan bearbeides, analyseres og tas i bruk i kommunens arbeid med skoleutvikling (s. 138). Dette kan oppsummeres slik:

- Et system for lærervurdering må ta hensyn til historiske, kulturelle og distriktspolitiske betingelser for norsk skole
- Et system for lærervurdering må ta hensyn til at det er svak tradisjon for styring og ledelse av utdanningssektoren og styrke skoleledelsen
- Et system for lærervurdering må ta hensyn til hva som forventes av lærerne og hva arbeidet deres faktisk består i
- Hvis et system for lærervurdering skal bidra til økt lærerprofesjonalitet og god skoleutvikling, må skoleledere og skoleeiere ha metode- og vurderingskompetanse
- Et system for lærervurdering må bygge på allmenngyldige prinsipper for vurdering

2. Den systematiske kunnskapsoversikten

I oppdragsbrev av 10. oktober 2013 bestilte Kunnskapsdepartementet en systematisk kunnskapsoversikt om temaet lærervurdering fra Kunnskapssenter for utdanning med frist 1. april 2014 (Vedlegg 1). På bakgrunn av dette ble det gjennomført prøvesøk og utformet en prosjektplan som ble godkjent av oppdragsgiver (Vedlegg 2). I prosjektplanen ble problemstillingen for kunnskapsoversikten avgrenset i følgende scope:

Hvilke former for lærervurdering kan ha positiv innvirkning på skolens kvalitet?

Gitt oppdragets korte tidshorisont, valgte Kunnskapssenter for utdanning å lage en «kort kunnskapsoversikt» *rapid evidence assessment*, også kalt *Quick Review*. De siste årene har andre kunnskapssentre publisert erfaringer de har gjort seg med korte kunnskapsoversikter. Det omtales som et format som særlig egner seg for politikkutforming (Thomas mfl. 2013). I arbeidet med denne rapporten har vi bygd på disse erfaringene. I tillegg har vi hentet erfaringer fra EPPI-senterets anbefalinger for utforming av systematiske kunnskapsoversikter som bygger på både kvantitativ og kvalitativ metode og som særlig egner seg til å informere politikkutformere og praksisfeltet.

Utgangspunktet for en systematisk kunnskapsoversikt er å finne all relevant litteratur som er publisert om et tema og utvikle en syntese som gir kunnskapsstatus på temaet. Et ideal for systematiske kunnskapsoversikter er at de skal være transparente, det vil si at de gjennomføres etter klart definerte prinsipper og prosedyrer. De bygger på en metode som beskriver hvordan de gjennomføres og viser hvordan konklusjoner blir nådd. At de er systematiske innebærer at det er et mål å finne så mange relevante studier som mulig, at utvalget av relevante studier skjer gjennom en åpen prosess, at reliabiliteten på studiene blir vurdert og at kvalitetsvurderingsmekanismer som handler om å inkludere og ekskludere studier er bygd inn i prosessen (Chalmers mfl. 2002; Gough mfl. 2012). For å klare dette, må det utvikles gode søkestrenger som gjør det mulig å finne det som er av forskning på temaet. I en systematisk kunnskapsoversikt er det de inkluderte artiklene som utgjør det empiriske grunnlaget, og kvaliteten på oversikten avhenger av kvaliteten på de arbeidene som blir identifisert og inkludert. Arbeidet med systematiske kunnskapsoversikter skiller seg fra grunnforskning på den måten at det ikke er nødvendig å definere forskningsobjektet på samme måte som i et vanlig forskningsprosjekt. En systematisk kunnskapsoversikt undersøker og innhenter heller ikke empiri om et saksfelt slik grunnforskning gjør.

Normalt tar det lang tid, gjerne inntil ett år, å lage en fullstendig oversikt over den forskningen som er tilgjengelig om et tema og presentere denne på en måte som tilfredsstiller

forskersamfunnets kvalitetskriterier. For de som skal utforme politikk eller andre som trenger kunnskap om et tema raskere enn dette, er det i det siste utviklet kortere formater som for eksempel «brief review« (Abrami mfl. 2010) eller «rapid evidence assessment» (REA) (Khangura mfl. 2012). I tillegg til at den teknologiske utviklingen forenkler søkeprosedyrer og at «mining»-funksjoner i tekst gjør det enklere å lage slike kortformat, foregår det en vedvarende metodeutvikling av systematiske kunnskapsoversikter med sikte på å gjøre resultater og innsikter fra forskning lettere tilgjengelig for de som skal bruke kunnskapen. Politikkutformere vil ikke bare ha svar på hva som ser ut til å virke, men også på spørsmål som «hva er mulig», «hva kan fungere», «hva trengs» samt hvorfor den ene løsningen ser ut til å være bedre enn den andre. Når målet er å besvare komplekse politiske spørsmål, har det derfor blitt vanligere å inkludere kvalitativ metode og surveydata i en kunnskapsoversikt. Kunnskapsoversikter som baserer seg på mixed methods, blir også stadig mer populære fordi de har mulighet til å bidra med svar på spørsmål om implementering.

Arbeidet med en kort kunnskapsoversikt følger de samme prosedyrene som for en fullstendig kunnskapsoversikt, og egner seg godt i de tilfellene da det er grupper av forskere som skal samarbeide om en rapport (Pope m.fl. 2000). En kort systematisk kunnskapsoversikt kan beskrives som et kompromiss mellom de strenge kvalitetskravene som stilles til en systematisk kunnskapsoversikt og policy-nivåets behov for å få forskningskunnskapen så hurtig som mulig. Søkene er systematiske, det er åpenhet om hvilke studier som inkluderes og ekskluderes (og etter hvilke kriterier), de inkluderte studienes reliabilitet blir vurdert, og det bygges mekanismer for kvalitetsvurdering (validering) inn i reviewprosessen. Når tempoet øker, vil det imidlertid alltid være en forhandlingsprosess mellom grundigheten på arbeidet og tempoet det skal gjennomføres i. Hva man eventuelt må renonsere på, varierer fra rapport til rapport og fra fagfelt til fagfelt. Derfor er det heller ikke tilrådelig å lage en «oppskrift» på formatet kort kunnskapsoversikt (Thomas mfl. 2013). Det er imidlertid viktig å ta hensyn til hvilke fremgangsmåter som kan brukes i ulike sammenhenger og være klar over hvilke fallgruver som finnes.

Prosess og forankring

Et problem som trekkes frem ved korte kunnskapsoversikter, er at *forankring* kan bli vanskelig når arbeidet skal gjøres så raskt. Når tiden er kort og mange aktører skal inn i arbeidet, kan bred forankring bidra til å gjøre fokus uklart (Thomas mfl. 2013 s. 15).

For å sikre forankring av kunnskapsoversikten om lærervurdering, ble det nedsatt en arbeidsgruppe fra GNIST-samarbeidet, med professor Eyvind Elstad (UiO) som leder. Elstad var med på det andre møtet om søkestrategi (28. november 2013), og hele gruppen deltok på møtene 17. januar 2014 og 17. februar 2014. På det siste møtet presenterte også arbeidsgruppen fra GNIST-samarbeidet resultater fra informasjon de har innhentet om praksiser i kommuner eller fylkeskommuner som har satt i gang ulike former for lærervurdering. Den parallelle prosessen har bidratt til å klargjøre kjernen av problemstillingen som skulle besvares og innvirket på utformingen av rapporten.

I komplekse systemer med mange aktører og bred deltakelse er det som regel ikke bare interessant å vite hva forskningen viser at kan være de beste løsningene eller hvilke anbefalinger for god praksis som ligger i artiklene. Politikere og forvaltningen trenger også svar på implementeringsutfordringer.

Avgrensing av problemstilling og inkludering av litteratur

En viktig del av arbeidet med en kunnskapsoversikt er å ramme inn problemstillingen. Hvordan problemstillingen er formulert og hva slags resultat man skal oppnå, avgjør fremgangsmåte og hvilken populasjon (utvalg) man skal avgrense søkene til. Godt forarbeid bidrar til at søke- og sorteringsprosessen går raskere.

Her var tematikken bred og flere spørsmål skulle besvares. Oversikten skulle inkludere studier om hvilke former for lærervurdering som kan ha positiv innvirkning på resultater og prosesser i skolen, og studier som kaster lys over prosesser som kan bidra til utvikling av kvalitet i skolen. Oppdragsgiver ønsket med andre ord å få belyst flere forhold som har betydning for en praksis (lærervurdering). Når oppdragsgiver vil vite hva forskning sier om hvilke former for lærervurdering det er som kan bidra til kvalitet i skolen, må både innsamling av data, kategorisering av artiklene og sammendragene vise hvordan forskningen – direkte og indirekte – diskuterer dette spørsmålet. I denne kunnskapsoversikten er derfor inklusjons- og eksklusjonskriterier brukt pragmatisk fordi det har vært et mål å belyse hele bredden av problemstillingen.

På unge forskningsfelt er det ofte få (eller ingen) studier som bygger på randomiserte kontrollerte forsøk. Slik er det i dette konkrete tilfellet. Tre av studiene som er inkludert er randomiserte kontrollerte forsøk, og alle tre handler om prestasjonslønn. Det viser seg at lærervurdering er en noe tilfeldig og spredt praksis, og det som finnes av forskning er følgelig enten svært lokal, gjort på små utvalg, eller i en avgrenset kontekst. Det kan for eksempel være innføring av prestasjonslønn som en av flere komponenter i en stor satsing på lærervurdering i skoler som ligger i svært fattige byområder i en amerikansk delstat. Svært mye av forskningen er dessuten publisert i rapporter som er bestilt på oppdrag fra nasjonale og lokale myndigheter.

«Både kvalitativ og kvantitativ forskning skal inkluderes»

En særlig utfordring i dette arbeidet har vært utvalg og avgrensing av den litteraturen som skal inkluderes. Oppdraget omfatter summativ og formativ vurdering, kvalitativ og kvantitativ forskning samt forskning som sier noe om prosess og produkt (resultat). Det har ikke vært enkelt å identifisere artikler som svarer direkte på prosjektets scope: *Hvilke former for lærervurdering kan ha positiv innvirkning på skolens kvalitet?* Slik informasjon fremkommer først og fremst indirekte, og det har derfor blitt nødvendig å lete «rundt« problemstillingen. Det viktigste har vært å få frem mest mulig kunnskap om ulike former for lærervurdering. Informasjon om hvilke former for lærervurdering som kan ha positiv innvirkning på skolens kvalitet er ofte implisitt i artiklene, og mest tydelig beskrevet i de artiklene som presenterer systemer for lærervurdering som er innført i forskjellige land (kapittel 3.1).

De siste tretti årenes forskning om vurdering – både skolevurdering og elevvurdering – kan bidra til å avklare spørsmålene som man søker svar på. I tillegg forskes det stadig mer på skoleledelsens betydning for skolens utvikling og kvalitetsarbeid – særlig om behovet for at skolelederen engasjerer seg *faglig* i spørsmål som angår skolens undervisning. Her er det imidlertid snakk om så mye forskningslitteratur at det er materiale til flere kunnskapsoversikter. Denne rapporten bygger på de kvalitetskriteriene som både EPPI-senteret og Dansk Clearinghouse for Uddannelsesforskning legger til grunn i sine rapporter. De utelater studier med lav evidensvekt eller lar disse studiene få liten plass i oversikten. I denne oversikten er følgende kvalitetskriterium satt: Artikkelen må ha en problemstilling som er klart formulert, den må gjøre rede for metodevalg og metodisk fremgangsmåte og det må være sammenheng mellom problemstilling, funn, drøfting og konklusjon. Kvaliteten på studiene er vurdert sammen med forskergruppen, og de er relevansvurdert med henblikk på om de kan besvare problemstillingen.

2.1 Inklusjons- og eksklusjonskriterier

I en systematisk kunnskapsoversikt er det nødvendig å spesifisere inklusjons- og eksklusjonskriterier for å avgjøre hvilke studier som skal inngå i kunnskapsoversikten og hvilke studier som skal ekskluderes. Klart definerte eksklusjonskriterier brukes aktivt i selve screening-prosessen, dvs. gjennomgangen og utvelgelsen av de ulike artiklene som skal inngå i oversiktene. Innledningsvis søkes det på tittel og sammendrag, og det er en fordel å ha klart definerte eksklusjonskriterier for å oppnå en effektiv screening-prosess. Følgende inklusjons- og eksklusjonskriterier ble definert og benyttet for å kode de ulike studiene basert på gjennomgang av tittel og sammendrag:

EKSKLUDER	INKLUDER
Dato	Artikler publisert i tidsrommet 1.1.2009 – 1.2.2014
Språk	Engelsk, norsk, svensk, dansk
Emne	Lærervurdering
Utdanningsnivå	Grunnskole og videregående skole
Ikke-vitenskapelige studier	Studier publisert i fagfellevurderte tidsskrifter
Bøker/bokkapitler, avhandlinger og rapporter	Utvalgte OECD-rapporter

Studier som ikke blir ekskluderte i henhold til eksklusjonskriteriene over, vil da være inkluderte. For å sikre transparens i prosessen, skal det i en systematisk kunnskapsoversikt gis en begrunnelse for valg av kriterier. Vedlegg 3 viser en fullstendig beskrivelse av inklusjons- og eksklusjonskriterier med begrunnelser.

Korte kunnskapsoversikter er et alternativ til tradisjonelle systematiske kunnskapsoversikter hvor det blir gjort kompromisser med hensyn til bredde og dybde i litteraturtilfanget. Bruken av inklusjons- og eksklusjonskriterier viser at de trenger ikke være mindre stringente når det gjelder å bestemme de konseptuelle grensene - som i sin tur bestemmer hvilke studier kunnskapsoversikten skal inneholde.

2.2 Søkestrategi

En kunnskapsoversikt er bare så god som de studiene som er inkludert i den. Hensikten med å søke bredt er å identifisere så mye som mulig av litteraturen som tilfredsstiller inklusjonskriteriene. Brede søk skal også sikre at kunnskapsoversikten reflekterer kunnskapsstatusen på feltet.

Denne rapporten baserer seg på en kombinasjon av søk med høy presisjon og søk med høy sensitivitet. Søk med høy presisjon gjennomføres ved hjelp av standardiserte emneord, og skal identifisere en høy andel av studier som møter inklusjonskriteriene, men med risiko for å miste andre relevante studier. Søk med høy sensitivitet gjøres ved fritekstsøk, som identifiserer en høy andel av totalt antall eksisterende artikler på feltet, men vil i tillegg produsere et høyt antall irrelevante artikler som må screenes og eventuelt ekskluderes. I litteraturen er det mange begreper som brukes om vurdering av lærere og om skolens kvalitet, slik at søkene som gjennomføres er komplekse.

Det ble foretatt en begrepsorientering i forhold til hvilke former for lærervurdering det var nødvendig å se nærmere på, samt hvilke mål som er satt for kvalitet i grunnopplæringen i Norge. Begreper for lærervurdering ble hentet fra oppdragsbrevet, men også fra OECD Education Working Paper No. 23 (Isoré 2009). Foruten begreper for prosesskvalitet og resultatkvalitet som fremkom i oppdragsbrevet fra KD, ble St. meld. Nr. 31 (2007-2008), *Kvalitet i skolen*, gjennomgått for en mulig utvidelse av aktuelle søkebegreper. Siden kunnskapsoversikten begrenses til å omfatte grunnskole og videregående skole, ble det i tillegg identifisert begreper fra databasene om lærere på det aktuelle utdanningsnivået. Dette ble gjort ved hjelp av databasenes synonymordbøker.

Basert på denne gjennomgangen av begreper ble det identifisert standardiserte emneord for de indekserte databasene Kunnskapssenteret har tilgang til (ERIC, ASSIA, IBSS og PQEJ). De standardiserte emneordene identifiseres i de ulike databasenes synonymordbok. Indekseringen er ikke lik for de enkelte databasene, slik at emneordene må oversettes mellom databasene. Det ble i tillegg utarbeidet en liste over fritekst søkeord til bruk i alle databasene Kunnskapssenteret har tilgang til (ERIC, ASSIA, IBSS, PQEJ, COS, PQDT A&I, PQDT UK&I). Et eksempel på en søkestreng med standardiserte emneord fra ERIC-databasen er vedlagt rapporten (Vedlegg 4).

For å finne den relevante litteraturen til kunnskapsoversikten har det blitt søkt i følgende kilder:

- Elektroniske databaser (blant annet ERIC, ASSIA og PQEJ)
- Databaser med skandinavisk litteratur (blant annet BIBSYS/ASK, IDUNN, KB)
- Håndsøk i utvalgte tidsskrifter og på internett (Google Scholar)
- Søk i såkalt grå-litteratur (blant annet OECD, RAND, NIFU)

Se Vedlegg 5 for en fullstendig oversikt over de søkekildene som er benyttet i arbeidet med denne kunnskapsoversikten.

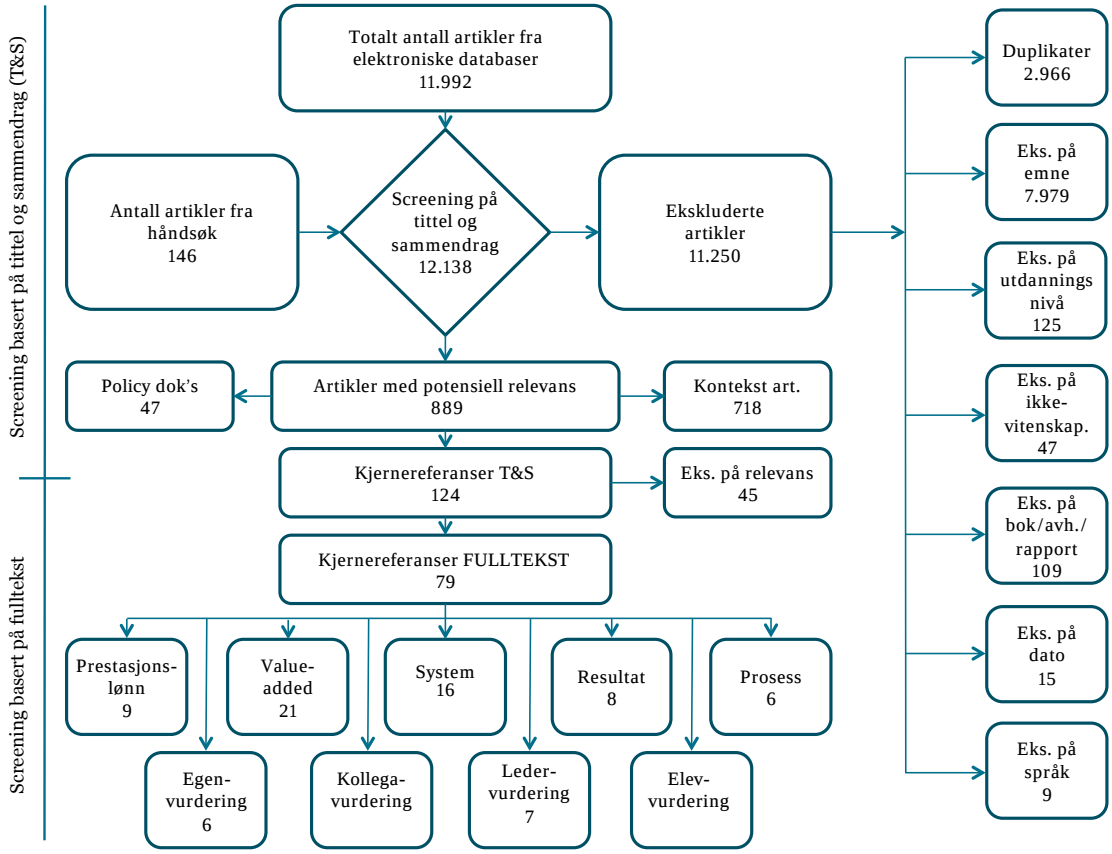
Etter at litteratursøkene var gjennomført i de ulike kildene, viste resultatene følgende:

- Søk i elektroniske databaser i ProQuest-portalen resulterte i 11.971 referanser.
- Søk i den elektroniske databasen Bibliotek.dk resulterte i 21 referanser.
- Håndsøk og søk i grå-litteratur resulterte i 146 referanser.

2.2.1 Referansehåndtering

For å håndtere en slik mengde med data, har Kunnskapssenter for utdanning benyttet programvaren EPPI-Reviewer 4 som er spesielt utviklet for å lage systematiske kunnskapsoversikter og syntetisere forskningen. Alle referansene (til sammen 12.138) ble importert til EPPI-Reviewer 4 programvare (ER4). EPPI-Reviewer 4 er utviklet av EPPI-senteret ved University of London. Referanser ble importert i ER4 som RIS-filer eller lagt inn manuelt.

Når alle referanser er importert i programvaren, starter arbeidet med å screene de ulike artiklene fra litteratursøkene for å bedømme om de skal inngå i kunnskapsoversikten eller ekskluderes. Flytdiagrammet i Figur 1 beskriver prosessen med screening av referansene, som foregår i to trinn. I det første trinnet screenes studiene basert på en gjennomgang av tittel og sammendrag, sett opp mot inklusjons- og eksklusjonskriteriene. I det andre trinnet screenes de inkluderte studiene ved at man gjennomgår fulltekst.



Figur 1: Resultat av screening

Resultat av screening basert på gjennomgang av tittel og sammendrag

I første trinn av screeningen ble det ekskludert 11.250 artikler basert på gjennomgang av tittel og sammendrag. Det ble identifisert 889 artikler med mulig relevans for kunnskapsoversiktens brede problemstilling. Disse artiklene ble sortert i tre kategorier: Kjernereferanser (124), kontekstartikler (718) og policydokumenter (47). Kategorien kontekstreferanser omfattet studier med relevans for lærervurdering og skoleutvikling. Det vil si at de omhandlet tema som lærerutdanning, klasseromspraksis, prosjektarbeid, lærerprofesjonalitet osv., men ikke koblet til spørsmål om lærervurdering.

Kjernereferansene ble så delt inn i ni tematiske kategorier: Prestasjonslønn, Value-Added, System, Resultat, Prosess, Egenvurdering, Kollegavurdering, Ledere som vurderer lærere, Elever som vurderer lærere. Kategoriene ble opprettet som sorteringshjelp i arbeidet med å gjennomgå kjernereferanser med potensiell relevans for kunnskapsoversikten.

Resultat av screening basert på gjennomgang av artikler i fulltekst

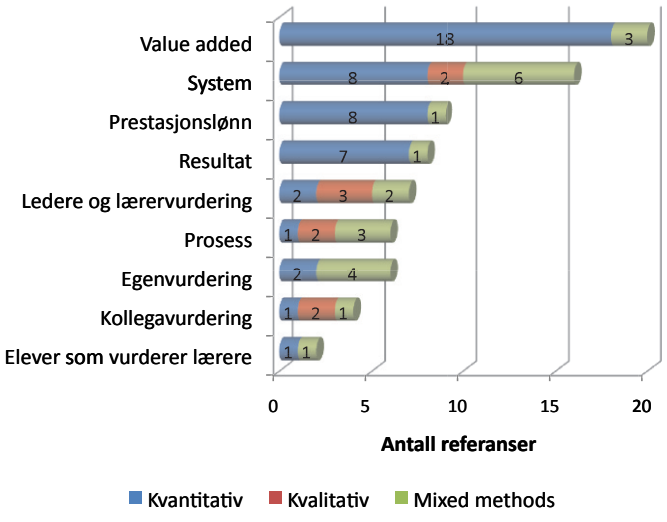
I det andre trinnet av screeningen tas det en gjennomgang av artiklene basert på fulltekst for å identifisere de artiklene som har størst relevans og som mest sannsynlig vil kunne svare på kunnskapsoversiktens problemstilling. Det ble skaffet artikler i fulltekst av alle kjernereferansene (124) som ble fordelt i de ni kategoriene som ble identifisert i første del av screeningen. Deretter ble fulltekstartikler sendt til medlemmene i forskergruppen. Under gjennomgangen av fulltekst ble det oppdaget flere artikler som enten handlet om elevvurdering eller om høyere utdanning. Disse ble derfor ekskludert på emne. Andre eksklusjonskriterier ble også anvendt i denne fasen. EKSKLUDER på relevans ble brukt i tilfeller da det viste seg at sammendraget ikke hadde gitt fullgod informasjon om innholdet i artikkelen. For eksempel var ikke lærervurdering det sentrale temaet i artiklene selv om det så slik ut i sammendraget. Noen av artiklene svarte heller ikke til de kvalitetskriteriene forskergruppen fulgte. Det kunne være manglende eller dårlig samsvar mellom forskningsspørsmål, metode og funn, uklar beskrivelse av datainnsamling, metode og analyse av data. I andre sammenhenger var det dårlig konsistens mellom det artikkelen innledningsvis tok mål av seg til å gjøre og det som faktisk ble gjort. Det skjedde også at en artikkel ikke bragte noe nytt til et tema som allerede var fyldig drøftet (mettet) eller at en artikkel tok opp tema med så spesifikk relevans for lokale forhold at den ikke var interessant i en norsk kontekst.

Vedlegg 6 gir et eksempel på hvordan medlemmer av forskergruppen systematisk har screenet artikler basert på gjennomgang av fulltekst.

Etter screening basert på gjennomgang av fulltekstartikler, ble 45 artikler ekskludert på relevans etter kriteriene beskrevet ovenfor. Til sammen 79 artikler gjenstår, fordelt på kategoriene: Prestasjonslønn (9), Value-Added (21), System (16), Resultat (8), Prosess (6), Egenvurdering (6), Kollegavurdering (4), Ledere som vurderer lærere (7), Elever som vurderer lærere (2).

2.2.2 Kartlegging («mapping»)

Mapping innebærer å kartlegge og kategorisere studier s om tilfredsstiller inklusjonskriteriene (Gough m. fl. 2012). Å kartlegge litteraturen er et første ledd i syntesearbeidet. Hens ikten med å gjennomføre en kartlegging er å sikre seg at kunnskapsoversikten konsentrerer seg om de områdene som er interessante for de som skal bruke funnene fra forskningen. En kartlegging tar utgangs punkt i bestemte kriterier (populasjon, intervensjon, resultat) og kan gjennomføres b åde på tittel og sammendrag og på fu lltekst. I denne kunnskapsoversikten har vi valgt å bruke kategori og metode i kartleggingen av de 79 kjerneartiklene. Dette gir et overblikk over hva slags forskningsmetode som er hyppigst benyttet i forskningen om lærervurdering. Figur 2 viser resultatet av kartleggingen.



Figur 2: Resultat av kartlegging etter kategori og metode av 79 kjerneartikler.

Kartleggingen viser at det er en overvekt av kvantitative studier i materialet (48) og at det finnes kvantitative studier innenfor alle de ni identifiserte kategoriene. Resultatet viser også hvilke forskningsmetoder som dominerer innenfor de ulike kategoriene. Dette er til hjelp i selve syntesearbeidet fordi det sier noe om hva slags type spørsmål forskerne kan besvare ved hjelp av de ulike metodene, for eksempel om tyngdepunktet i forskningen ligger på å måle resultater eller å beskrive prosesser.

Vedlegg 7 og 8 viser eksempler på kartlegging hvor studiene innenfor kategoriene System og Value-Added er kartlagt ved hjelp av metode og forskningsdesign. Fullstendige kartleggingsdata finnes i protokollen som er tilgjengelig elektronisk på Kunnskapssenterets hjemmeside. (www.kunnskapssenter.no)

2.2.3 Syntetisering av funnene

Når systematiske kunnskapsoversikter inkluderer artikler som bygger på flere metoder, er utfordringen å syntetisere dem på en god måte. Av de 79 inkluderte artiklene i denne syntesen er 48 kvantitative, 22 mixed-methods og 9 kvalitative. Bare i den kategorien artikler som omhandler ulike systemer for lærervurdering (16 inkluderte), er det brukt så ulike forskningsdesign som tverrstudier, casestudier, kritisk analyse, etnografisk analyse og reviews. En mulig fremgangsmåte i systematiske kunnskapsoversikter er å syntetisere kvalitative og kvantitative studier hver for seg, for deretter å kombinere syntesene. Metaanalyse er en statistisk analyse som er særlig nyttig når kunnskapsoversikten hovedsakelig inneholder RCT-studier (randomiserte kontrollerte undersøkelser), som gir svar på hvorvidt en intervensjon har (eller ikke har) effekt. I denne rapporten er det bare tre RCT-studier, som alle har undersøkt effekten av prestasjonslønn (Fryer 2013; Yuan m. fl. 2013; Springer m.fl. 2012) og fire eksperimentelle studier (Kleinknecht og Schneider 2013; van Diggelen m.fl. 2013; Nelson m. fl. 2013 og Verberg m. fl. 2013). Mange av de longitudinelle studiene og tverrstudiene som inngår i kategorien Value-Added er ikke empiriske, men designet for å teste ut modeller på store, forhåndsgenererte datasett, altså sekundæranalyser. Det vil si at en statistisk metaanalyse ikke er relevant her, fordi denne rapporten ikke inneholder et tilstrekkelig antall homogene effektstudier som kan danne grunnlag for en slik analyse. En anerkjent metode for å syntetisere kvalitative studier er metaetnografi (Noblit og Hare 1988). Denne metoden egner seg ikke i denne kunnskapsoversikten fordi det er få inkluderte kvalitative studier (9) og en stor andel av artiklene (22 av 79) har brukt såkalt mixed-method, det vil si at de kombinerer flere metoder for å få svar på sine forskningsspørsmål. Det heterogene forskningsdesignet og mangfoldet i metoder tilsier at det beste valget i denne kunnskapsoppsummeringen er å gjennomføre en narrativ syntese.

Narrativ syntese

Narrativ syntese egner seg særlig når de inkluderte primærstudiene bruker forskjellige metoder (Kløveager Nielsen 2014). Med en narrativ tilnærming kan syntesen både «fortelle historien» om resultatene av primærforskningen og gjennom tolkning og analyse av funnene i primærstudiene omdanne resultatene på måter som kan bidra til å skape ny kunnskap (Petticrew og Roberts 2006). Narrativ syntese kan både samle resultatene fra effektstudier som er for heterogene til å inngå i en statistisk metaanalyse og fra studier som ser på implementering av intervensjoner. I tillegg kan denne syntesemetoden få frem prosesser (kvalitative vurderinger) som kan forklare hvorfor noe ser ut til å virke, under hvilke betingelser det virker og for hvem det kan se ut til å virke. I litteraturen brukes betegnelsene *aggregerende* og *konfigurativ* for å betegne synteser som henholdsvis sammenstiller primærstudienes resultater og fortolker dem (Gough og Thomas 2012).

Syntesearbeid følger normalt fire trinn, ikke nødvendigvis i kronologisk rekkefølge, men gjerne som en iterativ prosess. Innledningsvis organiseres resultatene fra de inkluderte studiene i den hensikt å identifisere mønstre på tvers av studiene. I denne kunnskapsoversikten ble dette gjort gjennom den første kategoriseringen av artiklene (kapittel 2). På neste stadium begynner arbeidet med å sammenfatte (aggregere) resultatene fra primærstudiene (kapittel 3). Målet for syntesearbeidets tredje trinn er å identifisere mønstre på tvers av studiene (konfigurere) og utvikle en første beskrivelse av resultatene fra de inkluderte studiene (kapittel 4). Av særlig interesse på dette stadiet er å overskride primærstudiene ved å analysere funnene på tvers. Gjennom denne prosessen blir det mulig å avdekke mulige spenninger og motsetningsforhold i materialet, samt å få svar på spørsmål som hva som kan fremme eller hemme utvikling. Et formål med dette trinnet er også å undersøke hva som kan forklare hvorfor studiene konkluderer forskjellig. Dette kan være særlig viktig informasjon for politikkutformere som har ansvar for å designe tiltak som skal implementeres i et praksisfelt.

Til slutt besvares spørsmål om hvilke former for lærervurdering som ser ut til å ha positiv innvirkning på skolens kvalitet.

3 En systematisk gjennomgang av forskningslitteraturen

I dette kapittelet presenteres resultatene fra det andre trinnet i den narrative syntesen, altså den forberedende syntesen. Her sammenfattes resultatene fra primærartiklene og aggregeres i nye kategorier. Denne prosessen skal bidra til å identifisere mønstre på tvers av studiene. Etter å ha gjennomgått de 79 kjerneartiklene som gjenstod etter litteratursøk, kategorisering, eksklusjon, inklusjon, kartlegging, kvalitet- og relevansvurdering i fulltekst, har følgende hovedkategorier utkrystallisert seg i materialet:

1. Erfaringer med lærervurdering i ulike land,
2. Forskning med vekt på summativ vurdering,
3. Forskning med vekt på formativ vurdering,
4. Metodediskusjoner (spørsmål om validitet og reliabilitet)
5. Betydningen av ledelse i systemer for lærervurdering.

I disse fem hovedkategoriene sammenfattes og presenteres nå resultatene fra primærartiklene som er identifisert gjennom de systematiske søkene. Hensikten med denne sammenfatningen er å få frem ny informasjon som kan bidra til å besvare kunnskapsoversiktens forskningsspørsmål (scope).

3.1 Erfaringer med lærervurdering i ulike land

De systematiske søkene var avgrenset til å omfatte artikler som er publisert etter 2009. Det ble ikke funnet noen skandinaviske artikler som omhandlet lærervurdering i denne perioden, men en fra Belgia, en fra Portugal, to fra Kina og fire fra Chile. Disse artiklene bygger på empiriske undersøkelser, og gir fyldige beskrivelser av problemstillinger rundt innføring, gjennomføring og forbedring av systemer for lærervurdering. Gjennom eksemplene kommer det tydelige og konkrete råd som bidrar til å belyse problemstillingen om hvilke former for lærervurdering som kan bidra til god kvalitet i skolen.

