Utdannings-
direktoratet

Kunnskapsgrunnlag for evaluering av eksamensordningen

Eksamensgruppa har sammenstilt kunnskapsgrunnlaget vi har om eksamen i dag for å vurdere hvilken betydning fagfornyelsen og den teknologiske utviklingen bør ha for eksamensordningen. Dokumentet blir bearbeidet underveis med arbeidet til eksamensgruppa.

RAPPORT | SIST ENDRET: 27.02.2019

Innhold

Mer informasjon om den eksterne eksamensgruppa

Vurderinger og foreløpige anbefalinger fra eksamensgruppa

1. Innledning

Eksamensgruppa er nedsatt av Kunnskapsdepartementet, jf. brev fra departementet til Utdanningsdirektoratet (heretter Udir) datert 26. juni 2018. Gruppa skal bistå direktoratet i arbeidet med å utrede et helhetlig eksamensordningssystem for fagene som omfattes av fagfornyelsen. Dette dokumentet representerer delleveranse 1 fra eksamensgruppa «Kunnskapsgrunnlaget eksamen: foreløpig status og vurdering». Leveransen vil videreutvikles og utvides gjennom ytterlige to delleveranser i 2019 og inngå i eksamensgruppas sluttrapport til Kunnskapsdepartementet som skal sendes i 2020.

1.1 Medlemmer i eksamensgruppa og mandat

Eksamensgruppa

Sigrid Blömeke (leder), Universitetet i Oslo
Sissel Skillinghaug, Utdanningsdirektoratet
Marte Blikstad-Balas, Universitetet i Oslo
Per-Odd Eggen, NTNU – Norges teknisk-naturvitenskapelige universitet
Henning Fjørtoft, NTNU – Norges teknisk-naturvitenskapelige universitet
Siv Therese Måseidvåg Gamlem, Høgskulen i Volda
Tine Prøitz, Universitetet i Sørøst-Norge
Sverre Tveit, Universitetet i Agder
Rita Helgesen, Norsk Lektorlag
Stig Johannessen, Skolelederforbundet

Martin Minken, Utdanningsforbundet
Agathe Waage, Elevorganisasjonen
Mette Johnsen Walker, Skolenes landsforbund

Sekretariat

Cathrine Hjulstad, Hilde Hultin, Trude Saltvedt, Øyvind Pedersen og Per Kristian Larsen-Evjen (leder), Utdanningsdirektoratet

Det er i tillegg oppnevnt en referansegruppe med deltakere fra skolesektoren som skal gi innspill i arbeidet med leveransene. Referansegruppas innspill til kunnskapsgrunnlaget er tatt inn i dette dokumentet.

Referansegruppas medlemmer

Siri Halsan, KS
Marianne Lindheim, KS
Pål Georg Rødsten, Oslo kommune
Kjetil Stavø Høvig, Fylkesmannen i Vestland
Ragnhild Sperstad Lyng, Fylkesmannen i Trøndelag
Kajsa Kemi Gjerpe, Senter for samiske studier (UiT)
Karen-Inga Eira, Samisk høgskole (Kautokeino)

Eksamensgruppas oppdrag er å sammenstille kunnskapsgrunnlaget om eksamen, vurdere innspillene fra læreplangruppene og se på hvilken betydning fagfornyelsen og den teknologiske utviklingen bør ha for eksamensordningen. Eksamensgruppa kan også foreslå mulige endringer og nye eksamensformer innenfor følgende rammer:

1. Standpunktkarakterer og eksamenskarakterer som sluttvurderingsform skal bestå.
2. Sluttvurderingen skal fortsatt være individuell og faglig.
3. Dagens omfang og fordeling mellom eksamenskarakterer og standpunktkarakterer på elevenes vitnemål skal i hovedsak videreføres. Små justeringer kan vurderes.
4. Et eksamenssystem som lar seg gjennomføre med omtrent samme ressurser som i dag.
5. Eksamen skal fungere som kvalitetssikring for den enkelte elev gjennom ekstern vurdering og retten til å

klage.

6. Eksamen skal fortsatt kunne brukes som verktøy i kvalitetsutvikling og -sikring for skoleeier, skoler og lærere.

7. Eksamenssystemet skal fortsatt brukes som kilde til kompetanseutvikling for lærere (sensorskolering, felles kjennetegn på måloppnåelse og deltakelse i vurderingsfellesskap).

1.2 Bakgrunn for oppdraget

Fagfornyelsen skal lede til læreplaner som er mer relevante for framtiden. Målet er å styrke utviklingen av elevenes dybdelæring og forståelse samt få fram tydelige prioriteringer i fagene. Fagfornyelsen innebærer et utvidet kompetansebegrep og en ny overordnet del som framhever verdigrunnlaget. Blant annet skal elevene jobbe tverrfaglig, og kritisk tenkning og refleksjon samt kreativitet blir en viktig del av hva elevene skal lære i skolen. Samtidig skjer det en teknologisk utvikling og digitalisering på de aller fleste samfunnsarenaer. Bruken av digitale læremidler i opplæringen har økt, og teknologiutvikling er en driver for endring av skolens innhold. Denne utviklingen er tydelig gjenspeilet i de nye læreplanene.

Eksamen har høy legitimitet og aksept i samfunnet som en viktig del av elevens sluttvurdering, men når vi endrer læreplanene, er det nødvendig at vi også ser på vurderingsordningene, slik at vi sikrer god sammenheng mellom læreplanene og vurderingsformene. I NOU 2015: 8 ble det derfor anbefalt at det nedsettes et ekspertutvalg som går gjennom sluttvurderingssystemet og utreder hvordan standpunkt og eksamen samlet sett kan gi pålitelig og relevant informasjon om elevenes kompetanse. Eksamenssystemet har vært relativt stabilt, og prosedyrene har endret seg lite over de siste tiårene. Kunnskapsdepartementet har derfor bestemt seg for å gjennomgå eksamenssystemet i lys av fagfornyelsen og utvikling av digital teknologi.

Rammen nedenfor viser kompetansebegrepet definert i Kunnskapsløftet LK06 og det nye kompetansebegrepet i fagfornyelsen LK20. Forståelse og evne til refleksjon og kritisk tenkning er tatt inn i den nye definisjonen, i tillegg er det å anvende kunnskaper og ferdigheter i kjente og ukjente situasjoner lagt til. Det nye begrepet stiller dermed et høyt krav og forutsetter kognitiv overførbarhet, som igjen kan knyttes til dybdelæring (Kunnskapsdepartementet, 2016). Elevene skal i tillegg bli satt i stand til å selv tilegne seg kunnskaper og ferdigheter.

> **Kompetansedefinisjon - Kunnskapsløftet LK06**
>
> Kompetanse er evnen til å løse oppgaver og mestre komplekse utfordringer. Elevene viser kompetanse i konkrete situasjoner ved å bruke kunnskaper og ferdigheter til å løse oppgaver.

Kompetansedefinisjon - Fagfornyelsen LK20

Kompetanse er å *tilegne seg* og anvende kunnskaper og ferdigheter til å mestre utfordringer og løse oppgaver i *kjente og ukjente* sammenhenger og situasjoner. Kompetanse innebærer *forståelse og evne til refleksjon og kritisk tenkning*.

I Meld. St. nr. 28 (2015–2016) står det blant annet at vurderingsformer og kvalitetsvurderingssystemet må støtte opp under en opplæring som skal legge større vekt på dybdelæring og systematisk progresjon. En viktig del av oppdraget til eksamensgruppa er å se på i hvilken grad og hvordan fagfornyelsens nye kompetansebegrep kan gjenspeiles i eksamensoppgavene. Det må være et mål å sørge for at elevene opplever at det er et meningsfullt samsvar mellom læreplanen, opplæringen og eksamensordningene. I oppdragsbrev 03-17 del 5 (2017) ber Kunnskapsdepartementet Udir om

1. å vurdere status og beskrive pågående prosesser i arbeidet med å utvikle kvalitet på standpunktvurdering og eksamen i lys av anmodningsvedtak XII i Stortingets innst. 19 S (2016–2017) og
2. å vurdere behovet for nye tiltak som kan bidra til økt kvalitet.

I sitt svarbrev datert 01.06.2017 foreslår Udir blant annet en gjennomgang av eksamensordningene og å styrke kunnskapsgrunnlaget vi har om eksamen.

Eksamensgruppa ble nedsatt i siste del av september 2018 og avslutter etter planen sitt arbeid januar 2020. Arbeidet består av fire delleveranser:

1. Foreløpig status og vurdering av kunnskapsgrunnlaget om eksamen
2. Vurdering av læreplangruppenes forslag til fagenes eksamensordninger samt råd til læreplangruppa om fagenes eksamensordninger
3. Vurdering av hvilken betydning fagfornyelsen og den teknologiske utviklingen bør ha for eksamensordningen, samt forslag til endringer i eksamensordningen grunnet fagfornyelsen og den teknologiske utviklingen
4. Forslag til eventuelle nye eksamensformer og videre retning for arbeidet med sluttvurdering

1.3 Arbeidsprosess og formål med rapporten

Eksamensgruppa har jobbet med delleveranse 1 om kunnskapsgrunnlaget fram til fristen 17.12.2018 (foreløpig versjon) og videreutviklet rapporten innen publisering 27.02.2019 (sluttversjon). Det er gjennomført tre møter i gruppa i forbindelse med kunnskapsgrunnlaget etter oppstartsmøtet i oktober, og skolesektoren har gitt tilbakemeldinger gjennom referansegruppa. I tillegg har det blitt gjennomført en fagfellevurdering.

Denne leveransen bør ses på som en foreløpig dokumentasjon av kunnskapsgrunnlaget og vil utvides gjennom de neste delleveransene. Råd om fagenes eksamensordninger vil leveres læreplangruppene innen mars 2019, og anbefalinger til endringer i eksamensordningen grunnet fagfornyelsen og den teknologiske utviklingen vil leveres for beslutning 15.05.2019. Eksamensgruppas endelige rapport til Kunnskapsdepartementet med forslag om nye eksamensformer og videre retning for arbeid med sluttvurdering sendes i 2020.

Generelt debatteres eksamensoppgaver og eksamensordninger jevnlig i det offentlige ordskiftet, men eksamenssystemet og eksamenskvaliteten har med unntak av få småstudier eller spørreundersøkelser ikke vært gjenstand for tilsvarende forskning som nasjonale prøver eller de store internasjonale undersøkelsene. Dagens grunnlag for å kunne vurdere om standpunkt og eksamen samlet sett gir pålitelig og relevant informasjon om elevens kompetanse, er derfor begrenset og lite systematisk.

Så vidt oss bekjent finnes det bare to tidligere rapporter som har berørt eksamen, men heller ikke her var den hovedsak. En rapport fra Sjaastad, Carlsen og Wollscheid (2016) har sett på hvorvidt timetallet i fag på videregående skole blir oppnådd. Forfatterne kom fram til at eksamensperioden var den faktoren som førte til mest bortfall av undervisning. Forfatterne kom fram til at eksamensperioden var den faktoren som førte til mest bortfall av undervisning. Denne rapporten pekte også på andre utfordringer med eksamen, særlig trekkordningen. Lied-utvalget har også sett på eksamen som en del av sitt oppdrag. Delinnstillingen beskriver eksamensordningen på detaljert vis, og utvalget har varslet at de vil gå nærmere inn på den i hovedinnstillingen, men at de ønsker å avvente anbefalingene som kommer fra eksamensgruppa (NOU 2018: 15).

Kunnskap om eksamenssystemet foreligger ikke samlet i dag, men vidt fordelt over forskjellige typer kilder. Formålet med denne rapporten er å samle det vi vet om eksamensprosedyrer, kvalitet og resultater, på en strukturert måte for å legge et bedre grunnlag for politiske beslutninger. Vi har valgt en teoretisk-deskriptiv tilnærming på den ene siden og en empirisk-analytisk tilnærming på den annen side fordi begge perspektivene tilfører viktig informasjon. I tråd med eksamensgruppas mandat konsentrerer vi oss om eksamenssystemet i sin helhet, mens det fagspesifikke innholdet i eksamen (angående dets endringer over tid, se for eksempel Nygård Arntzen, 2015; Smestad og Fossum, 2019) blir bearbeidet av de forskjellige læreplangruppene i fagfornyelsen.

Rapporten baserer seg på ulike typer kunnskap som hver er relatert til både fordeler og ulemper. Teoretisk og empirisk utledet kunnskap, som målingsteori og validitetsforskning, kunne gi oss solid informasjon om eksamenssystemet enn så lenge det finnes. Siden omfanget av denne typen kunnskap er svært begrenset, har vi i stor grad også dratt inn erfaringsbasert kunnskap selv om denne typen kunnskap er av varierende

kvalitet og omfang. Vi er bevisst på at rekkevidden er begrenset, likevel kan brukererfaring formidle oss mye innsikt i eksamens*praksis*. Den erfaringsbaserte kunnskapen består blant annet av ulike former for dokumentert brukerinnsikt, resultater fra spørringer og tekniske rapporter. Et slikt kunnskapsgrunnlag innebærer en viss usikkerhet, og vi kan bare svært forsiktig konkludere basert på det.

Eksamensgruppa har lagt hovedvekt på å oppsummere de tekniske og organisatoriske rapportene utarbeidet på oppdrag av Udir samt forskning om det norske eksamenssystemet. Internasjonal forskning har blitt inkludert der det er formålstjenlig, men generelt sett finnes det også her lite forskning om eksamen. De studiene som finnes, har stort sett blitt gjennomført fra to ulike perspektiver: et målingsperspektiv med vekt på studier om klassiske kvalitetskriterier, for eksempel konstruktvaliditet og sensorreliabilitet, eller et skoleperspektiv med vekt på studier om hvordan skolepraksisen blir berørt av eksamen. I et moderne målingsperspektiv kunne det siste betegnes som konsekvensvaliditet (AEA Europe, 2017). Vår ambisjon er å inkludere begge perspektivene likeverdig fordi de tilføyer kunnskapsgrunnlaget viktige momenter. Vi ønsker også å være konsistente i begrepsbruken, selv om dette gir utfordringer der hvor forståelsen av begrepene kan være ulik og sammensatt.

Eksamensgruppa blir samkjørt med arbeidet med fagfornyelsen for å sikre at ordninger for sluttvurderinger er på plass når læreplanene er klare. Denne rapporten skal fungere som kunnskapsgrunnlaget for videre arbeid i eksamensgruppa og læreplangruppene i fagfornyelsen. Mandatet til eksamensgruppa er knyttet til læreplanene i fagfornyelsen, som dekker fag i grunnskolen, de gjennomgående fagene i videregående opplæring og enkelte programfag [1]. Endringer i eksamensordningene i fellesfagene i videregående opplæring vil også omfatte elever på yrkesfaglige utdanningsprogrammer. Denne rapporten omtaler ikke eksamensordninger på de yrkesfaglige utdanningsprogrammene, men dette utelukker ikke at disse kan være tatt med i betraktningene når eksamenssystemet i sin helhet omtales i rapporten.

Noen av spørsmålene eksamensgruppa kommer til å berøre i de fire delleveransene, er: Hvordan kan eksamen som sluttvurdering støtte opp om og bidra til å realisere intensjonene med fagfornyelsen, overordnet del i læreplanene og det nye kompetansebegrepet? Er for eksempel dagens trekkordning forenlig med nytt kompetansebegrep og overordnet del? Og hvordan kan den teknologiske utviklingen gi oss nye muligheter for å gjennomføre, videreutvikle og vurdere eksamen?

Formålet med denne delleveransen er i første del å oppsummere både rammebetingelsene for dagens eksamenssystem, bakgrunnen for dem (kap. 2) og eksamens formål som dokumentert i lovverket, samt en beskrivelse av retningslinjer og prosedyrer (kap. 3). I andre del av rapporten følger en oppsummering av hva vi vet om eksamenssystemets kvalitet: Vi definerer sentrale vurderingsfaglige begreper (kap. 4) og går dypere inn i eksamens validitet (gyldighet) (kap. 5) og reliabilitet (pålitelighet) (kap. 6), vi går inn på forholdet mellom eksamen og standpunkt, som samlet utgjør sluttvurderingssystemet (kap. 7), og ser på hva det betyr å vurdere kompetanse i fag (kap. 8). I kapittel 9 tar vi for oss elevenes opplevelser av eksamen. Tredje del av rapporten representerer en utredning knyttet til fagfornyelsen og oppsummerer hvilke råd forskningen kan gi oss når det gjelder prøving av det utvidede kompetansebegrepet i fagfornyelsen (kap. 10), og hvilke muligheter og begrensninger digital teknologi innebærer (kap. 11). I tillegg oppsummerer vi kort hva vi vet om

lærerutdanningens bidrag til lærernes vurderingskompetanse (kap. 12). I et eget sluttkapittel (kap. 13) trekker vi noen konklusjoner og peker på sentrale problemstillinger og spørsmål som vi skal utrede videre i de kommende månedene.

---

1) Engelsk programfag, fremmedspråk programfag, matematikk programfag for realfag, matematikk program for samfunnsfag.

Del 1 - Bakgrunn og formål ved eksamenssystemet i grunnopplæringen

2. Framveksten av dagens eksamenssystem

Dette kapitlet trekker opp et historisk perspektiv på framveksten av eksamenssystemet med henblikk på å forstå eksamens formål og roller i dagens grunnopplæring. Norge har, som våre naboland Danmark og Sverige, lange tradisjoner for at opptak til videre utdanning baseres på eksamener forvaltet av lærerprofesjonen selv i et tett samspill med nasjonale myndigheter. Den sterke stillingen læreren har i dagens eksamenssystem i Norge, er resultatet av en utvikling fra 1700-tallet, der eksamenssystemet har gjennomgått en rekke endringer.

Utviklingen av reguleringene av standpunktvurderingen vies også oppmerksomhet i denne historiske framstillingen, som identifiserer fem sentrale utviklingstrekk som har vært med på å forme dagens eksamenssystem:

- (2.1) Framveksten av lærerprofesjonen i takt med sekulariseringen av samfunnet på 1700- og 1800-tallet, der prestens rolle gradvis ble avløst av lærere som ble anerkjent som kvalifisert til å uteksaminere elever
- (2.2) Utbyggingen av «enhetsskolen» gjennom 1800- og 1900-tallet, som skapte nye behov for tilpasset opplæring til nye elevgrupper og etter hvert til utvalg av elever til videre utdanning, med de behovene som dermed fulgte for nye teorier og teknologier for å bedre validiteten og reliabiliteten i lærernes vurderinger (se del 2 av denne rapporten for definisjoner av de viktigste målingsbegrepene)
- (2.3) Etterkrigstidens diskusjoner rundt normrelatert versus målrelatert karaktersetting og andre

vurderingsprinsipper (se del 2 for definisjoner), inkludert de skarpe frontene i 1960- og 1970-årenes karakterstrid

- (2.4) 1990-årenes reformer og ambisjonen om å vurdere elevenes «helhetlige kompetanse», inklusiv de bredere målene for opplæringen i den generelle læreplanen
- (2.5) Kunnskapsløftet (2006) og tilhørende forskriftsendringers presiseringer av at kun den faglige måloppnåelsen skal danne grunnlag for fastsetting av standpunktkarakterer, og at vurderingen skal være basert på læreplanens kompetansemål

2.1 Framveksten av lærerprofesjonen gjennom 1700- og 1800-tallet

Opphavet til skriftlige eksamenssystemer kan dateres så langt tilbake som Han-dynastiet i Kina, som varte fra år 206 f.Kr til 220 etter e.Kr (Eckstein og Noah, 1993, s. 2). På det europeiske kontinentet oppsto imidlertid eksamenssystemer først mot slutten av 1700-tallet og begynnelsen av 1800-tallet – en tid kjennetegnet ved nære forbindelser mellom skolen og kirken. Ifølge Jarning og Aaas (2008) sammenfaller examen artium i Norge og Danmark (lovregulert i 1808) og Studenteksamen i Sverige og Finland (lovregulert i 1824) med det tyske Abitur (lovregulert i Preussia i 1788) og det franske Baccalauréat (lovregulert i 1808). Examen artium ble kontrollert av Universitetet i København, og senere Universitetet i Oslo (etablert i 1811), og fungerte som en opptaksprøve til høyere utdanning. Det var gjennom eksamensvesenet gymnasutdanningens innhold og kvalitet ble regulert og kontrollert (Lundahl og Tveit, 2014).

Byskoleloven og landskoleloven av 1848 fastslo at det skulle være en offentlig eksaminasjon overvåket av en prest (Lysne, 1999, s. 68). En fundamental endring skjedde i 1884 da Johan Sverdrup ble statsråd og startet arbeidet med enhetsskolen. Samme året ble examen artium overført fra universitetet til godkjente gymnasskoler (ibid., s. 81). At gymnaslærerprofesjonen fikk full kontroll med examen artium, var resultatet av en lang kamp for økt anerkjennelse av gymnaslærernes faglige forutsetninger for å vite hvilke kunnskaper som behøvdes på universitetet (Lysne, 1999, s. 47). Denne anerkjennelsen representerte en milepæl for den rollen lærerprofesjonen har i dagens eksamenssystem.

2.2. Utbyggingen av «enhetsskolen» gjennom 1800- og 1900-tallet

På slutten av 1800-tallet og begynnelsen av 1900-tallet bar skolesystemet preg av et klasseskille der borgerskapets barn etter fem år gikk over i middelskolen. Folkeskolen, lovfestet i 1889, hadde derimot til

hensikt å «virkeliggjøre en skoleordning som forutsatte at alle barna hadde et skolefellesskaptil 10-årsalderen» (Dale, 2008, s. 54). Visjonen om «enhetsskolen» var at den skulle gi mulighet for lik adgang for alle til å få den beste utdanningen. Denne utdanningspolitiske strategien medførte større oppmerksomhet om resultatet av opplæringen.

Progressive pedagoger ble pådrivere for testing fordi kunnskap om undervisningens resultater og elevenes forutsetninger for undervisning var viktige fundamenter for reformer og tiltak i utbyggingen av enhetsskolen. Dale (2008, s. 54) konkluderer at «testforskningen som var på offensiven på 1920- og 1930-tallet, ga grunnlaget for forestillinger om tilpasset opplæring og differensiering» (ibid.). Rektor Hans Eitrems analyse av karaktergivningen ved examen artium i perioden 1920–1924 bidro til enda større oppmerksomhet om sammenlignbarheten i lærernes karaktersetting (Lysne, 1999, s. 106). Sammen med at Bernhof Ribskog, leder av Læreskolerådet fra 1936 til 1957, avdekket ganske store avvik mellom lærernes karaktersetting og eksamenskarakterene. Med en tendens til overforbruk av de gode karakterene (Lysne, 1999, s. 112–131) var dette noe av bakgrunnen for at man i Normalplanen av 1939 innførte en normalfordelingsnorm i karaktersettingen og et minstekrav som en del av grunnlaget for standpunktkaraktersettingen og karakterkravene til eksamen (Dale, 2008).

2.3 Uklare vurderingsprinsipper fra etterkrigstiden til 1990-årene

Etter annen verdenskrig var videre utbygging av enhetsskolen det sentrale utdanningspolitiske prosjektet. I motsetning til Sverige (Sejersted, 2005) holdt Norge fast ved den kontinentaleuropeiske eksamenstradisjonen (Lundahl og Tveit, 2014). Eksamen fikk en stadig viktigere funksjon som verktøy for statlig styring med lærernes praksis for karaktersetting. «Når framhaldsskolen og realskolen ble innlemmet i den 9-årige grunnskolen, økte kravet om normrelatert vurdering på 1960-tallet» (Dale, 2008, s. 234). Dale peker på at «både skolepolitikere og framstående pedagoger var opptatt av at karakterene både i grunnskolen og i gymnaset måtte kunne sammenlignes på landsbasis» (ibid., s. 235).

På denne bakgrunn ønsket man å innføre et sentralt prøveinstitutt «for produksjon og administrasjon av normerte prøver» (ibid.). Man ville etablere ordninger som kunne sikre «ensartet karaktersetting i alle skoler for å skape like konkurransevilkår» (ibid.). Forsøksrådets (1969) forsøk med «standardiserte prøver» fikk imidlertid ikke gjennomslag som ordinær vurderingsordning. Kort tid etter ble Rådet for videregående opplæring (RVO) etablert, der arbeidet med eksamen sto sentralt.

I 1960-årene hadde det bygget seg opp mye frustrasjon og misnøye med den veiledende fordelingsnormen for standpunktkarakterer i Normalplanen av 1939, blant annet fordi mange lærere fulgte normalfordelingen mer enn det var grunnlag for. Dette medførte at det var lettere å få gode karakterer i en klasse med en

lavtpresterende elevgruppe, og tilsvarende ble det i en høytpresterende elevgruppe vanskeligere å få gode karakterer – noe som var en viktig grunn til at motstanden mot karaktersystemet tiltok etter hvert som det fikk større betydning i 1960- og 1970-årene.

Motstanden hadde også bakgrunn i 1970-årenes politiske diskurs om desentralisering og læring utenfor skolen. Tilsvarende diskusjoner var ofte av ideologisk karakter. Dale (2008, s. 236) peker på en holdning om at «et viktig middel for å vektlegge læring utenfor skolen var å nedbygge det formelle vurderingssystemet i skolen». I en NOU fra 1974 ble det lagt vekt på at en karakterfri skole ville «øke regionenes muligheter til å utnytte skolen på egne premisser» (ibid.).

Avviklingen av karaktersetting i ungdomsskolen ble imidlertid ikke gjennomført, og derfor ble det nødvendig å finne en annen relateringsmåte enn normrelatert vurdering. Målrelatert vurdering innebærer at målbeskrivelsene utledes i mer eksakte vurderingskriterier som angir grad av måloppnåelse for hvert karakternivå (Lysne, 1999, s. 39). Fra 1968 var det blitt utviklet et målrelatert vurderingssystem for vurdering i videregående skole. Målformuleringene var imidlertid «ikke blitt fulgt opp av utredningsarbeid for å utvikle felles vurderingskriterier» (Dale, 2008, s. 236). Lysne (1999) omtaler systemet for karaktersetting i gymnasene som «tillempet målrelatering». Det ble lagt vekt på valg av lærestoff, læremidler og arbeidsmåter med sikte på at læreprosessen hos den enkelte elev i større grad skulle innrettes mot elevenes personlige vekst mot en bredde av mål for opplæringen. Ifølge Lysne (2004) innebar dette i praksis en «skjønnsmessig helhetsvurdering» (s. 120). I 1981 ble prinsippet om «tillempet målrelatering» innført også i grunnskolen (ibid., s. 197).

2.4 1990-årenes reformer (Reform 94, L97)

I 1990-årenes reformer ble prinsippet tillempet målrelatering formalisert ved at både de generelle læreplanmålene og læreplanmålene for fag inngikk i det man kalte «helhetlig kompetanse» (Kirke-, utdannings- og forskningsdepartementet (1996), s. 13). I prinsipper og retningslinjer for L97 ble det gitt anvisninger om organisering og arbeidsmåter slik at den generelle læreplanen og fagplanenes innhold ble knyttet sammen. Tilsvarende ble læreplanens generelle del forsøkt integrert i fagene i Reform 94 gjennom en egen del i fagplanen kalt «felles mål for faget». Eksamenssystemet forble uendret i denne tidsperioden, slik at diskusjonene dreide seg om vurderingsprinsipper.

Lysne (1999) konkluderer med at tillempet målrelatering ga «unødig stort rom for individuelt skjønn og kan derfor gi lite sammenlignbare karakterer» (ibid., s. 39). Prinsippet bidro til store forskjeller i standpunktvurderingen mellom skoler og mellom lærere. Flere undersøkelser har påvist at lærerne manglet felles referanserammer a˚ relatere til på grunn av uklare eller utydelige nasjonale kriterier. Vurderingene bar isteden fremdeles preg av en normalfordeling relatert til elevgruppa (Hovdhaugen mfl., 2014;

Kommunerevisjonen i Oslo, 2013; Lie, Hopfenbeck og Turmo, 2005; Prøitz og Spord Borgen, 2010; Steffensen og Ziade, 2009), selv om denne gruppestørrelsen er altfor liten til at normalfordelingsprinsippet kunne anvendes.

2.5 Kunnskapsløftet (2006) og etterfølgende presiseringer av regelverket

I St.meld. nr. 30 (2003–2004), Kultur for læring, ble det anerkjent at «vurdering av elevenes 'helhetlige kompetanse' har vært en medvirkende årsak til at bestemmelser om individuell vurdering og læreplaner kan fremstå som uklare, særlig i videregående opplæring» (Utdannings- og forskningsdepartementet, 2004, s. 39). Kunnskapsløftet presiserte at kun den faglige måloppnåelsen skal danne grunnlag for fastsetting av standpunktkarakterer, og at «vurderingen skal være standardbasert» (ibid., s. 40). Karakterskalaen gikk bort fra normrelaterte konnotasjoner til målrelatert vurdering, og veiledende kjennetegn på måloppnåelse ble innført (se del 2 av denne rapporten for flere detaljer). I et skolehistorisk perspektiv representerte den målorienterte strukturen i LK06-læreplanen et brudd med tidligere læreplaners orientering mot aktiviteter og innholdsbeskrivelser.

Oppsummert viser dette korte historiske risset over utviklingen av eksamenssystemet blant annet at lærerprofesjonens sterke oppslutning om eksamen kan forstås i lys av at den profesjonalitet, autoritet og legitimitet lærerne har gjennom å anerkjennes som kompetente til å vurdere kvaliteten på elevers prestasjoner og slik kontrollere adgang til videre utdanning og yrkesliv. Eksamenssystemet i sin helhet og de viktigste prosedyrene har vært relativt stabile de siste tiårene, men vurderingsprinsipper og -kriterier har vært mye omdiskutert og har endret seg over tid. Karaktersetting har gått bort fra normrelatert og over til et målrelatert vurderingsprinsipp, mens vurderingskriteriene har vært uklare gjennom lange perioder av 1900-tallet med negative konsekvenser for vurderingenes kvalitet.

3. Eksamens formål og organisasjon

Dette kapitlet oppsummerer først eksamens og eksamenssystemets formelle funksjoner slik de er definert i opplæringsloven med forskrifter, og vi analyserer kort hvilke perspektiver på eksamen disse formelle definisjonene gjenspeiler.[2] Deretter beskriver vi hvordan eksamen er organisert for å oppnå formålene. Eksamen er en del av sluttvurderingssystemet og har dermed en spesifikk ramme samt noen særegne egenskaper som trekkordningen og privatistordningen. Til slutt dokumenterer vi endringer som har funnet

sted de seneste årene innenfor dagens eksamensordning. Endringene er en følge av innspill fra brukere, embetene og fagmiljøer og reflekterer særlig to utfordringer: at eksamensordningen i et fag kan begrense elevenes mulighet til å vise sin sluttkompetanse på en god måte, og at den økende tilgangen på hjelpemidler krever nye diskusjoner om hva som skal vurderes, og på hvilke måter.

---

2) For en detaljert analyse av eksamens mer implisitte og ikke-formaliserte roller i praksis, se kapittel 5.3.

3.1 Eksamens formelle formål som en del av sluttvurderingssystemet

Om eksamen fra forskrift til opplæringslova

§ 3-17. Sluttvurdering i fag

- Sluttvurderinga skal gi informasjon om kompetansen til eleven, lærlingen, praksisbrevkandidaten og lærekandidaten ved avslutninga av opplæringa i fag i læreplanverket, jf. § 3-3.
- Sluttvurderingar i grunnskolen er standpunktkarakterar og eksamenskarakterar.
- Sluttvurderingar i vidaregåande opplæring er standpunktkarakterar, eksamenskarakterar og karakterar til fag-/sveineprøve, praksisbrevprøve og kompetanseprøve.
- Sluttvurderingar er enkeltvedtak og kan klagast på etter reglane i kapittel 5.
- Elevar i grunnopplæringa som har individuell opplæringsplan, skal vurderast etter dei samla kompetansemåla i læreplanen for faget, jf. § 3-3.

§ 3-25. Generelle føresegner

- Eksamen skal vere i samsvar med læreplanverket.
- Læreplanen i det enkelte faget fastset om og eventuelt når i opplæringsløpet det skal vere eksamen i faget. Det er òg fastsett i læreplanen for fag om eleven skal opp til eksamen i faget, eller om elevar kan trekkjast ut, kva slags eksamensform som skal nyttast, og om eksamen skal vere lokalt gitt eller sentralt gitt. Departementet fastset kor mange eksamenar det skal vere på 10. årstrinnet og på kvart årstrinn innanfor utdanningsprogramma eller programområda i vidaregåande opplæring.

- Elevar på ungdomstrinnet som avsluttar eit fag tidlegare enn det som følgjer av læreplanverket, jf. § 1-15, skal vere med i trekket til eksamen i faget det opplæringsåret faget blir avslutta. Om eleven blir trekt til eksamen etter første punktum, kjem denne eksamen i tillegg til dei eksamenar som departementet har fastsett at eleven skal ha på 10. årstrinn.
- Eksamen skal organiserast slik at eleven eller privatisten kan få vist kompetansen sin i faget.
- Eksamenskarakteren skal fastsetjast på individuelt grunnlag og gi uttrykk for kompetansen til eleven eller privatisten slik denne kjem fram på eksamen.
- Fylkeskommunen har plikt til å informere elevar og privatistar i vidaregåande opplæring om kva for reglar som gjeld for ny eksamen, utsett eksamen og særskild eksamen.
- Det er ikkje ny eksamen, utsett eksamen eller særskild eksamen i grunnskolen.
- For fag- og sveineprøva og kompetanseprøva gjeld reglane i § 3-48 til § 3-68.

