Acta Didactica 1/2005

Svein Lie, Therese N. Hopfenbeck,
Elisabeth Ibsen og Are Turmo

NASJONALE PRØVER PÅ NY PRØVE

Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005

INSTITUTT FOR LÆRERUTDANNING OG SKOLEUTVIKLING
UNIVERSITETET I OSLO

ISSN: 1502-2013
ISBN: 82-90904-81-9

Utgiver og redaksjon for webpublikasjon: Institutt for lærerutdanning og skoleutvikling

Rapportserien distribueres av Unipub AS
Henvendelser om trykte publikasjoner i serien kan rettes til Unipub AS:
Telefon: 22 85 33 00
Telefaks: 22 85 30 39
E-post: post@unipub.no

Nasjonale prøver på ny prøve

Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005

Svein Lie, Therese N. Hopfenbeck, Elisabeth Ibsen og Are Turmo

Institutt for lærerutdanning og skoleutvikling
Universitetet i Oslo

Forord

Denne rapporten representerer en forskningsbasert evaluering av de nasjonale prøvene som ble gjennomført våren 2005. Rapporten er utarbeidet med utgangspunkt i et oppdrag gitt til Institutt for lærerutdanning og skoleutvikling (ILS), Universitetet i Oslo fra Utdanningsdirektoratet.

Forfatterne av denne rapporten har sin faglige bakgrunn i realfag og engelsk. I tillegg har vi trukket inn Lise Iversen Kulbrandstad, Høgskolen i Hedmark som ekstern ekspert i lesing. Hun har gitt verdifulle innspill til vurderingen av leseprøvene og har skrevet mesteparten av kapitlene om validitet av leseprøvene (kap. 5.1.2, 5.2.2 og 5.3.2).

Kapittel 1 i denne rapporten er et sammendrag av funn og anbefalinger. Våre vurderinger og anbefalinger er på to ulike nivåer. Flere steder har vi pekt på svakheter ved prøvene og hvordan slike forhold eventuelt kan forbedres. I tillegg har vi kommet med en mer overordnet vurdering av hovedtrekk og prinsipper for de nasjonale prøvene som system, særlig i kapittel 1.10. Vi vil understreke at det er viktig å se våre anbefalinger angående disse to nivåene i sammenheng.

Oslo, september 2005
Forfatterne

Innhold

1 Sammendrag, konklusjoner og anbefalinger

1.1 Oppdraget og metodisk tilnærming

1.1.1 Om oppdraget og denne rapporten

Denne rapporten er et svar på oppdraget fra Utdanningsdirektoratet om en forskningsmessig evaluering av de nasjonale prøvene i 2005. Oppdraget over tre år ble gitt til ILS med følgende beskrivelse:

> Evalueringen skal:
> 1. Belyse hvilke metoder som er brukt i utviklingen av prøvene
> 2. Foreta grunnleggende item-analyser
> 3. Vurdere prøvenes reliabilitet (indre konsistens)
> 4. Vurdere sensorreliabilitet for åpne oppgaver
> 5. Vurdere prøvenes validitet
> 6. Vurdere hvordan prøvene kan utvikles slik at de kan fungere som grunnlag for sammenlikninger med resultater fra år til år.
> 7. Foreta analyser av elevenes resultater i de ulike prøvene
> 8. Foreta vurderinger og analyser som kan gi grunnlag for blant annet utvikling av enkle kompetanseprofiler på sikt.

Dette første kapitlet i rapporten inneholder en innledning, samt sammendrag og overordnete konklusjoner. Kapittel 2 gir en nærmere beskrivelse av spesifikasjoner for vårt oppdrag. Kapittel 3 beskriver innhenting av data fra elevers besvarelser i form av skolenes interne læreres og de eksterne lærernes vurderinger av disse besvarelsene. Kapittel 4 består av en oversikt over hvilke metoder vi har brukt i våre analyser. Kapitlet inneholder også noen matematiske og tekniske forklaringer for de som eventuelt ønsker en en orientering om dette. Kapitlene 5-8 inneholder vurderinger og resultater av våre kvalitative og kvantitative analyser. Fagområdene lesing, skriving, engelsk og matematikk (i denne rekkefølgen) er omhandlet i hvert sitt kapittel.

Denne rapporten representerer våre svar på alle punktene ovenfor bortsett fra punkt 6. Når det gjelder å utvikle metoder for å sammenlikne resultater år for år, er det en meget omstendelig prosess. Vi har ikke funnet noe naturlig utgangspunkt for en slik diskusjon i forbindelse med årets prøver. Gitt kompleksiteten av en slik prosess mener vi at det bør grundig diskuteres om det virkelig skal være et prioritert formål for de nasjonale prøvene å gjøre dette. Vi anbefaler i den anledning å følge nøye de utredninger om dette som for tiden gjennomføres i Sverige.

1.1.2 Datainnhenting og metodisk tilnærming

Vår evaluering av prøvene har bestått av to komponenter. For det første har vi gjennomført en grundig analyse av data fra elevenes besvarelser. Dette har gitt oss informasjon om prøvenes psykometriske kvalitet når det gjelder forhold som reliabilitet og validitet. Vi har også gjennomført en kvalitativ analyse av oppgavene og vurderingskriteriene. Til sammen har dette tillatt oss å konkludere når det gjelder sterke og svake sider ved hver prøve.

MMI har gjennomført en parallell undersøkelse av lærernes, skoleledernes og de eksterne vurderernes syn på hvordan prøvene har fungert. Flere steder har vi trukket resultater fra den undersøkelsen inn i vår diskusjon. Heretter refererer vi til den som "MMI-rapporten" eller "MMI-undersøkelsen". I noen sammenhenger har vi også referert til rapporten om de nasjonale prøvene som TNS-Gallup utførte for Utdanningsforbundet (referert til som "Gallup-rapporten").

Data til våre analyser har vi skaffet oss via det systemet som var lagt opp fra Utdanningsdirektoratets side. Et tilfeldig utvalg av skoler ble bedt om å sende et tilfeldig utvalg av elevbesvarelser eller kopier av disse til ekstern vurdering. Både de eksterne og skolens egne vurderinger skulle deretter sendes til faggruppene for hvert fagområde. Fra faggruppene ble så dette sendt til oss ved ILS. Et slikt opplegg viste seg å fungere dårlig for vårt formål. Til tross for purring viste det seg vanskelig å få inn dataene innen rimelig tid for våre strenge tidsfrister. Vi ble derfor etter hvert tvunget til å renonsere på våre planlagte utvalg av elever og skoler og i stedet nøye oss med de data vi fikk inn i tide. Som vi også har beskrevet i kapittel 3 (se også tabell 3.1), er altså noen av utvalgene våre mindre enn ønskelig, i hovedsak på grunn av sen innsending fra skolene. Dette gjelder særlig for grunnkurs. Vi vil imidlertid understreke at det betyr at det *kan* ligge en skjevhet i våre utvalg, i og med at skoler og eksterne vurderere som *har* sendt inn data i tide, ikke nødvendigvis er helt representative for populasjonen. Det er imidlertid liten grunn til å tro at denne effekten er betydelig når det gjelder de fleste forhold som vi studerer i vår undersøkelse.

Også på andre måter var det problemer med å få lagt inn relevante data for våre analyser. Særlig viste det seg at identiteten til skoler og elever ofte var angitt på ulike måter, slik at det ikke var mulig å "matche" de to vurderingene av samme elevbesvarelse. Dette har betydd at tross et rimelig høyt antall besvarelser var antallet som ble registrert med minst to vurderinger, vesentlig lavere. Tabell 3.1 gir en oversikt over utvalgsstørrelsene våre.

Boikott av prøvene har i varierende grad vært et problem for årets prøver. Særlig gjelder dette for prøvene på grunnkurs, men det er også et betydelig innslag av boikott på 10. trinn (se tabell 3.2). I tillegg til selve fraværet kan det hende at det på noen skoler kan være spredt en viss svak motivasjon for å gjøre sitt beste ved prøvene. Slike effekter er det imidlertid umulig ut fra våre data å kunne si noe sikkert om, men MMI-rapporten indikerer at dette kan være tilfellet. Verre er det at vi har snakket med lærere som har overhørt samtaler der det er vedgått at elevene har fått betydelig hjelp av lærer under gjennomføringen. Det er umulig for oss å ta stilling til slike rykter, men bare at slike rykter går, er nok til å så tvil om gjennomføringen i praksis har foregått etter forutsetningene.

For elever på grunnkurs er det for de innsendte besvarelsene ikke oppgitt hvilke studieretninger elevene går på. Våre data for grunnkurselevene er derfor mer usikre og mindre informative enn for de andre trinnene. Timetallet i fagene matematikk, norsk og engelsk er lavere på yrkesfaglige studieretninger enn på allmennfaglig, og det faglige fokuset er annerledes. Elevene har derfor nokså ulike forutsetninger for å prestere godt på

prøvene i skriving og lesing på norsk og engelsk, siden disse prøvene er de samme for alle elever på grunnkurset. I matematikk er det tre ulike prøver avhengig av hva slage kurs i matematikk elevene følger, så der er ulike forutsetninger ikke noe problem i denne sammenheng.

De kvantitative og kvalitative metodene vi har brukt i vårt analysearbeid, er beskrevet i detalj i kapittel 4.

1.2 Lesing (Kapittel 5)

1.2.1 10. trinn og grunnkurs

Det var en felles prøve for 10. trinn og grunnkurs. Prøven var i hovedsak bygget over samme lest som leseprøven i PISA-prosjektet og bygger altså på en bred internasjonal tradisjon for hvordan funksjonell lesekompetanse kan måles. Oppgavene er organisert rundt åtte tekster med til sammen 44 oppgaver, hvorav 29 er åpne, i betydningen at de krever skjønnsmessig vurdering av elevenes egenformulerte svar for poengsetting. De resterende 15 oppgavene er flervalgsoppgaver. Særlig i lys av diskusjonene om belastningen lærerne får ved vurdering av de nasjonale prøvene, stiller vi oss uforstående til det lave antallet av flervalgsoppgaver, betydelig lavere enn året før.

Vår vurdering av prøven er at den som helhet har fungert meget bra som en reliabel og valid test av lesekompetanse på de to trinnene. Prøven spenner over flere sjangrer og flere emneområder. Faggruppa har kategorisert oppgavene etter tre delkompetanser, som de forenklet kaller *Finne, Tolke* og *Reflektere*, og de har tilstrebet en god balanse mellom disse tre oppgavetypene.

Så godt som alle oppgavene fungerer godt når det gjelder å diskriminere mellom svake og sterke lesere. Prøven som helhet har høy reliabilitet, både i form av høy indre konsistens (alfa = 0,93) og god overensstemmelse mellom ekstern og intern vurdering for de fleste åpne oppgavene. Imidlertid er avviket stort for noen få av oppgavene, særlig for "to-poengs-oppgavene", de av oppgavene som kan gi opptil to poeng. Det er videre en svak, men ikke foruroligende, tendens til at lærerne gir en litt høyere vurdering av egne elever enn den eksterne vurderingen.

Prøven framstår med rimelig vanskelighetsgrad, med gjennomsnittlige prestasjoner på litt over 60 % av fullt hus, omtrent identisk for de to klassetrinnene. Det hadde likevel vært en fordel om det hadde vært et noe større innslag av litt vanskelige flervalgsoppgaver (se kap. 5.1.3 og 5.1.9). At prøven ikke viser en forventet økning i prestasjoner fra 10. trinn til grunnkurs, legger vi her ikke noe vekt på, dertil er utvalget av elever på grunnkurset for lite representativt. Dette skyldes både at det ikke er skilt tydelig mellom studieretningene, og videre den betydelige effekten av elevenes boikott (se kap. 1.1.2).

Den foreslåtte inndelingen etter tre kategorier av kompetanse (*Finne, Tolke, Reflektere*), inviterer til å lage tre skalaer i tillegg til en overordnet skala for lesekompetanse. Vi kan imidlertid, både ut fra noe for lav reliabilitet og ut fra for sterk korrelasjon mellom kompetansene, ikke finne noen empirisk støtte for å bruke disse tre delkompetansene i

denne prøven (se kap. 5.1.4 og 5.1.7). Innholdsanalysen av validitet peker i samme retning (kap. 5.1.2). Vi anbefaler derfor at man ved en eventuell offentliggjøring nøyer seg med å rapportere resultater langs den overordnete skalaen. Det ligger i sakens natur at siden de tre delkompetansene framstår med lav validitet og reliabilitet, kan disse heller ikke anbefales brukt i det pedagogiske arbeidet på skolene.

1.2.2 4. og 7. trinn

De to prøvene for disse klassetrinnene har mange fellestrekk og er utviklet av en faggruppe som ikke er identisk med den som lager prøven for de eldre elevene. De tydelige fellestrekkene er disse:

- Prøvene består av få (henholdsvis to og tre) lange tekster. Til hver tekst er det knyttet rundt 15 oppgaver.
- Tekstene har et felles tema, men representerer ulike sjangrer. Én av tekstene er skjønnlitterær.
- Prøvene består bare, eller nesten bare, av flervalgsoppgaver, og disse er av en meget spesiell type: De har fire graderte svaralternativer, der det "riktigste" svaret gir 3 poeng, det nest beste 2 poeng osv.
- Faggruppa har foreslått en rapportering med en poengsum for hver tekst for seg.

På noen måter har prøvene fungert tilfredsstillende. De fleste oppgavene diskriminerer bra, og de åpne oppgavene på 7. trinn har høy sensorreliabilitet, noe som tyder på grundig utprøving. Vi har likevel flere innvendinger mot disse prøvene. Spesielt er vi sterkt kritiske til det graderte flervalgsformatet. En ting er at det er vanskelig ut fra dataene å forsvare de graderte poengene som er gitt, men verre er det at vi for de fleste oppgavene ikke kan forstå at det er gode argumenter for at det ene (ikke riktige) svaret er noe "bedre" enn det andre. I tillegg har vi pekt på at formatet inviterer til gjetting, i og med at dette automatisk fører til et betydelig antall poeng. I noen oppgaver er det et problem at svaralternativer som konstrueres for å være nesten rette, kan bidra til at elever som har forstått det de leste, likevel ikke får høyeste poengsum. Vi mener disse forholdene har ført til en vesentlig svekkelse av prøvens validitet, i den forstand at de poengene en elev oppnår, i større grad enn ønskelig henger sammen med noe annet enn elevens leseforståelse.

Faggruppa har ifølge informasjonsmaterialet til prøven valgt å bruke graderte svar-alternativer ut fra to forhold:

- Fjorårets prøver fikk en god del kritikk fordi noen av flervalgsoppgavene hadde "gale" svaralternativer som hadde for mye riktig i seg.
- Det hevdes at siden leseforståelse "*ikke er 0 eller 100*", er det en fordel med "*graderte svaralternativer der eleven også kan krediteres for delvis forståelse*".

Etter vår oppfatning burde svaret på det første punktet vært å lage distraktorer ("gale" svaralternativer) som definitivt *ikke* var riktige. Når det gjelder det andre punktet, vil vi hevde at kreditering for delvis forståelse langt fra gir mening for alle spørsmålene, spesielt ikke for slike som definitivt bare har ett riktig svar. Vi har kommentert mange slike eksempler i kapitlene 5.2.2 og 5.3.2.

Med graderte svaralternativer er det av matematiske grunner lettere å oppnå høyere korrelasjoner oppgavene imellom (høyere indre konsistens), og dermed høyere reliabilitet for prøven som helhet. Man kunne derfor ha ventet en høyere reliabilitet på årets prøve enn på fjorårets prøve for 4. trinn. Men faggruppa har heller ikke lykkes med dette, siden reliabiliteten ikke er blitt spesielt høy. For 4. trinn var reliabiliteten høyere for fjorårets prøve (se kap. 5.3.5), som etter vår mening på tross av kritiske innvendinger også hadde høyere validitet.

Faggruppas forslag om rapportering etter en egen skala for hver tekst får av to grunner ikke støtte gjennom våre analyser. For det første har ikke hver av disse foreslåtte skalaene høy nok reliabilitet til at vi kan anbefale dem som basis for rapportering. For det andre er det vanskelig å se at de ulike tekstenes oppgaver virkelig måler vesensforskjellige typer av kompetanser, selv om tekstene representerer ulike sjangrer. Verken rapportering på Skoleporten eller den pedagogiske tilbakemeldingen til lærer og elev er tjent med at Per rapporteres med høyere kompetanse i en tekstsjanger enn en annen når disse resultatene i så stor grad som her kan tilskrives de konkrete tekstene som har vært brukt. Å generalisere fra én konkret skjønnlitterær tekst (en fantastisk fortelling eller dagbokutdrag) til lesing av all skjønnlitteratur er ikke holdbart. Parallellen til internasjonale undersøkelser gir ikke noe godt argument. Slike undersøkelser har såkalt "rotated design", der hver elev riktignok har få tekster, men siden det i alt er mange forskjellige hefter med forskjellige tekster, blir resultater for store grupper av elever alltid basert på mange slike tekster.

Våre konklusjoner for 4. og 7. trinn er:

- For hver av disse prøvene bør det bare rapporteres etter én overordnet skala for kompetanse i lesing.
- Vi anbefaler innstendig at formatet med graderte flervalgsoppgaver unngås i framtidige prøver.