Eksemplene er fra land som ligger langt fra Norge, med andre historiske forutsetninger og kulturelle tradisjoner. Likevel kan det være mye å lære av å se hvilke problemer andre opplever når de skal utforme et system for lærervurdering som også skal bidra til kvalitet i skolen. Ettersom skolereformer siden 1980-årene har hatt stor oppmerksomhet på elevenes læringsutbytte, er det utvilsomt noen felles utfordringer. En fyldig beskrivelse av artiklene i denne kategorien får både frem gjenkjennelige lokale utfordringer og generiske problemstillinger rundt betingelser for lærervurdering.

Etter de systematiske søkene er 16 artikler som tilfredsstiller kvalitetskriteriene[23] og som beskriver og drøfter modeller for lærervurdering i ulike land inkludert. I det opprinnelige søket var det artikler fra flere land (blant annet Botswana og Malaysia), men disse ble ekskludert etter at de var lest i fulltekst fordi de enten ikke holdt kvalitetsmålene eller fordi de viste seg å omhandle feil utdanningsnivå.

3.1.1 Chile

Det mest veldokumenterte eksempelet på et system for lærervurdering er fra Chile. Taut og Sun (2014) har identifisert hele 17 publikasjoner som drøfter forskjellige sider ved det chilenske systemet for lærervurdering i de offentlige grunnskolene som ble lansert i 2003 og gjort obligatorisk i 2005 (Taut m.fl. 2011; Taut og Sun 2014). Innføringen av systemet må ses i sammenheng med overgangen fra militærregime til demokrati i 1990- årene. Under militærregimet ble lærerne sett på som en sentral opposisjonsgruppe. I 1981 ble derfor lærerutdanningen overført fra universitetene til tekniske fagskoler. Med denne manøveren ble læreryrket en teknisk karriere med lav lønn, lav status og få fordeler, og det ble økende bekymring i Chile for kvaliteten på lærernes arbeid og undervisning. Etter en periode med mye kritikk og debatt, gjorde militærregimet to av de tekniske fagskolene om til pedagogiske universiteter i 1987. Etter at den demokratiske regjeringen overtok i 1990-årene, tok den flere initiativ for å bedre lærernes status. I 1990 ble det lovfestet at lærerutdanningen skulle være en universitetsutdanning[24]. Videre ønsket den nye regjeringen å gjenopprette læreryrket som en attraktiv karrierevei, ved å øke lønningene og ansette lærere i faste stillinger. Deretter ville de innføre prestasjonsbaserte vurderinger av lærernes arbeid. Etter nesten ti år med forhandlinger mellom lærerorganisasjoner og kommunale ledere, ble partene i 2002 enige om en modell for lærervurdering.

I Chile finnes det tre kategorier skoler: a) offentlige (39,9 %), b) statssubsidierte private (52,9 %) og c) private (7,2 %). Det er landets 346 kommuner som administrerer de offentlige skolene, mens private (individer eller institusjoner) har ansvar for resten. Det er kun i de offentlige skolene at det systemet for lærervurdering som beskrives her, er innført.

Vurderingsmetoder som benyttes i systemet i Chile

Det nasjonale systemet for lærervurdering bygger på nasjonale standarder for god undervisning og insitamenter i form av belønninger for enkeltlærere og grupper av lærere. Vurderingsmetodene omfatter: 1) en vurderingsmappe (med en skriftlig tekst og videoopptak av egen undervisning), 2) egenvurdering (spørreskjema), 3) kollegavurdering (gjennomført av en lærer med samme fag fra en annen skole), og 4) vurdering som blir gjort av en veileder (skolelederen eller skolens pedagogiske ekspert). Til slutt får lærerne en samlekarakter basert på en vekting av resultatene fra de fem vurderingsmetodene. Vektingen er lovfestet og fordeler seg slik: mappevurdering 60 %, kollegavurdering 20 %, og vurdering av veileder/rektor samt selvevaluering 10 % hver.

Mappevurdering er tyngst vektlagt 60 %. Selve mappen består av to deler. I den første delen skal lærerne utvikle undervisningsmateriell for åtte undervisningstimer. De velger læringsmål fra den nasjonale læreplanen (etter fag og klassetrinn), og må levere inn det undervisningsmaterialet som faktisk ble brukt i timene. Dette inkluderer undervisningsplaner, skriftlige svar på spørsmål om hvordan materialet ble brukt,

23 Artikkelen må ha en problemstilling som er klart formulert, den må gjøre rede for metodevalg og metodisk fremgangsmåte og det må være sammenheng mellom problemstilling, funn, drøfting og konklusjon.

24 Avalos, B (2005) Secondary Teacher Education in Chile: An assessment in the light of demands of the knowledge society. Ministry of Education Chile. This paper is a modified version of earlier papers on the subject of teacher education in Chile prepared for UNESCO (Avalos, 2003) and the World Bank/DFFID Project «Learning to Teach in the Knowledge Society» (Avalos, 2004).

studentenes resultater fra vurderingen, og refleksjon rundt egen undervisningspraksis. I tillegg skal materiale som reflekterer pedagogisk praksis, for eksempel refleksjoner rundt elevens lærevansker og elevenes motivasjon leveres inn sammen med skjema som vurderer undervisningen. I den andre delen er det en ekstern spesialist som filmer en av undervisningstimene (40 min). Læreren vet på forhånd hvilken time som skal tas opp. I etterkant av opptaket svarer læreren på et kort spørreskjema om undervisningen. Lærerne har 12 uker til å fullføre mappen. I årene 2005-2010 har det i gjennomsnitt blitt evaluert ca. 13 000 lærere årlig.

Lærernes tilbakemelding på denne vurderingsmetoden er at de synes den skaper unødvendig mye tilleggsarbeid. Mer enn 70 % av respondentene rapporterer at de har «lite tid» til å fullføre mappen. På den annen side mener lærerne at mappevurderingen var nyttig og relevant for deres egen profesjonsutvikling fordi den gir dem mulighet for meningsfull interaksjon med kollegaer og til å reflektere over og revidere egen undervisningspraksis. Lærerne har tillit til at mappevurderinger gir et perspektiv på bredden av arbeidet deres. Som svakhet ved mappevurdering nevnes muligheten for kopiering og forfalskning hvis lærerne kjenner til skåringssystemet på forhånd.

Mappen blir vurdert av en evaluator, som har fått opplæring i å vurdere mapper, og som bruker et detaljert skåringssystem. De som vurderer mappene, må ha mer enn fire års erfaring og være praktiserende lærere i det aktuelle faget.

Den andre vurderingsmetoden som blir benyttet, er kollegavurdering. Det skjer i form av et intervju som gjennomføres av en lærer som arbeider på samme nivå, men på en annen skole. Intervjuet tar omkring 50 minutter. Hver vurderer gjennomfører i gjennomsnitt 12 intervju. En kollega kan bare arbeide som vurderer hvis han eller hun ikke skal vurderes selv det året. Intervjuet består av 6-8 spørsmål som er like for alle nivå og fag. Hvert år gjennomføres det tre pilotstudier for å utvikle spørsmålene og skåringssystemene. Omtrent 100 lærere og tekniske eksperter deltar i dette arbeidet. De evaluerer skåringssystemene, relevans og klarhet i spørsmålene, og hvor fokuserte svarene er. Over tid har spørsmålene blitt mer komplekse. De krever nå utfyllende svar der lærerne i større grad må demonstrere sine pedagogiske kunnskaper.

Intervjuere blir valgt ut på bakgrunn av et sett lovbestemte kriterier: a) de må være lærere i kommunen med minst fire års undervisningserfaring, b) de må ikke ha fått noen disiplinære reaksjoner, c) de må selv være «kompetent» eller «fremragende». Alle deltar i et to-dagers opplæringsprogram som tar opp tema som a) hvordan man gjennomfører et intervju, b) hvordan man skriver utfyllende notater, c) hvordan man skårer svarene basert på et detaljert skåringssystem og d) etiske spørsmål knyttet til det å intervjue en kollega. I løpet av opplæringen av de som skal intervjue, og på bakgrunn av observasjon av intervjugjennomføringer og oppgaveløsning, får omtrent 8 % beskjed om at de ikke får anledning til å vurdere kollegaer.

Svakheter ved denne vurderingsmetoden er knyttet til behovet for konfidensialitet vedrørende spørsmålene og skåringssystemet ettersom det er flere hundre evaluatorer involvert hvert år. Da OECD evaluerte det chilenske vurderingssystemet (OECD 2013), ble nettopp denne vurderingsmetoden kritisert fordi intervju av kollegaer ikke ble fulgt opp gjennom tilbakemeldinger. Med en litt annen innretning kunne den ha bidratt mer direkte til profesjonsutvikling.

Den tredje vurderingsmetoden er vurdering som blir gjort av en veileder. Denne metoden består av et spørreskjema som blir fylt ut av rektor og lederen av skolens pedagogiske avdeling. Hvert spørreskjema teller 5 % av sluttkarakteren, slik at de til sammen utgjør 10 %. Veilederne skårer lærerne på en skala fra 1 (ikke tilfredsstillende) til 4 (fremragende). For å motvirke inflasjon i høye skårer, ble det innført en rubrikk der veilederne måtte begrunne skåren «fremragende» når den ble brukt. Hvis ikke en slik begrunnelse forelå, ble karakteren automatisk senket til «kompetent». Det blir gitt en kort opplæring for rektorer om tema knyttet til personalvurdering generelt og selve vurderingsmetoden spesielt.

Vektingen av denne vurderingen har vært diskutert, fordi den kun teller 10 % av den totale karakteren. Hvis rektorer hadde en mer fremtredende rolle i vurderingen, kunne de ha koblet vurderingen mer direkte til utviklingen av lærerprofesjonalitet og i større grad brukt vurderingen i utviklingsøyemed.

Den fjerde vurderingsmetoden som blir benyttet, er egenvurdering. Denne metoden består av et strukturert spørreskjema med påstander hentet fra undervisningsstandarden. Lærerne skårer seg selv på en skala fra 1 - 4. Fra 2012 er et kommentarfelt knyttet til skåren «fremragende» slik at hver gang denne skåren blir brukt må lærerne beskrive atferd eller praksis som støtter en slik vurdering.

Ulempen med at egenvurdering inngår i vurderingssystemet og sluttskåren til lærerne, er at det kan gå inflasjon i høye skåreverdier. Forskerne spør derfor om denne vurderingsformen skal vektlegges i den totale karakteren, men siden vektingen er lovfestet, krever dette en lovendring.

Det nasjonale vurderingssystemet er formativt (basert på veiledning underveis), men er samtidig summativt med et klart resultatfokus. Lærerne får en av følgende fire sluttkarakterer etter at resultatene fra de ulike vurderingsmetodene er slått sammen: «fremragende», «kompetent», «grunnleggende» og «ikke tilfredsstillende». Lærere som skårer i de to første kategoriene, kvalifiserer til lønnsøkning etter at de også har bestått en faglig kunnskapstest, mens lærere som havner i de to siste kategoriene blir pålagt veiledning. Klarer de ikke å forbedre seg, risikerer de å miste jobben.

I 2011 vedtok Kongressen en lov som skjerpet konsekvensene av vurderingen. Lærere som får karakteren «grunnleggende», og som ikke klarer å forbedre karakteren til «kompetent» ved neste vurdering, står i fare for å miste retten til å undervise. Videre kan rektorer også fjerne 5 % av lærerne fra undervisningen basert på karakterene «grunnleggende» og «ikke tilfredsstillende». Bonusen for lærere som fikk de to beste karakterene, ble også økt. Videre ble den formative hensikten med systemet eksplisitt nevnt i lovteksten.

Det er opprettet en egen nettside for lærerne med informasjon om vurderingsprosessen. Dette omfatter både vurderingsmateriell, bakgrunn, lover og regler samt en «ofte stilte spørsmål»-kategori. Det er også opprettet et «callsenter» hvor lærere som arbeider med innholdet i mappen sin kan få svar på ulike spørsmål. Gjennomføringen av evalueringen tar et helt år. Arbeidet begynner med at man utarbeider listen over lærerne som skal evalueres det året og slutter når resultatene av vurderingen presenteres.

De som er ansvarlige for implementeringen i kommunene, får opplæring i systemet og de er også kontaktpersoner for lærerne som blir vurdert. Hvert år blir informasjonen fra vurderingssystemet registrert i et eget dataprogram utviklet for dette formålet. Deretter blir dataene analysert og rapporter blir generert på tre nivåer, for hver lærer, skole og fylke (Taut og Sun 2014).

Resultatene fra lærervurderingen viser at tallene fra 2003 til 2010 er relativt stabile. Majoriteten (52,4 % til 64 %) av lærerne fikk karakteren «kompetent». En tredjedel av lærerne fikk karakteren «grunnleggende». Det var kun en liten prosentandel (6 % i 2010) som fikk karakteren «fremragende» og 2,6 % i 2010 som fikk karakteren «ikke tilfredsstillende». Når resultatene fra de ulike vurderingsmetodene ble sammenlignet, var det skårene på mappen som trakk gjennomsnittet ned, mens andre metoder (spesielt egenvurdering), dro gjennomsnittet opp.

Resultat fra forskning knyttet til vurderingssystemet i Chile

Systemet for lærervurdering har nå vært brukt i ti år i Chile. Elevresultater er ikke en del av det nasjonale vurderingssystemet for lærere, men de brukes for å validere resultatene fra vurderingssystemet og de ulike metodene som benyttes i systemet.

Det blir særlig fremhevet at bred delaktighet og en bredt forankret implementeringsprosess var med på å sikre det nasjonale vurderingssystemet troverdighet og legitimitet. En annen viktig side ved det å utvikle politikk for lærervurdering, er at definisjon av formål og bruk av resultater er gjennomtenkt og velfundert. Videre bør den underliggende antakelsen om hva man vil oppnå gjennom utviklingen av vurderingssystemet være klart formulert, slik at det kan følges og evalueres over tid.

Den viktigste mekanismen som ble utformet for å innfri den formative ambisjonen, altså profesjonsutviklingen, er det som kalles profesjonelle utviklingsplaner. Alle lærere som får karakteren «ikke tilfredsstillende» og «grunnleggende» skal ha slike planer. Taut og Sun (2014), påpeker at dette er et av de viktigste tiltakene i utformingen av det nasjonale systemet for lærervurdering, men også det som er dårligst utviklet. Forskerne foreslår at en slik utviklingsplan heller burde vært utformet slik at den ble oppfattet som en nødvendig og attraktiv læringsmulighet i stedet for en administrativ, stigmatiserende og formell forpliktelse for lærere som kommer dårlig ut i vurderingen. Analyser av mappene viser at de fleste lærerne har forbedringsmuligheter i sitt pedagogiske arbeid. Et eksempel som trekkes frem er at flere lærere med samlekarakteren «kompetent» i 2012 kun fikk karakteren «grunnleggende» på mappevurderingen (N= 9249 (84,2 %) av N=10989). Derfor bør det vurderes å gi alle lærere profesjonelle utviklingsplaner.

Forskningen på kvaliteten på vurderingsmetodene som er benyttet i lærervurderingssystemet (validitet, reliabilitet og equity), samt ikke-intenderte konsekvenser av systemet har over tid ført til noen justeringer og forbedringer av systemet.

I en av artiklene som undersøker det chilenske lærervurderingssystemet (Taut m.fl. 2011), gjengis en undersøkelse som er gjennomført i 30 skoler i 10 forskjellige kommuner. 57 skoleledere er intervjuet om hvordan de har brukt resultater fra de årlige lærervurderingene i skolene sine og hvilke erfaringer– både positive og negative – de har med systemet. I 2009 fikk 63,1 % av lærerne «kompetent», mens 28,9 % fikk «grunnleggende». Bare 6,5 % havnet i kategorien «fremragende» og 1,5 % fikk «ikke tilfredsstillende» (Taut m.fl., 2011 p. 219). Artikkelen konkluderer med at skolelederne som er intervjuet, har delte oppfatninger om systemet. Jo mer aktivt lederne har vært engasjert i selve vurderingsprosessen, jo større legitimitet tilskriver de systemet for lærervurdering. Av positive virkninger nevner skolelederne at de har registrert mer teamarbeid og interne refleksjonsprosesser på skolen, noe de betrakter som økt profesjonalisering. Alle skolene rapporterer samtidig om at systemet har gitt lærerne en stor arbeidsbelastning og at enkelte lærere opplever vurderingen som emosjonelt belastende. Forskerne mener at systemet påfører lærerne svært mye arbeid som de ikke får kompensasjon for, og at den formative siden ved vurderingen – altså oppfølging av lærerne – er for dårlig ivaretatt og må styrkes.

En annen artikkel (Santelices og Taut 2011) spør om det chilenske lærervurderingssystemet er i stand til å identifisere de «riktige» lærerne som presterer henholdsvis «fremragende» og «ikke tilfredsstillende». Forskerne valgte ut 58 lærere som i 2005 ble evaluert som enten «fremragende» eller «ikke tilfredsstillende», for nærmere undersøkelse. Følgende data ble samlet inn om lærernes undervisning: tre klasseromsobservasjoner, ekspertvurderinger av mapper, samt resultater fra en faglig og en pedagogisk kunnskapstest. Resultatene viste at viktige forskjeller mellom lærere som fikk karakteren «fremragende» og de som fikk karakteren «ikke tilfredsstillende», handlet om hvordan lærere strukturerte timene, hvordan studentene oppførte seg i timene, hvordan lærerne utformet vurderingsmateriale i klasserommet og hvor

godt lærerne klarte å sikre at alle studentene arbeidet konsentrert mesteparten av tiden. Forskjellene her er statistisk signifikante og har stor praktisk betydning fordi det som er undersøkt er lærerpraksiser.

Når de vurderte lærernes praksis i klasserommet og rangerte lærernes undervisningsmateriell, fant forskerne både praktiske og signifikante forskjeller mellom lærerne i kategorien «fremragende» og kategorien «ikke tilfredsstillende». Det var også signifikante forskjeller mellom de to gruppene av lærere på den standardiserte kunnskapstesten.

Lærere som ikke vil bli vurdert

Noen lærere nekter å delta i den nasjonale vurderingen, og Tornero og Taut (2010) har undersøkt hva som er årsaken til dette. Det nasjonale systemet for lærervurdering ble gjort obligatorisk i 2005 og lærerne som blir vurdert får, som tidligere nevnt, en av følgende fire karakterer: «fremragende», «kompetent», «grunnleggende» og «ikke tilfredsstillende». Hvis en lærer får karakteren «ikke tilfredsstillende» tre ganger på rad, blir vedkommende gitt en økonomisk kompensasjon og får ikke lenger praktisere som lærer. Dette har nå blitt skjerpet inn til to ganger. Hvis en lærer nekter å delta i vurderingen, får også vedkommende karakteren «ikke tilfredsstillende». Disse lærerne får ingen økonomisk kompensasjon. I 2010 var det 2,6 % (N= 11 061) av de evaluerte lærerne som fikk karakteren ikke tilfredsstillende (Taut og Sun 2014).

Ved å gjennomføre dybdeintervju med 9 lærere fant forskerne følgende årsaker til at lærerne nektet å la seg vurdere: 1) ekstra tidsbruk og arbeidsmengde, 2) overskridelse av prinsipper, verdier og grunnsyn og 3) nært forestående pensjonsalder. I tillegg ble følgende faktorer trukket frem: konsekvenser av vurderingssystemet, ulike oppfatninger av vurderingssystemets legitimitet, andre kriterier for å bedømme undervisningspraksis, manglende informasjon og kunnskap om vurderingssystemet og lærernes profesjonskultur. Lærerne som nektet å la seg vurdere, mente for det første at vurderingen medfører en betydelig arbeidsinnsats og tidsbruk for den enkelte lærer. For det andre hadde vurderingen negative konsekvenser for de som skårer lavt, det vil si for de som får karakteren «ikke tilfredsstillende» eller «grunnleggende». For det tredje blir det problematisert at det nasjonale vurderingssystemet har et eksternt «mandat».

Taut og Sun (2014) mener at det chilenske systemet har vist at det er mulig å få til både summativ og formativ vurdering i et system for lærervurdering, men at dette ikke er uproblematisk og forutsetter komplekse og utfordrende forhandlinger.

Da OECD nylig evaluerte det chilenske systemet for lærervurdering (Santiago m. fl. 2013), var de atskillig mer kritiske. De fant at den opprinnelige ambisjonen om at systemet skulle fungere formativt, ha en klar utviklingskomponent og styrke lærernes kompetanse, ikke har fått nok oppmerksomhet. Systemet brukes stort sett til å holde lærerne ansvarlige for elevresultater. Det er store forskjeller mellom kommunene når det gjelder kompetanse og ressursinnsats. Lærerne får ofte ikke de tilbakemeldingene de trenger, og det gjennomføres for sjelden faglige, profesjonelle samtaler mellom lærere og vurderere om lærernes undervisningspraksis. Skolelederne deltar for lite i vurderingsarbeidet, og det blir for sjelden laget utviklingsplaner for lærerne. Oppfølgingen av undervisningspersonalet blir ofte noe tilfeldig fordi skolelederne heller ikke alltid har den nødvendige vurderingskompetansen eller tid til å prioritere dette arbeidet. I tillegg viser det seg at mange lærere mangler den kunnskapen som trengs både for å vurdere eget arbeid, og når de skal vurdere kollegaer. Lærerne har heller ikke tillit til at de som setter karakter på mappene deres har den nødvendige kompetansen. OECD anbefaler derfor at Chile styrker vurderingskompetansen på alle nivåer og gjør en innsats for å lage bedre sammenheng mellom de ulike delene i rammeverket.

3.1.2 Kina

Et annet land hvor lærervurdering har vært – og fortsatt er – et omdiskutert tema er Kina. Frem til 1980-årene var lærervurdering knapt nevnt i kinesisk utdanningspolitikk. I 1985 ble det satt i gang en reform av utdanningssystemet med lærerutdanning og vurdering av lærere i grunn- og videregående skole som prioriterte satsingsområder. I 1994 kom en lov som la det rettslige grunnlaget for vurdering av lærere, basert på følgende faktorer; politisk overbevisning, kompetanse, holdninger, og prestasjoner. Resultater fra vurderingen er med på å avgjøre hvorvidt lærerne får forlenget sine ansettelseskontrakter, om de får lønnsopprykk eller går ned i lønn, samt type ansettelsesforhold (Zhang og Ng 2011).

Fra summativt til et mer formativt vurderingssystem i Kina

Lærervurdering i Kina har hovedsakelig hatt summative formål, noe som også har preget ledelse og gjennomføring av vurderingene. I den senere tid har lærerprofesjonen blitt mer anerkjent, og lærerne oppfatter i økende grad lærervurdering som et virkemiddel for profesjonsutvikling. I 2001 kom nye retningslinjer for en lærerplanreform i grunnskolen. Her etterlyses etableringen av et vurderingssystem som tar sikte på å legge til rette for kontinuerlig utvikling av lærerkompetanse. Kina innførte også prestasjonslønn for lærere i 2009, noe som har ført til økt interesse for lærervurdering.

Utdanningssystemet i Kina er svært sentralisert. Skolene er underlagt sterk styring fra regjeringen. Policy og kriterier for vurdering blir ikke bestemt av skolen selv, men utvikles i tråd med gjeldende nasjonal og regional politikk. Lærervurdering blir også brukt til andre formål enn lokalt på skolen, som for eksempel i konkurranser om å vinne pris som «fremragende lærer» eller lignende. Videre står det kollektive sterkt i Kina. Lærere forventes å jobbe som del av en gruppe, og lærere blir derfor vurdert både individuelt og kollektivt.

Gjennomføring av lærervurdering og tilrettelegging for profesjonsutvikling

Zhang og Ng (2011) har undersøkt hvordan lærervurdering blir gjennomført og hvordan lærere og skoleadministrasjonen oppfatter sammenhengen mellom lærervurdering og lærernes profesjonsutvikling. Ved hjelp av en kvalitativ casestudie fra en skole i Shanghai samlet Zhang og Ng (2011) inn data gjennom dybdeintervju, dokumentanalyse og observasjon. 24 informanter ble intervjuet; rektor, viserektor, tre mellomledere, fire ledere av ulike fagavdelinger og 13 lærere med ulik fagbakgrunn samt en leder fra distriktets lærerhøyskole. Det ble samlet inn policydokumenter på skole- og systemnivå, regler, lover og manualer knyttet til lærervurdering og profesjonsutvikling, vurderingsmetoder, opptak, notater, skolens mål og utviklingsplaner. Interaksjon mellom ansatte ble observert under møter i forbindelse med gjennomføringen av selve vurderingen, samt timer og konferanser som ble holdt i etterkant.

På skolen i Shanghai ble lærerne vurdert ved hjelp av følgende fire kategorier; moralsk oppførsel og holdninger, kompetanse, prestasjoner og oppgaver. Selve vurderingen starter med lærernes egenvurdering. De fyller ut et vurderingsskjema hvor de skårer seg selv i forhold til en gitt standard. Deretter blir lærerne evaluert i egen avdeling av et vurderingspanel som består av tre til fire avdelingsmedlemmer, inkludert avdelingsleder, faglig gruppeleder og vanlige lærere. Vurderingspanelet gir lærerne en skåre basert på de samme kriteriene som brukes i lærernes egenvurdering. Skolens vurdering er siste steg i prosedyren. Vurderingskomiteen består av skoleledere og mellomledere som i hovedsak vurderer lærerne ut i fra elevenes

eksamensresultater, elevenes vurderinger av lærerne, klasseromsobservasjon, inspeksjon av lærernes daglige arbeid og resultater fra avdelingsvurderingen. Vurderingsdata samles inn gjennom hele året, mens egenvurdering, gruppevurdering og skolevurdering, gjennomføres på slutten av hvert semester.

Resultatene fra vurderingen blir brukt til å rangere lærerne på følgende skala; nivå A (fremragende), nivå B (kompetent), nivå C (bestått), nivå D (ikke bestått). Basert på resultatene fra vurderingen får lærerne utbetalt bonuslønn på slutten av hvert semester. Videre får lærere som presterer bra ulike former for anerkjennelse («awards» og «honors»). Analysen av casestudien viser at systemet til en viss grad støtter lærernes profesjonsutvikling. Det gir profesjonell anerkjennelse og virker motiverende gjennom bruk av prestasjonslønn. Gjennom de ulike kriteriene og mål som lærerne skal jobbe mot i sitt daglige arbeide gir vurderingen retningslinjer og tilbakemelding underveis i vurderingsprosessen. Videre viste resultatene at den obligatoriske klasseromsobservasjonen, som alle lærerne må igjennom, bidrar til å kvalitetssikre lærernes profesjonsutvikling. Det samme gjelder kravene om å delta i videreutdanning for å få mulighet til å oppnå en høyere profesjonstittel.

Zhang og Ng (2011) understreker at det er viktig å forstå lærervurdering og profesjonsutvikling i lys av kinesisk kultur for utdanning, og nevner særlig tre trekk som er viktige i forhold til læreres profesjonsutvikling. Det er: 1) høye forventninger til utvikling, 2) kontroll og 3) en sterk kollektiv orientering, som hindrer innbyrdes konkurranse blant lærerne. Zhang og Ng (2011) påpeker til slutt at et vurderingssystem som i høy grad baserer seg på kontroll og insentiver kan føre til konform praksis.

Innføring av prestasjonslønn og endring av vurderingspraksis

En annen artikkel som har sett på lærervurdering i Kina (Liu og Zhao 2013), analyserer innføringen av prestasjonslønn for lærere. Innføringen av lærervurdering kan analyseres i tre perioder: a) før 2001, b) 2001-2009 og c) 2009 frem til i dag. Prestasjonslønn ble innført i 2009, da det kinesiske utdanningsdepartementet vedtok «Retningslinjer for implementeringen av prestasjonslønn i den obligatoriske skolen». Dermed ble det behov for et nytt system for lærervurdering i Kina.

I artikkelen ser Liu og Zhao (2013) nærmere på perioden etter 2006. Artikkelens datagrunnlag omfatter internasjonale og kinesiskspråklige artikler fra vitenskapelige tidsskrift. I tillegg ble det samlet inn data fra to skoler i samme distrikt i Beijing for å belyse ulikheter i vurderingspraksisen etter innføringen av prestasjonslønn. Data bestod av dokumenter og fire intervju. De som ble intervjuet var rektor, viserektor som er ansvarlig for undervisning, en vanlig lærer og en «Banzhuren»[25]. I intervjuene ble informantene spurt om hvilke endringer de hadde observert i skolens praksis når det gjaldt lærervurdering etter innføringen av prestasjonslønn for lærere i 2009.

På skole A, som var en grunnskole i Beijing (40 klasser, 1883 elever og 132 lærere og andre ansatte), ble lærerne vurdert månedlig og årlig av skolens egen vurderingskomite. Denne komiteen bestod av ledere og lærerrepresentanter, der lærerrepresentantene utgjorde en tredjedel av komitemedlemmene. Disse bestemte om lærerne kvalifiserte til å få belønning basert på prestasjoner, oppgaver, undervisning og ekstra arbeid. Skole A vedtok to dokumenter ved innføringen av prestasjonslønn. Det ene var indikatorer knyttet til lærerens prestasjoner. Disse inneholdt a) politisk ståsted, b) yrkesmoral, c) holdning til arbeidet, d) undervisningsferdigheter, e) evne til forskning f) arbeidsprestasjon som inkluderte kvalitet og kvantitet på utført arbeid. Det andre dokumentet dreide seg om hvordan lærere skulle belønnes (eller utsettes for

25 Banzhuren er en vanlig lærer som har ansvar for klasseledelse som for eksempel: ansvarlig for studentenes oppførsel og sikkerhet, ansvarlig for å kontakt med foreldre og koordinere med andre lærer hvem det er som skal undervise klassene. En Banzhuren har også vanlig undervisning i tillegg til denne rollen.

sanksjoner). Den månedlige vurderingen rangerte lærerne på følgende skala: «fremragende», «tilfredsstillende», «akseptabelt» og «ikke akseptabelt». Resultatene fra vurderingen blir deretter benyttet for å kalkulere hver lærers månedslønn. Vurderingen på Skole A kan dermed karakteriseres som et system basert på belønning og straff.

Det nye systemet omfattet flere målinger av lærerne, ikke bare studentenes presentasjoner. Lærerne fikk månedlige tilbakemeldinger. Siden vurderingen var knyttet opp mot prestasjonslønn, tror forskerne at lærerne var ivrige etter å fikse problemene som hadde gjort at de ikke fikk belønning måneden før.

Skole B var en ungdomsskole i utkanten av Beijing (10 klasser, 197 elever og 59 lærere). Etter innføring av prestasjonslønn for lærere initierte skolen et nytt program for lærervurdering. Dette inkluderte 1) læreplan, 2) notater om profesjonsutvikling, 3) undervisningsplan, 4) daglig undervisning, 5) vurdering av elevenes arbeid, 6) demonstrasjon av undervisning, 7) undervisningskvalitet, 8) analyser av elevenes tester 9) tilbakemelding fra foreldre, 10) deltagelse i studier av undervisning, 11) antall besøk i andre læreres klasserom, 12) egenrefleksjon, 13) sammendrag av lærerens årlige arbeid 14) politisk ståsted, 15) yrkesmoral og 16) Banzhuren-arbeid. For å danne seg et inntrykk av lærernes undervisningsprofil, besøkte skolens vurderingskomite lærernes klasserom en eller to ganger i semesteret. Det ble gjort endringer etter at en spørreundersøkelse til foreldrene ble gjennomført på slutten av semesteret. Skolens rektor oppfattet endringene som forsøk på å etablere et flerdimensjonalt og mer prosessorientert system for lærervurdering.

Gjennom intervjurunden ble det identifisert noen problemer ved systemet. For det første ble alle lærerne vurdert årlig i stedet for etter hvert semester. På slutten av hvert skoleår måtte lærerne selv presentere sin egen profesjonsutvikling, samt undervisningsresultater og andre resultater. Basert på denne presentasjonen fikk lærerne karakteren «fremragende» eller «tilfredsstillende». I løpet av vurderingsprosessen viste deg seg at personlige bånd/vennskap var viktigere enn hardt arbeid for å oppnå en god karakter. Siden det kun er to karakterer å velge imellom, tar ikke lærerne dette seriøst. Et annet problem er at når vurderingen ikke var knyttet til prestasjonslønn, brydde heller ikke lærerne seg om resultatet av vurderingen.

Liu og Zhao (2013) bemerker at kinesiske ledere har en utfordring i forhold til hvordan de skal forholde seg til belønning og sanksjoner når målet er at systemet for lærervurdering i større grad skal virke formativt og ha innvirkning på lærernes profesjonsutvikling.