Formålet med alle sluttvurderinger er å gi informasjon om kompetansen til eleven, lærlingen og lærekandidaten ved avslutningen av opplæringen i faget, jf. forskrift til opplæringsloven § 3-17. Eksamenskarakterer er i likhet med standpunktkarakterer en sluttvurdering. De skal altså begge vise kompetanse ved avslutningen av opplæringen, men samspillet mellom de ulike karakterene er ikke nærmere avklart i regelverket.

Eksamen inkludert eventuelle forberedelsesdeler skal være i samsvar med læreplanverket og organiseres slik at eleven får vist kompetansen sin i faget. Eksamenskarakteren skal fastsettes på individuelt grunnlag. Til tross for at standpunktkarakterer og eksamenskarakterer er likestilte sluttvurderingsformer er eksamen i enkelte tilfeller gitt en større betydning enn standpunktkarakteren. Dersom en elev har fått karakteren 1 som standpunktkarakter vil eleven bestå i faget om han eller hun får 2 eller høyere på eksamen. Dette følger av forskrift til opplæringsloven § 3-4. Faget vil aldri være bestått dersom eksamen ikke er bestått.

Eksamenskarakterer skal sammen med standpunktkarakterer føres på vitnemålet og utgjøre grunnlaget for opptak til videre utdanning og arbeidsliv. Dette kommer fram både av reglene om inntak til videregående opplæring i forskrift til opplæringsloven kap. 6, og av forskrift om opptak til høyere utdanning.

De to formålene med eksamen – sertifisering av kompetanse og rangering av søkere – er begge direkte relatert til elever. I tillegg pålegger forskrift til opplæringsloven skoleeiere å «medverke til å etablere administrative system og å innhente statistiske og andre opplysningar som trengst for å vurdere tilstanden og utviklinga innanfor opplæringa» (§2-2). Eksamensresultatene er en del av dette, selv om dette formålet er

vesentlig mindre tydelig, og heller kan sees på som en kvalitetssikrende funksjon på systemnivå.

> **Kvalitetsvurderingssystemet** skal bidra til kvalitetsutvikling, åpenhet og dialog om skolens virksomhet. Det skal også gi grunnlag for kvalitetsutvikling av den enkelte skole.
>
> Utdanningsspeilet inneholder tall og analyse av barnehage og grunnopplæring i Norge og beskriver bl.a. elevenes læringsresultater ved standpunkt- og eksamenskarakterer.
>
> Karaktergjennomsnitt og fordeling på enkeltkarakterer på nasjonalt nivå endrer seg normalt lite fra år til år. Karaktergjennomsnitt for fylker kan derimot variere en del mellom år, spesielt i fag med få elever. Eksamenskarakteren er et uttrykk for den kompetansen eleven har vist på eksamen. Ettersom oppgavene er ulike fra år til år, er det normalt med noe variasjon i karaktergjennomsnitt og karakterfordeling. Dette innebærer at eksamensresultatene ikke er direkte sammenlignbare fra år til år. De kan derfor ikke brukes til å si noe om endringer i prestasjoner på tvers av kull.
>
> Utdanningsdirektoratet (2018)

3.2 Rammer for sluttvurdering i regelverket

Vurdering av elevenes sluttkompetanse er regulert i forskrift til opplæringsloven kap.3, og tilsvarende for forskrift til friskolelov kap.3. Læreplanene er også forskrifter. Det er en tydelig sammenheng mellom læreplaner og forskrift til opplæringsloven for eksempel ved at grunnlaget for vurdering etter forskriften § 3-3 er kompetansemålene i fag. Regelverket legger bl.a. føringer for vurderingspraksis, ansvar og roller, og eksamensformene er beskrevet i læreplanene. Ansvaret er fordelt mellom nasjonale og lokale myndigheter og skoler/lærere.

Udir har ansvaret for utvikling, gjennomføring og forvaltning av det sammenhengende prøve- og vurderingssystemet. Innen systemet har direktoratet ansvaret for sentralt gitt skriftlig eksamen på 10. trinn og i videregående opplæring. Kommunen og fylkeskommunen har ansvaret for lokalt gitt eksamen i henholdsvis grunnskolen og videregående opplæring.

Lokalt gitt eksamen i videregående opplæring kan være muntlig, skriftlig, muntlig-praktisk og praktisk. I tillegg er det en tverrfaglig praktisk eksamen (der de felles programfagene inngår) for de yrkesfaglige

utdanningsprogrammene på vg2. I grunnskolen er det kun muntlig eksamen som er lokalt gitt. I tillegg legger for eksempel læreplanen i naturfag på 10. trinn føringer for muntlig eksamen med praktiske innslag.

Eksamensformer i grunnopplæringen

Sentralt gitt eksamen
Skriftlig

Lokalt gitt eksamen
Skriftlig
Muntlig
Praktisk
Muntlig-praktisk

Oversikten nedenfor viser antall fastsatte karakterer i skoleåret 2015–2016 fordelt på lokalt gitt eksamen, sentralt gitt eksamen og standpunkt. Det må tas forbehold om at uttrekksordningen på det enkelte trinn i videregående opplæring gjør at andelen elever som trekkes ut til sentralt gitt og lokalt gitt eksamen fra år til år, kan variere.

Fastsatte eksamens- og standpunktkarakterer 2015–2016 [2]

	10. trinn		VGO studieforb.[3]		VGO yrkesfag		VGO samlet	
	Antall	%	Antall	%	Antall	%	Antall	%
Lokalt gitt eksamen	57131	5	62994	7	42816	11	105826	8
Sentralt gitt eksamen	74435	7	139051	15	2678	1	141757	11
Standpunktkarakterer	932854	88	736711	78	359488	89	1084480	81
Sum	1064420	100	938756	100	404982	100	1332063	100

Oversikten gir et bilde av omfanget av karakterer som fastsettes per år innenfor de ulike kategoriene. Både i grunnskolen og i videregående opplæring gjennomfører hver elev et begrenset antall eksamener, og standpunktkarakterer utgjør derfor et stort flertall av karakterene på elevenes vitnemål. Oversikten synliggjør også at fylkeskommunen, kommunen og skolen/lærer har vesentlige roller og stort ansvar i dagens sluttvurderingssystem, inkludert å sikre at karakterene gir pålitelig og relevant informasjon om elevens kompetanse.

Udir har utviklet et rammeverk for sentralt gitt skriftlig eksamen (Utdanningsdirektoratet, 2018c). Hensikten er

å etablere et felles grunnlag for kvalitetssikring og kvalitetsstyring av Udirs arbeid med eksamen.

Organiseringen av arbeidet med sentralt gitt skriftlig eksamen

Arbeidet med sentralt gitt skriftlig eksamen i Norge omfatter og involverer flere ulike instanser med ulike ansvarsområder.

Utdanningsdirektoratet har ansvar for utvikling, gjennomføring og forvaltning av det sammenhengende prøve- og vurderingssystemet. Dette inkluderer sentralt gitt eksamen med tilhørende informasjons- og veiledningsmateriell. Utdanningsdirektoratet kan også annullere eksamen.

Fylkesmannen oppnevner sensorer etter forslag fra skoleleder/rektor i de forskjellige fagene på grunnskolen og i videregående opplæring. I grunnskolen og på norsk vg3 velger de også oppmenn* for sin region.Fylkesmannen har ansvar for å gjennomføre fellessensur og administrere klagebehandlingen.

Kommuner og fylkeskommuner har det lokale ansvaret for gjennomføringen av sentralt gitt skriftlig eksamen i grunnskolen og i den videregående opplæringen. Dette gjelder både sentralt gitt skriftlig eksamen og de lokalt gitte eksamenene (skriftlig, muntlig, praktisk, muntlig-praktisk). De har også ansvaret for trekk av fag og kandidater ut fra de rammene som er satt av Utdanningsdirektoratet. Dette gjelder også for friskolene.For yrkesfagene har fylkeskommunene i tillegg ansvar for sensur, oppnevning av sensorer og klage.

Fagnemndene utarbeider eksamensoppgaver og har ansvar for at oppgavene er i tråd med læreplanverket og relevante bestemmelser, på oppdrag fra og i samarbeid med Utdanningsdirektoratet.

Eksterne konsulenter gir tilbakemelding til fagnemndene og Utdanningsdirektoratet på oppgaveutkastet. Eksterne konsulenter skal tale kandidatens sak, og er en viktig del av kvalitetssikringen av eksamensoppgaver.

Sensorer har ansvar for å sensurere eksamensbesvarelser i tråd med kompetansemålene i læreplanen og kjennetegnene på måloppnåelse i eksamensveiledningen på oppdrag fra Fylkesmannen.

* I videregående opplæring brukes begrepet formøteledere

(Utdanningsdirektoratet 2018c)

Regelverket omtaler sentralt gitt og lokalt gitt eksamen noe ulikt når det gjelder hvor omfattende eksamen skal være. Udir fastsetter hvordan eksamen i det enkelte fag skal organiseres, og hvordan eksamensoppgaven skal være ved sentralt gitt eksamen. Dette står i§ 3-28, og er i tråd med de overordnede rammene nevnt over. Tidsrammen for sentralt gitt skriftlig eksamen er normalt på fem timer, etter § 3-28a. Lokalt gitt eksamen har flere eksamensformer med ulike tidsrammer (se tekstboks).

Noen rammer for *lokalt gitt eksamen* (jf. §§ 3-29, 3-30):

- Skriftlig eksamen – inntil 5 timer
- Muntlig eksamen – inntil 30 minutter per kandidat
- Muntlig-praktisk eksamen – inntil 45 minutter per kandidat
- Praktisk eksamen – inntil 5 timer

Grunnskolen har kun muntlig eksamen.

Muntlig eksamen skal gjennomføres med forberedelsesdel der kandidaten skal få oppgitt et tema eller problemstilling 24 timer før selve eksamen. Denne delen skal ikke inngå i vurderingsgrunnlaget.

Fylkeskommunen bestemmer om privatistene skal ha forberedelse.

Fylkeskommunen bestemmer om de andre lokalt gitte eksamenene skal gjennomføres med forberedelsesdel. Denne kan være inntil to dager og skal normalt ikke inngå i vurderingsgrunnlaget.

For lokalt gitt eksamen gjelder de samme overordnede rammene som for sentralt gitt eksamen (nevnt over). I tillegg er det føringer i forskriften om at lokalt gitt eksamen skal eksamineringen gi kandidaten mulighet til å vise kompetanse i så stor del av faget som mulig (jf. §§ 3-29, 3-30). Under eksamen skal kandidaten prøves i flere relevante deler av læreplanen enn det som kan leses direkte ut av en eventuell forberedelsesdel. I eksamenstiden ved muntlig eksamen skal kandidaten presentere temaet eller problemstillingen som hun eller han har forberedt i forberedelsesdelen.

Ved muntlig eksamen skal eksaminasjon og vurdering skje i «sanntid», der dialogen mellom kandidat og sensorene blir en viktig del av eksamen. På skriftlig eksamen vil utformingen av eksamensoppgave og sensur skje uavhengig av eksamensavviklingen og være etterprøvbare. Hvor stor del av læreplanen som prøves til eksamen, vil derfor kunne variere, blant annet ut fra eksamensform, hvordan kompetansemålene er formulert, og fagets egenart. Eksamensformene legger føringer for hvorvidt kandidaten skal vise sin

kompetanse muntlig, skriftlig eller praktisk. For enkelte læreplaner er skriftlige og muntlige ferdigheter en del av kompetansen som skal prøves (f.eks. språkfag), men for et flertall av læreplanene er ikke dette tilfelle. I noen tilfeller kan skriftlige eller muntlig ferdigheter påvirke elevens mulighet for å vise kompetansen i faget.

---

2) For en detaljert analyse av eksamens mer implisitte og ikke-formaliserte roller i praksis, se kapittel 5.3.

3) Inkluderer Vg3 allmennfaglig påbygning, studieforberedende Vg3 naturbruk og studieforberedende Vg3 medier og kommunikasjon. Privatisteksamen og fag- og svenneprøven er ikke inkludert.

3.3 Trekkordningen

Trekkordningen er omtalt i Meld. St. 20 (2012–2013): «Trekkordningen innebærer at elevene ikke skal ha eksamen i alle fag, men at de skal være eksamensforberedt i de fagene der eksamen er en mulig sluttvurdering ved siden av standpunktkarakteren» (Kunnskapsdepartementet, 2013, s. 65–66). Dette betyr at elevene kan komme opp i eksamen i alle fag, men at de ikke vet hvilket fag eller hvilken eksamensform før trekket til eksamen er offentliggjort. Unntaket fra trekkordningen er at alle elevene som tar studieforberedende utdanningsprogrammer eller påbygging til generell studiekompetanse i yrkesfaglige utdanningsprogrammer, skal opp i sentralt gitt skriftlig eksamen i norsk hovedmål.

Trekkordningen innebærer at elever blir fordelt. I all hovedsak skal dette være basert på tilfeldige utvalg, i tråd med et randomiseringsprinsipp. Lokalt må man imidlertid ta hensyn til at lærere har bestemte elevgrupper i flere fag, at de har flere grupper med eksamen på samme dag, og at de kan være oppnevnt som ekstern sensor ved andre skoler de samme dagene. En konsekvens av trekkordningen er at antall eksamener per elev på vitnemålet i videregående opplæring kan variere, noe som kan få utslag på antall karakterer som ligger til grunn for gjennomsnittsberegningen for opptak til høyere utdanning. For eksempel er det kun 20 prosent av elevene på vg1 som trekkes ut til eksamen (se oversikten nedenfor).

At trekkordningen bygger på et randomiseringsprinsipp, har blitt utfordret i lang tid blant annet av Elevorganisasjonen[4] . Elevene opplever dagens trekkordning som urettferdig fordi den ikke gir alle samme mulighet til å vise sin kompetanse, noe som er i strid med et av de to formelle hovedformålene beskrevet i lovverket (se avsnitt 3.1 og 3.2 ovenfor).

Trekkordningen på 10. trinn

Alle elever skal trekkes til en skriftlig sentralt gitt eksamen (matematikk, norsk eller engelsk) og en lokalt gitt muntlig eksamen.

Trekkordningen på studieforberedende

- *Vg1* Om lag 20 prosent av elevene skal trekkes ut til eksamen i ett fag, skriftlig, praktisk, muntlig eller muntlig-praktisk.
- *Vg2* Alle elever skal trekkes ut til eksamen i ett fag, skriftlig, praktisk, muntlig eller muntlig-praktisk.
- *Vg3* Alle elever skal opp til obligatorisk skriftlig eksamen i hovedmålet sitt: norsk hovedmål eller samisk som førstespråk. For alle elever er eksamen i norsk sidemål et trekkfag. I tillegg til obligatorisk eksamen i norsk hovedmål eller samisk som førstespråk skal elevene innenfor studiespesialiserende utdanningsprogram trekkes ut i to skriftlige programfag i tillegg til én muntlig, praktisk eller muntlig-praktisk eksamen.

Trekkordningen i yrkesfaglige utdanningsprogrammer

- *Vg1 og vg2* Alle elever på vg2 skal opp til en obligatorisk tverrfaglig eksamen i programfag. I tillegg skal om lag 20 prosent av elevene på vg1 og vg2 trekkes ut til eksamen i ett fag. Andelen på 20 prosent skal ses over en toårsperiode.
- *Vg3 påbygging til generell studiekompetanse:* I tillegg til den obligatoriske eksamenen i norsk skal elevene trekkes ut til én skriftlig og én muntlig, praktisk eller muntlig-praktisk. For alle elever er eksamen i norsk sidemål et trekkfag.

I tillegg forbindes selve trekkdagen gjerne med press og stress hos elevene. I 2017 leverte en partssammensatt arbeidsgruppe nedsatt av Kunnskapsdepartementet en rapport som skulle vurdere mulige organiseringer av skoleåret i lys av utfordringer knyttet til undervisningstid, eksamen og eksamensforberedelser (Kunnskapsdepartementet, 2017). Udir har på oppdrag av Kunnskapsdepartementet sendt ut fire forslag til endringer på høring som en oppfølging av dette arbeidet. Høringsfristen er 21. mars 2019.

Konklusjonen til arbeidsgruppa var blant annet at en lengre forberedelsesperiode mellom offentliggjøring av trekket og eksamensdagen er viktig fordi ukene før eksamen kan planlegges på en bedre måte. De løftet også problemstillingene: Hva ville skjedd dersom elevene og lærerne i større grad kjenner til på forhånd hvilke fag elevene skal ha eksamen i? Er det argumenter som taler for at trekket til de skriftlige eksamenene bør være kjent lengre tid på forhånd dersom det gjør skoleåret lettere å planlegge slik at årstimetallet i fagene blir oppfylt? På den ene siden vil elevene få bedre tid til å forberede og fordype seg dersom trekket gjøres tidligere i skoleåret. På den annen side kan et tidligere trekk indirekte bidra til at eksamen vil styre undervisningen i enda større grad enn i dag. Dersom det er kjent hvilke elever som skal prøves i hvilke fag, tidlig i skoleåret, kan dette påvirke de faglige prioriteringene til elever og lærere og ha skadelig effekt på undervisningstid for

de fagene hvor elevene ikke trekkes til eksamen.

---

4) https://www.utdanningsnytt.no/nyheter/2018/november/elevorganisasjonen-vil-endre-eksamensordningen/

3.4 Privatistordningen

Privatisteksamen er en alternativ vei til yrkeskompetanse eller studiekompetanse. Det opprinnelige formålet med ordningen var at det skal være et tilbud om å dokumentere kompetanse i et fag man ikke tidligere har fått opplæring eller sluttvurdering i eller ønsket å forbedre karakterer i. Den siste gruppa blir kalt forbedringsprivatister, den førstnevnte førstegangsprivatister. Det er også elever som slutter med enkeltfag underveis eller tar opp fag opp igjen, samtidig som de avslutter opplæringen i faget som elev. Ordningen stammer fra før ungdom fikk lovfestet rett til videregående opplæring gjennom Reform 94.

Privatisteksamen er identisk med eksamenene i grunnopplæringen innholdsmessig, og de er sentralt og lokalt gitte, skriftlige og muntlige. Oppmelding til privatisteksamen i videregående opplæring i Oslo fra 2017 viser at flertallet av privatistene tar én eksamen, og nesten 40 prosent har meldt seg opp til to eller flere eksamener (Kunnskapsdepartementet, 2017).

Nesman og Kovač (2016) peker på at privatistordningen kan representere en god løsning for enkelte grupper, men at privatistene som gruppe har endret seg over tid. Deres studie viser at privatister er en sammensatt gruppe, der flertallet (52 %) i dag tar eksamen for å forbedre karakterer i allerede bestått fag. En stor del av privatistene er derfor samtidig elever i ordinær opplæring. Elever på ungdomstrinnet som har tilstrekkelig kompetanse i faget til å kunne følge opplæringen på videregående nivå, kan forsere fag etter forskrift til opplæringslova § 1-15. Privatistordningen medfører i tillegg noen uheldige bivirkninger. Mange av de som melder seg opp til privatisteksamen møter ikke fram; omvendt velger enkelte elever å ikke delta i opplæringen, men tar fag som privatist mens de har status som elever.

Endringene og utfordringene kan være grunn til å se på hvordan privatistordningen fungerer i praksis, på nytt. I svar på oppdragsbrev 13-12 til departementet foreslo direktoratet noen tiltak som vil bidra til å redusere omfanget av ordningen og bevare det opprinnelige formålet. Eksempler på tiltak er en økning i privatistgebyret og strengere krav til førstegangsvitnemål.

Noen endringer har allerede blitt iverksatt de seneste årene. Fra 01.08.2018 er privatistordningen i programfag i skole innen yrkesfaglige utdanningsprogrammer endret, noe som innebærer at alle privatister nå kan ta eksamen i enkelte programfag. Tidligere måtte privatistene ta eksamen i flere fag enn i de fagene de manglet karakter. Tall fra 2017 viste at nesten 40 prosent av privatistene skulle ta to eller flere eksamener

våren 2017.

3.5 Utvikling og endringer i eksamen

Administrative og innholdsmessige forhold ved eksamen videreutvikles kontinuerlig på bakgrunn av innspill fra brukere, embetene og fagmiljøer. De senere årene er det særlig to utfordringer som peker seg ut: at selve eksamensordningen i et fag kan begrense elevenes mulighet til å vise sin sluttkompetanse på en god måte, og at den økende tilgangen på hjelpemidler krever nye diskusjoner om hva som skal vurderes, og på hvilke måter. Udir har derfor foretatt noen endringer i eksamensordningen i enkelte fag, som blir beskrevet i dette avsnittet.

Om hjelpemidler til eksamen i forskriften

Alle hjelpemidler har vært tillatt under eksamen i de fleste fag etter Kunnskapsløftet, men det er forskjeller mellom hva som er tillatt til skriftlig og muntlig eksamen. Det er Udir som fastsetter hvilke hjelpemidler som er tillatt å bruke i hvert fag ved sentralt gitt eksamen, og skoleeier som fastsetter ved lokalt gitt eksamen (se § 3-31 i forskrift til opplæringslova, § 3-29 i forskrift til friskolelova). Ved muntlig eksamen er notatene til eleven eller privatisten fra forberedelsedelen eneste tillatte hjelpemiddel. Regelverket sier samtidig at tillatte hjelpemidler ikke må svekke grunnlaget for å vurdere kompetansen til eleven eller privatisten (§ 3-31). Direktoratet har fått tilbakemeldinger gjennom sensorrapporter fra og på sensorskoleringer for sentralt gitt skriftlig eksamen om at en del av sensorene synes det kan være utfordrende å vurdere i hvilken grad elevene selv har skrevet tekstene når elevene ikke viser til kilder de (antakelig) har brukt i besvarelsen sin.

Digitale hjelpemidler til eksamen

Bruken av digitale læremidler i opplæringen har økt, og dette har betydning for hjelpemidlene elevene kan bruke på eksamen. I flere fag er for eksempel det å kunne innhente, vurdere og bruke kilder på en relevant og etterprøvbar måte en del av kompetansen i faget, og dette inkluderer kilder fra internett.

Siden 2015 har elevene i tillegg til andre hjelpemidler hatt tilgang til et utvalg nettbaserte hjelpemidler på sentralt gitt eksamen. I 2017 kom en presisering for å bidra til økt likebehandling av kandidater innenfor samme fylkeskommune. Det er nå obligatorisk for fylkeskommunene og kommunene å tilby et utvalg nettbaserte hjelpemidler, og alle kandidater skal ha tilgang til de samme nettbaserte hjelpemidlene på eksamen i samme fylke. Det er som tidligere lagt vekt på at de nettbaserte hjelpemidlene skal være kjent for elevene fra opplæringen, slik at utvalget må skje i et samarbeid mellom skoler og skoleeier.

Forsøk med tilgang til åpent internett på eksamen

Det er også gjennomført et forsøk med åpent internett på selve eksamensdagen i fag der dette er relevant. I sluttrapporten fra forsøket med tilgang til internett på eksamen (2012–2015) skriver Rambøll (2015) at flertallet av de involverte elevene og lærerne var tilfredse med eksamensformen, og at eksamensformen bidro til å fremme nye, relevante pedagogiske praksiser og kompetanser i opplæringen. Eksamensformen medførte få tekniske og/eller praktiske utfordringer. Det framgår i sluttrapporten at det var få indikasjoner på at omfanget av fusk og plagiat øker, eller på at eksamensformen har en påviselig innvirkning på elevenes eksamensresultater, både i positiv eller negativ retning. Tilgangen til internett ble utvidet til å inkludere alle kandidater ved alle skoler i et utvalg fag fra våren 2018.

Sensorene var mer kritiske til bruken av åpent internett. Seks av ti sensorer syntes at internett på eksamen er godt egnet til å vurdere elevens fagkompetanse, men bare tre av ti sensorer mente ordningen bør videreføres, og enda færre (16 %) ønsket det som normalordning i alle fag. Selv om sensorene var delt i hvorvidt dette er en god eksamensordning i fagene, så de i liten grad forskjell på besvarelser med og uten internettilgang. 73 prosent av sensorene og 51 prosent av lærerne mente imidlertid at det er lettere å fuske på eksamen med tilgang til internett enn ved andre eksamener som gjennomføres på PC. Lærerne ga også tydelig uttrykk for at høytpresterende elever har mye større nytte av åpent internett enn det lavtpresterende elever har (Rambøll, 2015).

En metodisk begrensning det er viktig å framheve med disse studiene, er at skolene selv valgte å bli med på forsøket. Det vil si at hele evalueringen av tilgang til åpent internett under eksamen er gjennomført på skoler som på eget initiativ ønsket å teste ut en slik ordning. Skolene kan dermed sannsynligvis anses som et positivt utvalg av alle skolene, noe som betyr at funnene ikke nødvendigvis kan generaliseres til å gjelde andre skoler.

Todelt eksamen – eksempel fra matematikk

Todelt eksamen i matematikk, kjemi, fysikk og biologi ble innført med Kunnskapsløftet i 2008. Todelt eksamen i samfunnsøkonomi kom noe senere (2013) på bakgrunn av en evaluering i 2010 utført av Rambøll og Institutt for lærerutdanning og skoleforskning (ILS). Å innføre en todelt eksamen ble begrunnet med at det var utfordrende både å prøve og å vurdere elevenes brede kompetanse i fagene der ingen eller alle hjelpemidler ble tillatt på eksamen (med unntak av internett og verktøy som åpner for kommunikasjon).

Endringen la dermed til rette for en mer helhetlig prøving av elevenes kompetanse og burde kunne gi et bedre grunnlag for sensuren. I matematikk ble det for eksempel, med én del uten hjelpemidler (del 1) og én del med hjelpemidler (del 2), mulig å prøve elever i hode- eller overslagsregning i del 1, mens elevene blant annet kunne benytte seg av digitale verktøy for å løse andre typer oppgaver av mer kompleks art i del 2.

En evaluering av eksamen i matematikk på 10. trinn våren 2018 viser at eksamenskvaliteten ble gjennomgående vurdert som godt når det gjaldt samsvar mellom eksamensoppgaver og opplæringen, at oppgavene var forståelige og passe vanskelige, og at omfanget tilsvarte tiden elevene hadde til rådighet (Bjørnset mfl., 2018). Det kom imidlertid også fram at elever har ulik tilgang til og opplæring i digitale verktøy både på skolen og i hjemmet. Dette funnet kan bety at elevgrupper med større tilgang til og mer opplæring i

digitale verktøy hadde bedre forutsetninger for å lykkes på eksamen enn elever som hadde mindre gode muligheter. Bjørnset mfl. (2018) peker på dette som «en sentral ulikhetsskapende mekanisme», riktignok uten at de kunne konkludere noe om hvilken rolle mekanismen muligens hadde spilt for årets eksamen gitt datagrunnlaget.

Nivå I fremmedspråk

Det ble gjennomført en utprøving av ny eksamensordning i noen utvalgte fremmedspråk nivå I våren 2015, høsten 2015 og våren 2016 i Finnmark, Rogaland, Sør-Trøndelag og Troms. I utprøvingen ble eksamensordningen endret fra skriftlig 5-timers eksamen til en kombinert skriftlig-muntlig. Bakgrunnen for forslaget til utprøving var både et stort og systematisk avvik mellom standpunktkarakterer og eksamenskarakterer over tid og et ønske om å prøve ut en eksamensform som kunne gi elever en bedre mulighet til å vise sin kompetanse i faget, og som prøvde flere ferdigheter enn kun lesing og skriving. Både lærere og elever ga tilbakemelding om at de syntes at modellen bidro til at elevene fikk vist mer av sin samlede kompetanse i fremmedspråk. Karakterene som ble rapportert inn, viste ingen stor økning i karaktersnittet, men karakterene var noe bedre enn til den ordinære 5-timers skriftligeksamenen.

Modellen som ble prøvd ut i fremmedspråk nivå I, ville medført en del administrative og økonomiske utfordringer, spesielt for eksamenskontorene som skal gjennomføre eksamen for privatister. Endring i eksamensordningen ville ha betydd en økning fra en til to eksamener for alle privatister i fremmedspråk. Selv om både lærere og elever var positive til den nye modellen, ble eksamensordningen derfor ikke endret etter endt forsøk.

Lokalt gitt eksamen

Muntlig eksamen er lokalt gitt. Muntlige eksamener utgjør en større del av det totale antall eksamener en elev skal gjennom i sitt utdanningsløp. Alle elever har minst to muntlige eksamener i løpet av sin grunnopplæring. Reglene for lokalt gitt eksamen ble endret ved forskrift 26.september 2013. Formålet med endringene var blant annet å tydeliggjøre reglene for muntlig eksamen og sørge for en mer enhetlig nasjonal praksis. Reglene skal legge bedre til rette for at eksamen oppleves som forutsigbar og rettferdig for alle elever.

Endringene medførte en del henvendelser om hva som skulle inngå i vurderingsgrunnlaget. Tilbakemeldinger fra sektoren og de fylkeskommunale samlingene om eksamen tyder på at det fortsatt er skoler som strever med å forstå hvordan de skal tolke bestemmelsene slik de er beskrevet i forskriften, «førebuingsdelen skal ikkje inngå i vurderingsgrunnlaget», og i rundskriv Udir-2-2014, «Det er den kompetansen eleven viser under selve eksamenen som sensorene skal vurdere. De notatene som eleven har produsert i forberedelsedelen, for eksempel presentasjonen, er ikke en del av vurderingsgrunnlaget». Det siste har blitt presisert i avsnittet «Vurderingsgrunnlag», der det står at «den faglige kompetansen eleven viser gjennom måten temaet/problemstillingen presenteres på, er også en del av vurderingsgrunnlaget av elevens samlede kompetanse». Imidlertid har det vist seg at begrepet «måten» er gjenstand for tolkning og ikke-enhetlig praksis på tvers av skoler.

Utprøving av lokalt gitt eksamen i praktiske og estetiske fag

Som en del av Meld. St. 28 (2015–2016), *Fag – Fordypning – Forståelse,* ønsker Kunnskapsdepartementet å vurdere om de praktiske og estetiske fagene mat og helse, kroppsøving, kunst og håndverk og musikk skal bli en del av trekkfagsordningen til lokalt gitt eksamen i 10. trinn. Utdanningsdirektoratet har derfor fått i oppdrag å organisere en utprøving med en eksamenslignende prøve i disse fagene.

Formålet med utprøvingen er å innhente erfaringer for å kunne ta en beslutning om fagene skal bli en del av trekkfagsordningen, og om hvilken type eksamensform dette eventuelt skal bli:

- muntlig eksamen, med praktiske innslag, med 24 timers forberedelse og eksaminering på 30 minutter per elev
- muntlig-praktisk eksamen med forberedelsestid på inntil 48 timer og eksaminering på 45 minutter per elev

Skoleåret 2017–2018 deltok 4 fylker med til sammen 16 skoler i utprøvingen. Skoleåret 2018–2019 deltok 5 fylker med til sammen 19 skoler. De fikk selv velge eksamensform for utprøving.

Til sammen 40 lærere prøvde ut en *eksamenslignende prøve* i de praktiske og estetiske fagene. Ti lærere deltok i utprøvingen i hvert av fagene, og halvparten prøvde hver av de to eksamensformene. Skoler og lærere som valgte å delta i 2017–2018, er positive til utprøvingen. I en Questback-undersøkelse i juni i 2018 svarte lærerne at både de og elever er fornøyde med gjennomføringen. De hadde fått til en god fordeling mellom en praktisk og en muntlig del, og elevene fikk vist både praktisk og muntlig kompetanse på en god måte. I tillegg pekte flere på at utprøvingen har ført til en holdningsendring til fagene ved skolene ved at elever og lærere snakker mer om hvor viktige de praktiske og estetiske fagene er. Utprøvingen har tydelig ført til en større bevissthet om kompetansemålene i læreplanen og om vurdering. 14 av de 16 lærerne som gjennomførte utprøvingen, mente at en eksamen kan bidra til å styrke fagenes status.

Identifiserte utfordringer:

- Oppgaver som skal gi elevene anledning til å vise kompetanse, og vurderingen i selve eksamenssituasjonen
- Regelverk, tidspunkt for å sette standpunkt og trekk til eksamen

Del 2 - Kvalitet i dagens eksamenssystem

4. Sentrale begreper: kvalitetskriterier og relateringsprinsipper

Med dette kapitlet søker vi å klargjøre noen sentrale begreper som kan anvendes i utviklingen av det norske eksamenssystemet. I kapitlene deretter ser vi på kunnskapsgrunnlaget vi har, ut fra de forskjellige begrepene. Allerede i NOU 2015: 8 ble det etterspurt en gjennomgang av hvordan standpunktvurderingen og eksamenssystemet samlet sett kan gi rettferdig og relevant informasjon om elevenes kompetanse i et fag. Rapporten pekte på at lærere og sensorer trenger støtte i sine vurderinger gjennom tydelige mål, vurderingskriterier, veiledning og kvalitetssikring. Stoltenberg-utvalget anbefalte også å stille strengere kvalitetskrav til utformingen og utprøvingen av eksamensoppgaver (NOU 2019: 3). En slik vurdering av kvaliteten på sluttvurderingen bør ta utgangspunkt i testteoretiske og vurderingsfaglige begreper som validitet og reliabilitet, begreper som ivaretar ulike dimensjoner av kvalitet. Det er viktig å ha en helhetlig tilnærming som i tillegg ivaretar samsvar i kvalitetskriteriene.