1.2.3 Et felles rammeverk for leseprøvene

Vi vil sterkt understreke behovet for et felles grunnleggende rammeverk som teoretisk og operasjonelt definerer det som måles i leseprøvene på de ulike trinnene. I et slikt dokument bør tekst- og oppgaveformater beskrives og begrunnes, og spesielt bør forskjellene mellom prøvene gis en felles begrunnelse. Det bør også komme klart fram hvilke rapporteringskategorier det tas sikte på og begrunnelsen for disse. Endelig bør dette dokumentet referere tydelig til læreplanene på de ulike trinnene. En tydelig beskrivelse av progresjonen mellom trinnene kan med fordel også inngå.

Et rammeverk etter disse retningslinjene vil kunne heve den fagdidaktiske bevisstheten i diskusjonen rundt prøvene og bidra til klarhet når det gjelder prøvenes mening og mål. Dette vil også kunne bidra til et bedre overordnet perspektiv på leseprøvene, som hittil litt for tydelig bærer preg av å være utviklet av to forskjellige fagmiljøer med ulike definisjoner av leseforståelse og praktiske tilnærminger til måling av dette.

1.3 Skriving (Kapittel 6)

1.3.1 Generelt

Siden skriveprøven på grunnkurs var frivillig, og bare ca 4 % av elevene har gjennomført denne, har vi her nøyd oss med å diskutere prøvene for 4., 7. og 10. trinn.

Elevene ble prøvet i fri skriving, og prestasjonene ble vurdert etter en rekke spesifiserte kriterier. For hver variabel ble elevene vurdert etter en tredelt skala, hvorav de aller fleste elevene var ment å havne, og også faktisk havnet, i den midterste kategorien, "omtrent som forventet". Det understrekes her at "som forventet" relaterer seg til klassetrinnet. Skalaen representerer derfor ingen ambisjoner om å etablere en absolutt skala som basis for kriterierelatert vurdering, i retning av det som er gjort med CEF-skalaen for engelsk (se kap. 7).

1.3.2 Oppbygning og gjennomføring

Prøvene for de tre klassetrinnene var bygget opp på helt lik måte, så vi behandler dem her under ett. Detaljer om prøvene og våre resultater er gjengitt i kapittel 6. Prøvene består hver av to oppgaver, og hver av disse forutsetter at elevene skriver en sammenhengende tekst av betydelig lengde. På forhånd hadde elevene fått utlevert et hefte med informasjon om planeten Mars, og begge oppgavene var på ulik måte relatert til denne informasjonen.

Besvarelsene ble vurdert av egne lærere og eksterne sensorer etter en meget detaljert vurderingsmal, som er godt begrunnet i rammeverket for prøvene. Hver av de seks aspektene ved skrivekompetanse ble vurdert for hver oppgave, og for mange av aspektene ble det brukt mer enn én variabel. I tillegg var det en kategori for førsteinntrykket. De seks aspektene er:

1. Kommunikasjon
2. Innhold
3. Tekstoppbygging
4. Språkbruk
5. Rettskriving og tegnsetting
6. Grafisk utforming

Et stort problem har vært at nesten alle elevene har havnet på det midterste nivået, "omtrent som forventet", for alle variablene. Dette gir svært liten informasjon til skolens lærere om elevenes skrivekompetanse, og flere lærere har reagert på dette, ifølge MMI-undersøkelsen. En rektor på en ungdomsskole (se vedlegg) kom i et intervju med oss med følgende uttalelse om den pedagogiske verdien av dette:

> "Lærerne har derimot hatt ulike følelser knyttet til nasjonale prøver, og særlig norsklærerne har hatt mye frustrasjon knyttet til prøvene. Etter å ha brukt mye tid på 10. trinn på å lese informasjon om skriveprøvene i norsk, gjennomføre dem og rette dem, var lærerne svært skuffa da de fikk resultatene i form av "som forventet". To språklærere med 25 års erfaring følte det nærmest provoserende. Etter å ha lagt ned rimelig mye tid, opplevde de det som nedbrytende å få så lite informasjon tilbake. De uttalte: Hva tror de (fagmiljøet) om oss norsklærere og

vår kompetanse? De mener selv at de kjenner sine egne elever svært godt etter tre års undervisning, og derfor ikke fikk noe igjen for denne skriveprøven."

På den annen side er det åpenbart at prøvene har medført en ikke ubetydelig øking av vurderingskompetansen hos lærerne, spesielt på barnetrinnet, der vurderingstradisjonen er lite utviklet og systematisert. Det framgår av MMI-undersøkelsen at mange lærere har uttrykt seg positivt om dette. En rektor i barneskolen sier det slik (se vedlegg):

> "Jeg ser at barneskolelærerne har fått ord og begreper de ikke hadde før. I norsk skriftlig har jeg en lærer som har hatt en aha-opplevelse på 7. trinn, nettopp i forhold til ord og faglige begreper i vurderingsarbeidet."

Som redegjort for i rammeverket, var det ikke foreslått å komme fram til et samlet resultat for skrivekompetanse. Derimot ble det invitert til å bruke hele rekka med variabler som "resultater" for enkeltelever.

1.3.3 Våre analyser

Resultatene fra våre analyser underbygger tydelig punktene ovenfor:

- For hver kompetanse havner de aller fleste elevene i den midterste kategorien, "omtrent som forventet", så den pedagogiske informasjonen til skolene er lite informativ for de fleste elevene. Det synes å være lite pedagogisk informasjon å hente fra denne prøven ut over det lærerne allerede vet om elevene sine.
- Det er rimelig god overensstemmelse mellom sensorer angående *hvor mange* elever som fortjener "høyere" eller "lavere" enn forventet, men svært dårlig overensstemmelse når det gjelder *hvilke elever* dette gjelder. Med så lav sensorreliabilitet som er påvist her, vil det etter vår mening være meningsløst å rapportere noe på nettet fra denne prøven.

Den pedagogiske informasjonen er altså ikke bare lite informativ, men den er også lite pålitelig. Det er ut fra dette naturlig å konkludere at skriveprøven i sin nåværende form ikke fyller sin funksjon som nasjonal prøve for å måle skrivekompetanse(r). Men det paradoksale etter vårt syn er at dette heller ikke synes å være målet bak utviklingen av disse prøvene. Gjennom sitt meget teoretiske og faglig ambisiøse og grundige rammeverk for skrivekompetanse har faggruppa levert et viktig arbeid for å skape en bevissthet i norsk skole om hva skrivekompetanse går ut på, og ikke minst på en grundig måte gjort rede for hvorfor slik kompetanse ikke kan ses uavhengig av *hva* det skal skrives om. Faggruppa har rett og slett på en overbevisende måte godtgjort at skrivekompetanse som et endimensjonalt begrep målt ved *en* enkelt skriveprøve er meningsløst. Faggruppa har da også anbefalt at prøven er å betrakte som ett ledd i en større sammenheng der også elevenes mappe med allsidige skriftlige arbeider inngår. Det ligger imidlertid utenfor vårt mandat å diskutere dette perspektivet her.

Resultatene av vår analyse viser at selv om vi begrenser skrivekompetansen tematisk og legger til rette for at alle elevene har et faglig fundament å skrive ut fra, er det svært vanskelig å oppnå høy enighet om bruk av kvalitetskriterier i praksis. Det er langt fram til et godt tolkningsfellesskap når det gjelder de foreslåtte aspektene av skrivekompetanse. Vi anbefaler derfor ikke å fortsette med slike obligatoriske skriveprøver i årene framover.

Det går selvsagt an, og det kan diskuteres om det er aktuelt, å lage helt andre typer prøver, prøver som på en mer strukturert måte måler delferdigheter som syntaks, rettskriving og kommaregler. Vi tviler imidlertid på at sterkt fokus på slike ferdigheter vil være til gagn for skrivekompetansen i skolen, og vi vil ikke gjøre oss til talsmenn for en slik utvikling. Derimot kan det tenkes at noen innslag av slike oppgaver kan ha sin plass sammen med fri skriving i en framtidig prøveform, for derigjennom å øke reliabiliteten. Inntil det er dokumentert at vi kan oppnå høy enighet om vurdering av elevenes kompetanse i fri skriving, foreslår vi at skriveprøvene utgår og erstattes av en sterkere satsing på skriveopplæring og læreres kompetanseheving i vurdering av elevenes skriveferdigheter. I en slik satsing vil faggruppas arbeid gjennom en forenklet versjon av rammeverket kunne spille en viktig rolle.

Våre anbefalinger er derfor:

- Det bør ikke rapporteres noe på nettet fra disse skriveprøvene.
- Skriveprøvene som obligatoriske prøver i sin nåværende form opphører.
- Rammeverket for prøvene gjøres tilgjengelig for lærere i en enklere form og brukes som basis for en sterkere satsing på elevenes skriveopplæring og lærernes kompetanseheving når det gjelder vurdering.

1.4 Engelsk (Kapittel 7)

1.4.1 Engelsk skriving

Vurderingskriterier

Prøvene i engelsk skriftlig på 7. trinn, 10. trinn og på grunnkurset består av fri skriftlig produksjon basert på tre oppgaver i ulike sjangrer. Oppgavene skal vurderes på tvers av de ulike sjangrene for *Språk*, og for hver oppgave separat etter et oppgavespesifikt vurderingsskjema for *Innhold.* Faggruppa i engelsk har i år som i fjor tatt utgangspunkt i kompetansenivåene som er beskrevet i *Common European Framework of Reference for Languages: Learning, teaching and assessment* (heretter forkortet til *Rammeverket* og CEF-nivåer). Ferdigheter i fremmedspråk finnes her beskrevet på tre hovednivåer:

- Det laveste nivået *A - Basic user,* med oppdeling i *A1* og *A2.*
- Mellomnivået *B - Independent user,* med oppdeling i *B1* og *B2.*
- Det øverste nivået *C - Proficient user,* med oppdeling i *C1* og *C2*. Det høyeste nivået C2 anses imidlertid som uaktuelt for norske skoleelever.

Faggruppa i engelsk har i tillegg til nivåene beskrevet ovenfor laget mellomnivåer for å få en mer differensiert vurdering av elevenes kompetanse. Med mellomnivåer består den aktuelle skalaen av nivåene A1, A1/A2, A2, A2/B1, B1, B1/B2, B2, B2/C1 og C1. Denne nivåskalaen er den samme som ble benyttet for fjorårets prøver. B1 er det såkalte terskelnivået som anses som det språknivået man bør beherske for å kunne fungere sosialt og samfunnsmessig på et fremmedspråk.

Et hovedpoeng med en slik skala er at den er ment å representere en *absolutt* skala for kompetanse slik at elevene kan måle sin egen framgang over tid. Men vi finner ulike

beskrivelser for C1- nivået, for eksempel for grammatisk korrekthet på ulike trinn. Det kreves *"En meget høy grad av grammatisk korrekthet"* på 10. trinn, mens det for grunnkurset kreves *"Høy grad av grammatisk kontroll"*. Disse beskrivelsene er kanskje tilnærmelsesvis like i en praktisk vurdering, men det virker svært ulogisk at de ikke er identiske. Men det blir en enda mer uforståelig relativisering av de absolutte nivåene når 7. trinns B2- beskrivelse er *"Høy grad av grammatisk kontroll. Avvik kan forekomme i forbindelse med komplekse strukturer"*. Dette vil si at høy grammatisk korrekthet gir B2 på 7. trinn og C1 på grunnkurset. Fremskrittene som elevene gjør når de beveger seg oppover nivåene, blir på denne måten ikke tydelige.

I likhet med skriveprøvene i norsk medfører det en stor utfordring for lærere og de eksterne sensorene når det gjelder å enes om tolkninger av vurderingskriterier og vurdering av tekster. Dette ser vi tydeligere når nivåene i CEF skal operasjonaliseres, og nivåene brytes ned til A1, A1/A2, A2, A2/B1, osv. slik det er gjort i forbindelse med nasjonale prøver i engelsk skriving. Det blir utydelige forskjeller mellom nivåene. For å nå ett nivå, må elevene gjennom tekstene de har skrevet vise at de har nådd kriteriene for det gitte nivå. Det er her utfordringene for felles forståelse og vurdering oppstår. Et nytt vurderingssystem må etterprøves før man innfører dette som et nytt nasjonalt system. Vi ser fra dataanalysen av resultatene og fra vurderingseksperimentet (se kap.7.1.7) at lærerne er rimelig enige om hva som er gode og dårlige besvarelser, men de legger seg på ulikt nivå, slik at under halvparten av besvarelsene havner på samme CEF - nivå av to uavhengige sensorer.

Et annet problem med vurderingen av engelsk skriftlig er knyttet til selve innføringen av CEF, som fortsatt er ukjent for de fleste norske engelsklærere. Da mange lærere følger elevene sine over flere år, er det grunn til å regne med at vurderingen av elevers skriftlige besvarelser i 2005 antakeligvis også er utført av andre lærere enn i 2004. I følge MMI-undersøkelsen var det kun 14 % av lærerne som hadde vurdert nasjonale prøver tidligere. Selv om en oppnår et tolkningsfellesskap om CEF-nivåene over tid, er det ikke nok å ha flere eksterne sensorer, men de som vurderer oppgaver, må også ha tid til å snakke sammen. Det vil si en ordning tilnærmelsesvis lik den som eksisterer for skriftlig avgangsprøve i engelsk. Tendensen til at noen lærere systematisk vurderer sine egne elever høyere enn eksterne vurderere gjør, fører til at det er vanskelig å oppnå høy nok reliabilitet.

Det er rimelig godt samsvar mellom oppgaver/vurderingskriterier, rammene for prøven og det prøven skal måle, men dette gjelder ikke for den argumenterende teksten på de to øverste trinnene. Likeledes mener vi at oppgave 1 på 7. trinn ikke hadde god nok instruks for å gi grunnlag for å vurdere ordforråd (se kap. 7). Når det gjelder prøvenes relasjon til gjeldende læreplaner, er denne uklar. Språkdefinisjonen som prøvene baserer seg på, samsvarer godt både med L97 og R94, men det å kunne skrive en argumenterende tekst, er for eksempel ikke et krav i L97, men derimot i R94. En av tre lærere (på de tre trinnene som er testet) mener ifølge MMI-undersøkelsen at prøvene reflekterer læreplanen i "ganske stor grad". Men likevel mener både elever og lærere at det er vanskelig å få vist bredden i skriftlige engelskferdigheter gjennom de nasjonale prøvene.

På tross av de problematiske sidene vil vi understreke at bruk av CEF kan være verdifullt i en opplæringssituasjon, og særlig sammen med mappevurdering eller utvikling av språkpermer. Det er positivt at lærere utvikler et språk for vurdering ved hjelp av CEF, særlig på barnetrinnet hvor man ikke har hatt samme vurderingskultur som på ungdomstrinnet. Det at CEF ikke fungerer pålitelig nok som måleinstrument i nasjonale prøver, forhindrer ikke at det med fordel kan brukes i en opplæringssituasjon i skolen. Vi vil heller argumentere for at det er nettopp i praktisk arbeid med språklæring i skolen CEF vil kunne fungere som et godt redskap for lærere og elever.

Analyser av resultater for engelsk skriving

Med dataene både fra de nasjonale prøvene og fra vårt vurderingseksperiment (se kap. 7.1.7) har vi tydelig påvist at det er svært vanskelig å oppnå en pålitelig vurdering av elevenes prestasjoner. Vi har sett at problemet for sensorene særlig ligger i å bestemme hvor nivået skal ligge, mens det er noe enklere å rangere elevene innbyrdes. Det er altså stort sett enighet om hvilke elever er gode og hvilke er svake, men *hvor* gode eller *hvor* svake besvarelsene er, er langt mer usikkert. Resultatet av dette er blant annet at de skolevise resultatene (i form av gjennomsnittsverdier) i altfor liten grad reflekterer elevenes kompetanse, men snarere avhenger av hvordan lærerne har brukt skalaen; altså hvor strengt eller mildt de har vurdert besvarelsene.

Et annet gjennomgående trekk for alle trinn har vært at klassens egen lærer gir høyere nivå enn den eksterne sensoren (se kap. 7.1). Det ser videre ut til å være vanskeligst å vurdere 7. trinns besvarelser, idet her er spriket størst mellom ekstern og intern sensor. Ifølge studentene som fungerte som sensorer i vårt vurderingseksperiment, var det største problemet at det ikke fantes retningslinjer for hvordan man skulle vurdere innslag av norske ord og manglende besvarelser. For 10. trinn er reliabiliteten noe bedre, selv om de interne fortsatt tenderer mot å være noe "mildere" enn de eksterne. For kategorien *Formidling* er spriket størst, hvor hele 40 % av elevene vurderes høyere internt. I år hadde en del elever (om lag 500 i våre data) på 10. trinn to eksterne sensorer. Mellom de to eksterne sensorene er samsvaret mindre enn mellom ekstern og intern sensor. Idet det høyst sannsynlig er nye lærere som vurderer sine elever, kan bedring ikke forklares med økt tolkningsfellesskap. For grunnkurset er antall innsendte data så lavt (126) at det er vanskelig å si noe sikkert om reliabilitet mellom interne og eksterne vurderere, men den sammen tendensen gjør seg gjeldende her med lav reliabilitet.

