Sigrid Blömeke
Universitetet i Oslo
Rolf Vegar Olsen
Universitetet i Oslo

DOI: http://dx.doi.org/10.5617/adno.6278

På vei mot et sammenhengende nasjonalt kvalitetsvurderingssystem

Sammendrag
Artikkelen presenterer de mest sentrale verktøyene for kvalitetsmonitorering som har blitt innført som en del av det «Nasjonale kvalitetsvurderingssystemet for grunnopplæringen» (de nasjonale prøvene, kartleggingsprøvene og elevundersøkelsen), i all hovedsak ut fra et perspektiv om at dette er målinger som skal bidra med informasjon til skoler, skoleeiere og nasjonale beslutningstakere. Gjennom beskrivelsen legger vi et grunnlag for en drøfting av seks forhold som vi foreslår som spesielt viktige for å videreutvikle det nasjonale kvalitetsvurderingssystemet. Hovedkonklusjonen fra denne drøftingen er at norsk skole har utviklet mange målinger som hver for seg i stor grad har god kvalitet. Det legges imidlertid ikke til rette for en helhetlig analyseramme som gjør det mulig å se de ulike målingene i sammenheng med hverandre, og som i sterkere grad vektlegger et utviklingsperspektiv.

Nøkkelord: NKVS, læringsprogresjon, kartleggingsprøver, nasjonale prøver, elevundersøkelsen

Towards a coherent national quality monitoring system

Abstract
This article gives an overview of the core tools introduced for quality monitoring in the Norwegian school system since 2003 (the national assessments, the mapping tests and the student survey). It emphasizes a perspective that these tools are measurements with the purpose of providing information to schools, school owners and national stakeholders. The initial description is used to raise and discuss six aspects suggested to be of vital importance for the future development of the national quality monitoring system. The main conclusion derived from this discussion is that Norwegian schools have access to several measures which by themselves to a large degree have high quality and are fit for purpose. However, the system does not provide an analytical frame in which the

various measures relate to each other to support a more holistic perspective on development.

Keywords: Norwegian quality monitoring system, learning progression, mapping tests, national assessment, student survey

Innledning

Det nåværende vurderingssystemet i Norge består av deler som representerer lange tradisjoner (eksamener og lærerbaserte standpunktkarakterer), kombinert med nye komponenter som har eksistert i så kort tid at en enhetlig praksis for bruk av resultater ikke synes å eksistere ennå (eksempelvis kartleggingsprøver og nasjonale prøver). Året 2003 representerer et symbolsk tidsskille for norsk skole gjennom Stortingets vedtak om å innføre et Nasjonalt kvalitetsvurderingssystem for grunnopplæringen (NKVS) (St.meld. nr. 30, 2003–2004; St.prp. nr. 1, 2002–2003). Vedtakene bygget i stor grad på den første delrapporten fra et offentlig utvalg som var nedsatt noen år tidligere (NOU, 2002:10). Vedtaket ble fulgt opp raskt, blant annet gjennom å introdusere flere sentralt utviklede instrumenter for å kunne monitorere kvaliteten i utdanningssystemet. Ifølge dette vedtaket skal målingene som rapporteres, representere «det helhetlige læringsutbyttet med vekt på kunnskaper, ferdigheter og holdninger» og også «prosesskvaliteten for å kunne skape så gode læringsmiljøer som mulig» (St.prp. nr. 1, 2002–2003, kap. 2). De ulike instrumentene ble innført på ulike tidspunkt og med vektlegging av vekslende begrunnelser og formål. En slik stykkevis innføring kan være en av de underliggende årsakene til at NKVS ble karakterisert som fragmentarisk og lite sammenhengende i en OECD-rapport (Nusche, Earl, Maxwell & Shewbridge, 2011), en vurdering som også Utdanningsdirektoratet sluttet seg til (Norwegian Directorate for Education and Training, 2011).

Denne artikkelen har til hensikt å oppsummere og diskutere status for et utvalg av instrumentene som ble innført med kvalitetssystemet. Det er viktig innledningsvis å presisere at artikkelen ikke berører alle sider ved vurderingsarbeid i skolen, og den er heller ingen helhetlig drøfting eller evaluering av NKVS. I det følgende beskriver og drøfter vi de nasjonale prøvene (NP), kartleggingsprøvene (KP) og elevundersøkelsen (EU). Dette er tre instrumenter som har til felles at de alle ble innført med NKVS, og de er alle utviklet fra et målingsperspektiv. Enkelt sagt innebærer et målingsperspektiv at man piloterer og evaluerer instrumentene i lys av begreper som reliabilitet og validitet. Videre er viktige kjennetegn for disse instrumentene at de har funksjoner knyttet til monitorering av kvaliteter på systemnivå. Lærere og skoleledere rapporterer også at tallmålene som disse instrumentene leverer, blir brukt i skolers kvalitetsarbeid (se kap. 3 i Mausethagen, Prøitz & Skedsmo, 2018a).

I artikkelens første del presenterer og diskuterer vi egenskaper ved og status til disse tre instrumentene. Vi gir også korte oppsummeringer av bakgrunn og formål for de ulike instrumentene. I tillegg omtaler vi i den første delen kort de såkalte skolebidragsindikatorene fordi disse er basert på resultater fra de nasjonale prøvene, og fordi de representerer et viktig perspektiv på måling og monitorering av kvalitet på systemnivå. Innledningsvis nevner vi at alle de utvalgte instrumentene blir utviklet av faglig sterke miljøer tilknyttet universiteter på bestilling fra Utdanningsdirektoratet. For KP og NP er dette også miljøer som over tid har utviklet seg til å bli det som kan karakteriseres å være permanente og profesjonelle testutviklingsorganisasjoner. Både KP og NP er basert på innholdsmessige rammeverk for de begrepene som skal måles, og det finnes måletekniske kvalitetskrav som skal etterleves. For EU finnes ikke et tilsvarende detaljert rammeverk, men også for dette instrumentet har akademiske miljøer vært involvert i utformingen. Alle instrumentene blir grundig pilotert i flere steg, og tekniske rapporter fra utviklingen blir gjennomgått av en ekstern kvalitetssikrer. Rapportene blir, med unntak av elevundersøkelsen, i liten grad gjort offentlig tilgjengelige.

I artikkelens andre del drøfter vi noen av de utfordringene og mulighetene som den første delen har bidratt til å identifisere. Til sammen drøfter vi her seks betingelser for videre utvikling av dagens system for kvalitetsmonitorering på systemnivå.

Betydning av andre instrumenter i skolens kvalitetsutvikling: eksamener og internasjonale studier

Det er i tillegg en rekke andre instrumenter som blir brukt på ett eller flere nivåer i monitorering av skolekvalitet, men som vi likevel ikke drøfter i denne artikkelen. For det første er eksamen mye brukt som indikator for resultatkvalitet. For videregående skoler er dette, sammen med standpunktkarakterer, de eneste tilgjengelige indikatorene for læringsutbytte. Hovedformålet med eksamen er å dokumentere sluttkompetansen hos den enkelte elev i skolefag ved endt opplæring (Meld. St. 28, 2015–2016). Eksamens- og standpunktkarakterene brukes for seleksjon til videre utdanning og yrkesliv. Og som det blir pekt på i Meld. St. 28 (2015–2016, s. 59): «Det stiller krav til en rettferdig vurderingspraksis og til at karakterene er et så objektivt uttrykk for elevens faglige kompetanse som mulig.» Utfallet av eksamener har dermed store konsekvenser for enkeltpersoner. Enhver diskusjon av eksamen som ikke tar hensyn til dette perspektivet, vil dermed være lite nyttig. En slik drøfting vil imidlertid sprenge rammene for en artikkel som denne, som har som hovedformål å drøfte NKVS i et systemperspektiv.

For det andre er de internasjonale undersøkelsene tildelt en viktig rolle i NKVS, og dette er undersøkelser som er tuftet på tydelige måletekniske kvalitetskriterier. Inntil nylig har resultater fra de internasjonale undersøkelsene vært de eneste tilgjengelige indikatorene for hvordan norske elevers faglige dyktig-

heter har endret seg over tid. I tillegg har de internasjonale undersøkelsene et formål i NKVS ved at læringsutbyttet i norsk skole kan sammenliknes med andre land som det kan være relevant å sammenlikne seg med (se f.eks. Olsen & Björnsson, 2018). Resultatene fra undersøkelsene gir imidlertid ikke informasjon på skole- eller skoleeiernivå, og følgelig drøftes heller ikke disse undersøkelsene videre i denne artikkelen.