En stor utfordring for eksamen, hvor oppgaver må være hemmelige før gjennomføring, er at det er svært vanskelig å vite om en eksamensoppgave har de ønskede egenskapene før den tas i bruk. I Nederland blir for eksempel oppgaver for neste år pilotert i eksamen for et utvalg elever året før. Det finnes altså mulige løsninger for å få testet en eksamen på forhånd, men da trengs det konsens mellom alle involverte parter. For å sikre eksamenskvaliteten er det altså nødvendig å drøfte slike overordnede spørsmål i tillegg til å se på enkeltkriterier.

4.1 Validitet (gyldighet)

Validitet, eller gyldighet, bør regnes som det mest sentrale vurderingsteoretiske begrepet ved eksamen. Forskning påpeker at forståelsen av validitet er avhengig av ulike aktørers fortolkninger:

- hvorvidt en tolkning, avgjørelse eller handling er fornuftig,
- hva slags bevis, resonnementer eller kriterier som gjelder for å bedømme hvor fornuftig en tolkning er,
- hvordan vi kan utvikle mer fornuftige fortolkninger, avgjørelser eller handlinger (Moss, Girard og

Haniford, 2006)

Å validere en prøve eller eksamen innebærer å utvikle en argumentasjon om hva slags bevis som skal regnes som gyldige, og om hvordan tolkningen skal foregå (Markus og Borsboom, 2013). Det er nesten umulig å bedømme en eksamens kvalitet eller «rettferdighet» på generelt nivå, da dette alltid må diskuteres i lys av formålet. *Validering* er derfor en sentral prosess der man undersøker og dokumenterer en prøves gyldighet i lys av prøvens formål (Kane, 2015). Av dette følger at en endring i prøvens kontekst eller formål utløser behov for en ny validering av prøven. Dersom en prøve eller eksamen skal ha flere anvendelsesområder, må den valideres for hvert av disse formålene. Pellegrino, Chudowsky, Glaser og National Resarch Council (U.S.) (2001) påpeker at jo flere formål en enkelt prøve eller eksamen har, dess sterkere blir hvert enkelt formål truet.

Uintenderte negative konsekvenser av tester og prøver for visse befolkningsgrupper (f.eks. minoritetsspråklige elever), uønskede systemiske effekter (f.eks. stress, engstelse) og tilsiktede eller utilsiktede tilbakevirkende («washback») effekter på opplæringen er en del av diskusjonen om validitet (konsekvensvaliditet; Kane, 2015). Å sikre validitet i eksamen innebærer å ha et blikk på hele prosessen – fra utvikling av oppgavene, via administreringen av eksamen og tolkningen av resultatene, til måten disse tolkningene blir anvendt på.

Oppgaveutvikling er et veletablert område i vurderingsforskning, og det finnes gode rammeverk som beskriver viktige kvalitetskriterier for prøveoppgaver og hvilke trinn som bør gjennomgås under utviklingsprosessen (se f.eks. AEA Europe, 2017; Wilson, 2005). Prøveutviklingen begynner med tydelige definisjoner av prøveinnholdet og vurderingskriterier og inkluderer piloteringer for å sikre at en prøve har de ønskede egenskapene *før* den implementeres, særlig gjelder dette i tilfelle *high-stakes*-prøver. For muligheter når det gjelder kompetanseprøving, se kapittel 9.

I et forsøk på å konstruere en teoretisk modell også for de neste trinnene i kvalitetssikringsprosessen har vurderingsforskere utviklet den såkalte validitetskjeden (Crooks, Kane og Cohen, 1996). De åtte leddene er beskrevet og tilpasset her:

1. *Administrering* av oppgavene som elevene skal gjøre på eksamen
2. *Skåring* av elevenes prestasjoner på eksamensoppgavene
3. *Aggregering* av resultater fra enkeltoppgaver for å beregne del- eller totalskår på eksamen
4. *Generalisering* fra konkrete oppgaver og resultater på prøven til målområdet som skal vurderes (f.eks. ved å drøfte hva én lengre skriveoppgave eller en samling mindre skriveoppgaver kan si om forventningene til elevens skrivekompetanse slik de er uttrykt i kompetansemålene)

5. *Ekstrapolering* fra målene som vurderes på eksamen, til et større målområde (f.eks. generell skrivekompetanse), som omfatter alle oppgaver som kunne være relevante innenfor dette større området

6. *Evaluering* av elevens prestasjon. I eksamenssammenheng vil dette normalt innebære å treffe en beslutning om karakterresultat og eventuelt utforme en begrunnelse for beslutningen.

7. *Beslutning* om hvilke handlinger eller tiltak som er relevante i lys av resultatet. For eksempel kan en elev bestemme seg for å klage på en karakter, eller lærere og skoleledere kan bestemme seg for å se nærmere på skolens eksamenspraksis innenfor et visst område.

8. *Virkning* på elever og andre som blir berørt av eksamenspraksisens prosess, tolkninger og beslutninger (Crooks mfl., 1996)

Typiske trusler mot validitet i kjedens ulike ledd kan være: Noen elever får hjelp av lærere til å løse oppgavene i eksamenssituasjonen, andre ikke (administrering); lærere vektlegger det som er lett å vurdere i skåringen uten at mer komplekse dimensjoner av elevens prestasjon tillegges vekt (skåring); resultater fra svært ulike oppgavetyper sammenfattes på uheldig vis (aggregering); eksamen inneholder få oppgaver slik at man egentlig bare tester et lite utsnitt av elevens kompetanse (generalisering); eksamen inneholder ingen oppgaver fra viktige deler av målområdet (ekstrapolering); elevens prestasjon bedømmes ut fra læreplanens kompetansemål, men uten at det foreligger bevis for at eleven mestrer disse (evaluering); kravene som ligger til grunn for oppfølgingstiltak i etterkant av eksamen, er altfor høye eller lave (beslutning); eksamensprosessen påvirker mange elevers utvikling i negativ retning (virkning). Den som skal kvalitetssikre en test eller en eksamen, bør evaluere hva de svakeste leddene er, og forsøke å styrke disse.

4.2 Reliabilitet (pålitelighet)

Reliabilitet, eller pålitelighet, viser til hvorvidt resultatene fra gjentatte vurderinger samsvarer (Pellegrino mfl., 2001). Det kan dreie seg om flere vurderinger av samme konstruktet innen en eksamensprøve eller om vurderinger av en eksamensoppgave av flere sensorer. Reliabilitet regnes som et nødvendig, men ikke tilstrekkelig, vilkår for validitet.

Høy reliabilitet er en forutsetning for kvalitet i vurderingen slik at tilfeldigheter kan unngås. En elevs eksamensresultat, som i den tallfestede informasjonen om elevens kompetanse, bør være så uavhengig som mulig av sensoren som vurderer hans eller hennes eksamensbesvarelse, av eksamensformen som er brukt, av innholdet som har blitt valgt i nettopp denne eksamenen, eller av tidspunktet eksamenen har funnet sted.

Kravet er at verken andre sensorer eller en gjentakelse av eksamen en annen dag, med andre oppgaver eller andre eksamensformer, ville lede til et annet resultat, som i annen tallfestet informasjon om elevens kompetanse.

Det er opplagt at resultatvariasjon til en viss grad er uunngåelig. Denne variasjonen kalles målingsfeil. Jo større konsekvensene av et resultat er for en elev – noe som gjelder i høyeste grad til sluttvurderingen gitt at vitnemålet er grunnlag for opptak til høyere utdanning og yrkesliv – dess viktigere er det å redusere målingsfeil og å øke reliabiliteten så mye som mulig. Enkeltoppgaver er ofte lite reliable. Dette gjelder både til ustandardiserte og standardiserte oppgaver. Et godt råd er derfor å bruke så mange og så forskjellige typer oppgaver som mulig og å la disse vurderes av forskjellige sensorer.

En utfordring i en vurdering med begrenset tidsramme er at ambisjoner om å øke reliabiliteten kan medføre at oppgavene blir innsnevret i utforming og nedslagsfelt istedenfor å utvide antall og type oppgaver – med andre ord at vektleggingen av å sikre konsistent informasjon blir viktigere enn vektleggingen av å samle inn bevis for brede og viktige læringsmål (Broadfoot, 2007). Denne problemstillingen peker på nødvendigheten å ha et overordnet kvalitetsrammeverk, systematiske rutiner for å overvåke kvalitet og at resultatene så må bearbeides til dokumentasjon som gjøres tilgjengelig. For å utrede reliabilitet trengs det data på det mest konkrete nivået som mulig. Vanligvis ville det bety å lagre data fra sensuren på elevnivå om hver enkelt eksamensoppgave og hvert enkelt vurderingskriterium.

4.3 Rettferdighet (fairness)

Rettferdighet viser til at alle elever må ha den samme sjansen til å vise kompetansene sine under eksamen. Teknisk sett betyr det at eksamen er fri fra systematiske skjevheter for gruppen som skal ta prøven. Det betyr at eksamen ikke skal påvirkes av variabler som kjønn, språkbakgrunn, funksjonsgrad, bosted og lignende.

Det engelske *fairness* brukes om en rekke problemstillinger som kan knyttes til dette: hvorvidt eksamensoppgavene ikke gir fordeler til enkelte elevgrupper, hvorvidt alle elever blir likt behandlet i eksamineringsprosessen, og hvorvidt elever har hatt tilgang til å lære det de blir testet i (Pellegrino mfl., 2001). Også en rekke andre faktorer kan påvirke hvor rettferdig en eksamen er. For eksempel kan elevenes resultater påvirkes av språklige ferdigheter, motivasjon, tretthet, testengstelse, forhold i det fysiske miljøet i gjennomføringen eller ulik grad av eller uetisk forberedelse til eksamen (Haladyna og Downing, 2005).

4.4 Relateringsprinsipper (norm-, mål-, standard- og individrelatert vurdering)

Som omtalt i kapittel 2 har det teoretiske grunnlaget for vurdering historisk sett blitt utviklet fra normrelatering til målrelatering – i mange utdanningskontekster nylig videreutviklet til standardrelatering. Forskjellen mellom prinsippene går ut på hva man *relaterer* vurderingen *til*, altså hva man sammenligner med (Wiliam, 1996).

Ved en *normrelatert vurdering* blir et eksamenssvar fra en elev sammenlignet med andre elevers svar. Et helt prøvesystem vil som regel ha en innretning der en normalfordeling med symmetrisk klokkeformet kurve (også kjent som *Gauss-kurven*)gjør seg gjeldende. Når antallet elever er stort nok, kan det forventes at karakterene fordeler seg rundt en middelverdi, der de fleste elevene får en karakter nær denne verdien, mens de høyere eller lavere karakterene er sjeldnere. Imidlertid gjelder denne antakelsen ikke mindre enheter som en klasse eller skole. Likevel har tidligere mange lærere brukt en slik norm for å vurdere sine elevers læringsresultater, noe som innebærer at det er enklere å oppnå gode karakterer i en lavtpresterende klasse og omvendt (se kap. 2.3 og 2.4 for nærmere informasjon). Normrelatering er i tillegg problematisk fordi prinsippet egner seg for en rangering av læringsresultater, men ikke til å kommunisere forventningene og kravene til elevene. Normrelatering gir altså ikke lærerne et redskap for å kommunisere med elevene.

Denne kritikken av den normrelaterte evaluerings- og vurderingstradisjonen dannet utgangspunktet for utviklingen av det vi i Skandinavia kaller *målrelatert vurdering*. I amerikansk terminologi ble dette først kalt *criterion-referenced assessment* (Popham og Husek, 1969), i Norge gjerne omtalt som kriteriebasert vurdering. *Målrelatert vurdering* krever tydelige kriterier som grunnlag for å kunne vurdere måloppnåelsen. Glaser og Klaus (1962) uttrykte distinksjonen mellom målrelatert og normrelatert vurdering slik: «Criterion-referenced measures depend on an absolute standard of quality while norm-referenced measures depend on a relative standard» (ibid., s. 421). En fordel med å ha mål og kriterier som sammenligningsgrunnlag er at det bedre legger til rette for å gjennomføre vurderingen uten behov for et stort antall elever eller en representativ del av elevgruppa, slik det normrelaterte prinsippet forutsetter.

Sadler (1987) videreutviklet forståelsen av målrelatert vurdering til *standardrelatert vurdering*, der en standard definerer et bestemt kvalitetsnivå som en gruppe elever skal nå, og som blir etablert av myndighetene (Tveit, 2008; oversettelse utledet fra Sadler, 1987, s. 194). I en slik standardrelatert tilnærming blir vurderingskriteriene enda tydeligere spesifisert så at de på den ene siden beskriver høyere og lavere nivåer av måloppnåelsen. På den annen side blir et eller flere av disse nivåene definert som standarder alle (i tilfellet *minstestandard*) eller så mange elever som mulig (i tilfellet *regelstandard*) eller en spesifikk andel elever (i tilfellet *utmerket* standard) skal nå. Standardbegrepet inkluderer implisitt et ansvarsperspektiv på undervisningssiden gjennom å forplikte skolesystemet til å føre elever opp til forhåndsdefinerte nivåer.

For en helhetlig beskrivelse av sentrale vurderingsbegreper er det viktig å inkludere individrelatering som vurderingsprinsipp. Dette prinsippet er mye i bruk når man gir tilbakemeldinger til elever med utgangspunkt i elevenes tidligere måloppnåelse. *Individrelatert vurdering* er altså i overensstemmelse med *tilpasset opplæring* som grunnverdi i undervisningen og formålet med norsk grunnopplæring, og den kan brukes i

underveisvurderinger. Prinsippet er imidlertid ikke forenlig med et eksamenssystem som har rettferdig konkurranse om videre utdannings- og yrkesmuligheter som grunnverdi.

5. Validitet i dagens eksamenssystem

Det er sentralt for valideringsprosessen at det utvikles en argumentasjon om hva slags bevis som skal regnes som gyldige relatert til et eller flere formål med en prøve, og hvordan fortolkningen skal foregå. Dette legges til grunn i dette kapitlet for å oppsummere kunnskapsgrunnlaget om validitet i dagens eksamenssystem. Vi skal se nærmere på de to formålene med eksamen som direkte gjelder elevene, og som derfor har blitt identifisert som hovedformål i kapittel 3: å prøve elevenes individuelle kompetanse i faget som det er beskrevet i læreplanen, og å gi et grunnlag for opptak til høyere utdanning og yrkesliv. Hovdhaugen, Prøitz og Seland (2018) påpeker at man er avhengig av at karaktersystemet har høy legitimitet for å kunne ivareta formålene med eksamen.

5.1 Forholdet mellom eksamen og læreplanen

Eksamenssystemet skal sørge for validitet ved at spesielt kyndige fagpersoner samarbeider om oppgaver basert på nasjonale retningslinjer, med mulighet for systematisk tilbakemelding fra sensorkorpset. Det er generelt lite systematisk forskning på sammenhengen mellom læreplan og eksamen, men det finnes erfaringsbasert kunnskap og brukerinnsikt på feltet. Vi har bestemt oss for å inkludere dette i kunnskapsgrunnlaget selv om den er av varierende kvalitet og ikke systematisk dokumentert. I tillegg har bare et lite utvalg fag og eksamensformer blitt utredet. Og så er kunnskapen til en stor grad bare basert på spørreundersøkelser med ulike utvalgsstørrelser og svarprosenter. Det er vanskelig å kontrollere, eller i det minste vite om, mulige skjevheter, som gjerne er rettet mot det positive. Rekkevidden av kapitlet er dermed begrenset, og det må tas høyde for at det er usikkerhet på større områder, og at det dermed er vanskelig å konkludere på en presis måte. Det er ønskelig å undersøke elevenes og lærernes opplevelser og synspunkter på en mer systematisk måte, samt å gjøre faglige analyser av eksamensoppgaver der de blir sett i sammenheng med læreplaner og formålet for eksamen.

Som en del av arbeidet med å videreutvikle kvaliteten på sentralt gitt skriftlig eksamen blir skriftlig eksamen i matematikk for 10. trinn evaluert i perioden 2017–2019 av Fafo. Det blir undersøkt hvordan sensuren fungerer, og gitt en vurdering av eksamens innhold og utforming. Det blir også foretatt undersøkelser av

hvordan lærere og sensorer vurderer sammenhengen mellom læreplan, undervisning og eksamen i matematikk, og av hvordan elevene opplever eksamen.

Evalueringen våren 2017 viser at matematikkeksamen framstår som god og rettferdig (Andresen mfl., 2017). Dette er en oppfatning som er gjennomgående blant elever, lærere og sensorer. De fleste lærerne og sensorene mente det var godt samsvar mellom kompetansekrav og hva som ble prøvd til eksamen. Evalueringen våren 2018 bekreftet disse hovedfunnene (Bjørnset mfl., 2018). Prøvereliabiliteten er vurdert til å være høy, ifølge IRT-analyser, noe som ble tolket som at eksamen måler det den gir seg ut for å måle – elevenes matematiske kompetanse (Bjørnset mfl., 2018). Lærere som ble intervjuet, gir imidlertid uttrykk for at oppgaver med mye tekst hindrer elever i å få vist sin matematiske kompetanse, noe som særlig gjelder for minoritetsspråklige elever og elever med lese- og skrivevansker.

Et viktig validitetsspørsmål er om eksamen prøver det samme konstruktet over år gitt at eksamensoppgaver er forskjellige. I forbindelse med evalueringen av matematikkeksamen gjennomfører Udir i samarbeid med forskningsenheten Enhet for kvantitative utdanningsanalyser (EKVA) ved ILS derfor en kvantitativ undersøkelse av vanskegraden på eksamen ved hjelp av årlige kalibreringsprøver. Kalibreringsprøven gjennomføres april hvert år og består av de samme oppgavene hvert år, og elevens prestasjoner på kalibreringsprøven og eksamen skal sammenlignes over tre år. Ut fra dette kan forskerne etter hvert konkludere med om det er elevenes prestasjoner eller vanskegraden på eksamen som kan forklare eventuelle variasjoner i eksamensresultatene. Resultatene fra undersøkelsene så langt viser at elevresultatene er på samme ferdighetsskala både i 2017 og 2018.

At det er godt samsvar mellom eksamensoppgavene og kompetansemålene, og at oppgavene gir mulighet til å vise kompetanse på ulike nivåer, har generelt blitt bekreftet i den årlige sensorundersøkelsen som Udir gjennomfører til sentralt gitt eksamen (Utdanningsdirektoratet, Sensorrapporter 2018). Også IRT-analyser for våreksamen for biologi 2 i 2017 og 2018 peker på at det er god overensstemmelse mellom oppgavens vanskegrad og elevenes ferdigheter (Naturfagsenteret, IRT-analyse av biologieksamen 2017 og 2018), noe som kan tolkes slik at eksamen er i tråd med læreplanen.

Likevel er det viktig å se begrensningene ved disse undersøkelsene. Skoleledere og skoleeiere ble spurt om sine synspunkter og oppfatninger av eksamen i spørreundersøkelsen Spørsmål til Skole-Norge i 2017, blant annet om de oppfatter det slik at eksamen gir elevene mulighet til å vise sin kompetanse (Waagene mfl., 2018). Ifølge rapporten fra spørreundersøkelsen er skoleledere og skoleeiere i stor grad omforente om at muntlig og skriftlig eksamen gir elevene mulighet til å vise sin kompetanse (Waagene mfl., 2018). Det er imidlertid uenighet om eksamen er egnet til å vise kompetanse i *alle* fag eller bare i noen fag.

Det er også uenighet om det er klart *hvilken* kompetanse elevene skal vise til eksamen. Halvparten av skolelederne mener at det er helt klart, mens den andre halvparten svarer at det er noe uklart. Blant ungdomsskolelederne svarer en noe større andel at det er helt klart, sammenlignet med de andre skoletypene. Til sammenligning er skolelederne og skoleeierne i større grad enige om at det er helt klart hvilken kompetanse elevene skal vise til *standpunkt*. Her mener cirka tre av fire at det er helt klart hvilken kompetanse elevene skal vise. De minste skolene ser ut til å være noe mer positive til begge

eksamensformene enn de større skolene og har en større andel som svarer at muntlig og skriftlig eksamen gir elevene mulighet til å vise sin kompetanse i alle fag.

5.2 Læreplanforståelse i endring

Selv om det er lite forskning på sammenhengen mellom læreplan og eksamen, er det kommet flere studier om læreplaner og vurdering de siste årene som gir innsikt i klasseromspraksis. Det vil være naturlig å anta at det er en viss sammenheng mellom klasseromspraksis og praksis til eksamen.

Læreplanforståelse inkludert kompetansebegrepet er en forutsetning for å utvikle og vurdere eksamen i samsvar med læreplanverket i fag. FIVIS-studien påpekte at det kan være svak / mangel på kompetanse, samarbeid, fortolkningsfellesskap og planlegging i skolesektoren når det gjelder validitet i den løpende vurderingen i klasserommet (Buland, Engvik, Fjørtoft, Langseth, Sandvik, og Mordal, 2014). Gitt at fagfornyelsens kompetansebegrep er enda mer komplekst enn Kunnskapsløftets kompetansebegrep, kan det konkluderes at utfordringene sannsynligvis vil øke.

Sandvik mfl. (2012) finner at skoler har ulik forståelse av kompetansetenkningen i Kunnskapsløftet. Forskere påpeker utfordringen det er med bruk av lokale læringsmål som ikke gjenspeiler eller knyttes til kompetansemålene i læreplanen, og en fare ved at mange smale, lokale læringsmål som vurderes gjennom hyppig testing, kan føre til fragmentering og overflatelæring (Sandvik mfl., 2012; Hodgson mfl., 2011; 2012). En litteraturgjennomgang av forskningsrapporter viser at det er inkonsistens mellom kompetansemålene i LK06 og lokale læreplaner (Andreassen, 2016). Tilbakemeldinger fra embetene tyder på at det er en utfordring at en del lærere og skoleledere ikke ser de ulike delene av læreplanen i sammenheng ved standpunktvurderingen (Utdanningsdirektoratet, 2015). Enkelte embeter påpeker at lærere har problemer med å beskrive elevenes fagkompetanse ved klagesaker (Utdanningsdirektoratet, 2015).

Samtidig har det, som et resultat av lokale utviklingsprosesser og nasjonale tiltak (f.eks. satsingen Vurdering for læring), vært stor oppmerksomhet rettet mot vurderingsfeltet de senere år, og grunnopplæringen preges i stadig økende grad av en læringsfremmende vurderingskultur (Kunnskapsdepartementet, 2016). I en av NIFUs spørringer til Skole-Norge våren 2017 svarer 95 prosent av skolelederne at arbeidet med Vurdering for læring har økt bevisstheten om sammenhengen mellom vurdering og lokalt arbeid med læreplaner, har bidratt til mer aktiv bruk av læreplaner og til at skolen har utviklet en mer læringsorientert vurderingskultur (Federici mfl., 2017). Skolelederne har også i overveiende grad inntrykk av at et flertall av lærerne ser kompetansemålene i sammenheng, og at standpunktkarakterer settes på grunnlag av et bredt tilfang av kilder. Et stort flertall av skolelederne mener også at læreplanen gir god støtte til lærernes standpunktvurdering.

I hvilken grad denne utviklingen av praksis bidrar til å sikre eksamens validitet etter fagfornyelsen, trengs det mer kunnskap om. NOU 2015: 8 påpeker at utfordringene knyttet til forståelse av kompetansebegrepet sannsynligvis vil øke gitt den kompleksiteten som ligger i fagfornyelsens kompetansebegrep. Utredningen etterspør ulike tiltak for å kvalitetssikre sluttvurderingen, ikke minst det å tydeliggjøre kravene og kriteriene sluttvurderingen skal ta utgangspunkt i. Kompetansemål i læreplanene fordelt på trinn, helst med forskjellige nivåer av måloppnåelse, samt veilednings- og støttemateriell (som eksempler på elevbesvarelser) er kjerneelementer her. I overensstemmelse med krav fra skoleeiere, skoler og lærere ser utvalget i tillegg et behov for å styrke regelverket om standpunktvurderingen fordi dagens forskrift i liten grad spesifiserer kvalitetskrav eller vurderingsprosesser, noe som kan lede til forskjeller i vurderingsresultater (NOU 2015: 8; se også NOU 2019: 3).

5.3 Eksamens forskjellige roller i praksis

Som pekt på ovenfor (se kap. 4.1) er validering en prosess der man undersøker en prøves validitet i lys av dens funksjon. Empiriske studier viser imidlertid at eksamen og eksamenssystemer i praksis kan ha flere funksjoner enn dem som formelt sett er definert som formål i lovverket (Newton, 2007; Herman og Baker, 2009; Stobart, 2008). De ikke-definerte, implisitte funksjonene kalles «roller» i forskningen. Disse er ikke alltid ønskelige, men de finnes og må følges med på Det er et kjent problem at når eksamen har ulike formål og roller, vil det kunne oppstå spenninger og motsetninger dem imellom.

Det er derfor viktig å definere eksamens hovedformål og å avklare hvilke roller eksamen har utover dette i praksis. Dette bør gjøres for å unngå at disse rollene ikke kommer i veien for hovedformålene eller gir grunnlag for ulik tolkning av eksamensresultater, og fordi de kan representere en trussel mot eksamensvaliditeten. I Norge finnes det svært lite forskning om dette spørsmålet. I dette avsnittet presenterer vi et analytisk rammeverk som skiller mellom forskjellige eksamensformål og -roller, og beskriver disse nærmere. Rammeverket kan være et utgangspunkt for å utrede i hvilken grad eksamen i Norge har implisitte roller utover det eksplisitt definerte, formelle formålet.

Eksamens formål å sertifisere læring og rangere elevene

I overensstemmelse med internasjonal forskning skiller Tveit og Olsen (2018) mellom ulike formål og roller eksamen kan ha. En eksamen kan, for det første, brukes summativt til å sertifisere elevenes kompetanse og til å selektere elevene til videre utdanning og yrkesliv gjennom karaktersetting og rangering basert på den. Både eksamen og standpunktkarakter skal gi tallfestet informasjon om fagkompetansen til eleven ved slutten av opplæringen i faget. Karakterene fra sluttvurderingen har stor betydning for sertifisering av kompetanse og for opptak av elevene til høyere utdanning og yrkesliv eller på 10. trinnet for inntak til videregående opplæring. Disse to formålene er tydelig beskrevet i lovverket, og vi har identifisert dette som

hovedformålene med eksamen (se kap. 3.1).

Eksamens rolle i å kvalitetssikre elevenes resultater

Eksamen kan være med på å kvalitetssikre elevenes resultater fordi elevene får en ekstern vurdering av sin fagkompetanse (Meld. St. 28 (2015–2016)). I et sluttvurderingssystem som i stor grad er basert på faglærers vurdering, kan eksamen anses å være et viktig eksternt kvalitetselement. Fag med sentralt gitt eksamen har for eksempel identiske oppgavesett for alle som kommer opp i faget, noe som kan bidra til at elevene får et mer likeverdig vitnemål, idet eksamenskarakterene settes på det samme vurderingsgrunnlaget (Kunnskapsdepartementet, 2016, s. 62).[5] Den kvalitetssikrende rollen kommer mer implisitt også til uttrykk i St.meld. nr. 30 (2003–2004), der eksamen ble omtalt som «spesielt kvalitetssikrede prøver» (Utdannings- og forskningsdepartementet, 2004, s. 37) fordi de blir utviklet i tråd med tydelige kvalitetskriterier.

Eksamens rolle i å videreutvikle vurderingspraksisen

Eksamensresultatene kan bidra til at lærerne og skolen videreutvikler både egen praksis og arbeidet med vurdering. De to stortingsmeldingene Meld. St. 20 (2013) og Meld. St. 28 (2016) legger stor vekt på at en av eksamens funksjoner er kompetanseheving av sensorer. Lærere som fungerer som sensorer, kan delta i mange tiltak som forbedrer deres vurderingspraksis, for eksempel sensorskolering og i møter med andre sensorer for å utvikle tolkningsfellesskap. Denne kunnskapen/erfaringen tar de med seg tilbake til sine skoler, der de så har mulighet til å videreformidle den til andre lærere.

I tillegg er karakterer som gis til eksamen, en tilbakemelding til skolen om hvordan eksterne sensorer vurderer elevenes eksamensprestasjoner. Dette gjelder både sentralt og lokalt gitt eksamen. Tveit og Olsen (2018) peker på at videregående opplæring har få andre statistiske opplysninger, og det er derfor naturlig at eksamen blir en kunnskapskilde for å vurdere læringsresultater.

Ulike studier viser imidlertid at forholdet mellom eksamens- og standpunktkarakterer er uklart for mange lærere og skoleledere (Hovdhaugen mfl. 2014, 2018; Prøitz og Sport Borgen, 2010). Det er blant annet ulik oppfatning av om eller i hvilken grad de to karakterene bør harmonere, og om eksamen representerer en smalere prøving av kompetanse enn standpunkt. Ulik forståelse av hva som er eksamens rolle, kan for eksempel gi grunnlag for ulik tolkning og bruk av karakterstatistikk i lokale kvalitetsvurderingssystemer, noe som får konsekvenser for det lokale utviklingsarbeidet.

I Meld. St. nr. 28 (2015–2016) understrekes det imidlertid at en sammenligning av karakterer er mest hensiktsmessig når den brukes til å se om det er systematiske avvik fra den nasjonale, gjennomsnittlige differansen mellom standpunkt og eksamen over tid. Meldingen sier videre at dette bør kun være *én av flere kilder* til kunnskap om praksis i skolen. Hovdhaugen mfl. (2018) peker også på at det finnes flere svakheter ved ideen om at eksamen kan fungere som kalibreringsverktøy av standpunktkarakterene, for eksempel kan de to vurderingsformene

- være svært ulike og skille seg rent praktisk fra hverandre
- ha klare ulikheter slik de er juridisk definert

- ha helt forskjellige premisser i selve karaktersettingen

Eksamens rolle i å styre opplæringen

Eksamen kan også ha en rolle i å styre forståelsen og praktiseringen av læreplaner. Forskning viser at eksamen kan ha tilbakevirkende («washback») effekter på opplæringen, dette skyldes at gjennom eksamenssystemet anerkjennes tilsiktet eller utilsiktet hva som anses som viktig i læreplanen[6] (se Nordenbo mfl., 2009). Samtidig kan det argumenteres for at dette ikke trenger å være et problem så lenge eksamen gjenspeiler læreplanen.

I Utdanningsspeilet 2008 (Utdanningsdirektoratet, 2009) framgår det at vurderingsveiledningene til eksamenssensuren «skal ha ein læringsfremjande effekt ved at lærarar kan formidle kjenneteikna til elevane før eksamen» (s. 105). I tråd med dette kan tidligere gitte eksamensoppgaver og vurdering av disse være eksempler som skoler, skoleledere og enkeltlærere kan bruke som utgangspunkt for å tolke og analysere kompetansebegrepet, læreplanen og kompetansemålene i det enkelte faget. Kapittel 2, om framveksten av dagens eksamenssystem, viste at eksamen opprinnelig var et slikt instrument i styringen av utdanningssystemet, mens dagens hovedformål, sertifisering og seleksjon, etter hvert har blitt viktigere.

Eksamens roller i å støtte læring og undervisning

Eksamen kan få en formativ rolle ved at lærere bruker tidligere eksamensoppgaver for å eksemplifisere/synliggjøre hva som er forventet kompetanse i sluttet av opplæringen, og som et utgangspunkt for å diskutere kompetanse i fag, progresjon og kjennetegn på måloppnåelse med elevene. På denne måten kan eksamen være til støtte i læringsprosesser og brukes til å tilpasse opplæringen. Denne rollen gjelder riktignok først og fremst til underveisvurderinger og prosesser i klasserommet, ikke til eksamen som en del av sluttvurderingen. Likevel kan eksamen brukes formativt ved at læreren tar for seg eksamensresultatene og analyserer styrker og svakheter i eksamensbesvarelsene fra sine elever og ser dette i sammenheng med opplæringen som er gitt, og kan da bruke dette som utgangspunkt for å justere opplæringen i faget for neste skoleår.

Noen ganger blir det framhevet at også eksamen kan ha en formativ rolle ved å bli brukt som ekstern motivasjon så at elevene opprettholder innsatsviljen mot slutten av skoleløpet. I denne konteksten bør igjen trekkordningen drøftes – muligens kan den bidra tilsiktet eller utilsiktet til å sikre denne effekten i flere fag. I tillegg kan det antas at forberedelsen til eksamen har en egen læringseffekt utover undervisningstiden.

Oppsummert viser dette analytiske rammeverket at dagens eksamenssystem kan ha flere implisitte roller utover formål definert i lovverket (se kap. 3.1 om disse) i sertifisering, seleksjon, kvalitetssikring, videreutvikling av vurderingspraksis, styring av undervisningen og til og med støtte av læring i norsk grunnopplæring. Det er sannsynlig at eksamen har flere implisitte roller enn det som beskrives som formål med eksamen i regelverket. Flere og ulike formål og roller kan føre til ulike tolkninger av eksamensresultater og ulike bivirkninger ved endringer. Det er derfor svært viktig å avklare de implisitte rollene eksamen har i praksis. Imidlertid finnes det lite forskning som utreder dette feltet, det er følgelig vanskelig å komme med

tydeligere konklusjoner her.

---

5) I det svenske systemet forventes til og med at lærere tillegger resultatene fra de nasjonale prøvene betydelig vekt når de setter karakter (Gustafsson og Erickson, 2018).