3.1.3 Belgia

I 2007 ble det innført en ny lærervurderingspolicy i flamske videregående skoler. Det nye systemet for lærervurdering skal ha både summative og formative formål. Skolene er pålagt å gjennomføre prestasjons- og vurderingsintervju med hver lærer hvert fjerde år. Fireårsperioden starter med at det oppnevnes en evaluator som er en overordnet, f. eks rektor. Prestasjonskriteriene som læreren skal vurderes etter, er nedfelt i individuelle jobbeskrivelser, og danner grunnlaget for vurderingssystemet. I løpet av fireårsperioden skal lærer og evaluator ha minst ett prestasjonsintervju eller en samtale som skal fungere lærende (ha en formativ funksjon). I løpet av denne samtalen får læreren tilbakemelding på prestasjonene sine og råd om hvordan de kan forbedres. Den fireårige vurderingsperioden avsluttes med en vurderingssamtale mellom læreren og hans/hennes evaluator hvor oppmerksomheten rettes mot de resultatene læreren har oppnådd. Deretter skriver evaluator en rapport som konkluderer med enten «tilstrekkelig» eller «utilstrekkelig». Hvis den endelige konklusjonen er «utilstrekkelig», kan læreren anke. Fører ankebehandlingen til at konklusjonen blir stående, må læreren gjennomgå en ny formell vurdering innen utgangen av neste år. Får læreren «utilstrekkelig» to ganger på rad, kan det føre til oppsigelse. Det flamske utdanningsdepartementet

poengterer at dette kun skal skje i ekstremtilfeller. Det overordnede målet med systemet er at lærervurderingen skal fungere formativt og bidra til å forbedre undervisningen på de flamske skolene.

Lærervurderingens innvirkning på læreres profesjonsutvikling

Delvaux og Vanhoof (2013) har undersøkt hvilke deler av det belgiske systemet for lærervurdering som – sett fra lærernes perspektiv – bidrar til profesjonsutvikling. I studien har de sett nærmere på hensikten med vurderingen, kjennetegn ved ledelse og oppfølging, og spesielle trekk ved vurderingssystemet. Særlig har de undersøkt hvor klart formulert den overordnede hensikten med systemet og vurderingskriteriene er, om systemet oppleves som rettferdig, og hvor fornøyde lærerne var med prestasjons- og vurderingsintervjuene. Et utvalg på 1983 lærere på 65 skoler i Flandern svarte på spørreskjema. I utvalget hadde 28,3 % (N=560) mindre enn 5 års erfaring, 47,1 % (N=930), hadde mellom 5-20 års erfaring og 24,6 % (N=486) hadde mer enn 20 års erfaring. Det ble benyttet en flernivåanalyse på resultatene fra spørreskjemaet.

Analysene viser at lærerne generelt i liten grad opplever at vurderingssystemet styrker deres profesjonsutvikling, men standardavviket er høyt og det er stor variasjon i de rapporterte effektene. Flere lærere med kort undervisningserfaring (mindre enn 5 år), rapporterer at de har nytte av vurderingssystemet for egen profesjonsutvikling. Videre trekkes det frem at tilbakemeldinger oppleves som nyttige og at skolelederens holdning til systemet er viktig for om det skal fungere etter intensjonene. Gitt disse tre forholdene ser det ut som om lærervurderingen kan virke positivt inn på lærernes opplevelse av at vurdering bidrar til profesjonsutvikling

Delvaux og Vanhoof (2013) finner ikke effekt av systemets formative intensjon på lærernes profesjonsutvikling. Vurderingssystemets summative formål har en liten, men signifikant positiv innvirkning på lærernes opplevde profesjonsutvikling. Videre trekkes det frem noen kjennetegn ved skoleledelsen som har betydning for lærernes profesjonsutvikling. Det omhandler at rektor legger til grunn at vurderingssystemet skal fremme profesjonsutvikling og om rektors engasjement som undervisningsleder. Forskerne finner at lærerne opplever at noen trekk ved systemet har en positiv innvirkning på deres profesjonsutvikling: At vurderingskriteriene er tydelige har en liten, men signifikant effekt. Hvor fornøyde lærerne var med intervjuene og opplever at de har en positiv effekt på deres profesjonsutvikling, avhenger av at forholdet mellom evaluator og lærer er positivt og at lærerne tilskriver evaluator legitimitet. Det trekket som har størst betydning er hvor stor nytte lærerne synes de har av tilbakemeldingene.

Artikkelen konkluderer med at følgende faktorer kan betraktes som suksesskriterier i implementering og gjennomføring av lærervurdering: undervisningsledelse bør være et sentralt tema og rektor bør ha en positiv holdning til den formative siden av vurderingen, tilbakemeldinger må oppleves som nyttige og de som blir vurdert, må ha tillit til de som vurderer. Legitimitet og troverdighet er nøkkelord her.

3.1.4 Portugal

Portugal har de siste tiårene innført flere reformer som har hatt til hensikt å bedre elevenes prestasjoner og heve kvaliteten på undervisningen. Et ledd i dette arbeidet har vært implementeringen av et system for lærervurdering med dobbelt formål; både forbedring og karriereprogresjon. Flores (2012) har sett nærmere på endringer knyttet til lærervurdering og lærerkarrierer – spesielt implementeringen av vurderingssystemet, altså hvilke oppfatninger lærere har av politikken og det nye systemet for lærervurdering, hva de mener om implementeringen og om den nye politikken har påvirket dem og skolen.

Oppbygging av systemet for lærervurdering i Portugal.

I 2007 utstedte regjeringen en ny vedtekt, hvor målet var å utvikle et system med karriereveier for å identifisere, fremme og belønne lærernes prestasjoner. Systemet for lærervurdering ble tilpasset lærernes mandat og arbeidsoppgaver langs følgende fire dimensjoner; 1) profesjonell, sosial og etisk dimensjon, 2) utvikling av undervisning og læring, 3) deltakelse i skoleaktiviteter og forhold til samfunnet, og 4) læring og utvikling i et livslangt perspektiv. Disse fire dimensjonene er igjen delt opp i ulike domener og indikatorer.

I 2008 og 2009 ble systemet for lærervurdering forenklet og revidert for å imøtekomme skepsis og motstand blant lærere og lærernes fagforeninger. I 2010 formulerte også regjeringen en nasjonal prestasjonsstandard som skiller mellom fem nivå: «fremragende», «svært godt», «godt», «tilfredsstillende» og «ikke tilfredsstillende». Et nøkkelelement i lærervurderingssystemet er egenvurdering som bidrag til profesjonsutvikling. På hver skole (eller klynge av skoler) blir det nedsatt en komité som skal koordinere hele prosessen rundt lærervurderingen. Selve vurderingsprosessen er utviklet av komiteen og den som skal gjennomføre vurderingen. Den som skal vurdere (evaluator), er ansatt i departementets læreplanavdeling. En koordinator i departementet har ansvar for å koordinere vurderingen og ha tilsyn med vurdererens arbeid. Den som vurderer er ansvarlig for utviklingen av vurderingsprosessen og må være i kontinuerlig dialog med de som blir vurdert for å ivareta det formative aspektet ved vurderingsprosessen. Vurderingsmetodene som blir benyttet er egenvurderingsrapport og et observasjonsskjema.

Lærernes oppfatning av systemet for lærervurdering

I studien har Flores (2012) undersøkt hvordan respondentene betrakter selve hensikten med systemet for lærervurdering. Datagrunnlaget for studien er 150 spørreskjema, 45 semi-strukturerte intervju og 10 fokusgrupper som hver hadde tre lærere. 74 % av lærerne var sertifiserte, og de hadde fra 0 til 36 års erfaring i yrket.

Hun finner at lærervurdering, etter lærernes mening, bør konsentrere seg om å identifisere behovet for profesjonsutvikling (67,9 %) samt at det bør ha som siktemål å gi nyttig informasjon som hjelper lærere til å forbedre arbeidet sitt (86,9 %). Respondentene var også opptatt av at lærervurdering bør ta sikte på å øke lærernes refleksjon over egen praksis (90,4 %). På den annen side viser undersøkelsen at lærerne er ganske negative til implementeringen av systemet og skeptiske til hvilke effekter den nye politikken vil få for skolen. Dette er spesielt knyttet til formål og hensikt med systemet, selve implementeringsprosessen, samt at de ikke alltid anerkjenner fagligheten til de som vurderer. De synes at de har fått mangelfull informasjon og opplæring, systemet fører med seg mer byråkrati og det blir ikke satt av nødvendig tid til å gjennomføre alle de tiltakene som et slikt komplekst system fører med seg.

Et hovedtema som løftes frem i artikkelen er lærernes skepsis og den manglende tilliten til det nye systemet. Til tross for intensjonen med systemet om profesjonsutvikling, god skoleutvikling og økt samarbeid mellom lærerne, har implementeringen av systemet ført til en praksis som er summativ, byråkratisk og preget av formaliteter.

I følge Flores (2012) har også noen av respondentene positive oppfatninger av deler av systemet for lærervurdering. Dette handler om at et system for vurdering kan øke lærernes image som profesjonelle i offentligheten, at det kan avdekke hvor krevende læreryrket er og at et system for vurdering kan bidra til å «vekke» noen lærere som kan trenge det. Artikkelen konkluderer med at det aller viktigste i et system for lærervurdering er at det er klart orientert mot utvikling, at det gjennomføres etter formative prinsipper og at tilbakemeldingene gis i rett tid.

Hva viser eksemplene?

Resultatene fra forskningen som har sett på ulike system for lærervurdering i landene Chile, Kina, Belgia og Portugal, viser at alle systemene hadde doble intensjoner. De skulle både være formative, dvs. bidra til lærernes profesjonsutvikling og summative dvs. måle resultater. Forskningen avdekket at i samtlige system tok resultatfokuset overhånd og de formative sidene ved systemet ble ikke godt nok fulgt opp. Dette resulterte gjerne i at systemene ble redusert til omfattende byråkrati i stedet for å bidra til den faglige utviklingen ved skolene.

3.2 Forskning med vekt på summativ vurdering

Etter at det på 1980-90-tallet ble innført nye styringsmodeller i utdanningssektoren, har interessen for elevenes læringsutbytte kommet i sentrum for utdanningspolitikken i store deler av verden. Det vanlige er å måle elevenes læringsutbytte gjennom symboler, karakterer, tall og prosenter, men hva elevene har lært kan også dokumenteres og beskrives på andre måter. Lærere gir for eksempel elever løpende tilbakemeldinger på innsats, arbeidsmåter, sosiale ferdigheter, initiativ og samarbeid. Hittil er det den delen av elevenes læringsutbytte som kan tallfestes, måles og rangeres (den summative vurderingen) som har fått størst oppmerksomhet. I dette kapitlet presenteres studier som har undersøkt konsekvenser av at elevenes (målbare) læringsresultater blir lagt til grunn i lærervurdering. Først presenteres noen studier som har sett på resultater og skolens læringsmiljø, deretter studier som har undersøkt effekt av prestasjonslønn og til slutt studier i tradisjonen Value-Added.

De siste ti årene har svakt læringstrykk blitt nevnt som et av de store problemene i norsk skole. Dette er påpekt som en mulig årsak til at norske elever ikke skårer bedre enn de gjør på internasjonale undersøkelser. I følge Stortingsmelding nr. 30 (2003-04) *Kultur for læring* er det viktig at skolen og lærerne sørger for tilstrekkelig ytre trykk i undervisningen hvis elevens indre motivasjon er svak (s. 55). Antakelsen er at hvis man øker læringstrykket, vil også elevenes læringsutbytte øke. Spørsmålet blir da hvordan man kan skape økt læringstrykk. En beslektet antakelse er at organisasjonskultur og skoleledelsens handlinger har betydning for elevenes læringsutbytte. Dermed må man kunne anta at ledere som er opptatt av elevenes læringsutbytte, bidrar til en resultatorientert kultur. Derfor er det interessant å undersøke om en resultatorientert ledelseskultur får lærere til å gjøre en ekstra innsats for å få elevene til å arbeide hardere. Med andre ord: Vil lærere utsette elevene for større læringstrykk hvis ledelsen ved skolen er resultatorientert? I og med at dette handler om trekk ved skolers organisasjonskultur, har Christophersen m. fl. (2012) undersøkt hvordan en

resultatorientert kultur påvirker lærernes innsats for å få elevene til å anstrenge seg mer (altså øke læringstrykket).

Studien har sammenlignet 236 lærere som arbeider i skoler preget av mange tester og resultatpress med 366 lærere i norske folkehøyskoler – hvor det verken finnes prøver eller tester, og hvor det følgelig er en annen resultatorientering. Ved å sammenligne holdningene i disse to lærergruppene, skal det være mulig å finne ut om økt resultatorientering er et kulturtrekk som motiverer lærerne til å øke læringstrykket i undervisningen sin. Forskerne finner at kvaliteten på relasjonene mellom ledere og lærere har betydning for kvaliteten på lærernes arbeid, men at det ikke er mulig å se noen målbar effekt av en resultatorientert kultur på lærernes holdninger. Artikkelen konkluderer med at det slett ikke er sikkert at det er en lineær relasjon mellom eksternt press, holdninger og atferd. Derimot er relasjoner preget av tillit en sterk kvalitetsindikator i skolene som er undersøkt – uavhengig av ledelsens resultatorientering. Det er altså ikke nødvendigvis en lineær relasjon i praksis, hvor det ikke alltid er samsvar mellom holdninger og handlinger. Det er heller ikke sikkert at eksternt press alltid virker motiverende på innsats.

I en annen artikkel som også ser på forholdet mellom resultater og læringsmiljø tar Barile m. fl. (2012) utgangspunkt i at stater i USA nå prøver ut mange forskjellige modeller for lærervurdering som bygger på ulike oppfatninger av læreryrket. Det er imidlertid lite forskning som ser nærmere på resultatene av de forskjellige tiltakene. Ved bruk av flernivåanalyse undersøker de mulige langsiktige sammenhenger mellom lærervurdering og belønningssystemer på den ene siden og elevenes matematikkresultater og frafall på den andre. Utvalget er 7.779 elever fra 431 offentlige videregående skoler. Det var like mange gutter som jenter i utvalget; elevene var rundt 16 år og 53 % av dem var hvite. Studien så på betydningen av lærer-elevrelasjoner for sammenhenger mellom tiltak og elevenes læringsutbytte. Forskerne har tre konklusjoner: For det første var det bedre læringsmiljø på skoler hvor elevene svarte at de fikk anledning til å vurdere lærerne sine. For det andre svarte studentene at det var et dårlig læringsmiljø på skoler som hadde innført en form for belønningspolitikk som gikk ut på å la de beste lærerne undervise de flinkeste elevene. For det tredje var det lavere frafall på skoler hvor elevene rapporterte om et godt læringsmiljø.

I en undersøkelse av kulturens betydning for elevenes læringsutbytte bruker Fang m. fl. (2013) Hofstedes seks kulturelle dimensjoner og to dimensjoner fra World Value Survey til å undersøke sammenhenger mellom nasjonal kultur og elevens læringsutbytte. De gjennomførte flere regresjonsanalyser for å avdekke hvilke variabler som kunne forutsi elevenes resultater, slik de ble målt i PISA 2009. Analysen viser at de viktigste variablene for elevenes læringsutbytte på alle tre læringsområdene (lesing, matematikk og naturfag), inkluderte følgende kulturelle dimensjoner: 1) Kulturell holdning til utholdenhet, fremtidsrettet orientering og måloppnåelse og 2) Kulturell oppmerksomhet om sekulære og rasjonelle verdier vs. tradisjonelle verdier.

For å undersøke hva som kan forklare de store variasjonene mellom elevenes læringsutbytte i ulike klasserom, kombinerer Kane m. fl. (2011) data fra observasjoner av læreres undervisning med mål på lærernes bidrag til å bedre elevenes læringsutbytte. De finner at noen undervisningspraksiser kan forutsi resultatoppnåelse bedre enn andre og at observasjonsmål på lærerdyktighet absolutt er koblet til økninger i elevenes læringsutbytte. Å koble observasjon av undervisning til tall som viser økninger i elevenes læringsutbytte, kan gi verdifull informasjon for tiltak som tar sikte på å fremme lærernes profesjonsutvikling. For eksempel viser denne undersøkelsen at det er lærerens totalskåre som har størst betydning. Om man skal oppnå bedring av elevenes læringsutbytte, må man tenke helhetlig og satse på hele bredden av lærerkompetanse. For det andre viser det seg at det hjelper godt på elevenes læringsutbytte om lærerne gjør en innsats for å bedre læringsmiljøet. For det tredje vil lærere som bruker mye dialog i undervisningen kunne øke elevenes leseferdigheter, men dette gir ikke utslag på deres regneferdigheter. Selv om forskerne sier at

resultatene må brukes med forsiktighet, er de ganske sikre på at det er *summen av ferdigheter* som inngår i læreres kompetanse som gir utslag på elevenes læringsutbytte. At det er den samlede effekten av lærerens kompetanse som gir størst utslag, viser også en metaanalyse gjennomført av Finn m. fl. (2009). Der analyseres funn fra 51 studier. Forskerne ser særlig på sammenhengene mellom læreres kompetanse og troverdighet (credibility), hva lærere gjør når de underviser, og elevenes læringsutbytte. I undersøkelsen måles tre dimensjoner av troverdighet (kompetanse, pålitelighet og omsorg), og forskerne kommer frem til at det er en moderat sammenheng mellom læreres kompetanse, troverdighet og elevenes generelle læringsutbytte. En moderat sammenheng ble funnet for alle de tre dimensjonene hver for seg, selv om den samlede effekten var større. Grossman m. fl. (2013) finner også, gjennom en stort anlagt undersøkelse, fordeler ved å kombinere data fra elevenes læringsutbytte med observasjoner av læreres undervisning. De argumenterer for at man på denne måten får frem informasjon som har potensial til å vise hva som faktisk kan gjøres for å forbedre undervisningen.

Koble data om elevenes læringsutbytte til læreren

Gjennom de siste utdanningsreformene No Child Left Behind (NCLB) fra 2001 og Race to the Top (RTTT) fra 2009 er skolene i USA pålagt økte krav til kvaliteten på data, noe som har ført til mer standardisering i amerikanske skoler. I 2001 var ambisjonen at man skulle kunne måle elevenes ferdighetsnivå på slutten av et skoleår (uavhengig av hvor de var da de begynte), og det ble bestemt at innen 2014 skulle alle elever ha et læringsutbytte som var tilfredsstillende (proficient). Offentligheten skulle føle seg trygg på at det ble lagt press på lærere og skoler slik at alle amerikanske barn ble sikret en tilfredsstillende utdanning. Etter hvert har det imidlertid blitt klart at det ikke er umiddelbart enkelt å definere hva som er tilfredsstillende. Ikke bare er «tilfredsstillende» en uklar betegnelse. Uansett hvordan det defineres, strever noen elever mer for å nå dette målet enn andre[26].

Tidlig på 1990-tallet prøvde lokale skolemyndigheter i statene Tennessee og Texas ut såkalte vekstmodeller. Man begynte med å måle skoleeffekter og ble senere opptatt av lærereffekter. Målet var å finne ut hva som lå utenfor og innenfor skolens kontroll (hva skolen kunne/ikke kunne gjøre noe med). For å identifisere effekten av lærerens og tidligere læreres innsats ble det brukt longitudinelle prestasjonsmål, og elevenes faktiske læringsutbytte ble målt mot deres forventede læringsutbytte. Vekstmodeller skal kunne avgjøre i hvilken grad en faktor (for eksempel skolen eller læreren) har påvirket elevens læringsutbytte. De som vurderer håper å finne ut om elever som blir «utsatt» for en bestemt lærer (eller skole) presterer over eller under forventet nivå. Schafer m. fl. (2012) har sammenlignet seks mye brukte psykometriske effektmodeller. Forskerne finner at det hefter mange måleproblemer ved metodene, og at det er særlig grunn til å se kritisk på avgjørelser som fattes basert på slike modeller når de får store konsekvenser (high stakes) for elevene og for enkeltlæreres ansettelsesforhold.

Flere stater og distrikter i USA har tatt i bruk Student Growth Modeling (SMG) og Value Added Modeling (VAM). Mariano m. fl. (2010) mener det er problematisk at de multivariate modellene som benyttes går ut fra at hver enkelt lærer har en (identifiserbar) effekt på elevenes læringsutbytte og at denne består (og forblir uforandret) i all fremtidig testing. Modellene legger dessuten til grunn at sammenhengene vil vedvare selv om det er enighet om at «lærereffekten» kan avta over tid. Når vurderingssystemer som baserer seg på slike modeller ikke bruker en vertikal skala eller ser på utviklingen på tvers etter hvert som elevene flytter seg oppover i klassene, er disse antakelsene om stabilitet ikke nødvendigvis konsistente med de målene man har på læringsutbytte. De fleste lærere arbeider for eksempel ikke alene – de har andre lærere og støttepersonale

26 https://www.aft.org/pdfs/teachers/whatsproficient0504.pdf

som bibliotekarer, helsepersonale, frivillige foreldre, assistenter og de arbeider i tolærersystem. Alt dette er faktorer som igjen påvirker lærefaktoren. Det kan altså være grunn til å spørre hvilken effekt man faktisk måler. Også Newton m. fl. (2010) påpeker at faktorer som forstyrrer «lærereffekten» vanskeliggjør modelleringen. Det er et problem at slike faktorer ikke bare er effekter som lett kan kontrolleres ved å måle elevenes prestasjonsnivå ved begynnelsen av skoleåret. Faktorene vil også interagere med lærerens evne til å prestere optimalt hele skoleåret og de interagerer i tillegg med ulike elevfaktorer.

3.2.1 Kort historisk bakgrunn for «Value-Added»

Her presenteres forskning om «Value-Added», en tradisjon som først og fremst har fått fotfeste i USA og England. På norsk vil Value-Added kunne oversettes som *tilleggs-* eller *merverdi*. Det er en betegnelse som varehandelen bruker for å angi den ekstra verdien (eller nytten) en selger kan tilføre et produkt før produktet er hos kunden. I utdanningssammenheng er Value-Added den tilleggsverdi skolen – eller læreren – bidrar med til elevens læringsutbytte i løpet av ett år.

For i det hele tatt å forstå hva forskerne diskuterer i de artiklene som tar opp problemstillinger rundt Value-Added, er det nødvendig med et historisk tilbakeblikk. Røttene til Value-Added-tradisjonen går ca. femti år tilbake i tid og har ulike kilder. Et viktig bidrag kommer fra sosiologen James Coleman, som, sammen med kollegaer, publiserte rapporten *Equality of Educational Opportunity* i 1966. Rapporten påviste at familiebakgrunn betyr mer for den enkelte elevs læringsutbytte enn det som kan måles av skolefaktorer, og at det finnes kjennetegn ved enkelte lærere som forklarte mer av variasjonen i elevenes læringsutbytte enn noen annen skolefaktor. En noe fortegnet versjon av rapportens funn ble at skolen ikke har noen betydning for elevenes læring (noe forskerne ikke konkluderte med). Mange interessante og viktige funn i rapporten «druknet» dessuten i den ene påvisningen av at sosio-økonomisk status (SØS), altså det elevene har med seg til skolen (og rundt seg gjennom hele skolegangen) har større betydning for det de kommer ut av skolen med enn det skolen kan bidra med – altså en Matteuseffekt[27]. Dette var problematisk for alle som hadde et sterkt håp om at utdanning kan redusere forskjeller mellom elevene og bidra til sosial utjevning.

I 1971 publiserte økonomen Eric Hanushek en artikkel i American Economic Review[28] hvor han påviste store forskjeller i elevenes læringsutbytte avhengig av hvilken lærer de hadde. Den mest siterte artikkelen av Hanushek i forbindelse med Value-Added er fra 1992[29], hvor han påviser at elevene som hadde de dårligste lærerne kunne ligge et helt skoleår bak sine jevnaldrende i målbart læringsutbytte. Hanushek er også kilden til urokkelige påstander om at klassestørrelse ikke har noen betydning for elevenes læringsutbytte, selv om hans egne konklusjoner refererer til statistiske effekter og kanskje ikke er like kategoriske[30]. I den samme tradisjonen som Eric Hanushek arbeider William Sanders, en statistiker som utviklet The Tennessee-Value-Added Assessment System (TVAAS). Siden 1993 har dette systemet utviklet metoder for å måle effekten av lærerens undervisning på elevenes læringsutbytte ved å følge elevenes utvikling fra et år til neste (og gjennom hele skoletiden) avhengig av hvilken lærer de har. Sanders og Horn (1998) og Sanders og Rivers (1996)[31] argumenterer for at den viktigste faktoren som påvirker elevenes læringsutbytte, er læreren, og at effekten av lærerne på elevene læringsutbytte både er additivt og kumulativt.

27 For en interessant diskusjon om rapporten 40 år etter, se Gamoran og Long (2006) ***http://www.wcer.wisc.edu/publications/workingPapers/Working_Paper_No_2006_09.pdf***
28 Hanushek, E. (1971): Teacher characteristics and gains in student achievement; estimation using micro data, ***American Economic Review***, 61, 280-288
29 Hanushek, E. (1992): The trade-off between child quantity and quality, Journal of Political Economy, 100, 84-117
30 Hanushek, E. (1999): The Evidence on Class Size, in Mayer, S. E. and Peterson, P. E. (Eds.): ***Earning and learning: How schools matter***, Washington D.C., Brookings Institution, 131-168
31 Sanders, W. L. and Rivers, J. C. (1996): Cumulative and Residual Effects of Teachers on Future Student Academic Achievement. http://www.cgp.upenn.edu/pdf/Sanders_Rivers-TVASS_teacher%20effects.pdf

Mens James Coleman og kollegaene på 1960-tallet var interessert i hvilke kjennetegn ved lærerne det kunne være som hadde positiv (eller negativ) innvirkning på elevenes læringsutbytte, er Hanushek og Sanders i 1990-årene mer opptatt av å måle ressursinnsats mot utbytte og undersøke kostnad-nytte-spørsmål. I flere arbeider tidlig på 2000-tallet argumenterer Linda Darling-Hammond for at god lærerkvalitet utvikles gjennom god lærerutdanning og lærerkvalifisering som gir det beste grunnlag for å bedre elevenes læringsutbytte[32]. I motsetning til dette, definerer Rivkin, Hanushek og Kain omtrent samtidig lærerkvalitet som identisk med elevenes målbare læringsutbytte[33]. Det skjer altså en bevegelse fra en input-tenkning om utdanning til en output-tenkning.

Selv om forskningen har ulik innretning, forskjellige forskningsspørsmål, og bygger på forskjellige forutsetninger, konkluderer alle med at kvaliteten på læreren, eller hva læreren gjør, er det som har størst betydning for elevenes læringsutbytte. Uenigheten går på om det skal satses på lærerutdanning og læreren (input) eller om lærerressursene bør utnyttes bedre (output). Når slike forskningsfunn kobles mot en utdanningspolitikk som er opptatt av målinger, standarder, prediktorer og indikatorer, blir politikerne svært opptatt av å få svar på hva det er som kjennetegner lærerkvalitet, god undervisning og godt lærerarbeid. De som kan levere raske svar om effekt, får stor oppmerksomhet. De siste 10-15 årenes forskning på lærervurdering reflekterer dette. I neste avsnitt presenteres forskning som har sett på hvordan ulike modeller for Value-Added i USA har vært koblet til lærervurdering.

3.2.2 Value-Added (VAM)

Value-Added modeller (VAM) varierer i forhold til hvor mange bakgrunnsvariabler de kontrollerer for og hvordan de analyserer vekst/endringer i elevenes testprestasjoner. Felles for alle modeller er at de undersøker de målbare effektene skolen eller lærerne har på elevenes læringsutbytte. «Effekt» er i denne sammenheng knyttet til forskjeller i elevenes testprestasjoner fra ett år til et annet. Ved å bruke bestemte korrelasjons- og regresjonsteknikker i analyse av forandringene er det mulig å beregne påvirkninger fra elevene selv, læreren i den enkelte klasse, eller skolen som helhet. I denne typen studier er man opptatt av å måle lærerproduktivitet og lærereffektivitet.

VA-modellene er tatt i bruk i mange amerikanske stater og brukes både til å vurdere kvaliteten på lærernes arbeid og innsats, bestemme lønnsnivå for lærere og sette standarder for akseptabel kvalitet på undervisningen. Lærere som ikke oppnår en minstestandard gjennom VA-målinger, kan stå i fare for å miste jobben, få redusert lønn eller bli forflyttet (jf. Bastian m.fl. 2013; Berliner 2014; Goldhaber m. fl. 2013a og 2013b; Winters og Cowen 2013). VA-målinger er gjort mulig gjennom en omfattende oppbygning av databaser for ressursbruk i skolen og akkumulering av elev- og lærerdata over flere år. De er longitudinelle og blir ofte utført som sekundæranalyser av store datasett. Følgende to studier viser størrelsen og kompleksiteten i slike undersøkelser:

Buddin og Zamaro (2009) har sett på data fra over 300 000 elever og 16 000 lærere på 2-5 klassetrinn i barneskolen i Los Angeles skoleårene 2000-2006. Målet var å undersøke sammenhengen mellom lærerkvalitet og elevprestasjoner i lesing og matematikk (California Achievement tests) på tvers av klasserom og skoler, lærernes utdanning og erfaring, samt lærernes eksamensresultater ved den pålagte yrkessertifiseringen i California. Analysene baserte seg på en rekke bakgrunnsvariabler for elever og lærere som var tilgjengelig i databasene. VA-beregningene ble kontrollert for viktige bakgrunns- og effektvariabler. Resultatene viste at det var store forskjeller i lærerkvalitet (målt med den aktuelle VA-modellen) på tvers av

32 For eksempel Darling-Hammond, L. (2000): Teacher Quality and Student Achievement, Educational Policy Analysis Archives, Vol. 8, No. 1 Open Access: *http://epaa.asu.edu/ojs/article/view/392/515*
33 Rivkin, Hanushek and Kain (2005): *http://www.econ.ucsb.edu/~jon/Econ230C/HanushekRivkin.pdf* (lastet ned 01.03.14)

skoledistrikter, men at de lærerkarakteristikkene som ble målt forklarte lite av forskjellene. Staten Californias modell for lærersertifisering slo ikke ut på prestasjonsforskjellene i de aktuelle klassene. Elevenes prestasjoner var videre uavhengig av hvorvidt lærerne hadde høyere grads utdanning eller ikke. Det viste seg at elevprestasjonene økte noe med lærererfaring, men sammenhengen var svak og avspeilet i stor grad problemene som nye lærere opplever i de to første årene av sin yrkeskarriere (jf. også Muñoz m.fl. 2013).

I et annet arbeid (Sass m. fl. 2012) var målet å anslå effekten av læreres innsats i skoler som hovedsakelig hadde elever fra familier med lav inntekt, sammenlignet med skoler der elevene tilhørte familier med høyere inntekt. Dataene omfattet 9000 lærere i Florida og 8000 lærere i North Carolina, og rundt 50 000 elever i hver av delstatene på 3-5 klassetrinn i barneskolen fra årene 2001-2005. Det ble benyttet data fra standardiserte tester i lesning og matematikk (literacy in reading and mathematics). Dessuten ble det hentet data fra store databaser i de to delstatene og etablert en rekke bakgrunnsvariabler. En Value-Added modell med data om elevers testresultater, elev- og familiekarakteristikker (f eks etnisitet eller sosioøkonomiske forhold), samt informasjon man ellers hadde om undervisning og lærere ble tatt i bruk. Åtte ulike VA-beregningsmodeller ble testet ut for å minimere feilmålinger. Som et generelt funn viser studien lavere skåre for lærereffekt i skoler hvor elevene i stor grad kommer fra lavinntekts familier enn i skoler hvor elevene kom fra familier med høyere inntekt. Variasjonen i lærereffekt er imidlertid større i skoler med lave inntektsgrupper enn i andre skoler, noe som skyldes en forholdsvis større andel lærere med lavere effektmål i disse skolene. Sass m. fl. (2012) konkluderer blant annet med at effektforskjellene mellom lærere og skoler har sammensatte årsaker, og at blant annet rekruttering av lærere spiller en rolle.

VA-modeller for å måle lærereffekt kan settes opp på flere måter alt etter hvilken informasjon som er tilgjengelig for elever, skoler og lærere. Sass m. fl. (2014) bygger på korrelasjonsberegninger mellom aktuelle variabler og avanserte regresjonsanalyser for å redusere feilmålinger og kontrollere for indre sammenhenger mellom variablene. Ved en kritisk gjennomgang av eksisterende forskning viser Schochet og Chiang (2013) at det hefter betydelig usikkerhet ved beregning av lærereffektivitet ved VA-modeller. De registrerer konsistente funn på at gjennomsnittlige økninger i testverdier både på lærer- og skolenivå er ustabile over tid, og at det bare er moderate korrelasjoner mellom VA-effektmål (0.2-0.6) for enkeltlærere fra år til år. En konsekvens av dette er at en finner betydelige forandringer fra år til år i hvordan lærere blir rangert. Dette er vurderinger som også støttes av Berliner (2013, 2014), Koedel (2009), Kinsler (2012) og Newton m. fl. (2010). Broatch og Lohr (2012) hevder at Value-Added-mål bare indirekte kan måle lærens bidrag til elevenes læringsutbytte og ikke fanger opp lærernes langsiktige bidrag til elever i «den virkelige verden». Etter å ha gått gjennom forskning på området viser Berliner (2014) at det er svært krevende å kontrollere for alle ytre variabler som påvirker testresultater og effektmål. Gjennom avanserte analyser av data over tre år viser Schochet og Chiang (2013) at det i ett av fire tilfeller kan skje at lærere blir feilrangert ved bruk av vanlige VA-modeller og longitudinelle data. Dersom en ikke bruker datasett, men skolen som enhet for effektmålinger, kan feilraten senkes med fem til ti prosent. De mener dette kan tyde på at modellene ikke er konsistente og fortsatt befinner seg på utprøvingsstadiet.