En beskrivelse av målingene i NKVS

I denne delen gir vi en beskrivelse av de mest sentrale prøvene og verktøyene i det nasjonale kvalitetsvurderingssystemet: Kartleggingsprøvene (KP), de nasjonale prøvene (NP) og elevundersøkelsen (EU).

Kartleggingsprøvene

Kartleggingsprøvene (KP) er de første standardiserte målingene en norsk elev møter. De gjennomføres i løpet av de fire første skoleårene. Det er prøver på første, andre og tredje trinn i lesing og regning. I tillegg finnes en kartleggingsprøve i engelsk for tredje trinn og i grunnleggende digitale ferdigheter for fjerde trinn[1]. Formålet med KP er å identifisere elever som kan trenge ekstra ressurser eller støtte, definert som elevene i den nedre femtedelen av fordelingen. Den kritisk viktige måleegenskapen ved prøven er altså hvor godt dette skillet (grenseverdien) fungerer for å identifisere de elevene som har behov for forsterket opplæring i en kortere eller lengre periode. Denne egenskapen ved prøvene er gitt prioritet i utviklingen, noe som fører til at prøvene for de aller fleste elever er svært lette og ikke gir god informasjon.

I dag er deltakelse i leseprøvene på 1., 2. og 3. trinn, samt regneprøven for 2. trinn, obligatorisk, mens deltakelse i de andre prøvene blir avgjort på kommune- og/eller skolenivå. De samme prøvene blir brukt i en periode på om lag fem år. Det samles jevnlig inn prøveresultater fra et begrenset nasjonalt representativt utvalg av skoler/elever for å kunne gjøre vurderinger av om prøvene fungerer slik de skal, men utover dette finnes det ingen sentral registrering av data fra prøvene (se også Nortvedt, 2018). Det er kommunene og/eller skolene som avgjør hvordan data fra disse prøvene blir samlet og brukt. Tidligere rapporterte noen kommuner indikatorer på skolenivå, gjerne i form av andel elever under grenseverdien, en praksis som vi (gjennom kontakt med sektoren) har inntrykk av er mindre utbredt nå. Evalueringen av NKVS viser for øvrig at prøvene i all hovedsak blir brukt lokalt ved den enkelte skole i arbeidet med enkeltelevene. Kartleggingsprøvene er populære, og om lag ¾ av skolene deltar også i de frivillige prøvene (Utdanningsdirektoratet, 2017).

[1] https://www.udir.no/eksamen-og-prover/prover/kartlegging-gs/

KP er først og fremst ment å tjene som et pedagogisk verktøy[2]. Lærere, skoleledere og skoleeiere får generell informasjon fra Utdanningsdirektoratet om mulig oppfølging av resultatene fra prøvene. Her påpekes blant annet viktigheten av at en enkeltstående prøve ikke har perfekt reliabilitet på elevnivå, og at lærerne og skolene derfor må bruke resultatene fra disse prøvene sammen med annen informasjon om elevene. Det finnes imidlertid lite kunnskap om hvordan skoleeiere, skoler og lærere tolker og anvender prøveresultater, om de blir brukt til å styre undervisning eller som informasjon for andre pedagogiske eller organisatoriske beslutninger. Spørreundersøkelser viser at lærerne stort sett anerkjenner at prøvene treffer formålet (Allerup, Kovac, Kvåle, Langfeldt & Skov, 2009).

Med unntak av oppstarten av KP i regning (Alseth, Throndsen & Turmo, 2009) vet vi lite om reliabiliteten til kuttskåren som definerer det kritiske skillet mellom den nedre femtedelen og alle andre elever. Det finnes heller ikke, så vidt vi vet, undersøkelser som søker å validere prøvenes grunnleggende formål, altså som et verktøy for identifisering og oppfølging av enkeltelever som sliter i fagene. Når det gjelder identifisering, bør det eksempelvis gjennomføres studier hvor elever følges over tid for å kunne si noe om prøvene bidrar til dette formålet. Når det gjelder oppfølging, er det avgjørende å kunne presentere et argument om at arbeidet med tidlig innsats støttes av prøvene. Det finnes så vidt vi vet heller ikke informasjon om prøvenes faktorstruktur (der det finnes delprøver) eller prøvenes egenskaper når det gjelder å fungere likt på tvers av ulike elevgrupper.

Kommunene og skolene er som sagt de eneste som har tilgang til data fra prøvene. Tanken bak dette er nok å sikre at prøvene ikke blir brukt til andre formål enn de er tenkt, nemlig å identifisere elever som trenger hjelp. Ideen er nok også at lokalt eierskap til dataene skal gi en sterkere forpliktelse til å bruke resultatene i oppfølgingen. Innsamling av data, spesielt dersom resultatene blir offentliggjort, vil kunne føre til at de oppfattes som kvalitetsindikatorer for skoler. En risiko med dette vil være at kvalitetssystemet gir insentiver for å realisere et lavt ambisjonsnivå knyttet til å løfte elever over den kritiske grensen. Det at data ikke samles systematisk fra prøvene, medfører imidlertid at det ikke er helt enkelt å gjennomføre gode valideringsstudier.

Nasjonale prøver

På 5. og 8. trinn gjennomføres det hvert år obligatoriske nasjonale prøver (NP) i lesing, regning og engelsk. Prøvene i lesing og regning gjennomføres også for 9. trinn samme år[3]. Eierskap og ansvar for disse prøvene ligger på nasjonalt nivå, det vil si hos Utdanningsdirektoratet. Det er derfor en sentral datainnsamling gjennom et elektronisk prøveadministrativt system (PAS). Prøvenes hovedformål er å evaluere elevenes oppnåelse av grunnleggende ferdigheter slik disse

[2] https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-kartleggingsprover-pa-1.-4.-trinn/
[3] https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/

beskrives i læreplanene. Resultatene fra prøvene rapporteres detaljert for den enkelte elev og i form av gjennomsnitt for aggregerte nivåer (klasse, skole, skoleeier/kommune, fylke, land). NP er dermed først og fremst ment å tjene som styringsverktøy på systemnivå med «system» definert som skole, kommune eller fylke (Seland, Vibe & Hovdhaugen, 2013). Imidlertid ligger det også et annet formål for prøvene, nemlig at de skal fungere som en formativ indikator for planlegging av undervisning og læring. I omtalen av NP finner man åpenbare selvmotsigelser i beskrivelser av disse to formålene. I Utdanningsdirektoratets rapport til OECD om det norske målingssystemet sies det at «The tests provide useful information about groups of students (school and municipal level), but not detailed information on individual students» (Norwegian Directorate for Education and Training, 2011, s. 70), men den samme rapporten fastslår også at «Students and their parents will be told the results from their mapping tests and national tests and the results will be used to follow up the students" (ibid., s. 73).

Nasjonale prøver i skriving som grunnleggende ferdighet var også opprinnelig inkludert, men disse prøvene ble avsluttet etter en evaluering som konkluderte med at prøvene ikke tilfredsstilte grunnleggende psykometriske kriterier (Lie, Hopfenbeck, Ibsen & Turmo, 2005)[4]. Denne evalueringen avdekket for øvrig mangelfull kvalitet ved alle prøvene. Delvis på bakgrunn av denne rapporten ble det vedtatt en full stans i gjennomføringen av alle de nasjonale prøvene i noen år. Når prøvene kom i gang igjen i 2007, ble tidspunktene for prøvene flyttet fra slutten av 4. og 7. trinn til høsten 5. og 8. trinn, og prøvene for 10. trinn ble nedlagt. Det å flytte prøvene til begynnelsen av mellom- og ungdomstrinnet signaliserte også sterkere et læringsstøttende formål for prøvene. De nye prøvene var basert på et utvidet teoretisk og metodisk grunnlag gjennom rammeverk for de ulike fagprøvene og et felles psykometrisk rammeverk. Det finnes også egne nasjonale prøver i lesing for tre samiske språk (se Henriksen, Eira, Keskitalo & Øzerk, 2018).