6) Trekkordningen bør muligens også drøftes i denne konteksten fordi den skal sikre at elevene «skal være eksamensforberedt i de fagene der eksamen er en mulig sluttvurdering ved siden av standpunktkarakteren» (Kunnskapsdepartementet, 2013, s. 65–66). Det finnes imidlertid ingen empiriske studier som sier at trekkordningen faktisk har en slik styringsrolle.

6. Reliabilitet i dagens eksamenssystem

Det er et viktig kvalitetskjennetegn at en eksamensoppgave får samsvarende vurderinger av flere sensorer slik at karaktersettingen ikke er preget av tilfeldigheter. Dette krever tydelige oppgaver med gode instrukser, tydelige vurderingskriterier (i.e. kjennetegn på måloppnåelse) og omfattende sensorskolering for å sikre tolkningsfellesskap. Samtidig vil det til en viss grad alltid være forskjeller i vurderinger blant sensorer. Tilstrekkelig reliabilitet er imidlertid en forutsetning for kvalitet i vurderingsarbeidet. Dette kapitlet beskriver dagens rammer for å sikre god reliabilitet og oppsummerer kunnskapsgrunnlaget vi har på dette området.

En utfordring med dagens datagrunnlag med tanke på å kunne forske på eksamensreliabilitet er at sensorinformasjon om hvert fag bare finnes samlet på elevnivå, men ikke på oppgavenivå innen en elevs eksamen. Det gjør det vanskelig å utrede årsaker til mulige problemer med sensorsamsvar i etterkant. Som pekt på tidligere (kap. 5) er det i tillegg en utfordring å kunne sikre høy reliabilitet gjennom pilotering før en eksamen/oppgave blir tatt i bruk, ettersom eksamensoppgaver må holdes hemmelige.

6.1 Rammer for eksamenssensuren

Udir har i samarbeid med fylkesmennene ansvaret for sensur til sentralt gitt skriftlig eksamen, og kommunen/fylkeskommunen har ansvaret for sensuren til lokalt gitt eksamen (§§ 3-28, 3-29, 3-30). Eksamen sensureres av to eksterne sensorer, ved lokalt gitt eksamen kan den ene sensoren være elevens faglærer. Ved uenighet om karakteren skal karakteren avgjøres av en oppmann til sentralt gitt skriftlig eksamen og av den eksterne sensoren ved lokalt gitt eksamen.

Regelverket gir i dag føringer og rammer for sluttvurderingen generelt (jf. 3.2). Kravene til prosessene rundt sensur av eksamen har ikke tilsvarende innramming, for eksempel er det ikke beskrevet hvem som stiller krav til kvaliteten på sensuren. Prosessen rundt sensuren vil avhenge av hvilken eksamensform det er snakk om, og foregår på forskjellige måter, for eksempel om det er tale om sentralt gitt skriftlig eksamen eller muntlig eksamen. Mens det for sentralt gitt skriftlig eksamen utvikles felles oppgaver, vurderingskriterier og gjennomføres felles sensorskolering, vil det for muntlig eksamen være ulike oppgaver, vurderingskriterier og sensorskoleringer. Uavhengig av eksamensform baserer sensuren seg på et system der sensorene skal «diskutere seg fram til en karakter», der det å utvikle et tolkningsfellesskap mellom sensorene blir viktig for å øke reliabiliteten til eksamen i dagens system.

Et godt kvalitetssikringssystem bør inkludere systematiske tilnærminger som ivaretar sensorsamsvar og reliabilitet på en god måte uavhengig av eksamensform. Vurderingen ved muntlig eksamen skjer i sanntid og er ikke etterprøvbar på samme måte som for skriftlig eksamen per i dag. Muntlig eksamen gir imidlertid en viktig mulighet for elevene til å vise kompetanse på en annen måte enn til skriftlig eksamen. Det kan derfor være behov for ulike tilnærminger for å sørge for høy prøvekvalitet og pålitelig sensur ved de ulike eksamensformene.

6.2 Kjennetegn på måloppnåelse

Det er ingen føringer for at det skal utvikles kjennetegn på måloppnåelse eller tydelige vurderingskriterier knyttet til eksamen. Elevene har rett til å kjenne til hva som blir vektlagt i vurderingen av hans eller hennes kompetanse (jf. forskriften § 3-1). I hvilken utstrekning det finnes slike kjennetegn eller vurderingskriterier, og hvordan de blir brukt, er ulikt for sentralt gitt skriftlig eksamen og lokal gitte eksamener og varierer i tillegg på tvers av skoler og skoleeiere.

Ved sentralt gitt skriftlig eksamen utvikler Udir eksamensveiledninger med kjennetegn på måloppnåelse knyttet til alle eksamener med sentral sensur. Disse kjennetegnene skal brukes ved sensur og er utgangspunkt for diskusjon på sensorskoleringer og sensurmøtene. Udir har også utviklet veiledende kjennetegn på måloppnåelse i utvalgte fag på 10. trinn for å støtte standpunkt- og underveisvurderingen. Disse er det frivillig å bruke. Ved å tilby et felles utgangspunkt for å vurdere kompetanse i fag kan man bidra til å fremme en mer lik og rettferdig vurdering i hele landet. Kjennetegnene tar utgangspunkt i læreplanene og er beskrivelser av kvaliteten på kompetanse i fag på tvers av hovedområdene. Kompetansen er beskrevet på ulike nivåer; per i dag er kjennetegnene formulert for karaktergruppene 2, 3–4 og 5–6. Det er forventet at lærere ved en skole drøfter kjennetegnene og på denne måten utvikler en felles forståelse. Disse kjennetegnene kan også brukes som utgangspunkt for å utvikle kjennetegn til lokalt gitt eksamen.

Ulike undersøkelser viser at eksamensveiledninger med kjennetegn på måloppnåelse blir mye brukt og

oppleves som nyttige i skolenes vurderingsarbeid (Hovedhaugen mfl., 2014; Gjerustad mfl., 2015). 65 prosent av skolelederne oppgir også å bruke vurderte eksamenssvar (Waagene mfl., 2018), og halvparten bruker eksamensrapporter. I motsetning til dette svarer bare et mindretall av skolelederne at de bruker forhåndssensurrapporter for å utvikle et felles vurderingsgrunnlag på skolen. Spørsmål til Skole-Norge høsten 2014 viser at så godt som alle skoleledere og skoleeiere oppgir å ha utarbeidet og å ha brukt lokale kjennetegn i vurderingsarbeidet (Gjerustad, Waagene og Salvanes, 2015). Det finnes ingen systematisk informasjon om innholdet i og kvaliteten på disse kjennetegnene. Ved alle studiene må det i tillegg tas hensyn til at enten utvalgsstørrelsen eller svarprosenten var begrenset. Dermed finnes det – som ofte i spørreundersøkelser – en viss fare for at svarene har noen skjevheter, oftest rettet mot det positive.

Udirs kjennetegn på måloppnåelse gir et visst rom for tolkning som må utvikles og diskuteres i samarbeid med andre lærere. Hovdhaugen mfl. (2014) finner at lærerne synes det er lettest å vurdere besvarelser som ligger i ytterpunktene av skalaen, mens det krever mer arbeid å grunngi hvorfor man setter karakteren 3 fram for karakteren 4 enn for eksempel karakteren 5 heller enn 6. Fordi karakterene 3 og 4 utgjør en særlig stor andel av karakterene, har lærere etterlyst tydeligere vurderingskriterier på ulike typer oppgaver samt svareksempler som skal gjøre det enklere å skille mellom en 3-er og en 4-er (Krogh, 2016).

Som en oppsummering kan det slås fast at det finnes lite systematisk forskning, unntatt spørreundersøkelser, om hvordan skoleledere og lærere arbeider med kjennetegn på måloppnåelse. Kunnskapsgrunnlaget gir grunn til å anta at de har ulik erfaring med kjennetegn og vurderingskriterier, noe som kan føre til forskjeller i vurderingsprosessen.

6.3 Betydningen av tolkningsfellesskap

Det er mange fylkeskommuner og kommuner som har utviklet retningslinjer for muntlig eksamen, men disse har ulik innretning og sier i ulik grad noe om fagspesifikke forhold. Det er lite av systematisert kunnskap om hvordan kommuner og fylkeskommuner arbeider kvalitativt med sensuren av lokalt gitt eksamen. Fylkeskommunene har etablert samarbeid om lokalt gitt eksamen og har ulike samarbeidsarenaer og -områder, for eksempel samarbeides det om å utvikle felles eksamensoppgaver til lokalt gitt skriftlig eksamen i enkelte fag.

I NIFUs spørreundersøkelse til skoleeiere og skoleledere våren 2017 oppgir et stort flertall av kommunene (68 %) og fylkeskommunene (87 %) at de legger til rette for arenaer for læring og deling der lærere kan videreutvikle vurderingspraksis (f.eks. nettverk / faste samlinger / møteplasser) (Federici mfl., 2017). Det er færre skoleeiere, henholdsvis 47 prosent av kommunene og 53 prosent av fylkeskommunene, som oppgir å legge til rette for diskusjoner om innholdet i læreplanene.

Hovdhaugen mfl. (2014) finner at det er ulike former for fagsamarbeid mellom lærerne, og at skoleledelsen noen steder har satt inn konkrete tiltak for å utvikle og forme fagsamarbeidet mellom lærerne, andre steder er dette overlatt til seksjonene. Et stort flertall av skolelederne i NIFUs spørreundersøkelser gir uttrykk for at det på skolen i stor grad diskuteres på hvilken måte lærernes vurderingspraksis kan hjelpe elevene i å lære og å nå målene (Federici mfl., 2017). Skolelederne har også i stor grad inntrykk av at alle eller de aller fleste lærerne i samme fag/fagområde jobber sammen om å få en felles forståelse for hva kompetanse i faget er. Det er en sterk oppslutning om at både skoleeiere og skoleledere oppfattes som pådrivere for å utvikle vurderingspraksis, men det er særlig skolelederne som oppfattes å ha rollen som pådrivere.

De ulike undersøkelsene viser at det er et stort omfang av ulike former for samarbeid i forbindelse med vurdering lokalt, men sier ikke noe om kvaliteten i samarbeidsarenaene og i hvilken grad eller på hvilken måte dette arbeidet er knyttet til lokalt gitt eksamen. Vi kan imidlertid anta at samarbeid om vurdering indirekte også vil påvirke lokalt gitt eksamen. Samtidig er geografisk avstand i fylket eller kommunen en faktor som kan påvirke muligheten for samarbeid på tvers av skoler og deltakelse på eventuelle sensorskoleringer knyttet til lokalt gitt eksamen.

Udir arrangerer fellessensur i samarbeid med fylkesmennene ved alle sentralt gitte skriftligeksamener (med unntak av sentralt gitt eksamen med lokal sensur) og har ulike tiltak som samlet skal bidra til tolkningsfellesskap på sensuren (se tekstboksen under). Sensorskoleringene er en del av fellessensuren og er per i dag ikke obligatorisk. Alle sensorer oppfordres til å delta, og det er generelt stor oppslutning fra sensorene på disse møtene. Skolene er heller ikke pålagt å ha med lærere i sentralt gitt sensur, noe som kan bety at det er skoler (gjennom år) som ikke har hatt lærere som har hatt sensoroppdrag til sentralt gitt eksamen. Disse skolene vil da ikke ha lærere som kan tilbakeføre erfaringer fra skoleringen og sensurmøter.

Lærere som har deltatt i sensorskolering, opplever dette som svært nyttig (Hovdhaugen mfl., 2014), noe som samsvarer med direktoratets erfaringer. Lærere med sensorerfaring uttrykker tillit til det rette- og tolkningsfellesskapet som oppstår i arbeidet med fellessensur (Hovdhaugen mfl., 2014). Ifølge forskerne kan sensurmøtene bli en slags nøytral grunn for lærerne hvor selve faget står i sentrum, og hvor hver besvarelse er anonym og det kun er kjennetegn/vurderingskriterier som kommer til anvendelse. Sensurmøtene styrker lærernes opplevelse av trygghet og etterprøvbarhet ved eksamenssensuren.

> Utdanningsdirektoratet har flere tiltak som samlet sett skal bidra til et tolkningsfellesskap ved sensuren ved sentralt gitt eksamen (unntak: sentralt gitt eksamen med lokal sensur):
>
> - **Forhåndssensur** i grunnskolen og for norsk på vg3 for alle oppmenn: Retningsgivende for sensorskoleringene som ledes av oppmennene.
> - **Sensorskolering og felles sensur:** På sensorskoleringene, med utgangspunkt i et utvalg reelle eksamensbesvarelser, diskuteres hvilken kompetanse eksamensbesvarelsene viser og karakteren til den enkelte eksamensbesvarelsen. Tolkningsfellesskapet fra

sensorskoleringen legger føringer for karaktersetting av alle eksamensbesvarelsene i faget.

- **Eksamensveiledninger med kjennetegn på måloppnåelse** gir informasjon om eksamen og hvordan denne skal vurderes. Kjennetegn på måloppnåelse skal bidra til å sikre en samlet vurdering av kompetansen. Sensorene skal bruke veiledningen som en felles referanseramme i arbeidet sitt. Veiledningen skal være kjent i god tid før eksamen.
- **Eksamensbesvarelser med begrunnelser for ulike karakter** publiseres på Udir.no i ulike fag på grunnskolen og videregående skole. For hver besvarelse er det en begrunnelse til karakteren som er gitt. Brukes som referanse ved sensur, og kan brukes som utgangspunkt for å utvikle tolkningsfellesskap lokalt.
- **Eksamensrapporter** i et utvalg fag. Formålet er å gi lærere og kandidater bedre innsikt i hvordan eksamensoppgavene er forankret i læreplan og om erfaringene fra eksamensgjennomføringen og fellessensur, rapportene inkluderer også karakterstatistikk.

(Utdanningsdirektoratet 2018c)

Forskerne peker for øvrig på flere positive effekter av sensorskoleringer/-møter (uavhengig av hvem som arrangerer):

- Sensorskolering er viktig kompetanseheving i vurdering og for noen kanskje den eneste «opplæringen» i å sette karakterer.
- Mange framhever at samarbeid om sensur har tilført en måte å tenke helhetlig vurdering på og gitt dem redskaper for å jobbe med vurdering på i et tolkningsfellesskap.
- Lærerne oppfatter dette som verdifullt også i egen vurderingspraksis og at det kan bidra til å styrke vurderingsfellesskapet på skolen.
- Flere foreslår obligatorisk skolering fordi det kan komme hele fagfellesskapet på skolen til nytte og kan bidra til en mer ensartet vurdering på hele skolen.
- Sensur kan skape møteplasser og fagfellesskap, som mange lærere sier er en styrke for profesjonalitet i læreryrket generelt og vurdering spesielt.

Generelt opplever skolelederne at sensorenes erfaringer bidrar til å heve vurderingskompetansen ved skolen til muntlig og skriftlig eksamen. Åtte av ti skoleledere mener de bidrar i noen eller stor grad til skriftlig

eksamen og ni av ti til muntlig eksamen (Waagene mfl., 2018).

6.4 Sensorsamsvar

Høyt samsvar mellom sensorene i vurderingen av eksamen er viktig for kvaliteten. Selv om det er urealistisk å forvente at sensorene alltid vurderer en besvarelse likt, bør ambisjonen være å unngå større forskjeller i vurderingen. Det er viktig å notere seg at standardiserte oppgaver ikke nødvendigvis har høyere sensorsamsvar enn ustandardiserte oppgaver. Sensorsamsvar er ofte knyttet til om det var mulig å utvikle tydelige oppgaver, tydelig instruks og tydelige vurderingskriterier på forhånd. Reliabilitet er også knyttet til hvor omfattende sensorskolering har vært. Selv om vi ikke har empirisk forskning om dette, er det i tillegg ikke usannsynlig at antall eksamensbesvarelser per sensor påvirker både kvaliteten på sensuren og sensorsamsvaret fordi antall besvarelser kan påvirke hvor mye tid sensorene faktisk har til å diskutere og vurdere skriftlig besvarelser.

Felles sensurmøter og sensorskoleringer er lagt inn som en del av kvalitetssikringen ved sentralt gitt skriftlig eksamen. Sensorene gjør en foreløpig vurdering av oppgavene før sensurmøtene og foretar en endelig vurdering basert på tolkningsfellesskapet.

Fafos evalueringen av eksamen i matematikk for 10. trinn i 2017 viser at det var godt samsvar mellom sensorene i deres karakterforslag før fellessensurmøtet, selv om noen sensorer etterlyste bedre veiledning i sensur av enkelte oppgaver. Dette gjaldt særlig å få klarere retningslinjer for sensurering av oppgaver som krever digitale hjelpemidler (Andresen mfl., 2017). Forskerne konkluderer med at det var høyt sensorsamsvar ved eksamen våren 2018, det vil si at sensorene vurderer noenlunde likt (Bjørnset mfl., 2018).

> *Profesjonalisering av vurderingen:*
>
> Ekstern sensur medfører at lærere må diskutere læreplan, vurdering og karaktergivning med andre lærere både før og etter eksamen. Til sentralt gitt skriftlig eksamen rekrutteres sensorene fra hele landet, og det gjennomføres sensorskoleringer for å profesjonalisere vurderingen av besvarelsene og bidra til tolkningsfellesskap og rettferdig sensur. Muntlig, muntlig-praktisk og praktisk eksamen er på den annen side eksamensformer som innebærer en bred involvering av lærere lokalt. Både gjennom å være eksaminator for egne elever og ved å være sensor på andre skoler vil lærere få et eksternt blikk på egen praksis. Dette kan være et utgangspunkt for å diskutere, justere og videreutvikle egen opplærings- og vurderingspraksis.

Selv om prosjektet ligger litt tilbake i tid, må KAL-prosjektet (Kvalitetssikring av læringsutbyttet i norsk skriftlig) nevnes når det er snakk om eksamensvurdering i Norge (Berge mfl., 2005). Dette er en studie av 3300 eksamenstekster fra 1998–2001, og KAL har fremdeles status som den mest omfattende studien av elevers skriving på norskeksamen og sensuren av disse oppgavene. Blant annet fant KAL at elevene er ganske gode skrivere, selv de lavestpresterende skriverne var i stand til å produsere enkle fortellende tekster. Det ble samtidig avdekket store kjønnsforskjeller i jentenes favør og at elevene foretrekker å skrive subjektive fortellende tekster framfor saktekster. Denne tendensen blir direkte utfordret gjennom Kunnskapsløftet, der sakprosa og skjønnlitteratur sidestilles, og gjennom eksamensoppgaver som ikke har gjort det valgfritt å vise at en kan skrive sakpregede tekster.

Et spesielt viktig funn i denne sammenheng er knyttet til sensorenes vurdering av eksamensoppgavene. Her konkluderer KAL-forskerne med at samsvaret mellom sensorer i grunnskolen ikke er så høyt som ønskelig, men bedre enn hva mange har antatt. De framhever at grunnskolen utvikler en samtalekultur om elevers prestasjoner og elevteksters kvalitet. Videre kommer KAL-rapporten med en tydelig føring om at lærernes samtalekultur er en strategisk hovednøkkel for videre kvalitetsutvikling i grunnskolens skriveopplæring (Berge mfl., 2005).

Basert på en undersøkelse i regi av NIFU ser det ut som om behovet for tydeligere vurderingskriterier og større tolkningsfellesskap er særlig stor i videregående skole (Seland, Lødding og Prøitz, 2015). En metodisk undersøkelse EKVA gjorde på oppdrag fra Udir, tyder på at sensorsamsvar på eksamen er en utfordring særlig i norskfaget. Undersøkelsene viste at enigheten om vurderingen av besvarelsene fra sensor 1 og sensor 2 før etablering av tolkningsfellesskap på fellessensuren ikke var særlig god. At sensorsamsvar i fag som norsk er lavere enn i fag som matematikk, kan blant annet ha sammenheng med de vidt forskjellige oppgaveformatene kandidatene prøves i i disse fagene, og i hvilken grad besvarelsene gir rom for og behov for faglig skjønn. En masteroppgave om sensorreliabilitet i norsk ved årskullet 2015 støtter denne tolkningen og peker i tillegg på uklarheter angående vekting av vurderingskriteriene og vekting av kort- og langsvarsoppgavene (Krogh, 2016).

Bøhn (2017) har i sin doktoravhandling spesielt sett på hvordan vurderingen av muntlig eksamen i fellesfaget engelsk i videregående skole fungerer. Denne undersøkelsen konkluderer med at det jevnt over er akseptabelt sammenfall i bruk av overordnede kriterier. Reliabiliteten hos sensorene i undersøkelsen som inkluderte 80 informanter, var stort sett god. Men Bøhn peker samtidig på utfordringer med vurdering i engelskeksamen. Disse er relatert til vurdering av uttale og innhold og nivåfastsetting knyttet til enkelte kriterier. Her kunne det være nyttig å utvikle tydeligere kjennetegn for måloppnåelse. Ytterlig et funn er at felles forståelse av vurderingskriterier ikke automatisk betyr at lærerne vurderer prestasjoner likt. Det er også viktig at de er enige i hvordan en prestasjon skal nivåplasseres på karakterskalaen. I denne undersøkelsen framgår det at lærerne på engelsk for de yrkesfaglige utdanningsområdene hadde en tendens til å vurdere elevene «snillere» enn lærere på studiespesialiserende.

Carlsen (2003) har også sett på sensorbasert vurdering av muntlige språkferdigheter, i dette tilfellet i norsk som andrespråk. Hennes funn bekrefter Bøhns. Hun konkluderer at sensorene bør være enige med hverandre

om karaktersettingen og legge vekt på de samme trekkene i sine vurderinger, ellers står vurderingen i fare av å være preget av tilfeldigheter og dermed ikke til å stole på.

Samtidig viser forskningen oss hvordan det er mulig å oppnå større tolkningsfellesskap og høyere reliabilitet – til og med når det gjelder konstrukter som i utgangspunktet er dårlig definert som skriving. Mye av forskningen som finnes i Norge om vurdering av skriving, har blitt gjennomført i konteksten til de nasjonale skriveprøvene og i Normprosjektet. Erfaringer fra denne konteksten og hovedfunnene kan muligens overføres til eksamen.

Kvistad og Smemo (2015) fant ut at elevenes tekster og deres vurdering profitterte mest på eksplisitte forventninger, særlig på presisjon av formål (Otnes, 2015) samt detaljerte krav til innhold og struktur (Smemo og Solem, 2015). Upresist formulerte oppgaver var ikke bare avgjørende for elevenes prestasjon, men også vanskelige å vurdere (Solheim og Matre, 2014). Forfatterne fant ut at bruk av eksempeltekster var godt egnet til å utvikle tolkningsfellesskap blant sensorer fordi de ulike vurderingsnormene ble synlige på denne måten. Når det gjelder antall sensorer, gjennomførte Borgström og Ledin (2014) en studie i Sverige og konkluderte at i tekstvurdering trengs det tre sensorer for å sikre god reliabilitet.

For å komme fram til felles forventninger og vurderingskriterier («standarder») valgte Normprosjektet en *bottom-up*-prosess som involverte et større antall erfarne lærere (Solheim og Matre, 2014; Evensen mfl., 2016). Gjennom å utrede hvordan disse vurderte elevtekster, kunne det utvikles en matrise med flere dimensjoner som tydelig spesifiserte vurderingsnormene. En intervensjon der andre lærere fikk informasjon om skriving og vurdering, hjalp videre med å utvikle læreres vurderingskompetanse betydelig.

7. Forholdet mellom eksamen og standpunkt

Eksamensresultater som publiseres i Skoleporten og i statistikkportalen, gir Udir, fylkesmennene, skoleeiere og skolene et visst kunnskapsgrunnlag om karakterfordelinger og gjennomsnitt for eksamen. Karakterer og karakterforslag fra sensuren til sentralt gitt skriftlig eksamen, som blir registrert i PAS, er en informasjonskilde som kan brukes i arbeidet med å videreutvikle eksamen.

Standpunkt- og eksamenskarakter i et fag bør være uttrykk for det samme, men ofte stilles det spørsmål ved forskjellen mellom eksamenskarakterer og standpunktkarakterer. Forskjeller i karakterene er ikke nødvendigvis problematiske i seg selv. Det kan imidlertid være et spørsmål om hvor store forskjellene kan være før karakterene ikke lenger er et uttrykk for den samme kompetansen i læreplanen. Samtidig er det enighet om at eksamen og standpunkt dekker ulike perspektiver.

Forskjeller tyder altså ikke nødvendigvis på en over- eller undervurdering av elevene og er ikke alene nok til å så tvil om, eller legitimere, verken standpunkt- eller eksamenskarakterer. Men ulikheter i eksamens- og

standpunktkarakterer bør ikke være systematisk relatert til årskull, fag, klasser, skoler, geografiske egenheter eller andre type grupperinger. Hvis forskjellene kan relateres systematisk til andre ytre forhold enn elevenes kompetanse, representerer de skjevheter som ikke er forenlige med rettferdig vurdering.[7] Dette utreder vi nedenfor.

Forskning dokumenterer tydelig at det finnes forskjeller mellom eksamen og standpunkt (se f.eks. Hovdhaugen, Prøitz, og Seland, 2018), og at disse har eksistert i lang tid (Hægeland mfl., 2005). Dette er noen av funnene fra forskningen når det gjelder forholdet mellom eksamens- og standpunktkarakterer:

- Nasjonalt ligger de gjennomsnittlige eksamenskarakterene for sentralt gitt eksamen vanligvis noe under de gjennomsnittlige standpunktkarakterene. Hva disse forskjellene skyldes, har vi fortsatt lite kunnskap om. En studie viser for eksempel at differansen mellom standpunkt og eksamen avhenger av eksamensform og fag, og at det er større differanse i norsk enn i matematikk.1 Det er størst forskjell mellom standpunkt- og eksamenskarakteren i praktisk matematikk, der det skiller en hel karakter mellom gjennomsnittlig karakter til standpunkt og til eksamen. 78 prosent av elevene får lavere karakter til eksamen enn til standpunkt i praktisk matematikk for vg1. (Utdanningsdirektoratet, 2017)
- Standpunktkarakterene i fagene med skriftlig eksamen er forholdsvis konstante over tid, mens eksamenskarakterene varierer mer (Hovdhaugen mfl., 2014). Dermed er eksamenskarakteren den minst stabile av de to. Med utgangspunkt i dette funnet stiller forskerne spørsmål ved eksamens funksjon som objektivt målepunkt. Det spørs om endringer i eksamenskarakterer representerer endringer i kompetansenivået eller endringer i eksamenens vanskelighetsgrad.
- Små skoler (mindre enn 50 elever over 7 skoleår) og skoler med lave gjennomsnittlige eksamenskarakterer gir elevene bedre standpunktkarakterer enn det større skoler (omtrent 40 elever eller flere i gjennomsnitt per årskull) og skoler med høye gjennomsnittlige eksamenskarakterer gjør. Mellom 40–50 prosent av skolene peker seg ut med spesielt høye eller lave standpunktkarakterer sammenlignet med eksamenskarakterer. Karakterpraksisen på hver skole er i stor grad stabil mellom fag. Det vil si at hvis en skole gir høye standpunktkarakterer i ett fag, gir de også høye standpunktkarakterer i andre fag ved skolen. Karakterpraksisen er også stabil over år (Galloway, Kirkebøen og Rønning, 2011). Det ser ut som om lærere implisitt bruker en skoleintern sosial norm når de setter standpunkt. Det betyr at de orienterer seg mot det generelle nivået på sin skole. Standpunktkarakteren satt av en lærer ved en skole med høytpresterende elever evaluerer sannsynligvis derfor den samme sluttkompetanse som litt lavere enn det en lærer ved en skole med en høy andel lavtpresterende elever gjør. Fordi eksamen er gitt sentralt med de samme oppgavene og den samme sensureringspraksisen overalt i landet, resulterer dette gjerne i litt høyere eksamenskarakterer sammenlignet med standpunkter ved høytpresterende skoler enn ved lavtpresterende skoler.
- Det er også systematiske forskjeller i avviket mellom eksamens- og standpunktkarakterer når man sammenligner høytpresterende og lavtpresterende elever. Omtrent halvparten av elevene får en annen eksamenskarakter enn standpunktkarakter. Mens 75 prosent av elevene som hadde fått 5

eller 6 i standpunkt, går ned, er det færre enn 50 prosent av elevene som hadde fått 2, 3 eller 4, som får en lavere karakter ved eksamen (Utdanningsdirektoratet, 2013). Dette kan riktignok være en effekt som er drevet av naturlige svingninger fordi elever som har fått 5 eller 6, kan gå ned, men knapt nok opp, mens de som har karakterer i midten av skalaen, kan gå både opp og ned.

- Jenter gjør det – relativt sett – bedre på standpunkt enn på eksamen (se Utdanningsspeilet gjennom mange år), og dette gjelder særlig i fagene norsk og fremmedspråk (+ 0,4 til 0,5 karakterpoeng mer i fordel til jentene enn guttene sammenlignet med eksamen) (Wollscheid mfl., 2018; Borgonovi, Ferrara og Maghnouj, 2018). Stoltenberg-utvalget har undersøkt kjønnsforskjeller i skoleprestasjoner og mener at vurderingssystemet ser ut til å være en ulempe for gutter ettersom det er mange lærervurderte standpunktkarakterer og få eksamenskarakterer (NOU 2019: 3, 2019).
- Et annet eksempel er offentlige og private skoler hvor det viser seg at forskjellene mellom eksamens- og standpunktkarakterer er større ved private enn ved offentlige videregående skoler (Hovdhaugen, Seland, Lødding, Prøitz, og Vibe, 2014). Det skyldes trolig at elever ved private skoler får høyere standpunktkarakterer (gitt samme faglige dyktighet vist i eksamen) (Utdanningsdirektoratet, 2013).
- Rapporten Skoleresultater 2008, som Statistisk sentralbyrå har gjort på oppdrag for direktoratet, presenterer en kartlegging av skoleresultater i grunnskolen og den videregående skolen. Den viser en sterk sammenheng mellom elevenes karakterer i grunn- og videregående skole (Steffensen og Ziade, 2009). Fagkarakter fra grunnskolen gir generelt en god pekepinn på karakteren i tilsvarende fag på videregående, selv når det kontrolleres for forskjeller i familiebakgrunn. Rapporten inneholder også en analyse av strykprosent på tvers av utvalgte fag og elevgrupper i videregående. Det er lavere strykandel i språkfagene norsk og engelsk enn i matematikk, og for matematikkfagene er andelen som stryker, klart lavere i teoretisk enn i praktisk matematikk. Videre er andelen som stryker, lavere i fag på studieforberedende enn på yrkesfaglige utdanningsprogrammer.

De forskjellene mellom eksamen og standpunkt dokumentert her (mellom gutter og jenter, private og offentlige, store og små skoler samt mellom høytpresterende og lavtpresterende skoler eller elever) viser at det dreier seg om *systematiske* avvik mellom eksamens- og standpunktkarakterer. Dette igjen viser at forskjellene ikke med rimelighet kan knyttes til elevenes faglige sluttkompetanse. Flere studier har dokumentert at lærerne muligens legger vekt på andre faktorer enn læreplanmålene ved fastsetting av standpunktkarakterer i fagene: for eksempel elevenes innsats eller orden og oppførsel (Dale og Wærness, 2006; Prøitz og Spord-Borgen, 2010; Sjyvollen, 2007; Tveit, 2007b).

I tillegg ser det ut som om lærerne til dels muligens følger en normrelatert vurdering der de sammenligner elever i en klasse eller skole med hverandre istedenfor å utelukkende gjennomføre en målrelatering (Galloway, Kirkebøen, og Rønning, 2011; se kap. 4.4 for definisjoner av begrepene). Slike forskjeller skaper en fare for at ulike elever gis ulike muligheter. Når det er sagt, er det riktignok viktig også å peke på at en

målorientert vurdering vil forutsette entydig mål og operasjonalisering, hvilket det nok ikke finnes i dag på en tilstrekkelig måte.

I tillegg til de grupperelaterte forskjellene dokumentert ovenfor er det forskjeller mellom eksamen og standpunkt knyttet til årskull. Det betyr at variasjon i nivået av karaktersetting over år kan slå ut som en kilde til ikke-fair konkurranse om de samme studieplassene.

Det finnes også variasjoner i karaktersetting ved eksamen versus standpunkt på tvers av fag. Elever som har spesialisering i realfag, har høyere karaktergjennomsnitt i fellesfagene enn elevene som for eksempel tar samfunnsfag. Likevel får realfagelevene lavere karakterer på vitnemålet sitt i studiespesialiseringsfagene enn det de andre elevgruppene får (Angell, Lie, og Rohatgi, 2011). Det betyr at karakterkravene ser ut til å variere på tvers av studiespesialiseringsfagene, og at for eksempel en karakteren 5 ikke har samme betydning i realfag sammenlignet med i samfunnsfag. Noe av det samme kan observeres når fremmedspråkene sammenlignes med samfunnsfag. Hovdhaugen (2014) peker på at slike fagspesifikke forskjeller muligens skyldes forskjellige tilnærminger til vurdering. I tillegg kan forskjellene kan hende forklares ut fra epistemologiske aspekter i fagene, særlig fagenes spesifikke struktur (for flere detaljer, se kap. 8).

Et annet fenomen som kan få konsekvenser for opptak til høyere utdanning og yrkesliv, er skjevheter på bakgrunn av fagenes uttelling på vitnemålet. Fag med lavt timetall vektes likt til opptak som fag med høyt timetall. Antall karakterer per fag stemmer ikke nødvendigvis overens med antall opplæringstimer i videregående opplæring. Norskfaget, for eksempel, kan ha opptil seks karakterer på vitnemålet i vg3. Ettersom kjønnsforskjellene er store i språkfagene, vil antallet språkkarakterer gagne jenter, påpeker Stoltenberg-utvalget (NOU 2019: 3, 2019). Utvalget anbefaler derfor å utrede vekting av karakterer etter timetall eller andre modeller.