Resultatene fra engelsk skriftlig er tenkt å rapporteres etter tre kategorier, *Språk, Formidling* og *Totalt.* De to første kategoriene er ment å gi diagnostisk informasjon om elevenes faglige profil. Analyser har vist at det er svært liten forskjell mellom kategoriene *Språk* og *Formidling,* og at kategoriene *Språk* og *Totalt* framstår som nesten like, kanskje ikke uventet når språklige forhold skal telle mest. Når nivåene for *Språk* og *Formidling* viser seg å være nærmest identiske, er det liten diagnostisk verdi i en slik nyansert tilbakemelding.

I en kommentar fra faggruppa i engelsk til rapporten *"Nasjonale prøver på prøve"* i fjor, var ett av argumentene at fagmiljøet var redd for at dersom ikke prøvene ble offentliggjort på Skoleporten, ville man gå glipp av en positiv tilbakevirkning, det vil si

hvis ikke fri skriftlig produksjon ble testet og publisert, ville lærerne slutte med å vektlegge denne type fri skriving. For det første har fri skiving vært praktisert lenge i norsk skole, ikke minst på grunn av tilbakevirkningseffekt av skriftlig eksamen på 10. trinn og på grunnkurset. For det andre er ikke formidling vektlagt i særlig grad i vurderingen av de nasjonale prøver, idet språket skal telle mest. Som påvist i kap. 7 har også de såkalte oppgavespesifikke nivåbeskrivelsene for *Formidling* liten vekt på innhold.

I vår tenkning er en offentliggjøring av resultatene i seg selv ikke det viktigste spørsmålet. Mest uheldig er det at man informerer elever og foreldre om nivåer og fagprofiler som er usikre. Motivasjonsgevinsten med å operere med absolutte nivåer og muligheter for å spore framgang, blir også uklar. Elever kan til og med risikere nedgang i nivå, og informasjonsverdien blir liten. Gallupundersøkelsen for Utdanningsforbundet og MMI-undersøkelsen for Utdanningsdirektoratet dokumenterer også at de nasjonale prøvene ser ut til å ha liten informativ verdi for skolen og lærerne. Når det gjelder analyser av resultatene, ser vi ikke uventet at det er en økning i gjennomsnittsnivå for de trinnene som har tatt skriftlig prøve fra 3,5 (mellom A2 og A2/B1) for 7. trinn, til 4,9 (nesten B1) for 10. trinn, til 5,2 (solid B1) for grunnkurset. I så måte kan man si at prøven kan påvise samlet framgang for hele elevgrupper i engelsk skriftlig, men for den enkelte elev er det for stor usikkerhet knyttet til fastsetting av CEF-nivå.

I brev til UFD i januar 2003, advarte Kjell Lars Berge, Lars Sigfred Evensen og Frøydis Hertzberg mot det de kalte "…*naive forestillinger om at det var enkelt å oppnå valide og reliable resultater av skriveprøver*". Vi kan ikke se annet enn at engelsk skriving står overfor den samme utfordringen, og vi er også i tvil om det er mulig å oppnå dette på sikt så lenge den samme modellen ligger til grunn.

1.4.2 Engelsk lesing

Prøvene i engelsk lesing er elektronisk basert for både 4., 7., 10. trinn og for grunnkurset. Alle prøvene har ulike former for flervalgsoppgaver, med alt fra to til fem svaralternativ. I noen oppgaver skal man markere hvilken tekst som passer til et bilde, i andre oppgaver skal elevene lese en tekst for deretter å vurdere om en påstand er "true or false". Det er naturlig nok noen flere bildeoppgaver på de to laveste trinnene, mens man for de to øverste trinnene har noen flere oppgaver som tester ordbegreper gjennom å lese tekster for deretter å finne ord som kan bety det samme som begreper hentet fra tekstene.

Prøvene er satt sammen av ulike antall oppgaver, med noen færre for de laveste trinnene. Elevene møter ulike teksttyper som beskjeder, postkort, brev, beskrivelser, historier og faktatekster, og blir deretter bedt om å løse ulike flervalgsoppgaver. BITE-IT rapporten, som er faggruppens rapport om utviklingen av leseprøven og resultater fra 2004, refererer til følgende grunnmodell for prøvene:

- Forprøve – et oppgavesett bestående av om lag 20 oppgaver som målte elevenes omtrentlige nivå for å finne riktig hovedprøve
- Hovedprøve – et oppgavesett som inneholder om lag 35 oppgaver som skal måle elevenes faktiske nivå relatert til CEF.

Prøvene er altså adaptive, og det er utviklet en utvelgelsesalgoritme som sørger for at hver enkelt elev får en hovedprøve på riktig nivå basert på elevenes prestasjoner i forprøven. *"Leseprøvene i engelsk er laget for datamaskiner og er adaptive, det vil si at de tilpasser seg elevenes ferdighetsnivå automatisk" (...) "Alle elevene vil få oppgaver som er individuelt tilpasset deres ferdighetsnivå" (BITE-IT-rapporten 2004).*

I vurderingsveiledningen kan vi lese: "*De nasjonale leseprøvene i engelsk skal måle elevenes grunnleggende ferdigheter i å lese engelsk tekst. Det vil si at prøvene måler en del av det elevene lærer i engelskfaget, mens andre deler av faget må vurderes på andre måter*". Leseferdigheter blir videre forklart med at elevene skal kunne:

- trekke konklusjoner
- forstå hovedinnholdet
- forstå detaljer, hente ut informasjon

Leseprøvene i engelsk er også knyttet til det europeiske *Rammeverket* (se kap. 1.4.1), og etter å ha tatt prøven vil elevene umiddelbart få tilbakemelding om hvilket nivå de er på. I 2004 var resultatet overraskende godt, idet 90% av elevene på 10. trinn fikk B1, B1/B2 eller B2. For 2005 er innslaget over B2 en god del høyere (se for øvrig kap.7). Dersom vi ser på vurderingskriteriene for B1, sies det blant annet følgende om kriterier på B1- nivå for 10. trinn:

- I can understand the main points in most straightforward factual texts with a clear structure, if the theme is familiar.
- I can understand most stories written specially for learners my age.
- I can understand the main point from a narrative written for native speakers my age, with some background knowledge or support, e.g. from pictures.
- I can follow sets of straightforward instructions.
- I can scan straightforward, clearly-structured, longer texts to find specific information.

Det virker for oss urimelig at det bare er rundt én prosent av elevene som ikke mestrer dette terskelnivået for lesing (se kap 7.2.3). Når man ser på CEF- kriteriene og resultatet fra 2004 og 2005, er det et åpent spørsmål om leseoppgavene i engelsk har truffet riktig nivåplassering idet så få elever er på et A-nivå.

Det at læreren ikke har innsyn i selve oppgavene, hvordan elevene har svart på dem, hvilken type leseferdighet som har voldt problemer, samt hva som er riktige og gale svar, gjør at vi ikke kan se at denne leseprøven kan brukes pedagogisk i undervisningen. Et nivå i seg selv er et usikkert utgangspunkt for individualisert leseopplæring.

Ved analyse av elevbesvarelser oppgave for oppgave på leseprøven, ser vi at mange av oppgavene diskriminerer godt, men det er også en god del oppgaver som burde vært luket ut i forkant, siden de framstår med så lav diskriminering at de bidrar til å svekke prøvenes reliabilitet. Grunnen til lav diskriminering er i mange tilfeller at det er brukt oppgaver av formatet riktig/galt i kombinasjon med at elevene må svare for å komme videre i prøven. Vi mener prøvene ved enkle grep kan forbedres.

Det er lagt ned et verdifullt arbeid med utvikling av de nasjonale lesetestene. Det er positivt at det er laget prøver som lærere slipper å rette, og hvor elever kan få resultatet umiddelbart. Dette kan være motiverende for elever.

I kapittel 7.2.3 har vi gitt og kommentert resultater for de fleste hovedprøvene. Men det er fortsatt noe uklart for oss hvordan de ulike hovedprøvene på samme trinn fungerer sammen når det gjelder nivåplassering. Vi betviler at resultatene er pålitelige nok til å kunne publiseres.

1.5 Matematikk (Kapittel 8)

1.5.1 Generelt

Matematikkprøvene for de ulike trinnene har mye til felles. De består av en rekke uavhengige oppgaver, som nesten alle er i et åpent format, altså at elevene skriver svarene selv. Besvarelsene må vurderes etter en detaljert veiledning, der det er angitt oppgavespesifikke koder for kategorier av svar. Flervalgsoppgaver finnes nesten ikke, og de meget få slike oppgavene blir kodet på linje med de åpne oppgavene.

1.5.2 Validitet og oppgavenes kvalitet

Etter vår vurdering inneholder prøvene i hovedsak svært gode oppgaver, og til sammen framstår prøvene med høy validitet i forhold til læreplanene. Faggruppa har ikke gitt noen vurdering av forholdet mellom fagmålene i læreplanen og de grunnleggende ferdighetene. Det er, slik vi forstår det, uklart i hvor stor grad de nasjonale prøvene er ment å konsentrere seg om disse grunnleggende ferdighetene. Et spørsmål er for eksempel om de nasjonale prøvene skal ha et mer begrenset faglig siktemål enn vi er vant til at skriftlig eksamen har. Matematikkprøvene har tydeligvis hatt som mål å dekke hele ”pensum” på en god måte. En avklaring av dette for framtidige prøver vil være viktig. Dette er særlig aktuelt for grunnkurs. I den grad prøven er ment å i hovedsak teste grunnleggende ferdigheter, kan man tenke seg at alle grunnkurselevene får samme prøve, slik som delprøve 1 fungerte i år. De spesifikke kravene for hvert kurs kan naturlig dekkes gjennom skriftlig eksamen.

Det er en fin balanse mellom ulike faglige emner og mellom ferdig strukturerte regneoppgaver og mer logisk krevende problemer. I tillegg er mange av oppgavene svært gode i et diagnostisk perspektiv, idet de på en fin måte måler elevenes grunnleggende forståelse og avdekker typiske misoppfatninger. Det er mye å lære for elevene av en tilbakemelding av resultater, vel å merke ved en gjennomgang oppgave for oppgave. Diagnostisk sett er dette gode prøver, men gode rapporteringsskalaer krever noe mer enn en samling av ”gode” oppgaver. Kvaliteten til de foreslåtte rapporteringsskalaene kommer vi tilbake til nedenfor.

1.5.3 Vurdering av besvarelser

I hovedsak har vurdering av elevenes besvarelser foregått med høy sensorreliabilitet. For noen av oppgavene er det svært mange koder, helt opp til 12 for en oppgave, og mange

av disse viser seg å representere helt marginale elevsvar. Etter vår mening har det gått inflasjon i koder, noe som har gjort det vanskelig å få oversikt. Det anvendte kodesystemet framstår videre som ulogisk, idet samme tallkode kan representere ulik poengverdi fra oppgave til oppgave. Av disse grunnene tror vi at vurderingsprosessen har vært unødig komplisert og tatt for lang tid. MMI-undersøkelsen viser at lærerne i gjennomsnitt har brukt 25 minutter per besvarelse på vurderingen. Den høyeste tidsbruken rapporteres på 7. trinn; 34 minutter per besvarelse. Vi anbefaler derfor for eventuelle framtidige prøver:

- å skifte til et kodesystem som følger prinsippene i studier som TIMSS og PISA. Her anvender man kodesystemer hvor første siffer alltid angir poeng, mens det andre sifferet bare har en kategorisk funksjon. Dette skaper bedre oversikt og gjør det lettere å sette seg inn i systemet.
- å redusere antall koder, basert på pilotering, ved at bare de kodene inkluderes som forekommer rimelig hyppig. Eventuelt kan man vurdere for noen av eller alle oppgavene å kun kode etter poeng.

Oppgavene varierer når det gjelder hvor mange poeng elevene kan få. De fleste kan gi ett eller to poeng, men det er også noen få "trepoengsoppgaver". For de fleste av disse sistnevnte finner vi poenggivingen dårlig motivert og vanskelig å forsvare ut fra elevdata. At oppgaver er vanskelige eller tidkrevende, er psykometrisk sett ikke noe godt argument for at de skal tillegges flere poeng. Det viktige er at hvert poeng må kunne defineres slik at det diskriminerer på en pålitelig måte. Vi anbefaler derfor for eventuelle framtidige prøver:

- i hovedsak å nøye seg med å gi ett eller to poeng for den typen åpne oppgaver som er gitt i årets prøver.
- at eventuelle skiller mellom ett og to poeng prøves grundigere ut gjennom pilotering.

Vi mener videre at et sterkere innslag av flervalgsoppgaver vil være en betydelig fordel for framtidige prøver. Vurderingen vil for det første bli mye enklere, og i tillegg vil det bli en bedre variasjon i formatet, antall blanke svar vil gå betydelig ned, og reliabiliteten vil trolig bli høyere.

Vurderingen av elevenes besvarelser har bestått av to trinn. Først har lærerne satt koder for type svar, og deretter har de overført denne informasjonen fra kode til poeng. Faggruppa har laget et regneark i Excel som gjør dette på en enkel måte og samtidig beregner summer av poeng med grafikk for delskalaer og for hele prøven. Imidlertid er det tydeligvis ikke alle lærere som mestrer slike verktøy. Rapporteringen av resultater til oss viste seg å bestå av en blanding av Excel-filer, utskrift av disse, samt av håndskrevne lister. De nevnte poengsummene med grafikk representerer for øvrig et annet problem, idet de innbyr til ukritisk bruk av upålitelig informasjon, se mer om dette nedenfor.

1.5.4 Rapportering og egenskaper til foreslåtte skalaer

Hver av prøvene består av en rekke enkeltoppgaver, og hver av disse ble vurdert med koder, som deretter ble overført til poeng. Faggruppa har tatt sikte på å rapportere etter

tre kompetansekategorier, og hver oppgave er plassert i en av disse. De tre kompetanseområdene er:

1. Representasjoner, symbolbruk og formalisme (av oss kalt RSF)
2. Matematisk resonnement, tankegang og kommunikasjon (RTK)
3. Matematisk anvendelse, problembehandling og modellering (APM)

For hver av de tre kategoriene får hver elev en poengsum som er ment å representere elevens kompetanse på dette området. På det tidligere omtalte regnearket framkommer disse summene som grafiske framstillinger som viser oppnådde poeng i forhold til fullt oppnåelig. Det er videre angitt en prosedyre for hvordan disse poengsummene skal overføres til kompetansenivåer fra 1 til 5. Imidlertid er det helt uklart hvordan disse kompetansenivåene skal brukes videre.

Vi er av flere grunner sterkt kritiske til disse prosedyrene. For det første har vi i vår analyse av data fra prøvene påvist at disse skalaene i de aller fleste tilfeller ikke har høy nok reliabilitet til at de kan sies å representere pålitelig informasjon om elevenes kompetansenivå. Dette medfører at det ikke er aktuelt å rapportere disse tre delskalaene i tillegg til den totale skalaen på Skoleporten (med et par mulige unntak som vi kommer tilbake til). Et større problem er imidlertid at de ovenfor omtalte diagrammene som kommer ut av regnearket lett kan gi opphav til alvorlige feiltolkninger om elevenes kompetanse. Hvor høyt en elev skårer i form av prosent av fullt hus på en delprøve, sier like mye om hvor vanskelige oppgavene har vært, som hvor dyktige elevene er. På flere av prøvene, særlig for 10. trinn, er det stor forskjell i vanskelighetsgrad mellom de tre skalaene, og da forteller antall prosent riktige svar i seg selv ingen ting om elevenes kompetanse. Et slikt prosenttall blir først meningsfullt når det sammenliknes med andre elever. Av slike grunner er også de foreslåtte fem kompetansenivåer nokså meningsløse, fordi de innbyr til de samme feiltolkningene. I tillegg har vi motforestillinger mot å introdusere nivåer på en måte som likner på karakterer uten å være det. Vi kan heller ikke se hvordan disse kompetansenivåene er tenkt brukt i rapportering og i det pedagogiske arbeidet på skolene.

Et annet og like viktig spørsmål er om de tre skalaene virkelig framstår som forskjellige, og om denne ulikheten kommuniseres på en forståelig måte. Vår analyse av samsvaret mellom de tre skalaene viser for det første at det ikke synes å være mange nok oppgaver til at tre skalaer får høy nok reliabilitet. Videre viser det seg at de to skalaene RTK og APM nesten ikke lar seg skille og derfor med fordel kan slås sammen. Også begrepsmessig virker dette fornuftig, idet de to beskrivelsene etter vår mening ikke kommuniserer noen tydelig forskjell mellom de to. Dersom skalaen RSF har forholdsvis høy reliabilitet, slik som under tvil for 10. trinn og grunnkurs (1M), kan det forsvares å ha to rapporteringskategorier, henholdsvis RSF og RTK/APM.

For øvrig tror vi at kompetansebetegnelsene er for kompliserte til å kunne formidle noen pedagogisk mening. Vi tviler på at foreldrene, og til og med lærerne, blir særlig godt informert om Pers sterke og svake sider i matematikk ved å få vite at han har ”høyere representasjonskompetanse og kompetanse i symbolbruk og formalisme enn resonnements-, tankegangs- og kommunikasjonskompetanse samt anvendelse-,

problembehandlings- og modelleringskompetanse". Dersom de to siste kategoriene hadde vært slått sammen, ville vi sittet igjen med to kategorier som i hovedsak dreier seg om innøvet ferdighet i regning med tall og symboler for den ene, og resonnering og løsning av problemer for den andre. En slik todeling tror vi kan formidles nokså enkelt med få ord.