Rapportering til elever, foreldre, skoler og kommuner skjer i form av mestringsnivåer (tre og fem nivåer for prøvene på henholdsvis 5. og 8. trinn), og disse nivåene er knyttet til overordnete beskrivelser av hva som er typiske kjennetegn ved kompetanse for elever på de ulike nivåene. I tillegg har det i de siste årene også blitt rapportert en nasjonalt standardisert skår[5]. Oppgavene til NP blir utviklet på nytt hvert år, men siden 2014 har prøvene i regning og engelsk (og siden 2016 leseprøvene) inkludert et sett med oppgaver som gjentas over tid, såkalte ankeroppgaver. Gjennom innføringen av ankeroppgaver er det i dag mulig å analysere trender over tid for skoler, kommuner og høyere nivå. Dette har dermed styrket mulighetene for å kunne monitorere endringer i elevenes faglige kompetanser (se for øvrig Björnsson, 2018).

[4] Prøvene fortsetter imidlertid å eksistere som læringsstøttende prøver (se Skar & Aasen, 2018).
[5] Nasjonalt gjennomsnitt satt til 50 og ett standardavvik satt til 10 det første året prøven ble gjennomført med ankeroppgaver.

Elevundersøkelsen

Elevundersøkelsen (EU) gjennomføres årlig og samler informasjon om det som generelt kan kalles for transversale ferdigheter[6] (motivasjon, målorientering, selvbilde, utholdenhet osv.). I tillegg inneholder EU spørsmål hvor elevene beskriver sine opplevelser av det som skjer i klasserommet. Denne typen målinger er viktig for å kunne vurdere kvaliteten til et utdanningssystem. Forskning over mange tiår har vist at enkeltelevers utvikling i motivasjon, selvbilde og mestringsopplevelse er svært viktige prediktorer for utvikling i faget og vice versa (Elliot, Dweck & Yeager, 2017; Petersen & Hyde, 2017).

Det finnes så vidt vi kjenner til, ikke et grundig teoretisk rammeverk som ligger til grunn for EU, men det foreligger flere publiserte rapporter hvor formål og psykometriske egenskaper er nevnt (blant annet Wendelborg, Røe, Federici & Caspersen, 2015). I tillegg har det blitt publisert vitenskapelige artikler basert på analyser av data fra undersøkelsen (blant annet Federici, Caspersen & Wendelborg, 2016). Basert på denne tilgjengelige informasjonen er det derfor mulig å konkludere at indeksene som utvikles fra spørsmålene i elevundersøkelsen er tematisk godt forankret i forskningen om psykososiale forhold i læringsprosesser. Skalaene ser ut til å ha god reliabilitet, men når målene aggregeres til skolenivå, er det tydelig at flere av indikatorene har store gulv- og takeffekter – noe som vil si at nesten alle skolene ligger svært lavt eller svært høyt på skalaene. Flere av indikatorene har derfor lite følsomhet for eventuelle endringer på skolenivå. Målingen av læringskonteksten (som indikator for undervisningskvalitet) er mindre godt relatert til pedagogisk forskning (se for eksempel Nilsen & Gustafsson, 2016). Vi vet for eksempel at det er viktig å knytte spørsmål om undervisning til spesifikke fag (Opheim, Gjerustad & Sjaastad, 2013). De konkrete spørsmålene og svarskalaene i EU har også blitt endret så ofte at det er vanskelig å bruke denne informasjonen til å følge utviklingen over tid. Trendanalyser er bare mulig siden 2014 (Wendelborg, 2018).

Dataene fra EU er lagret hos SSB, men undersøkelsene gjennomføres, med gode grunner, slik at elevene er anonyme (men skoletilhørighet lagres). Undersøkelsen har som formål at elevene skal kunne gi oppriktige og ærlige svar, blant annet om hvordan de opplever støtte fra lærerne og om mobbing. Det er derfor ikke mulig å koble dataene med annen informasjon på individnivå. Elevers selvrapporterte karakterer på spørreskjemaet fra elevundersøkelsen er dermed eneste mulighet til å koble de psykososiale variablene til kognitiv utvikling (Wendelborg & Caspersen, 2016). I og med at skoleinformasjonen er lagret, kan man i noen grad gjøre analyser på skolenivå gjennom å lenke til annen informasjon om eksempelvis andel minoritetsspråklige elever, skolens størrelse, geografisk beliggenhet eller lærertetthet for å nevne noen variabler (Wendelborg et al., 2015).

[6] Soft skills, 21st century skills og ikke-kognitive størrelser er noen merkelapper som inkluderer mange av de samme begrepene.

Elevundersøkelsen er obligatorisk bare for 7. og 10. trinn, men skoler/ skoleeiere kan velge å gjennomføre undersøkelsen for alle trinn fra 5. og oppover. Dette velger mange skoler å gjøre, og mer enn halvparten av alle norske elever utenfor de obligatoriske trinnene deltar dermed i undersøkelsen. Selv om dette gir en stor database som inkluderer svar fra mange titusen elever på alle trinn, gir et slikt design med frivillighet metodiske utfordringer knyttet til selvseleksjon og manglende informasjon om hvilke elever som ikke deltok. Ved å inkludere flere bakgrunnsvariabler kunne det vært mulig å kontrollere for mulige skjevheter i utvalget (Wendelborg et al., 2015).

Skolebidragsindikatorer

Gjennom prøveresultater får skoler informasjon om det faglige nivået til elevene på skolen for noen utvalgte sentrale områder, og dette er nyttig informasjon for å kunne sammenlikne med andre (tilsvarende) skoler og for å identifisere tiltak for videre skoleutvikling. Imidlertid er det vanskelig å tolke prøveresultatene som en indikator på skolens kvalitet. Disse resultatene sier vel så mye om hvilke elever som går på skolen.

For å fange inn skolekvalitet har det vært gjennomført to runder med utvikling av såkalte skolebidragsindikatorer (SBI). Dette er ikke et instrument i seg selv, men heller en metodikk for å sette sammen data om elevers prestasjoner over tid. For å være mer presis er SBI et såkalt *value-added*-resultatmål (se f.eks. OECD, 2008) som tar i bruk tidsserier for enkeltelever – inkludert nasjonale prøver og eksamenskarakterer for 10. trinn og i videregående skole (Falch, Bensnes & Strøm, 2016; Hægeland, Kirkebøen, Raaum & Salvanes, 2005; Steffensen, Ekren, Zachrisen & Kirkebøen, 2017). SBI utvikles gjennom en analyse av resultatene for elevene på et gitt tidspunkt ved å kontrollere for elevenes tidligere prestasjoner (og eventuelt andre kjennetegn ved elevene som ikke kan knyttes til skolens kvalitet, eksempelvis indikatorer for elevenes hjemmebakgrunn). Tanken er da at noen skoler vil prestere bedre (positivt skolebidrag) eller svakere (negativt skolebidrag) enn man kunne forvente ut fra kjennetegn ved elevgruppen som går på skolen.

Drøfting av seks betingelser for et godt kvalitetsvurderingssystem

I denne delen løfter vi fram noen temaer som har vært berørt i beskrivelsene ovenfor. I denne sammenhengen er det viktig å gjenta at artikkelen i all hovedsak er en betraktning om ambisjonen som NKVS har om å legge til rette for god monitorering for kvalitetsutvikling på systemnivå.

Kvalitetsutvalgets rapporter banet som nevnt veien for dagens kvalitetsvurderingssystem slik det etter hvert har utviklet seg (NOU, 2002:10; NOU, 2003:16). Utvalget opererte med begrepene resultat-, prosess- og strukturkvalitet. Vi har i denne artikkelen lagt vekt på å drøfte noen av de sentrale

indikatorene for resultatkvalitet, i noe mindre grad omtalt indikatorer for prosesskvalitet, og vi har valgt å se helt bort fra indikatorer for strukturkvalitet. Å utvikle gode indikatorer for kvaliteter i prosesser og strukturer er utfordrende, og mange av tiltakene som vi i det følgende antyder som relevante framover for å videreutvikle NKVS, er knyttet til utvikling av indikatorer som reflekterer kvaliteten til prosessene enda bedre.

Monitorering på systemnivå

Et system forstås her som et organisatorisk nivå som koordinerer, regulerer og legger føringer for utdanningen. Det vil si at det for et system tas beslutninger om innhold, form, kapasitet og ressurser for utdanningen. Et skolesystem kan være et helt land, som for eksempel Norge som har et sentralstyrt og nasjonalt skolesystem. I andre land med mer desentralisert myndighet kan systemet være et fylke eller en delstat, eksempelvis i Tyskland eller USA (Jones & Olkin, 2004). På systemnivå trenger politiske og administrative myndigheter informasjon som gir tilbakemelding på områder de kan påvirke (OECD, 2013). På systemnivå skal man eksempelvis ha en effektiv økonomisk håndtering og bevilge penger til mange ulike formål. Med andre ord er det på dette nivået behov for høyt aggregerte indikatorer.