Oppsummert viser kunnskapsgrunnlaget at forholdet mellom eksamenskarakterer og standpunktkarakterer er preget av systematiske forskjeller som er relatert til ytre forhold, men muligens ikke til elevenes kompetanse – noe som ville svekke rettferdigheten til vurderingene. Et viktig spørsmål er på hvilken måte det er mulig å motvirke eller kompensere for slike skjevheter. Her viser vår gjennomgang så langt at det både i Norge og internasjonalt finnes lite forskning om dette.

[1] Det er viktig å huske at rangeringene bare kan være «noenlunde» de samme fordi det alltid er tilfeldig variasjon, særlig ved eksamen som blir tatt bare en gang, og gjerne for et lite utvalg av elever ved skolen. Tilfeldige forhold som eksamensengstelse, sykdom, dårlig dagsform, uflaks med de spesifikke oppgavene som ble gitt akkurat denne dagen, osv. vil føre til at en enkeltstående måling ikke er perfekt. Men dette vil ikke føre til systematiske forskjeller.

8. Vurdering i fag – fagforskjeller

Kapittel 7 dokumenterer at vurdering og vurderingsresultater varierer mellom skolefag. En forklaring på variasjonene knyttes til skolefagenes ulike og bestemte egenskaper, i norsk kontekst ofte kalt fagenes egenart. En annen forklaring er at lærere og sensorer i vurderingssituasjoner trekker på bestemte og forskjellige epistemologiske og ideologiske forestillinger om vurdering i ulike fag. Å prøve kompetanse i tråd med fagfornyelsen og basert på de nye læreplanene vil innebære å være nødt til å anerkjenne skolefagenes innhold og struktur. Dette kapitlet oppsummerer derfor hvilke oppfatninger forskere og lærere har om de forskjellige fagene, og hvordan de implisitte og/eller eksplisitte ideologiene påvirker holdninger til vurdering.

8.1 Oppfatninger av fag

Muller (2009) skiller mellom fag med ulik konseptuell og kontekstuell sammenheng. Der fag med sterkere konseptuell sammenheng er tydeligere disiplinært forankret i fagdomener i høyere utdanning (forskningsdisiplinen som faget refererer til), har de en strammere struktur med hierarkisk og sekvensiell oppbygning som gir lærere tydeligere rammer for vurdering. På den annen side finnes det fag med sterkere kontekstuell sammenheng som har svakere kobling til fagets referansedisiplin, er mindre hierarkiske, mer segmentert og som krever stadig utvikling av felles rammer for fagenes kunnskapsområder og hva som er viktig kunnskap i faget, som dermed bør vurderes. Skolefag utgjør således grunnlaget for læreres og sensorers konstruksjoner av rammeverk for vurdering av prestasjoner og tilhørende praksiser for karaktersetting (Wiliam, 1996).

Basert på sammenligninger av læreres utsagn om vurdering i engelsk, naturfag og matematikk hevder Black mfl. (2003, s. 68) at lærere i matematikk og naturvitenskap anser sine fag for å ha unike og objektivt definerte mål, mens lærere i engelsk (i en engelskspråklig kontekst) anser at det finnes en rekke mål som det kan være aktuelt for elever å nå på et bestemt tidspunkt (Black mfl. 2003, s. 68). Dette finner vi også i norske studier med gjentatte runder med intervjuer fra 2009 og fram til i dag og som totalt omfatter over 100 norske lærere i ungdomsskole og videregående opplæring, om vurdering i deres fag (Prøitz og Borgen 2010; Prøitz 2013; Hovdhaugen mfl., 2014; Seeland mfl. 2018; Prøitz 2018).

8.2 Oppfatninger av vurdering i fag

I nasjonale og internasjonale studier av lærerrapporterte betraktninger finner vi også at forståelser av fagenes egenart påvirker læreres vurderingspraksis. For eksempel blir vurdering i fag som engelsk ofte karakterisert som holistisk, intuitiv, ikke-numerisk og gjerne basert på observasjon og dialog, mens i fag som matematikk blir vurdering karakterisert som rasjonell-analytisk med fastsatte standarder og kriterier og med verdifrie og stabile indikatorer (Wyatt-Smith og Klenowski, 2013). Vi kan se beskrivelser av vurdering i fag som basert på smalere eller bredere grunnlag, der den smale tilnærmingen er dominert av bruk av bare en vurderingsform, gjerne skriftlig, eller en veldig kort prøvesituasjon. Den brede tilnærmingen domineres av et bredere utvalg av vurderinger og spesielt en kombinasjon av skriftlig, muntlig og/eller praktisk prøving ved eksamen som gjør det mulig å prøve kompetanse bredere. I norsk kontekst er det mye som tyder på at det er eksamensformen for det enkelte fag som bidrar til å definere disse mer smale eller brede rammene for vurdering (Prøitz, 2018).

Vi vet fra forskning at lærere og sensorer i stor grad er lojale overfor regelverk og retningslinjer om vurdering, men forskningen viser også at det kan by på problemer å følge nye ordninger og regler innenfor skolefagets ramme dersom politikken bak nye vurderingsordninger ikke harmonerer like godt med skolefagets rammer (Prøitz, 2014). For eksempel vet vi at noen skolefag synes å passe bedre med dagens kompetansebaserte tenkning enn andre skolefag (Muller, 2009; Prøitz, 2014).

Tidligere studier i Norge har vist at det kan være svakere sammenheng mellom fag, innhold og nasjonalt regelverk for vurdering, spesielt i mer kontekstuelt forankrede skolefag som norsk og kunst og håndverk (Prøitz og Borgen, 2009; Prøitz, 2013). Dette kan gjenspeile en svakhet knyttet til vurdering i Norge i forbindelse med utvikling eller revisjon av faginnhold i den nasjonale læreplanen der vurdering gjerne kommer inn i diskusjonene for sent eller «henges på til sist» og dermed ikke er en integrert del av arbeidet med læreplandokumentene (Lysne 2006; Gjone, 1983). Dette fører gjerne til et omfattende etterarbeid og tilpasninger for å sikre god vurdering.

Fagenes egenart har i svært begrenset utstrekning vært i vurderingsforskningens sentrum. Nasjonal og internasjonal vurderingsforskning har i stor grad tatt mål av seg å bidra til økt kunnskap om og definere god vurderingspraksis på et mer generelt og universelt grunnlag til tross for at forskningen som oftest skjer innenfor rammen av skolefag. Dette handler derfor ikke om flere studier på hvordan universelle prinsipper for god vurdering kan utvikles eller støttes (Brookhart 2013; Wyatt-Smith og Klenowski, 2013), men om skolefagenes innhold og struktur har vært tilstrekkelig anerkjent som sentrale faktorer innen vurderingsforskning.

Fagfornyelsen prøver å møte denne utfordringen ved å definere kjerneelementer som skal dekke det viktigste i fagene og gi en tydelig prioritering av hva elevene skal lære. Kunnskapsområder, metoder, begreper, tenkemåter og uttrykksformer som har blitt definert som viktigst, skal prege innholdet og progresjonen i læreplanene og bidra til at elevene over tid utvikler forståelse av innhold og sammenhenger i faget. På denne måten kan kjerneelementene bidra til at fagenes innhold og struktur anerkjennes, men om det faktisk skjer, er

et empirisk spørsmål og bør utredes grundig (inkludert utilsiktede bivirkninger).

9. Elevers opplevelse av eksamen

For å danne et helhetlig bilde av hvordan eksamen fungerer, er det avgjørende å lytte til hva elevene sier om hvordan de opplever eksamensordningen. Elever har gjennom Norsk Gymnasiastsamband og Elevorganisasjonen påpekt mangler ved eksamenssystemet helt tilbake til 1963. Det handler blant annet om at elevene ikke opplever å få vist sin fulle kompetanse, at dagsform i betydelig grad påvirker elevenes prestasjoner, og at det i stor grad er tilfeldig hvilket fag eleven blir trukket opp i.

Eksamen er en del av en kompleks og psykologisk virkelighet. Den kan være spennende og krevende, men samtidig være noe man gruer seg til. Eksamen kan føre til engstelse, men den kan samtidig ruste eleven til videre arbeidsliv og studier. Harris og Brown (2016) påpeker at sosiale og psykologiske faktorer påvirker en rekke aspekter i et skolesystem: avgjørelser om hvordan læring skal bedømmes, læreres og elevers deltakelse i vurderingspraksiser samt hvordan vurderingsresultater fortolkes, forstås og anvendes. Elever kan for eksempel bekymre seg for å få dårlige karakterer; lærere kan påvirkes av tidspress, humørsvingninger, fordommer og lignende når de setter karakterer på eksamensbesvarelser. De sosiale, historiske og kulturelle rammene i og rundt utdanningssystemet påvirker elevers syn på eksamen, motivasjon, selvbilde og selvfølelse og muligheter for samarbeid. De politiske og juridiske rammene for eksamen kan på lignende vis harmonere med eller komme i konflikt med læreres antakelser, verdier, holdninger og lignende. Harris og Brown (2016) påpeker at de menneskelige vilkårene for vurdering derfor bør ligge til grunn for hvordan vi forstår utforming, implementering og skåring av eksamen og andre vurderingssituasjoner.

Forskrift til opplæringslova § 3-32 gir rom for å tilrettelegge lokalt for elever som har behov for det, slik at de får vist kompetansen sin i faget. Tilretteleggingen skal ikke føre til at de får fordeler de andre elevene ikke har.

Det er lite forskning å finne på elevperspektivet i sluttvurderingssystemet generelt, og dette gjelder også elevenes egen opplevelse av eksamen. Dette kapitlet oppsummerer noen funn som er hentet fra internasjonal forskning og noen tilbakemeldinger direktoratet har fått gjennom spørreundersøkelser.

9.1 Elevstemmen, motivasjon, prøveengstelse, stress og prestasjon

En måte å få større innsikt i hvordan elevene tenker rundt for eksempel oppgavetyper, tidspress og stress knyttet til eksamen, er å få elevstemmen inn i arbeidet med eksamen og sensur. Som en utprøving har Udir, i samråd med Elevorganisasjonen, inkludert elevstemmene ved å gjennomføre spørreundersøkelser blant elever som gjennomførte eksamen i engelsk etter 10. trinn og i fellesfag engelsk vg1/vg2 i 2016 og 2017. Våren 2017 deltok også elever på forhåndssensuren i engelsk for 10. trinn for første gang og forklarte hva de tenkte om oppgavene.

Elev-/brukerperspektivet omhandles i oppmannsrapportene og i eksamensrapportene fra eksamen i engelsk 10. trinn våren 2017. De ble blant annet spurt om begrunnelsen for valg av oppgaver, formålet med forberedelsesdelen til eksamen og eksamenslengden (Utdanningsdirektoratet, 2018d). Nedenfor vises kun 2 av de 500 elevstemmene som uttalte seg om forberedelsedelen til eksamen i engelsk 10. trinn 2017. De aller fleste var positive til å «komme inn i faget og temaet» før de gikk i gang med selve eksamen:

> *Ærlig, man trengte ikke forberedelseshefte, nesten så det ikke er vits ...*

> *Fint, for da kommer man inn i arbeidsmodus og*
> *får tenke på faget i en dag før man har en såpass stor vurdering ...*

Sammenhengen mellom motivasjon og prestasjon er viktig å ta med inn i utviklingen av eksamensfeltet (Eccles, 1983). Både svært høye og svært lave nivåer av motivasjonsvariabler kan være mindre ønskelige enn mellomnivåer. For eksempel, hvis elever opplever at eksamensoppgavens betydning er lav, kan de velge å ikke bruke energi og anstrenge seg for å oppnå mestring (Natriello og Dornbusch, 1984). Hvis elever derimot oppfatter prestasjonskravet ved en oppgave som svært høyt, kan engstelse hemme ytelse (Tobias, 1985). På samme måte, hvis elevene har et svært lavt nivå av mestringsforventning for en oppgave, er det lite sannsynlig at de vil angripe oppgaven med mye entusiasme eller utholdenhet. Har de et høyt nivå av mestringsforventning, kan de risikere å ikke gi oppgaven tilstrekkelig oppmerksomhet for å oppnå gode resultater (Schunk, 1984).

Når det gjelder elevers prøveengstelse, rapporterer studier (Hill, 1984) at elevers nervøsitet er økende når elever opplever at testen vil ha stor betydning, når den er forventet å være vanskelig, og når forholdene rundt prøvesituasjonen er påtrengende (f.eks. rigide tidsrammer og assosiert tidspress, spesielle testinstruksjoner og ukjent prøveform). Selv om elevers feil på tidligere oppgaver påvirker utviklingen av engstelse, oppstår ikke nervøsiteten bare ved mangel på kunnskap eller ferdigheter som kreves for å svare på oppgavene. Studier har vist at elever med høy prøveengstelse gjør det bedre og utfører på nivåer nærmere sine mindre engstelige jevnaldrende på de samme kognitive oppgavene når prøvene administreres under mindre stressfulle forhold (Hill, 1984; Hill og Wigfield, 1994).

9.2 Elevers opplevelse av eksamensformer

Det finnes svært lite forskning om hvordan elever i Norge opplever de ulike eksamensformene. Vi har derfor valgt å se etter internasjonal forskning om vurdering i *high-stakes*-kontekster og -eksamen for å få en pekepinn på mulige utfordringer norske elever kunne streve med. Men det må vises til at vi ikke kan vite med sikkerhet om den internasjonale forskningsstatusen også gjelder for Norge. Det som riktignok kan sies, er at eksamen er en prøving med store konsekvenser så at det kan anses som sannsynlig at slike utfordringer finnes.

Internasjonal forskning samlet sett viser at elever foretrekker vurderingsformater som reduserer stress og engstelse (Nassar, Qaraeen, og Naba'h, 2011; van de Watering, Gijbels, Dochy, og van der Rijt, 2008). Birenbaum og Feldman (1998) finner i sin studie at elever med liten/ingen nervøsitet for prøver foretrekker åpne oppgaver. Elever med høy testengstelse foretrekker derimot i større grad flervalgsoppgaver fordi de assosierer dem med mer sikkerhet i vurderingssituasjonen. Dette funnet stemmer overens med studien til Nassar mfl. (2011), der elevene mente at flervalgsoppgaver i eksamen er mindre vanskelig, klarere og mer rettferdig enn en langsvarsoppgave i eksamen. Studentene mente imidlertid at begge typer av eksamen er verdifulle. Birenbaum og Feldman antar således at dersom elevene får den typen vurderingsform de foretrekker, vil de være motivert til å utføre sitt beste.

En tidligere studie av Ben-Chaim og Zoller (1997) finner at elever i naturvitenskapelige fag på videregående skole foretrekker eksamener som er skriftlige, med ubegrenset tid, og hvor de kan bruke støttende materiell. Tidsavgrensninger blir opplevd som stressende og fører til uro og press. Vurderingsformer som reduserer stress, vil i henhold til Ben-Chaim og Zoller (1997) øke sjansen for suksess, og elever foretrekker fortrinnsvis eksamener som legger vekt på forståelse i stedet for overflatelæring. Baeten mfl. (2008) finner at preferanse for ulike eksamensformer ser ut til å være relatert til ulike læringsstrategier og tilnærminger til læring; elever med en dybdetilnærming ser ut til å foretrekke langsvarsoppgaver, mens elever med en overflatetilnærming foretrekker flervalgsoppgaver ved prøver (Baeten, Struyven og Dochy, 2008; Birenbaum og Feldman, 1998).

En metastudie av Beller og Gafni (2000) fant at det er kjønnsforskjeller i preferanse for prøveformer og eksamen. Der man fant kjønnsforskjeller, foretrekker jenter langsvarsoppgaven, og gutter viser en liten preferanse for flervalgsoppgaver (f.eks. Gellman og Berkowitz, 1993). Videre finner Beller og Gnafni (2000) at gutter skårer bedre på flervalgsoppgaver enn jenter, og at jenter skårer bedre enn guttene på åpne oppgaver enn på flervalgsoppgaver (f.eks. Ben-Shakhar og Sinai, 1991). Sant nok finner man samtidig enkelte studier som påpeker det motsatte knyttet til kjønnsforskjeller for eksamensformer. Evidensen er dermed litt uklar.

Nassar mfl. (2011) sin studie finner et skille mellom lavt- og høytpresterende elever når det gjelder preferanse for bruk av langsvarsoppgave som prøveformat til eksamen. De fant at elever med høye prestasjoner foretrekker langsvarsoppgaver ved eksamen mer enn moderat- og lavtpresterende elever.

McDowell (1995) antyder at elever synes nye vurderingsformer i skolen er interessante og motiverende. Elevene er fortsatt klare over behovet for å oppnå gode karakterer, men det vil variere i hvilken grad de går inn

for dette. Alternative vurderingsformer (eksamen) kan bidra til å forandre en testkultur som er styrt av en retningen på en tradisjonell eksamensform, til en vurderingskultur som vektlegger sammenheng mellom undervisning og vurdering (Birenbaum og Dochy, 1996; Dochy og McDowell, 1997). Forskning viser at alternative vurderingsmetoder (f.eks. mappevurdering, gruppeprosjekt, bruk av caser) er mindre truende for de fleste elever enn tradisjonell testing. Disse alternativene oppfattes også som rettferdige prøveformater (Sambell, McDowell og Brown, 1997).

Dochy og McDowell (1997, s. 292) peker på at endring av vurderingsformer er en effektiv måte å oppmuntre elevene til å endre sine læringsmetoder. Videre blir det framhevet at vurdering er et av de mest effektive verktøyene for innovasjon både i undervisning og læring. «When assessment stays the same, students often will not accept the need to change their approaches to learning; for example, students often prepare for exams by rote learning even if this is not appropriate.» (Dochy og McDowell, 1997, s. 292). Forskerne advarer likevel mot en tro på at nye vurderingsformer automatisk er til det bedre, da de mener at det finnes ingen ideell enkelt vurderingsform. En enkelt vurderingsform kan ikke tjene flere forskjellige formål, og hver vurderingsform har sin egen metodevariasjon som samhandler med personer.

Oppsummert viser dette kunnskapsgrunnlaget store forskningshull angående elevenes opplevelse av eksamen og eksamensformer i Norge. De få spørreundersøkelsene som finnes, tyder på at det er mye å hente ved å lytte til elevene fordi resultatene kan bidra til økt validitet. Elevene har over tid gitt uttrykk for mangler ved eksamenssystemet i sin helhet og ved enkelte eksamensformer, for eksempel at de oppfatter oppgaver eller instrukser som utydelige. Ut fra internasjonal forskning ser det ut som om det er viktig å variere eksamensformer så mye som mulig så at ulike elevgrupper har mulighet til å prestere på den best mulige måten.

Del 3 - Frampek mot fagfornyelsen

10. Fagfornyelsens utvidede kompetansebegrep og eksamen

De kompetansebaserte læreplanene kom med Kunnskapsløftet LK06. Evalueringer av Kunnskapsløftet pekte på lokale forskjeller når det gjelder læreplanforståelse ute på skolene (jf. pkt. 6.3), men senere rapporter (Utdanningsdirektoratet, 2018e) viser samtidig at lærerne har fått økt bevissthet og forståelse for kompetansebegrepet og læreplanene. Fagfornyelsens nye kompetansebegrep som ligger til grunn for utforming av læreplanene, vektlegger – slik vi har vist i innledningskapitlet – elevenes anvendelse av

kunnskap og ferdigheter i både kjente og ukjente situasjoner og at det å forstå, å reflektere og å tenke kritisk er en viktig del av elevenes kompetanse.

> **Kompetansebegrepet i fagfornyelsen:**
>
> Kompetanse er å tilegne seg og anvende kunnskaper og ferdigheter til å mestre utfordringer og løse oppgaver i *kjente* og *ukjente sammenhenger* og situasjoner. Kompetanse innebærer *forståelse og evne til refleksjon og kritisk tenkning.*

Retningslinjene for utforming av læreplaner i fag i fagfornyelsen (LK20 og LK20S; Utdanningsdirektoratet, redigert 11.10.2018) gir føringer for å lage læreplaner som beskriver relevant kompetanse, tydelige prioriteringer, tydelig progresjon og god sammenheng i og mellom fag. Læreplanene skal være gode verktøy for støtte og styring for lærere, skoleledere og skoleeiere. Det står videre at læreplanene skal legge til rette for varierte undervisningsformer og vurderingsmåter som fremmer dybdelæring. Dybdelæring i fagfornyelsen er definert som «å gradvis utvikle kunnskap og varig forståelse av begreper, metoder og sammenhenger i fag og mellom fagområder. Det innebærer at elevene reflekterer over egen læring og bruker det de har lært på ulike måter i kjente og ukjente situasjoner, alene eller sammen med andre» (Utdanningsdirektoratet, 2018f).

Dybdelæring og kompetansebegrepet har elementer som både overlapper og samsvarer med hverandre. Begge begrepene framhever det å forstå, å anvende kunnskap og ferdigheter i kjente og ukjente sammenhenger. Det å lære å lære og reflektere over egen læring ligger som en del av både kompetansebegrepet og dybdelæring. Dybdelæring kan dermed anses som en forutsetning for å utvikle kompleks kompetanse som uttrykt i fagfornyelsen.

Å på den ene siden lage åpne og overordnede mål for at elevene skal kunne overføre det de har lært til både kjente og ukjente sammenhenger, og på den annen side gi tydelig uttrykk for hva elevene skal lære, og hvilken kompetanse som skal være gjenstand for sluttvurdering, er en vanskelig balansegang (se kap. 10.1). Det står dessuten i retningslinjene at «kompetansemålene også i noen tilfeller kan utformes noe smalere og uttrykke en avgrenset kompetanse». Kapittel 10.2 presenterer hva internasjonal forskning sier blant annet om hva som kan bidra til å utvikle eksamener i retning av kompleks kompetanse, og hvilke forhold det er viktig å ta med i betraktningen når vi skal prøve og vurdere denne sammensatte kompetansen til eksamen. På hvilken måte kan elevperspektivet ivaretas til eksamen, og hvilke konsekvenser får fagfornyelsen for involvering av elevene i forkant og/eller under eksamen (se kap. 10.3)?

Det finnes lite forskning om kompetanseorienterte eksamener, i særdeleshet knyttet til deres fordeler og ulemper, psykometriske kvaliteter og styringsfunksjon, inkludert utilsiktede effekter. Et unntak er innenfor utdanning i medisin, spesielt den kliniske formen *Objective Structured Clinical Examination* (OSCE). I kapittel 10.4 redegjøres det derfor også for kunnskap og erfaringer med prøving og vurdering av kompleks kompetanse som er hentet fra universitets- og høgskolesektoren.

10.1 Muligheter og utfordringer ved å måle kompetanse til eksamen

Kompetanseorienterte eksamener tar sikte på å måle mer komplekse evner og kunnskap. Slike eksamener kan føre til dybdelæringsprosesser allerede i forberedelsesfasen både med tanke på fagkompetanse og evnen til å benytte fagkompetanse i ulike kontekster. Eksempler på dette er problemløsning som involverer analyse og evaluering. I tillegg kan kompetanseorienterte eksamener gjøre det enklere for elevene å se relevansen av deres kunnskap og ferdigheter, noe som kan stimulere til dybdelæring og utholdenhet. Kompetanseorienterte eksamener som gjenspeiler en sammenheng mellom vurdering, undervisningspraksis og ønsket læringsutbytte i tråd med Biggs (2003) modell om *constructive alignment*, kan bidra til å styre tidlig læring i retning av kompleks kompetanse framfor kunnskap løsrevet fra kontekst.

Utviklingen av kompetanseorienterte eksamener er ofte mer krevende enn ved tradisjonelle eksamener (Schaper, Hilkenmeier, og Bender, 2013). Det er gjerne vanskeligere å vurdere kompetanse fordi mer komplekse evner og kunnskap som regel er mindre presist definert, og fordi det ikke alltid er mulig å utvikle klare kriterier som definerer hvorvidt et svar er riktig eller galt. Slike eksamener krever i tillegg kriterier som kan ivareta kvalitative forskjeller i besvarelsen samt i hvilken grad et kriterium er møtt. Dette kan føre til reduksjon i objektivitet og/eller reliabilitet. I alle tilfeller krever kompetanseorienterte eksamener mer faglig skjønn og utstrakt opplæring av prøveutviklere/eksamensnemnder og sensorer samt tilrettelegging for erfaringsutveksling og refleksjon (ibid.).

Det er ikke bare Norge som har utfordringer med å redefinere vurderingssystemet slik at det ivaretar et nytt og utvidet kompetansebegrep. Schleichers bok *World Class: How to build a 21st-century School System* (2018) omtaler utfordringene slik: «The dilemma for educators is that routine cognitive skills, the skills that are easiest to teach and easiest to test, are exactly the skills that are also easiest to digitise, automate and outsource.» Hvordan vi forholder oss til blant annet denne problemstillingen, vil ha avgjørende betydning for om vi lykkes i å møte kravene som ligger i fagfornyelsen.

10.2 Utvikling av eksamener som måler kompetanse

Før eksamensform velges – det være seg langsvarsoppgaver, flervalgsprøver, muntlige prøver eller mappevurderinger – er det nyttig å forestille seg i hvilke situasjoner elevene vil ha bruk for denne kompetansen senere i livet (Schaper, Hilkenmeier, og Bender, 2013). Deretter vil det være til hjelp å tenke

gjennom hva slags oppgaver som kan vurdere denne kompetansen, før beslutningen om eksamensform tas. I kompetanseorienterte eksamener vil oppgavene typisk være å løse og evaluere problemstillinger hentet fra det virkelige liv, med større eller mindre grad av kompleksitet (case- eller scenariobaserte eksamener). Ren reproduksjon av kunnskap vil være mindre aktuelt. En velkjent ulempe med slike eksamener er økt usikkerhet hos elevene om hvorvidt de har funnet den «korrekte» løsningen på oppgaven (ibid.).

Kompleksiteten som ligger i fagfornyelsens kompetansebegrep, er nesten umulig å prøve med en enkel eksamen eller eksamensform. Kompleksiteten krever å tenke helhetlig på sluttvurderingen som et system. Innenfor en enkelt eksamen kan i så tilfellet hver oppgave (eller grupper av oppgaver) konsentrere seg om enkeltaspekter (eller grupper av aspekter) innenfor kompetansen. Alle oppgavene bør imidlertid til sammen dekke kompetansen i sin fulle bredde og dybde og – dersom det er mulig – være integrert i en større case eller et mer vidtrekkende scenario (ibid.). Ettersom alle eksamensoppgaver inkluderer målingsfeil, er det bedre å ha flere små oppgaver enn én stor oppgave.

Siden kompetanseorienterte eksamener gir mer rom for tolkning enn tradisjonelle eksamensformer, blir det nødvendig å forhåndsdefinere hva som er høy og lav måloppnåelse samt ha klare terskler for disse nivåene og utvikling av mulige løsninger på oppgaven for å sikre riktig vurdering (ibid.). Eksamensutviklere må utvikle progresjonsbeskrivelser for kompetansen og hvilke innholdsdimensjoner den består av.

Det er vanskelig å se for seg at alle elementer ved det nye kompetansebegrepet kan prøves gjennom de tradisjonelt etablerte eksamensformene eller gjennom eksamen alene. Begrensninger følger for eksempel av bredden i kompetansebegrepet, når eleven skal lære å lære og å reflektere over egen læring, samt når eleven skal jobbe langsiktig med et område. Mappeeksamen har blitt diskutert som en ny eksamensform i denne konteksten fordi den kunne kompensere for at dagens eksamensform har preg av å være et øyeblikksbilde eller en stikkprøve og ville styrke mangfoldigheten i prøveformer, noe som kan komme ulike elevgrupper til gode. Den har i tillegg blitt pekt på som en vurderingsform som kan gjøre det mulig å inkludere elevperspektivet ved å tilby valgmuligheter (se kap. 10.3 for nærmere utredning av andre muligheter til å involvere elevene).

Imidlertid finnes det en spenning mellom fleksibilitet og mulighet for sammenligning, som identifisert av Koretz (1998, s. 332):

"Portfolio assessment has attributes that make it particularly appealing to those who wish to use assesment to encourage richer instruction – for example, the "authentic" nature of some tasks, the reliance on large tasks, the lack of standardization, and the close integration of assessment with instruction. But some of these attributes may undermine the ability of the assessments to provide performance data of comparable meaning across large numbers of schools."

Black, Harrison, Hodgen, Marshall og Serret (2011) utredet nødvendige komponenter i en elevmappe for å sikre validiteten til denne med tanke på kompetansekravene. Forfatterne konkluderte med at en samling av flere oppgaver var nødvendig. Validiteten av summative vurderinger var avhengig av omfanget og balansen mellom innholdet i hver elevs mappe, idet innholdet skulle gjenspeile omfanget og målene til faget og burde

være variert i stil (form). En mulig bekymring i denne konteksten var å vite hvem det er som faktisk svarer på oppgavene som inngår i den typer oppgaver som gjøres hjemme.

10.3 Elevinvolvering mot eksamen

Elevens aktive rolle i læringsprosessen er en kjerne i fagfornyelsens utvidede kompetansebegrep og vektleggingen av dybdelæring. Ifølge Meld. St. 28 (2015–2016) må vurderingsordninger og kvalitetsvurderingssystemet støtte opp under en opplæring som skal legge større vekt på dybdelæring og systematisk progresjon (s. 123). Som en følge av de nye elementene i fagfornyelsen vil det være naturlig å se nærmere på elevens rolle i forkant av og/eller under eksamen.

Å involvere elevene i eget læringsarbeid, inkludert elevenes vurdering av sine faglige prestasjoner, er en del av underveisvurderingen og har vært et viktig fokusområde i de siste årene, blant annet gjennom den nasjonale satsingen Vurdering for læring (2010–2018). Spørsmålene knyttet til egenvurdering og elevmedvirkning i Elevundersøkelsen viser imidlertid at det fortsatt er et stykke igjen før det er en innarbeidet praksis, og undersøkelsen viser også at skoler har ulik praksis. Samtidig har det vært en relativ god utvikling på disse spørsmålene sammenlignet med andre spørsmål om vurdering i perioden 2013–2017.

Eksamensordningene i dag gir til en viss grad mulighet for å involvere elevene i eksamen. For eksempel kan dette gjøres gjennom forberedelsesdelen til eksamen som gir elevene mulighet til å forberede seg alene og/eller i samarbeid med andre, eller gjennom oppgavetyper som åpner mer for å velge ulike tilnærminger enn andre.

10.4 Reliabilitet og validitet i vurderinger av kompleks kompetanse

Fra forskning om eksamen i medisinfaget framkommer det at et egnet utvalg eksamensoppgaver av ulik type, kontekst og flere sensorer kan sikre høy reliabilitet (Wass, Van der Vleuten, Shatzer, og Jones, 2001). Forskningsfunnene viste at alle eksamensformene kan oppnå tilstrekkelig reliabilitet – selv om det ikke er standardiserte tester – forutsatt at det inngår et passende utvalg oppgaver av flere typer og i ulike kontekster, og som rettes av ulike sensorer (Norcini, J. mfl., 2018).

Den viktigste anbefalingen er å ha flere oppgaver per eksamen og at hver av dem rettes av forskjellige sensorer. Forskningen fra medisinutdanningen viste at et adekvat utvalg av oppgaver hadde en større

påvirkning på reliabiliteten enn standardisering, slik at et klokt eksamensdesign kan generere reliable resultater innenfor en rimelig tid (ibid.).

Reliabiliteten er i tillegg knyttet til tidsaspektet ved at kortere eksamener er mindre pålitelige enn de som varer lenger. Uavhengig av eksamensform vil elevers prestasjon på én oppgave ikke nødvendigvis forutsi hvordan elevene presterer på andre oppgaver (Wass, Van der Vleuten, Shatzer, og Jones, 2001). Dessuten kan noen eksamensformer være mindre pålitelige enn andre, for eksempel langsvarsoppgaver og muntlige eksamener. En konsekvens av dette er dermed at prøven må være av en viss lengde og dekke tilstrekkelig bredde i kompetansen for å kunne gi resultater som rettmessig kan brukes til eksamensformål. Å ta i bruk et bredere tilfang av eksamensformater innebærer å inkludere former som – når de står alene – muligens er mindre pålitelige, men å aggregere ulike metoder og kontekster ivaretar denne bekymringen (Van der Vleuten, og Schuwirth, 2005).

Eksempel fra medisinutdanningen

Objective Structured Clinical Examinations (OSCE):

- Et multicaseformat bestående av en serie oppgaver og situasjoner (stasjoner)
- Introdusert for å måle høyere kognitiv kapasitet og øke eksamens validitet
- Kandidatene møter simulerte, realistiske utfordringer på datamaskin eller i laboratorium for å styrke autentisitet
- Oppgavene er kontekstuelle og problemorienterte slik at de krever resonnerende ferdigheter
- Bruk av teknologi kan heve kvaliteten av slike eksamener ved å tilby en mer realistisk framstilling av kliniske funn

Oppsummert kan det pekes på at det finnes eksempler på hvordan kompleks kompetanse kan prøves uten fare for kvalitetskrav som validitet, reliabilitet og rettferdighet. Likevel er det en utfordring å ha en eksamen som er standardisert med et bredt kompetansebegrep fordi det er vanskelig å presisere kompetanse på en slik måte at den kan måles reliabelt nok. I medisinutdanningen har de lykkes med det – men systemet har blitt utviklet over lang tid og ved hjelp av betydelige ressurser. Om denne tilnærmingen passer til et så stort system som eksamen på 10. trinnet og etter videregående skole, er ytterligere et spørsmål. Her trengs det en nøye utredning og grundig diskusjon. Samtidig kan sluttvurderingen anses som et helhetlig system der standpunkter tar over viktige oppgaver i kompetanseprøvingen så at robustheten i dagens eksamenssystem som tilfredsstiller krav til reliabilitet, validitet og overordnet sett rettferdighet, kan opprettholdes og styrkes.