På tross av slike svakheter som er påpekt over mener Corcoran og Goldhaber (2013) at det er lite uenighet mellom forskerne i denne tradisjonen om VA-modellenes evne til å estimere egenskaper ved læreres effektivitet. Interessen for dem må ses i forhold til at lærervurderinger tradisjonelt ikke har koblet læreres arbeid og elevenes prestasjoner. I tillegg til metodediskusjoner er den store utfordringen hvordan VA-målingene brukes politisk og administrativt til å fremme elevenes læring og utvikling i skolen, f eks til å si opp lærere som ikke klarer å bedre elevenes læringsutbytte (Goldhaber og Theobald 2013), bruke lønnsinsentiver overfor dyktige lærere (Yuan m.fl. 2013), endre skolestrukturer eller stimulere til lokalt utviklingsarbeid i kombinasjon med andre tiltak (Wang m. fl. 2013).

3.2.3 Prestasjonslønn

Ofte kobles Value-Added-modeller til lønnsinsentiver for lærere. Flere forskere har sett på om lønnsinsentiver kan bidra til økt lærereffektivitet, altså om lærere klarer å bedre elevenes målbare læringsutbytte fra ett år til neste dersom de får lønnsøkning. Gjennomgående ser det ut som om forskningen finner liten eller ingen effekt av lønnsinsentiver for lærere på elevenes læringsutbytte.

Yuan m.fl. (2013) har brukt randomiserte eksperimenter til å studere tre forskjellige program for prestasjonslønn. Spørreskjema ble brukt for å undersøke i hvilken grad programmene motiverte lærere til å bedre elevenes resultater og hvordan det enkelte program påvirket lærernes undervisning, antall timer de arbeider, stress på jobben og opplevelse av kollegialitet. Det viser seg at flertallet av lærerne ikke fant programmene motiverende. Resultatene viser dessuten at ingen av de tre programmene fikk lærerne til å jobbe flere timer eller til å forandre undervisningen sin. Programmene innvirket heller ikke på lærernes opplevelse av stress på jobben eller deres opplevelse av kollegialitet. Artikkelen konkluderer med at man, i stedet for å koble lønn til elevenes læringsutbytte, kan teste ut belønningsmodeller som honorerer lærere når de tar på seg ekstra ansvar eller gjør en ekstra arbeidsinnsats.

Leigh (2013)[34] har analysert tre forskjellige datasett som ser spørsmålet om prestasjonslønn fra ulike vinkler: a) Studier som har målt effekten av modeller for prestasjonslønn for lærere, b) Studier som har frembragt kunnskap om læreres holdning til prestasjonslønn og c) Spørreskjema som viser hvilke holdninger offentligheten har til prestasjonslønn. I analysen kommer det frem at systemer for lærervurdering basert på prestasjonslønn ofte er svært kompliserte, og at de fleste bare har kortsiktig effekt. Lærerne har blandede følelser for prestasjonslønn – yngre lærere er mer positivt innstilt enn de mer erfarne. Særlig negativt innstilt er lærerne til systemer for prestasjonslønn som kobler læreres prestasjoner til elevenes testresultater. Fryer (2013) finner ingen effekt av prestasjonslønn på elevenes målbare læringsutbytte eller på lærernes motivasjon, arbeidsinnsats og undervisningsform. Heller ikke Goodman og Turner (2013) kunne dokumentere effekt av et system for prestasjonslønn i New York. I en analyse av data fra USA i tidsrommet 2003-2007 (School and Staffing Survey), finner Jones (2013) at menn var mer positive til prestasjonslønn enn kvinner, og at prestasjonslønn fikk lærere til å jobbe mindre og engasjere seg mindre i skolens fellesaktiviteter.

I en annen studie blir det imidlertid funnet en liten effekt av prestasjonslønn. Alfaro m.fl. (2013) fulgte over tre år et prosjekt i tiltakspakken The Texas Educator Excellence Grant (TEEG). Delstaten Texas satte av inntil 320 millioner dollar årlig til skoler i de fattigste områdene som klarte å øke elevenes læringsutbytte. Utvalgsskolene hadde i utgangspunktet (relativt sett) gode elevresultater på de statlige prøvene. Forskerne undersøkte elevenes prøveresultater i 4. klasse og så på sammenhengen mellom bevilgningene og elevresultatene gjennom tre årssykluser. Studien viser statistisk signifikante forskjeller på prosentandelen elever som bestod i matematikk, og lesing, men ikke i skriving, og konkluderer med at prestasjonslønn under gitte betingelser kan virke positivt.

Det har også vært interesse for å undersøke om prestasjonslønn til grupper av lærere kan ha effekt. Gjennom en randomisert kontrollert studie har Springer m. fl. (2012) analysert et program som baserte seg på bonuslønn til lærerteam på ungdomstrinnet. Teamet ble belønnet for sin kollektive innsats for å bedre elevenes læringsutbytte. Effekten av bonusprogrammet ble målt etter ett skoleår og målt igjen ett år senere. Det ble ikke funnet noen signifikant effekt på elevenes læringsutbytte eller på lærernes holdninger/praksis. Mangelen på effekt av teamlønn for prestasjoner bekrefter funn fra andre undersøkelser som har sett på effekt av bonuslønn for enkeltindivider eller hele skolens innsats for elevenes læringsutbytte.

34 Leigh, Andrew (2013): The economics and politics of teacher merit pay, *CESifo economic studies* 59(1), 1-33

3.2.4 Kritikk av Value-Added

Mange av de inkluderte artiklene diskuterer og kritiserer måten tall fra Value-Added brukes på. Det finnes også en del forskning som ser på spesielle sider ved VAM-modellene (for eksempel hvordan man skårer resultater og utformer prosedyrer for målinger). Hill m.fl. (2011) undersøker samsvar mellom Value-Added-skårer og andre indikatorer på undervisningskvalitet. Fra et stort datasett som omfattet 222 informanter, ble 24 matematikklærere valgt ut til undersøkelsen, som tok sikte på å sammenligne disse lærernes Value-Added skårer med indikatorer på lærerkvalitet fra spørreundersøkelser, observasjonsmateriale, undervisnings- og elevkarakteristikker. Analysen finner at lærers Value-added skårer korrelerer, ikke bare med deres matematikkunnskaper og undervisningskvalitet, men også med de elevgruppene som de underviser. Casestudien illustrerer problemer som kan oppstå når man bruker tall fra Value-added til å planlegge ordninger for prestasjonslønn.

Jesse M. Rothstein (2010) er blant de som kritiserer selve grunnlaget for VAM. Han mener at metoden kommer skjevt ut, fordi den har som grunnleggende premiss at elevene blir tilfeldig plassert i skoler og klasser (slik at lærerne underviser et tilfeldig utvalg elever). I den virkelige verden er det imidlertid aldri slik at elever blir tilfeldig plassert verken på skoler eller i klasser. Han mener at dette grunnleggende premisset i Value-Added modellen kan føre til skjevfordelinger i estimatene av den såkalte «lærereffekten», og påpeker dessuten at måten modellene faktisk blir praktisert på er en årsak til alvorlige feilkilder. Data fra North Carolina viser for eksempel at man ikke hadde overholdt noen av eksklusjonskriteriene i VAM. Det fikk som resultat at modellene viste «effekter» av at lærere på 5. trinn hadde økt elevenes testskårer på 4. trinn.

I en studie av årsvariasjonen i Value-Added-estimater for matematikklærere i grunnskolen i fem skoledistrikt i Florida, har McCaffrey m. fl. (2009) kritisert premissene og grunnlaget for modellen. De finner at en stor del av variasjonen i målt lærerprestasjon (ca. 30-60 %) kan skyldes stikkprøvefeil forårsaket av «støy», i elevenes testskårer. Vedvarende lærereffekt som ikke skyldes «støy» kan tilskrives rundt 50 % av variasjonen på barnetrinnet og 70 % på ungdomstrinnet. Resten av variasjonen skyldes faktorer som varierer over tid på lærernivå, men lite av dette kan forklares av observerte karakteristikker ved lærerne.

3.2.5 Prediktorer

Noen av de inkluderte artiklene handler om å kunne forutsi (predikere) hvem som kan bli dyktige lærere i fremtiden. Ved å samle data om lærerne fra rekruttering, via utdanningen og ut i yrkeslivet, vil det være mulig å tegne profiler hvor noen trekk ved lærerne slår sterkere ut enn andre. Forskerne har for eksempel vært interessert i å finne ut om det er sammenheng mellom læreres eksamensresultater og elevenes skoleprestasjoner. Hill m. fl. (2012c) har sammenlignet observasjonsskårer som målte kvaliteten på lærernes matematikkundervisning med hvordan de samme lærernes presterte på en matematikkprøve (multiple choice), og elevenes value-added-skårer fra en nasjonal test. Funnene viste at dårlige resultater på en skriftlig prøve kunne forutsi dårlige prestasjoner i klasserommet. Gode resultater på en skriftlig prøve predikerte gode prestasjoner i klasserommet. Når det gjaldt lærernes prestasjoner blant de som skåret middels, var det imidlertid betydelige variasjoner i fordelingen av skårene.

Darling-Hammond m. fl. (2013) undersøker hvor godt PACT (the Performance Assessment for California Teachers) kan predikere hvem som kommer til å bli gode lærere. PACT er et verktøy utviklet for å måle nyutdannede læreres evne til å planlegge, undervise, vurdere og reflektere rundt undervisning i klasseromssituasjoner. Studien undersøker hvor godt PACT faktisk kan måle og forutsi (predikere) lærerdyktighet. Resultatene viser at skårene fra PACT–testen er signifikante predikatorer når det gjelder å

forutsi hvem som blir gode engelsk- og matematikklærere. I tillegg rapporterer et stort flertall av PACT kandidatene at de ved å fullføre selve vurderingen tilegnet seg kunnskaper og bedret sine undervisningsferdigheter. Utslaget var imidlertid størst når de også følte at de fikk støtte fra programmet «learning to teach» og hjelp til å gjennomføre vurderingsprosessen.

For å undersøke i hvilken grad læreres testskårer og resultater fra lærerutdanningen predikerer senere undervisningskompetanse, har D`Agostino og Powers (2009) gjennomført en metaanalyse. Hovedfunnet fra metaanalysen er at lærerstudenters eksamensresultater predikerer undervisningsferdigheter signifikant bedre enn lærernes testskårer som kun var moderat relatert til undervisningskompetanse.

3.3 Forskning med vekt på formativ vurdering

I dette kapitlet presenteres studier hvor forskerne først og fremst har ønsket å få svar på hvordan kvaliteten på ulike utdanningspraksiser kan gjøres bedre. De fleste artiklene har også sett på forhold mellom resultat og prosess, altså hvordan resultater fra summativ vurdering kan brukes formativt. Artiklene bidrar til å kaste lys over de formative sidene ved vurdering gjennom undersøkelser av ulike spørsmål rundt tema egenvurdering, altså lærere som vurderer sin egen praksis, kollegavurdering og elever som vurderer lærere. Først presenteres noen studier som har sett mer generelt på formative prosesser for vurdering.

3.3.1 Generelt om prosesser for vurdering

For å finne ut hva lærere kan og vet om vurdering, har Howley m. fl. (2013) dybdeintervjuet 26 lærere med svært variert fagbakgrunn på tre videregående skoler som hadde fra 400 til 800 elever. Målet var å få dybdekunnskap om lærernes vurderingskunnskap, hva som kjennetegner vurderingskulturen på skolene og hvordan lærere betraktet vurderingskompetansen til andre grupper i eller rundt skolen. Dataanalysen viste at det på tvers av skolene er mulig å identifisere en felles kunnskapsbase om vurdering og vurderingspraksis. Lærerne bruker formativ vurdering til å gi tilbakemeldinger til elevene og til å justere sin undervisning. De tror at godt samarbeid utgjør grunnlaget for å forbedre vurderingspraksiser og mener at mange elever, foreldre, skoleledere og administrasjon ofte har et ganske naivt syn på hvor mye viktigere og bedre summativ vurdering er enn formativ. Howley m. fl. (2013) konkluderer med at lærere både har svært mye erfaring med og kunnskap om vurdering og at de gjerne deler den praktiske kunnskapen de har på dette området med andre. De vet hvordan de skal koble vurdering til ulike mål for utdanningen, hva som kjennetegner ulike former for vurdering og hva som forener og skiller summativ og formativ vurdering. I tillegg har lærerne stor respekt for hvordan vurdering skal brukes på en fornuftig måte til å støtte elevenes læring og utvikling – både på kort og lang sikt. Den viktigste konklusjonen hos Howley m. fl. (2013) er derfor at det finnes mye kompetanse i skolen som det er viktig å bygge på når man skal utforme systemer for vurdering.

I en studie fra Vietnam har Pham og Stacey (2012) undersøkt prosedyrer for undervisningsvurdering i videregående skole. Vurderingen tok sikte på å avgjøre hvordan veiledning kan bidra til lærernes profesjonsutvikling. Det ble samlet intervjudata fra 34 deltakere: 24 lærere og 10 evaluatorer, både fra skoler i byer og på landsbygda. Data viste a) at lærerne betraktet «vurderingskonferanser» og analyser etter konferansene som nyttige, b) det var stort engasjement i diskusjonene og vilje til å dele ideer samt c) evne til å gi og ta imot tilbakemeldinger. Studien konkluderer med at det bør legges større vekt på diskusjoner om undervisningsaktiviteter før og etter observasjonene i stedet for å innføre «byråkratiske» og standardiserte prosedyrer.

Noen forskere påpeker at det er overraskende lite forskning på de følelsesmessige sidene ved læreryrket. Mange lærere opplever – i alle fall i perioder – stressende arbeidsdager, og de bruker mye energi ikke bare på å kontrollere sine egne følelser, men også elevenes. I en studie som har undersøkt forholdet mellom følelsesregulering, jobbtilfredshet og utbrenthet blant 123 lærere i engelsk videregående skole, finner Brackett m. fl. (2010) en positiv sammenheng mellom lederstøtte og lærernes positive følelser og evne til følelsesregulering. Engelske lærere har i en årrekke jobbet i et system basert på accountability, hvor elevenes læringsutbytte, slik det måles gjennom tester, brukes til å rangere skoler og lærere. Taber m. fl. (2011) har intervjuet 17 lærerstudenter som sier at de blir litt forvirret av det de opplever når de er ute i praksis. Det er uklart for dem hva som «egentlig» er vurderingens formål. På den ene siden blir de fortalt at forskningen er entydig på at vurdering for læring bør prege undervisningen, men konteksten dette budskapet formidles i er preget av storstilt testing og summativ praksis.

Det er også mange forskere som er opptatt av læres profesjonsutvikling. Noen undersøker undervisning i fag, og ser på hvordan det kan etableres bedre sammenheng mellom læreres fagkunnskaper og deres pedagogiske kunnskaper. Lee Shulmans begrep «pedagogical content knowledge» (PCK)[35], eller undervisningens «hvordan», kobles med undervisningens «hva», altså fagkunnskap (content knowledge, CK). Forskerne forsøker både å måle disse to kunnskapsformene, vise hvordan de henger sammen og gi anvisninger for hvordan undervisningen kan forbedres. Jüttner m. fl. (2013) mener at vi vet lite om forholdet mellom lærernes fagkunnskaper (CK), pedagogiske kunnskaper (PCK) og elevenes læringsutbytte og beskriver en metode som, i kombinasjon med klasseromsobservasjoner, kan brukes til å undersøke undervisningskvalitet og øke vår forståelse av hvilke aktiviteter som særlig kan støtte elevenes læring. Graves m. fl. (2009) påpeker at lærere trenger tilbakemeldinger fra flere kilder enn sporadisk elevvurdering. Hvis målet med vurderingen er å bedre lærernes undervisning, må den som vurderer se på andre forhold enn de vanlige (planlegging og gjennomføring av undervisningen, valg og bruk av læremidler og hvordan vurdering skjer gjennom tilbakemeldinger og vurdering av timen). Det som må komme i tillegg, er å vurdere hvorvidt læreren klarer å utvikle et positivt læringsmiljø, om læreren faktisk hjelper elevene til å forstå hvordan de lærer (metakognisjon) og om elevene får vite hvilke læringsressurser som er tilgjengelige for dem.

3.3.2 Elever som vurderer lærere

Flere forskere påpeker at det har vært liten interesse for å undersøke praksiser hvor elever vurderer lærere. Dette kan skyldes at slike praksiser er tilfeldige og lite utbredt, eller at de mest oppfattes som informasjon til læreren der og da og ikke har vært systematisert. I søkeprosessen dukket det opp mange artikler om elever som vurderer lærere, men de aller fleste var fra universiteter og høyskoler (på engelsk brukes student både om elever og studenter). Tre artikler tilfredsstilte kvalitetskriteriene[36], og gjengis under.

35 Shulman, L. S. (1987). Knowledge and teaching: Foundations of the new reform. *Harvard Educational Review*, 57(1), 1-22.
36 Artikkelen må ha en problemstilling som er klart formulert, den må gjøre rede for metodevalg og metodisk fremgangsmåte og det må være sammenheng mellom problemstilling, funn, drøfting og konklusjon

Etter å ha gjennomført en metaanalyse av forskningslitteraturen, påpeker Clayson (2009) at forskerne stiller spørsmål ved kvaliteten på tiltak som baserer seg på at elever vurderer læreres undervisning. Et stridsspørsmål har vært forholdet mellom vurdering og læring. Det er rimelig å gå ut fra at hensikten med å la elever vurdere lærerens undervisning er at den skal bli bedre. Hvis god undervisning fører til at elevenes læringsutbytte blir bedre, må det være en relasjon mellom det som vurderes og læringen. Litteraturgjennomgangen viser imidlertid at alle forsøk på å avdekke en slik relasjon vanskeliggjøres på grunn av forskjeller i praksis, metode og fortolkning. Clayson (2009) konkluderer med at det er svak sammenheng mellom vurdering og læring fordi det er snakk om situasjonsbestemte sammenhenger som vanskelig kan generaliseres til større populasjoner, flere fag, eller på tvers av klassenivå.

Aldridge m. fl. (2012) har undersøkt utvikling og implementering av et spørreskjema som ble brukt til å la elever på ni skoler vurdere læringsmiljøet. Data fra 2043 elever i 147 klasser på 9 skoler (11. og 12. klasse) ble brukt til å vurdere skjemaets reliabilitet og validitet. Intensjonen var at elevenes tilbakemeldinger skulle få lærerne til å reflektere over egen undervisningspraksis og gi retning for implementering av strategier som kunne bedre læringsmiljøet. I tillegg ble det utformet et aksjonsforskningsdesign som bestod av refleksjonsnotater, skriftlige tilbakemeldinger, diskusjonsfora og intervjuer med åtte lærere. Artikkelen ser særlig på hvordan lærerne tenkte rundt bruken av spørreskjemaet og analyserer i detalj hvordan en av lærerne brukte skjemaet som et redskap for refleksjon og praktisk hjelp til å gjøre forandringer i klassen. Casestudien viser hvordan aksjonsforskning basert på elevvurdering kan være til nytte i utviklingen av et bedre læringsmiljø.

Basert på data fra 225 lærere ved syv skoler og 538 lærerstudenter i Tyskland har Hinz (2011) undersøkt læreres og lærerstudenters erfaring med og holdninger til offentlig nettbasert vurdering av undervisningskvalitet. Det ble gjennomført fokusgruppeintervjuer før spørreskjemaet ble utformet. Svarprosenten var henholdsvis 87 og 99 prosent. Resultatene viser at lærerne var mer reserverte i forhold til denne typen undervisningsvurdering enn elevene, at eldre lærere var mer skeptiske enn yngre og at kvinner var mer skeptiske enn menn. Det konkluderes med at alder synes å være den faktoren som har størst betydning for holdning til nettbasert undervisningsvurdering.

3.3.3 Egenvurdering

I de landene hvor det er innført systemer for lærervurdering, er ofte egenvurdering en av komponentene i systemet. Det er derfor interessant å se på hvilke konklusjoner forskning trekker etter å ha undersøkt denne praksisen. Seks artikler som spesielt har sett på egenvurdering er inkludert i rapporten. Van Diggelen m. fl. (2013) har undersøkt hvordan lærere på yrkesfaglig studieretning vurderer sin egen kompetanse i å veilede elever. Lærerne som deltok i studien fulgte en prosedyre hvor de først måtte vurdere seg selv ved hjelp av bestemte kriterier. De fikk så tilbakemeldinger fra en kollega med utgangspunkt i de samme kriteriene, og skrev deretter en rapport hvor de reflekterte rundt sine egne læringsopplevelser. Egenvurderingsskjemaene, kollegavurderingene, video-opptak av veiledningssamtaler med kollegaer og 24 læreres refleksjonsnotater ble analysert. Forskerne kom frem til at lærernes prosedyre for egenvurdering kjennetegnes ved: 1) litt (for) positive egenvurderinger om prestasjonsnivå, 2) konstruktive tilbakemeldinger fra kollegaer som i stor grad ble akseptert av lærerne som ble vurdert og 3) klare og informative refleksjonsrapporter som viste at lærerne hovedsakelig var opptatt av at egenvurderingen fikk dem til å handle annerledes. Forskerne konkluderer med at lærere kan tjene på prosedyrer for egenvurdering og at prosessen også øke lærernes vurderingskompetanse og hjelpe dem til å bli flinkere til å gi tilbakemelding til elevene.

Digitale undervisningsmapper betraktes som nyttige for læreres profesjonsutvikling, men det mangler

empirisk belegg for å kunne slå fast hvor nyttige de er. Sung m. fl. (2009) har tatt i bruk digitale mapper som inneholdt flere vurderingsformer (f.eks. egenvurdering, kollegaveiledning, diskusjon og dagbokskriving), og fulgte 44 lærervikarer som fikk kurs i hvordan disse mappene skulle brukes. Ved hjelp av et rammeverk for lærerefleksjon, utviklet av Sparks-Langer, ble det avdekket at bare en tredjedel av lærerne kom på det høyeste refleksjonsnivået, og at flertallet av lærerne skåret gjennomsnittlig på sine refleksjonsnotater. Gjennom arbeidet med mappene bedret lærernes vurderingspraksis seg betydelig. Konklusjonen er at digitale mapper som består av ulike vurderingsprosedyrer gir god støtte til lærernes refleksjon og profesjonsutvikling. Også Tinoca og Oliveira (2013) har undersøkt om en nettbasert vurderingsstrategi kan støtte læreres profesjonsutvikling. Målet med studien var mer spesifikt å finne ut hvordan et bestemt nettbasert vurderingsdesign kan hjelpe de deltagende lærerne til å se nytten av formativ vurdering, og få dem til selv å ta i bruk formativ vurdering i egen undervisning. Vurderingsdesignet består av ulike ikke-standardiserte nettbaserte vurderingsmetoder. Data fra 494 spørreskjema ble analysert, sammen med deltagernes refleksjoner over sin egen læring. Studien konkluderer med at designet bør forandres slik at den formative intensjonen kan bli bedre ivaretatt.

Roby (2012) har latt 142 lærere i Ohio, som alle har mer enn tre år høyere utdanning, vurdere sine relasjonsevner. Studien baserer seg på en relasjonsundersøkelse (human relations survey), og resultatene ble sammenlignet med vurderinger gjort av deltakernes kollegaer. Resultatene ble analysert med sikte på å finne likheter og forskjeller mellom vurderingene. Undersøkelsene avdekket ingen statistisk signifikante forskjeller mellom lærernes egenvurderinger og kollegaenes oppfatninger.

Til tross for den utbredte bruken av videoopptak av undervisning, finnes det lite kunnskap om hvilken innvirkning ulike typer videopptak kan ha på lærernes motivasjon og deres kognitive og emosjonelle utvikling. Kleinknecht og Schneider (2013) fulgte ti matematikklærere på ungdomstrinnet mens de analyserte opptak av sin egen og andre læreres undervisning. Det viste seg at det var lettere å få lærere engasjert i å analysere problematiske hendelser når de så opptak av andre lærere. Det å se opptak av andre læreres undervisning ga høyere engasjement. Resultatene indikerer at det er mindre belastende å vurdere opptak av andre enn av seg selv, og at lærere som skal vurdere opptak av sin egen undervisning kan trenge støtte (scaffolding) og hjelp i analyseprosessen.

Friedrich m. fl. (2012) presenterer en undersøkelse som bygger på data samlet inn for programmet SINUS-Transfer, et prosjekt for profesjonsutvikling i Tyskland. I prosjektet samarbeider lærere om å forbedre sin undervisning gjennom samtaler om pedagogiske problemstillinger. Det ble prøvd ut en team-mappe som skulle støtte lærernes profesjonsutvikling. Ettersom metodens legitimitet avhenger av lærernes tillit til metoden, er hovedspørsmålet i denne artikkelen hvordan lærere vurderer team-mappen som redskap til profesjonsutvikling. Resultatene viser at det er store forskjeller blant lærerne med hensyn til om og hvordan de anerkjenner mappen som vurderingsverktøy. Grovt sett havner lærerne i to kategorier: de som er kritiske og en litt større gruppe som er positive. Når forskerne ser nærmere på de to gruppene, viser det seg at i gruppen med kritiske lærere er det både de som avviser hele prosjektet og noen som mener at team-mapper er bortkastet tid. De mener at det finnes bedre måter å drive profesjonsutvikling på. Blant de positive lærerne er det noen som aksepterer mål og metode for prosjektet og andre som gjennomfører vurderingen som et rent pliktløp. Konklusjonen er at det på overflaten ser ut som om noen lærere er «for» og andre er «mot», mens det faktisk er store nyanser innad i disse to kategoriene.

3.3.4 Kollegavurdering

Brantlinger m. fl. (2011) har fulgt en gruppe lærere i videregående skole som møttes 16 ganger for å diskutere videopptak som de hadde gjort av sin egen undervisning. Videoopptakene skulle sendes inn til National Board for Professional Teaching Standards, og hensikten med møtene var at lærerne skulle se på videoene sammen. Diskusjonene i møtene ble transkribert, og analyser av det transkriberte materialet viser at lærerne engasjerte seg i dyptpløyende samtaler om matematiske spørsmål. Lærerne snakket særlig om tre tema: teknikker man kan bruke i undervisningen for å snakke om matematiske spørsmål, kontekstuelle forhold som påvirker måten det snakkes om matematikk på i undervisningen og kriterier for å vurdere måten det snakkes på. Møtene førte til at det utviklet seg et profesjonsfellesskap mellom deltakerne etter hvert som de samarbeidet om å undersøke og vurdere hverandres praksis. Studien gir særlig innsikt i hvordan det å forberede seg til sertifisering kan styrke læreres profesjonslæring og utviklingen av profesjonelle lærerfellesskap. Man får et objekt som kan undersøkes i fellesskap.

Brown og Crumpler (2013) tar utgangspunkt i følgende paradoks: Selv om det er bred enighet blant forskere om at det viktigste vi kan gjøre for å heve kvaliteten i utdanning er å bedre undervisningen, er det ikke enighet om hvordan vi skal vurdere kvaliteten på lærernes arbeid. Selv om lærervurdering har blitt et sentralt spørsmål, er det heller ikke enighet om hva som utgjør en god vurderingsmetode. De argumenterer for at kollegavurdering må utgjøre kjernen i lærervurderingen hvis hensikten med vurdering er å forbedre kvaliteten på undervisningen. Lærere trenger mange former for tilbakemeldinger, også faglige. Å plassere kollegaveiledning sentralt i lærervurderingen vil kunne styrke lærerens profesjonslæring og bidra til å forbedre kvaliteten på undervisningen.

Det er bred enighet blant forskerne om at skoler som presterer bra tar kollektivt ansvar for elevenes læring, etablerer og utvikler sterke kollegiale relasjoner mellom lærerne. Få undersøkelser har imidlertid undersøkt om dette påvirker lærernes læring. I et kontrollgruppedesign som tester effekten av skriftlige kommentarer (praise notes) fra lærer til lærer, undersøker Nelson m. fl. (2013) samhold og kollegialitet i videregående skole. Deltakerne fylte ut et spørreskjema som tok sikte på å måle samarbeidsinteraksjoner i en gruppe lærere. Resultatene viser en statistisk signifikant sammenheng mellom testgruppe og kontrollgruppe, med moderate effektskårer. Etter at lærerne hadde fått de skriftlige kommentarene fra sine kollegaer, økte opplevelsen av kollegialitet og støtte betraktelig. Effekten var større blant yngre enn blant eldre lærere.

Verberg m. fl. (2013) har sett på en spesiell prosedyre for lærers profesjonslæring som kalles forhandlet vurdering. Forhandlinger mellom den som vurderer og den som blir vurdert antas å virke svært læringsfremmende. Hittil har antakelsen imidlertid ikke blitt undersøkt empirisk, så den mangler dokumentasjon. I denne studien deltok 27 yrkesfaglærere på en studieretning for helsefag (23 kvinner og 9 menn), 18 av dem ble vurdert, de 9 andre var vurderere. Begge gruppene fikk opplæring i teknikken. Deltakerne ble spurt om hvor nyttig de opplevde de forskjellige elementene i en forhandlet vurdering for sin profesjonslæring og hvilket læringsutbytte de mente det hadde å delta i undersøkelsen. På spørsmål om hvilke deler av prosessen de opplevde som mer og mindre nyttige, sa respondentene at de hadde fått et mer nyansert syn på sin egen kunnskap fordi de ble tvunget til å tenke gjennom hvordan de fungerte som lærere. Noen rapporterte at forhandlingene hadde utfordret grunnleggende antakelser om undervisning og læring. Dette hadde ført til at de forandret sin undervisningspraksis, og at elevene deres ble mer reflekterte.

De artiklene som har undersøkt formativ vurdering ser på veiledning, profesjonsutvikling og ulike sider ved vurderingsprosessen, samt de følelsesmessige implikasjonene av vurdering. Få studier har undersøkt ordninger hvor elever vurderer lærernes undervisning. Fordi de kan bidra til å øke vurderingskompetansen og styrke profesjonaliteten i skolen, betraktes dessuten kollegavurdering og egenvurdering som nyttige praksiser i systemer for lærevurdering.

3.4 Metodediskusjoner – spørsmål om validitet og reliabilitet

De systematiske søkene avdekket flere artikler som diskuterer validitet (gyldighet) og reliabilitet (pålitelighet) ved de ulike vurderingsmetodene som blir brukt i lærervurdering. Det er flere grunner til at det er viktig å diskutere om metodene er gyldige og pålitelige. Både må man forsikre seg om at når forskere påstår at de kan dokumentere effekt av variabler, så er tall og effektstørrelser korrekte, altså at målemetodene er pålitelige. Ulike målemetoder kan komme frem til forskjellige resultater, og det er en viktig del av forskningen å undersøke om man kan bekrefte eller avkrefte tidligere resultater. Jo flere studier som kan bekrefte et funn, jo sikrere kan man være på at resultatet er riktig. Like viktig kan det være å få avkreftet resultater. Det er i tillegg viktig at man måler det man faktisk har tenkt å måle, og ikke noe annet. Metodediskusjoner er nødvendige fordi de fører forskningen fremover og gir oss stadig bedre kunnskapsgrunnlag.

Metodediskusjon knyttet til modeller

Det elektroniske søket har fanget opp flere artikler som diskuterer Value-Added, men som først og fremst er metodediskusjoner. Noen bruker binomial logistisk regresjonsanalyse for å måle forholdet mellom læreres Value-Added skårer og elevresultater eller forholdet mellom læreres Value-Added skårer og lærernes tendens til å bli i yrket (Ingle 2009). Andre drøfter hvordan målefeil kan oppstå i longitudinelle design (Boyd m. fl. 2013; Karl m.fl. 2013), gir eksempler på målemetoder som kan gi bedre prediksjon (Lefgren og Sims, 2012), viser hvor vanskelig (og nødvendig) det er å kontrollere for målefeil i Value-Added estimater (Briggs og Weeks 2011; Lockwood og McCaffrey 2014) og anbefaler at man utviser forsiktighet ved bruk av Value-Added modeller (Koedel 2009; Ehlert m.fl. 2014). Everson m. fl. (2013) foreslår en alternativ fremgangsmåte som kalles «propensity score matching», som gjør det mulig å måle hvor godt en lærer presterer relativt til lærere som underviser sammenlignbare klasser. Argumentet er at en slik målemetode er bedre egnet til å måle hvor godt ansatte gjør det i den jobben de faktisk skal gjøre.

Metodediskusjon knyttet til observasjon som metode for å vurdere undervisning

En mye brukt metode for lærervurdering er observasjon av undervisning og aktiviteter i klasserom og grupper. Bell m.fl. (2012), Casabianca m.fl. (2013), Hill m.fl (2012a), Hill m.fl. (2012b), Hill og Grossman (2013), Ing og Shih (2013) har alle sett nærmere på observasjon som metode for å vurdere undervisning.

Både Hill m.fl. (2012a) og Bell m.fl. (2012) har undersøkt validitet ved observasjon som metode. Validitet sier noe om gyldigheten av data, altså om man har observert (og samlet informasjon) om det som faktisk skulle undersøkes. Hill m. fl. (2012a) argumenterer for at de som bruker ulike observasjonsverktøy og andre typer vurdering av undervisning eller klasseromspraksis må undersøke validiteten på de verktøyene som blir benyttet. Særlig er det viktig å se på effekten av å bruke forskjellige personer til å skåre observasjonsskjema, undersøke innholdet, se kritiske på prosedyren for observasjon og den lokale konteksten som observasjonene utføres i. Bell m.fl. (2012) har også sett nærmere på hvordan det er mulig å øke validiteten ved observasjoner. Dette er gjort ved å teste ut det som kalles «validity argument approach»[37], en argumentasjonsteknikk som kan brukes til å kritisk undersøke (og validere) kvaliteten på observasjonsprotokoller.