1.5.5 Kort om de enkelte prøvene

Resultatene for hver av prøvene er presentert og diskutert i detalj i kapittel 8. Her vil vi gjengi hovedpunktene.

Prøven for 4. trinn har høy validitet, de fleste oppgavene diskriminerer bra eller tilfredsstillende, og det er få blanke elevsvar. Det er generelt godt samsvar mellom intern og ekstern vurdering. Totalt sett framstår prøven som lett, idet gjennomsnittet av oppnådde poeng ligger så høyt som 67 % av fullt oppnåelig skåre. En såpass lett prøve har motivasjonsmessige fordeler for så unge elever. Men en ulempe med en så skjev fordeling (se figur 8.1) er at de få elevene med lavt poengtall vil telle urimelig mye og trekke ned gjennomsnittsverdier for klasser og skoler i for stor grad. Ingen av de tre foreslåtte delskalaene har høy nok reliabilitet til å være aktuelle for rapportering på Skoleporten (se kap. 8.1.4).

Prøven for 7. trinn har stort sett de samme egenskapene som den for 4. trinn. Men av en eller annen grunn er den blitt forholdsvis mye vanskeligere, elevene oppnår i gjennomsnitt bare halvparten av oppnåelig poengsum. En slik vanskelighetsgrad medfører en god og symmetrisk fordeling, men ulempen er mange blanke svar. Ingen av de tre foreslåtte delskalaene har høy nok reliabilitet til å være aktuelle for rapportering på Skoleporten (se kap. 8.2.4).

På 10. klassetrinn er prøven enda vanskeligere, elevene oppnår bare 45 % av "fullt hus". Men vanskelighetsgraden er svært forskjellig mellom de tre kompetansekategoriene, fra 58 % for RSF til så lavt som 33 % for APM. Oppgavene har gjennomgående høy diagnostisk informasjonsverdi, men for mange av oppgavene er det et betydelig innslag av blanke svar. Vi har påpekt at det synes å være noe uheldig og umotivert bruk av så mye som 3 poeng for enkelte oppgaver (se kap. 1.5.3 og kap. 8). Som nevnt ovenfor, kan det være aktuelt å rapportere etter to skalaer i tillegg til den totale skalaen på Skoleporten, henholdsvis RSF og RTK/APM (se kap. 8.3.4).

Generelt tilfredsstiller også de tre prøvene for grunnkurs i videregående skole grunnleggende psykometriske krav. Det er likevel et viktig poeng at særlig prøvene for elever på allmennfaglig studieretning, 1MX og 1MY, har mange oppgaver med lav diskriminering. Flere av disse oppgavene er enten så enkle at "alle" får poeng eller så vanskelige at "ingen" får poeng. Vi vil anbefale å se nærmere på disse oppgavene fram mot eventuell utvikling av nye prøver. Særlig prøvene for 1M (yrkesfaglige studieretninger) og 1MY har falt altfor vanskelig ut. Gjennomsnittene ligger på kun henholdsvis 35 % og 31 % av fullt hus. Prøven for 1MX har falt noe lettere ut med et gjennomsnitt på 47 % av fullt hus, men også denne prøven er for vanskelig sett i forhold til hva som er ideelt. Særlig i 1M- og 1MY-prøvene er høye andeler blanke svar et stort

problem. Hvor mye som eventuelt kan tilskrives lav motivasjon i prøvesituasjonen, er det selvsagt umulig for oss å avgjøre. Selv om de psykometriske analysene av de foreliggende dataene isolert sett kunne gitt grunnlag for rapportering, gjør utbredt boikott (36 % for 1MX, 43 % for 1MY og 45 % for 1M) at vi ikke kan anbefale noen rapportering av resultater for grunnkurs.

1.6 Hva kan publiseres fra de nasjonale prøvene i 2005?

I denne rapporten tar vi ikke stilling til spørsmålet om hvorvidt skolevise resultater fra nasjonale prøver skal publiseres på Skoleporten eller ikke. Det oppfatter vi som et politisk spørsmål. Men *hvis* man velger slik publisering, vil vi i det følgende summere opp hvilke skåreverdier som tilfredsstiller kravene man må stille for slik publisering. Slike krav er ikke mindre viktig som grunnlag for pedagogisk bruk overfor elever. Selv om man av politiske grunner skulle velge å ikke publisere noen resultater på Skoleporten, er allikevel listen nedenfor en viktig oversikt over hvilke skåreverdier som tilfredsstiller nødvendige kvalitetskrav. Det er også alvorlig å gi tilbakemelding til enkeltelever som ikke inneholder pålitelig informasjon.

Detaljert argumentasjon for våre vurderinger finnes i kapitlene 5-8. Vi konkluderer med følgende:

Lesing

- 4. trinn: En samlet skåre for hele prøven kan under tvil publiseres. Vi er imidlertid bekymret for validiteten til prøven. Skåreverdier for de to delskalaene etter tekstsjanger tilfredsstiller ikke kravene til publisering (og derved heller ikke til pedagogisk bruk).
- 7. trinn: En samlet skåre for hele prøven kan publiseres. Vi er imidlertid også her noe bekymret for validiteten til prøven. Skåreverdier for de to delskalaene etter tekstsjanger tilfredsstiller heller ikke her kravene til publisering.
- 10. trinn: Man kan publisere en samleskåre for hele prøven, men de tre foreslåtte delskalaene tilfredsstiller ikke kvalitetskrav til rapportering.
- Grunnkurs: På grunn av høy andel boikott anbefaler vi ikke publisering av resultater fra denne prøven.

Skriving

- Vi anbefaler å ikke publisere noen resultater fra skriveprøvene. Til det er sensorreliabiliteten altfor lav.

Engelsk

- Vi anbefaler å ikke publisere noen resultater fra skriveprøvene i engelsk. Sensorreliabiliteten er for lav, og uenigheten går på mange skoler systematisk i favør av egne elever.
- Vi har fått begrenset adgang til oppgaver og resultater for leseprøvene i engelsk, som foregikk på PC. Basert på resultatene som prøvene genererer, som virker urimelig gode, er vi skeptiske til validiteten. Heller ikke reliabiliteten er høy nok.

Vi er derfor skeptiske til å anbefale eventuell publisering. På grunn av boikott er det uansett uaktuelt å publisere resultater for grunnkurs.

Matematikk

- 4. trinn: Man kan publisere kun én samlet skåre for hele prøven.
- 7. trinn: Det kan også her publiseres kun én samlet skåre for hele prøven.
- 10. trinn: Det kan publiseres kun én samlet skåre for hele prøven. Det kan imidlertid også være grunnlag for å publisere resultater for to delskalaer, *Representasjoner, symbolbruk og formalisme* og en sammenslåing av de to andre. Navnene på skalaene bør i så fall endres slik at de kommuniserer bedre hva skalaene faktisk måler.
- Grunnkurs: På grunn av høy andel boikott anbefaler vi ikke publisering av noen resultater fra disse prøvene.

1.7 Om strategi og ledelse

De nasjonale prøvene varierer sterkt når det gjelder utforming, vurderingskriterier og rapporteringskategorier. Faggruppene har tydeligvis hittil fått følge egne ideer, med lite forsøk på kritisk overprøving og styring. Vi tillater oss å etterlyse en sterkere felles strategi i tråd med de overordnede formålene med nasjonale prøver. Etter vårt syn kan nasjonale prøver bare fungere positivt etter intensjonene med en betydelig sterkere styring i retning av et bedre samsvar mellom mål og virkemidler. For å oppnå dette tenker vi her ikke minst på at det synes å være et behov for bedre testteoretisk innsikt blant alle involverte, og spesielt blant den sentrale ledelsen av de nasjonale prøvene i Utdanningsdirektoratet.

Årets nasjonale prøver bærer etter vår mening generelt et preg av altfor stor tro på ”profiler” for kompetanse. Slik prøvene er utformet, er det et stort problem at flere foreslåtte skalaer har for store målefeil, altså at reliabiliteten ikke er høy nok. Vi anbefaler at før framtidige prøver gjennomføres i full skala, bør det gjennom pilotering *dokumenteres* at de planlagte skalaene trolig vil fungere. Hvor *mange* profiler man kan måle med høy nok reliabilitet, kan i stor grad forutsies ut fra oppgavenes psykometriske egenskaper.

Vi etterlyser videre mer enhetlige skalaer for måling av kompetanse. I prøvene brukes både poeng, prosent riktig og nivåer, noe som gjør det vanskelig å forstå hva verdiene betyr. Uansett om slike resultater skal rapporteres på Skoleporten eller ikke, bør man sørge for å gjøre skåreverdiene mer meningsfulle og eventuelt bedre sammenliknbare.

Vi mener det er uheldig at faggruppene i så stor grad har fått lage sine egne systemer for vurdering og innrapportering. Her burde det vært mye sterkere styring. Spesielt vil vi kritisere bruk av regneark som gir grafikk og ”profiler” basert på prosent riktige svar. Antall riktige svar avhenger av oppgavenes vanskelighetsgrad og kan ikke i seg selv gi meningsfull informasjon om sterke og svake sider hos enkeltelever eller klasser. Her har man trolig kommet i skade for å gi informasjon til lærere, elever og foreldre som kan være direkte misvisende.

Det burde etter vår oppfatning vært gjort mer fra sentralt hold for å begrense omfanget av vurderingsarbeidet for lærerne. Dette er ikke minst viktig for å skape en positiv holdning til prøvene ute i skolen. Dette kan for eksempel gjøres ved å inkludere flere flervalgsoppgaver i framtidige prøver. Det kan også gjøres ved å legge opp til en enklere prosedyre for innrapportering, gjerne ved at selve innrapporteringen via regneark gjøres av en egen person på skolen som har god erfaring med dataregistrering.

Det synes også å være et problem for skolene å holde orden på all informasjonen. Informasjonen burde i mye større grad vært styrt av Utdanningsdirektoratet. At informasjonsstrømmen har vært et problem framgår også tydelig av MMI-undersøkelsen. En rektor ved en ungdomsskole sa dette i et intervju med oss (se vedlegg):

> "Uansett hva som skjer, er det viktigste fremover å få orden på informasjonsstrømmen til skolene, og i første rekke strømmen av brev og mailer til rektorene. Det er en formidabel flyt av informasjon gjennom brev og mail fra henholdsvis fagmiljøene og Utdanningsdirektoratet. Slik kan det ikke fortsette – for meg har det vært utmattende som skoleleder."

1.8 Sprikende formål

I de opprinnelige oppdragsbrevene til faggruppene er formålene ved de nasjonale prøvene gitt slik (vår nummerering):

> "Formålet med prøvene skal være:
> 1. å gi beslutningstakere på ulike nivå informasjon om tilstanden i utdanningssektoren og dermed gi grunnlag for iverksetting av nødvendige tiltak for sektoren
> 2. å gi informasjon til brukere av utdanning om kvaliteten i opplæringen på det enkelte lærested og dermed blant annet gi bedre grunnlag for å gjøre valg og stille krav om forbedringer
> 3. å gi informasjon til skoleeier, skoleledere og lærere som grunnlag for forbedrings- og utviklingsarbeid på det enkelte lærested
> 4. å gi informasjon til den enkelte elev/elevens foresatte som grunnlag for elevens læring og utvikling
> 5. å kunne registrere utviklingen over tid, både på systemnivå og individnivå"

Slik vi ser det, kan vi i hovedsak her skille mellom to hovedgrupper av formål. Punktene 1, 2, 3 og 5 peker tydelig på at prøvene skal gi pålitelig og relevant informasjon om kompetanse hos elever enkeltvis eller i grupper. Punkt 4 gir en like klar understreking av at det *diagnostiske* eller pedagogiske formålet er viktig. Vi vil understreke at vi har oppfattet kravspesifikasjonen til vår undersøkelse slik at den særlig går på validitet og reliabilitet for årets prøver, noe som for oss setter det første formålet i fokus. Ut fra dette handler denne rapporten særlig om hvor godt prøvene har fungert som pålitelig kartlegging av kompetanse.

Det er i mange sammenhenger blitt understreket at den pedagogiske bruken av prøvene skulle være et sentralt formål, kanskje det viktigste. Dette skulle tilsi at det burde vært lagt mye vekt på at prøvene har god diagnostisk informasjonsverdi, og at det er gitt god

veiledning om hvordan prøven kan brukes i det pedagogiske arbeidet. Imidlertid synes dette ikke å ha skjedd i særlig grad. På tross av at flere av prøvene utvilsomt har et diagnostisk potensial, synes det å mangle mye på at dette er kommunisert av faggruppene og utnyttet av lærerne. Den samme rektoren som er nevnt ovenfor, fortsatte slik i vårt intervju (se vedlegg):

> "På tross av informasjonsstrømmen, er det så vidt jeg har kunnet se, ingen informasjon om hvordan vi skal bruke resultatene. All informasjon dreier seg om hvordan prøvene har vært utviklet, forarbeidet, registrering av data, sikring av resultatene, men ingenting om hvordan vi skal ta disse resultatene inn i skolen for videre organisasjonslæring. For oss kom dessuten resultatene sent, slik at ikke alle 10. trinns elever fikk snakket med lærerne sine om profiler, før de sluttet skolen. Noen lærere har klart å snakke med foreldre og elever, men ikke alle."

Her peker rektoren etter vår mening på et nøkkelpunkt, og mye tyder på at dette kan være et typisk inntrykk fra skole-Norge. I MMI-undersøkelsen var det spørsmål til lærerne om i hvor stor grad prøvene ga informasjon om elevene som de ikke visste på forhånd. Ikke for noen av prøvene på noe klassetrinn er det mer enn 12% av lærerne som svarer "i meget stor grad" eller "i ganske stor grad" på dette. Så mange som mellom 40% og 60% av lærerne svarte "i ingen grad" på dette spørsmålet. Også i Gallup-undersøkelsen er de ulike faglærerne jevnt over enige i at prøvene i liten grad gir dem ny informasjon om elevenes kunnskap og ferdigheter. Og når det gjelder i hvor stor grad resultater fra prøvene *faktisk* er blitt brukt pedagogisk overfor elever og foreldre, er resultatene fra MMI-undersøkelsen rett og slett nedslående lesing, uten at vi her går i detalj.

Etter vår mening bør man være varsom med å overbetone de diagnostiske sidene ved de nasjonale prøvene. I så fall står man i fare for å nedprioritere valide og reliable kompetansemål, noe årets prøver synes å være noe preget av. En helt annen sak er at diagnostiske prøver finnes allerede, og det vil etter vår mening være en bedre idé å bygge videre på disse. Det kan lages flere slike prøver, og i flere fag, som kan legges på nettet til fri benyttelse på skolene.

Det synes som om det har vært en oppfatning at resultater i form av "profiler" av kompetanse, altså sett av delkompetanser, i seg selv skulle ha diagnostisk verdi. Vi er enig i at kjennskap til sterke og svake sider vil hjelpe elever i pedagogisk henseende. Men det er et påfallende trekk ved de foreslåtte delskalaene at de i så liten grad imøtekommer dette formålet. Vi har påvist at de aller fleste av de foreslåtte delskalaene verken er særlig reliable eller valide, og da er den pedagogiske verdien liten. Faktisk er vi bekymret for alle de detaljerte resultatene som allerede er rapportert inn på "nasjonaleprover.no", og som i mange tilfeller kanskje vil bli lagret for å følge elevens videre gang i skolen. Vår rektor på barneskolen sier det slik:

> "Nå er det kommunal enighet om at profilene fra 7. trinn skal overføres til ungdomsskolene, slik at lærerne kan bruke dem videre."

Vi frykter at disse resultatene senere kan gi opphav til ikke holdbare slutninger om elevenes framgang. Det er i det hele tatt et problem hvis resultater med dårlig validitet og reliabilitet skal følge elevene og brukes som grunnlag for det pedagogiske arbeidet.

Vi mener at de diagnostiske elementene som ligger i prøvene best ivaretas på to måter. Det er mye å lære av en gjennomgang av og diskusjon om *hvilke* konkrete feil og mangler som hefter ved en besvarelse. Dette vil fungere best ved en gjennomgang oppgave for oppgave. Resultater for delskalaer *kan* også gi god diagnostisk informasjon, men da må de ha høy reliabilitet og validitet, og videre må de kommunisere tydelig hva de representerer. Vi finner dessverre ingen delskalaer i årets prøver som tilfredsstiller dette. For framtidige nasjonale prøver vil vi anbefale at det heller legges vekt på å utarbeide veiledningsmateriale som kan hjelpe lærerne i å bruke selve elevbesvarelsene (oppgave for oppgave) i kombinasjon med skåre på prøven som helhet i det pedagogiske arbeidet.

1.9 Omfattende vurderingsprosess

MMI-undersøkelsen viser at lærerne har brukt betydelig tid på å vurdere elevbesvarelsene. Det varierer noe, men tiden brukt per elev har ligget mellom litt over 20 og litt under 40 minutter. Å vurdere skriving har tatt lengst tid, gjennomsnittlig 33 minutter for alle klassetrinnene. Med slike høye tall er det åpenbart at vurderingen representerer en stor ressursinnsats, og denne synes ikke å stå i forhold til det man får ut av informasjon.