11. Teknologiens betydning for eksamen

Den teknologiske utviklingen får betydning for eksamen på ulike måter. Det kan strekke seg fra å distribuere og levere eksamen i et digitalt system til å utvikle eksamensoppgaver på en digital plattform og utnytte mulighetene som ligger i det, noe som i stor grad berører også det vurderingsfaglige aspektet og innholdet i eksamen. Vurdering av besvarelsene kan også støttes av teknologi. Et tilbakevendende diskusjonstema er i hvor stor grad eksamen skal speile den teknologiske utviklingen og den store digitaliseringen som finner sted på de aller fleste samfunnsarenaer, og på hvilke måter dette kan skje. Den teknologiske utviklingen innebærer nye muligheter for vurdering, noen utfordringer og ikke minst forutsetninger både med hensyn til kompetanse og tilgang til digitalt utstyr.

Dette kapitlet oppsummerer det foreløpige kunnskapsgrunnlaget om teknologiens betydning for eksamen med stor vekt på norsk forskning og bærer preg av at vi per dags dato har få erfaringer med å prøve kompetanse digitalt på eksamensfeltet. I tillegg viser det seg en gang til at kunnskapsgrunnlaget i stor grad bare baserer seg på spørreundersøkelser. Det kan settes spørsmåltegn ved om denne tilnærmingen er den riktige om man ønsker å utrede effektene av teknologibruk, eller om vi ikke trenger annen type forskning, for eksempel intervensjonsstudier.

Kapitlet er delt inn i områder som påvirkes av digitalisering, forutsetninger for endring og foreløpige erfaringer med digital eksamen.

11.1 Områder som påvirkes av digitalisering

I dette delkapitlet har vi valgt å trekke fram følgende områder som digital teknologi kan påvirke eller endre ved eksamen:

- Administrasjon og gjennomføring av eksamen
- Teknologistøtte og hjelpemidler til eksamen
- Innhold til eksamen
- Sensurering av eksamen

Administrasjon og gjennomføring av eksamen

Ett aspekt av digitaliseringen handler om å gjøre selve innleveringen eller prøvegjennomføringen gjennom en

digital prosess. Formålet med slik digitalisering er i hovedsak økt effektivitet, informasjonssikkerhet og personvern. Teknologi åpner også for nye formater på produktet som skal vurderes. Lydfiler, video, multimodale tekster og programvare er bare noen få eksempler på digitale produkter som kan være relevante for sluttvurdering. Dagens eksamenssystem er digitalt i den forstand at elevene kan laste ned eksamensoppgavene og levere digitalt. Utviklingen på nasjonalt hold de senere årene har primært handlet om å fornye administrasjonsløsningene for eksamen.

Utdanningsdirektoratets eksamenstjeneste

Prøveadministrasjonssystemet PAS og prøvegjennomføringssystemet PGS ble utviklet av Udir for å utarbeide, gjennomføre og administrere både prøver og sentralt gitte eksamener.

Innføringen av PAS-/PGS-systemene skjedde gradvis fra 2008 og har bidratt vesentlig til å heve kvaliteten på gjennomføringen av sentralt gitt skriftlig eksamen gjennom økt effektivitet og bedre sikkerhet.

Systemene utgjør i dag en digital tjeneste for eksamen som brukes til å utarbeide eksamensoppgaver, hente materiell og til å melde på kandidater til eksamen. Dessuten brukes de under selve eksamensgjennomføringen og til sensur og klagebehandling. Fra høsten 2015 ble systemene også tatt i bruk ved lokalt gitt skriftlig eksamen.

Udir har fornyet administrasjonsløsningen og har startet på en prosess for å kunne anskaffe en ny gjennomføringsløsning innen 2021. Den nye løsningen for utvikling og gjennomføring av eksamen og prøver skal kunne tilby nye oppgaveformater og ny funksjonalitet som legger til rette for å prøve kompetanse på nye måter og gi støtte til sensur.

Teknologistøtte og hjelpemidler til eksamen

Teknologistøtte handler om å benytte digitale verktøy i opplæringen for å understøtte og berike undervisnings-, lærings- og vurderingsprosesser. For eksamen vil det først og fremst handle om å bruke ulike digitale hjelpemidler i eksamenssituasjonen. Slike hjelpemidler kan for eksempel være tilgang til åpent internett, lese-/skrivestøtte eller fagspesifikk programvare. Som nevnt i kapittel 3 har forsøk med eksamen med tilgang til internett i utvalgte fag i videregående opplæring blitt gjennomført årlig fra 2012 til og med 2015. Disse forsøkene er evaluert på oppdrag fra Udir (Rambøll, 2012; Rambøll, 2013; Rambøll, 2014; Rambøll, 2015). Evalueringsrapportene ser på blant annet forberedelse til og gjennomføring av eksamen med internett, opplevd nytte og tilfredshet samt resultater og implikasjoner. Under delkapitlet om erfaringer fra digital eksamen oppsummeres hovedfunn fra sluttrapporten som ble publisert i januar 2019.

Innhold til eksamen

At teknologi gir mulighet for nye oppgavetyper og vurderingsprodukter, har ført til en diskurs om hvorvidt digital vurdering har potensial til å måle kompetanser som tidligere har vært vanskelige å fange, for eksempel knyttet til metakognisjon (Erstad, 2008; Redecker og Johannessen, 2013). Det er imidlertid vanskeligere å finne kunnskap om hvordan dette konkret kan gjøres, og påstandene i diskursen er i liten grad bygget på evidens. Denne økende oppmerksomheten om hvilke kompetanser som kan måles, betraktes imidlertid som et paradigmeskifte i digital vurdering, fra tidligere å være mest opptatt av å benytte teknologi til å effektivisere vurderingsprosesser og øke reliabilitet i skåring (Redecker og Johannessen, 2013).

Hvordan teknologi kan benyttes for å videreutvikle eksamensordningen, er spesielt aktuelt i lys av det nye kompetansebegrepet i fagfornyelsen (Kunnskapsdepartementet, 2016), og det er behov for mer evidensbasert kunnskap på dette området. Teknologiutvikling er også en driver for endring av skolens innhold og derigjennom hvilke kompetanser det er relevant å måle (NOU 2015: 8, 2015). Dette kan vise seg ved at nye områder eller temaer innføres i skolefagene, ved at vektingen mellom innholdsområdene endres, og ved at nye tverrfaglige temaer eller fagovergripende kompetanser finner sin plass i læreplanene. Eksempler på dette er innføring av programmering i matematikkfaget, digitale tekstformer og tekstlige uttrykk i norskfaget, kildekritisk kompetanse og digitale ferdigheter som en grunnleggende ferdighet (Hultin og Berge, 2014).

Sensurering ved eksamen

Teknologi gir mulighet for automatisk skåring av oppgaver og kan dermed være til støtte for sensor ved vurdering av eksamensbesvarelser. Kvaliteten på slik automatisk skåring vil variere med oppgavetype, men for egnede oppgavetyper vil automatisk skåring kunne innebære betydelig tidsbesparelse ved sensur samt gi mindre risiko for skåringsfeil.

En annen mulighet som ligger i digital sensurering, er å lagre data for å kunne utrede sensorreliabiliteten. Hvis data fra alle sensorer ble lagret på elevnivået og per oppgave, ville tilsvarende studier ha tilgang til mer informasjon enn i dag og føre til at en utredning av sensorreliabilitet ville blitt mer innholdsrik og oppklarende.

11.2 Digital kompetanse og forutsetninger

Rapporten *Teknologi og programmering* for alle beskriver hvordan digital teknologi kan brukes til å skape nye muligheter for å bedre kvalitet og effektivitet i lærings- og undervisningsprosesser, men understreker at disse mulighetene har noen forutsetninger og endringsbehov, spesielt knyttet til elevers og læreres digitale kompetanse (Sanne mfl., 2016).

Læreplanen forutsetter at lærere tar i bruk digitale verktøy i undervisningen samt bidrar til å utvikle elevenes

digitale ferdigheter i fag. Dette har vært et premiss i alle fag siden innføringen av de grunnleggende ferdighetene gjennom LK06. For å undervise elever i digital kompetanse trenger lærerne å inneha en profesjonsfaglig digital kompetanse (Utdanningsdirektoratet, 2018a). En del av å ha profesjonsfaglig digital kompetanse er å ha kunnskap om digitale vurderingsformer og ferdigheter til å benytte dem i undervisnings- og læringsprosesser. Dersom elevene er fortrolige med ulike former for digital vurdering, har de et bedre grunnlag for å håndtere en digital eksamenssituasjon, men det forutsetter at lærerne har kompetanse til å inkludere slike vurderingsformer i sin undervisning. De norske forskerne som arbeidet med den internasjonale komparative studien ICILS (International Computer and Information Literacy Study), beskrev at den faglig-pedagogiske kompetansen blant lærerne til å ta i bruk digitale hjelpemidler på kvalifisert vis var mangelfull (Hatlevik og Throndsen, 2015).

ICILS-studien fant også at nærmere en fjerdedel av de norske elevene på 9. trinn har så svake digitale ferdigheter at de vil ha problemer med å kunne delta fullt ut i utdanning, arbeids- eller samfunnsliv (Hatlevik og Throndsen, 2015). En tredjedel av de norske elevene er i stand til å søke etter informasjon, utøve kildekritikk og lage digitale presentasjoner etter nærmere spesifiserte kriterier. Omtrent halvparten av elevene viser at de kan bruke datamaskinen som et redskap og er i stand til å bruke digitale ressurser til å løse enkle oppgaver. De har en viss bevissthet omkring personvern, men viser samtidig mangelfull kritisk vurderingsevne til hvordan personinformasjon på nett kan brukes. En fjerdedel av elevene har kun kjennskap til elementær filhåndtering og tekstredigering. De har bare en overflatisk forståelse av datasikkerhet og nettvett. Fra SMIL-studien ser man at elever i videregående generelt har for lav kompetanse knyttet til faglig bruk av IKT og digitale læringsstrategier (Krumsvik mfl., 2013).

Vi har observert at det er store forskjeller mellom skoler når det gjelder tilgangen til digitale ressurser, og hvorvidt opplæring i disse prioriteres i undervisningen, noe som også framkommer i de nasjonale Monitor-undersøkelsene som kartlegger skolens digitale tilstand (Egeberg, Hultin og Berge, 2016; Hatlevik, Egeberg, Gudmundsdottir, Loftsgarden og Loi, 2013). Tilgangen på ulike former for digitalt utstyr er generelt sett høy i norske skoler. Kvaliteten på maskinene og tilhørende infrastruktur er imidlertid noe variabel, og det er store forskjeller på tilgangen skolene imellom, viser funn fra Monitor-undersøkelsene og ICILS (Egeberg mfl., 2016; Hatlevik mfl., 2013; Hatlevik og Throndsen, 2015). Monitor skole 2016 undersøkte grunnskolers digitale modenhet på organisasjonsnivå og fant at blant de undersøkte faktorene var det på utstyrsområdet det ble rapportert størst spredning i opplevd kvalitet blant skolene som deltok (Egeberg mfl., 2016).

I SMIL-studien som undersøker sammenhengen mellom IKT-bruk og læringsutbytte i videregående opplæring, finner man imidlertid at digitale skiller på dette nivået primært oppstår med grunnlag i elevgruppers bruksmønster og ikke lenger er basert på ulik tilgang til teknologi (Krumsvik mfl., 2013). Dette funnet er i tråd med en generell utvikling, som gjerne beskrives som en overgang fra første- til andregenerasjons digitale skiller (Hatlevik og Throndsen, 2015). Delrapporten Digitale skillelinjer i evalueringen av eksamen i matematikk for 10. trinn undersøker hva slags undervisning elevene har fått i bruk av digitale hjelpemidler som er relevante for matematikkeksamen, og hvordan de har blitt forberedt på å bruke disse på eksamen (Bjørnset, Fossum, Rogstad, Smestad og Talberg, 2018). Rapporten omtaler visse elevgrupper som digitalt privilegert, i den forstand at de har bedre forutsetninger for å lykkes på eksamen enn

andre elever. Dette fortrinnet kan være knyttet til tekniske forhold, som tilgang til utstyr og infrastruktur, eller undervisningsforhold, som omfang av og kvalitet på opplæringen i digitale ferdigheter.

11.3 Erfaringer fra digital eksamen

Fra og med 2012 har det vært normalordningen for skriftlig eksamen i grunnskole og videregående skole å levere eksamensbesvarelsen elektronisk i et digitalt prøvegjennomføringssystem (Utdanningsdirektoratet, 2016). Evalueringer av disse gjennomføringene dreier seg stort sett om tilgang til åpent internett og bruk av digitale hjelpemidler. Våren 2017 inkluderte den halvårlige omnibusundersøkelsen *Spørsmål til Skole-Norge* spørsmål om bruk av nettbaserte hjelpemidler til sentralt gitt eksamen (Federici, Gjerustad, Vaagland, Larsen, Rønsen og Hovdhaugen, 2017). Undersøkelsen viste at omtrent to av tre skoleeiere og skoleledere svarer at de tilbyr nettbaserte hjelpemidler. Nettbaserte hjelpemidler er mest utbredt i videregående opplæring, hvor 88 prosent svarer at de tilbyr dette. Blant grunnskolene svarer 62 prosent at de tilbyr dette. SMIL-studien viser imidlertid at sentrale digitale læremidler i fagene og elevens multimediale og multimodale læringsarbeid i liten grad er fanget opp av eksamensformene i videregående (Krumsvik, Egelandsdal, Sarastuen, Jones og Eikeland, 2013).

Digitale vurderingsformer framhever betydningen av ulike digitale kompetanser, som produksjonskompetanse, verktøykunnskap og sjangerforståelse. Som vi kan se i evalueringen av forsøk med bruk av åpent internett på eksamen i videregående skole, gir en slik eksamensform en «washback-effekt» på opplæringen (Rambøll, 2014). Lærerne ved skolene som deltok i forsøket, bruker i større grad enn lærerne ved referanseskolene internett i undervisningen. Disse lærerne gjennomfører prøver og heldagsprøver hvor elevene har tilgang til internett, og er opptatt av kildebruk og kildekritikk i undervisningen. Denne effekten kan også strekke seg utover det som er definert som kompetansemål i læreplanen. Den kvalitative Monitor-rapporten fra 2010 forteller om lærere på ungdomstrinnet som prioriterer å undervise i formatering av tekst for å mestre digitale formkrav til eksamen, og at dette oppfattes som kompetanse som måles utover det som er skissert i læreplanen (Hatlevik, Tømte, Skaug og Ottestad, 2010).

I januar 2019 ble det publisert en sluttrapport fra evalueringen av åpent internett til eksamen i syv fag fra studiespesialiserende utdanningsprogram i videregående (Rambøll, 2019). I rapporten beskrives i hovedsak funn fra spørreundersøkelser til elever, lærere, eksamensansvarlige og IT-ansvarlige på skoler og i fylkeskommunen. Undersøkelsen ble gjennomført i mai–juni 2018. Sentrale temaer i rapporten er teknisk modenhet og gjennomføringen av eksamen med åpent internett, sensur og regelverk med henblikk på avdekking av fusk ved årets eksamen, autentisitet og relevans knyttet til eksamensordningens samsvar med undervisningspraksis, oppgaveformulering og vurdering samt støtte som viser til skolenes tilrettelegging for elever med særskilte behov.

Oppsummering av hovedfunn fra evalueringen av eksamen med åpent internett (Rambøll, 2019):

- Det er få tekniske eller praktiske utfordringer knyttet til eksamensgjennomføringen.
- De fleste skoler har iverksatt forberedende tiltak som overvåking av internettbruk under eksamen, opplæring av eksamensvakter samt rekruttering av flere og mer digitalt kompetente eksamensvakter.
- 90 prosent av eksamensansvarlige har informert elevene om fusk og plagiat i forkant av eksamen. Det er imidlertid en lavere andel elever som oppgir at de har mottatt denne informasjonen.
- Kvalitative intervjuer indikerer at elevene har god forståelse for fusk og plagiat, men at det finnes gråsonetilfeller som krever avklaring.
- Det er kun rapportert om ett tilfelle av fusk på eksamen i de aktuelle fagene.
- 93 prosent av lærerne i målgruppa oppgir at bruken av internett inngår som en viktig del av elevenes læring i deres undervisning.
- 96 prosent av lærerne i målgruppa oppgir at elevene har fått opplæring i kildebruk.
- Lærerne i målgruppa oppgir i større grad at de gjennomfører andre prøver med tilgang til internett enn det lærere i kontrollgruppa gjør.
- Elevene opplever det som nyttig å ha tilgang til internett på eksamen, men både sensorer og lærere er mer usikre på utbyttet av tilgangen.
- Vårens eksamensoppgaver oppleves som godt egnet for eksamensformen. Samtidig oppgir 62 prosent av sensorene og 36 prosent av lærerne at tilgang til internett fordrer nye oppgaver.
- Én av fem elever opplever eksamensformen som mer stressende enn eksamen uten internett.
- Det gjelder særlig jenter og særlig elever i fagene «Politikk og menneskerettigheter» og «Samfunnsfaglig engelsk».
- Eksamensveiledning og vurderingskriterier oppleves som tydelige blant sensorene.
- En del sensorer rapporterer at de vurderer besvarelser der elever har hatt tilgang til internett, strengere enn besvarelser fra elever uten tilgang til internett, uten at dette kan påvises i karaktergivningen.

- Omfang av støtte til gjennomføring av eksamen er omtrent lik for elever i målgruppa og kontrollgruppa.

12. Lærerutdanningene og vurderingskompetanse

Det hører ikke direkte til eksamensgruppas mandat å utrede lærerutdanningen eller å foreslå endringer om den, men fordi lærerutdanningen har den klart beste muligheten til å bygge opp høy vurderingskompetanse både på formativ og summativ vurdering hos alle lærere så at de også kan sikre validitet, reliabilitet og rettferdighet i sine summative vurderinger på den best mulige måten, har vi bestemt oss for også å omtale lærerutdanningen. Vi vet at lærere er dypt involvert i sluttvurderingen på 10. trinn og i videregående opplæring gjennom å utvikle eksamensoppgaver, å jobbe som sensorer og ikke minst å gjennomføre standpunktvurderingen samt karaktersettingen. Til disse oppgavene trengs det utvidet vurderingskompetanse.

I rammeplan for lærerutdanning 1–7, 5–10 og 8–13 er vurdering omtalt under § 2, som beskriver krav til studieprogrammenes læringsutbytte i tråd med det nasjonale kvalifikasjonsrammeverket. I rammeplanene for trinn 1–7 og 5–10 som ble vedtatt i 2013, er vurdering omtalt i to punkter under henholdsvis temaene kunnskap og ferdigheter. Punktene vektlegger at lærere etter endt utdanning skal ha inngående kunnskap om blant annet vurderings- og kartleggingsverktøy samt vurdering av elevenes læring. Lærere skal også kunne vurdere elevers læring og gi læringsfremmende tilbakemeldinger. Lektorutdanningens rammeplan har blitt vedtatt i 2018 og refererer til det nye kompetansebegrepet samt kjennetegn på måloppnåelse under «Ferdigheter», men nevner ingen kunnskapsområder som lærerstudentene skal undervises i.

> **Forskrift om rammeplan for grunnskolelærerutdanning for trinn 1–7 og om rammeplan for grunnskolelærerutdanning for trinn 5–10**
>
> **§ 2 Læringsutbytte**
>
> Kunnskap
>
> - har inngående kunnskap om begynneropplæring, grunnleggende ferdigheter, vurderings- og kartleggingsverktøy, klasseledelse og vurdering av elevers læring og hva som fremmer læring i fagene
>
> Ferdigheter

- kan analysere, vurdere og dokumentere elevers læring, gi læringsfremmende tilbakemeldinger, tilpasse opplæringen til elevenes forutsetninger og behov, bruke varierte undervisningsmetoder og bidra til at elevene kan reflektere over egen læring og utvikling

Forskrift om rammeplan for praktisk-pedagogisk utdanning allmennfag og om rammeplan for lektorutdanning for trinn 8–13

§ 2 Læringsutbytte

- Kunnskap

Ferdigheter

- kan beskrive kjennetegn på kompetanse, vurdere og dokumentere elevers læring, gi læringsfremmende tilbakemeldinger og bidra til at elevene kan reflektere over egen læring og egen faglige utvikling

Forskrift om rammeplan for praktisk-pedagogisk utdanning for yrkesfag og for trinn 8–13

§ 2 Læringsutbytte

- Kunnskap

Ferdigheter

- kan vurdere og dokumentere elevers læring og utvikling, gi læringsfokuserte tilbakemeldinger og bidra til at elevene/lærlingene kan reflektere over egen læring

Lærerutdanningen har endret seg betydelig de siste årene. Dette gjelder særlig for grunnskolelærerutdanningen der studentene tar femårig masterutdanning fra og med høsten 2017. Vi forholder oss til de nyeste rammeplanene og retningslinjene i dette kapitlet, og begrenser oss til de fem største programmene. I rammeplan for grunnskolelærerutdanning 1-7 og 5-10, praktisk-pedagogisk utdanning allmennfag og yrkesfag (PPU-A, PPU-Y) og lektorutdanning 8-13 er vurdering omtalt under § 2 som beskriver krav til studieprogrammenes læringsutbytte i tråd med det nasjonale kvalifikasjonsrammeverket.

I rammeplanene for trinn 1-7 og 5-10, som ble vedtatt i 2016, er vurdering omtalt i to punkter under henholdsvis temaene *kunnskap* og *ferdigheter*. Punktene vektlegger at lærerstudenter etter endt utdanning skal ha inngående kunnskap om blant annet vurderings- og kartleggingsverktøy samt vurdering av elevenes læring. Lærerstudenter skal også kunne vurdere elevers læring og gi læringsfremmende tilbakemeldinger. Det siste inngår også rammeplanen for PPU-Y, som ble vedtatt i 2013. PPU-As rammeplan og lektorutdanningens

rammeplan ble vedtatt i henholdsvis 2015 og 2013. Begge to referer til kompetansebegrepet samt kjennetegn på måloppnåelse under ferdigheter, men nevner ingen kunnskapsområder relatert til vurdering som lærerstudentene skal undervises i.

Karaktersetting eller sensurering er ikke eksplisitt nevnt i noen av de rammeplanene som et område der lærere trenger spesifikk kompetanse. Det er en rekke andre momenter i disse punktene og andre læringsutbytter i rammeplanene som forutsetter vurderingsfaglig kompetanse, som å tilpasse opplæringen, vite hva som fremmer læring og sikrer progresjon, men som i mindre grad er synliggjort direkte som «vurdering».

De nye nasjonale retningslinjene for lærerutdanningen som ble vedtatt i Universitets- og Høgskolerådet – Lærerutdanning (UHR-LU) i 2017 – sier lite spesifikt når det gjelder vurdering i fellesdelen for alle programmene, det eneste som kreves, er at «lærarutdanningane skal kvalifisere studentane til å kunne foreta etisk grunngjevne vurderingar». Retningslinjene for lektorutdanningen gir derimot tydeligere uttrykk for kunnskap om og ferdigheter i vurdering (NRLU, 2017). Hovedvekten ligger på vurdering for læring og underveisvurdering, men sluttvurderingen er eksplisitt nevnt som et kompetanseområde i den fagdidaktiske delen der studentene skal «lære å gi elevene underveisvurdering og sluttvurdering, bruke faglige kjennetegn på måloppnåelse og gi gode begrunnelser for vurdering i faget». I tillegg er det praksisopplæringen som får i oppgave å sikre at studenten «har erfaringsbasert kunnskap om elevers læringsprosesser og vurdering for og av læring». Retningslinjene for PPU-A nevner vurdering som et gjennomgående tema som må ivaretas av institusjonene og tilføyer «bred kunnskap om undervisnings-, arbeids- og vurderingsformer generelt og fagspesifikt» som læringsutbytte. I tråd med lektorutdanningen skal lærerstudentene i PPU kunne «gi underveis- og sluttvurdering» samt begrunnelser, og de skal få en sjanse til å «prøve ut formativt og summativt orienterte vurderings- og eksamensformer som de selv kan bruke som lærere» (NRLU, 2017b). PPU-Y legger også vekt på det siste og nevner i tillegg praktisk eksamen (NRLU, 2018).

Det er svært begrenset med forskning og oppdatert systematisk informasjon eller kunnskap om hvordan disse kvalifikasjonskravene knyttet til vurdering ivaretas i lærerutdanningene. Vi vet heller ikke mye om læringsutbytte eller effekten av lærerutdanningen når det gjelder vurderingskompetanse. Dette gjelder i særlig høy grad i forbindelse med summativ vurdering, karaktersetting og sensurering.

Fra en eldre spørreundersøkelse kommer det fram at lærerutdanningene får laveste karakterer på resultatkvaliteten blant annet når det gjelder vurderingskompetanse (Finne mfl., 2011). Særlig skoleledere evaluerer denne delen av utdanningen betydelig mindre positivt enn for eksempel utdanning i sosialkompetanse og profesjonsidentitet, men denne forskjellen i evalueringen av resultatkvalitet gjelder også for lærerstudenter og lærerutdannere. Universitets- og Høgskolerådet (2011) refererer i en egen rapport fra samme året til disse resultatene og krever klarere nasjonale føringer og kontroll av om en faktisk retter seg etter disse føringene for blant annet utdanning i vurdering. En undersøkelse fra 2013 viste at de fleste lærerutdannere rapporterer god kunnskap om vurdering *for* læring[8] men det er uklart hvilke slutninger vi kan trekke om deres kunnskap om summativ vurdering ut fra dette.

En litt nyere rapport om endringer i lærerutdanningene omtaler ikke endringer på detaljnivået, det er dermed ikke kjent om de nye lærerutdanningsmodellene vil skåre bedre når det gjelder vurderingskompetanse

(Munthe, mfl., 2014). Men programmet til ProTed – Senter for fremragende lærerutdanning – kan muligens forstås som en indikasjon på at hovedoppmerksomheten i lærerutdanningen er rettet andre steder enn på summativ vurdering, karaktersetting og sensurering. ProTed er Norges første senter for fremragende utdanning og et langsiktig samarbeidsprosjekt mellom Universitet i Oslo og Universitetet i Tromsø, finansiert av NOKUT. Senterets oppgave er å fremme kvalitet i høyere utdanning, og i tråd med denne oppgaven har ProTed utviklet imponerende tiltak innen lærerutdanning. Flere prosjekter inkluderer også utdanning i vurdering, men disse dreier seg – så vidt det er mulig å se basert på årsrapporter og andre dokumenter – om formativ vurdering, også kalt vurdering for læring (se f.eks. ProTed, 2016; 2017).

Generelt sett finnes det også lite kunnskap om hvorvidt lærerstudenter tar del i arbeidet med vurdering når de er ute i praksis. Praksisperioder harmonerer for eksempel tidsmessig i liten grad med når sluttvurdering skjer, altså mot semesterslutt høsten og våren da lærerstudenter som regel selv er i intense eksamens- og vurderingsperioder. Det er også et spørsmål om hvor tilgjengelig skolens arbeid med vurdering er for studentene, da mye av arbeidet for eksempel knyttet til standpunktsetting og fram mot eksamen skjer individuelt blant lærere og/eller i deres kontortid/møtetid – som ikke nødvendigvis er like tilgjengelig for lærerstudentene i praksis. Dette er imidlertid forhold som vi vet lite om, og som det er behov for å samle mer systematisk kunnskap om.

I UHR-LU[9] blir det tidvis arbeidet med vurdering i lærerutdanningene, men da er det spesielt vurdering av lærerstudentene som vektlegges. Grunnideen er at utdanningen kan bidra til å gi gode eksempler på arbeid med undervisning og vurdering, som følge blir det viktig at man er svært bevisst på hvordan vurdering utøves også innenfor lærerutdanningene. Eksempler på denne ideen er prosjekter fra UiOs og UiTs nevnte samarbeid, ProTed – Senteret for fremragende lærerutdanning, en rapport fra karakterundersøkelsen i matematikk i GLU-utdanningene i 2014 (Arbeidsgruppe, 2015) og Lærerutdanningskonferansen 2019 om framtidsrettet vurdering i lærerutdanningene. Særlig har det hos ProTed blitt utviklet nye vurderingsmåter ved å bruke nettbrett til vurdering av lærerstudenter i praksisfasen og automatisk tilbakemelding ved eksamen (NOKUT, 2015).

Generelt sett er det viktig å peke på at lærerutdanning bare er en grunnutdanning, og at læreres læring også skjer gjennom uformell etterutdanning og gjennom formell videreutdanning, for eksempel i regi av Kompetanse for kvalitet. Flere nyere politiske dokumenter understreker behovet for samspill mellom grunnutdanning, videreutdanning og samarbeid i profesjonsfellesskapet, for eksempel strategien Lærerutdanning 2025[10] . Det finnes regionalt arbeid med kompetanseutvikling for lærere i skolen, for eksempel SKUV-prosjektet[11] i Trøndelag ved NTNU. Tiltaket er et eksempel på et partnerskap mellom skoleeier og universitet initiert fra praksisfeltet.

Ellers får lærerutdanningene og lærerne lite støtte i jobben med summativ vurdering gjennom allmennpedagogiske eller fagdidaktiske lærebøker. Det fins innføringsbøker på engelsk, men disse er dårlig tilpasset norske forhold, særlig det norske standpunktvurderingssystemet. Det er lite systematisert kunnskap om utvikling av lærernes vurderingskompetanse i praksis, unntatt de store utviklingsprogrammene som har dreid seg om formativ vurdering, for eksempel nasjonale satsinger på vurdering for læring og *Ungdomstrinn i*

utvikling (som hadde vurdering for læring som et gjennomgående tema). Flere fylkesmenn arrangerer sensorskoleringer og samlinger på standpunktvurdering, både på eget initiativ og i samarbeid med Udir. Det er imidlertid påfallende at det ikke finnes tilsvarende store kompetanseutviklingsprogrammer når det gjelder summativ vurdering og karaktersetting som det har vært på formativ vurdering de siste årene.

---

8) https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/2013/ntnu_a_bidra_til_skolebasert_kompetanseutvikling.pdf

9)https://www.uhr.no/temasider/nasjonale-retningslinjer-for-larerutdanningene/

10) https://www.regjeringen.no/contentassets/d0c1da83bce94e2da21d5f631bbae817/kd_nasjonal-strategi-for-larerutdanningene_nett.pdf

11) https://www.ntnu.no/ilu/skuv

13. Status for kunnskapsgrunnlaget og problemstillinger ved eksamenssystemet i Norge

Dette sluttkapitlet av rapporten presenterer eksamensgruppas vurdering av kunnskapsgrunnlaget om eksamen. Den første delen presenterer hovedkonklusjonene og hva som følger av den. Den andre delen oppsummerer kort de viktigste funnene av hvert kapittel. Basert på disse funnene identifiseres det i tredje kapittel problemstillinger og spørsmål som bør stilles om eksamenssystemet, og som eksamensgruppa ønsker å utrede i det videre arbeidet. Ytterlige to delleveranser skal utvide og komplettere disse første drøftingene slik at det skal bli mulig å gi råd om fagenes læreplaner i delleveranse 2 i løpet av mars 2019 og å komme med anbefalinger til endringer i eksamensordningen som følge av fagfornyelsen og den teknologiske utviklingen som skal leveres for beslutning 15. mai 2019.

13.1 Status for kunnskapsgrunnlaget og hovedkonklusjoner

I denne rapporten har vi sammenstilt et kunnskapsgrunnlag om eksamenssystemet i Norge. Sammenstillingen har tatt utgangspunkt i dagens eksamenssystem, hvordan det har vokst fram, og dets

offisielt definerte formål. Vi har inntatt et bredt perspektiv på kvaliteten i dagens eksamenssystem, dette inkluderer kriterier som validitet, reliabilitet og rettferdighet samt elevperspektivet. Vi har også sett på forholdet mellom eksamenskarakterer og standpunktkarakterer som to ulike sluttvurderinger. Læreplanverket og forskrift til opplæringslova gir rammer for innhold, organisering og vurdering av eksamen og føringer for kvaliteten i eksamenssystemet og er derfor tatt med som et viktig perspektiv.

Med et forbehold om at dette er en foreløpig dokumentasjon av kunnskapsgrunnlaget, er en hovedkonklusjon at det finnes en del brukerinnsikt og erfaringsbasert kunnskap om eksamen, men at det er store forskningshull på feltet. Vi må få understreke at mangel på forskning ikke nødvendigvis betyr at det ikke skjer mye godt kvalitativt arbeid på eksamensfeltet. Men det mangler likevel et systematisk forskningsbelegg om det. Der det finnes studier dreier det seg nesten utelukkende om spørreundersøkelser som gjenspeiler hva de involverte mener eller husker, men som ellers har noen mangler når det gjelder en systematisk utredning av prosesser og effekter samt langsiktige konsekvenser. Robust informasjon om prosesser, effekter og konsekvenser trenger eksperimenter og designforsøk i skoler og bør belyses både kvantitativt og kvalitativt.