Bell m. fl. (2012) fant både styrker og svakheter ved metoden. En av styrkene er at forskere ved å bruke en

37 «Validity argument approach» består av to stadier 1) Det formative stadiet hvor det utvikles et fortolkende argument der antakelser og slutninger i tolkningen av teksten er eksplisitt forklart. 2) Det summative stadiet hvor det fortolkende argumentet blir vurdert og muligens reformulert i lys av empirisk evidens.

slik tilnærming i større grad må tydeliggjøre hvilke forestillinger de har om sammenhenger mellom begrep, instrument og slutninger, altså gjenstanden for vurderingene, målemetodene og hvilke konklusjoner man kan trekke. Metoden kan bidra til en mer etterrettelig bruk av observasjon fordi den forutsetter detaljerte beskrivelser av hvordan observatørene er samsnakket og hvordan undervisningstimer er valgt ut over et helt skoleår. Tilnærmingen skjerper også argumentasjonen og sikrer validitet ved at vurderingsmetoden blir klarere og tydeligere.

Hill m. fl (2012b) og Hill og Grossman (2013) ser nærmere på observasjon av undervisning og utfordringer knyttet til å utvikle gode observasjonssystemer. Hill m. fl (2012b) diskuterer den økende interessen for klasseromsobservasjon som et middel for å nå flere mål samtidig (både profesjonsutvikling, lærervurdering og vurdering av intervensjoner i klasserommet) og argumenterer for at det trengs bedre observasjonsmetoder om man skal lykkes med observasjoner. I tillegg til gode observasjonsskjema må det være mulig å få frem skårer som gjør det mulig å kalkulere pålitelige kost/nytte-vurderinger og prosesser for rekruttering, opplæring og sertifisering av de som skal vurdere. Hill og Grossman (2013) diskuterer observasjon som en del av et nytt system for lærervurdering. De argumenterer for at hvis et observasjonsverktøy skal innfri målet om å støtte lærerne i å forbedre egen undervisningspraksis, må skjemaer være fagspesifikke, eksperter på innholdet bør involveres i observasjonsprosessen, og skjemaet bør gi informasjon som er både presis og nyttig for lærerne. Forfatterne diskuterer utforming av skjemaet, hva som skjer rundt utfyllingen av observasjonsskjema, utvikling av rangeringsskala, designet av systemet samt timing av og formen på tilbakemeldinger fra observasjonene. Forfatterne påpeker at politikere møter noen utfordringer ved innføringen av et slikt vurderingssystem. Hvis det skal lykkes, må de motstå fristelsen til å forenkle undervisningens iboende kompleksitet. Dette innebærer å akseptere at undervisning handler om faginnhold og pedagogisk tilrettelegging, hvilke personer som blir valgt ut til å vurdere, hvordan de læres opp og hvilken ekspertstatus de får i arbeid med observasjon og profesjonsutvikling, samt hvordan ressurser allokeres til veiledere som kan bidra til å forbedre undervisningspraksis.

3.5 Betydningen av ledelse i systemer for lærervurdering

Forskning om lederes vurdering av læreres arbeid handler om vurderinger som får konsekvenser for personalbeslutninger og ansettelsespraksiser (Cohen-Vogel 2011, Master 2013), studier som ser på hvordan vurdering kan brukes til å utvikle enkeltlæreres undervisningspraksis og profesjonsutvikling (Grissom m.fl. 2013, Moreland 2009, Tuytens og Devos 2011) og studier som mer generelt omhandler skoleutvikling, der vurdering av lærere inngår som en komponent (Maslow og Kelly 2012, Nehring og O'Brien 2012) . Studiene varierer også med hensyn til hvor sentral skoleledelsen er i vurderingen, om en ser på skolelederes bruk av vurderingsdata som er definert og innhentet av aktører utenfor skolen eller om de kommer fra skolelederens egne vurderinger av de ansatte.

Skoleledere bruker vurderinger av læreres arbeid til å foreta personalprioriteringer, men det kan diskuteres om vurderingsformene som brukes får frem et fullstendig bilde av hvordan læreren fungerer i jobben sin. For å undersøke om skoleledere bruker informasjon fra testresultater i sin rekrutteringspolitikk har Cohen-Vogel (2011) analysert funn på tvers av flere casestudier. Artikkelen diskuterer implikasjoner av «evidensbasert ansettelse». Det høye accountability-presset fra føderale og statlige myndigheter i USA får skoleledere til å eksperimentere med både læreplaner, skolemat og tidsutnyttelse når målet er å øke elevenes læringsutbytte. Ved å undersøke skoleledernes ansettelsespraksis i en skole med gode elevresultater og en skole med lavere elevresultater i hvert av de fem skoledistriktene i Florida, ble det avdekket at skoleledere bruker elevenes

testresultater i personalprioriteringer (staffing to the test). De ansetter, omplasserer og gir lærere etter- og videreutdanning strategisk for å øke skolenes gjennomsnittlige resultater. Når de kan sette slike strategiske overlegninger til side, gir imidlertid skolelederne mer nyanserte vurderinger av kvaliteter ved lærernes arbeid.

En studie av Master (2013) går videre inn i denne problematikken ved å ta utgangspunkt i at vi vet forholdvis lite om hva skoleledere legger vekt på når de bruker informasjon fra ulike typer vurderinger til å ta avgjørelser som får konsekvenser for personalet. Med bakgrunn i kvantitative data fra charterskoler (som er offentlig finansierte, og privat administrert gjennom treårige kontrakter og som har større autonomi enn statlige skoler), undersøker Master i hvilken grad skolelederes formative vurderinger av lærere midtveis i året samstemmer med andre data som elevresultat og surveydata der foreldre og lærere har uttalt seg om de samme lærerne. Funnene viser at skoleledernes formative vurderinger samstemmer med andre data og har konsekvenser for beslutninger om oppsigelser og forfremmelser på et senere tidspunkt. I analysen av data viser det seg imidlertid at formative rangeringer som blir gjort midt i året, og som er gjenstand for diskusjon mellom skoleleder og lærer, er mer preget av varierte vurderingskriterier og en mer helhetlig vurdering av lærerne. Enkelte kriterier for god undervisningspraksis ble verdsatt høyt i alle typer vurderinger, for eksempel klasseledelse, mens andre kriterier, for eksempel vurderinger av forholdet til elever og foreldre, ble vurdert ulikt. Grunnen til dette er at surveydata og gjennomsnittskårer fra elevresultat ikke kan si noe om hvordan læreren fungerer i et kollegium. Lederen sitter på mer informasjon og kan faktisk vurdere hvilken rolle læreren har på skolen. En konklusjon er derfor at evalueringssystemer som bare fokuserer på enkeltlærere overser at lærere er avhengige av hverandre og spiller ulike roller i et kollegium. Studien konkluderer med at det er viktig å inkludere ulike sider ved læreres arbeid når lærervurdering skal brukes som grunnlag for ulike personalavgjørelser fordi mer generelle vurderingsformer ikke fanger opp kompleksiteten.

Både Cohen-Vogel 2011 og Master 2013 ser på konsekvenser av summativ vurdering av lærere, i og med at resultatene blir brukt til å plassere lærere på rett sted i skolesystemet, samt til å si opp, omplassere eller forfremme lærere. Maslow og Kelley (2012) argumenterer for at lærervurdering også må være formativ. Når lærere får tilbakemeldinger gjennom strukturert tilbakemelding, kan man styrke organisasjonens behov for fornyelse og forbedret undervisningspraksis. Studien undersøker ulike former for tilbakemeldinger, med spesiell oppmerksomhet på skolelederes vurdering av erfarne lærere i videregående skoler med stort elevmangfold. Gjennom intervju med lærere og ledere fremkom data som viste hvordan lærere i videregående skoler bruker informasjon fra lærervurderinger til å forbedre sin undervisningspraksis. Det viser seg at lærernes utbytte av vurderingen formes av skolekonteksten i vid forstand, kulturen på skolen, samt hvorvidt skoleleder oppfatter evaluering som et meningsfylt verktøy for lærer- og organisasjonslæring. Studien peker altså på forholdet mellom vurdering av den enkelte lærer og vurdering som en del av skolens kultur.

Nehring og O'Brien (2012) har undersøkt hvilke individuelle faktorer og skolefaktorer som virker fremmende eller hindrer utvikling i høyt presterende skoler. Dermed flyttes oppmerksomheten bort fra enkeltlæreren, over til vurdering og utvikling av organisasjonen. Basert på handlingsplaner og refleksjonsnotater fra 28 lærere i 14 skoler i 10 skoledistrikt (USA), analyserte forskere progresjonen til enkeltlærere og ledere som deltok i opplegget som var utviklet ved et universitet. Lederne og lærerne gikk inn i rollen som endringsagenter i skoler, og forskerne registrerte hva de rapporterte om utviklingen på skolene, det vil si hva som skjedde de fem første månedene i implementeringen av skolenes handlingsplaner. Basert på en omfattende litteraturgjennomgang presenterer studien også en modell for systemarbeid som tar utgangspunkt i trekk ved skoler som blir betegnet som sterke system og har gode resultater. Disse skolene har ledere som utvikler en felles visjon, støtter og ansvarliggjør sine ansatte, mens lærerne er aktive i sin egen

læring, blir veiledet, har en felles forståelse av målene med arbeidet, undersøker praksis, eksperimenterer i praksis, veileder hverandre og involverer eksterne endringsagenter. Funnene indikerer at 1) Når skoler har endringsagenter (ledere og lærere) som er kunnskapsrike og kompetente, er det mulig å skape endring i svake system 2) For å få til varige forbedringer trengs kunnskapsrike og høyt kompetente ledere i skoler og skoledistrikt 3) Endringsagentenes evne til å initiere forbedringer var ikke nødvendigvis koblet til posisjonsmakt, noe som bekreftes i kompleksitetsforskning som påpeker at utvikling ikke er lineær. 4) Når de arbeider med kollegaer, fokuserer dyktige endringsagenter på kultivering av intellektuell kapasitet og komplekse ferdigheter. Slike aktiviteter er det for lite av i arbeid med profesjonsutvikling når lærere trenes i forhåndsdefinerte prosedyrer. Derfor anbefaler studien at arbeid med profesjonsutvikling dreies bort fra opplæring i tekniske prosedyrer mot å legge mer vekt på å oppøve komplekse ferdigheter og intellektuell kapasitet. Forfatterne argumenterer for tett samarbeid mellom skoler og lærerutdanningsinstitusjoner. Institusjonenes kompetanse i forskningsmetoder, analyse og refleksjon kan støtte læreres profesjonslæring.

Få studier har undersøkt spesifikke lederaktiviteter som bidrar til konstruktiv lærervurdering av mer formativ karakter der vurderingen kobles til lærerens profesjonsutvikling. Tuytens og Devos (2011) undersøker hvordan transformativ ledelse og undervisningsledelse kan brukes til å stimulere læreres profesjonslæring gjennom lærervurdering. I studien analyserer forskerne data fra et spørreskjema som ble sendt til 32 videregående skoler i Belgia (640 lærere, 65 % svarte). Lærerne som deltok i studien hadde hatt minst en vurderingssamtale med en leder. De to variablene som ble studert var knyttet til hvilke aktiviteter læreren satte i gang med etter vurderingssamtalen de hadde hatt med sin leder. Den ene variabelen dreide seg om i hvor stor grad læreren eksperimenterte og reflekterte over sin praksis etter evalueringssamtalen og den andre om i hvilken grad læreren etter samtalen hentet inn ny kunnskap om praksis innenfor sitt felt, for eksempel ved å lese faglitteratur. Studien viser at ledere direkte påvirker hvordan tilbakemeldinger blir brukt og at de dermed indirekte påvirker læreres profesjonslæring. Skoleledere kan utvikle gode former for lærervurdering som blir til nytte for læreres profesjonslæring. For mange skoleledere er det imidlertid en vanskelig oppgave å gi klare og læringsfremmende tilbakemeldinger i forbindelse med lærervurderinger.

Spørsmålet om rektorer bør være undervisningsledere har lenge blitt diskutert blant forskere, men få studier har koblet undervisningsledelse til skoleprestasjoner. Grissom m. fl. (2013) undersøker hvordan lederatferd innvirker på elevenes måloppnåelse. Studien bygger på en omfattende datainnsamling fra skoler i Florida. Trente observatører skygget rektorer gjennom en skoledag og fylte ut et skjema med 50 definerte arbeidsoppgaver. Ca. 100 rektorer ble observert i tre vårsemester (2008, 2011 og 2012). Observasjonsdata ble så koblet opp mot administrative data om elevmassen og om elevenes prestasjoner. I tillegg til dette bygger studiene på strukturerte intervju og survey data fra det samme utvalget av rektorer, og det ble samlet administrative data om rektorene og deres ansettelsesforhold. Forskerne finner at rektorer som bruker tid på undervisningsfunksjoner generelt, altså at de bruker mye tid og har sin oppmerksomhet rettet mot lærernes undervisning, ikke bidrar til å bedre elevenes læringsutbytte eller fremmer skoleutvikling. Når forskerne undersøker mer inngående ledernes handlinger, viser det seg at hva de gjør med hensyn til å engasjere seg i lærernes undervisning gir utslag år etter år. Det som særlig gir positive utslag på elevenes prestasjoner er den tiden som blir brukt til å veilede lærere, og arbeid som blir lagt ned i å utvikle skolens undervisning (både hva det undervises i, hvordan det undervises og vurderes). Tid som blir brukt på det vi kan kalle skolevandring, der ledere går inn og observerer for å støtte læreren med det han eller hun trenger, viser seg å gi negative utslag på elevprestasjoner, spesielt i videregående skoler. Survey og intervjudokumentasjon viser at denne negative sammenhengen gjerne oppstår fordi rektorer ofte ikke bruker besøk i klasserommet som del av en større skoleutviklingsstrategi.

Mens mye av forskningen om lærervurdering og prestasjonsledelse (perfomance management) har undersøkt lærernes erfaringer, tar Moreland (2009) utgangspunkt i hva ledere tenker om lærervurdering og

prestasjonsledelse. Han undersøker hvordan ledere i videregående skoler forstår hensikten med prestasjonsledelse ved å sammenligne metodene ledere bruker for å implementere lovpålagte forandringer, hvordan de reflekterer over det å være leder og sammenhengen mellom egen ideologi og oppfatning av prestasjonsledelse. Videre blir koblinger mellom prestasjonsledelse og strategisk ledelse undersøkt. Ved hjelp av et balansert «scorecard» får man frem lærernes stemmer og en profil på ledelsens undervisnings- og ledelseskompetanse. Moreland finner at det er en sammenheng mellom den daglige praksisen i skoler, strategiske prioriteringer, standarder som en skal jobbe etter, kompetanse, og måten lederen involverer seg i lærerens karriereplanlegging. Prestasjonsledelse ble imidlertid gjennomført på ulike måter. Noen ledere la vekt på å anerkjenne læreres resultatoppnåelse, å vurdere deres bidrag til profesjonen, hjelpe dem å utvikle seg og gi råd om karriereveier. Andre ledere forsikret seg om at lærerne utviklet seg profesjonelt slik at det kunne komme elevene til gode, og andre igjen var opptatt av skolens rykte og den organisatoriske utviklingen. Denne studien viser at det åpner seg et stort forskningsfelt som handler om å undersøke mulige sammenhenger mellom hvordan prestasjonsledelse blir praktisert innad i skoler, rektors lederstil og hans eller hennes relasjon til skolens mellomledere.

Forskningslitteraturen avdekker et bilde som viser at i spørsmål om lærervurdering ser det ut til å være store forskjeller på hvilket ledelsesnivå man ser på. System- og policy-nivået har andre behov for data enn operative skoleledere. Likevel kan det oppstå problemer når sentrale eller lokale myndigheter forventer at den enkelte skoleleder skal ha kompetanse til å tolke og «oversette» data som genereres fra store databaser til å forbedre skolens undervisning slik at det gir seg utslag i bedre læringsutbytte for elevene. Skoleledere er forskjellige og har ulike oppfatninger av hva det betyr å være leder. Noen får til gode rutiner for formativ vurdering, og forskningen viser hva disse lederne gjør.

4 Tverrgående tema i forskningslitteraturen

I dette kapitlet presenteres resultatene fra det tredje trinnet i den narrative syntesen. Her er målet å identifisere mønstre i materialet og å samle og komprimere de resultatene som slår sterkest ut i de inkluderte studiene. Det viktigste på dette stadiet er å overskride primærstudiene og undersøke funnene på tvers av de opprinnelige kategoriene (kapittel 2) og den forberedende syntesen (kapittel 3). Gjennom en slik prosess blir det mulig å avdekke spenninger og motsetningsforhold i materialet og få svar på spørsmål som hva som kan fremme eller hemme utvikling.

Ved å lese materialet på tvers ble det identifisert tre hovedtema som har stor relevans for arbeid med lærervurdering. Det første temaet handler om ledelse på ulike nivåer og om forskjellige former for vurdering. Lærere forholder seg normalt til tre ledelsesnivåer: sentrale myndigheter som utformer den nasjonale skolepolitikken, skoleeiere (kommuner og fylker) som forvalter ressursene og skoleledelsen på den enkelte skole. Det andre temaet tar opp den gjennomgående spenningen mellom de to formålene med lærervurdering, det å måle resultater (summativ vurdering) og det å bruke resultater til læring og utvikling (formativ vurdering). Det tredje temaet drøfter de praktiske implikasjonene av at det er ulike intensjoner og formål med lærervurdering og forskjellige måter å organisere lærervurdering på.

4.1 Tema 1: Ledelse av ulike former for lærervurdering

I så godt som alle studiene som har undersøkt spørsmål med relevans for lærervurdering, finnes det informasjon om betydningen av ledelse på forskjellige nivåer. Hvordan lærevurdering struktureres og arbeidet gjennomføres, ledes, evalueres og forbedres, er av stor betydning for hvor vellykket den skal bli.

Nesten alle artiklene som ser på systemer for lærervurdering med hovedvekt på summative data er fra USA. I de fleste tilfellene har forskerne gjennomført sekundæranalyser av data som brukes i store programmer i delstater som California (Darling-Hammond m.fl. 2013); Florida (McCaffrey m.fl. 2009; Sass m. fl. 2014; Winters og Cowen 2013); North Carolina (Rothstein, 2010; Kinsler, 2012; Sass m. fl. 2012; Goldhaber m.fl. 2013a); Ohio (Kane m. fl. 2011) og Texas (Alfaro m. fl. 2013) samt byer som Chicago (Wang m. fl. 2013); Los Angeles (Buddin og Zamarro 2009); New York (Goodman og Turner 2013; Boyd m. fl. 2013); San Diego (Koedel 2009) og San Francisco (Newton m. fl. 2010). Forskningen viser at sentrale myndigheter hovedsakelig er opptatt av summativ vurdering som gir oversikt over og mulighet for kontroll med ressursinnsats og ressursbruk. Sentrale og lokale myndigheter kobler data om elevresultater mot data om ressursinnsats, for eksempel bygg og oppvarming, spesialundervisning, antall lærere og assistenter i skoler og kommuner, utgifter til skolemat, antall elever som har krav på gratis lunsj etc. På dette nivået er det ønskelig å finne effekter av ulike initiativ og tiltak for å målrette ressursinnsatsen bedre. Mange som har ansvar for politikkutforming er også interessert i å få svar på om prestasjonslønn kan få lærere til å yte mer

slik at elevenes læringsutbytte øker. Da prøves prestasjonslønn ut i noen skoler, effektene av tiltaket måles, og så bruker man skoler hvor lærerne ikke får prestasjonslønn til å kontrollere for resultatene.

I følge forskerne er et resultat av en slik summativ orientering at noen skoleledere bruker resultatdata fra lærervurdering strategisk i personalavgjørelser som ansettelser, forfremmelser og oppsigelser (Goldhaber og Theobald 2013; Cohen-Vogel 2011). Hvis lederens mål primært er å bedre skolens gjennomsnittlige skåre, kan det for eksempel skje at ledere baserer seg på summativ informasjon og ansetter lærere som kan øke elevenes testresultater («staffing to the test»). I slike tilfeller rettes interessen mot ethvert tiltak som kan bedre skolens samlede testresultat – også de menneskelige ressursene.

Når de blir spurt om hva som er viktig for skolen, viser det seg at forskjellige ledere har ulike vurderinger og prioriteringer. I en undersøkelse av hvordan skoleledere reflekterer rundt temaet lærervurdering og prestasjonsledelse (performance management) finner Moreland (2009) at ledere tenker ulikt når de blir bedt om å gi innhold til begrepet prestasjonsledelse. Mens noen ledere er mest opptatt av å knytte prestasjoner til de resultatene lærerne oppnår og mener at lederens ansvar først og fremst er å gi lærerne råd om etterutdanning og karriereveier, er andre mer opptatt av hvordan de best kan støtte lærernes profesjonsutvikling. En tredje gruppe ledere er mer orientert mot organisasjonsutvikling og hvordan omverdenen oppfatter skolen. De mener at det er viktig at skolen fremstår som utviklingsorientert og har et godt rykte.

I andre studier blir det påpekt at det påfallende ofte er sammenfall mellom hvordan skoleledere vurderer lærerne, hva resultater fra surveydata og testresultater viser, og hvordan foreldre og elever vurderer de samme lærerne. Roby (2012) finner også sammenfall mellom læreres egenvurdering og kollegavurderinger av de samme lærerne. I flere av studiene understrekes betydningen av lederstøtte, altså at skolelederne støtter lærernes profesjonsutvikling.

Både Tuytens og Devos (2011), Delvaux og Vanhoof (2013) og Grissom m. fl. (2013) finner at skoleledere som tar sitt ansvar som undervisningsledere alvorlig, har bedre forutsetninger for å lykkes med å styrke skolen som organisasjon og bedre elevenes læringsutbytte. Master (2013) mener i tillegg at det har betydning for resultatkvaliteten hva ledere gjør og hvilke konkrete initiativ de tar for å forbedre undervisningen på skolen. Det er ikke tilstrekkelig at de gir uttrykk for å være interessert i lærernes undervisning. Selv om holdninger er viktige, er det handlinger som gir utslag på elevenes læringsutbytte. Noen handlinger gir større utslag enn andre, for eksempel tilbakemeldinger som kommer til riktig tid og at lærerne opplever følelsesmessig støtte fra ledelsen (Brackett m. fl. 2010). Det viser seg også at jo mer aktivt skolelederne har vært involvert i selve vurderingsprosessen, jo større legitimitet tilskriver både lærere og ledere et system for lærervurdering (Taut m. fl. 2011).

Master (2013) omtaler undervisningsledelse som aktiviteter som har profesjonslæring som mål og systematisk bruker formativ vurdering for å bedre undervisningen. Surveydata og prosenttall som viser gjennomsnittlige elevresultater gir ikke informasjon om hvordan den enkelte lærer fungerer i et kollegium. Derfor kan vurderingssystemer som er for opptatt av å måle enkeltlæreres prestasjoner komme til å overse at lærere har ulike oppgaver, er individuelle personligheter og ivaretar forskjellige funksjoner i et kollegium. En viktig ressurs for skolen er den daglige innsatsen hver enkelt lærer gjør ved å støtte, avlaste, motivere og supplere hverandre. Hvis man flytter blikket fra den enkelte og tenker helhetlig (Grissom m. fl., 2013), kan skolen betraktes som et kollektivt potensial av muligheter som undervisningsledere kan ha stor nytte av i arbeidet med å styrke og utvikle hele organisasjonen. Betydningen av at ledelsen tenker helhetlig blir også fremhevet i andre artikler. Både Finn m. fl. (2009); Kane m. fl. (2011) og Grossmann m. fl. (2013), understreker at det er læreres samlede ferdigheter som gjør dem til gode lærere. Selv om enkelte handlinger kan ha statistisk effekt, og noen større effekt enn andre, finner de, ved å kombinere resultater fra

testresultater med resultater fra observasjoner av undervisning og aktiviteter i klasserommet, at det er summen av det læreren gjør som gir størst utslag på elevenes læringsutbytte.

Det som særlig gir positive utslag på elevenes prestasjoner, er den tiden ledere bruker til å gi lærerne tilbakemeldinger og den tiden lærerne bruker til å gi elever tilbakemeldinger, samt arbeid som skoleledere og lærere legger ned i å utvikle skolens undervisning (både hva det undervises i, hvordan det undervises og vurderes). Tilbakemeldinger er en del av didaktikken og dermed en del av lærerens undervisning. I et profesjonsperspektiv lærer man best ved å gi tilbakemeldinger som en naturlig del av undervisningen. Derfor understreker også mange av de inkluderte artiklene hvor viktig lederen er som undervisningsleder. En viktig oppgave blir å identifisere og bli enige om kvalitetskriterier for ulike aktiviteter i skoler og klasserom. Master (2013) finner at det er lettere å få lærere til å bli enige om kriterier for god klasseledelse enn kriterier for gode relasjoner i skolen.

Ledelsens betydning i arbeidet med lærervurdering på alle nivåer diskuteres også i de artiklene som beskriver systemer for lærervurdering (kapittel 3). I Chile, Kina, Belgia og Portugal, er det bestemt nasjonalt at det skal innføres lærervurdering. I varierende grad beskriver artiklene arbeidet med involvering av de aktørene vurderingen angår, men alle understreker betydningen av at det blir satt av tilstrekkelig tid til aktiv deltakelse i utformingen av systemet. Det må videre være enighet om hvorfor systemet skal innføres (hvilken hensikt det skal tjene), hvilke mål man tar sikte på å nå gjennom systemet og hvordan resultatene som genereres skal brukes. Hvis disse helt grunnleggende prinsippene ikke overholdes, kan tilliten til både system og ledelse svikte blant de som skal arbeide med vurderingen. Etter at sentrale myndigheter har bestemt at det skal innføres et system for lærervurdering, og systemet er utformet, blir implementeringen overlatt til lokale myndigheter som får ansvaret for å gjennomføre det praktiske arbeidet.

4.2 Tema 2: Spenning mellom summativ og formativ vurdering

Det andre temaet som avtegnet seg ved å lese artiklene på tvers, er at det finnes en sterk spenning mellom summativ og formativ vurdering i materialet. Denne spenningen går igjen i flere sammenhenger og ytrer seg på forskjellige måter, for eksempel i metodediskusjoner som er gjengitt i rapporten, men også ved at lærerstudenter som opplever motstridende praksiser lurer på hva som «egentlig» er vurderingens formål (Taber m. fl. 2011). Hvis bedre praksis er intensjonen med lærervurdering, må ledere på alle nivåer forstå hvordan de skal håndtere spenningen mellom summativ og formativ vurdering.

Resultater fra summative vurderinger kan brukes på tre måter: Strategisk, administrativt og utviklende. I en OECD-rapport påpeker Looney (2011) at i systemer som har ambisjon om både å ivareta summative og formative mål, strever man med å ivareta den formative delen av vurderingen på en god måte. Prinsippet er at informasjon fra data som genereres gjennom ekstern vurdering, skal brukes til å forbedre undervisning og læring lokalt. Det viser seg imidlertid å være svært vanskelig å få til sømløs integrasjon av summativ og formativ vurdering. Dels skyldes dette at data fra tester ikke er på et detaljeringsnivå som gjør at de kan brukes til å diagnostisere enkeltelevers behov. De kommer heller ikke på det tidspunktet de trengs. I tillegg er det store utfordringer knyttet til å utvikle pålitelige mål for metakognisjon (higher order skills) og kompetanse i problemløsning og samarbeid som forventes i nyere læreplaner.

Det kan se ut som om de problemene som skaper spenninger i forholdet mellom summativ og formativ vurdering særlig kommer til uttrykk i forbindelse med overføringen av ansvar fra sentralt til lokalt nivå (kapittel 3.1). Hvem som har ansvar for hva er ikke klart nok uttrykt, og det er heller ikke gode nok

mekanismer for å kontrollere at de som har ansvaret faktisk tar det. For eksempel er et av de viktigste formative tiltakene i systemet i Chile en plan for profesjonsutvikling (Taut og Sun 2014), noe myndighetene og lærerorganisasjonene sentralt har blitt enige om. Når oppfølgingen av det som oppfattes som den viktigste formative komponenten i systemet svikter lokalt, får heller ikke systemet den tiltenkte effekten. I stedet for å fungere formativt (lærende og utviklende), får systemet andre formål som forskere (Flores 2013; Pham og Stacey 2012) beskriver som administrative og byråkratiske. De kan bli unødig komplekse og kompliserte (Leigh 2013; Flores 2013; Pham og Stacey 2012; Taut og Sun 2014), og vanskelige å håndtere. Om systemet skal ha legitimitet avhenger også av kvaliteten på metoder og verktøy som blir tatt i bruk. Det er heller ikke nødvendigvis slik at lærere er enten for eller mot lærervurdering. Det kan være store nyanser innad i lærergruppen med hensyn til hvordan de betrakter tiltak for lærervurdering (Friedrich m. fl. 2012). Å forstå hva nyansene handler om er viktig for ledere som nettopp trenger slik informasjon om vurderingen skal fungere formativt.

Det er også store lokale forskjeller i selve implementeringen av systemer for lærervurdering; det vil si hvordan tiltak blir presentert, forankret og fulgt opp og hvor godt man klarer å synliggjøre sammenhenger mellom systemets ulike deler. En erfaring fra Chile er for eksempel at skolelederne engasjerer seg for lite i selve vurderingsarbeidet. Dermed går man ofte glipp av formative muligheter som kunne ha bidratt til læring både for den enkelte og organisasjonen. Det er grunn til å tro at jo mer komplekst og ambisiøst systemet er, jo mer kreves det av lokal ledelse (skoleeier og skoleleder) som skal tenke helhet og sammenheng og koordinere det hele. Hvis det er for mange mål som skal innfris samtidig, blir det også en spenning mellom målene. I systemer som er for ambisiøst lagt opp, blir utfordringen lokalt å håndtere kompleksitet, både horisontalt og vertikalt. Det er tidkrevende, involverer mange aktører og forutsetter hyppig samhandling.

Knyttet til denne spenningen mellom summativ og formativ vurdering er det to forhold som er viktige. For det første ser det ut til å være av betydning hvorvidt skoleledelsen bruker vurderingsdata som er generert av andre (utenfor skolen) eller om lederne selv deltar aktivt i innsamlingen av de dataene som skal brukes i vurderingsprosessen. Dette kan oppfattes som en spenning mellom «andres» data og «egne» data. De som har deltatt aktivt i prosessen med å generere data har mindre problemer med å bruke resultatene formativt enn de som forholder seg til andres data. Egne data bidrar til større forståelse av hva som er problemet og gjør det lettere å forstå hva som må gjøres med problemet. Man skjønner hva problemet bunner i og har utviklet et «eierforhold» til det, noe som også gjør det lettere å finne ut hvordan det skal løses.

Det andre forholdet er knyttet til om utvikling av læreres kompetanser skal være preget av tekniske prosedyrer eller om det skal handle om å utvikle intellektuell kapasitet (Nehring og O'Brien 2012). Dette kan tolkes som en spenning mellom å betrakte undervisning som en profesjonskompetanse eller undervisning som en rekke teknikker. Grossman m.fl. (2013) mener at forskning om lærervurdering hittil ikke har vært tilstrekkelig opptatt av at undervisning faktisk er et intellektuelt, profesjonelt arbeid som består av ulike deler som henger sammen og forutsetter hverandre. Det er sammenheng mellom fagstoffet som presenteres, hvordan det presenteres og hvordan læreren får elevene til å bli interessert i lærestoffet og engasjere seg i undervisningsaktivitetene.

4.3 Tema 3: Formål med og innretting av systemet

Med den store interessen for vurdering som nå preger diskusjoner om utdanning, kan man komme til å tro at dette er noe nytt. Lærere har imidlertid «alltid» arbeidet med vurdering – både av elevers og eget arbeid. I dag har imidlertid vurdering fått ny relevans som en sentral ferdighet i nye styringssystemer. Forskningen som er gjennomgått, viser imidlertid at man må bygge på den kompetansen og kunnskapen som allerede finnes i sektoren når målet er å styrke og fornye praksis.

I et system for kvalitetsvurdering er lærervurdering en av flere komponenter. Hvis lærervurdering skal bli vellykket, må den forankres i systemet og ha berøringspunkter med de andre delene. Det betyr at ledelsen får mange variabler å forholde seg til. Store mengder kvalitative og kvantitative data gir komplekse datasett. Her er det omfanget av data som er problemet, ikke om de er kvantitative eller kvalitative. Selv om ambisjonen er at vurderingen skal føre til læring og utvikling, oppstår problemer når informasjon fra resultater skal omdannes til kunnskap om hvordan man skal bruke resultatinformasjonen til å utvikle kvaliteten i skolen. Dette forutsetter metode- og implementeringskompetanse.

Det kan være vanskelig å se koblingen mellom informasjon som fremkommer gjennom summativ vurderingsinformasjon og formative vurderingstiltak. Måling og utvikling følger ulike logikker. Utfordringen blir å få summativ informasjon omdannet til handlingsrelevant kunnskap som er forståelig for de som skal forbedre resultatene. Mens tallene som genereres på systemnivå gir abstrakt informasjon om gjennomsnittlige resultater, skal forbedringer skje gjennom konkret praksis i skoler, klasser og på individnivå. Tilbakemeldinger er mest effektive når de er konkrete og kommer til riktig tid. Derfor må det være kort vei mellom resultatene og handlingene i skolen. Men resultatene kan ikke komme i form av rådata. Det er en lederoppgave å analysere og presentere dem i den konteksten de skal komme til nytte.