For et slikt formål gir derfor godt gjennomførte utvalgsundersøkelser god og tilstrekkelig informasjon, noe som gjør det mulig å skaffe til veie et rikholdig og bredt informasjonsgrunnlag uten at alle elever deltar på alle målinger (Greaney & Kellaghan, 2008). Dette er et tilbakevendende moment i debatten om eksempelvis de nasjonale prøvene. En fordel med utvalgsprøver er at kun en liten andel av elever/skoler deltar hver gang de gjennomføres. Man kan derfor også tenke seg at dette gir mulighet til å gjennomføre prøver i flere ulike fagområder enn i dag[7]. Dette er også en løsning som hindrer at monitorering oppleves som kontroll og overvåking.

Det er mindre vanlig å betrakte en kommune som et system, fordi mange avgjørelser er tatt på høyere nivå og ikke kan påvirkes av kommunen. Opplæringslovas § 13-10[8] pålegger imidlertid skoleeierne å dokumentere kvaliteten i sine skoler, og Forskrift til opplæringslova spesifiserer dette i mer detalj i kapittel 2[9]. Dette medfører en ansvarliggjøring av skoleeiere, og det er dermed rimelig å hevde at kommunene utgjør egne mikrosystemer med sine egne behov for informasjon om kvalitet. De fleste kommuner og skoler i Norge er imidlertid så små at *utvalgs*undersøkelser ikke er nyttige. For å få gode mål for læringsresultatene for skolens elever, må derfor alle elevene inkluderes. Dette gir data som spesielt skoleeierne og skolelederne rapporterer at de har nytte av (Allerup et al., 2009). Mausethagen, Prøitz og Skedsmo (2018a) identifiserer at såkalte

[7] I tillegg til slike utvalgsundersøkelser kan det sentrale nivået skaffe seg informasjon for å monitorere, styre og skaffe til veie beslutningsgrunnlag gjennom aktiv bruk av registerdata og ved å utlyse målrettede oppdrag (forskning og evaluering).
[8] https://lovdata.no/dokument/NL/lov/1998-07-17-61#KAPITTEL_15
[9] https://lovdata.no/dokument/SF/forskrift/2006-06-23-724/KAPITTEL_3#KAPITTEL_3

«resultatdialoger» nærmest er blitt et allment fenomen. Dette er gjerne årlige møter hvor representanter fra kommunen og skolene har en strukturert samtale om kvalitet, og i disse samtalene har blant annet resultater fra NP, KP og EU en tyngde. Selv om slike dialoger nå skjer i alle eller de fleste kommuner, varierer de i form og hva som vektlegges (Mausethagen, Prøitz & Skedsmo, 2018b). En omlegging til utvalgsundersøkelser, til tross for de fordelene slike undersøkelser kan ha, vil dermed sannsynligvis redusere skoleeiernes og skoleledernes mulighet til å ta ansvar for det lokale arbeidet med å utvikle sin egen virksomhet.

Standardiserte målinger av elevenes kunnskap og motivasjon skal imidlertid som regel tjene flere formål (Clarke, 2012). Tveit og Olsen (2018) identifiserer tre formål: sertifisering, styring og støtte av læring og undervisning. Lærere og skoler skal ta i bruk informasjonen fra prøvene for å støtte det videre arbeidet med å gi elevene god undervisning og læring. Dette formålet med de nasjonale prøvene (og de andre verktøyene) blir ikke vektlagt i denne artikkelen. Men det er opplagt at i motsetning til monitoreringsfunksjonen vil et slikt formål peke mot noen andre valg i utformingen av kvalitetssystemet: i stedet for aggregerte mål vil et slikt perspektiv vektlegge å skaffe til veie mer detaljert informasjon om enkeltelevers prestasjoner. Dette formålet peker også i retning av at utvalgsundersøkelser ikke er tilstrekkelig. Videre vil vektlegging av et slikt formål peke i retning av å gi lærerne tilgang til mer fleksible instrumenter som de kan benytte når de mener de trenger denne informasjonen[10]. Dette siste vil være i konflikt med et monitoreringsperspektiv som er avhengig av å ha en standardisert gjennomføring av prøven, og begrensninger i tidspunktet for gjennomføringen av prøven er en viktig standardisering. Utover dette er imidlertid de fleste av de momentene som vi diskuterer videre i denne artikkelen (fra et målings- og systemperspektiv), være like relevante for en diskusjon om bruk av prøver for å støtte elevers læring og utvikling.

Måling av progresjon

Ludvigsen-utvalget hadde som mandat å beskrive utviklingstrekk for norsk skole framover. I sin rapport peker de blant annet på behovet for å ha et sterkere fokus på elevers progresjon i fagene (NOU, 2015:8). Dersom man skal ta et slikt progresjons- eller forløpsperspektiv på alvor, blir det viktig å ha gode verktøy for å kunne observere og måle elevenes kompetanse på flere tidspunkter, noe dagens prøvesystem for øvrig allerede gjør. Dagens verktøy gir imidlertid ikke gode muligheter for å se alle de ulike kvalitetsindikatorene i sammenheng, og det er dermed heller ikke opplagt hvordan man kan bruke informasjonen i et forløps- eller læringsprogresjonsperspektiv. Det er tre momenter som har en avgjørende betydning dersom man skal vektlegge en omlegging av dagens prøvesystem til å gi informasjon om progresjon: det er nyttig å ha et *tidlig målepunkt*, man må ha et design for å kunne *lenke* resultater fra ulike prøver til hverandre,

[10] De læringsstøttende prøvene er tenkt brukt nettopp slik (se https://www.udir.no/eksamen-og-prover/prover/laringsstottande-prover/).

og man må ha individuelle målinger med *høy presisjon*. I det følgende skisserer vi elementer i en løsning for å oppnå dette. Et viktig utgangspunkt for oss er at dette skal kunne la seg realisere med det samme antallet prøver som i dag.

Av dagens prøver er de nasjonale prøvene på 5. trinn det første mulige startpunktet for en tidsserie for progresjon i faglige kompetanser. På dette tidspunktet er mye av grunnlaget for videre læring allerede lagt. Dersom man ønsker å vektlegge målinger av elevers (og skolers) utviklingsløp, er det derfor interessant å reflektere over hvordan dagens kartleggingsprøver for 1.–3. trinn eventuelt kan inngå som de første målepunktene av faglig progresjon. Som nevnt er dette prøver som er utviklet for å identifisere de elevene som sliter mest i fagene og som en konsekvens er prøvene svært lette for de fleste elevene. For en stor andel elever er derfor skårene på KP upresise. En mulig løsning som ivaretar det opprinnelige formålet (identifisere elever med så store lærevansker at de kan trenge spesiell oppfølging) og som samtidig legger til rette for å kunne bruke disse prøvene som startpunkt for målinger av læringsprogresjon, er å gjennomføre kartleggingsprøvene som såkalte *adaptive* prøver (Magis, Yan & von Davier, 2017), en prøveform man eksempelvis har i de danske nasjonale prøvene (Bundsgaard, in press). Kort fortalt gjennomføres en adaptiv prøve (som regel) på en datamaskin. Prinsippet for adaptive prøver er at man på forhånd har utviklet et stort antall oppgaver med kjent vanskegrad. Disse lagres i en oppgavebank, og ut fra elevens svar på et innledende sett med oppgaver begynner systemet å tildele oppgaver som er tilpasset elevens nivå.

Det neste elementet som må være på plass dersom man ønsker å måle elevers progresjon, er lenking. Lenking betyr at prøvene er utformet slik at det er meningsfullt å sammenlikne resultater på tvers av prøvene. De nasjonale prøvene er allerede lenket over tid gjennom et design hvor det benyttes ankeroppgaver. Dette gjør at skoler kan sammenlikne prøveresultater fra ett år til et annet. Resultatene på tvers av trinn kan i dag imidlertid ikke sammenliknes i absolutt forstand.