Tatt i betraktning den betydningen eksamen har for den enkelte elev og statusen i et samfunnsperspektiv, har det vært relativt lite forskning på kvaliteten på eksamen. Dette står i kontrast til den oppmerksomheten og forskningen som har vært utført på de store internasjonale undersøkelsene og de nasjonale prøvene og den offentlige oppmerksomheten eksamen får hvert år. For eksamensgruppas del innebærer et mangelfullt kunnskapsgrunnlag at det blir vanskelig å svare utfyllende på de mest nærliggende utfordringene. Det kan være krevende å generalisere ut fra resultater fra andre land for å si noe om Norge – eller fra UH-sektoren til grunnopplæringen – fordi konteksten og rammebetingelsene er ulike. Men denne forskningen gir likevel indikasjoner på hva som kan være relevant for det videre arbeidet i dette oppdraget.

Gjennomgangen av kunnskapsgrunnlaget gir et sterkt fundament for å etterspørre en mer helhetlig tilnærming til sluttvurderingen. Dette gjelder både til forholdet mellom eksamen og standpunkt og til kvalitetssikringen:

- Når sluttvurderingen blir planlagt som et helhetlig og koordinert system, kan det komme tydeligere fram hvilken kompetanse som skal prøves til eksamen og hvilken kompetanse som skal ivaretas gjennom standpunkt. Vi trenger derfor et helhetlig rammeverk for sluttvurdering, som relaterer de ulike sluttvurderingsordningene til det nye kompetansebegrepet, slik at hele kompetansebegrepet prøves systematisk.
- Kvalitetssikring av eksamen bør også ses på og planlegges på en helhetlig måte. Det finnes noen rutiner for å overvåke kvalitet, men ikke alle data er bearbeidet til dokumentasjon som er gjort tilgjengelig. Hvis det ikke ligger et rammeverk til grunn, er det vanskelig å se om kvalitetssikringen skjer på en helhetlig måte. Det kan være nyttig å bruke et av de etablerte rammeverkene for å oppnå dette formålet (f.eks. AEA Europe, 2017; Stobart, 2009).
  En utfordring vil være å prioritere mellom ulike hensyn. Det er for eksempel vanskelig og tids- og ressurskrevende å sikre både høy validitet, høy reliabilitet og høy rettferdighet (karaktersetting uten systematiske avvik for enkelte grupper) fordi tiltak som skal styrke validiteten, kan svekke

reliabiliteten (og omvendt). Noen kvalitetskriterier er dessuten tydelig relatert til selve eksamen (validitet, reliabilitet og rettferdighet), mens andre er relatert til større sammenhenger som er vanskeligere å kontrollere (tolkning av resultater og konsekvenser ved eksamen i praksis). Det er i tillegg sannsynlig at prioriteringene ser annerledes ut i forskning om eksamen ut fra et målingsperspektiv – som ofte vektlegger kvalitetssikring av selve prøvene (f.eks. konstruktvaliditet, sensorreliabilitet), helst i forkant av implementeringen – sammenlignet med forskning ut fra et skoleperspektiv – som ofte ser på hvordan eksamen regulerer handlinger og kunnskap i praksis (konsekvensvaliditet). En tilnærming i slike tilfeller kan være å samkjøre de to perspektivene for å bedre balansen mellom dem.

En grundig kvalitetssikring trenger dessuten en ny tilnærming til datalagring. Når det gjelder for eksempel sensurering, finnes det i dag bare data på et overordnet nivå (karakterene fra sensorene på en eksamen), mens gode analyser trenger data på minste nivå, altså en registrering av poeng per eksamensoppgave eller vurderingskriterium og sensor. Slik data gjør det mulig å utrede nøyere mulige årsaker til enighet/uenighet blant sensorer. Det er for eksempel mulig at sensorer vurderer ulikt på en oppgave enn på en annen (ustabil sensurering, lav intra-sensorreliabilitet), eller det er mulig at størrelsen av uenighet varierer med oppgavetype fordi sensorene vektlegger ulike aspekter. Og så er det mulig at sensorenes vurderinger korresponderer per oppgave eller vurderingskriterium, men resulterer i ulike karakterer basert på forskjeller i den holistiske vurderingen. Hvis det bare finnes data på karakternivået, er detaljene som inngår i karakterene ikke synlig slik at det blir vanskelig å ta tak i årsaker for uenigheten, for eksempel gjennom sensorskolering.

Gjennomgangen av kunnskapsgrunnlaget viser også at det er et stort kunnskapsbehov om tre helt sentrale temaer i rapporten:

- Med bakgrunn i at *validitet* (gyldighet) er det viktigste kvalitetskriteriet i eksamenssammenheng, har dette perspektivet hatt høy prioritet i dette kunnskapsgrunnlaget. Selv om det er mye kunnskap å finne om den teoretiske tilnærmingen, og selv om kunnskapsgrunnlaget utvikler en tydelig forståelse av begrepet validitet, har det vært vanskelig å lete fram studier som utreder gyldigheten i dagens eksamenssystem i Norge. Eksamensnemndene er i dag en viktig del av dagens system for sikring av validitet (eksamens innhold), men vi har lite systematisert forskning for eksempel om eksamens innholdsvaliditet per fag eller om eksamen som gis ulike elevkull. Det er en vesentlig rettferdighetsdimensjon at eksamen måler tilsvarende kompetanse, all den tid elevene konkurrerer på tvers av årskull i forbindelse med studieopptakene.

  Eksamens *reliabilitet* (pålitelighet) er bedre undersøkt, i det minste når det gjelder matematikk og norsk skriftlig og karaktersetting. Det finnes likevel i liten grad studier om muntlig eller andre prøveformer, om andre fag enn norsk og matematikk eller om studier der sensurering av hver oppgave, og ikke bare karaktersetting som siste trinnet, utredes.

  Forholdet mellom eksamens- og standpunktkarakterer er godt undersøkt, når det gjelder avvik i karakterene på tvers av utvalgte kriterier, men det finnes lite forskning om hvordan det er mulig,

eller om det er nødvendig, å motvirke eller kompensere systematiske avvik, for eksempel på tvers av kjønn, regioner eller fag. Hvis forskjellene er knyttet til noe annet enn elevenes kompetanse, kan de svekke rettferdighetsdimensjonen. Overordnet gjenspeiler dette forskningshullet en generell uklarhet rundt forholdet mellom eksamen og standpunkt.

- Elevperspektivet er også lite undersøkt. Det finnes nesten ingen forskning som direkte utreder elevens oppfatning av eksamensformer eller eksamensoppgaver i Norge. Ofte rapporterer bare lærerne om sine inntrykk av elevenes subjektive vurdering av eksamensformer og oppgaver. De få studiene vi har, tyder på systematiske forskjeller i oppfatninger av og mestring av ulike eksamensformer og at oppgaver samt instruksene noen ganger blir oppfattet som utydelige. Det trengs etter alt å dømme flere og mer systematiske studier av hvordan elevene tolker og mestrer oppgavene. Det er i tillegg vanskelig å skille mellom årsakene som muligens ligger bak eksamensengstelse eller press, fordi individuelle forutsetninger for å takle stress og forhold utenfor skolen er sjelden tatt med i de eksisterende studiene.
- Selv om utdanningssektoren har fått erfaring og kompetanse på å utvikle og vurdere eksamen ut fra kompetansebaserte læreplaner i Kunnskapsløftet, vil det nye kompetansebegrepet stille nye krav til prøveutvikling og sensurering. Forskning fra både andre land og UH-sektoren kan gi oss en første pekepinn på hvordan disse kravene kan møtes, men om disse anbefalingene virkelig passer til det norske eksamenssystemet, er et åpent spørsmål og må utredes. De store endringene i kjølvannet av fagfornyelsen skal evalueres, og det samme bør gjelde ved mulige endringer i eksamenssystemet og effektene av dette.

De viktigste konklusjonene på bakgrunn av denne rapporten er derfor at sluttvurderingen og dens kvalitetssikring trenger en helhetlig tilnærming, og at det er et stort behov for forskning som utreder fagfornyelsens og eksamenssystemets forutsetninger, prosessene og resultatene. Begge tiltakene kan bidra til en annen type diskusjoner om eksamen enn den vi har i dag. For å sikre validitet, reliabilitet og et rettferdig eksamenssystem er det behov for utprøvinger og tid til å tenke gjennom om slutningene som trekkes fra vurderingen kan anses som legitime. Dette er spesielt viktig når det gjelder *high-stakes*-situasjoner som eksamener.

13.2 Oppsummering av kunnskapsgrunnlaget

Om framveksten av dagens eksamenssystem

Norge har lange tradisjoner for at opptak til videre utdanning baseres på eksamener forvaltet av lærerprofesjonen selv i et tett samspill med nasjonale myndigheter. Eksamenssystemet har historisk representert statlige myndigheters viktigste verktøy for å styre og kontrollere lærernes karaktersetting. Den

historiske gjennomgangen viser at lærerne har vært anerkjent som kompetente til å vurdere kvaliteten på elevenes prestasjoner og på denne måten hatt et stort ansvar for å kontrollere adgang til videre utdanning og yrkesliv. Eksamenssystemet i sin helhet og de viktigste prosedyrene har vært relativt stabile de siste tiårene, men vurderingskriterier har vært mye omdiskutert, og man har gått bort fra normrelatert og over til et målrelatert vurderingsprinsipp.

Om eksamens formål og organisasjon

Eksamens formål fremgår av forskrift til opplæringslova, der eksamenskarakterer, på lik linje med standpunktkarakterer, skal være et uttrykk for elevenes kompetanse ved avslutningen av opplæringen i et fag. Eksamenskarakterene, sammen med standpunktkarakterene, gir grunnlag for inntak til både videregående og høyere utdanning. Dette gir eksamen en formell funksjon utover å være et uttrykk for elevens sluttkompetanse i fag. Det kan argumenteres for at legitimiteten til en eksamen som en del av et rangeringssystem står og faller på at eksamenskarakteren er et uttrykk for elevens kompetanse og er et likeverdig uttrykk uavhengig av fag. Samtidig kan det argumenteres for at legitimiteten til kravet om at eleven viser sin kompetanse og at den i tillegg tallfestes står og faller på at eksamenskarakterene blir brukt til noe meningsfullt (som for eksempel opptak til videregående og høyere utdanning). Rollen i inntakssystemet vil derfor være viktig i diskusjonen i hva eksamen skal være og ikke være.

Eksamens- og standpunktkarakterene er i dag begge et uttrykk for elevens kompetanse ved avslutningen av opplæringen i faget, men de må anses som ulike. Det kan framstå som uklart i forskrift til opplæringslova om eksamen skal prøve hele bredden i læreplanen. Denne uklarheten kan medføre at de ulike aktørene i systemet tolker forholdet mellom eksamen og standpunkt ulikt.

Trekkordningen innebærer at elever blir fordelt, i all hovedsak basert på tilfeldige utvalg. Kunnskapsgrunnlaget viser at denne ordningen kanskje ikke gir alle elevene mulighet til å vise kompetanse på en valid måte. Den er heller ikke forenlig med tenkningen om sluttvurdering som et helhetlig system. Elevene selv kan oppfatte fordelingen på ulike eksamener som urettferdig fordi de ikke får den samme sjansen til å vise kompetansene sine. Attpåtil kan fag og antall eksamener på vitnemålet variere, noe som kan få utslag på gjennomsnittsberegningen for opptak til videregående opplæring og høyere utdanning.

Privatistordningen er et tilbud om å dokumentere kompetanse i et fag man ikke tidligere har fått opplæring eller sluttvurdering i, eller et tilbud dersom man ønsker å forbedre karakterer. Antallet forbedringsprøver har vokst over tid og utgjør i dag en betydelig andel med privatister, noe som setter spørsmåltegn ved formålet med denne ordningen. I tillegg er ordningen administrativt utfordrende å gjennomføre, noe som i mange tilfeller får konsekvenser for muligheten til videreutvikling av eksamen.

Selv om eksamenssystemet har vært relativt stabilt, har noen utprøvinger og endringer funnet sted de seneste årene som en følge av innspill fra brukere, embetene og fagmiljøer. Tiltak som nye eksamensformer eller tilgang til hjelpemidler gjenspeiler at eksamensordningen i et fag kan begrense elevenes mulighet til å vise sin sluttkompetanse på en god måte, og at den økende tilgangen på hjelpemidler krever nye diskusjoner om hva som skal vurderes, og på hvilke måter.

Kvalitet i dagens eksamenssystem

Sluttvurderingen skal gi rettferdig og relevant informasjon om elevenes kompetanse i fag. For å oppnå dette trenger lærere og sensorer støtte i sine vurderinger gjennom tydelige mål, vurderingskriterier, veiledning og kvalitetssikring. Denne rapporten utreder ulike kvalitetskriterier (særlig validitet, reliabilitet, rettferdighet) og bruker en helhetlig tilnærming som ivaretar samsvar i kvalitetskriteriene og ser på hele prosessen, fra utvikling av eksamensoppgavene, via administreringen av eksamen og fastsetting av resultatene, til måten resultatene blir tolket på og anvendt i praksis. En utfordring for eksamen, hvor oppgaver må være hemmelige før gjennomføring, er at det er vanskelig å vite at en eksamensoppgave har den ønskede kvaliteten før den tas i bruk, for eksempel å teste gjennom pilotering.

Om eksamens validitet

Det er nesten bare validiteten til matematikkeksamen som har blitt utredet til en viss grad. Her er de fleste lærerne enige i at det er godt samsvar mellom kompetansemålene og hva elevene blir prøvd i. Funn fra KAL-prosjektet forteller noe om validitet i norsk skriftlig eksamen, selv om denne undersøkelsen ligger en del tilbake i tid (Berge mfl., 2005). For øvrig er det altså lite forskning på sammenhengen mellom læreplan og eksamen og hvorvidt det er forskjeller mellom fag og de ulike eksamensformene. Blant skoleledere og skoleeiere er det uenighet om eksamen er egnet til å vise kompetanse i allefag. Det er også uenighet om hvorvidt det er klart hvilkenkompetanse elevene skal vise til eksamen.

Kunnskapsgrunnlaget viser at dagens eksamenssystem kan ha flere implisitte roller utover formål definert i lovverket, for eksempel i sertifisering, seleksjon, kvalitetssikring, videreutvikling av vurderingspraksis, styring av undervisningen og til og med støtte av læring i norsk grunnopplæring. Flere og ulike formål og roller kan føre til ulike tolkninger av eksamensresultater og ulike bivirkninger ved endringer. Det er derfor viktig å avklare de implisitte rollene eksamen har i praksis. Imidlertid finnes det lite forskning som utreder dette, og noen av rollene eksamen har i utdanningssystemet, er derfor muligens underkommunisert. Dette gjelder for eksempel eksamens rolle i kompetanseheving av sensorene og eksamens bidrag til profesjonalisering av vurderingen. Forskningen dokumenterer at skoleledere og lærere oppfatter deltakelse i sensur som viktig for den profesjonelle utviklingen til både skolen som organisasjon og den individuelle lærer.

Om eksamens reliabilitet

Det er et viktig kvalitetskjennetegn at en eksamensoppgave får samsvarende vurderinger av flere sensorer, slik at karaktersettingen ikke er preget av tilfeldigheter. Dette krever tydelige oppgaver med gode instrukser, tydelige vurderingskriterier (i.e. kjennetegn på måloppnåelse) og omfattende sensorskolering for å sikre tolkningsfellesskap.

Kjennetegn på måloppnåelse fra eksamener med sentral sensur blir mye brukt og oppleves som nyttig i skolenes vurderingsarbeid. Skoleledere og skoleeiere har i tillegg utarbeidet lokale kjennetegn. Kjennetegnene er formulert på karaktergruppene 2, 3–4 og 5–6. Fordi karakterene 3 og 4 utgjør en særlig stor andel ved karakterene, har lærere etterlyst klarere vurderingskriterier som skal gjøre det enklere å skille mellom en 3-er

og en 4-er.

Lærere som har deltatt i sensorskolering, opplever dette som svært nyttig, og skolelederne opplever at sensorenes erfaringer bidrar til å heve vurderingskompetansen ved skolen. Mangel på samsvar mellom sensorene i vurderingen av eksamen ser likevel ut til å være et problem i flere fag. Elevenes besvarelser og sensorenes vurdering profitterer på eksplisitte forventninger, tydelige formål og detaljerte krav til innhold og struktur og vekting av kriterier. At eksamen består av et større antall oppgaver, er det viktigste for å oppnå høy reliabilitet og er viktigere enn å standardisere oppgaver. I tillegg bør oppgavene innen en eksamen vurderes av forskjellige lærere og sensorer.

Det er lite systematisk forskning på hvordan kommuner og fylkeskommuner arbeider kvalitativt med sensuren av lokalt gitt eksamen. Brukerinnsikt og spørreundersøkelsene viser at det er ulike former for samarbeid når det gjelder vurdering, men studiene sier ikke noe om kvaliteten i samarbeidsarenaene og i hvilken grad / måte dette arbeidet er knyttet til lokalt gitt eksamen.

En utfordring med dagens datagrunnlag fra forskning om eksamens reliabilitet er at sensorinformasjon om hvert fag bare finnes samlet på elevnivå, men ikke på oppgavenivå innen en elevs eksamen. Det gjør det vanskelig å i etterkant utrede hva som kan være årsaker til mulige problemer med mangel på sensorsamsvar.

Om systematiske forskjeller mellom eksamen og standpunkt

Forskjeller mellom eksamen og standpunkt trenger ikke i seg selv være en grunn til bekymring med mindre det dreier seg om systematiske forskjeller mellom grupper. Kjønnsrelaterte forskjeller mellom eksamen og standpunkt, forskjeller mellom private og offentlige skoler, store og små skoler, høytpresterende og lavtpresterende skoler samt forskjeller på tvers av regioner viser at det er systematikk i ulikheten i hva standpunkt og eksamen måler som ikke med rimelighet kan knyttes til elevenes faglige sluttkompetanse. Slike forskjeller skaper en situasjon der de ulike elevene gis ulike muligheter, noe som ikke er forenlig med tanken om en rettferdig vurdering.

Det er lite forskning om årsakene til de systematiske forskjellene. Eksempler tyder på at det kan være svakheter ved prosedyrer eller fagkulturelt betingede normer i sluttvurderingen i norsk skole. Eksamen og ekstern sensur kan tenkes å svekke de uheldige bieffektene hvis standpunktkarakterer gjenspeiler faktorer utover læreplanmålene, for eksempel elevenes innsats eller orden og oppførsel.

I tillegg til systematiske forskjeller på tvers av elevgrupper kommer forskjeller mellom eksamen og standpunkt knyttet til årskull og fag. Slik variasjon i nivået av karaktersetting kan slå ut som en kilde til ikke-fair konkurranse om de samme studieplassene. Et annet fenomen som kan få konsekvenser for opptak til høyere utdanning og yrkesliv, er skjevheter på bakgrunn av fagenes uttelling på vitnemålet ut fra timetall.

Om vurdering i fag

Fagenes egenart har i svært begrenset grad vært i vurderingsforskningens sentrum. Når vi skal prøve kompetanse i tråd med de nye læreplanene i fagfornyelsen, vil det bli viktig å anerkjenne skolefagenes

innhold og struktur, noe som er ofte et undervurdert forhold. Det finnes fag som er tydelig disiplinært forankret i fagdomener i høyere utdanning, og de har samtidig en strammere struktur med hierarkisk og sekvensiell oppbygning. På den annen side finnes det fag som har svakere kobling til sine akademiske referansedisipliner, er mindre hierarkiske og mer segmentert. Fagfornyelsens idé med å definere kjerneelementer kan bidra til å anerkjenne fagenes innhold og struktur.

Det er skolefag som utgjør grunnlaget for læreres og sensorers vurdering. Vurdering i fag er basert på smalere eller bredere grunnlag, der den smale tilnærmingen gjerne er dominert av bruk av bare en vurderingsform, for eksempel skriftlig eksamen. Den brede tilnærmingen domineres av et bredere utvalg av vurderinger, for eksempel skriftlig og muntlig eller muntlig og praktisk prøving ved eksamen. I norsk kontekst er det mye som tyder på at det er eksamensformen for det enkelte fag som bidrar til å definere disse mer smale eller brede rammene for vurdering.

Om elevers opplevelse av eksamen

Elever har gjennom Norsk Gymnasiastsamband og Elevorganisasjonen påpekt mangler ved eksamenssystemet helt tilbake til 1963. Manglene som er påpekt, er blant annet at elevene ikke opplever å få vist sin fulle kompetanse, at dagsform i betydelig grad påvirker elevenes prestasjoner, og at det i stor grad er tilfeldig hvilket fag eleven blir trukket opp i. Forskning tyder på at det å lytte til elevene, for eksempel om de oppfatter oppgaver eller instrukser som klare eller uklare, kan bidra til økt validitet i utviklingen av eksamensoppgaver.

Kunnskapsgrunnlaget viser at elevers prøveengstelse er økende når elever opplever at testen vil ha stor betydning, når den er forventet å være vanskelig, og når forholdene rundt prøvesituasjonen er stressende. Elever foretrekker vurderingsformater som reduserer stress og nervøsitet, men det finnes ikke en ideell vurderingsform. Elevenes preferanser varierer blant annet med graden av oppgavenes åpenhet, kjønn, prestasjonskrav og læringsstrategier. Dette tyder på at det er viktig å variere eksamensformer så mye som mulig så at ulike elevgrupper har mulighet til å prestere på den best mulige måten.

Om prøving av fagfornyelsens utvidede kompetansebegrep

Kompetanseorienterte eksamener tar sikte på å måle komplekse evner og kunnskap. Fagfornyelsens kompetansebegrep framhever det å forstå, å anvende kunnskap og ferdigheter i kjente og ukjente sammenhenger, det å lære å lære og å reflektere over egen læring. Dybdelæring kan anses som en forutsetning for å utvikle kompleks kompetanse. Utviklingen av kompetanseorienterte eksamener er imidlertid ofte krevende fordi komplekse evner og kunnskap som regel er mindre presist definert, og fordi det ikke alltid er mulig å utvikle klare kriterier som definerer hvorvidt et svar er riktig eller galt.

Erfaring fra medisinutdanningen viser at før eksamensformen velges, er det nyttig å forestille seg i hvilke situasjoner elevene skal ha bruk for denne kompetansen senere i livet, og hva slags oppgaver som egner seg til å vurdere denne kompetansen. Kompleksiteten som ligger i fagfornyelsens kompetansebegrep, er nesten umulig å prøve med en enkel eksamen eller eksamensform, men krever at man tenker helhetlig på sluttvurderingen som et system. Mappevurdering har blitt pekt på som et mulig nyttig element i prøving av

kompleks kompetanse fordi den kunne kompensere for at eksamen har preg av å være et øyeblikksbilde eller en stikkprøve, og ville styrket mangfoldigheten i prøveformer og gitt mulighet til å inkludere elevperspektivet ved å tilby valgmuligheter. Men med denne vurderingsformen følger det også noen utfordringer knyttet til vurderingsarbeidet. Et eksempel på det er manglende «kontroll av» om det er elevene som har utført arbeidet som inngår i mappa.

Å på den ene siden lage åpne og overordnede mål for at elevene skal kunne overføre det de har lært, til nye sammenhenger, og på den annen side gi tydelig uttrykk for hva elevene skal lære, og for hvilken kompetanse som skal være gjenstand for sluttvurdering, er en vanskelig balansegang. Det finnes svært lite forskning om kompetanseorienterte eksamener, i særdeleshet om hvilke fordeler og ulemper de har, og om psykometriske kvaliteter og styringsfunksjon, inkludert utilsiktede effekter. Fra forskning om eksamen i medisin framkommer det at et større antall eksamensoppgaver av ulike typer, kontekst og sensorer kan sikre god reliabilitet.

Om betydningen av teknologiske muligheter for eksamen

Digital teknologi kan påvirke eller endre forskjellige områder ved eksamen: eksamens administrasjon, bruk av hjelpemidler, innholdet til eksamen og sensurering. Tilgangen på ulike former for digitalt utstyr i norske skoler er generelt høy, men det å utnytte mulighetene for å få bedre kvalitet og effektivitet ved eksamen krever visse forutsetninger, og det oppstår et endringsbehov, spesielt knyttet til elevers og læreres digitale kompetanse. Den faglig-pedagogiske kompetansen blant lærerne til å ta i bruk digitale hjelpemidler ser ut til å variere mye. Elevgrupper som har bedre forutsetninger knyttet til tekniske forhold (tilgang til utstyr og infrastruktur) eller undervisningsforhold (omfang av og kvalitet på opplæring i digitale ferdigheter), har større sjanse å lykkes på eksamen enn andre elever.

Dagens eksamenssystem er digitalt i den forstand at elevene kan laste ned eksamensoppgavene og levere digitalt. Formålet er i hovedsak økt effektivitet, informasjonssikkerhet og personvern, men det åpner også for nye formater på produktet som skal vurderes, for eksempel lydfiler, video eller multimodale tekster. Den nye løsningen for å utvikle og gjennomføre eksamen som skal anskaffes innen 2021, skal kunne tilby slike nye oppgaveformater og i tillegg gi støtte til sensur. Den teknologiske utviklingen kan dermed gi muligheter for at eksamensoppgavene gjenspeiler bredden i kompetansebegrepet og følgelig blir mer valide. Gjennom å gi tilgang til automatisk skåring av oppgaver innebærer teknologi en betydelig tidsbesparelse ved sensuren.

Teknologiutvikling er også en driver for endring av skolens innhold og derigjennom hvilke kompetanser det er relevant å måle. Eksempler på dette er innføring av programmering i matematikkfaget, digitale tekstformer og tekstlige uttrykk i norskfaget eller kildekritisk kompetanse og digitale ferdigheter som en grunnleggende ferdighet. Teknologistøtte handler i tillegg om å bruke ulike digitale hjelpemidler i eksamenssituasjonen, for eksempel tilgang til åpent internett, lese-/skrivestøtte eller fagspesifikk programvare.

Samtidig kan den teknologiske utviklingen gi nye utfordringer: De nye mulighetene må balanseres med de faglige tradisjonene og kravene til at elevene skal kunne vise hvordan de mestrer grunnleggende kunnskaper og ferdigheter i det enkelte faget. I tillegg tar det tid for elever å lære å bruke hjelpemidler (både på papir og

digitale) på en hensiktsmessig måte, og sensorer har behov for et tolkningsfellesskap for å sikre en felles forståelse av hva som kjennetegner god bruk av kilder. En spørreundersøkelse i etterkant av forsøket med åpent internett på eksamen i videregående skole tyder på at de fleste elevene opplever det som nyttig å ha tilgang til internett på eksamen, mens både sensorer og lærere er mer usikre på utbyttet av tilgangen. Samtidig opplever særlig jenter eksamen med åpent internett som mer stressende enn eksamen uten internett.

Om lærerutdanningen og vurderingskompetanse

Rammeplanene for lærerutdanning er ulike for trinn 1–7 og 5–10 på den ene siden og PPU og lektorprogrammet på den annen side. I rammeplanene for trinn 1–7 og 5–10 som ble vedtatt i 2013, vektlegges at lærerstudenter etter endt utdanning skal ha inngående kunnskap om blant annet vurderings- og kartleggingsverktøy samt om vurdering av elevenes læring. Lærerstudentene skal også kunne vurdere elevers læring og gi læringsfremmende tilbakemeldinger. PPUs og lektorutdanningens rammeplan refererer til det nye kompetansebegrepet samt kjennetegn på måloppnåelse under «Ferdigheter», men nevner ingen kunnskapsområder som lærerstudentene skal undervises i. Retningslinjene for PPU og lektorutdanningen nevner derimot eksplisitt sluttvurderingen og vurdering av læring. Karaktersetting eller sensurering er ikke uttrykkelig nevnt i noen av disse rammeplanene eller retningslinjene.

Det er svært begrenset med forskning og oppdatert systematisk informasjon eller kunnskap om hvordan disse kvalifikasjonskravene knyttet til vurdering ivaretas i lærerutdanningene. Vi vet heller ikke mye om læringsutbytte eller effekten av lærerutdanningen når det gjelder vurderingskompetanse. Det finnes noe regionalt arbeid med kompetanseutvikling for lærere i skole, men det er i iøynefallende at det ikke finnes tilsvarende store kompetanseutviklingsprogrammer når det gjelder summativ vurdering og karaktersetting, som det gjør for formativ vurdering og vurdering for læring, noe som muligens kan tolkes som en gjennomgående rød tråd fra lærerutdanningen til etter- og videreutdanning.

13.3 Problemstillinger og spørsmål i det videre arbeidet

Kunnskapsgrunnlaget som vi har sammenstilt her, selv om det er til dels begrenset, leder til noen grunnleggende spørsmål som eksamensgruppa skal utrede i det videre arbeidet. Disse områdene framgår som særdeles viktige:

- å tydelig definere eksamens formål
- å diskutere muligheter til å prøve det utvidede kompetansebegrepet i fag på eksamen
- å se på forholdet mellom standpunkt og eksamen
- å vurdere om trekkordningen er hensiktsmessig

- å videreutvikle kvalitetssikringen av eksamen ut fra validitet, reliabilitet og rettferdighet
- å vurdere betydningen av ny teknologi for eksamen

Noen drøftinger som mulig utgangspunkt for det videre arbeidet

Gjennomgangen av kunnskapsgrunnlaget viser at eksamenssystemet har flere roller utover formelt definerte formål, som heller ikke alltid er like synlige. Eksamen skal samtidig imøtekomme validitetskrav, dette gjør det viktig å tydeliggjøre formålet med eksamen og følge med på implisitte roller den kan ha, slik at disse kan ligge til grunn for valideringsprosesser knyttet til eksamens utforming og gjennomføring.

Hvis eksamenskarakteren skal ha som hovedformål å være en ekstern vurdering i tillegg til standpunktkarakteren, kan vi spørre oss hvorfor de fleste standpunktkarakterene ikke følges opp med en eksamensvurdering. Samtidig er det ikke opplagt hvorvidt det er formålstjenlig å kombinere eksamenens kvalitetssikrende funksjon med standpunktvurderingen.

I dag utgjør standpunktkarakterer omtrent 80 prosent og eksamenskarakterer 20 prosent av vitnemålet. Derfor er følgende spørsmål aktuelle: Er det rimelig at eksamen og standpunkt teller like mye på vitnemålet, og at skriftlig og muntlig eksamen teller likt gitt at skriftlige eksamener kan kvalitetssikres på en annen måte enn muntlig? Er det rimelig at noen fag kan vektes mer på vitnemålet enn antall timer gjennom videregående skole tilsier, mens andre fag kan vektes mindre?

Trekkordningen må drøftes i et større perspektiv som inkluderer systematisk tenkning om eksamens rettferdighet, forutsigbarhet og hvordan man organiserer eksamen. Det kan være utfordrende å se for seg hvordan trekkordningen kan inngå i et helhetlig system for eksamen og standpunkt som ivaretar disse nevnte perspektivene på en god måte.

Det er en relevant problemstilling om eksamen kan eller bør prøve bredden i elevenes kompetanse, eller om eksamen bare skal prøve visse deler av kompetansen. Hvis eksamen og standpunkt utfyller hverandre som deler av et helhetlig system for vurdering og blir planlagt deretter, kan problemene unngås. Samtidig skaper en slik helhetlig tilnærming en mulighet til å inkludere et bredt spekter av prøveformer som samlet sett prøver kompleks kompetanse.

Et spørsmål som kan stilles i denne konteksten, er om alle eksamensformer er like formålstjenlige. Noen kompetansemål kan egne seg i mindre grad til å prøves i skriftlig eller muntlig eksamen. Dersom en større andel kompetanser ikke er egnet for eksamensformen som tradisjonelt brukes, bør det tenkes nytt. En mulig innfallsvinkel er å i større grad velge eksamensformene ut fra kompetansen/ arbeidsformen elevene vil trenge i hverdags- eller yrkeslivet samt i videre og høyere utdanning.

Å prøve elevenes samarbeidsevne og/eller løsninger og produkter de har kommet fram til i fellesskap, kan være utfordrende ved hjelp av en individuell eksamen. Dagens regelverk åpner for ulike løsninger så lenge vurderingen fortsatt er individuell. Samtidig er det mulig å se på eksamen og standpunkt som et helhetlig system der standpunkt bedre ivaretar noen av dimensjonene enn eksamen.

Eksamensformer og vurderingsprosesser er forankret i fagenes innhold og struktur, og dette forholdet bør ikke tas for gitt. Med fagfornyelsens tverrgående temaer og arbeid med dybdelæring i flere parallelle fag blir dette en aktuell problemstilling som bør tas inn i den videre diskusjonen av eksamen og sluttvurdering. Forskning om validitet vil kunne belyse hvordan en eksamen utformes og anvendes i ulike kontekster og til ulike formål.

Kunnskapsgrunnlaget viser forskjeller i karaktersetting mellom skoler, fag, kjønn, over år osv. Forskning bør se på hvordan forskjellene kan forstås, og i hvilken grad de kan forsvares eller endres. Hvordan skal for eksempel nivået på ulike fag avstemmes mot hverandre? I hvilken grad måler eksamen i et fag det samme som eksamen i samme fag året etter? Rettferdigheten svekkes om kravene er systematisk høyere i enkelte fag eller år framfor andre.