To eksempler vil belyse dette. I matematikk er det brukt et omfattende kodesystem for registrering av type svar på hver eneste oppgave. Dette gjøres senere om til poeng. For en lærer er det ikke naturlig å bruke kodene for en detaljert gjennomgang med en elev (i den grad dette i det hele tatt skjer), men heller å ta utgangspunkt i selve besvarelsen. Rapportering foregår etter poeng, så ikke i noen av disse sammenhengene blir kodene faktisk relevante. For forskningsformål kan det være fruktbart å analysere resultater på kodenivå, men vi har hittil ikke sett noe slikt publisert fra fjorårets prøve. Og uansett kan man stille spørsmålstegn ved om dette formålet kan forsvare tidsbruken. Når dette er sagt, vil vi likevel peke på at det har vært en betydelig forenkling av vurderingsprosedyrene i matematikk i forhold til fjorårets prøver. Tidsbruken har da også gått betydelig ned, ifølge MMI-undersøkelsen.

Et annet eksempel gjelder lesing i 4. trinn, der det bare er flervalgsoppgaver. Slike resultater burde det være svært enkelt å registrere. Med et egnet regneark med omregning til poeng innlagt som formler, kunne registreringsarbeidet for flervalgsoppgavene skjedd på under 5 minutter, og dette kunne gjerne gjøres av andre enn læreren selv. Lærerne har brukt lang tid (gjennomsnitt 24 minutter ifølge MMI-undersøkelsen) på å vurdere hele prøven, noe som tyder på både at det er brukt en svært lite rasjonell prosedyre, og/eller at ordkjedeprøven må ha tatt mye tid å telle opp. Uansett representerer dette lite effektiv tidsbruk i forhold til informasjonen man får ut. En registreringsmåte som nevnt ovenfor vil også ha den store fordelen at det er lett for skolen å lage oversikter over svarfordeling for hvert spørsmål. Det er også lett å framstille enkeltelevers ”profil” oppgave for oppgave.

1.10 Avsluttende oppsummering og konklusjon

Vi har påvist at det er mange problemer med årets nasjonale prøver. Til sammen framstår disse problemene som en nokså grunnleggende systemsvikt. Ifølge MMI- og Gallup-

undersøkelsen hersker det stor usikkerhet og misnøye rundt om på skolene, og dette kan vi faktisk forstå.

Noen formuleringer av formålene ved prøvene synes å beskrive dem som at de *i hovedsak* skal ha diagnostiske formål. Men etter vår mening fungerer prøvene ikke særlig godt slik. I hvert fall er det usikkerhet ute i skolene, og lærerne synes særlig å savne hjelp til *hvordan* prøvene kan brukes i det pedagogiske arbeidet. Hvis det virkelig er meningen at prøvene skal være et slikt pedagogisk hjelpemiddel, må det etter vår mening vises tydeligere hvordan de skal kunne være det.

Vi har i våre analyser lagt *avgjørende* vekt på at de kompetansemålene som resultater rapporteres etter (enten det er på Skoleporten eller ikke), må være av høy kvalitet. Med dette mener vi særlig høy validitet og reliabilitet, altså at det er de relevante kompetansene som faktisk måles, og at disse måles med ikke for store målefeil. Vi har påvist betydelige svakheter ved de fleste av prøvene i så måte, og vi hevder at de fleste av de resultatmålene som er rapportert inn, ikke holder mål. Hovedproblemet er at ønsket om å ha *mange* slike resultatmål (”profiler”) har ført til at alle er blitt lite pålitelige. Det burde etter vår mening være innlysende at når elevene ved skriftlige eksamener i vårt land sitter i fem timer og får et resultat (etter svært omstendelige prosedyrer) i form av *ett* tall, så kan man ikke regne med at det skal være naturlig å få tre eller flere resultater av høy kvalitet fra en liknende prøve på en time eller to.

Etter vår mening er det etter årets prøver viktig for alle involverte å ha en gjennomgripende diskusjon av alle sider ved prøvene. I den forbindelse vil vi advare mot å bortforklare problemer med at det har tatt tid å få prøvene til å fungere. Det kom betydelig kritikk etter fjorårets prøver, og det er vanskelig å se at det for årets prøver har vært noen vesentlig forbedring. Samlet sett mener vi kvaliteten på årets prøver framstår som dårligere enn fjorårets, så det dreier seg om noen fundamentale grep som bør gjøres, og som ikke bare ”går seg til” med tiden til hjelp. Fjorårets evalueringsrapport sluttet med noen klare råd som vi her sammenfatter slik:

1. Det bør være mye sterkere styring og koordinering når det gjelder design, oppgaveformater, vurderingskriterier og rapporteringsskalaer.
2. Det anbefales en mer enhetlig utprøving av oppgaver med konkrete krav til enkeltoppgaver og prøven som helhet når det gjelder vanskelighetsgrad, diskriminering og reliabilitet. Utprøving bør gjennomføres med mange flere oppgaver enn det endelige antallet for at det som ikke fungerer bra, kan lukes bort.
3. Det er avgjørende at prøvene er påvist å tilfredsstille grunnleggende krav før de sendes ut til hele elevkullet.
4. Det må ikke legges opp til å rapportere kompetanse i form av ”profiler” uten at hver delkompetanse på forhånd kan påvises å ha god validitet og reliabilitet.
5. Det bør gis tydeligere beskjed til lærerne og skolene om hva som skal skje med vurdering og innsending av resultater, og ikke minst hvordan de kan og bør bruke resultater i pedagogisk sammenheng.
6. Vi anbefaler at det utarbeides et grundig faglig rammeverk for hvert av fagområdene. Et slikt rammeverk kan inneholde et rasjonale for prøvene på ulike klassetrinn, samt en oversikt over design, fordeling av oppgaver etter oppgavetype, samt hvilke rapporteringskategorier som er tilstrebet. Spesielt bør forholdet til de

grunnleggende ferdighetene (i motsetning til hele læreplanen) avklares for hver prøve.

7. Dersom man virkelig mener alvorlig å skulle måle utvikling i elevenes kompetanse over tid, både på individnivå og på nasjonalt nivå, er det nødvendig å gjøre en egen grundig planlegging av hvordan man skal få til dette.

Vi kan ikke se at det på noen av disse punktene har skjedd noen vesentlig bedring ved årets prøver sammenliknet med fjorårets. En *gradvis* endring av slike grunnleggende forhold er kanskje heller ikke mulig. Men i hovedsak er våre anbefalinger de samme denne gangen. Derfor foreslår vi at det snarest tas en grundig diskusjon om primære formål og virkemidler ved prøvene *før* man forlenger prøvene og kanskje også problemene videre til neste år. Helt uavhengig av eventuell publisering av resultatene bør kvaliteten på prøvene forbedres.

Vi anbefaler derfor at det ikke gjennomføres noen nasjonale prøver i året 2006, og at tiden brukes til en gjennomgripende utredning og debatt. Det er skapt mye skepsis og motstand i forhold til de nasjonale prøvene, og det er etter vår mening avgjørende at neste runde med prøver framstår med betydelig høyere og mer enhetlig kvalitet.

Etter vårt beste skjønn vil en videreføring av de nasjonale prøvene etter de retningslinjene som er brukt de to første årene, ikke være til beste for norsk skole.

2 Forutsetninger og kravspesifikasjoner

2.1 Forutsetninger

I forbindelse med at de nasjonale prøvene for første gang ble gjennomført i fjor, ble det gjennomført en utvalgsundersøkelse som studerte prøvenes kvalitet ut fra testteoretiske og pedagogiske kriterier (”Nasjonale prøver på prøve”). Det foreliggende arbeidet er en oppfølging av dette, idet det var et ønske om en vitenskapelig basert vurdering også av prøvene i 2005.

Emnene i denne rapporten følger nøye de kravspesifikasjonene for vurderingens del 1 som Utdanningsdirektoratet sendte ut (se kapittel 3). Vi har også tatt et viktig utgangspunkt i formålene for de nasjonale prøvene, slik de er beskrevet i prosjekt-beskrivelsene til faggruppene:

Formålet med prøvene skal være:

- å gi beslutningstakere på ulike nivå informasjon om tilstanden i utdanningssektoren og dermed gi grunnlag for iverksetting av nødvendige tiltak for sektoren
- å gi informasjon til brukere av utdanning om kvaliteten i opplæringen på det enkelte lærested og dermed blant annet gi bedre grunnlag for å gjøre valg og stille krav om forbedringer
- å gi informasjon til skoleeier, skoleledere og lærere som grunnlag for forbedrings- og utviklingsarbeid på det enkelte lærested
- å gi informasjon til den enkelte elev/elevens foresatte som grunnlag for elevens læring og utvikling
- å kunne registrere utviklingen over tid, både på systemnivå og individnivå

Vi vil understreke at vi har oppfattet kravspesifikasjonen til undersøkelsen slik at undersøkelsen særlig går på validitet og reliabilitet for årets prøver. Vi oppfatter det ut fra dette at et hovedmål for prøvene er å måle elevenes faglige kompetanse på en god måte. Det innebærer at det diagnostiske elementet ved prøvene, altså hva elevene kan lære av den faglige tilbakemeldingen, uansett hvor viktig dette er, ikke må være til hinder for dette, som vi i vår sammenheng oppfatter som det primære målet.

De nylig framlagte rapportene fra MMI (heretter referert til som ”MMI-rapporten” eller ”MMI-undersøkelsen”) og TNS-Gallup (”Gallup-undersøkelsen”) diskuterer funn fra en spørreundersøkelse ute i skolene. Flere steder bruker vi funn i disse rapportene som bakgrunn for diskusjoner av våre egne resultater.

Innholdet i denne rapporten er skrevet av en gruppe forskere ved ILS, Universitetet i Oslo. Det må nevnes at kolleger ved ILS er engasjert i faggrupper for nasjonale prøver. Dette gjelder både i matematikk og i lesing. Vi har dermed et erkjent habilitetsproblem, noe som vi selvsagt har diskutert med oppdragsgiver. Etter beste evne har vi forholdt oss til disse kollegene som til de andre faggruppene, kommunisert om innhenting av data og andre forhold ved prøvene, men selvsagt ikke diskutert våre vurderinger med dem. Vi har

for øvrig hatt god kontakt med faggruppene, og vi takker dem for hjelp og velvillighet for å skaffe fram data fra elevbesvarelsene og vurderingene av disse. Vi har videre diskutert våre vurderinger med Anne-Berit Kavli og Annette Qvam fra Utdanningsdirektoratet som representanter for oppdragsgiverne. Alle vurderinger og anbefalinger står forfatterne alene ansvarlig for.

2.2 Kravspesifikasjoner

I det følgende gjengir vi kravspesifikasjonen for undersøkelsen, slik den ble formulert av Utdanningsdirektoratet i anbudsutlysningen:

> **Fra prosjektbeskrivelsen:**
> Beskrivelsen er todelt. Del 1: Det skal gjennomføres årlige evalueringer av prøvenes kvalitet når det gjelder reliabilitet og validitet, og det skal utvikles analyser av resultatene fra elevbesvarelser som inngår i utvalget.
>
> **Fra Konkurransegrunnlaget**:
> Evalueringen skal:
> - Belyse hvilke metoder som er brukt i utviklingen av prøvene
> - Foreta grunnleggende item-analyser
> - Vurdere prøvenes reliabilitet (indre konsistens)
> - Vurdere sensorreliabilitet for åpne oppgaver
> - Vurdere prøvenes validitet
> - Vurdere hvordan prøvene kan utvikles slik at de kan fungere som grunnlag for sammenlikninger med resultater fra år til år.
> - Foreta analyser av elevenes resultater i de ulike prøvene
> - Foreta vurderinger og analyser som kan gi grunnlag for blant annet utvikling av enkle kompetanseprofiler på sikt. (Her må det være tett dialog med fagmiljøene som lager prøvene.)

Denne rapporten er en beskrivelse av vårt arbeid når det gjelder vår evaluering av de nasjonale prøvene i 2005 samt noen synspunkter på hvordan arbeidet med prøvene kan forbedres for de kommende årene.

3 Utvalg og innhenting av data

Som i 2004 var det allerede lagt opp en strategi for å "sjekke" hvor pålitelig rettingen av åpne oppgaver foregikk. Et antall uttrukne skoler hadde fått beskjed om å sende inn kopier av et visst antall (opp til 20) tilfeldig uttrukne elevbesvarelser til ekstern vurdering. Denne uttrekkingen skulle ideelt skje uten at lærerne visste *hvilke* elever det dreide seg om, for at de ikke skulle være påvirket av dette i sin egen vurdering. De eksterne sensorene sendte så sine data tilbake til skolen for at skolen skulle få en vurdering av kvaliteten på egen retting. Det var imidlertid ikke lagt opp til at denne informasjonen skulle nå Utdanningsdirektoratet eller faggruppene direkte. Dette ble det informert om til skolene og de eksterne vurdererne i ettertid, med frister for innsending.

Etter å ha brukt elevbesvarelsene og vurderingen av disse i pedagogisk øyemed, skulle de uttrukne skolene sende følgende til den aktuelle faggruppa for hver prøve (litt forskjellig fra fag til fag):

- De uttrukne elevbesvarelsene (i noen fag)
- Liste (eventuelt datafil) med skolens egen vurdering av disse
- Liste (datafil) med ekstern vurdering av disse
- Liste (datafil) for ekstern vurdering nr. 2 (i noen fag)

Faggruppene skulle deretter sende alt dette materialet til oss. Det viste seg at det tok lang tid før vi mottok alt materialet. På tross av pålegg om å påskynde prosessen viste det seg at tempoet av innsendingen ble for svak. På et tidspunkt ble vi derfor nødt til å sette strek og nøye oss med de skolene vi hadde materiale fra. Vi vurderer dette slik at det her foreligger en mulighet for at vårt endelige utvalg kan være litt skjevt. Men det er liten grunn til å tro at dette utgjør en betydelig effekt. Vi vil kommentere mer om dette for de enkelte prøvene senere.

Det aktuelle antall skoler og elever i våre datafiler er vist i tabell 3.1. For engelsk lesing har vi ikke hatt tilgang på elevdata på samme måten, da denne prøven foregikk på PC, og datafilene ikke enkelt kunne overføres.

Strategien gikk som nevnt ut på å hente inn data fra de samme elevene både fra de eksterne og fra skolene, slik at vi kunne sammenlikne vurderingene gjort av de to sensorene. Imidlertid var det ikke alle skolene som sendte inn slik de var bedt om, så i tabell 3.1 har vi i parentes ført opp antall der vi har data fra to uavhengige sensorer som vi kunne sammenlikne, enten en intern og en ekstern, eller to eksterne sensorer.

Tabell 3.1: Utvalgsstørrelse for hver prøve

Trinn	Fag	Totalt antall elever	Antall elever med minst to uavhengige vurderinger*
Grunnkurs	Matematikk, 1M	164	118
	Matematikk, 1MY	148	89
	Matematikk, 1MX	108	63
	Lesing	528	525
	Skriving	-	-
	Engelsk lesing	900	0
	Engelsk skriving	616	383/450 *
10. trinn	Matematikk	290	154
	Lesing	514	467
	Skriving	512	0/512 *
	Engelsk lesing	900	0
	Engelsk skriving	899	201/512 *
7.trinn	Matematikk	537	537
	Lesing	450	438
	Skriving	603	0/603 *
	Engelsk lesing	600	0
	Engelsk skriving	450	383/450 *
4. trinn	Matematikk	374	198
	Lesing	253	-
	Skriving	158	0/158*
	Engelsk lesing	600	0

* *Det første tallet gjelder intern mot ekstern vurdering, det andre tallet gjelder to eksterne vurderinger*

Som det framgår av tabell 3.1, er det for flere av prøvene et lavt antall besvarelser vi bygger våre analyser på. Dette var en helt nødvendig avgjørelse, siden vi ikke hadde noen muligheter for å utsette arbeidet til alle de aktuelle skolene hadde sendt inn. Også det faktum at skolene skulle sende til faggruppene, betydde en forsinkelse. På slutten av juni måned valgte vi å bruke de dataene vi da hadde for å kunne holde tidsfristen for denne rapporten. De skolene som skulle sende besvarelser til ekstern vurdering, var trukket tilfeldig. Og vi har altså et strengt tatt ikke-tilfeldig utvalg av disse skolene med i vår undersøkelse. Vi vurderer det slik at det derved er en mulighet for noen systematiske feil på grunn av at det ikke er et tilfeldig utvalg. Det imidlertid all grunn til å tro at når det gjelder kvaliteten av vurderingsarbeidet, er det noe høyere i utvalget enn for alle de aktuelle skolene. Vi tror at skoler som følger de oppgitte tidsfristene, også gjennomgående har vurdert elevbesvarelsene på en samvittighetsfull måte. Vi tror derfor at våre resultater om samsvar i vurderingene tenderer svakt til å vise noe bedre samsvar enn det som er gjennomgående rundt i landet. Når det derimot gjelder prestasjonsnivået, tror vi d ikke det er noen tydelig tilsvarende effekt, men utvalgsfeilene er selvsagt økende med lavere antall elevbesvarelser.