Tenk for eksempel på en elev som får skåren 45 på NP på 5. trinn og skåren 55 tre år senere på prøven på 8. trinn. Det man kan si om denne elevens progresjon, er at hun presterte et halvt standardavvik lavere enn gjennomsnittet nasjonalt i femte klasse, mens hun presterte et halvt standardavvik bedre enn det nasjonale gjennomsnittet tre år senere. Relativt til det nasjonale gjennomsnittet for sin alderskohort har hun altså hatt en forbedring[11]. Men det er ikke mulig ut fra dette å si noe om hva som kjennetegner denne elevens progresjon. Vi vet strengt tatt ikke om de to prøvene måler det samme underliggende begrepet, selv om prøvene har samme merkelapp, eksempelvis «lesing». Å være en dyktig

[11] Beskrivelsen i dette eksemplet er basert på antakelsen om at skalaen holder seg stabil med et gjennomsnitt på 50 og et standardavvik på 10. Så langt har det vært slik i de nasjonale prøvene, men siden prøvene har et ankerdesign, kan disse verdiene endre seg over tid (se Björnsson, 2018), og denne tolkningen er dermed en forenkling for å få fram prinsippet om forbedring som en relativ størrelse vs. faktisk og absolutt progresjon.

leser på 8. trinn vil sannsynligvis inkludere andre aspekter ved lesing enn på 5. trinn.

For å kunne måle en progresjon i mer absolutte termer trenger vi å kunne plassere disse to skårene på den samme skalaen. Og dersom vi i tillegg har en god kvalitativ forståelse av skalaen, kan den gi innblikk i typiske kjennetegn ved elevenes kompetanse på 8. trinn sammenliknet med 5. trinn. En vanlig metode for å lenke to eller flere prøver er nettopp ved å bruke ankeroppgaver som nevnt ovenfor. Man kan eksempelvis bruke noen av de vanskeligste oppgavene fra prøven på 5. trinn i prøven for 8. trinn. Tilsvarende prinsipper kan brukes for å lenke resultater fra kartleggingsprøvene til de nasjonale prøvene, forutsatt at de endres som beskrevet over, til å bli bedre målinger også for elever med høyere dyktighet.

En annen fordel med et slikt lenket design er at man også vil kunne gi skolebidragsindikatorer gjennom en enkel aritmetisk differanse mellom to av målepunktene – uten behov for å bygge lite gjennomsiktige statistiske modeller, slik det gjøres i beregningene av dagens skolebidragsindikatorer. Dette gjør også at skolebidragsindikatorer kan rapporteres regelmessig, uten ekstra kostnader eller ressurser – og raskere enn i dag. Skoler med ulike utgangspunkt kan på denne måten få målinger som i større grad enn i dag reflekterer skolekvalitet.

En gunstig bieffekt av å utvikle prøver som er lenket sammen for å måle progresjon, er for øvrig at dette vil bidra med en instrumentering som kan benyttes i forskning og evalueringer som sikter mot å gi kausale tolkninger av tiltak på systemnivå (Raudenbush & Liu, 2000; Raudenbush, Martinez & Spybrook, 2007).

Helhetlig måling av elevers kompetanse

Lenking av prøver for å kunne måle progresjon er et svar på spørsmålet om hvordan man kan få et nasjonalt kvalitetsvurderingssystem som henger mer sammen og gir muligheter for mer helhetlige analyser. Et annet viktig element i et helhetlig målesystem ville være å gi aktører på ulike nivåer muligheter til å kunne se progresjon i læringsresultater i sammenheng med andre utviklingstrekk for elevene. Det finnes et godt teoretisk og empirisk grunnlag for å hevde at elevenes utvikling av en rekke transversale ferdigheter (sosiale ferdigheter, selv-regulering, utholdenhet, motivasjon osv.) er svært viktig også for elevenes faglige utvikling (se for eksempel Duckworth & Yeager, 2015).

Det finnes derfor mange som ut fra ulike ståsteder argumenterer for at dette er ferdigheter som kan og bør utvikles i skolen, og den offentlige utredningen Fremtidens skole (NOU, 2015:8) har fanget opp disse argumentene og foreslått et kompetansebegrep med fire dimensjoner: i tillegg til «fagspesifikk kompetanse» inkluderes kompetanse i «å lære», i «å kommunisere, samhandle og delta» og i «å utforske og skape». Utvalget påpeker videre at skolen vil ha behov for andre former for vurderingsverktøy for å fange inn et slikt flerdimensjonalt kompetansebegrep.

Utvalget har ikke vurdert om eller hvordan det kan legges til rette for at disse kompetansene skal kunne monitoreres på systemnivå. Gitt at de er viktige forutsetninger for læring, er det logisk at skoler og skoleeiere som er pålagt et ansvar for kvalitetsutvikling, også har behov for informasjon om status (og progresjon) for flere av dimensjonene i et slikt mangefasettert perspektiv på læringsutbytte eller resultatkvalitet. Elevundersøkelsen inkluderer målinger av motivasjon og mestring som kan sies å falle inn under denne kategorien av kompetanser.

En annen forutsetning for å kunne gjøre mer helhetlige analyser av kvalitet er å gi aktørene på de ulike nivåene verktøy for å kunne følge utviklingen over tid også for de begrepene som fanges inn av elevundersøkelsen. Vi ser imidlertid at dette er utfordrende. For at dette skal være mulig må en eller annen instans, eksempelvis SSB, ha tillatelse til å knytte elevers identitet til svar på elevundersøkelsen, noe som opplagt gir behov for grundige etiske og juridiske vurderinger knyttet til lovverk om personvern. I tillegg til etiske, juridiske og tekniske aspekter ved håndtering av data er også enhver nyttig og gyldig bruk av resultater fra EU helt avhengig av at elevene kan være trygge på at svarene på EU behandles konfidensielt, og at elevenes responser ikke er kjent for lærere, skoler og skoleeiere.

La oss her (for resonnementets skyld) anta at det er mulig å løse disse utfordringene. I så fall kan det produseres indikatorer for utvikling over tid også for disse viktige kvalitetene og læringsutbyttene i skolen. Dette vil videre gjøre det mulig å lage indikatorer på skolenivå for *sammenhenger* med resultater fra de faglige prøvene. Dette er viktig informasjon for å støtte opp under læreres og skolers arbeid med å utvikle gode læringsforløp. Elevers motivasjon for mange skolefag ser blant annet ut til å synke på mellom- og ungdomstrinnet (se blant annet Wendelborg, 2018). Dette er bekymringsfullt fordi vi vet at elevenes videre valg av utdanning i stor grad påvirkes av motivasjon og selvbilde i fag (Eccles, Barber, Updegraff & O'Brien, 1998). Nettopp derfor identifiserer også den nasjonale strategien «Tett på realfag» tiltak rettet mot elevenes motivasjon og affektive forhold til realfagene som helt sentrale for å øke andelen elever som velger seg et videre utdanningsløp i en realfaglig retning (Kunnskapsdepartementet, 2015). Imidlertid er det vanskelig å kunne evaluere om denne strategien (og andre tiltak) er effektive dersom man ikke samtidig kan inkludere representative målinger av hvordan elevers motivasjon og forhold til realfagene utvikler seg over tid.

Måling av hva som skjer i klasserommet: Undervisningskvalitet

Det finnes i dag svært lite informasjon på systemnivå om læringskontekstene. Det eneste som finnes, er flere enkle indikatorer for lærertetthet. I tillegg finnes det noen vurderinger av læringsmiljøet og undervisningen i elevundersøkelsen. Prosessene i klasserommet er dermed usynlige for beslutningstakere på nasjonalt nivå – og i noen grad også for skoleledere. Det gjennomføres riktignok flere større videobaserte klasseromsstudier (se f.eks. Klette, Blikstad-Balas & Roe,

2017), og det finnes spørreskjemaer til elever og lærere i de internasjonale undersøkelsene, og da spesielt i TIMSS og PIRLS, som til sammen gir kunnskap om prosesser i klasserommet på nasjonalt nivå. Men denne typen informasjon samles ikke som en del av den systematiske kunnskapsinnhentingen for alle skoler. Generelt er undervisningskvalitet et mangefasettert begrep (Fauth, Decristan, Rieser, Klieme & Büttner, 2014; Kuger, Klieme, Jude & Kaplan, 2016) som inkluderer direkte målinger av forekomsten av det som forskning har vist er kjennetegn ved god undervisning (f.eks. kognitiv utfordring, støttende lærer og klasseledelse). Både lærere og elever kan gi relevant informasjon for å fange inn slike begreper. Elevundersøkelsen inkluderer i dag målinger av «Faglig utfordring» og «Støtte fra lærerne» (for skolen som helhet på tvers av fag), men dette er en dimensjon i spørreskjemaet som også kan vurderes å bli utvidet. Tross faglige gode grunner for å samle også slik informasjon om det enkelte klasserommet, bør nok rapporteringen av denne typen informasjon like-vel avgrenses til skolenivået (og høyere) for å unngå at dataene blir brukt i ulike former for utilsiktet lokal evaluering med konsekvenser for enkeltlærere.