Det er to forhold som fremstår som viktige her. Det første omhandler betydningen av å begynne i det små. Det har vist seg at systemer som baserer seg på summativ vurdering ofte begynner i bredden og forsøker å håndtere store mengder data samtidig. Forskningen har blant annet vist at det er viktig å bygge på den kunnskapen som allerede finnes i skolen (Howley m. fl. 2013). Dermed kan man også bedre ivareta de følelsesmessige sidene ved vurdering (Kleinknecht og Schneider 2013). Dessuten kan enkle metoder som egenvurdering fungere bra hvis de inngår i en sammenheng og har en formativ funksjon (van Diggelen m. fl. 2013). Det anbefales å koble egenvurdering og kollegavurdering, både fordi kollegaer er konstruktive, og fordi det ikke er store sprik mellom egenvurdering og kollegavurdering. Graves (2009) mener at elever som vurderer lærere kan gi verdifull informasjon som lærere kan bruke til å forbedre sin praksis. Vurdering er også en praksis som hele tiden kan bli bedre. Lærere som jevnlig får tilbakemeldinger fra elevene og systematisk gjennomfører egenvurdering og kollegavurdering øker sin vurderingskompetanse og kan bli flinkere til å gi elevene læringsfremmende tilbakemeldinger.

Det andre forholdet omhandler betydningen av delaktighet i både utforming og implementering. Utvikling er ikke alltid lineære prosesser, og ikke alle lærere blir motivert til ekstra innsats av eksternt press (Christophersen m. fl. 2012; Nehring og O'Brien 2012). At lærerne kjenner seg igjen i måten ting blir gjort på er også av betydning. Lærervurderingen, metodene som brukes og kompetansen hos de som vurderer må ha legitimitet og troverdighet blant lærerne som skal vurderes. Forskerne påpeker at lærervurdering må ha lederstøtte og at lederen må utvikle kompetanse i undervisningsledelse. Mappevurdering innebærer mye arbeid, både for lærere som skal lage mappene og for de som skal vurdere innholdet i dem. Forskerne mener imidlertid at det er et stort læringspotensial i bruk av ny teknologi – for eksempel digitale mapper (Hinz 2011; Tinoca og Oliveira 2013), eller videopptak av undervisning (Brantlinger m. fl. 2012; Kleinknecht og Scneider 2013). Læringspotensialet knyttes til at deltakerne får et felles objekt å analysere (Brantlinger m. fl. 2012). Det er nødvendig å skille oppgaver fra hverandre når man skal studere og analysere dem. Profesjonskompetanse handler imidlertid om å bringe kunnskap fra ulike kilder *sammen*, med gjensidig berikelse som mål.

5 Prinsipper for vurdering i systemer for lærervurdering

Det er gode argumenter for at det bør være konsistens i skolens faglige arbeid. Hvis man skal innføre lærervurdering, er det derfor fornuftig å bygge på gjeldende prinsipper for skole- og elevvurdering. Selv om ikke alle skoler arbeider systematisk med egenvurdering, er det mange som gjør det. De har etablert en vurderingskultur og utviklet en vurderingskompetanse som det må bygges videre på. De siste årene har også mange skoler systematisk utviklet kunnskap om vurdering gjennom arbeid med elevvurdering (vurdering for læring). Det finnes følgelig en vurderingskompetanse blant skoleeiere, skoleledere og lærere som må brukes når det skal gjennomføres vurderinger i andre deler av utdanningssystemet.

Gjennomgangen av forskningslitteraturen har vist at systemer for lærervurdering ofte er initiert nasjonalt. : I rapportens innledning presenteres noen generelle prinsipper og retningslinjer for vurdering som er et resultat av tretti års forskning, og som bør ses i sammenheng med erfaringer fra systemer for lærervurdering. Det handler for eksempel om betydningen av å bli enige om hva som skal vurderes. Fordi lærerens praksis er kompleks og sammensatt, blir det særlig viktig å avgrense vurderingsobjektet. Forskningslitteraturen gir eksempler på at det kan bli uklart hva som skal vurderes når man skal dekke svært mange sider ved lærernes arbeid (Taut og Sun 2014, Flores 2012, Delvaux og Vanhoofen 2013). Eksemplene fra forskningslitteraturen (kapittel 3.1) viser forsøk på å kombinere mange ulike vurderingsmetoder. I Chile brukes for eksempel mappevurdering, kollegavurdering, vurdering av veileder og leder samt egenvurdering (Taut og Sun 2014). I Kina benyttes egenvurdering, avdelingens vurdering av læreren, skolens vurdering av læreren, klasseromsobservasjon, elevenes eksamensresultater samt inspeksjon av lærernes daglige arbeid (Zhang og Ng 2011, Liu og Zhao 2013).

Kombinasjonen av uklart vurderingsobjekt og mange vurderingsmetoder fører til at systemer for lærervurdering blir vanskelige å håndtere. Alle systemene hadde som intensjon både å være formative, det vil si å bidra til lærernes profesjonsutvikling og summative, det vil si å måle resultater. Likevel tok resultatfokuset overhånd og de formative ambisjonene ble ikke godt nok fulgt opp. Dette resulterer gjerne i at systemene blir redusert til administrasjon og byråkrati i stedet for å bidra til den faglige utviklingen ved skolene, slik hensikten var.

Prinsippene for vurdering som er utviklet av Assessment Reform Group (se side 12) gir retningslinjer for arbeidet med god vurderingspraksis. Her analyseres erfaringene fra systemene for lærervurdering i Chile, Kina, Portugal og Belgia i lys av fem prinsipper. Ved å analysere artiklene på tvers identifiseres problemer som avdekkes når prinsippene for vurdering blir fulgt og når de blir brutt. En tabelloversikt (vedlegg 9) viser hvordan de ulike systemene følger eller bryter med prinsippene for vurdering.

Prinsipp 1:
De ressursene som trengs for å gjennomføre vurderingen må stilles til rådighet (ekspertise, økonomi, tid) og innsatsen må balanseres mot det man forventer å få ut av aktiviteten.

I systemer for lærervurdering som benytter mange ulike vurderingsmetoder, kreves det store ressurser for å gjennomføre og følge opp vurderingen. I eksemplene kommer det tydelig frem at det ikke er satt av nok ressurser til vurderingen. Det er særlig mangel på tid som blir rapportert. Lærerne opplever at de verken har tid nok til å gjennomføre selve vurderingsarbeidet eller følge opp tiltakene i etterkant av vurderingen. Når vurderingen er omfattende, fører den også til mye byråkrati og tilleggsarbeid for lærerne (Taut og Sun 2014, Tornero og Taut 2010, Flores 2012). Det er dessuten store lokale variasjoner når det gjelder ressursinnsats og kompetanse i vurdering i kommuner og regioner, noe som fører til ulik praksis og varierende kvalitet i vurderingsarbeidet (Santiago m.fl 2013, Liu og Zhao 2013). I utviklingen av et system for lærervurdering blir det viktig å avgjøre om lærerne først og fremst skal bruke tiden på formative eller summative aktiviteter.

Prinsipp 2:
Vurderingen må planlegges og gjennomføres på en slik måte at den eller de som blir vurdert, opplever resultatet av vurderingen som gyldig og troverdig.

Delaktighet i både utforming og implementering er viktig. Videre må et system for lærervurdering, metodene som brukes og de som vurderer ha legitimitet og troverdighet blant lærerne som blir vurdert. Lærere forventer solid faglighet av de som vurderer. Verken i Chile, Portugal eller Belgia hadde lærerne tillit til vurderingskompetansen hos de som gjennomførte vurderingen (Santiago m.fl. 2013, Flores 2012, Delvaux og Vanhoofen 2013). Heller ikke i Kina opplever lærerne at gjennomføringen lokalt er troverdig (Liu og Zhao 2013). Dette viser at disse systemene for lærervurdering bryter med det andre prinsippet for vurdering, noe som for eksempel kan føre til motstand blant lærerne (Tornero og Taut 2010). Forskerne påpeker at et system for lærervurdering forutsetter kunnskap om vurdering og at ledere må utvikle kompetanse i undervisningsledelse.

Prinsipp 3:
Vurderingen må avgrenses til og konsentrere seg om bestemte sider ved arbeidet, men likevel ta hensyn til at det som vurderes inngår i en større sammenheng

Innledningsvis ble det påpekt at vurderingsobjektet må være klart definert. Mange av problemene som ble avdekket i studiene kan settes i sammenheng med størrelsen på og kompleksiteten ved systemene (Taut og Sun 2014, Santiago m.fl. 2013). Når vurderingen ikke er godt nok avgrenset og det samles inn svært mye data, kan det også bli vanskelig å bruke resultatene på en god måte. En ting er å samle inn tall, data og informasjon. Noe annet er det når denne informasjonen skal omdannes til kunnskap og brukes til å forbedre praksis. Da gjelder det å ha et faglig blikk for hvordan delene henger sammen og informerer hverandre. I et system for kvalitetsutvikling er lærervurdering som regel en blant flere komponenter, men vurderingen av lærernes arbeid må ses i sammenheng med de andre komponentene.

Prinsipp 4:
Vurdering bør gjennomføres på måter som ikke bare måler prestasjoner, men som også generelt bidrar positivt til arbeidet som utføres i skolen og styrker læringsmiljøet

Alle systemene hadde det som kan kalles doble intensjoner. Det vil si at de både skulle måle prestasjoner (summativ) og føre til profesjonsutvikling og forbedringer i lærernes undervisningspraksis (formativ). Det viser seg at det er vanskelig å oppfylle systemenes formative intensjon. De fleste opplever problemer når vurderingen først og fremst skal føre til læring og utvikling – hos den enkelte og på organisasjonsnivå (Taut og Sun 2014, Santiago m.fl. 2013, Liu og Zhao 2013, Delvaux og Vanhoofen 2013, Flores 2012). De får ikke dette til slik det var tenkt. Forskningen viser at tilbakemeldinger må gis på riktig tidspunkt, og at de må være konkrete og oppleves som nyttige av de som får dem. Videre kan det være et godt grep å lage utviklingsplaner for alle lærerne (Santiago m.fl. 2013), ikke bare de som får dårlige vurderinger.

Prinsipp 5:
Vurdering har generelt stor innvirkning på praksis i et felt, så vurderingen må planlegges og gjennomføres på en slik måte at uønskete virkninger av vurderingen minimeres

For å unngå uønskede virkninger av vurderingen, er det nødvendig med forankring og delaktighet i både utviklingen og implementeringen av et vurderingssystem. En enkel måte å sikre forankring er å ta utgangspunkt i den vurderingskompetansen som finnes blant lærere og elever. De ulike systemene som er gjennomgått, har alle hatt ikke-intenderte virkninger som for eksempel omfattende byråkrati. I Kina ble vennskap (personlige bånd) viktigere enn hardt arbeid for å oppnå gode resultater (Liu og Zhao 2013). I Chile brukes systemet stort sett til å holde lærerne ansvarlige for elevenes resultater (Santiago m.fl. 2013) og fungerer i liten grad formativt. I Portugal måtte systemet gjennom en omfattende revisjon for å imøtekomme kritikk fra lærerne og deres organisasjoner (Flores 2012).

6 Former for lærervurdering som kan ha positiv innvirkning på skolens kvalitet

Den systematiske gjennomgangen av forskningslitteraturen viser at hva som oppfattes som kvalitet i skolen avhenger av aktørenes posisjon i forhold til utdanningssystemet og av hvordan de oppfatter skolens funksjon i samfunnet. Noen ser pragmatisk på skolen som et sted hvor fagkunnskaper overføres fra lærer til elev. Andre har en mer holistisk og systemisk forståelse og ser ikke bare skolen som et sted for kunnskapstilegnelse og læring, men også som en viktig møteplass for sosialisering og omsorg, hvor barn skal oppleve mestring, vennskap, anerkjennelse og aksept. Skolen kan dessuten betraktes som et bidrag til samfunnsmessig inklusjon, integrasjon, demokratiutvikling, dannelse og utjevning av sosioøkonomiske forskjeller. Hva som er kvalitet i skolen, er altså et spørsmål som kan resultere i et mangefasettert svar.

Hensikten med å innføre lærervurdering er først og fremst å forbedre kvaliteten på elevenes utdanningstilbud. Det er enighet blant forskerne om at medvirkning er nødvendig for at lærervurdering skal fungere godt. Elever kan for eksempel delta med tilbakemeldinger om gode og dårlige læringsopplevelser og om hva som kan gjøres annerledes. Dette kan skje både muntlig og skriftlig, individuelt og i grupper. Selv om det finnes lite forskning om elevers vurdering av læreres arbeid, er det rimelig å anta at hvis elever får anledning til å komme med innspill om gode og dårlige læringsopplevelser, og innspillene fra elevene tas alvorlig, vil dette styrke deres opplevelse av delaktighet og medvirkning i skolen.

Resultatene fra den systematiske kunnskapsoversikten viser at følgende fire forutsetninger må være på plass for at lærervurdering skal kunne bidra til god kvalitet i skolen:

Metodekompetanse

For at lærervurdering skal ha positiv innvirkning på skolens kvalitet, trengs metodekompetanse på alle nivåer. Lærere, skoleledere og skoleeiere må ha et godt grunnlag for å forstå hva slags informasjon man får fra ulike typer vurderingsresultater (kvalitative og kvantitative), og hvordan resultater fra vurderingen kan brukes i arbeid med å forbedre praksis i skolen. Generelt må skoleeiere, skoleledere og lærere kunne vurdere om det som gir seg ut for å være kunnskap faktisk er kunnskap. Metodekompetanse er også en nødvendig hjelp i arbeidet med å dimensjonere et system for kvalitetsvurdering hvor lærervurdering er en av komponentene. Innsikt i metode øker forståelsen av hva man vinner ved summative og formative vurderingspraksiser, hvor mye (og hva slags) data man trenger å samle inn og hvordan metoder kan kombineres og data gjøres om til kunnskap som kan brukes til å forbedre praksis i skolen. Metodekompetanse er nødvendig for å kunne tolke prosenter og statistikk. Det er viktig å ha metodekompetanse for å vite at man vurderer det man skal (validitet) og for å avklare om de resultatene man forholder seg til er pålitelige (reliabilitet).

Medvirkning, ansvar og tillit

For at lærervurdering skal ha positiv innvirkning på skolens kvalitet, er det nødvendig å sikre delaktighet og medvirkning på alle nivå. Vurdering må være noe man sammen tar del i. Hvis medvirkning er et premiss i vurderingen, styrkes relasjonene horisontalt og vertikalt; mellom nivåene (sentrale og lokale myndigheter og den enkelte skole) samt internt i skolen (for eksempel lærer-leder, lærer-lærer, lærer-elev, elev-elev). Det er lettere å forstå betydningen av data som man selv har vært med på å generere. Medvirkning øker følelsen av ansvar for de resultatene som foreligger og har betydning for hvordan vurderingstiltakene blir forankret i organisasjonen. Delaktighet skaper tillit og kan gjøre det lettere å få frem den kunnskapen som allerede finnes i skolen. Målet er å få til møter mellom eksterne tiltak og kjent praksis, slik at det kan skje innovasjon og fornyelse i skolen. Medvirkning bidrar også til at tiltakene blir forankret hos både lærere, skoleledere, skoleeiere, lærerorganisasjoner og myndigheter. Bred medvirkning gir sterkere ansvarsfølelse og eierskap til aktivitetene. Hvis systemet skal styrke lærerprofesjonen, må både lærere og ledere ha en klar oppfatning av hvordan profesjonslæring bør skje.

Tydelighet og enkelthet

For at lærervurdering skal ha positiv innvirkning på skolens kvalitet, må den være strukturert slik at den motvirker kompleksitet. Dette forutsetter tydelighet og enkelhet. Det må skapes en felles forståelse av 1) hva det er som skal vurderes, altså at man avklarer vurderingsobjektet, 2) hvordan det skal vurderes, altså hvilke metoder som skal benyttes og 3) hvordan resultatene skal brukes. I utviklingen av et system for lærervurdering bør man starte i det små med å samle inn så mye data som det er mulig å håndtere og som er på et slikt nivå at informasjonen kan brukes i forbedringsarbeid. Forskningen viser at det kan være vanskelig å avgrense vurderingsobjektet i forbindelse med lærervurdering. Læreres arbeid er langt mer komplisert enn det umiddelbart virker. Forskningen anbefaler derfor at man i stedet for å starte bredt og omfattende begynner i det små og heller utvider etter hvert som man høster erfaringer.

Ansvarsplassering, dialog og oppfølging

For at lærervurdering skal ha positiv innvirkning på skolens kvalitet, har det betydning hvordan ansvar kommuniseres og overføres fra nasjonalt til lokalt nivå. Det er viktig at skoler og kommuner har lokalt handlingsrom, men like viktig er det å ha mekanismer på plass for å følge opp at de faktisk tar det ansvaret de har. Vellykket implementering forutsetter tydelig ansvarsplassering, dialog og kontinuerlig oppfølging. I stedet for å betrakte implementering som en top-down, lineær prosess, kan den oppfattes som en sum av flere parallelle, iterative aktiviteter. Implementering er verken fullført når nasjonale myndigheter har bestemt at det skal innføres et system for vurdering eller når vurderingsresultatene foreligger. Implementering forstås bedre som en kontinuerlig lærings- og forbedringsprosess som går ut på å bruke resultater fra vurdering i arbeidet med å forbedre praksis. Praksis kan alltid bli bedre. Implementering i komplekse systemer forutsetter at den oppfattes som et vedvarende forbedringsarbeid.

Denne systematiske kunnskapsoversikten har vist at hvis lærervurdering skal ha positiv innvirkning på skolens kvalitet, må den basere seg på en kombinasjon av flere metoder. Den må bestå av flere komponenter og ha som hovedhensikt å bruke summative resultater formativt for å styrke lærerprofesjonen. Samtidig er det viktig å unngå at lærervurderingen blir unødig kompleks.

Dette forutsetter at ledelsen både har kunnskap om vurdering og den kompetansen som trengs for å kunne håndtere kompleksitet. Hvis lærervurderingen skal føre til læring og utvikling, må ledelsen kommunisere enkelt og tydelig og sørge for delaktighet og involvering.

Hvis målet med lærervurdering er at den skal bidra til å øke kvaliteten i skolen, må det tenkes helhetlig – både i utformingen av vurderingen og i gjennomføringen av vurderingsaktivitetene.

Litteraturliste 1: Fullstendig referanseliste til rapporten

Abrami, P.C., Borokhovski, E., Bernard, R.M, Wade, A. C., Tamim, R., Persson, T. Bethel, E. C., Hanz, K and Surkes, M. A. (2010): Issues in conducting and disseminating brief reviews of evidence, *Evidence & Policy*, 6 (3): 371-89.

Aldridge, J., Fraser, B. J., Bell, L. and Dorman, J. (2012): Using a New Learning Environment Questionnaire for Reflection in Teacher Action Research, *Journal of Science Teacher Education* 23(3), 259-290.

Alfaro, M., Jones, D., Holland, G. and Mundy, M-A. (2013): Effect of a teacher incentive program on 4th grade state assessment scores, *Journal of Case Studies in Education* 4, 1-16.

Assessment Reform Group (2006): The role of teachers in the assessment of learning, Nuffield Foundation *http://www.nuffieldfoundation.org/sites/default/files/files/The-role-of-teachers-in-the-assessment-of-learning.pdf*

Barile, J. P., Donohue, D. K., Anthony, E. R., Baker, A. M., Weaver, S. R. and Henrich, C. C. (2012) Teacher-Student Relationship Climate and School Outcomes: Implications for Educational Policy Initiatives, *Journal of Youth and Adolescence* 41, 256-267.

Bastian, K.C., Henry G.T. and Thompson, C.L. (2013): Incorporating Access to More Effective Teachers into Assessments of Educational Resource Equity, *Education Finance and Policy* 8(4), 560-580.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C. and Qi, Y. (2012): An Argument Approach to Observation Protocol Validity, *Educational Assessment* 17(2), 62-87.

Berliner, D. (2013): Problems with Value-Added Evaluations of Teachers? Let Me Count the Ways!, *The Teacher Educator* 48(4), 235-243.

Berliner, D. (2014): Exogenous Variables and Value-Added Assessments: A Fatal Flaw, *Teachers College Record* 116(1), http://www.tcrecord.org/library ID Number 17293, Date Accessed: 1/22/2014.

Black, P. and Wiliam, D. (1998): *Assessment and Classroom Learning. Assessment in Education: Principles, Policy and Practice* 5 (1), 7-74.

Black, P. and Wiliam, D. 2009. Developing the theory of formative assessment, *Educational assessment, evaluation and accountability* 21(1), 5–31.

Boyd, D. H. L., Lankford, H., Loeb, S. and Wyckoff, J. (2013): Measuring Test Measurement Error: A General Approach, *Journal of Educational and Behavioural Statistics*, 38 (6), 629-663.

Brackett, M. A., Palomera, R., Mojsa-Kaja, J., Reyes, M. R. and Salovey, P. (2010): Emotion, regulation ability, burnout and job satisfaction among British secondary-school teachers, *Psychology in the Schools* 47(4), 406-417.

Brantlinger, A., Sherin, M. G. and Linsenmeier, K. A. (2011) - Discussing Discussion: A Video Club in the Service of Math Teachers' National Board Preparation,*Teachers and Teaching: Theory and Practice*, 17 (1), 5-33.

Briggs D.C. and Weeks J.P. (2011): The Persistence of School-Level Value-Added, *Journal of Educational and Behavioral Statistics* 36(5), 616- 637.

Broatch, J. og Lohr, S. (2012): Multidimensional Assessment of Value Added by Teachers to Real-World Outcomes, *Journal of Educational and Behavioral Statistics*, 37 (2) 256-277.

Brown, I. I. and Crumpler, T. (2013): Assessment of Foreign Language Teachers: A Model for Shifting Evaluation Towards Growth and Learning, *The High School Journal*, 138-151.

Bruno, R., Ashby, S., & Manzo, F., IV (2012): *Beyond the classroom: An analysis of a Chicago public school teacher's actual workday*. School of Labor and Employment Relations, University of Illinois at Urbana-Champaign Web site *http://www.ler.illinois.edu/labor/images/Teachers%20Activity- Time%20Study%202012%20(1)-Final.pdf* (downloaded 22.02.14).

Buddin, R. and Zamarro, G. (2009): Teacher qualifications and student achievements in urban elementary schools, *Journal of Urban Economics*, 66, 103-115.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K. and Pianta, R. C. (2013): Effect of Observation Mode on Measures of Secondary Mathematics Teaching, Educational and Psychological Measurement 73(5), 757-783.

Chalmers, I., Hedges, L. and Cooper, H. (2002): A brief history of research synthesis, *Education and the Health Professions*, 25: 12-37.

Christophersen, K-A., Elstad, E. and Turmo, A. (2012) Antecedents of Teachers Fostering Effort within Two Different Management Regimes: An Assessment-Based Accountability Regime and Regime without External Pressure on Results, *International Journal of Education Policy and Leadership* 7(6).

Clayson D. E. (2009): Student Evaluations of Teaching: Are They Related to What Students Learn?: A Meta-Analysis and Review of the Literature, *Journal of Marketing Education*, 31 (1), 16-30

Cohen-Vogel, L. (2011): «Staffing to the Test»: Are Today's School Personnel Practices Evidence Based?, *Educational Evaluation and Policy Analysis* 33(4), 483-505.

Corcoran, S. and Goldhaber D., (2013): Value Added and Its Uses: Where You Stand Depends on Where You Sit, *Education Finance and Policy* 8(3), 418-434.

D'Agostino, J. V. and Powers, S. J. (2009): Predicting Teacher Performance With Test Scores and Grade Point Average: A Meta-Analysis, *American Educational Research Journal* 46 (1), 146-182.

Darling-Hammond, L., Newton, S. P., Wei, R. C. (2013): Developing and assessing beginning teacher effectiveness: the potential of performance assessments, *Educational Assessment, Evaluation and Accountability* 25(3), 179-204.

Delvaux, E. Vanhoof, J. (2013): How may teacher evaluation have an impact on professional development? A multilevel analysis, *Teaching and Teacher Education* 36, 1-11.

Ehlert, M., Koedel, C., Parsons, E. and Podugursky, M., J. (2014): The Sensitivity of Value-Added Estimates to Specification Adjustments: Evidence from School and Teacher level Models in Missouri, *Statistics and Public Policy* 1(1), 19-27.

Everson, K. C., Feinauer, E., Sudweeks, R. R. (2013) Rethinking Teacher Evaluation: A Conversation about Statistical Inferences and Value-Added Models, *Harvard Educational Review* 83(2), 349-370.

Fang, Z., Grant, L. W., Xu, X., Stronge, J. H. and Ward, T. J. (2013) An international comparison investigating the relationship between national culture and student achievement, *Educational Assessment, Evaluation and Accountability* 25, 159-177.

Fevolden, T. og Lillejord, S. (2005): *Kvalitetsarbeid i skolen*. Oslo: Universitetsforlaget.

Finn, A., Schrodt, P. N., Witt, P. L., Elledge, N., Jernberg, K. A. and Larson, L. M. (2009): A Meta-Analytical Review of Teacher Credibility and its Associations with Teacher Behaviors and Student Outcomes, *Communication Education* 58(4), 516-537.

Flores, Maria Assunção (2012): The implementation of a new policy on teacher appraisal in Portugal: how do teachers experience it at school?, *Educ Asse Eval Acc* (2012) 24:351–368

Friedrich, A., Ostermeier, C., Diercks U., Krebs, I. og Stadler, M. (2012): The Team Portfolio: A Support and Evaluation Tool? Findings from a Teacher Professional Development Programme in Germany, *Professional Development in Education* 38(3), 377-394.

Fryer, R. G. (2013): Teacher Incentives and Student Achievement: Evidence from New York City Public Schools, *Journal of Labor Economics* 31(2), 373-407.

Goldhaber, D., Cowan, J. and Walsh, J. (2013a): Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review* 36, 216-228.

Goldhaber, D. Goldschmidt, P. and Tseng, F. (2013b): Teacher value-added at the high-school level: Different models, different answers. *Educational Evaluation and Policy Analysis* 35(2), 220-236.

Goldhaber, D. and Theobald, R. (2013): Managing the Teacher Workforce in Austere Times: The Determinants and Implications of Teacher Layoffs. *Education Finance and Policy* 8(4), 494-527.

Goodman, S. F. and Turner, L. J. (2013): The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program, J*ournal of Labor Economics* 31 (2) 409-420.

Gough, D., Olivier, S. and Thomas, J. (2012): *An introduction to systematic reviews.* London: Sage publications.

Gough, D. and Thomas, J. (2012): Commonality and diversity in reviews, in: Gough, D., Oliver, S. and Thomas, J. (eds): *An Introduction to systematic reviews*. London: SAGE publishing.

Granheim, M. K. og Lundgren, U. P. (1990): *Målstyring og evaluering i norsk skole. Sluttrapport fra EMIL-prosjektet*. NORAS/LOS-i utdanning. Norges råd for anvendt samfunnsforskning.

Grant, L. W., Stronge, J. H. and Xianxuan, X. (2013): A cross-cultural comparative study of teacher effectiveness: Analyses of award-winning teachers in the United States and China, *Educational Assessment and Evaluation*, 25, 251-276.

Graves, G. H., Sulewski, C. A., Dye, H. A., Deveans, T. M., Agras, N. M. and Pearson, M. J. (2009): How Are You Doing? Assessing Effectiveness in Teaching Mathematics, *Primus: Problems, Resources, and Issues in Mathematics Undergraduate Studies* 19 (2), 174-193.

Grissom, J. A., Loeb S. og Master, B. (2013): Effective Instructional Time Use for School Leaders: Longitudinal Evidence From Observations of Principals, *Educational Researcher* 42(8) 433.

Grossman, P., Loeb, S., Cohen, J., Wyckoff, J. (2013): Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores, *American Journal of Education* 119 (3) 445-469.

Gunter, H., Rayner, S., Thomas, H., Fielding, A., Butt, G., and Lance, A. (2005): Teachers, time and work: findings from the Evaluation of the Transforming the School Workforce Pathfinder Project. *School Leadership and Management*, 25(5), 441-454,http://dx.doi.org/10.1080/13634230500340781

Hayward, L. and E. Spencer (2010): «The complexities of change: formative assessment in Scotland.» *Curriculum Journal* 21(2): 161-177.

Hill, H., Kapitula, L. and Umland, K. (2011): A Validity Argument Approach to Evaluating Teacher Value-Added Scores, *American Educational Research Journal* 48(3), 794-831.

Hill, H. C., Charalambous, C. Y, Blazar, D., McGinn, D., Kraft, M.A., Beisiegel, M., Humez, A., Litke, E., and Lynch, K. (2012a): Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation, *Educational Assessment* 17(2-3), 1-19.

Hill, H. C., Charalambos, Y., and Kraft, M. A. (2012b): When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study, *Educational Researcher* 41(2), 56-64.

Hill, H.C., Umland, K., Litke, E., and Kapitula, L.R. (2012c) Teacher quality and quality teaching: Examining the relationship of a teacher assessment to practice. American Journal of Education, vol. 118, No. 4 pp. 489-519

Hill, H. and Grossman, P. (2013): Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, Vol. 83 no. 2, 371-384.

Hinz (2011) Attitudes of German Teachers and Students towards Public Online Ratings of Teaching Quality, *Electronic Journal of Research in Educational Psychology* 9(2), 745-764.

Hopfenbeck, T. N., Tolo, A., Florez, T. og El Masri, Y. (2013): *Balancing Trust and Accountability? The Assessment for Learning Programme in Norway. A Governing Complex Education Systems Case Study.* OECD Education Working Papers, No 97. *http://dx.doi.org/10.1787/5k3txnpqlsnn-en.*

Hopfenbeck, T. N. og Lillejord, S. (2013): Vurdering etter Kunnskapsløftet i: Krumsvik, R. og Säljö, R. (red.): *Praktisk pedagogisk utdanning: En antologi*. Bergen: Fagbokforlaget, 229-253

Howley, M. D., Howley, A., Henning, J. E., Gillam, M. B. and Weade, G. (2013): Intersecting Domains of Assessment Knowledge: School Typologies Based on Interviews with Secondary Teachers, *Educational Assessment* 18(1), 26-48.

Imsen, G. (2009): Lærernes profesjonalitet og nye styringsregimer, Bedre skole 1/2009, 42-49.

Ing, M. and Shih, J. C. (2013) Using Generalizability Theory as a Framework for Informing Measurement Issues in Middle School Settings, *Middle Grades Research Journal* 8(2), 25-39.

Ingle, W. K. (2009): Teacher quality and attrition in a US school district, *Journal of Educational Administration* 47(5), 557-585.

Isoré, M. (2009): Teacher Evaluation: Current Practices in OECD Countries and a Literature Review. OECD Education Working Papers No. 23. OECD Publishing. *http:dx.doi.org/10.1787/223283631428*

Jones, M. D. (2013): Teacher behavior under performance pay incentives, *Economics of Education Review* 37, 148-164.

Jüttner, M., Boone, W., Park, S. and Neuhaus, B. J. (2013): Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK), *Educational Assessment, Evaluation and Accountability* 25(1), 45-67.

Kane T. J., Taylor, E. S., Tyler, J. H and Wooten, A. L. (2011): Identifying Effective Classroom Practices Using Student Achievement Data *Journal of Human Resources* 46(3), 587-613.

Karl, A. T., Yang, Y. and Lohr, S. L. (2013): A Correlated Random Effects Model for Nonignorable Missing Data in Value-Added Assessment of Teacher Effects, *Journal of Educational and Behavioral Statistics* 38(6), 577-603.

Khangura, S., Konnuy, K. Cushman, R., Grimshaw, J. and Moher, D. (2012): Evidence summaries and the evolution of a rapid review approach, *Systematic Reviews*, 1-10.

Kinsler, J. (2012): Beyond Levels and Growth: Estimating Teacher Value-Added and Its Persistence, *Journal of Human Resources* 47(3), 722-753.

Kleinknecht, M. and Schneider, J. (2013): What do teachers think and feel when analyzing videos of themselves and other teachers teaching? *Teaching and Teacher Education* 33,13-23.

Klette, K. (2013): Hva vet vi om god undervisning? Rapport fra klasseromsforskningen, i: Krumsvik, R. og Säljö, R. (red.): *Praktisk pedagogisk utdanning: En antologi*. Bergen: Fagbokforlaget, 173-201.

Kløveager Nielsen, T. (2014): *Teori og praksis I professionsbacheloruddannelserne. Et systematisk review*. Ph.D.-afhandling, Aarhus Universitet, Institut for uddannelse og pædagogik (DPU).

Koedel, C. (2009): An empirical analysis of spillover effects in secondary schools, *Economics of Education Review*, 28, 682-692.

Lefgren, L. and Sims, D. (2012): Using Subject Test Scores Efficiently to Predict Teacher Value Added, *Educational Evaluation and Policy Analysis* 34(1), 109-121.

Lillejord, S. og Hopfenbeck, T. N. (2013): Vurdering og læring i skolen, i: Lillejord, S., Manger, T. og Nordahl, T.: Livet i skolen 2. *Grunnbok i pedagogikk og elevkunnskap: Lærerprofesjonalitet.* Bergen: Fagbokforlaget, 231-260.

Lindblad, S. and Popkewitz, T. eds. (2004): *Educational restructuring: International Perspectives on Traveling Policies*, Greenwich, Conn: Information Age Publ.