Vi har for øvrig valgt i denne rapporten å ikke gi detaljerte feilmarginer for prestasjoner eller samsvar i vurderingen, det ville bli for mange detaljer. I og med at vi ikke har et

tilfeldig trukket utvalg av enkeltelever, vil det også være vanskelig å estimere slike feilmarginer.

Det var et betydelig, og i noen tilfeller foruroligende, frafall av enkeltelever i forhold til det antallet som var trukket ut på skolene (maksimalt 20). Denne informasjonen er summert opp i tabell 3.2 og er hentet fra Utdanningsdirektoratets landsoversikt for innrapportering Vi har ikke helt presise data om hva som ligger bak "innvilget fritak" og boikott, så det er vanskelig å konkludere for sterkt ut fra dette. Vi kommer tilbake til boikottproblemet senere i denne rapporten.

Et spesielt problem gjelder besvarelser fra grunnkurs på videregående skole. Det har ikke vært et system for å angi hvilken studieretning elevene går på, noe som gjør at alle sammenlikninger skoler imellom blir nokså lite meningsfulle. Som det framgår av tabell 3.2, har det vært et betydelig innslag av boikott på dette trinnet. Blant annet av disse grunnene har vi i vår rapport lagt lite vekt på resultatene for dette trinnet. For engelsk lesing er informasjonen om deltakelse mangelfull på hjemmesiden til nasjonale prøver.

Tabell 3.2: Oversikt over deltakelse og frafall av enkeltelever (fra nasjonaleprover.no)

	Fag	**% som deltok på prøven**	**% ikke gjennomført**	**% som har fått innvilget fritak / syk**
Grunnkurs	Matematikk 1M (YF)	45	45	4
	Matematikk 1MY	45	43	1
	Matematikk 1MX	58	36	1
	Lesing	52	40	3
	Skriving	4	6*	*
	Engelsk skriving	45	45	4
10. trinn	Matematikk	81	14	3
	Lesing	84	11	3
	Skriving	74	20	3
	Engelsk skriving	85	10	4
7.trinn	Matematikk	95	2	4
	Lesing	95	1	4
	Skriving	93	2	4
	Engelsk skriving	94	1	4
4. trinn	Matematikk	96	1	3
	Lesing	96	1	3
	Skriving	94	2	4

* Denne prøven var frivillig

I tillegg til å skaffe de kvantitative dataene har vi selvsagt snakket med mange elever, lærere og rektorer om hvilke erfaringer de har hatt med prøvene. I tillegg ba vi to rektorer, den ene gjennomgående positiv og den andre nokså kritisk, om å skrive et notat til oss der de kommenterte det som for dem var viktigst. Disse notatene er gjengitt i vedlegg, og sitater derfra er brukt noen steder i vår tekst.

4 Strategi og metoder for undersøkelsen

I dette kapitlet vil vi gi en oversikt og beskrivelse av de analysene som er gjennomført, og bakgrunnen for disse. På flere punkter er det lagt inn en forklaring på de begrepene og metodene som er brukt. Disse skiller seg ut typografisk, så lesere kan gå inn i dette etter behov uavhengig av den logiske framdriften i teksten for øvrig.

4.1 Hvilke metoder ble brukt i utviklingen av prøvene?

Svaret på dette vil vi søke delvis ved å studere faggruppenes egne rapporter om dette. Dernest har vi gjennom item-analysene (se nedenfor) påvist hvordan hver oppgave har "fungert" testteoretisk sett, noe som har gitt indikasjoner på hvordan krav til god diskriminering og svarfordeling er blitt ivaretatt gjennom piloteringen.

4.2 Grunnleggende item-analyser

Basert på de innkomne data har vi gjennomført en tradisjonell item-analyse oppgave for oppgave. Dette innebærer en analyse av:

- Prosentfordelingen for hvert svaralternativ (flervalgsalternativ eller type svar).
- Gjennomsnittlig skåre for hele prøven for hvert svar- eller poengalternativ. ("Dyktighet" til de elevene som har gitt et bestemt svar). Spesielt har vi sett på om noen svaralternativer tydelig ikke fungerer etter forutsetningene, for eksempel noen "gale" svar som særlig gis av flinke elever.
- Oppgavens diskriminering, her kalt D (Pearson korrelasjon mellom skåre på oppgaven og på prøven totalt sett).

Om korrelasjon: Ofte er det gunstig å kunne angi et tall som et uttrykk for i hvor stor grad to variabler varierer sammen. Vi snakker da om graden av samvariasjon eller korrelasjon. En korrelasjonskoeffisient er et mål på i hvor stor grad de to variablene varierer "i takt", altså i hvor stor grad den ene variabelen har en høy verdi samtidig med (for eksempel for samme elev) når den andre har det, og omvendt. Den vanligste korrelasjonskoeffisienten er den såkalte Pearsons korrelasjonskoeffisient (ofte symbolisert med r), og den måler i hvor stor grad de målte dataene faller langs en rett linje når de avtegnes i et koordinatsystem.

I vårt tilfelle har den ene variabelen ofte bare to verdier (for eksempel riktig-galt for en oppgave), og da forteller r i hvor stor grad det er de som har svart riktig,, som har høyest verdi på testen som helhet. Korrelasjonskoeffisienter kan ha verdier fra –1 (perfekt negativ korrelasjon) via 0 (ingen korrelasjon) til 1 (perfekt positiv korrelasjon). En oppgaves diskriminering, D, bør som en tommelfinger-regel være større enn 0,30 for at den skal bidra positivt til å øke en prøves reliabilitet. Vi har derfor i våre analyser brukt dette kriteriet som et kvalitetskrav for oppgavene.

4.3 Prøvens reliabilitet (indre konsistens)

Vi har beregnet prøvens indre konsistens reliabilitet (Cronbachs alfa) og dermed svart på om oppgavene fungerer godt nok sammen til at tilfeldighetene knyttet til oppgaveutvalget ikke er for stort. Videre har vi pekt på problematiske oppgaver som har bidratt til å trekke

reliabiliteten ned. For foreslåtte delkompetanser er reliabilitetsanalyser gjennomført for hver rapporteringskategori for seg i tillegg til for hele prøven samlet. Vi har lagt stor vekt på disse spørsmålene, siden enhver rimelig og presis kvalitetsmåling er helt avhengig av høy reliabilitet, og dette gjelder enten resultatene skal brukes til å sammenlikne skoler eller til å informere elever om deres spesifikke kompetanser.

Om reliabilitet På de nasjonale prøvene beregner vi skåreverdier for prestasjoner ved hjelp av en rekke oppgaver (eller delferdigheter) for derved å oppnå en tilstrekkelig høy reliabilitet. Hadde vi bare brukt noen få oppgaver, ville det vært altfor store tilfeldigheter når det gjaldt hvor godt oppgavene passet den enkelte elev eller elevgruppe. Det er et ufravikelig krav at enkeltoppgavene må støtte opp om hverandre, at de viser en rimelig høy indre konsistens. Jo lavere konsistens (eller om vi vil, jo mer forskjellig enkeltoppgavene er), jo flere oppgaver må vi ta med for å få tilstrekkelig høy reliabilitet.

Men hva er "god nok" reliabilitet? For å svare på det vil vi først gi en kvalitativ beskrivelse av en reliabilitetskoeffisient i form av det som kalles Cronbachs alfa. Vi deler først testen i to deler og lager en samlevariabel for skåre for hver del. Så beregner vi korrelasjonskoeffisienten mellom de to delene. Denne inndelingen i to deler kan vi gjøre på mange måter, og vi får derfor mange slike koeffisienter. Gjennomsnittet av alle disse (korrigert for at halvdelene er kortere enn hele testen) gir oss alfa. Vi kan også si at alfa forteller oss hvor mye av samlevariabelen som virkelig representerer det vi måler, og hvor mye som simpelthen er tilfeldigheter (i valg av oppgaver). En høy alfa betyr at resultatet for enkeltelever i liten grad bestemmes av nøyaktig **hvilke** oppgaver som er med i samlevariabelen, så da ville resultatene blitt omtrent det samme om vi byttet ut en oppgave med en annen. En verdi på 0,70 for alfa regnes i mange sammenhenger som en nedre grense for en samlevariabel som skal brukes til å sammenlikne store grupper av elever. En slik verdi forteller oss at 70% av variansen (som representerer den informasjonen samlevariabelen gir oss) er "sann varians", mens resten (30%) er "feilvarians". Begrepet "feilvarians" indikerer ikke at noe er gjort feil, men at det representerer noe annet enn det som er felles for variablene som inngår. Populært sagt: Vi har 70% sann varians og 30% tilfeldigheter.

I tilfeller der man tilstreber å sammenlikne enkeltpersoner og små grupper av personer med høy presisjon, og hvor resultatene skal gi viktig informasjon til disse, ligger alfa vanligvis mye høyere enn 0,70. For prøver som får en viss betydning for enkeltpersoner, kan vi ofte oppfatte 0,85 som et naturlig nedre grense for alfa. Vi har brukt dette som et kvalitetskriterium i våre undersøkelser.

Betydningen av høy reliabilitet Reliabilitetskoeffisienten for en prøve, ofte betegnet som r_{xx}, er altså definert som korrelasjonen mellom to parallelle prøver eller to versjoner av den "samme" prøven (se ovenfor). Mangel på perfekt reliabilitet medfører at enhver måling har en viss målefeil, og denne målefeilen (SE, standardfeilen eller "standard error" til målingen) kan beregnes ut fra reliabilitetskoeffisienten. Det er en enkel sammenheng:

$$SE_{måling} = S\,(1- r_{xx})^{1/2}$$

Eller i ord: Standardfeilen til enkeltmålingen ("Standard Error of the measurement") er lik standardavviket for fordelingen multiplisert med kvadratrota av (1 – reliabilitets-koeffisienten). Denne målefeilen kan angis som feilmarginer, idet det er 95 % sannsynlighet for at en målt verdi "egentlig" svarer til en "sann" verdi innenfor intervallet målt verdi + eller – 2 standardfeil. Et

eksempel vil vise hvordan dette fungerer: Med en reliabilitet på 0,85 vil faktoren $(1- r_{xx})^{1/2}$ utgjøre 0,39, og følgelig vil feilmarginen (2 X SE) være 0,78 eller 78% av et standardavvik. Når vi i denne undersøkelsen vurderer en reliabilitet på 0,85 som en nedre grense for hva som er forsvarlig, henger det sammen med at med lavere reliabilitet enn dette vil feilmarginene på målingene være så store at de målte verdiene vil inneholde for stor grad av usikkerhet.

Tabell 4.1 viser hvordan dette forholder seg ved andre verdier av reliabilitet. Vi ser her at ved en reliabilitet på for eksempel 0,65 vil feilmarginen utgjøre omtrent 120 % av et standardavvik. Dette kan vi konkretisere ved et eksempel: En prøve tenkes målt med en slik reliabilitet langs en vanlig karakterskala (1-6). Fordelingen på karakterer på prøven kan typisk ha et standardavvik på 1,4 målt i karakternivåer. Hver elevs måleresultat har derfor en feilmargin på ca 120 % av 1,4 og det utgjør nesten 1,7 målt i nivåer. Hvis Per har fått karakteren 4, burde vi egentlig si at Pers kompetanse er målt til en verdi som med 95 % sannsynlighet "egentlig" ligger mellom 2,3 og 5,7, eller med andre ord ligger på nivå enten 2, 3, 4, 5 eller 6 (2 og 6 riktignok med liten sannsynlighet). Og det er jo en svært upresis og nokså lite verdifull informasjon.

På den annen side er selvsagt ikke høy reliabilitet isolert sett et tilstrekkelig kriterium på kvalitet. Hvis en prøve har veldig høy reliabilitet, for eksempel en alfa over 0,95, kan dette være et tegn på at de enkelte oppgavene korrelerer så høyt at de rett og slett er for like og i for stor grad måler det samme. I slike tilfeller er det ekstra viktig å vurdere kritisk om oppgavene til sammen virkelig dekker hele det fagområdet de gir seg ut for å dekke, eller om vi vil: om validiteten er god nok.

Tabell 4.1: Hvordan feilmarginer avhenger av reliabiliteten

Reliabilitets-koeffisient	***Feilmargin uttrykt i prosent av ett standardavvik***
(Ingen prøve)	*200 %*
0,6	*126 %*
0,65	*118 %*
0,7	*110 %*
0,75	*100 %*
0,8	*89 %*
0,85	*77 %*
0,9	*63 %*
0,95	*22 %*

4.4 Sensorreliabilitet for åpne oppgaver

For åpne oppgaver har vi foretatt en analyse av om rettingen er foregått på tilstrekkelig lik måte fra person til person slik at vi kan stole på de angitte poengene. For hver enkelt åpen oppgave har vi beregnet overensstemmelsen mellom to uavhengige vurderinger av samme besvarelser. Vi har bare i liten grad kunnet gå inn på i hvor stor grad hver person har rettet konsistent, men har konsentrert oss om å måle inter-sensor reliabiliteten. Vi har i denne analysen ikke betraktet de eksterne vurderingene som mer "riktig" enn lærernes. Men likevel er det en viktig forskjell når det gjelder utgangspunktet for vurderingen. Lærerne kjenner sine egne elever, og de har også en viss egeninteresse av at resultatene blir gode.

Ved å sammenholde de to uavhengige vurderingene av de samme elevene har vi besvart disse spørsmålene:

- Hvor stor overensstemmelse er det mellom de to vurderingene oppgave for oppgave? Vurdert sammen med alfa (se kap 4.3), er dette tilfredsstillende for å forsvare å publisere skåreverdiene?
- Er det noen spesielle oppgaver der overensstemmelsen er særlig svak?
- Hvordan varierer vurderingen fra skole til skole? Er det tydelige tendenser til konsistent for streng eller for mild retting generelt eller på noen skoler?

Om sensorreliabilitet: Et enkelt mål for overensstemmelse mellom sensorer er hvor mange prosent av besvarelsene som er vurdert likt. I våre datatabeller (kolonne merket "R") har vi gitt resultatene og "flagget" dårlig overensstemmelse ut fra kriteriene < 85 % og også < 75 %.

Imidlertid er den såkalte Cohen's Kappa ("coefficient of agreement") i mange sammenhenger et bedre mål for dette, så vi har også inkludert dette (kolonne merket "K" i tabellene). Verdien 1 betyr perfekt overensstemmelse, og 0 betyr like god overensstemmelse som det som vil skje bare ved en tilfeldighet. En Kappa over 0,8 regnes som god overensstemmelse. Kappa har den fordelen over prosent overensstemmelse at den er korrigert for tilfeldigheter (den er "chance corrected") og dette er meget viktig når man vurderer overensstemmelse på oppgaver som har få svaralternativer. For eksempel kan man se at hvis en oppgave har enten riktig eller galt som svar, ville fullstendig tilfeldig vurdering gi 50% samsvar, men en Kappa på 0. Et viktig poeng er også at Kappa bare kan beregnes for de variablene der nøyaktig de samme kategoriene i praksis er brukt av begge sensorene. Dette er bakgrunnen for at Kappa noen få ganger ikke kan angis.

Siden ekspertene i flere tilfeller har sendt sine vurderinger tilbake til skolene, er det en mulighet for at lærerne kan ha endret sin vurdering. Dette kan være gjort fordi de har foretatt en revidert vurdering og blitt "overbevist" av eksperten. Vi har imidlertid grunn for å tro at dette har foregått i mindre grad i 2005. Fjorårets rapport understrekte at hvis skoler skal sammenliknes, så er det viktig at skolene får en tydeligere beskjed om hvilken strategi som skal følges når det gjelder dette. For inneværende år, i motsetning til i 2004, fant vi for eksempel bare en skole der hver eneste besvarelse i engelsk skriving var vurdert identisk med den eksterne vurderingen.

Når det gjelder prøver med mange oppgaver, og noen eller alle disse oppgavene krever skjønnsmessig vurdering, er det naturlig å studere sensorreliabiliteten for hver oppgave for seg. Hvis det da viser seg at det bare er noen få oppgaver med dårlig samsvar mellom sensorene, vil dette bare i liten grad gi svekket reliabilitet for prøven som helhet. Og det vil være oppgavenes indre konsistens reliabilitet (Cronbachs alfa) som er avgjørende for påliteligheten av resultatene. For prøver som krever stor grad av fri skriving, forholder det seg annerledes. For slike prøver består prøven kanskje bare av noen svært få oppgaver, eller til og med bare én. Da blir sensorreliabiliteten helt avgjørende, men det har vist seg svært vanskelig å oppnå god enighet når vurderingen må baseres så mye på skjønn. All erfaring har vist at det ikke er tilstrekkelig med et sett av tydelige kriterier for vurderingen. Pålitelige resultater kan bare oppnås ved at

- man gjennom bred erfaring har kommet fram til et godt tolkningsfellesskap blant alle sensorene, og at
- samme besvarelser vurderes av minst to uavhengige sensorer, som ideelt sett deretter møtes og diskuterer uoverensstemmelsene.

Angående dette vil vi henvise til foredrag av Kjell Lars Berge ved Nasjonal konferanse om Nasjonale prøver i februar 2005.