Kvalitetssikring og validering

Dagens prøver går gjennom flere kvalitetssikringssteg. Hos prøveutviklerne arbeides det med forankring til skolevirkeligheten, hvor oppgavene prøves ut i flere steg, og som også inkluderer bruk av lærere eller tidligere lærere med god kontekstuell innsikt som testutviklere. Videre gjennomgår prøvene en ekstern kvalitetssikring av prøvenes kvalitet i lys av det gjeldende rammeverket for psykometrisk kvalitet. I stor grad er dette kriterier som er viktige for å lage indikatorer som er gode for hovedformålet med kvalitetsmonitorering. Hvis rapportene i tillegg offentliggjøres, vil de grundige kvalitetsprosessene også bli kjent og det vil være mulig å få en saklig kritisk debatt om prøvene – noe som i seg selv vil styrke muligheten for validering av prøvene.

Prøvenes formål som en pedagogisk ressurs er imidlertid i mindre grad kvalitetssikret eller validert. Et generelt akseptert prinsipp for validering er at man gjennomfører studier som er spesielt utformet for å finne evidens som støtter eller ikke støtter opp under de spesifikke tolkningene som man ønsker at det skal være mulig å gjøre fra testskårene (Haertel, 2013; Kane, 2013). Validering av prøvenes kvalitet og hvordan prøveresultatene kan støtte ulike typer tolkninger, er tid- og ressurskrevende. I tillegg er dette forsknings- eller evalueringsaktivitet som krever spesiell kompetanse. Det er behov for studier som har fokus på å validere at testene måler det de skal måle (konstruktvalidering), men det er også et stort behov for studier som undersøker om og hvordan prøvene blir brukt for de intenderte formålene. Flere artikler har kommet ut i det siste som peker på utfordringene både skoleeiere og lærere har med å trekke ut informasjon fra NP som er relevant for bruk på den enkelte skole eller for å gi informasjon om undervisningen (Allerup et al., 2009; Hovdhaugen, Vibe & Seland, 2017; Mausethagen et al., 2018a; Monsen, 2014; Mausethagen, 2013;

Roald, 2010). Det er derfor et behov for stadig å utvikle bedre måter å rapportere data på – innrettet for ulike grupper og for ulike formål.

Rapportering og kommunikasjon av usikkerhet og endring

Skoleporten[12] og Utdanningsspeilet[13] er i dag de viktigste kildene til å presentere resultater fra de nasjonalt koordinerte datainnsamlingene som utgjør NKVS. I rapporteringen legges det vekt på å gjøre det mulig å sammenlikne enheter (skoler, kommuner og fylker). Det er viktig å kombinere rapporteringer som tilfredsstiller strenge vitenskapelige krav, med offentlig tilgjengelige rapporter som framstiller resultater i mer tilgjengelig språk og form.

I denne konteksten er det relevant å løfte fram viktigheten av å gi forståelige presentasjoner av usikkerhet og spredning (Greaney & Kellaghan, 2008). Det er store målefeil knyttet til alle mål på elevnivå, og for små skoler er det også en betydelig målefeil på skolenivå. Det er imidlertid viktig å ikke skape et feilaktig inntrykk av usikkerheter. Data i Skoleporten rapporteres eksempelvis med usikkerhet knyttet til utvalgsfeil – altså usikkerheter knyttet til at elevgruppen anses å være et tilfeldig utvalg (fra en mye større populasjon). Men fra et skoleperspektiv er elevgruppen nettopp *ikke* et tilfeldig utvalg, men heller populasjonen man ønsker å ha informasjon om[14]. Når det er sagt, er det likevel fornuftig å rapportere utvalgsfeil for gjennomsnittsresultater når disse skal brukes som indikatorer for endringer fra en alderskohort til den neste. Jo mindre en skole er, jo mer vil gjennomsnittsresultatene svinge fra en alderskohort til den neste.

En generell anbefaling er derfor at rapporteringsverktøyet legger til rette for to typer tolkninger: a) tolkninger knyttet til den spesifikke elevgruppen som var inkludert i en måling, og b) tolkninger knyttet til skolens langsiktige arbeid med utvikling av kvalitet. For arbeidet med elevgruppen er det viktigste å få på plass verktøy som gjør det mulig å følge elevenes progresjoner over tid – og for dette bør man rapportere feilmarginer som skyldes at prøveskårene ikke har perfekt reliabilitet. Videre bør man rapportere målefeil for bestemte viktige kuttskårer som gis en bestemt tolkning, eksempelvis kritisk grense for KP. For skolens arbeid med kvalitet på et overordnet nivå er det også viktig at man legger til rette for å se data over større tidsrom. Skoler er relativt små enheter, noe som skaper naturlige svingninger i resultater for skolen over tid. En vanlig løsning for slike problemer er at man bruker såkalte glidende gjennomsnitt[15] som vil være langt mindre påvirket av tilfeldige svingninger fra ett år til det neste.

[12] https://skoleporten.udir.no/
[13] Se siste utgave http://utdanningsspeilet.udir.no/2017/
[14] Den viktige kilden til usikkerhet i dette tilfellet er målefeil, ikke utvalgsfeil, og dette er en kilde til usikkerhet som minker raskt når resultater aggregeres opp.
[15] Glidende gjennomsnitt består i at man rapporterer gjennomsnitt for lengre tidsperioder. For eksempel kan man rapportere tre-års glidende gjennomsnitt på følgende måte: Resultatet for 2017 rapporteres som gjennomsnittet fra prøvene i 2015, 2016 og 2017; resultater for 2018 rapporteres som gjennomsnittet fra prøvene i 2016, 2017 og 2018, osv.

Aggregerte resultater på kommune- og fylkesnivå er mer presise. Her er utfordringen at det i mange tilfeller er veldig små forskjeller, og man bør derfor også kommunisere betydningen av størrelsen på forskjellene. Her kan en hjelpestørrelse være at man for mange prøver observerer at en forskjell på 0,4–0,7 standardavvik[16] (eller en forskjell på 4–7 poeng for de nasjonale prøvene) svarer til omtrent den samlede effekten av at elever har gått ett år ekstra på skolen og blitt ett år eldre. Skoleporten har en omfattende hjelpeside med hjelp til tolkning av resultater, som også oppfordrer til edruelighet i tolkninger. Her blir usikkerheter knyttet til gjennomsnitt for skoler og høyere nivåer rapportert i form av konfidensintervaller, og spredningen rapporteres som en angivelse av den 20. og den 80. prosentilen. Ut fra den offentlig tilgjengelige informasjonen ser det imidlertid ikke ut som om det oppgis målefeil for enkeltelevers skårer, og heller ikke for kuttskårene som definerer kompetansenivåene[17].

Oppsummering og konklusjon

Denne artikkelen har diskutert status for noen utvalgte instrumenter som ble innført som en del av NKVS. Perspektivet for drøftingene har vært at dette er instrumenter som brukes for å måle sentrale kvaliteter i norsk skole som redskap for kvalitetsarbeid på systemnivå. Instrumentene som er diskutert, er de nasjonale prøvene, kartleggingsprøvene og elevundersøkelsen.

Videre har vi presentert seks sentrale betingelser for at målesystemet skal gi god støtte til langsiktig og evidensbasert skole- og politikkutforming:

1. Et system for måling av kvalitet på et nasjonalt nivå kunne vært gjennomført ved hjelp av utvalgsundersøkelser, men øvrige formål som prøvene har (som støtte for lokalt kvalitetsarbeid og lærernes arbeid med elevene i klasserommet), tilsier at hele elevpopulasjoner må inkluderes.
2. Vi har ikke vurdert hvor mange prøver og målinger vi bør ha i norsk skole, men tar som utgangspunkt at det er mulig å forbedre kvalitetssystemet med de målingene som eksisterer. Et viktig steg i så måte er å utvikle dagens utforming av systemet for å kunne lenke prøveresultater over tid for å få informasjon om elevers progresjon – sammen med utvikling av kvalitative kjennetegn for progresjon i ferdighetene som måles. I dette har vi også foreslått at ved å gjøre kartleggingsprøvene adaptive vil man kunne beskrive læringsprogresjoner helt fra første klassetrinn.
3. Det er behov for å utvikle målinger av andre typer kompetanser som forskning viser er viktige, og som Ludvigsen-utvalget har foreslått som

[16] Dette er en effekt som avtar noe over tid. For NP5 kan man bruke 0,7 som et estimat for denne ettårseffekten, mens man for NP8 kan bruke 0,4.
[17] Se et eksempel på hvordan resultatene publiseres for lærere her: https://prover.udir.no/eksempelvisning#/

sentrale i framtidens skole. Dette er målinger som også bør inngå i kvalitetsarbeidet på systemnivå.