Liu., S. and Zhao, D. (2013): Teacher evaluation in China: latest trends and future directions, *Educational Assessment, Evaluation and Accountability* 25, 231-250.

Lockwood, J. R. and McCaffrey, D. F. (2014): Correcting for Test Score Measurement Error in ANCOVA Models for Estimating Treatment Effects, *Journal for Educational and Behavioral Statistics* 39(1), 22-52.

Looney, J. (2011): *Integrating formative and summative assessment: Progress toward a seamless system?* OECD Education Working Paper No. 58
http://search.oecd.org/officialdocuments/displaydocumentpdf/?cote=edu/wkp(2011)4&doclanguage=en

Mariano, L. T., McCaffrey, D. F. and Lockwood J. R. (2010): A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling *Journal of Educational and Behavioral Statistics* 35(3), 253-279.

Maslow, V. J. and Kelley, C. J. (2012): Does Evaluation Advance Teaching Practice? The Effects of Performance Evaluation on Teaching Quality and System Change in Large Diverse High Schools, *Journal of School Leadership* 22(3), 600-632.

Master, B. (2013): Staffing for Success: Linking Teacher Evaluation and School Personnel Management in Practice *Educational Evaluation and Policy Analysis.* Published online before print October 4, 2013, doi: 10.3102/0162373713506552

McCaffrey, D. F., Sass, T. R., Lockwood, J. R. and Mihaly, K. (2009): The intertemporal variability of teacher effect estimates, *Educational Finance and Policy* 4(4), 572-606.

Moreland, J. (2009): Investigating Secondary School Leaders' Perceptions of Performance Management, *Educational Management Administration & Leadership* 37(6), 735-765.

Munthe, E. (2013): Planlegging av undervisning, i: Krumsvik, R. og Säljö, R. (red.): *Praktisk pedagogisk utdanning: En antologi*. Bergen: Fagbokforlaget, 203-227.

Muñoz, M. A., Scoskie, J. R. and French, D. L. (2013): Investigating the «black box» of effective teaching: the relationship between teachers' perception and student achievement in a large urban district, *Educational Assessment, Evaluation and Accountability*, 25 (3), 231-250.

Nehring, J. H. and O'Brien, E. J. (2012): Strong Agents and Weak Systems: University Support for School Level Improvement, *Journal of Educational Change* 13(4) 449-485.

Nelson, J. A. P., Caldarella, P., Adams, M. D. and Shatzer, R. H. (2013) Effect of Peer Praise Notes on Teachers' Perception of School Community and Collegiality, *American Secondary Education* 41 (3), 62-77.

Newton, X. A., Darling-Hammond, L., Haertel, E. and Thomas, E. (2010): Value-added modelling of teacher effectiveness: An exploration of stability across models and contexts, *Education Policy Analysis Archives*, 18 (23), 1-22.

Noblit, G.W. and Hare, R.D. (1988): *Meta-ethnography: synthesizing qualitative studies*. Newbury Park: Sage, 1988.

Nusche, D., Earl, L. Maxwell, W. and Shewbridge, C. (2011) Norway country review
http://www.oecdilibrary.org/docserver/download/9111271e.pdf?expires=1356104025&id=id&accname=ocid57003439&checksum=4B92853145C5AF90C396D69AED7604B7

OECD (1989): *OECD-vurdering av norsk utdanningspolitikk. Norsk rapport til OECD. Ekspertvurdering fra OECD*. Kirke- og undervisningsdepartementet/Kultur- og vitenskapsdepartementet. Oslo: Aschehoug.

OECD (2011): *Education at a Glance. OECD Indicators.* OECD Publishing doi:10.1787/eag-2011-en

OECD (2013): *Synergies for Better Learning: An International Perspective on Evaluation and Assessment in Education* (sammendrag lastet ned 27.02.14):
http://www.oecd.org/edu/school/Synergies%20for%20Better%20Learning_Summary.pdf

Petticrew, M. and Roberts, H. (2006): *Systematic Reviews in the Social Sciences: A Practical Guide*. Oxford: Blackwell.

Pham, H. Q. and Stacey, R. (2012): Classroom Performance Evaluation: Stages and Perspectives for Professional Development of Secondary Teachers in Vietnam, *International Journal of Progressive Education* 8 (2), 6-24.

Philipp, A. and Kunter, M. (2013): How do teachers spend their time? A study on teachers' strategies of selection, optimisation, and compensation over their career cycle, *Teaching and Teacher Education*, 35, 1-12.

Pope, C., Ziebland, S. and Mays, N. (2000): Qualitative research in health care: analysing qualitative data, *British Medical Journal*, 320, 114-6.

Roald, K. (2010): *Kvalitetsvurdering som organisasjonslæring mellom skole og skoleeigar.* PhD-avhandling. Det psykologiske fakultet, Universitetet i Bergen (BORA).

Roby, D. E. (2012): Teacher Leader Human Relations Skills: A Comparative Study, *Education* 132(4), 898-906.

Rothstein, J. (2010): Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement, *The Quarterly Journal of Economics* 125(1), 175-214.

Sanders, W. L. and Horn, S. P. (1998): Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research, *Journal of Personnel Evaluation in Education* 12:3 247-256.

Santelices, M. V. and Taut, S. (2011) Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system: *Assessment in Education: Principles, Policy and practice*, 18:1, 73-93.

Santiago, P., Benavides, F., Danielson, C., Goe, L. and Nusche, D. (2013): *Teacher Evaluation in Chile. Main conclusions*. OECD Reviews of Evaluation and Assessment in Education.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N. and Feng, L. (2012): Value added of teachers in high-poverty schools and lower poverty schools, *Journal of Urban Economics* 72(2-3), 104-122.

Sass, T. R., Semykina, A. and Harris, D. N. (2014): Value-added models and the measurement of teacher productivity, *Economics of Education Review* 38, 9-23.

Schafer, W. D., Lissitz, R. W., Zhu, X. Zhang, Y., Hou, X. and Li, Y. (2012) Evaluating Teachers and Schools Using Student Growth Models, *Practical Assessment, Research & Evaluation* 17(17). Available online: *http://pareonline.net/getvn.asp?v=17&n=17*

Schochet, P. Z. and Chiang, H. S. (2013): What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models? *Journal of Educational and Behavioral Statistics* 38(2), 142-171.

Springer, M. G., Pane, J. F., Le, V-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S. and Stecher, B. (2012): Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives, *Educational Evaluation and Policy Analysis* 34(4), 367-390.

Stobart, G. (2008): *Testing Times. The uses and abuses of assessment*. Abingdon: Routledge.

Stortingsmelding nr. 30 (2003-04): *Kultur for læring.* Utdannings- og forskningsdepartementet.

Sung, T. Y., Chang, E. K., Yu., C. W. og Chang, H. T. (2009): Supporting Teachers' Reflection and Learning through Structured Digital Teaching Portfolios, *Journal of Computer Assisted Learning* 25(4), 375-385.

Taber, K. S., Riga, F., Brindley, S., Winterbottom, M., Finney, J. and Fisher, L. G. (2011): Formative conceptions of assessment: trainee teachers' thinking about assessment issues in English secondary schools, Teacher Development 15(2), 171-186.

Taut, S., Santelices, M. V., Araya, C. and Manzi, J. (2011): Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools, *Studies in Educational Evaluation*, 37, 218-229.

Taut, S., Sun, Y. (2014): The Development and Implementation of a national, standards-based, Multi-method Teacher Performance Assessment System in Chile. *Education Policy Analysis Archivs*, 22 (58) *http://epaa.asu.edu/ojs/*

Thomas, J., Newman, M. and Oliver, S. (2013): Rapid evidence assessment of research to inform social policy: taking stock and moving forward, *Evidence & Policy* vol. 9 no. 1, pp 5-27 *http://dx.doi.org/10.1332/174426413X662572*

Tidsbruksutvalget (2009): Rapport fra Tidsbruksutvalget *http://www.regjeringen.no/upload/KD/Vedlegg/Grunnskole/Tidsbrukutvalget/Rapport_Tidsbrukutvalget.pdf* (lastet ned 22.02.14).

Tinoca L. and Oliveira, I. (2013): Formative assessment of teachers in the context of an online learning environment, *Teachers and Teaching: Theory and Practice* 19(2), 214-227.

Tolo, A. og Lillejord, S. (2009): For en offensiv kunnskaps- og kompetansepolitikk. UNIO *http://www.unio.no/kunder/unio/mm2011.nsf/lupgraphics/kunnskaps_kompetansepolitik.pdf/$file/kunnskaps_kompetansepolitik.pdf* (lastet ned 27.02.14).

Tolo, A. (2011): *Hvordan blir lærerkompetanse konstruert? En kvalitativ studie av PPU-studenters kunnskapsutvikling.* Avhandling for graden PhD. Det psykologiske fakultet, Universitetet i Bergen.

Tornero, B., Taut, S. (2010): A mandatory, high-stakes National Teacher Evaluation System: Perceptions and attributions of teachers who actively refuse to participate, *Studies in educational evaluation* 36, 132-142.

Tuytens, M. and Devos, G. (2011): Stimulating Professional Learning through Teacher Evaluation: An Impossible Task for the School Leader?, *Teaching and Teacher Education: An International Journal of Research and Studies* 27(5) 891-899.

Uline, C. L., Miller, D. M., Tschannen-Moran, M. (1998): School Effectiveness: The underlying dimensions, *Educational Administration Quarterly* Vol 34, no 4, 462-483.

Van Diggelen, M., den Brok, P. and Beijaard, D. (2013): Teachers' use of a self-assessment procedure: the role of criteria, standards, feedback and reflection, *Teachers and Teaching: Theory and Practice* 19(2), 115-134.

Verberg, C. P. M., Tigelaar, D. E. H, and Verloop, N. (2013): Teacher learning through participation in a negotiated assessment procedure, *Teachers and Teaching: theory and practice*, 19 (2), 172-187.

Wang, A. H.; Walters, A. M. and Thum, Y. M. (2013): Identifying highly effective urban schools: comparing two measures of school success, *The International Journal of Educational Management* 27(5), 517-540.

Winters, M. A. and Cowen, J. M. (2013): Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies, *Educational Researcher* 42(6), 330-337.

Yuan, K., Le, V-N., McCaffrey, D. F., Marsh, J. A., Hamilton, L. S., Stecher, B. M., Springer, M. G. (2013): Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies, *Educational Evaluation and Policy Analyses* 35(1), 3-22.

Zhang, X. F. and Ng, H. M. (2011): A Case Study of Teacher Appraisal in Shanghai, China: In Relation to Teacher Professional Development, *Asia Pacific Education Review* 12, 569-580.

Litteraturliste 2:

Inkluderte artikler i den systematiske kunnskapsoversikten

Aldridge, J., Fraser, B. J., Bell, L. and Dorman, J. (2012): Using a New Learning Environment Questionnaire for Reflection in Teacher Action Research, *Journal of Science Teacher Education* 23(3), 259-290.

Alfaro, M., Jones, D., Holland, G. and Mundy, M-A. (2013): Effect of a teacher incentive program on 4th grade state assessment scores, *Journal of Case Studies in Education* 4, 1-16.

Barile, J. P., Donohue, D. K., Anthony, E. R., Baker, A. M., Weaver, S. R. and Henrich, C. C. (2012) Teacher-Student Relationship Climate and School Outcomes: Implications for Educational Policy Initiatives, *Journal of Youth and Adolescence* 41, 256-267.

Bastian, K.C., Henry G.T. and Thompson, C.L. (2013): Incorporating Access to More Effective Teachers into Assessments of Educational Resource Equity, *Education Finance and Policy* 8(4), 560-580.

Bell, C. A., Gitomer, D. H., McCaffrey, D. F., Hamre, B. K., Pianta, R. C. and Qi, Y. (2012): An Argument Approach to Observation Protocol Validity, *Educational Assessment* 17(2), 62-87.

Berliner, D. (2013): Problems with Value-Added Evaluations of Teachers? Let Me Count the Ways!, *The Teacher Educator* 48(4), 235-243.

Berliner, D. (2014): Exogenous Variables and Value-Added Assessments: *A Fatal Flaw, Teachers College Record* 116(1), *http://www.tcrecord.org/library ID Number 17293, Date Accessed: 1/22/2014 8:28:13 AM.*

Boyd, D. H. L., Loeb, S. and Wyckoff, J. (2013): Measuring Test Measurement Error: A General Approach, *Journal of Educational and Behavioural Statistics,* 38 (6), 629-663.

Brackett, M. A., Palomera, R., Mojsa-Kaja, J., Reyes, M. R. and Salovey, P. (2010): Emotion, regulation ability, burnout and job satisfaction among British secondary-school teachers, *Psychology in the Schools* 47(4), 406-417.

Brantlinger, A., Sherin, M. G. and Linsenmeier, K. A. (2011) - Discussing Discussion: A Video Club in the Service of Math Teachers' National Board Preparation, *Teachers and Teaching: Theory and Practice*, 17 (1), 5-33.

Briggs D.C. and Weeks J.P. (2011): The Persistence of School-Level Value-Added, *Journal of Educational and Behavioral Statistics* 36(5), 616- 637.

Broatch, J. og Lohr, S. (2012): Multidimensional Assessment of Value Added by Teachers to Real-World Outcomes, *Journal of Educational and Behavioral Statistics*, 37 (2) 256-277.

Brown, I. I. and Crumpler, T. (2013): Assessment of Foreign Language Teachers: A Model for Shifting Evaluation Towards Growth and Learning, *The High School Journal*, 138-151.

Buddin, R. and Zamarro, G. (2009): Teacher qualifications and student achievements in urban elementary schools, *Journal of Urban Economics*, 66, 103-115.

Casabianca, J. M., McCaffrey, D. F., Gitomer, D. H., Bell, C. A., Hamre, B. K. and Pianta, R. C. (2013): Effect of Observation Mode on Measures of Secondary Mathematics Teaching, *Educational and Psychological Measurement* 73(5), 757-783.

Christophersen, K-A., Elstad, E. and Turmo, A. (2012) Antecedents of Teachers Fostering Effort within Two Different Management Regimes: An Assessment-Based Accountability Regime and Regime without External Pressure on Results *International Journal of Education Policy and Leadership* 7(6).

Cohen-Vogel, L. (2011): «Staffing to the Test»: Are Today's School Personnel Practices Evidence Based? Educational Evaluation and Policy Analysis 33(4), 483-505.

Corcoran, S. and Goldhaber D., (2013): Value Added and Its Uses: Where You Stand Depends on Where You Sit, *Education Finance and Policy* 8(3), 418-434.

D'Agostino, J. V. and Powers, S. J. (2009): Predicting Teacher Performance With Test Scores and Grade Point Average: A Meta-Analysis, *American Educational Research Journal* 46 (1), 146-182.

Darling-Hammond, L., Newton, S. P., Wei, R. C. (2013): Developing and assessing beginning teacher effectiveness: the potential of performance assessments, *Educational Assessment, Evaluation and Accountability* 25(3), 179-204.

Delvaux, E. Vanhoof, J. (2013): How may teacher evaluation have an impact on professional development? A multilevel analysis, *Teaching and Teacher Education* 36, 1-11.

Everson, K. C., Feinauer, E., Sudweeks, R. R. (2013) Rethinking Teacher Evaluation: A Conversation about Statistical Inferences and Value-Added Models, *Harvard Educational Review* 83(2), 349-370.

Fang, Z., Grant, L. W., Xu, X., Stronge, J. H. and Ward, T. J. (2013) An international comparison investigating the relationship between national culture and student achievement *Educational Assessment, Evaluation and Accountability* 25, 159-177.

Finn, A., Schrodt, P. N., Witt, P. L., Elledge, N., Jernberg, K. A. and Larson, L. M. (2009): A Meta-Analytical Review of Teacher Credibility and its Associations with Teacher Behaviors and Student Outcomes *Communication Education* 58(4), 516-537.

Flores, Maria Assunção (2012): The implementation of a new policy on teacher appraisal in Portugal: how do teachers experience it at school? Educ Asse Eval Acc (2012) 24:351–368

Friedrich, A., Ostermeier, C., Diercks U., Krebs, I. og Stadler, M. (2012): The Team Portfolio: A Support and Evaluation Tool? Findings from a Teacher Professional Development Programme in Germany, *Professional Development in Education* 38(3), 377-394.

Fryer R. G. (2013): Teacher Incentives and Student Achievement: Evidence from New York City Public Schools, *Journal of Labor Economics* 31(2), 373-407.

Goldhaber, D., Cowan, J. and Walsh, J. (2013a): Is a good elementary teacher always good? Assessing teacher performance estimates across subjects. *Economics of Education Review* 36, 216-228.

Goldhaber, D. Goldschmidt, P. and Tseng, F. (2013b): Teacher value-added at the high-school level: Different models, different answers. *Educational Evaluation and Policy Analysis* 35(2), 220-236.

Goldhaber, D. and Theobald, R. (2013): Managing the Teacher Workforce in Austere Times: The Determinants and Implications of Teacher Layoffs. *Education Finance and Policy* 8(4), 494-527.

Goodman S. F. and Turner L. J. (2013): The Design of Teacher Incentive Pay and Educational Outcomes: Evidence from the New York City Bonus Program. *Journal of Labor Economics* 3(2), 409-420.

Graves, G. H., Sulewski, C. A., Dye, H. A., Deveans, T. M., Agras, N. M. and Pearson, M. J. (2009): How Are You Doing? Assessing Effectiveness in Teaching Mathematics, *Primus: Problems, Resources, and Issues in Mathematics Undergraduate Studies* 19(2), 174-193.

Grissom, J. A., Loeb S. og Master, B. (2013): Effective Instructional Time Use for School Leaders: Longitudinal Evidence From Observations of Principals, *Educational Researcher* 42(8) 433.

Grossman, P., Loeb, S., Cohen, J., Wyckoff, J. (2013): Measure for Measure: The Relationship between Measures of Instructional Practice in Middle School English Language Arts and Teachers' Value-Added Scores, *American Journal of Education* 119(3) 445-469.

Hill, H., Kapitula, L. and Umland, K. (2011): A Validity Argument Approach to Evaluating Teacher Value-Added Scores, *American Educational Research Journal* 48(3), 794-831.

Hill, H. C., Charalambous, C. Y, Blazar, D., McGinn, D., Kraft, M.A., Beisiegel, M., Humez, A., Litke, E., and Lynch, K. (2012a): Validating Arguments for Observational Instruments: Attending to Multiple Sources of Variation, *Educational Assessment* 17(2-3), 1-19.

Hill, H. C., Charalambos, Y., and Kraft, M. A. (2012b): When Rater Reliability Is Not Enough: Teacher Observation Systems and a Case for the Generalizability Study, *Educational Researcher* 41(2), 56-64.

Hill, H. C., Umland K., Litke, E. and Kapitula L. R. (2012c): Teacher Quality and Quality Teaching: Examining the Relationship of a Teacher Assessment to Practice. *American Journal of Education* 118(4), 489-519.

Hill, H. and Grossman, P. (2013): Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, Vol. 83 no. 2, 371-384.

Hinz (2011) Attitudes of German Teachers and Students towards Public Online Ratings of Teaching Quality, E*lectronic Journal of Research in Educational Psychology* 9(2), 745-764.

Howley, M. D., Howley, A., Henning, J. E., Gillam, M. B. and Weade, G. (2013): Intersecting Domains of Assessment Knowledge: School Typologies Based on Interviews with Secondary Teachers, *Educational Assessment* 18(1), 26-48.

Ing, M. and Shih, J. C. (2013) Using Generalizability Theory as a Framework for Informing Measurement Issues in Middle School Settings, *Middle Grades Research Journal* 8(2), 25-39.

Jones, M. D. (2013): Teacher behavior under performance pay incentives, *Economics of Education Review* 37, 148-164.

Jüttner, M., Boone, W., Park, S. and Neuhaus, B. J. (2013): Development and use of a test instrument to measure biology teachers' content knowledge (CK) and pedagogical content knowledge (PCK), *Educational Assessment, Evaluation and Accountability* 25(1), 45-67.

Kane T. J., Taylor, E. S., Tyler, J. H and Wooten, A. L. (2011): Identifying Effective Classroom Practices Using Student Achievement Data, *Journal of Human Resources* 46(3), 587-613.

Kinsler, J. (2012): Beyond Levels and Growth: Estimating Teacher Value-Added and Its Persistence, *Journal of Human Resources* 47(3), 722-753.

Kleinknecht, M. and Schneider, J. (2013): What do teachers think and feel when analyzing videos of themselves and other teachers teaching?, *Teaching and Teacher Education* 33, 13-23.

Koedel, C. (2009): An empirical analysis of spillover effects in secondary schools, *Economics of Education Review*, 28, 682-692.

Liu., S. and Zhao, D. (2013): Teacher evaluation in China: latest trends and future directions, *Educational Assessment, Evaluation and Accountability* 25, 231-250.

Mariano, L. T., McCaffrey, D. F. and Lockwood J. R. (2010): A Model for Teacher Effects From Longitudinal Data Without Assuming Vertical Scaling *Journal of Educational and Behavioral Statistics* 35(3), 253-279.

Maslow, V. J. and Kelley, C. J. (2012): Does Evaluation Advance Teaching Practice? The Effects of Performance Evaluation on Teaching Quality and System Change in Large Diverse High Schools *Journal of School Leadership* 22(3), 600-632.

Master, B. (2013): Staffing for Success: Linking Teacher Evaluation and School Personnel Management in Practice Educational *Evaluation and Policy Analysis.* Published online before print October 4, 2013, doi: 10.3102/0162373713506552

McCaffrey, D. F., Sass, T. R., Lockwood, J. R. and Mihaly, K. (2009): The intertemporal variability of teacher effect estimates, *Educational Finance and Policy* 4(4), 572-606.

Moreland, J. (2009): Investigating Secondary School Leaders' Perceptions of Performance Management, *Educational Management Administration & Leadership* 37(6), 735-765.

Nehring, J. H. and O'Brien, E. J. (2012): Strong Agents and Weak Systems: University Support for School Level Improvement, *Journal of Educational Change* 13(4), 449-485.

Nelson, J. A. P., Caldarella, P., Adams, M. D. and Shatzer, R. H. (2013) Effect of Peer Praise Notes on Teachers' Perception of School Community and Collegiality, *American Secondary Education* 41(3), 62-77.

Newton, X. A., Darling-Hammond, L., Haertel, E. and Thomas, E. (2010): Value-added modelling of teacher effectiveness: An exploration of stability across models and contexts, *Education Policy Analysis Archives*, 18(23), 1-22.

Pham, H. Q. and Stacey, R. (2012): Classroom Performance Evaluation: Stages and Perspectives for Professional Development of Secondary Teachers in Vietnam, *International Journal of Progressive Education* 8 (2), 6-24.

Roby, D. E. (2012): Teacher Leader Human Relations Skills: A Comparative Study, *Education* 132(4), 898-906.

Rothstein, J. (2010): Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement, *The Quarterly Journal of Economics* 125(1), 175-214.

Santelices, M. V. and Taut, S. (2011) Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system: *Assessment in Education: Principles, Policy and practice*, 18(1), 73-93.

Sass, T. R., Hannaway, J., Xu, Z., Figlio, D. N. and Feng, L. (2012): Value added of teachers in high-poverty schools and lower poverty schools, *Journal of Urban Economics* 72(2-3), 104-122.

Sass, T. R., Semykina, A. and Harris, D. N. (2014): Value-added models and the measurement of teacher productivity, *Economics of Education Review* 38, 9-23.

Schafer, W. D., Lissitz, R. W., Zhu, X. Zhang, Y., Hou, X. and Li, Y. (2012) Evaluating Teachers and Schools Using Student Growth Models, *Practical Assessment, Research & Evaluation* 17(17). Available online: *http://pareonline.net/getvn.asp?v=17&n=17*

Schochet, P. Z. and Chiang, H. S. (2013): What Are Error Rates for Classifying Teacher and School Performance Using Value-Added Models? *Journal of Educational and Behavioral Statistics* 38(2), 142-171.

Springer, M. G., Pane, J. F., Le, V-N., McCaffrey, D. F., Burns, S. F., Hamilton, L. S. and Stecher, B. (2012): Team Pay for Performance: Experimental Evidence from the Round Rock Pilot Project on Team Incentives, *Educational Evaluation and Policy Analysis* 34(4), 367-390.

Sung, T. Y., Chang, E. K., Yu., C. W. og Chang, H. T. (2009): Supporting Teachers' Reflection and Learning through Structured Digital Teaching Portfolios, *Journal of Computer Assisted Learning* 25(4), 375-385.

Taber, K. S., Riga, F., Brindley, S., Winterbottom, M., Finney, J. and Fisher, L. G. (2011): Formative conceptions of assessment: trainee teachers' thinking about assessment issues in English secondary schools, *Teacher Development* 15(2), 171-186.

Taut, S., Santelices, M. V., Araya, C. and Manzi, J. (2011): Perceived effects and uses of the national teacher evaluation system in Chilean elementary schools, *Studies in Educational Evaluation*, 37, 218-229.

Taut, S., Sun, Y. (2014): The Development and Implementation of a national, standars-based, Multi-method Teacher Performance Assessment System in Chile. *Education Policy Analysis Archivs*, 22(58) http://epaa.asu.edu/ojs/

Tinoca L. and Oliveira, I. (2013): Formative assessment of teachers in the context of an online learning environment *Teachers and Teaching: Theory and Practice* 19(2), 214-227.

Tornero, B., Taut, S. (2010): A mandatory, high-stakes National Teacher Evaluation System: Perceptions and attributions of teachers who actively refuse to participate, *Studies in educational evaluation* 36, 132-142.

Tuytens, M. and Devos, G. (2011): Stimulating Professional Learning through Teacher Evaluation: An Impossible Task for the School Leader? *Teaching and Teacher Education: An International Journal of Research and Studies* 27(5) 891-899.

Van Diggelen, M., den Brok, P. and Beijaard, D. (2013): Teachers' use of a self-assessment procedure: the role of criteria, standards, feedback and reflection, *Teachers and Teaching: Theory and Practice* 19(2), 115-134.

Verberg, C. P. M., Tigelaar, D. E. H, and Verloop, N. (2013): Teacher learning through participation in a negotiated assessment procedure, *Teachers and Teaching: theory and practice*, 19(2), 172-187.

Wang, A. H.; Walters, A. M. and Thum, Y. M. (2013): Identifying highly effective urban schools: comparing two measures of school success, *The International Journal of Educational Management* 27(5), 517-540.

Winters, M. A. and Cowen, J. M. (2013): Who would stay, who would be dismissed? An empirical consideration of value-added teacher retention policies, *Educational Researcher* 42(6), 330-337.

Yuan, K., Vi-Nhuan, McCaffrey, D. F., Marsh J. A., Hamilton, L. S., Stecher, B. M. and Springer M. G. (2012): Incentive pay programs do not affect teacher motivation or reported practices: Results from three randomized studies. Educational Evaluation and Policy Analysis 35(1), 3-22.

Zhang, X. F. and Ng, H. M. (2011): A Case Study of Teacher Appraisal in Shanghai, China: In Relation to Teacher Professional Development, *Asia Pacific Education Review* 12, 569-580.

Vedlegg

I. Oppdragsbrev fra Kunnskapsdepartementet

DET KONGELIGE
KUNNSKAPSDEPARTEMENT

Kunnskapssenter for utdanning
Postboks 2700 St. Hanshaugen
0131 OSLO

Deres ref	Vår ref	Dato
	13/4522-	10.10.13

Bestilling av systematisk kunnskapsoversikt om lærervurdering

Kunnskapsdepartementet ønsker å inngå en avtale med Kunnskapssenter for utdanning (KSU) om å lage en systematisk kunnskapsoversikt om lærervurdering. Kunnskapsoversikten skal belyse og vurdere hvilke former for lærervurdering som fremmer prosesskvalitet og resultatkvalitet i skolen.

Oppdraget

Flere land har de senere årene innført ulike former for lærervurdering, men begrepet lærervurdering er ikke innarbeidet i norsk utdanningspolitikk. Selv om mange har en oppfatning av norske læreres kompetanse og dyktighet, er det sannsynligvis store variasjoner i hvor presise og nyttige tilbakemeldinger lærerne får fra elever, foresatte, fra skoleeier og skoleledelse. Ett av hovedfunnene fra den internasjonale TALIS-undersøkelsen var at tilbakemeldingskulturen i norsk skole er svak sammenliknet med andre land.

Kunnskapsdepartementet ønsker forskningsbasert kunnskap om hvilke former for lærervurdering som kan ha positiv effekt på skolens kvalitet. Etter departementets vurdering er det i denne sammenheng naturlig å skille mellom to ulike former for kvalitet:

- *Prosesskvalitet* defineres som undervisningens og læringens innhold, didaktikk, metodisk tilnærming, og lærernes profesjonsutvikling og anvendelse av egen kompetanse. Læringsmiljøet, elevenes samspill seg imellom og foreldreinvolvering inngår også i prosesskvalitetsbegrepet.
- *Resultatkvalitet* kjennetegner det man oppnår med det pedagogiske arbeidet – først og fremst elevenes utbytte av opplæringen i form av kunnskaper, ferdigheter og holdninger.

Postadresse
Postboks 8119 Dep
NO-0032 Oslo
postmottak@kd.dep.no

Kontoradresse
Akersg. 44

http://www.kd.dep.no/

Telefon*
22 24 90 90*
Org no.
872 417 842

Avdeling for analyse, internasjonalt arbeid og kompetansepolitikk

Saksbehandler
Marie Wien Fjell
22247488

Kunnskapsoversikten skal omfatte alle former for lærervurdering, for eksempel testresultater / eksamenskarakterer, klasseromsobservasjoner, lærerporteføljer, elevvurdering og egenvurdering. Både kvalitative og kvantitative studier skal inkluderes. Kunnskapsdepartementet ønsker at ulike formål med lærervurdering inkluderes i kunnskapsoversikten

Gjennomføring og fremdrift

Departementet ber Kunnskapssenter for utdanning utarbeide en prosjektplan for hvordan oppdraget skal gjennomføres og at prosjektplanen oversendes departementet for kommentarer innen 25.10.2013.

Arbeidet med kunnskapsoversikten skal sees i nær sammenheng med oppdraget til arbeidsgruppen for lærervurdering i GNIST-partnerskapet.

Det skal avholdes møter mellom departementet som oppdragsgiver og KSU som oppdragstaker:

- Om prosjektplanen og avtalen om oppdraget
- Møte i etterkant av litteratursøk
- Midtveisrapportering (gi en indikasjon av funnene)
- Om utkast til sluttrapport

Departementet ser for seg at arbeidet starter opp 1. november 2013 og avsluttes 1. april 2014.

Leveranser

KSU skal ved oppdragets slutt - 1. april 2014 – overlevere Kunnskapsdepartementet en sluttrapport. Offentliggjøringen av sluttrapporten foretas av KSU etter avtale med departementet.

Finansiering

Oppdraget har en ramme på inntil 500 000 NOK ink mva. Beløpet skal dekke *merkostnader* utover Kunnskapssenterets ordinære ramme. Med merkostnader menes honorar til ekstern forskergruppe, reiser, teknisk bistand, trykking og spredning av rapport m.m. KSU leverer Kunnskapsdepartementet et detaljert budsjett i tilknytning til prosjektplanen. KSU er oppdragsgiver for ekstern forskergruppe.

Med hilsen

Morten Rosenkvist (e.f.)
avdelingsdirektør

Marie Wien Fjell
seniorrådgiver

Dokumentet er elektronisk signert og har derfor ikke håndskrevne signaturer.

II. Prosjektplan

Notat

Emne:	**Prosjekt Lærervurdering**
Til:	**Kunnskapsdepartementet**
Kopi:	
Fra:	**Kunnskapssenter for utdanning**
Saksbehandler:	**Peder Fischer-Griffiths**
Vår referanse:	**13/9566**
Dato:	**05.11.2013**
Oppdragsgiver:	**Kunnskapsdepartementet**
Oppdragstaker:	**Kunnskapssenter for utdanning**
Prosjektleder:	**Sølvi Lillejord**
Tidsplan:	**Prosjektet er planlagt ferdigstilt 01.03. 2014**

Premisser for kunnskapsoversikten

Kunnskapsdepartementet (oppdragsgiver) har bestilt en systematisk kunnskapsoversikt som viser hvilke former for lærervurdering som kan ha positiv effekt på skolens kvalitet. Prosjektet skal ta hensyn til dette og samle studier som både rapporterer om forhold som gjelder prosesskvalitet og resultatkvalitet. Både kvantitativ og kvalitativ forskning skal inkluderes, og prosjektet skal omfatte både formativ og summativ vurdering.

Oppdraget er gitt til Kunnskapssenter for utdanning (KSU). Følgende krav er stilt til oppdraget:

- kunnskapsoversikten skal inkludere forskningslitteratur om alle former for lærervurdering
- ulike formål med lærervurdering som er omtalt i forskningslitteraturen skal inkluderes i kunnskapsoversikten

Kunnskapsdepartementet ønsker svar på følgende spørsmål:

- *Hvilke former for lærervurdering fremmer prosesskvalitet i skolen?*
- *Hvilke former for lærervurdering fremmer resultatkvalitet i skolen?*

Med bakgrunn i dette er den systematiske kunnskapsoversiktens scope slik formulert:

Hvilke former for lærervurdering kan ha positiv innvirkning på skolens kvalitet?

Scopet kan bli justert når resultatene fra de innledende søkene foreligger.