For de nasjonale skriveprøvene i engelsk og norsk har vi vurderinger fra to eksterne sensorer. Spesielt for engelsk skriving er sensorreliabiliteten et avgjørende punkt, men her er det to forhold som gjør seg gjeldende. For det første er det et spørsmål om sensorene er enige om hvor gode besvarelsene er *i forhold til hverandre,* og for det andre om sensorene bruker skalaen likt, altså at gjennomsnitt og spredning for de felles besvarelsene er omtrent like. Siden det er aktuelt å rapportere skoleresultater på nettet, er det særlig viktig for oss å se systematisk på forskjeller mellom "strenge" og "milde" sensorer.

Innrapporteringen fra skoler og eksterne sensorer har ikke gitt oss tilstrekkelige opplysninger om hvem som har vært sensor, til at vi har kunnet studere samspillet mellom de to nevnte forholdene. Derfor har vi for engelsk skriving gjennomført et eksperiment med fire innleide sensorer som alle har vurdert de samme (omtrent 50) elevbesvarelser. Dette er beskrevet i kapittel 7.1.7.

4.5 Validitet: Om prøven og underkategoriene måler det den gir seg ut for å måle

Dette er et vanskelig spørsmål å svare på, så lenge prøvens formål er så mangfoldig. Vi har konsentrert oss om å diskutere i hvilken grad prøvene virkelig måler den type fagkompetanse som de er ønsket å måle, gir seg ut for å måle, eller som den blir oppfattet å måle. Vi vil presisere at det ikke finnes noen objektive svar på slike spørsmål. Vi har gitt vår vurdering av dette i forhold til hver foreslåtte rapporteringskategori. For å gjøre dette har vi studert oppgavene selv i detalj, prøvd å analysere nøyaktig hva slags kompetanse oppgaven krever for et riktig svar. I noen tilfeller går vi svært grundig inn på dette spørsmålet.

En viktig del av validitetsspørsmålet kan bare besvares i lys av læreplaner og fagdidaktisk innsikt i hva kompetanse innen faget "egentlig" består av. Nettopp dette er faggruppenes sterke side, og de har brukt sin innsikt og beste skjønn ved utvikling av prøvene. Vi har bare i begrenset grad gått inn på en analyse av overensstemmelse mellom prøvene og innholdet i læreplanene. Spesielt når det gjelder lesing, er det tvilsomt om vi i det hele tatt kan snakke om noen "læreplan", siden dette så definitivt handler om en ferdighet på tvers av skolens fag. Vi har imidlertid referert lærernes syn på slike spørsmål, slik det er kommet til uttrykk i MMI-undersøkelsen.

Videre har vi noen steder vurdert om beskrivelsen og navnet eller "merkelappen" til kompetansen kommuniserer til lærere og skolepolitikere det aktuelle innholdet slik at innholdet blir forstått.

Det er åpenbart at skal det være noen vits i å rapportere flere enn én kompetanse for hver prøve, så må de foreslåtte kompetansene være så forskjellig fra hverandre at de virkelig gir separat informasjon. Dette vurderer vi best ved å beregne korrelasjonskoeffisienter mellom kompetansene og sammenholde dette med reliabiliteten for hver av dem.

> **Latente korrelasjoner**: Det som er viktig når vi skal avgjøre om to variabler er "forskjellig nok" i denne forstand, er som nevnt ovenfor å studere korrelasjonen mellom de to variablene i lys av hvor "nøyaktig" hver av variablene er målt, og dette siste uttrykkes ved reliabiliteten. Dersom korrelasjonen er omtrent like stor som reliabiliteten til hver av dem, må vi bare konkludere at de to variablene korrelerer så høyt det er mulig teknisk sett, og vi sier da at den "latente" korrelasjonen er tilnærmet perfekt. I så fall er det ikke empirisk grunnlag for å hevde at de to variablene faktisk måler forskjellige kompetanser.
>
> Den latente korrelasjonen mellom to samlevariable a og b er gitt ved den vanlige korrelasjonen dividert på kvadratrota av produktet av alfa for de to variablene:
>
> $r_{a,b,\,latent} = r_{a,b}\,(alfa_a \times alfa_b)^{-1/2}$

Vi har i noen tilfeller også vurdert om prøvene har en faktorstruktur som samsvarer med de skalaer som ble brukt. En eksplorerende faktoranalyse har gitt oss en indikasjon på i hvor stor grad dette er oppfylt.

> **Om faktoranalyse**: Dette er en matematisk metode til å finne den underliggende "strukturen" i et omfattende datamateriale. Ut fra elevenes svar på hver oppgave, kan man med en faktoranalyse be om en "naturlig" måte å gruppere oppgavene på ut fra hvilke oppgaver som besvares mest likt. Ideelt sett bør det være samsvar mellom de delkompetansene som er foreslått og de "faktorene" som dataprogrammet foreslår ut fra hvordan elevene faktisk svarte. Dette kan enten gjøres eksplorerende ved å la programmet velge ut grupper, eller konfirmerende ved å prøve ut hvor godt den på forhånd utvalgte inndelingen stemmer overens med dataene

I diskusjoner om eksamen og prøver har vi ofte hørt uttalelser i retning av at det ikke er så farlig om reliabiliteten er lav, fordi "det er validiteten som teller". I et testteoretisk perspektiv kan ikke et slikt standpunkt forsvares. En god måling krever både høy validitet og god reliabilitet, det ene kan ikke på noen måte erstatte det andre. Uten god reliabilitet er høy validitet for en prøve ikke mulig. På den ene siden vil vi fullt ut støtte at i en viss forstand er høy validitet det viktigste. En prøve med et bestemt formål må selvsagt ha et faginnhold som dekker dette formålet. For de nasjonale prøvene er det en selvfølge at prøvene først og fremst dekker det faglige innholdet og den fagdidaktiske tilnærmingen i læreplanen på en god måte. Dette hensynet er vektlagt meget sterkt ved at faggruppene er knyttet til landets fremste fagdidaktiske miljøer. Faggruppene har utformet oppgaver som viser god sammenheng med læreplanene, og de har også demonstrert god innsikt i

fagdidaktiske utviklingstrekk nasjonalt og internasjonalt. Vi har i det hele tatt i vårt land svært gode tradisjoner gjennom eksamen i å lage nasjonale prøver med høy validitet.

Et viktig punkt gjenstår å kommentere. Høy validitet er ikke bare et spørsmål om oppgavene i seg selv er egnet. Like mye handler det om vurderingskriteriene, siden disse er en integrert del av oppgavene. Mye av validitetsdiskusjonene i denne rapporten handler om i hvor stor grad vurderingskriteriene og skalaene som de måles etter, fungerer etter forutsetningen, eller mer presis: på en egnet måte.

Imidlertid hjelper ikke de beste intensjoner om høy validitet hvis ikke den mer tekniske siden av prøven også er av høy kvalitet. Med lav reliabilitet blir resultatene mer eller mindre tilfeldige, uansett hvor godt oppgavene dekker læreplanen. Vi tillater oss en analogi for å belyse dette. En vellykket jakt er avhengig at jegeren har fellingstillatelse, at han kjenner sitt byttedyr og klarer å komme på skuddhold og vet hvor han skal sikte. Men alt dette er verdiløst hvis han ikke også er en god skytter. Det spiller ingen rolle hva han sikter på hvis skuddene sendes ut i "hytt og vær". For å holde oss til analogien: Det spørs altså om jegerne har gode nok skyteferdigheter til å felle det byttet de har klart å komme på skuddhold av. Vårt utgangspunkt i denne rapporten er at det er viktig å undersøke om faggruppenes mål for prøven *faktisk* er gjenspeilet i de målte resultatene.

4.6 Absolutt mål for sammenlikninger?

Kan de foreslåtte kompetansenivåene fungere som absolutte (kriterierelaterte) skalaer for kompetanse?

Kan kompetansenivåene fungere som grunnlag for sammenligninger med resultater et annet år?

Hvordan kan prøvene utvikles så sammenlikninger fra år til år skal være mulig?

Disse spørsmålene henger sammen og kommenteres her samlet. Dersom kompetanser er målt langs en skala med klart beskrevne nivåer, kan vi si at vi har å gjøre med en **kriterierelatert vurdering**. Med et slikt instrument kan vi ut fra resultatene beskrive i absolutt forstand hvor gode elevene er. Og man kan da tenke seg at man året etter kan lage en ny, men tilsvarende, prøve ut fra de samme kriteriene, der man kan studere eventuell framgang eller tilbakegang. Forutsetningen for dette er at elevene kan vurderes entydig etter klare kriterier. Initiativet i engelsk (se kapittel 7) med bruk av "Common European Framework" (CEF) er et eksempel på at man har prøvd å innføre en slik kriterierelatert vurdering, og vi skal gå nøye inn på i hvilken grad dette har fungert etter forutsetningen. Disse prøvene tar utgangspunkt i og er bygget opp rundt internasjonale kompetansebeskrivelser. Ut fra resultatene på disse prøvene har vi gitt en grundig analyse av hva utfordringene består i, og i hvilken grad dataene bekrefter eller avkrefter hypotesene. Analysene har prøvd å besvare i hvor stor grad de foreslåtte nivåbeskrivelsene bekreftes av empirien. Spesielt har vi kommentert hvorvidt det er mulig å lage et nytt prøvesett etter de samme kriteriene neste år og kunne si noe troverdig om hvorvidt elevene er bedre eller dårligere enn året før.

Motsatsen til dette er å bruke en relativ eller **normrelatert vurdering**, der vurderingen foregår ved sammenlikning mellom elevene. En normrelatert skala kan med fordel standardiseres, slik at gjennomsnitt og standardavvik settes som gitte verdier, for eksempel som i den vanlige T-skalaen, henholdsvis 50 og 10. Det sier seg selv at det ikke er mulig å sammenlikne slike resultater år for år, gjennomsnittet vil per definisjon alltid måtte settes som det samme. Det vil heller ikke gå an å bruke endring i prosent riktige svar for hver av prøvene som et mål for endring. Det er jo i prinsippet umulig å vite om en tilsynelatende framgang betyr at elevene er blitt flinkere, eller om oppgavene er blitt lettere. Skal man virkelig kunne sammenlikne oppgaver år for år uten å ha kriterierelatert vurdering, må i hvert fall *noen* av oppgavene besvares av et utvalg av elever minst to ganger. Da går det an å "linke" de to oppgavesettene til hverandre på en slik måte at resultatene kan justeres i forhold til hverandre, og sammenlikning blir mulig.

Vi er bedt om å gi en vurdering av hvordan prøvene generelt kan utvikles til å bli måleinstrumenter for faglig framgang. Vi konstaterer fra årets prøver at det foreløpig er få tilløp til at en slik målsetting er lagt vekt på. Det eneste er det nevnte initiativet i engelsk. For de andre prøvene finner vi ingen slike initiativ.

4.7 Råd for publisering av resultatene i 2005 og for videreutvikling av prøvene

Undersøkelsen fokuserer på hva som har fungert bra og dårlig, sett i de perspektivene som er beskrevet ovenfor. Et gjennomgående spørsmål er hvilke konsekvenser dette bør få for eventuell publisering på Skoleporten og for utvikling av neste års prøver. Slike spørsmål vil vi komme grundig inn på i konklusjonene, se kap. 1.

4.8 Presentasjon av analyser av data fra elevbesvarelsene

I kapitlene 5-8 vil vi beskrive resultatene av undersøkelsen i detalj. For hver prøve vil vi først gi resultatene oppgave for oppgave (eller kompetanse for kompetanse) i tabellform. Disse dataene er beregnet og gjengitt i tabellene:

- Svarfordeling i prosent. For flervalgsoppgaver oppgis prosent for hvert svaralternativ (eller bare for riktige og gale svar), samt for blanke svar. For åpne oppgaver oppgis svarfordeling på de ulike skåreverdier.
- Neste sett av data gjelder hvor dyktige elevene er som har gitt hver av disse svarene. Elevenes dyktighet er her gitt som deres skåre på testen som helhet.
- Neste informasjon gjelder oppgavens diskriminering (D), og det er korrelasjonen mellom skåre/nivå/poeng på den aktuelle oppgaven og for testen som helhet (se kap. 4.2).
- For åpne oppgaver kommer det deretter en kolonne med data om hvor likt oppgaven er rettet av de to uavhengige retterne (R). Dette er gjort i form av prosent overensstemmelse (R) og eventuelt verdien på koeffisienten Kappa (se kap. 4.4).
- I en egen kolonne er eventuelle problemer "flagget" i form av henvisning til en fotnote.

Videre oppgir vi for hver prøve disse opplysningene:

- Prøvens reliabilitet (Cronbachs alfa, se 4.3)
- Tilsvarende for hver av delene eller foreslåtte rapporteringskategoriene
- Korrelasjon mellom hver av delene eller de foreslåtte rapporteringskategoriene

Ut fra disse dataene har vi diskutert hvordan prøven har fungert, og særlig har vi gitt en vurdering av hva vi kan anbefale å publisere, ut fra vanlige krav til god testpraksis. Vi har også gitt en vurdering av hvordan utprøvingen ser ut til å ha fungert.

5 Lesing

5.1 Lesing på 10. trinn og grunnkurs

5.1.1 Struktur og vurderingskriterier

Den samme prøven ble gitt til siste år i grunnskolen og til grunnkurs i videregående skole. Denne prøven framstår nesten som en kopi av lestesteten i PISA, både når det gjelder design, balansen mellom åpne oppgaver og flervalgsoppgaver, kategorier foreslått for rapportering, og når det gjelder kriterier for retting av de åpne oppgavene. Spesielt vil vi peke på at oppgavene er organisert rundt åtte tekster med til sammen 44 oppgaver. Hver enhet og hver oppgave er klassifisert etter et sett av kriterier, og det er tilstrebet å rapportere etter de samme tre kategoriene som i PISA: *Finne, Tolke* og *Reflektere* (se mer om dette nedenfor).

Omtrent to tredeler av oppgavene (29 av 44) er åpne oppgaver, en overraskende høy andel, betydelig høyere enn i fjorårets prøve. Vi undrer oss over dette, både i lys av den økte rettebyrden for lærerne og det nødvendigvis økende innslag av målefeil på grunn av den skjønnsmessige vurderingen (se mer om dette senere).

Imidlertid konstaterer vi at det åpenbart er et stort behov for at faggruppa, i samarbeid med faggruppa for lesing blant de yngre elevene, lager et tydelig rammeverk for de nasjonale leseprøvene. Et slikt rammeverk kan nå bygges i tråd med de nye læreplanene i Kunnskapsløftet (jf. læreplanen i norsk med sitt særskilte ansvar for leseopplæringen og beskrivelsen av lesekompetanse i andre fag). Et rammeverk bør inneholde en klar definisjon av den lesekompetansen og de delkompetansene man bestreber seg på å måle på de ulike trinnene gjennom de nasjonale prøvene. Videre bør det gis et rasjonale for den foreslåtte strukturen på prøvene på de forskjellige trinnene.

Tabell 5.1A viser fordelingen av oppgaver for hver av de tre kategoriene *Finne, Tolke* og *Reflektere*. Som det framgår av tabellen, er det en ujevn fordeling av oppgaveformat etter hvilket av disse tre aspektene ved kompetansen som måles. Dette kan være en konsekvens av egenarten til kompetansene og derfor framstå som naturlig. Det er imidlertid viktig at slike ting begrunnes, så man ikke kommer i skade for å trekke feil slutninger fra resultatene. I dette tilfellet kan det tenkes at en eventuell målt forskjell mellom kompetansekategoriene egentlig bare reflekterer en forskjell etter oppgave-formater. For eksempel kan en elev som er særlig svak til å skrive svar med egne ord, lettere skåre høyt innenfor kategorien *Tolke,* siden det nesten bare er flervalgsoppgaver der.

Tabell 5.1A: Fordeling av oppgaver etter kategorier. Lesing, 10. trinn

	Finne	Tolke	Reflektere	Totalt
Flervalg	4	11	0	15
Åpne	10	5	14	29
Totalt	14	16	14	44

5.1.2 Vurdering av prøvens kvalitet og validitet

En vurdering av prøvens validitet er ikke lett, så lenge det ikke finnes noen læreplan i "lesing" i L97. Vi konstaterer at målt med internasjonale mål, slik de er gjenspeilet i PISA-prosjektet, representerer denne prøven nettopp det som forstås som "reading literacy", eller på norsk, lesekompetanse. Å kopiere PISA-prosjektets design har flere fordeler, men det bør begrunnes hvis dette som en selvfølge brukes år etter år. Vi mener at prøven framstår med høy validitet, ut fra hva man med rimelighet kan forlange. En nærmere analyse av validiteten er gitt nedenfor.

Vi har ved gjennomgangen i det følgende valgt å ta utgangspunkt i versjonen av prøven der hovedmålet er bokmål. De to versjonene er for øvrig ekvivalente, siden de samme to tekstene er på sidemålet i begge versjonene.

Det er en styrke ved oppgavesettet at elevenes lesekompetanse prøves i et bredt spekter av emner og tekster. Bredden illustreres godt i følgende oversikt.