4. For å få et mer helhetlig kvalitetssystem kan det være en fordel å utvikle indikatorer for sammenhenger mellom ulike typer mål (resultatkvalitet, prosesskvalitet og strukturkvalitet) for skoler og høyere nivåer.
5. Selv om de fleste av instrumentene er utviklet fra rammeverk for faglig innhold og psykometrisk kvalitet, finnes det lite forskning/evaluering av om instrumentene har de egenskapene som formålene tilsier, eksempelvis om grenseverdiene som benyttes for kartleggingsprøvene på en god og reliabel måte identifiserer elever som er i en risikosone.
6. Rapportering og kommunikasjon av prøveresultater er et viktig element for å sikre at skoler og skoleeiere får en god forståelse av hva prøvene kan si noe om – og hvilke begrensninger som ligger i tolkninger av resultatene.

Til sammen framhever disse seks betingelsene at det er behov for en tydelig og kontinuerlig drøfting av formålene med de ulike målingene som er inkludert i det nasjonale kvalitetssikringssystemet (Tveit, 2018). I tillegg må det ikke underslås at også eksamener er en viktig del av kvalitetsvurderingssystemet. Gjennomsnittlige eksamensresultater brukes lokalt av de enkelte skolene (Mausethagen et al., 2018a), og nasjonalt brukes eksamensresultater for å drøfte tilstanden i norsk skole (se f.eks. de fleste utgaver av Utdanningsspeilet). Eksamensordningen har en lang tradisjon, det finnes et generelt rammeverk[18], og oppgavene kvalitetssikres i form av at mange personer er involvert i eksamensnemndene. Men verken rammeverket for eksamen eller andre dokumenter angir eksplisitte kvalitetskriterier eller testspesifikasjoner som sier noe om eksamen som målinger. Det at resultater fra eksamen blir brukt til en rekke formål, fører til dilemmaer og utfordringer som identifiseres og drøftes av blant annet Hovdhaugen, Prøitz og Seland (2018) og av Tveit og Olsen (2018) i andre artikler i dette spesialnummeret av Acta Didactica.

Status for systematisk kvalitetsvurdering i norsk skole er likevel en annen nå enn da en ekspertgruppe fra OECD (1988–89) påpekte at et slikt system var totalt fraværende. Med innføringen av det nasjonale kvalitetsvurderingssystemet har man fått utviklet mange relevante og gode verktøy som tilbyr skoler, skoleeiere og det nasjonale nivået informasjon om prioriterte kvalitetsområder. I tillegg har det blitt etablert et rapporteringssystem som gir tilgang til data for ulike grupper. En kortfattet oppsummering av gjennomgangen av dette kvalitetssystemet i denne artikkelen, er likevel: Til tross for at norsk skole for så vidt måler nok, så er de enkelte målingene løsrevet fra hverandre. Ambisjonen bør være å utvikle et kvalitetssystem som tilbyr en helhetlig analyseramme. I et slikt system bør det være mulig å se de ulike målingene i sammenheng med

[18] https://www.udir.no/eksamen-og-prover/eksamen/rammeverk-eksamen/

hverandre, og det bør i sterkere grad vektlegge et utviklingsperspektiv gjennom å legge til rette for måling av progresjon.

Om forfatterne

Sigrid Blömeke er professor ved Universitetet i Oslo og leder for Centre for Educational Measurement. Hennes forskningsinteresser omfatter blant annet måling av læreres kompetanse og sekundære analyser av data fra internasjonale storskalastudier.
Institusjonstilknytning: Centre for Educational Measurement, Universitetet i Oslo, Postboks 1161 Blindern, 0318 Oslo.
E-post: sigrid.blomeke@cemo.uio.no

Rolf Vegar Olsen er professor ved Universitet i Oslo. Hans forskningsinteresser omfatter blant annet prøveutvikling og sekundære analyser av data fra internasjonale storskalastudier.
Institusjonstilknytning: Centre for Educational Measurement, Universitetet i Oslo, Postboks 1161 Blindern, 0317 Oslo.
E-post: r.v.olsen@cemo.uio.no

Referanser

Allerup, P., Kovac, V., Kvåle, G., Langfeldt, G, & Skov, P. (2009). *Evaluering av det Nasjonale kvalitetsvurderingssystemet for grunnopplæringen* (FoU-rapport 8/2009). Kristiansand: Agderforskning.

Alseth, B., Throndsen, I. & Turmo, A. (2009). *Rapport fra kartleggingsprøver i tallforståelse og regneferdighet for 2. årstrinn og Vg1*. Oslo: Institutt for lærerutdanning og skoleutvikling.

Björnsson, J. K. (2018). Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid. *Acta Didactica Norge, 12*(4), Art. 16.

Bundsgaard, J. (in press). Pædagogisk brug af test. *Sakprosa*.

Clarke, M. (2012). *What Matters Most for Student Assessment Systems: A Framework Paper*. SABER Student Assessment Series Working paper no 1. Washington, DC: World Bank. http://siteresources.worldbank.org/INTREAD/Resources/7526469-1335214323234/WP1_READ_web_4-19-12.pdf

Duckworth, A. L. & Yeager, D. S. (2015). Measurement Matters: Assessing Personal Qualities Other Than Cognitive Ability for Educational Purposes. *Educational Researcher*, *44*(4), 237–251. doi: https://doi.org/10.3102/0013189x15584327

Eccles, J. S., Barber, B. L., Updegraff, K. & O'Brien, K. M. (1998). An expectancy-value model of achievement choices: The role of ability self-concepts, perceived task utility and interest in predicting activity choice and course enrolment. I L. Hoffmann, A. Krapp, K. A. Renninger & J. Baumert (red.), *Interest and learning: Proceedings of the Seeon Conference on Interest and Gender* (s. 267–281). Kiel: University of Kiel, Institute for Science Education.

Elliot, A. J., Dweck, C. S. & Yeager, D. S. (red.) (2017). *Handbook of Competence and Motivation: Theory and Application*. New York: The Guilford Press.

Falch, T., Bensnes, S. & Strøm, B. (2016). *Skolekvalitet i videregående opplæring: Utarbeidelse av skolebidragsindikatorer og mål på skolekvalitet* (SØF-rapport 01/16). Trondheim: Senter for økonomisk forskning.

Fauth, B., Decristan, J., Rieser, S., Klieme, E. & Büttner, G. (2014). Student ratings of teaching quality in primary school: Dimensions and prediction of student outcomes. *Learning and Instruction, 29*, 1–9.

Federici, R. A., Caspersen, J. & Wendelborg, C. (2016). Students' Perceptions of Teacher Support, Numeracy, and Assessment for Learning: Relations with Motivational Responses and Mastery Experiences. *International Education Studies, 9*. http://www.ccsenet.org/journal/index.php/ies/article/view/56543

Greaney, V. & Kellaghan, T. (2008). *Assessing National Achievement Levels in Education.* Washington, DC: World Bank.

Haertel, E. (2013). How is testing supposed to improve schooling? *Measurement: Inter-disciplinary Research and Perspectives, 11*, 1–18.

Henriksen, M., Eira, K. I., Keskitalo, J. H. & Øzerk, K. (2018). Nasjonale prøver i lesing på samisk – på hvilke premisser? *Acta Didactica Norge, 12*(4), Art. 4.

Hovdhaugen, E., Prøitz, T. S. & Seland, I. (2018). Eksamens- og standpunktkarakterer – to sider av samme sak? *Acta Didactica Norge, 12*(4), Art. 17.

Hovdhaugen, E., Vibe, N. & Seland, I. (2017). National test results: Representation and misrepresentation. Challenges for municipal and local school administration in Norway. *Nordic Journal of Studies in Educational Policy, 3*, 95–105.