En grundig kvalitetssikring trenger systematisk planlegging av tiltak basert på et rammeverk (se vedlegg for et eksempel), en ny tilnærming til datalagring og dokumentasjon av resultater som gjøres tilgjengelig. I noen hensyn er det relativt enkelt å ta grep som forbedrer kvaliteten til eksamen. Dette gjelder særlig reliabilitet ved å utvikle kjennetegn på måloppnåelse på alle karakterer, ved å bruke et større antall oppgaver som innen en elevs eksamen vurderes av forskjellige sensorer, og ved å utvikle på forhånd eksplisitte vurderingskriterier som tydeliggjør forventninger og detaljerte krav til innhold og struktur og vekting av kriterier.

Fra elevenes perspektiv er det naturlig å reise spørsmål om de får tilstrekkelig anledning til å påvirke eksamensinnholdet og hvordan de prøves. En annen del av drøftingen i denne konteksten kunne dreie seg om eksamen kan følges opp med en mer utfyllende tilbakemelding enn karakteren. Slike tilbakemeldinger vil være ressurskrevende hvis de ikke gis automatisk, så nytteverdien må utredes først.

Det finnes i dag ingen samlet oversikt over de totale kostnadene for både lokalt gitt og sentralt gitt eksamen. Selv om Udir har oversikt over sine kostnader ved sentralt gitt skriftlig eksamen, herunder oppgaveutvikling og produksjon (ca. 34 mill.), systemstøtte og IT-forvaltning (ca. 27 mill.), sensur og klagebehandling (128 mill.), er det ingen samlet framstilling som viser kostnadene lokalt. Her vil Udir gjennomføre en utredning i 2019 for å gi et bedre grunnlag for å vurdere føringen om å overholde gjeldende kostnadsramme i eksamensgruppas mandat.

Det bør åpnes for dialog med lærerutdanningen for å styrke vurderingskompetansen hos nyutdannede lærere. Lærere har et stort ansvar ved å vurdere og eksaminere elever. Lærerutdanningen har den beste muligheten for å bygge opp et godt grunnlag slik at alle lærere er kompetente til å utføre vurderingsoppgaven på en valid, reliabel og rettferdig måte. Økt publisering av norskspråklig faglitteratur/lærebøker på summativ vurdering, spesielt lærebøker tilpasset de ulike lærerutdanningene vil kunne være ønskelig.

Vedlegg

Rammeverk for kvalitetssikring av eksamen (basert på Stobart, 2009; oversatt til norsk)

Kvalitetskriterium	Spørsmål	Mulige trusler
Formål	• Hva er formålet med denne vurderingen? • Er det flere formål?	• manglende klargjøring • konkurrerende formål • uoppnåelige formål
Gyldighet	• Hva blir vurdert?	• uklart definert (elevene forstår det ikke) • omstridt (vi er ikke enige)
Gyldighet	• Gjør vurderingen det den hevder å gjøre / det vi tror den gjør?	• utilstrekkelig grunnlag • uforutsett prøving av andre (og irrelevante) kompetanser
Pålitelighet	• Hvor pålitelig er vurderingssystemet?	• svikt i sikkerhetsrutiner • inkonsistent administrering av prøver • upassende endringer/tidsbegrensninger • upålitelige elever i testsituasjonen
Pålitelighet	• I hvilken grad kan vi forsvare resultatene?	• inkonsistent karaktersetting • upålitelig karaktersetting • upålitelig sammenfatning og aggregering av karakterer • utilstrekkelig informasjon for beslutninger • upassende vektinger • inkonsistente grenser mellom karakterer • liten bruk av prosedyrer for vurdering
Tolkning av	• Hvordan står resultatene i forhold til	• svak pålitelighet i enkeltelevers

resultater	formålet som de anvendes til?	resultater • feilklassifisering • forenklet eller upresis tolkning av aggregerte resultater
Konsekvenser og effekter	• Hvor effektiv er denne vurderingen i lys av formålet?	• begrenset tillit til resultatene
Konsekvenser og effekter	• Hadde vurderingen uforutsette konsekvenser?	• konflikt om tolkning av resultatene • upassende beslutninger på grunnlag av resultatene • negative konsekvenser for undervisning og læring

14. Litteraturliste

Andresen, S., Fossum, A., Rogstad, J., Smestad, B. (2017). *På prøve. Evaluering av matematikkeksamen på 10. trinn våren 2017*. Fafo.

Arbeidsgruppe nedsatt av nasjonalt råd for lærerutdanning: Christiansen, Enge, Lode (2015). Rapport frå karakterundersøkelsen i matematikk i GLU-utdanningene i 2014. Hentet 11.02.2019 fra: https://docplayer.me/6669704-Rapport-fra-karakterundersokelsen-i-matematikk-i-glu-utdanningene-i-2014.html

Association of Educational Assessment – Europe (AEA Europe) (2017). European Framework of Standards for Educational Assessment 1.0. Hentet 06.02.2019 fra: https://www.aea-europe.net/wp-content/uploads/2017/07/SW_Framework_of_European_Standards.pdf

Backmann, K., og Sivesind, K. (2012). Kunnskapsløftet som reformprogram: fra betingelser til forventninger. I T. Englund, E. Forsberg, og D. Sundberg (red.), *Vad räknas som kunskap? Läroplansteoretiska utsikter ock inblickar i lärarutbildningen ock skola* (s. 240–260). Stockholm: Liber.

Baeten, M., Struyven, K., og Dochy, F. (2008). *Students assessment preferences and approaches to learning in new learning environments: A replica study*. New York: AERA (Paper presented at AERA March 2008).

Baird, J.-A., og Hopfenbeck, T.N. (2016). Curriculum in the Twenty-First Century and the Future of Examinations. I D. Wyse, L. Hayward, og J. Pandya (red.), *The SAGE handbook of curriculum, pedagogy and assessment* (s. 821–837). Los Angeles; London; New Delhi; Singapore; Washington, DC: SAGE.

Beller, M., og Gafni, N. (2000). Can item format (multiple choice vs. open-ended) account for gender differences in mathematics achievement? *Sex Roles*: A Journal of Research, 42, 1–21.

Ben-Chaim, D., og Zoller, U. (1997). Examination-type preferences of secondary school students and their teachers in the science disciplines. *Instructional Science*, 25(5), 347–367.

Ben-Shakar, G., og Sinai, Y. (1991). Gender differences in multiple choice tests: The role of differential guessing. *Journal of Educational Measurement*, 28, 23–35.

Biggs, J.B., (2003). *Teaching for quality learning at university*. Buckingham: Open University Press/Society for Research into Higher Education. (Second edition).

Birenbaum, M., og Dochy, F. (1996). Introduction. I: M. Birenbaum og F. Dochy (red.). *Alternatives in assessment of achievements, learning processes and prior knowledge* (s. xiii–xv). Boston: Kluwer.

Birenbaum, M., og Feldman, R.A. (1998). Relationships between learning patterns and attitudes towards two assessment formats. *Educational Research*, 40(1), 90–97.

Bjørnset, M., Fossum, A., Rogstad, J., Smestad, B., og Talberg, N. (2018). *Digitale skillelinjer: Evaluering av matematikkeksamen på 10. trinn våren 2018*. Fafo-rapport 2018:36.

Black, P., Harrison, C., Lee, C.S., Marshall, B., og Wiliam, D. (2003). *Assessment for learning, putting it into practice*. Open University Press.

Black, P., Harrison, C., Hodgen, J., Marshall, B., og Serret, N. (2011). Can teachers summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy og Practice*, 18(4), 451–469.

Borgonovi, F., Ferraram, A., og S. Maghnouj (2018): The gender gap in educational outcomes in Norway, *OECD Education Working Papers, No. 183*, OECD Publishing, Paris. http://dx.doi.org/10.1787/f8ef1489-en

Broadfoot, P. (2007). *An introduction to assessment*. [London]; New York: Continuum.

Brookhart, S.M. (2013). Grading. I J.H. McMillan (red.), *SAGE Handbook of Research on Classroom Assessment* (s. 257–272). USA: Sage.

Buland, T., Engvik, G., Fjørtoft, H., Langseth, I., Sandvik, L.V., og Mordal, S. (2014): *Vurdering i skolen. Utvikling av kompetanse og fellesskap. Sluttrapport fra prosjektet Forskning på individuell vurdering i skolen* (FIVIS). NTNU.

Bøhn, H. (2017): *What is to be assessed? Teachers' understanding of constructs in an oral English examination in Norway* (doktoravhandling). Universitetet i Oslo.

Carlsen, C. (2013): *Guarding the Guardians. Rating scale and rater training effects on reliability and validity of scores of an oral test of Norwegian as a second language* (doktoravhandling). Universitetet i Bergen.

Crooks, T.J., Kane, M.T., og Cohen, A.S. (1996). Threats to the Valid Use of Assessments. *Assessment in Education: Principles, Policy og Practice, 3(3),* 265–286.

Dale, E.L. (2008). *Fellesskolen – reproduksjon av sosial ulikhet*. Oslo: Cappelen akademisk forlag

Dale, E.L., og Wærness, J.I. (2006). *Vurdering og læring i en elevaktiv skole*. Oslo: Universitetsforlaget.

Dochy, F., og McDowell, L. (1997). Introduction assessment as a tool for learning. *Studies in Educational Evaluation, 23(4),* 279–298.

Duncan, C.R., og Noonan, B. (2005). Factors Affecting Teachers' Grading and Assessment Practices. *The Alberta Journal of Educational Research, 53(1),* 1–21.

Eccles, J. (1983). Expectancies, values and academic behavior. I: J.T. Spence (red). *Academic and achievement motives*. San Francisco: Freeman.

Eckstein, M.A., og Noah, H.J. (1993). *Secondary School Examinations. International Perspectives on Policies and Practice.* New Haven: Yale University Press.

Egeberg, G., Hultin, H., og Berge, O. (2016). *Monitor skole 2016: Skolens digitale tilstand.* Oslo: Senter for IKT i utdanningen.

Eggen, A.E. (2004). *Alfa and Omega in Student Assessment; Exploring Identities of Secondary School Science Teachers* (ph.d.-avhandling). Department of Teacher Education and School Research, University of Oslo.

Erstad, O. (2008). Changing Assessment Practices and the Role of IT. I J. Voogt og G. Knezek (red.), *International Handbook of Information Technology in Primary and Secondary Education* (Bind 20, s. 181–194): Springer US.

Evensen, L.S., Berge, K.L., Thygesen, R., Matre, S., og Solheim, R. (2016). Standards as a tool for teaching and assessing cross-curricular writing. *The Curriculum Journal*, 27, 229–245.

Federici, R.A., Gjerustad, C., Vaagland, K., Larsen, E.H., Rønsen, E., og Hovdhaugen, E. (2017). *Spørsmål til Skole-Norge våren 2017*. NIFU-rapport 2017:12. Oslo.

Finne, H., Jensberg, H., Aaslid, B.E., Haugsbakken, H., Holth Mathiesen, I., og Mordal, S. (2011). *Oppfatninger av studiekvalitet i lærerutdanningen blant studenter, lærerutdannere, øvingslærere og rektorer* (=SINTEF rapport; A18011). Trondheim: SINTEF.

Forsøksrådet for skoleverket (1969). *Standardiserte prøver i skolen. Forsøk og reform i skolen – nr 16.* Oslo: Universitetsforlaget.

Galloway, T.A, Kirkebøen, L.J., og Rønning, M. (2011): *Karakterpraksis i grunnskoler: sammenheng mellom standpunkt- og eksamenskarakter.* SSB.

Gellman, E., og Berkowitz, M. (1993). Test-item type: What students prefer and why. *College Student Journal, 27(1),* 17–26.

Gjerustad, C., Waagene, E., og Salvanes, K.V. (2015). *Spørsmål til Skole-Norge høsten 2014.* NIFU.

Gjone, G. (1993). Types of problems and how students in Norway solve them. I N. Mogens (red.), *Cases of assessment in Mathematics Education: An ICMI Study* (s. 107–118). Amsterdam: Kluwer Academic Press.

Gustafsson, J.-E., og Erickson, G. (2018). Nationella prov i Sverige – tradition, utmaning och förändring. *Acta Didactica Norge, 12(4).* DOI: http://dx.doi.org/10.5617/adno.6434

Haladyna, T.M., og Downing, S.M. (2005). Construct-Irrelevant Variance in High-Stakes Testing. *Educational Measurement: Issues and Practice, 23(1),* 17–27.

Harlen, W. (2005). Teachers' Summative Practices and Assessment for Learning – Tensions and Synergies. *The Curriculum Journal, 16(2),* 207–223.

Harris, L.R., og Brown, G.T.L. (2016). *The Human and Social Experience of Assessment: Valuing the Person and the Context.* I G.T.L. Brown og L.R. Harris (red.), Handbook of human and social conditions in assessment (s. 1–17). New York: Routledge, Taylor og Francis Group.

Hatlevik, O.E., Tømte, K., Skaug, J.H., og Ottestad, G. (2010). *Monitor 2010: Samtaler om IKT i skolen.* Oslo: Senter for IKT i utdanningen.

Hatlevik, O.E., Egeberg, G., Gudmundsdottir, G.B., Loftsgarden, M., og Loi, M. (2013). *Monitor skole.* Oslo: Senter for IKT i utdanningen.

Hatlevik, O.E., og Throndsen, I. (red.) (2015). *Læring av IKT : Elevenes digitale ferdigheter og bruk av IKT i ICILS 2013.* Oslo: Universitetsforlaget.

Herman, J.L., og Baker, E.L. (2009). Assessment policy: Making sense of the babel. I G. Sykes, B. Schneider og D. Plank (red.). *Handbook of Education Policy Research,* Newbury Park, London: Sage.

Hill, K.T. (1984). *Debiliating motivation and testing: A major educational problem – Possible solutions and policy applications.* I: R.E. Ames og C. Ames (red.), Research on motivation in education: Vol. 1. Student motivation. New York: Academic Press.

Hill, K.T., og Wigfield, A. (1984). Test anxiety: A major eduactional problem and what can be done about it. *Elementary School Journal, 85,* 105–126.

Hovde, P., og Olsen, S.O. (2015). *Utredning – Digital eksamen NTNU 2015–2019.* NTNU.

Hovdhaugen, E., Seland, I., Lødding, B., Prøitz, T., og Vibe, N. (2014). *Karakter i offentlige og private videregående skoler. En analyse av eksamens- og standpunktkarakter i norsk og matematikk og rutiner for standpunktvurdering i offentlig og private videregående skoler.* NIFU. Rapport 24/2014

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

Hultin, H., og Berge, O. (2014). *Notat til utvalgsarbeid om digital kompetanse.* Oslo: Senter for IKT i utdanningen.

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

Jarning, H., og Aas, G.H. (2008). Between Common Schooling and the Academe: The International Examinations Inquiry in Norway, 1935–1961. I: *An Atlantic Crossing? The Work of the International Examination Inquiry, its Researchers, Methods and Influence,* 181–204, redigert av M. Lawn. Oxford, UK: Symposium Books.

Kane, M.T. (2015). Explicating validity. Assessment in Education: *Principles, Policy og Practice, 23(2),* 1–14.

Kirke-, utdannings- og forskningsdepartementet (1996). *Om elevvurdering, skolebasert vurdering og nasjonalt vurderingssystem* (St.meld. nr. 47 (1995–1996)). Oslo: Departementet.

Kommunerevisjonen (2013). *Standpunktkarakterer i videregående skole – likebehandles elevene?* Oslo: Oslo kommune kommunerevisjonen.

Koretz, D. (1998). Large-scale portfolio assessments in the US: Evidence pertaining to the quality of measurement. *Assessment in Education: Principles, Policy og Practice, 5(3),* 309–334.

Krogh,L.C.(2016). *Kreativitet og ambivalens: En undersøkelse av variasjoner i vurdering og kjennetegn ved sprikvurderte tekster fra eksamen i norsk hovedmål 2015.* MasteroppgaveHøgskoleniSørøst-Norge.

Krumsvik, R.J., Egelandsdal, K., Sarastuen, N.K., Jones, L.Ø., og Eikeland, O.J. (2013). *Sammenhengen mellom IKT-bruk og læringsutbytte (SMIL) i videregående opplæring.* Bergen.

Kunnskapsdepartementet. (2013). *På rett vei.* (Meld. St. 20 (2012–2013)).

Kunnskapsdepartementet. (2016). *Fag – Fordypning – Forståelse — En fornyelse av Kunnskapsløftet.* (Meld. St. 28 (2015–2016)).

Kunnskapsdepartementet (2017). *Organisering av skoleåret i videregående opplæring,* rapport fra

arbeidsgruppa oppnevnt av KD. Hentet 01.11.2018 fra: https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/2017/rapport---organisering-skolearet.pdf

Lawn, M. (2008). Red. *An Atlantic Crossing? The work of the International Examination Inquiry, its researchers, methods and influence.* Oxford: Symposium Books.

Lekholm, A.K., og Cliffordsson, C. (2008). Discrepancies between School Grades and Test Scores at Individual and School Level: Effects of Gender and Family Background. *Educational Research and Evaluation, 14(2)*, 181–199.

Lekholm, A.K., og Cliffordsson, C. (2009). Effects of Student Characteristics on Grades in Compulsory School. *Educational Research and Evaluation, 15(1),* 1–23.

Lysne, A. (1999). *Karakterer og kompetanse. Stridstema i norsk skolehistorie.* AVA forlag.

Lysne, A. (2004). *Karakterer og kompetanse. Kampen om skolen.* AVA forlag.

Lysne, A. (2006). Assessment Theory and Practice of Students' Outcomes in the Nordic Countries. *Scandinavian Journal of Educational Research, 50(3),* 327–359.

Lundahl, Ch., og Tveit, S. (2014): Att legitimera nationella prov i Sverige och i Norge – en fråga om profession och tradition. *Pedagogisk Forskning i Sverige, 19(4–5),* 297–323.

Markus, K.A., og Borsboom, D. (2013). *Frontiers of test validity theory: measurement, causation and meaning.* New York, N.Y: Routledge / Taylor og Francis Group.

McDowell, L. (1995). The impact of innovative assessment on student learning. *Innovations in Education and Training International, 32(4),* 302–313.

McMillan, J.H. (2003). Understanding and improving teachers' classroom assessment decision-making: Implications for theory and practice. *Educational Measurement: Issues and Practice, 2(4),* 34–43.

Moss, P.A. (2007). Reconstructing Validity. *Educational Researcher, 36(8),* 470–476.

Moss, Pamela A., Girard, B.J., og Haniford, L.C. (2006). Validity in Educational Assessment. *Review of Research in Education, 30* (Special Issue on Rethinking Learning: What Counts as Learning and What Learning Counts), 109–162.

Muller, J. (2009). Forms of Knowledge and Curriculum Coherence. *Journal of Education and Work, 22(4),* 205–226.

Munthe, E., Solbakken, J.I., Hjetland, H., og Hustad, B.C. (2014). Lærerutdanninger i endring: Indre utvikling – ytre kontekstuelle og strukturelle hinder (Følgegruppen for lærerutdanningsreformen; Rapport Nr. 4).

Nassar, Y.H.B., Qaraeen, K., og Naba'h, A.A. (2011). Secondary School Students' Perceptions of Essay and

Multiple-Choice Type Exams. Dirasat, *Educational Sciences, 38(1),* 345–358.

Natriello, G., og Dornbusch, S.M. (1984). *Teacher evaluative standards and student effort.* New York: Longman.

Nesman, M., og Kovač, V.B. (2016): Privatister – hvem er de, og hva motiverer dem til å lykke på eksamen? Kartlegging av bakgrunnsvariabler og deres intensjon i lys av en utvidet versjon av teorien om planlagt adferd. *Nordisk tidsskrift for pedagogikk og kritikk.*

Newton, P.E. (2007). Clarifying the purposes of educational assessment. *Assessment in Education, 14(2),* 149–170.

NOKUT (2015). Centre for Professional Learning in Teacher Education (ProTed): Mid-term evaluation – Centre of Excellence in Higher Education. Hentet 12.02.2019 fra: https://www.uv.uio.no/proted/om/arsrapporter/proted_mid-term_evaluation_report_2015.pdf

Nordenbo, S.E., Allerup, P., Andersen, H.L., Dolin, J., Korp, H., Larsen, M.S., . . . Østergaard, S. (2009). *Pædagogisk brug af test – Et systematisk review.* København: Danmarks Pædagogiske Universitets Forlag

Norcini, J., Anderson, M., Brownell, Bollela, V., Burch, V., Costa, M.J., Duvivier, R., Hays, R., Palacios Mackay, M.F., Roberts, T., og Swanson, D. (2018). 2018 Consensus framework for good assessment, *Medical Teacher, 9,* 1–8. https://doi.org/10.1080/0142159X.2018.1500016

NOU (Norges offentlige utredninger) 2015: 8. (2015). *Fremtidens skole: Fornyelse av fag og kompetanser.* Oslo: Kunnskapsdepartementet.

NOU (Norges offentlige utredninger) 2018: 15. (2018). *Kvalifisert, forberedt og motivert — Et kunnskapsgrunnlag om struktur og innhold i videregående opplæring.* Oslo: Kunnskapsdepartementet.

NOU (Norges offentlige utredninger) 2019: 3. (2019). *Nye sjanser – bedre læring: Kjønnsforskjeller i skoleprestasjoner og utdanningsløp.* Oslo: Kunnskapsdepartementet.

Norgesuniversitetet (2015). *Sluttrapport fra Ekspertgruppa for digital vurdering og eksamen per februar 2015:* https://diku.no/rapporter/digital-tilstand-2014

Nasjonalt råd for lærerutdanning (NRLU) (2017). Nasjonale retningslinjer for lektorutdanning for trinn 8–13. Hentet 12.02.2019 fra: https://www.uhr.no/_f/p1/i4d4335f1-1715-4f6e-ab44-0dca372d7488/lektorutdanning_8_13_vedtatt_13_11_2017.pdf

Nasjonalt råd for lærerutdanning (NRLU) (2017b). Nasjonale retningslinjer for praktisk pedagogisk utdanning – allmennfag. Hentet 18.02.2019 fra: https://www.uhr.no/_f/p1/i13d351d8-d4a8-4c93-ac64-f0d2fbbdc6c6/godkjente-retningslinjer-ppu.pdf

Nasjonalt råd for lærerutdanning (NRLU) (2018). Nasjonale retningslinjer for praktisk-pedagogisk utdanning

for yrkesfag trinn 8-13. Hentet 19.02.2019 fra: https://www.uhr.no/_f/p1/i6c34f03d-e46c-4ce8-8c90-9c47b488bbc1/nasjonale-retningslinjer-for-praktisk-pedagogisk-utdanning-for-yrkesfag-trinn-8-13_ferdig.pdf

Nygård Arntzen, H. (2015). *Matematikkeksamen gjennom tre reformer: En analyse av avgangseksamen på høyeste nivå i den videregående skolen.* Masteroppgave, UiT.

Pellegrino, J.W., Chudowsky, N., Glaser, R., og National Research Council (U.S.) (Red.). (2001). *Knowing what students know: The science and design of educational assessment.* Washington, DC: National Academy Press.

Popham, W.J., og Husek, T.R. (1969). Implications of criteriion-referenced measurement. *Journal of Educational Measurement, 6(1),* 1–9.

ProTed (2016). Centre for Professional Learning in Teacher Education: Annual report for 2016. Hentet 12.02.2019 fra: https://www.uv.uio.no/proted/om/arsrapporter/annual_report_2016_proted.pdf

ProTed (2017). Centre for Professional Learning in Teacher Education: Annual report for 2017. Hentet 12.02.2019 fra: https://www.uv.uio.no/proted/om/arsrapporter/annual-report-for-proted_2017.pdf

Prøitz, T.S., og J.S. Borgen (2010). *Rettferdig standpunktvurdering – det (u)muliges kunst?* NIFU STEP report 16/2010.

Prøitz, T.S., (2013a). Variations in grading practice—subjects matter. *Education Inquiry, 4(3),* 1–22. (in press, forthcoming September 2013).

Prøitz, T.S., (2018). *Ten years later – variations in grading practices revisited*, paper presented at LOaPP-project meeting 30.11.18 USN.

Rambøll (2012). *Forsøk med internett til eksamen: Sluttrapport.* Hentet 27.11.2108 fra: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Eksamen-med-tilgang-til-Internett/

Rambøll (2013). *Sluttrapport: Evaluering av eksamen med tilgang til internett.* Hentet 27.11.2018 fra: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Eksamen-med-tilgan-til- internett/

Rambøll (2014). *Forsøk med tilgang til internett på eksamen.* Hentet 27.11.2018 fra: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Forsokt-med-Internett-pa-eksamen/

Rambøll (2015). *Evaluering av forsøk med tilgang til internett på eksamen 2014–2015.* Hentet 27.11.2018 fra: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/forsok-med-tilgang-til-internett-pa-eksamen/

Rambøll (2019). *Evaluering av åpent internett til eksamen: Sluttrapport.*

Redecker, C., og Johannessen, Ø. (2013). Changing Assessment: Towards a New Assessment Paradigm Using ICT. *European Journal of Education, 48(1),* s. 79–96. Blackwell.

Resh, N. (2009). Justice in Grades Allocation: Teachers' Perspective. *Social Psychology of Education, 12(3),* 315–325.

Sambell, K., McDowell, L., og Brown, S. (1997). «But is it fair?»: An exploratory study of student perceptions of the consequantial validity of assessment. *Studies in Educational Evaluation, 23(4),* 349–371.

Sandvik, L.V., Engvik G., Fjørtoft, H., Langseth, I.D., Aaslid, B.E., og Buland, T. (2012): *Vurdering i skolen. Intensjoner og forståelse. Delrapport 1 fra prosjektet Forskning på individuell vurdering i skolen (FIVIS).* Trondheim: NTNU.

Sandvik, L.V., og Buland, T. (2013). *Vurdering i skolen. Operasjonaliseringer og praksiser Delrapport 2 fra prosjektet Forskning på individuell vurdering i skolen (FIVIS).* Trondheim: NTNU/SINTEF.

Sanne, A., Berge, O., Bungum, B., Jørgensen, E.C., Kluge, A., Kristensen, T.E., Mørken, K.M., Svorkmo, A., og Voll, L.O. (2016). Teknologi og programmering for alle: *En faggjennomgang med forslag til endringer i grunnopplæringen.* Oslo. Hentet 19.11.2018 fra: https://www.udir.no/globalassets/filer/tall-og-forskning/forskningsrapporter/teknologi-og-programmering-for-alle.pdf

Schaper, N., Hilkenmeier, F., og Bender, E. (2013). *Umsetzungshilfen für kompetenzorientiertes Prüfen: HRK-Zusatzgutachten.* Bonn, Germany. Hentet 06.02.2019 her: https://www.hrk-nexus.de/fileadmin/redaktion/hrk-nexus/07-Downloads/07-03-Material/zusatzgutachten.pdf

Schunk, D. (1984). Self-efficacy perspective on achievement behavior. *Educational Psychologist, 19,* 48–58.

Sejersted, F. (2005). Sosialdemokratiets tidsalder. *Norge og Sverige i det 20. århundre. Andre del av historieverket Norge og Sverige gjennom 200 år.* Oslo: Pax forlag.

Seland, I., Lødding, B., og Prøitz, T.S. (2015). *Delrapport 1 fra evaluering av forsøk med halvårsvurdering med én eller to karakterer i norsk. Litteraturstudie.* NIFU-rapport 33/2015. Oslo: NIFU.

Seland, A., Hovdhaugen, E., Lødding, B., Prøitz, T., og Rønsen, E. (2018): *Sluttrapport fra evaluering av forsøk med halvårsvurdering med én eller to karakterer i norsk.* Oslo.

Sjaastad, J., Carlsten, T.C., og Wollscheid, S. (2016). *Får elevene den opplæringen de har krav på? Kartlegging av undervisningstimer med kvalifiserte lærere i videregående opplæring. Rapport 26/2016.* Oslo: NIFU.

Smestad, B., og Fossum, A. (2019). Primary school exams in calculations/mathematics in Norway 1946–2017: Content and form. CERME 2019.

Solheim, R. Og Matre, S. (2014). Lærarsamtalar om elevtekstar: Mot eit felles fagspråk om skriving og vurdering.

Steffensen, K., og Ziade, S.E. (2009): *Skoleresultater 2008. En kartlegging av karakterer fra grunnskoler og videregående skoler i Norge.* Rapporter 2009/23, Statistisk sentralbyrå.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51(2),* 161–179.

Tobias, S. (1985). Test anxiety: Interference, defective skills, and cognitive capacity. *Educational Psychologist, 20,* 135–142.

Tveit, S. (2007). Elevvurdering i Kunnskapsløftet. I: H. Hølleland (red.). *På vei mot Kunnskapsløftet.* Oslo: Cappelen akademiske forlag.

Tveit, S. (2018). *(Trans)national Trends and Cultures of Educational Assessment: Reception and Resistance of National Testing in Norway and Sweden during the Twentieth Century.* I "Assessment Cultures", edited by Cristina Alarcon and Martin Lawn. Book series Studia Educationis Historica. Berlin: Peter Lang.

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

Universitets- og Høgskolerådet (2011). *En helhetlig tilnærming til lærerutdanning:* Rapport fra en arbeidsgruppe nedsatt av Nasjonalt råd for lærerutdanning. Oslo: UHR.

Universitets- og Høgskolerådet – Lærerutdanning (UHL-LU) (2017). Felleskapittel – Nasjonale Retningslinjer for Lærarutdanningane. Hentet 12.02.2019 fra: https://www.uhr.no/_f/p1/i4fbd09e0-6a5f-4a13-9e89-3971c57cfa5d/fellestekst-for-retningslinjene-for-alle-typer-av-larerutdanning.pdf

Utdanningsdirektoratet (2009). *Utdanningsspeilet 2008.* Hentet 06.02.2019 fra: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Utdanningsspeilet-2008-ei-analyse-av-grunnopplaringa-2009/

Utdanningsdirektoratet (2013). *Utdanningsspeilet 2013.* Hentet 06.02.2019 fra: https://www.udir.no/globalassets/filer/tall-og-forskning/rapporter/utdanningsspeilet_2013/us2013.pdf

Utdanningsdirektoratet (2015). *Rapport om utviklingen i klager på standpunktkarakterer fra 2010 til 2015.* Oslo.

Utdanningsdirektoratet (2016). *Erfaringer og vurderinger av eksamen våren 2012 og 2013.* Oslo. Hentet 14.11.2018. Hentet 06.02.2019 fra: https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Erfaringer-og-vurderinger-av-eksamen-varen-2012-og-2013/

Utdanningsdirektoratet (2017). *Utdanningsspeilet (2017).* Hentet 06.02.2019 fra: http://utdanningsspeilet.udir.no/2017/

Utdanningsdirektoratet (2018). *Utdanningsspeilet 2018.* Hentet 16.12.2018. https://www.udir.no/tall-og-forskning/finn-forskning/tema/utdanningsspeilet/

Utdanningsdirektoratet (2018a). *Rammeverk for lærerens profesjonsfaglige digitale kompetanse (PfDK).* Oslo. Hentet 29.11.2018. Hentet 06.02.2019 fra: https://www.udir.no/kvalitet-og-kompetanse/profesjonsfaglig-

digital-kompetanse/rammeverk-larerens-profesjonsfaglige-digitale-komp/

Utdanningsdirektoratet (2018b). *Trekkordning ved eksamen for grunnskolen og videregående opplæring Udir-2-2018.* Hentet 17.12.2018 fra: https://www.udir.no/regelverkstolkninger/opplaring/eksamen/trekkordning-ved-eksamen-for-grunnskole-og-videregaende-opplaring-udir-2-2018/

Utdanningsdirektoratet (2018c). *Rammeverk for eksamen.* Hentet 16.12.2018 fra: https://www.udir.no/eksamen-og-prover/eksamen/rammeverk-eksamen/5.-analyse-av-eksamen-og-bruk-av-resultatene/

Utdanningsdirektoratet (2018d). *Eksamensundersøkelse engelsk 10. trinn.* Utdanningsdirektoratet, 31.5.2018.

Utdanningsdirektoratet (2019). *Sluttrapport vurdering for læring.* https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/erfaringer-fra-nasjonal-satsing-pa-vurdering-for-laring-2010-2018/

Utdanningsdirektoratet (2018f). *Retningslinjer for læreplanutvikling.* (Upublisert).

Utdannings- og forskningsdepartementet (2004). *Kultur for læring.* (St.meld. nr. 30 (2003–2004)).

van de Watering, G., Gijbels, D., Dochy, F., og van der Rijt, J. (2008). Students' assessment preferences, perceptions of assessment and their relationships to study results. *High Educ, 56,* 645–658.

van der Vleuten, C., og Schuwirth, L.W.T (2005). Assessing professional competence: from methods to programmes. *Medical Ecucation (39),* 309–317.

Waagene, E., Larsen, E., Vaagland, K., og Federici, R.A. (2018). *Spørsmål til Skole-Norge høsten 2017: Analyser og resultater fra Utdanningsdirektoratets spørreundersøkelse til skoler og skoleeiere.* NIFU.

Wass, V., Van der Vleuten, C., Shatzer, J., og Jones, R. (2001). Assessment of clinical competence. *The Lancet (357),* 945–949.

Wiliam, D. (1996). National Curriculum Assessments and Programmes of Study: validity and impact. *British Educational Research Journal, 22(1),* 129–141.

Wilson, M. (2005). *Constructing Measures:* An Item Response Modeling Approach, Volume 2. Lawrence Erlbaum Associates.

Wollscheid, S., Hjetland, H.N., Rogde, K., og Skjelbred, S.V. (2018): *Årsaker til og tiltak mot kjønnsforskjeller i skoleprestasjoner. En kunnskapsoversikt.* NIFU 2018:25.