Bakgrunn

Allerede i 1988 påpekte OECD at norske myndigheter burde vurdere å innføre et system for kvalitetsvurdering som kunne gi nødvendig informasjon for utforming av utdanningspolitikken. Etter flere stortingsmeldinger og utredninger, og med utgangspunkt i arbeidene fra Kvalitetsutvalget, (NOU 2002:10): *Førsteklasses fra første klasse* og (NOU 2003:16): *I første rekke*, vedtok Stortinget å etablere et nasjonalt system for kvalitetsvurdering (NKVS). Da OECD i 2011 analyserte det nasjonale kvalitetsvurderingssystemet, kom organisasjonen med en rekke forslag til forbedringer (Nusche mfl. 2011), og foreslo blant annet at Norge bør integrere lærervurdering (*teacher appraisal*) i det nasjonale kvalitetsrammeverket, med en profil som sikrer lærere tilbakemeldinger som støtter deres profesjonsutvikling, hjelper dem til karriereutvikling samt bidrar til utvikling av skolen som organisasjon (Isoré, 2009).

Mange land i OECD-området har lenge hatt systemer for lærervurdering, og det finnes mye erfaring og en stor mengde forskningsrapporter og vitenskapelige artikler som drøfter problemstillinger med relevans for temaet. The Tennessee Value-Added Assessment System (Sanders og Horn, 1998) har siden 1993 brukt mixed method og longitudinell, multivariat analyse for å anslå effekten av skole, klassestørrelse og lærer. Hensikten med arbeidet er å fremskaffe summativ informasjon om skolens, systemets, eller lærerens bidrag til elevenes læringsutbytte. Diskusjoner om «school effectiveness»- retningen som har vært opptatt av å måle skolens tilleggsverdi (value-added) (Uline mfl., 1998), har vist at det er når summativ informasjon brukes formativt at den kan bidra til læring på individ- og organisasjons- eller systemnivå. Skolen er en kompleks organisasjon og undervisning en mangefasettert aktivitet som det er vanskelig å vurdere på en slik måte at den enkelte lærer, grupper av lærere eller hele skoleorganisasjonen kan lære av tilbakemeldingene hvordan undervisningen kan gjøres bedre. Hill og Grossman (2013) viser for eksempel hvilke problemer man kan få dersom observasjonsskjema og rapporter er for generelle og ikke tilpasset utdanningsnivå, elevgruppe og særtrekk i det enkelte undervisningsfag. Litteraturen de har gått gjennom viser at det er et problem hvis tilbakemeldinger er for allmenne, at det ikke alltid er rektor som er best egnet til å vurdere alle lærerne på skolen, at det er viktig å skjelne mellom hva som er felles profesjonskompetanser og hva som er fagspesifikk kunnskap (å øke elevenes forståelse i matematikk fordrer andre metoder enn i historie), at gode observasjonsskjema er presise, at det er viktig å finne riktig detaljeringsnivå, at det er en fordel om flere observerer samme lærer og snakker sammen om det de har observert (validerer), at observasjonen må foregå over tid og at ikke alle lærere trenger like hyppige tilbakemeldinger.

Arbeidsprosess: Søkestrategi, kvalitets- og relevansvurdering

Arbeidet med den systematiske kunnskapsoversikten er etablert som et prosjekt, med Kunnskapssenterets direktør som prosjektleder. Arbeidsfordelingen mellom aktørene som tar del i dette prosjektet er slik:

- Kunnskapsdepartementet er oppdragsgiver og deltar i oppsatte kontaktmøter (5. november 2013; 28. november 2013; 17. januar 2014 og 17. februar 2014).
- Kunnskapssenter for utdanning er oppdragstaker og deltar i alle møter. Arbeidet med kunnskapsoversikten er organisert som et prosjekt med Kunnskapssenterets direktør som prosjektleder.
- EPPI-senteret skal bistå Kunnskapssenter for utdanning med søkestrategi. Det innebærer leasing av software fra EPPI, tre dagers opplæring i London, anslagsvis 2-3 uker innledende søk som gjennomføres i samarbeid mellom EPPI og KSU. EPPI vil også bistå med rådgiving med hensyn til hvilke søkeord som skal/kan benyttes for å oppnå ønskede treff i tidsskriftbasene. Råd om søkemetoder for å redusere antall treff (duplikater, ikke-relevant litteratur). Råd om søkemetoder for alternative søk (i og med at oppdraget er så bredt er det snakk om å søke i store mengder litteratur i tillegg til den som spesifikt omhandler lærervurdering, særlig for å undersøke hva forskningen viser har innvirkning på kvalitetsutvikling i skolen).
- En arbeidsgruppe med deltakere fra GNIST-partnerskapet deltar i møte etter litteratursøk (28. november 2013, møte for midtveisrapportering 17. januar 2014 samt møte i forkant av publisering 17. februar 2014)
- En gruppe forskere følger prosjektet og skal bistå med formulering (og eventuell reformulering) av forskningsspørsmål (scope) samt kvalitetssikring av rapporten underveis. Kunnskapssenteret er oppdragsgiver for forskergruppen. I og med at oppdraget er vidt, er forskergruppen satt sammen slik at den *til sammen* skal dekke bredden av problemstillinger som oppdraget skal belyse. Det vil si at noen av forskerne vil bli bedt om å se særlig på litteraturen om vurdering, noen vil bli bedt om å se på i hvilken grad profesjonsutviklingsperspektivet er ivaretatt, noen vil ta seg av skoleutviklings- (organisasjons-) perspektivet og skoleledelse. Helheten vil bli i varetatt ved at hele gruppen leser midtveisrapport og utkast til sluttrapport. Arbeidsfordelingen vil bli tydeligere når vi ser hvilken kvalitet det er på litteraturen og hvor mange artikler vi klarer å fange opp gjennom systematiske søk med relevans for prosjektets scope.

Følgende fem forskere inngår i prosjektets forskergruppe:

Karen Hammerness, Bard College. NY. Lærerutdanning som spesialfelt. Bakgrunn fra Stanford University. Fulbrightstipend og forskningsopphold ved UiO. Arbeider med observasjon av undervisning og har skrevet artikler som sammenligner norsk lærerutdanning med amerikansk.
Therese N. Hopfenbeck, OUCEA, Oxford University, vurdering som spesialfelt. Hovedforfatter på OECD-rapporten Governing Complex Education Systems (2013). OUCEA gjennomfører, på oppdrag for Kunnskapssenteret, en review om vurdering som vil være til stor nytte for prosjekt lærervurdering.
Trond E. Hauge, professor em (UiO). Har arbeidet med skolen som organisasjon, evalueringsproblematikk og spørsmål om styring og ledelse. Han var aktiv i utforming av søknad og etablering av ProTed, Senter for fremragende utdanning (UiO/UiT), og har erfaring fra arbeid med systematiske kunnskapsoversikter.

Astrid Tolo, UiB, førsteamanuensis ved Institutt for pedagogikk. Arbeider med lærerutdanning (PPU) og skolelederutdanning (rektorskolen). Avhandling (2011) om lærerkompetanse. Medforfatter på rapporten Governing Complex Education Systems.
Jens Chr. Smeby HiOA, professor og stedfortredende senterleder ved Senter for profesjonsstudier. Arbeider med profesjonskvalifisering i utdanning og arbeidsliv og kjennetegn ved profesjonskunnskap.

Det planlegges et møte for forskergruppen på Universitetet i Oxford den første uken i februar 2014. På denne samlingen vil også OUCEA presentere arbeidet med den systematiske litteraturgjennomgangen om tema vurdering. Den innsamlede forskningslitteraturen kvalitets- og relevansvurderes av Kunnskapssenter for utdanning, forskergruppen og arbeidsgruppen i GNIST. Et kvalitetskriterium i syntesearbeid er transparens. Dette kriteriet vil også bli fulgt og overholdt i dette prosjektet.

Protokoll

Det opprettes en protokoll som spesifikt definerer søkestrategien for hvordan man skal skaffe seg oversikt over forskningsfeltet. Protokollen inneholder også en beskrivelse av metodevalg, samt refleksjoner omkring styrker og svakheter ved innsamlingsopplegget. Protokollen skal være tilgjengelig for interessenter ved behov.

Tidsplan

Rapport skal leveres 1. april 2014.

Det skal avholdes minimum fire møter mellom oppdragsgiver og oppdragstaker:

1. Oppstartsmøte 5. november 2013, for å avklare innretning, søkestrategi m.m.
2. Møte i etterkant av litteratursøk 28. november 2013
3. Midtveisrapportering 17. januar 2014
4. Rapport presenteres før publisering 17. februar 2014

Budsjett

Det er bevilget kr. 500.000 til prosjektet. Budsjett ligger ved.

Litteratur

Hill, H. and Grossman, P. (2013): Learning from Teacher Observations: Challenges and Opportunities Posed by New Teacher Evaluation Systems. *Harvard Educational Review*, Vol. 83 no. 2, 371-384.

Isoré, M. (2009): Teacher Evaluation: Current Practices in OECD Countries and a Literature Review. OECD Education Working Papers No. 23. OECD Publishing. *http:dx.doi.org/10.1787/223283631428*

Nusche, D., Earl, L. Maxwell, W. and Shewbridge, C. (2011) Norway country review *http://www.oecdilibrary.org/docserver/download/9111271e.pdf?expires=1356104025&id=id&accname=ocid57003439&checksum=4B92853145C5AF90C396D69AED7604B7*

Sanders, W. L. and Horn, S. P. (1998): Research Findings from the Tennessee Value-Added Assessment System (TVAAS) Database: Implications for Educational Evaluation and Research, *Journal of Personnel Evaluation in Education* 12:3 247-256.

Uline, C. L., Miller, D. M., Tschannen-Moran, M. (1998): School Effectiveness: The underlying dimensions, *Educational Administration Quarterly* Vol 34, no 4, 462-483.

III. Inklusjons- og eksklusjonskriterier med begrunnelse.

EKSKLUDER på dato: Ekskluder studier publisert før 1. januar 2009.

Begrunnelse: Det fokuseres på nyere litteratur for å begrense antall referanser slik at tidsfristen for levering av kunnskapsoversikten kan overholdes.

EKSKLUDER på språk: Ekskluder studier som ikke er skrevet på engelsk, norsk, svensk eller dansk.

Begrunnelse: Oversettelse av vitenskapelige artikler er svært tidkrevende, og er ikke mulig innenfor prosjektets tidsramme.

EKSKLUDER på emne: Ekskluder studier som ikke handler om lærervurdering.
Begrunnelse: Det publiseres langt flere studier om elevvurdering enn om lærervurdering. Mange studier som ble identifisert handlet om elevvurdering, ikke lærervurdering, og disse ble ekskludert.

EKSKLUDER på utdanningsnivå: Ekskluder studier som ikke omhandler lærere i grunnskole og videregående skole.

Begrunnelse: Forskning om barnehager og høyere utdanning ligger utenfor prosjektets scope.

EKSKLUDER ikke-vitenskapelige studier: Ekskluder studier som ikke er publisert i tidsskrifter med fagfellevurdering.

Begrunnelse: Ved å begrense kunnskapsoversikten til bare å omfatte studier publisert i fagfellevurderte tidsskrifter sikres høyere kvalitet av de inkluderte studiene.

EKSKLUDER bøker/bokkapitler, avhandlinger og rapporter
Studier publisert som bøker/bokkapitler, avhandlinger og rapporter er ikke fagfellevurderte. Ved å begrense kunnskapsoversikten til bare å omfatte studier publisert i fagfellevurderte tidsskrift, sikres høyere kvalitet av de inkluderte studiene.

INKLUDER studie basert på tittel og sammendrag: Inkluder alle studier som ikke er ekskludert basert på eksklusjonskriteriene over.

IV. Eksempel på søkestreng med standardiserte emneord

Søk utført 25.11.2013 I ERIC-databasen

(SU.EXACT(«Elementary school teachers») OR SU.EXACT(«Mathematics teachers») OR SU.EXACT («Middle school teachers») OR SU.EXACT(«Reading teachers») OR SU.EXACT(«Science teachers») OR SU.EXACT(«Secondary school teachers«) OR SU.EXACT(«Writing teachers»)) AND (SU.EXACT(«Evaluation methods») OR SU.EXACT(«Formative evaluation») OR SU.EXACT(«Occupational tests») OR SU.EXACT(«Parent surveys») OR SU.EXACT(«Peer evaluation») OR SU.EXACT(«Performance based assessment») OR SU.EXACT(«Personnel evaluation») OR SU.EXACT(«Portfolio assessment») OR SU.EXACT(«School surveys») OR SU.EXACT(«Self evaluation (Individuals)») OR SU.EXACT(«Student evaluation of teacher performance») OR SU.EXACT(«Student teacher evaluation») OR SU.EXACT(«Summative evaluation») OR SU.EXACT(«Teacher competency testing») OR SU.EXACT(«Teacher evaluation») OR SU.EXACT(«Teacher surveys»)) AND (SU.EXACT.EXPLODE(«Academic achievement») OR SU.EXACT(«Accountability») OR SU.EXACT(«Achievement gains») OR SU.EXACT(«Bullying») OR SU.EXACT(«Career readiness») OR SU.EXACT(«Dropout prevention») OR SU.EXACT(«Dropout rate») OR SU.EXACT(«Educational attitudes») OR SU.EXACT.EXPLODE(«Educational environment») OR SU.EXACT.EXPLODE(«Educational improvement») OR SU.EXACT(«Educational quality») OR SU.EXACT («Excellence in education») OR SU.EXPLODE(«Feedback (Response)») OR SU.EXACT(«Goal orientation») OR SU.EXACT(«Graduation») OR SU.EXACT(«Graduation rate») OR SU.EXACT(«Knowledge level») OR SU.EXACT(«Language skills») OR SU.EXACT(«Mathematics skills») OR SU.EXACT(«Outcomes of education») OR SU.EXACT(«Professional development») OR SU.EXACT («Reading attitudes») OR SU.EXACT(«Reading improvement») OR SU.EXACT(«Reading skills») OR SU.EXACT(«School attitudes») OR SU.EXACT(«School effectiveness») OR SU.EXACT(«Student attitudes») OR SU.EXACT(«Student development») OR SU.EXACT(«Student educational objectives») OR SU.EXACT(«Student improvement») OR SU.EXACT(«Student school relationship») OR SU.EXACT(«Student teacher attitudes») OR SU.EXACT(«Student welfare») OR SU.EXACT(«Teacher effectiveness») OR SU.EXACT(«Teacher improvement») OR SU.EXACT(«Writing attitudes») OR SU.EXACT(«Writing improvement»))

V. Kilder for litteratursøk

Følgende kilder har blitt benyttet for litteratursøk i arbeidet med kunnskapsoversikten.

Elektroniske databaser tilgjengelige gjennom ProQuest-portalen
Education Resources Information Center (ERIC)
Applied Social Sciences Index and Abstracts (ASSIA)
International Bibliography of the Social Sciences (IBSS)
ProQuest Dissertions and Thesis A & I (PQDT A&I)
ProQuest Dissertions and Thesis - UK & Ireland (PQDT UK&I)
COS Conference Papers Index (COS)
ProQuest Education Journals (PQEJ)

Skandinavisk litteratur
Forskningsdatabasen.dk (DEFFnet)
REX-Det kongelige bibliotek og Københavns Universitets biblioteksservice (REX)
Bibliotek.dk
Roskilde universitetsbibliotek (RU)
Kungliga biblioteket-National Library of Sweden (KB)
IDUNN
BIBSYS/ASK

Håndsøk [38]
Individuelle fagfellevurderte tidsskrifter
Referanselister i inkluderte studier
Internett (Google Scholar)
Personlige kontakter (prosjektets forskergruppe og identifiserte forskere)

Grålitteratur [39]
OECD Library
RAND Corporation
Nordisk institutt for studier av innovasjon, forskning og utdanning (NIFU)
Arbeidsforskningsinstituttet (AFI)

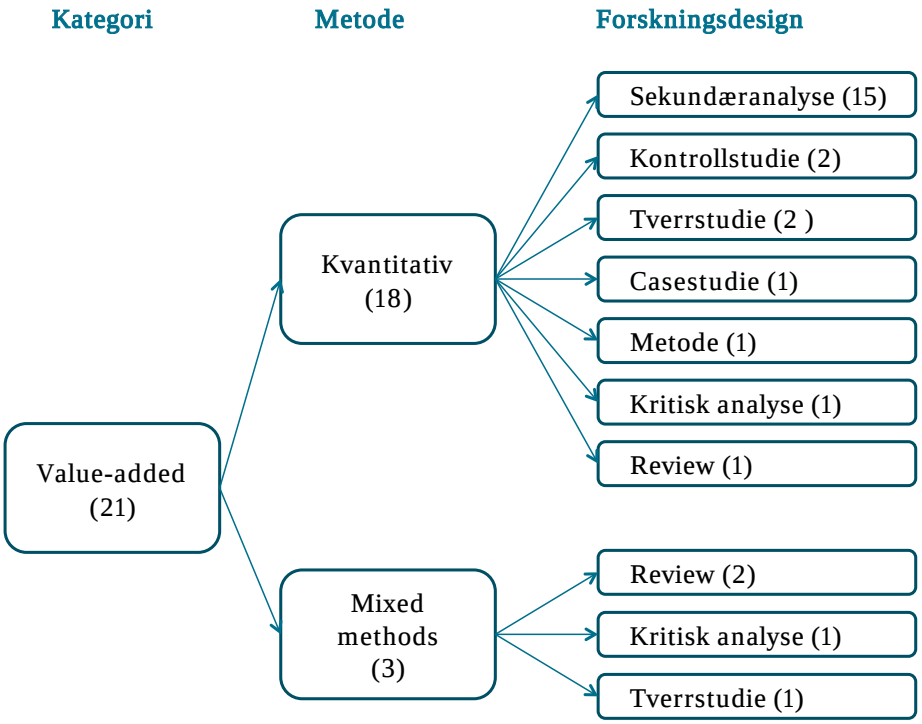
38 Programvaren som ble brukt til de elektroniske søkene fanger ikke opp all litteratur publisert i 2013. Det ble derfor gjort håndsøk i de tidsskriftene som publiserer artikler med relevans for prosjektets problemstilling (scope), det ble lett etter nyere publikasjoner av forfattere som tidligere har publisert innenfor tematikken og spesifikt søkt etter artikler fra 2013 og 2014 som bruker begrepet lærervurdering i tittel og sammendrag.

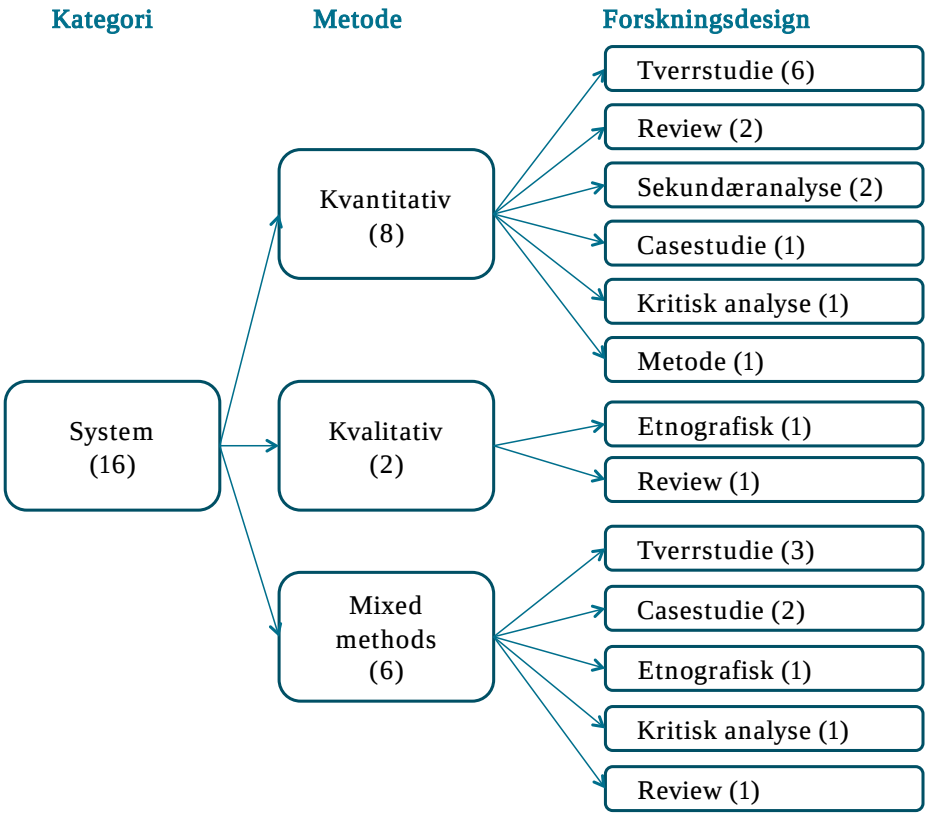
39 Søk i grå-litteratur er begrenset til følgende:

VI. Eksempel på skjema benyttet til kvalitetsvurdering av artiklene

STUDIE	LAND/ SPRÅK	FORSKNINGSSPØRSMÅL/ FORMÅL MED UNDERSØKELSEN	METODE	FORSKNINGS-DESIGN	METODE FOR DATA-INNSAMLING	FUNN
Casabianca m.fl. (2013)	USA	1) Vil man skåre vurderingsprotokoller fra klasser ulikt pga observasjonsmodus (observasjon vs video).. 2) Vil man rangere klasser ulikt ut i fra observasjonsmodus?	Kvantitativ	Tverrstudie	Observasjon og video	Det var små eller ingen utslag når det gjaldt observasjonstype i undervisning over et år. Tidsanalyser av den som observerte og hvordan hun eller han brukte observasjonsskjemaet var imidlertid signifikant for to CLASS-S domener og ga typeforskjeller i pålitelighet og sammenfall i data om de individuelle undervisningsøktene
Darling-Hammond m.fl. (2013)	USA	Undersøker hvor godt PACT (the Performance Assessment for California Teachers) kan predikere hvem som vil bli gode lærere.	Kvantitativ	Sekundæranalyse	Sekundære data	Skårene fra PACT–testen er signifikante predikatorer når det gjelder å forutsi hvem som blir gode engelsk- og matematikklærere. Et stort flertall av PACT kandidatene rapporterer at de ved å fullføre selve vurderingen tilegnet seg kunnskaper og bedret sine undervisningsferdigheter.
Delvaux og Vanhoof (2013)	Belgia	1) Hvilken påvirkning har lærervurderingssystemer på profesjonsutvikling sett fra lærernes perspektiv? 2) Hvilke komponenter i vurderingssystemet påvirker de effektene vurderingen har på profesjonsutvikling set fra lærernes ståsted? Hvilken innvirkning har disse komponentene?	Kvantitativ	Survey	Spørreunder-søkelse	Resultatene viser at begrensent undervisningserfaring, nyttig tilbakemelding, og positiv holdning fra rektor var de viktigste karakteristikkene i evalueringssystemet. Disse karakteristikkene er positivt korrelert med utfall på profesjonell utvikling.
Flores (2012)	Portugal	Hvilke oppfatninger har lærerne av den nye policyen vedrørende lærervurdering? Hvordan rangerer lærerne implementeringen av denne?	Mixed methods	Tverrstudie og etnografisk	Spørreskjema, intervju og fokusgruppe	Lærernes oppfatning av systemet for lærervurdering er preget av skepsis og usikkerhet. Kritiske momenter som løftes frem er mangelen på tillit til vurderere, den byråkratiske og summative dimensjonen og mangelen på nødvendige betingelser for implementering.
Hill m.fl (2012a)	USA	Forfatterne undersøker et sett antagelser knyttet til et validitetsargument og tester om hypotetiske forskjeller i implementeringen av observasjonsinstrumentet fører til ulike vurderinger.	Kvantitativ	Metode-diskusjon	Sekundære data	Argumenterer for at de som bruker ulike observasjonsverktøy og andre typer vurdering av undervisning/utdanning må undersøke validiteten på slike verktøy når de blir benyttet. Særlig er det viktig å se på effekten av å bruke forskjellige personer til å skåre observasjonsskjema, se nærmere på innholdet, prosedyren for observasjon og den lokale konteksten som observasjonene utføres i
Hill og Grossman (2013)	USA	Forfatterne diskuterer behovet for forbedrede observasjonsverktøy og hvordan dette kan oppnås.	Mixed methods	Kritisk analyse		Hvis et observasjonsverktøy skal innfri målet om å støtte lærerne i å forbedre egen undervisningspraksis, må skjemaer være fagspesifikke, eksperter på innholdet bør involveres i observasjonsprosessen, og skjemaet bør gi informasjon som er både presis og nyttig for lærerne.

VII. Eksempel på kartlegging av studier





VIII. Prinsipper for vurdering og system for lærervurdering

5 Prinsipper for vurdering:

1. De *ressursene* som trengs for å gjennomføre vurderingen må stilles til rådighet (ekspertise, økonomi, tid), og innsatsen må balanseres mot det man forventer å få ut av aktiviteten.
2. Vurderingen må planlegges og gjennomføres på en slik måte at den eller de som blir vurdert opplever resultatet av vurderingen som gyldig og *troverdig.*
3. Vurderingen må *avgrenses* til og konsentrere seg om bestemte sider ved arbeidet, men likevel ta hensyn til at det som vurderes inngår i en større sammenheng.
4. Vurdering bør gjennomføres på måter som *ikke bare måler prestasjoner*, men som også generelt *bidrar positivt* til arbeidet som utføres i skolen og styrker læringsmiljøet.
5. Vurdering har generelt stor innvirkning på praksis i et felt, så vurderingen *må planlegges og gjennomføres* på en slik måte at uønskete *virkninger av vurderingen minimeres.*

Chile

(Taut og Sun 2014, Tornero og Taut 2010, Santiago m.fl 2013, Taut m.fl. 2011 Santelices og Taut 2011)

Hvem: Initiert nasjonalt

Hva: Vurderingsobjektet er lite avgrenset og omfatter fire dimensjoner i standarden for god undervisning 1) planlegging og forberedelse, 2) skape et læringsmiljø, 3) profesjonsansvar og 4) undervise for å sikre læring hos hver enkelt student

Hvordan: Vurderes ved hjelp av mappevurdering, kollegavurdering, vurdering utført av veileder/leder og egenvurdering

Hvorfor: Forbedring og kontroll

Prinsipper for vurdering	Positivt	Negativt
1. Ressurser	- Svært omfattende støtteapparat rundt vurderingssystemet og opplæring av vurderere	- Lærerne opplever vurderingen som belastende fordi den skaper mye tilleggsarbeid (Taut og Sun 2014) - Det settes ikke av nok tid for lærerne til å gjennomføre selve vurderingsarbeidet (mappevurdering) (Tornero og Taut 2010) - Store lokale forskjeller i kommunene når det gjelder kompetanse og ressursinnsats (Santiago m.fl 2013)
2. Troverdig	- Vurderingssystemet oppleves som troverdig og implementeringen var godt forankret (Taut og Sun 2014) - Lærerne har tillitt til at egenvurderingen viser bredden i arbeidet (Taut og Sun 2014)	- Lærerne har ikke tillit til vurderingskompetansen til de som vurderer mappen (Santiago m.fl. 2013)
3. Avgrenset		- Systemet er stort og svært komplekst og omfatter mange sider ved lærerarbeidet (Taut og Sun 2014)
4. Ikke kun måle prestasjoner – bidra positivt til arbeid i skolen og styrke læringsmiljøet	- Systemet er formativt (basert på veiledning underveis), men samtidig summativt med et klart resultatfokus.	- Den formative siden er ikke godt nok ivaretatt (Taut m.fl. 2011, Taut og Sun 2014 og Santiago m.fl. 2013) - Manglende vurderingskompetanse og tid hos skoleledere gjør at oppfølgingen av undervisningspersonale blir tilfeldig (Santiago mfl. 2013) - Lærerne får ofte ikke de tilbakemeldingene de trenger (Santiago mfl. 2013) - Det lages sjelden utviklingsplaner for lærerne (Santiago m.fl. 2013)
5. Planlegges og gjennomføres slik at uønskede virkninger minimeres		- Systemet brukes stort sett for å holde lærerne ansvarlige for elevresultater (Santiago m.fl. 2013) - Skoleledere deltar for lite i vurderingsarbeidet (Santiago m.fl. 2013)

Kina

(Zhang og Ng 2011, Liu og Zhao 2013)

Hvem: Initiert nasjonalt

Hva: Vurderingsobjektet er knyttet til politisk ståsted, kompetanse, holdninger og prestasjoner

Hvordan: Egenvurdering, avdelingens vurdering av læreren, skolens vurdering av læreren, klasseromsobservasjon, elevens eksamensresultater, inspeksjon av lærernes daglige arbeid (lokale variasjoner)

Hvorfor: Forbedring og kontroll

Prinsipper for vurdering	Positivt	Negativt
1. Ressurser		- Lokale praksiser og gjennomføring gjør at ressursbruk og innsats varierer (Liu og Zhao 2013)
2. Troverdig		- Styrt og besluttet på nasjonalt plan - Oppleves ikke troverdig av lærere (Liu og Zhao 2013)
3. Avgrenset		- Vurderingen er avgrenset og definert av nasjonale myndigheter
4. Ikke kun måle prestasjoner – bidra positivt til arbeid i skolen og styrke læringsmiljøet	- Systemet endrer seg fra summativt til et mer formativt vurderingssystem (Zhang og Ng 2011) - Gjennom kriterier og mål for lærernes arbeid gir vurderingen retningslinjer og tilbakemelding underveis i vurderingsprosessen (Zhang og Ng 2011) - Kombinasjonen med prestasjonslønn rapporteres av forskerne som positivt (Liu og Zhao 2013) - Lærerne blir vurdert individuelt og kollektivt (Zhang og Ng 2011) - Vurderingsarbeidet bidrar til å kvalitetssikre lærernes profesjonsutvikling (Zhang og Ng 2011)	- Utfordring å forholde seg til belønning og sanksjoner når målet er at systemet skal virke formativt og ha innvirkning på lærernes profesjonsutvikling (Liu og Zhao 2013).
5. Planlegges og gjennomføres slik at uønskede virkninger minimeres		- Ulik praksis (vennskap viktigere enn hardt arbeid) for å oppnå godt resultat (Liu og Zhao 2013) fordi lærerne beskriver selv sin egen profesjonsutvikling. - Prestasjonslønn kan føre til et system basert på belønning/straff (Liu og Zhao 2013)

Belgia

(Delvaux og Vanhoofen 2013)

Hvem: Initiert nasjonalt

Hva: Individuelle prestasjonskriterier

Hvordan: Vurderingsintervju gjennomføres med hver lærer hvert fjerde år med utgangspunkt i et sett individuelle prestasjonskriterier som utgjør grunnlaget i vurderingssystemet. Lærerne har minst en formativ samtale der de får tilbakemelding om hva som kan forbedres.

Hvorfor: Forbedring og kontroll

Prinsipper for vurdering	Positivt	Negativt
1. Ressurser		
2. Troverdig		- Viktig med en positiv relasjon mellom lærer og vurderer for å oppnå troverdighet og lærerne må tilskrive de som vurderer legitimitet
3. Avgrenset		
4. Ikke kun måle prestasjoner – bidra positivt til arbeid i skolen og styrke læringsmiljøet	- Systemet har både summative og formative hensikter. Overordnet mål er at systemet skal fungere formativt og bidra til å forbedre undervisningen. (Delvaux og Vanhoof 2013)	- Forskerne finner ikke effekter av systemets formative funksjon (Delvaux og Vanhoofen 2013) - Systemet fungerer formativt gitt visse betingelser som at tilbakemeldingene er nyttige, og skoleleders holdning til systemet.
5. Planlegges og gjennomføres slik at uønskede virkninger minimeres	- Antall år undervisningserfaring ser ut til å spille en rolle i forhold til hvor stor nytte lærerne tilskriver systemet (Delvaux og Vanhoof)	

Portugal

(Flores 2012)

Hvem: Initiert nasjonalt

Hva: Vurderes i forhold til fire dimensjoner 1)profesjonell, etisk og sosial, 2) utvikling av undervisning og læring, 3) deltakelse i skoleaktiviteter og forhold til samfunnet og 4) læring og utvikling i et livslangt perspektiv

Hvordan: Egenvurdering, observasjon av undervisning (ikke obligatorisk), globalt skjema

Hvorfor: Forbedring og måling i forbindelse med karriereprogresjon

Prinsipper for vurdering	Positivt	Negativt
1. Ressurser		- Systemet fører med seg mye byråkrati og det er ikke satt av tilstrekkelig med tid til å gjennomføre tiltakene - Fører til stor ekstra arbeidsmengde for lærerne
2. Troverdig	- Vurderere må være fra samme fagfelt som lærerne de vurderer.	- De som vurderer har lav troverdighet og liten tillit blant lærerne
3. Avgrenset		
4. Ikke kun måle prestasjoner – bidra positivt til arbeid i skolen og styrke læringsmiljøet	- Systemet har en formativ målsetting, men måler prestasjoner i form av kvalitetsvurderinger.	- Har ført til en summativ praksis selv om formålet i hovedsak var formativt
5. Planlegges og gjennomføres slik at uønskede virkninger minimeres		- Det opprinnelige systemet måtte forenkles og revideres for å imøtekomme skepsis blant lærer og lærerforbund.

KUNNSKAPSSENTER
FOR UTDANNING

Kunnskapssenter for utdanning,
Norges forskningsråd

Telefon: +47 22 03 70 00

Epost: kunnskapssenter@forskningsradet.no

Interrnett: www.kunnskapssenter.no