Hægeland, T., Kirkebøen, L. J., Raaum, O. & Salvanes, K. G. (2005). *Skolebidragsindikatorer: Beregnet for avgangskarakterer fra grunnskolen for skoleårene 2002–2003 og 2003–2004* (Rapporter SSB 2005/33). Oslo: Statistisk sentralbyrå.

Jones, L. V. & Olkin, I. (red.) (2004). *The Nation's Report Card: Evolutions and Perspectives.* Bloomington: Phi Delta Kappa International.

Kane, M. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Klette, K., Blikstad-Balas, M. & Roe, A. (2017). Linking Instruction and Student Achievement. A research design for a new generation of classroom studies. *Acta Didactica Norge, 11*(3), Art. 10. doi: http://dx.doi.org/10.5617/adno.4729

Kuger, S., Klieme, E., Jude, N. & Kaplan, D. (2016). *Assessing contexts of learning: An international perspective*. Rotterdam: Springer.

Kunnskapsdepartementet (2015). *Tett på realfag: Nasjonal strategi for realfag i barnehagen og grunnopplæringen (2015–2019).* https://www.regjeringen.no/contentassets/869faa81d1d740d297776740e67e3e65/kd_realfagsstrategi.pdf

Lie, S., Hopfenbeck, T. N., Ibsen, E. & Turmo, A. (2005). *Nasjonale prøver på ny prøve. Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Magis, D., Yan, D. & von Davier, A. A. (2017). *Computerized Adaptive and Multistage Testing with R: Using Packages catR and mstR.* Rotterdam: Springer.

Mausethagen, S. (2013). Talking about the test. Boundary work in primary school teachers' interactions around national testing of student performance. *Teaching and Teacher Education, 36*, 132–142.

Mausethagen, S., Prøitz, T. S. & Skedsmo, G. (2018a). *Elevresultater. Mellom kontroll og utvikling*. Bergen: Fagbokforlaget.

Mausethagen, S., Prøitz, T. S. & Skedsmo, G. (2018b). Teachers' use of knowledge sources in "result meetings": Thin data and thick data use. *Teachers and Teaching, 24,* 37–49.
Meld. St. 28 (2015–2016). *Fag – Fordypning – Forståelse: En fornyelse av Kunnskapsløftet*. Oslo: Kunnskapsdepartementet.
Monsen, M. (2014). *Store forventninger? Læreroppfatninger om eksterne leseprøver*. Doktoravhandling, Universitetet i Oslo.
Nilsen, T. & Gustafsson, J. E. (red.) (2016). *Teacher Quality, Instructional Quality and Student Outcomes. Relationships Across Countries, Cohorts and Time*. Cham: Springer International Publishing.
Nortvedt, G. A. (2018). «Det er et verktøy, ikke sant, for oss?» – Erfaringer fra fire gjennomføringer med kartleggingsprøver i regning 2014–2017. *Acta Didactica Norge, 12*(4), Art. 8.
Norwegian Directorate for Education and Training (2011). *OECD Review on Evaluation and Assessment Frameworks for Improving School Outcomes: Country Background Report for Norway*. http://www.oecd.org/education/school/47088605.pdf
NOU (2002:10). *Førsteklasses fra første klasse. Forslag til rammeverk for et nasjonalt kvalitetsvurderingssystem av norsk grunnopplæring*. Oslo: Statens forvaltningstjeneste, Informasjonsforvaltning.
NOU (2003:16). *I første rekke – Forsterket kvalitet i en grunnopplæring for alle: Utredning fra et utvalg oppnevnt ved kongelig resolusjon av 5. oktober 2001. Avgitt til Utdannings- og forskningsdepartementet 5. juni 2003*. Oslo: Statens forvaltningstjeneste, Informasjonsforvaltning.
NOU (2015:8). *Fremtidens skole. Fornyelse av fag og kompetanser*. Oslo: Departementenes sikkerhets- og serviceorganisasjon, Informasjonsforvaltning.
Nusche, D., Earl, L., Maxwell, W. & Shewbridge, C. (2011*). OECD Reviews of Evaluation and Assessment in Education: NORWAY*. Paris: OECD.
OECD (1988–89). *Ekspertvurdering fra OECD. OECD-vurdering av norsk utdanningspolitikk* (A.-M. Smith, Trans.). Oslo: Aschehoug.
OECD (2008). *Measuring improvements in learning outcomes: Best practices to assess the value-added of schools.* Paris: OECD Publishing.
OECD (2013). *Synergies for Better Learning: An International Perspective on Evaluation and Assessment.* Paris: OECD.
Olsen, R. V. & Björnsson, J. K. (2018). 20 år med internasjonale skoleundersøkelser i Norge: Bakgrunn, læringspunkter og veien videre. I J. K. Björnsson & R. V. Olsen (red.), *Tjue år med TIMSS og PISA i Norge: Trender og nye analyser*. Oslo: Universitetsforlaget. https://www.idunn.no/tjue_aar_med_timss_og_pisa_i_norge
Opheim, V., Gjerustad, C. & Sjaastad, J. (2013). *Jakten på kvalitetsindikatorene: Sluttrapport fra prosjektet "Ressursbruk og læringsresultater i grunnopplæringen"* (Rapport 23/2013). Oslo: NIFU.
Petersen, J. L. & Hyde, J. S. (2017). Trajectories of self-perceived math ability, utility value and interest across middle school as predictors of high school math performance. *Educational Psychology, 37*(4), 438–456.
Raudenbush, S. W. & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods, 5*, 199–213.
Raudenbush, S. W., Martinez, A. & Spybrook, J. (2007). Strategies for Improving Precision in Group-Randomized Experiments. *Educational Evaluation and Policy Analysis, 29*, 5–29.
Roald, K. (2010). *Kvalitetsvurdering som organisasjonslæring mellom skole og skoleeigar*. Doktoravhandling, Universitetet i Bergen.

Seland, I., Vibe, N. & Hovdhaugen, E. (2013). *Evaluering av nasjonale prøver som system* (Rapport 4/2013). Oslo: NIFU.

Skar, G. B. U. & Aasen, A. J. (2018). Å måle skriving som grunnleggende ferdighet. *Acta Didactica Norge, 12*(4), Art. 10.

Steffensen, K., Ekren, R., Zachrisen, O. O. & Kirkebøen, L. J. (2017). *Er det forskjeller i skolers og kommuners bidrag til elevenes læring i grunnskolen? En kvantitativ studie* (Rapporter 2017/2). Oslo: Statistisk sentralbyrå.

St.meld. nr. 30 (2003–2004). *Kultur for læring*. Oslo: Utdannings- og forsknings-departementet.

St.prp. nr. 1 (2002–2003). *Tillegg nr. 3 (2002–2003). FOR BUDSJETTERMINEN 2003: Om tilleggsforslag i statsbudsjettet for 2003 under kapitler administrert av Utdannings- og forskningsdepartementet.* Oslo: Utdannings- og forskningsdepartementet.

Tveit, S. & Olsen, R. V. (2018). Eksamens mange roller i sertifisering, styring og støtte av læring og undervisning i norsk grunnopplæring. *Acta Didactica Norge, 12*(4), Art. 18.

Tveit, S. (2018). Ambitious and ambiguous: shifting purposes of national testing in the legitimation of assessment policies in Norway and Sweden (2000–2017). *Assessment in Education: Principles, Policy & Practice*, *25*(3), 327–350. doi: https://doi.org/10.1080/0969594X.2017.1421522

Utdanningsdirektoratet (2017). *Oppfølging av kvalitetsgjennomgangen på prøvefeltet*. Brev til Kunnskapsdepartementet 18.12.2017.

Wendelborg, C. (2018). *Analyser av indekser på Skoleporten 2017: Analyser på fylkes- og nasjonalt nivå for 7. trinn, 10. trinn og Vg 1.* Trondheim: NTNU Samfunnsforskning.

Wendelborg, C. & Caspersen, J. (2016). *Høyt presterende elevers vurdering av læringsmiljøet: Analyser av Elevundersøkelsen 2013 og 2014.* Trondheim: NTNU Samfunnsforskning.

Wendelborg, C., Røe, M., Federici, R. A. & Caspersen, J. (2015). *Elevundersøkelsen 2014. Analyse av Elevundersøkelsen 2014*. Trondheim: NTNU Samfunnsforskning.