Tabell 5.1B: Tekster brukt i leseprøven for 10. trinn og grunnkurs. (Bokmålsversjon)

Tittel (hovedmål-sidemål)	Sammen-hengende tekst	Ikke sammen-hengende tekst	Beskrivelse og kilde
Lesevaner blant norske 10.-klassinger i 2003 og i 2004 (hovedmål)		X	Sakprosa: Søylediagram (prosent-tall) med innledning. Kilde ikke oppgitt
Ansvar for eiga dumping (sidemål)	X (1 s.)		Sakprosa: Avisartikkel, Aftenposten 2004
Fem på topp blant jentenavn i Norge fra 1970 til 2003 (hovedmål)		X	Sakprosa: Tabell som lister opp de fem mest populære navnene i perioden. Kilde: Statistisk sentralbyrå
Hjemmelekse til fru Bishop (hovedmål)	X (3 s.)		Skjønnlitteratur: oversatt novelle av Sue Townsend.
Medium Roxy (hovedmål)	X (1 s.)	X	Sakprosa: Artikkel med tabell (tallvurdering) Kilde: Forbrukerrapporten 2002
Ulv, Ulv… (sidemål)	X (2 s.)		Sakprosa: Innledende forklaring, tre leserinnlegg. Vag kildeangivelse (lokalavis Montana)
En god latter forlenger livet …(hovedmål)	X (1 s.)		Sakprosa: Internett-tekst fra NRK. Kilde: nettsidene til tv-programmet Newton
Tyskerne ventr på dom i lakrisstrid (hovedmål)	X (1s.)		Sakprosa: avisartikkel. Kilde: Aftenposten 2004

Tekst- og emnevalget er altså variert og, etter vår mening, godt egnet for å måle lesekompetanse på de to trinnene. De forbedringsmulighetene vi ser, er noe mer bruk av bilder og arbeid med en noe mer innbydende layout. Dermed vil tekstene i høyere grad samsvare med den vektleggingen av multimodalitet og estetiske dimensjoner som finnes i det tekstuniverset elevene møter til daglig i skole, fritid og samfunn. Når det gjelder rekkefølgen på tekstene, kan man diskutere hvor heldig det er å innlede et prøvesett i lesing med lesing av tall i en tabell.

Prøven bygger som nevnt på samme design som PISA-prøvene i lesing. Stort sett virker dette relevant, og det ser ut til å fungere godt i forhold til å gi et mål for elevenes lesekompetanse. Vi vil imidlertid peke på to utfordringer knyttet til validitet i

bestrebelsene på å kunne måle eventuelle delkompetanser i lesing. I faggruppas informasjonmateriell heter det at oppgavene er laget "for å måle ulike former for lesekompetanse". Og videre at oppgavene kan deles inn etter "hensikt/lesemåte" i følgende tre kategorier:

- å finne fram til konkret informasjon i tekstene
- å forstå hva en tekst handler om, og kunne tolke og trekke slutninger på bakgrunn av informasjon i teksten
- å reflektere over eller vurdere tekstens form eller innhold kritisk eller analytisk. (Fra: *Nasjonale prøver i lesing for 10. trinn og grunnkurs videregående*. www. utdanningsdirektoratet.no)

Det er disse tre kategoriene som forkortes til *Finne, Tolke* og *Reflektere*. En innholdsanalyse av oppgavene illustrerer at det ikke alltid er entydig i hvilken kategori enkelte oppgaver skal plasseres. Her må vi legge til at problemer med å plassere oppgaver entydig i gitte kategorier er velkjent også i andre tilsvarende prøver, for eksempel i PISA-undersøkelsen. Videre mener vi at man kan diskutere om noen av refleksjonsoppgavene er velegnet til bruk i nasjonal prøver i *lesing*. Den følgende analysen rokker ikke ved hovedinntrykket vårt, nemlig at dette er en prøve der elevenes lesekompetanse prøves ved ulike lesemåter av et bredt sammensatt tekstutvalg. Men analysene viser at det ennå må arbeides med å utvikle prøvekonseptet før man eventuelt kan greie å skille ut ulike delkompetanser (dvs. skille mellom *Finne*, *Tolke* og *Reflektere,* slik ambisjonen her har vært).

Vi ser først på noen eksempler på at spørsmål er plassert i en kategori, men kanskje like gjerne hører hjemme i en annen: Spørsmål 22 er kategorisert som refleksjon, men kunne også regnes som tolking. Spørsmålet stilles til den skjønnlitterære teksten og lyder slik: "*På linje 130 står det om bestemora til Steven: 'Ansiktet hennes ble forandret, og hun så gammel ut'. Hva er grunnen til at hun er beskrevet på denne måten?"*. Dette er eksempel på et tekstnært spørsmål som krever at elevene leser mellom linjene. De må "forstå hva teksten handler om og kunne tolke og trekke slutning på bakgrunn av informasjon i teksten", jf. tolkekategorien.

Spørsmål 13 til tabellen over de mest populære jentenavnene er kategorisert som et tolke-spørsmål, men det kan argumenteres for at det heller er et finne-spørsmål. Spørsmålet lyder slik:

> "Katrine var blant de fem mest populære navnene fra 1976 til 1993. Cathrine, som er enda vanligere, er ikke å se. Hva er grunnen til det?"

Svaret finner man her i en merknad til tabellen. Merknaden er markert med en stjerne når "Katrine" dukker opp i tabellen. Forklaringen står under tabellen og lyder slik: "**Navnet kan skrives på flere måter*". For å svare riktig her må man ikke tolke i vanlig forstand, men vel heller lese nøyaktig og kjenne konvensjonene for at tabeller ofte har en slik type merknader.

Spørsmål 7 er et flervalgsspørsmål til artikkelen ”Ansvar for egen dumping”. Det lyder slik:

> Hva er ifølge artikkelen den viktigste årsaken til at forskjellene mellom sterke og svake elever har økt i perioden 1992 til 2002?
>
> A: De flinkeste elevene vil ikke bruke tid til å hjelpe andre.
> B: De svakaste elevane leser ikke det de må for å klare seg på skolen,
> C: Bare de flinkeste elevene klarer å ta ansvar for egen læring.
> D: Skolen har ikke klart å lære de svakaste elevene nok.

Spørsmålet er kategorisert som finne, men det kan diskuteres om ikke dette spørsmålet dreier seg mer om å tolke både de ulike alternativene i spørsmålene og det aktuelle avsnittet i teksten som gir svaret på spørsmålet.

Det neste perspektivet vi vil trekke fram, er spørsmålet om noen av refleksjonsoppgavene egentlig kan sies å måle kompetanser som går ut over det vi vanligvis vil regne til en lesekompetanse. Det dreier seg om oppgaver som opplagt ville hørt hjemme i en nasjonal prøve i *norsk*, men der det kan diskuteres hvor relevante de er i den sammenhengen de nå inngår i, nemlig en nasjonal prøve i *lesing* som står sammen med en nasjonal prøve i skriving, der elevene altså har egen prøvedag og en mappevurdering av skrive-kompetansen sin i tillegg.

I spørsmål 28, ”*Tror du denne testen vil påvirke salget av dongeribukser*?”, som stilles i til teksten ”Medium Roxy”, kan man for eksempel ikke si at elevene prøves direkte i hva de har lest, men mer i hvordan de kan gå videre fra det de har lest og bruke det i en skriveoppgave. Det etterfølgende spørsmålet (nr. 29) er eksempel på en refleksjonsoppgave som er sentral i forhold til forståelsen av tekstinnholdet. Det lyder slik*: Hva kan være grunnen til at forfatteren har valgt overskriften ”Medium Roxy” på denne teksten?*”. Denne er kategorisert som refleksjon, men kunne kanskje like gjerne vært regnet som tolking.

Spørsmål 4 stilles i tilknytning til en tabell om lesevaner blant ungdom, en tabell som ikke har forklarende tekst. I dette spørsmålet inviteres elevene til å være medieforskere og forklare funn man kan gjøre ved å lese tallene i diagrammet: ”*Se på det du svarte på spørsmål 3. Hvorfor tror du det er så store kjønnsforskjeller når det gjelder disse to typene lesestoff? Gi en begrunnelse for hver av de to*”. I vurderingsveiledningen heter det at oppgavens hensikt er ”*å reflektere over tekstens innhold kritisk eller analytisk*”, men dette er en oppgavetype som vanskelig kan sies å være basert på forståelsen av en tekst. Denne oppgaven er for øvrig den eneste av oppgavene til denne teksten som gir to poeng ved tilfredsstillende svar.

Et siste eksempel på refleksjonsoppgaver man bør vurdere om kan forsvare sin plass som mål på lesekompetanse, er spørsmål 36 til tekstene om ulv. Det lyder slik: ”*Hvilken av brevskriverne er du enig med? Begrunn svaret med dine egne ord ved å vise til det som står skrevet i det ene eller i flere av brevene*.” Det som kobler denne oppgaven til de leste tekstene, er altså at man skal velge en brevskriver å være enig med, og at man skal vise til noe som står i brevene. Selv om oppgaven forutsetter at man har forstått det leste, dreier

selve oppgaven seg om å gjøre seg opp egne meninger og skrive dem. Det er en oppgave som avgjort ville ha vært velegnet om det var en bred skriftkyndighetskompetanse (literacy) eller kompetanse i norsk som skulle prøves, men spørsmålet er om den passer like godt som en oppgave i en egen prøve i lesing.

Oppsummerende kan vi si at vurderingen av innholdsvaliditeten understreker behovet for at faggruppene i lesing arbeider sammen om å utarbeide et rammeverk for prøvene. Selv om sider ved validiteten i disse oppgavene kan diskuteres i forhold til ambisjonen om å dokumentere ulike *del*kompetanser i lesing, rokker det altså ikke ved hovedinntrykket av prøven som helhet, nemlig at validiteten er høy som et mål på kompetanse i lesing. Men innholdsanalysen peker i samme retning som item-analysen nedenfor, nemlig at det ikke er grunnlag for å skille ut egne mål for delkompetanser i lesing gjennom denne prøven.

5.1.3 Item-analyse av oppgavene for 10. trinn

Resultatene av item-analysene er gitt i tabell 5.2 sortert etter delkompetanse (*Finne, Tolke, Reflektere*). Som mål på elevenes dyktighet har vi i vår analyse brukt antall poeng eleven har oppnådd på hele prøven. For de 15 flervalgsoppgavene har riktig alternativ gitt ett poeng, mens galt eller blankt svar har gitt 0 poeng. De 29 åpne oppgavene er i de fleste tilfellene vurdert til 1 poeng (riktig) eller 0 poeng (galt). For åtte av disse oppgavene er det brukt 2 (riktig), 1 (delvis riktig) eller 0 (galt) poeng.

Med utgangspunkt i resultatene i tabellen har vi disse kommentarene:

- Disse oppgavene har i stor grad fungert etter hensikten, og faggruppa må ha gjort en meget god jobb med utprøvingen**.** Særlig er det påfallende at alle oppgavene unntatt den aller første diskriminerer godt (D>0,30). Vi ser da også at for hver eneste oppgave er totalpoengene ”riktig ordnet” etter dyktighet, altså at poengene for prøven som helhet øker mot høyre i tabellen (i kolonnen ”Dyktighet”).
- Det er 6,4 % blanke svar i gjennomsnitt for alle oppgavene, 7,9 % for de åpne oppgavene og 3,3 % for flervalgsoppgavene. Dette er ikke særlig høyt, og det er heller ikke en påfallende sterk økning mot slutten av prøven. Dette tyder på at elevene har fått rimelig god tid til å svare og har vært motivert for å svare på prøven.
- Gjennomsnittlig er det 70 % riktige svar på flervalgsoppgavene, mens elevene i gjennomsnitt skårer 59 % av ”fullt hus” på de åpne oppgavene. Til sammen oppnår elevene i gjennomsnitt 62 % av det som er oppnåelig poengsum (52 poeng).
- Graden av overensstemmelse mellom de to vurderingene av åpne oppgaver er gjennomgående tilfredsstillende. Spesielt ligger gjennomsnittet av overensstemmelser for de 21 ”1 poengs-oppgavene” så høyt som 90 %. Bare fire av disse oppgavene ligger under det vi har satt som en kvalitetsgrense på 85 %, og de er alle oppgaver innen kategorien ”Reflektere”.
- Adskillig dårligere overensstemmelse er det når det gjelder de åtte ”2 poengs-oppgavene”. I gjennomsnitt er det bare 76 % overensstemmelse for disse. For 3 % av besvarelsene er avviket på hele 2 poeng. Det er grunn til å uttrykke en viss bekymring for innslaget av oppgaver med store vurderingsproblemer, dette

svekker den totale reliabiliteten for prøven. Med fordel kunne vurderingen av disse oppgavene ha nøyd seg med å skille mellom 0 og 1 poeng.

- Generelt utgjør usikkerheten knyttet til vurderingen omtrent 6 % av variansen i observert skåre. Dette kan sammenliknes med andelen av variansen knyttet til tilfeldigheter i oppgaveutvalget, som er på 9 % (fordi alfa er 0,91).
- Det er samlet sett en liten tendens til at lærerne gir en litt høyere vurdering enn de eksterne. Forskjellen utgjør i gjennomsnitt omtrent ett poeng i favør av egne elever. Av de 32 skolene i utvalget har 16 gitt gjennomsnittlig ett poeng *høyere* til egne elever, mens bare én skole har gitt over ett poeng *dårligere* til egne elever. Denne tendensen til "mild" retting er likevel ikke urovekkende stor og vil ikke påvirke forskjellene mellom skoler i stor grad.

Tabell 5.2: Item-analyse for leseoppgavene for 10. trinn (N=514).
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering, og R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm) er det henvist til ulike fotnoter.

Opp-gave nr.	Åpen eller Fler-valg	Svarfordeling i %				Dyktighet			D	R	Komm
		Blank	0 poeng	1 poeng	2 poeng	Blank/ 0 poeng	1 poeng	2 poeng			
Finne											
1	Å	1	17	82		30	33		,11	97	a
3	Å	2	23	75		23	36		,54	97	
5	Å	3	27	71		26	35		,40	97	
7	F	3	19	78		26	34		,32		
8	F	1	26	72		24	36		,49		
12	Å	3	22	75		27	34		,31	99	
14	Å	3	62	35		29	38		,42	94	
15	Å	5	3	92		18	34		,41	99	
25	F	2	12	87		22	34		,39		
27	Å	4	44	52		27	38		,50	94	
32	Å	12	37	33	19	27	36	40	,49	79	b
34	Å	13	14	26	47	20	33	39	,73	80	b
38	Å	9	15	77		23	35		,50	92	
39	F	5	25	70		24	36		,54		
Tolke											
2	F	3	37	60		28	36		,36		
9	Å	10	16	20	54	23	31	38	,59	69	b!
10	F	3	41	56		29	35		,26		
13	Å	7	34	59		26	37		,50	88	
17	Å	4	9	10	77	18	31	35	,52	85	
18	F	3	20	78		23	35		,48		
19	F	2	44	54		29	36		,32		
20	Å	10	42	48		27	38		,52	97	
21	F	3	32	66		27	35		,40		
26	F	2	11	87		19	34		,49		
30	F	2	12	87		20	34		,47		
31	F	2	19	79		23	35		,44		
35	F	5	40	56		26	38		,57		
40	Å	4	57	31	8	29	37	42	,42	85	
41	F	7	47	45		28	37		,43		
43	F	7	19	74		26	35		,38		
Refl.											
4	Å	5	12	23	60	21	30	37	,54	73	b!
6	Å	5	19	77		24	35		,44	85	
11	Å	11	18	28	43	22	33	38	,64	66	b!
16	Å	7	20	73		24	36		,50	89	
22	Å	5	35	59		26	37		,48	93	
23	Å	8	65	27		30	39		,36	87	
24	Å	16	32	20	32	27	35	39	,49	69	b!
28	Å	9	22	69		24	36		,51	81	b
29	Å	10	34	56		27	37		,48	87	
33	Å	15	41	44		28	38		,47	82	b
36	Å	11	18	71		22	37		,64	88	
37	Å	12	35	54		27	37		,51	81	b
42	Å	10	15	75		23	36		,53	89	
44	Å	16	32	52		27	37		,48	77	b

a) Svak diskriminering (<0,30)
b) Dårlig overensstemmelse mellom rettere (< 85 %, b! betyr < 75 %)

5.1.4 Oversikt over de foreslåtte kategoriene for 10. trinn

Tabell 5.3 viser hvordan hver av de foreslåtte rapporteringskategoriene har fungert.
Vi legger merke til at innenfor alle tre kategoriene har oppgavene vært lette, men det gjelder særlig de to første.

Dessverre viser det seg at etter vår vurdering har alle tre skalaene noe for lav reliabilitet til å kunne rapporteres. Videre er det slik at korrelasjonene mellom de tre kategoriene innbyrdes er omtrent like store (0,76) og omtrent like store som reliabilitetskoeffisientene (se tabell 5.3). Dette betyr at de *latente* korrelasjonene (se kap. 4.5) er nesten 1,0. Det er derfor vanskelig ut fra empiri å se at de foreslåtte kategoriene virkelig representerer *forskjellige* kompetanser. Vi må derfor konstatere at de tre kategoriene hver for seg har tvilsom (diskriminerende) validitet. Slik dette ser ut, tyder alt på at den beste måten å rapportere resultatene fra denne prøven på, er å bruke bare én skala, generell kompetanse i lesing. På den annen side vil vi også peke på at dersom målet i neste omgang blir å rapportere bare én kompetanse i lesing, kan man vurdere om ikke prøven som helhet kan gjøres noe kortere og likevel få god reliabilitet.

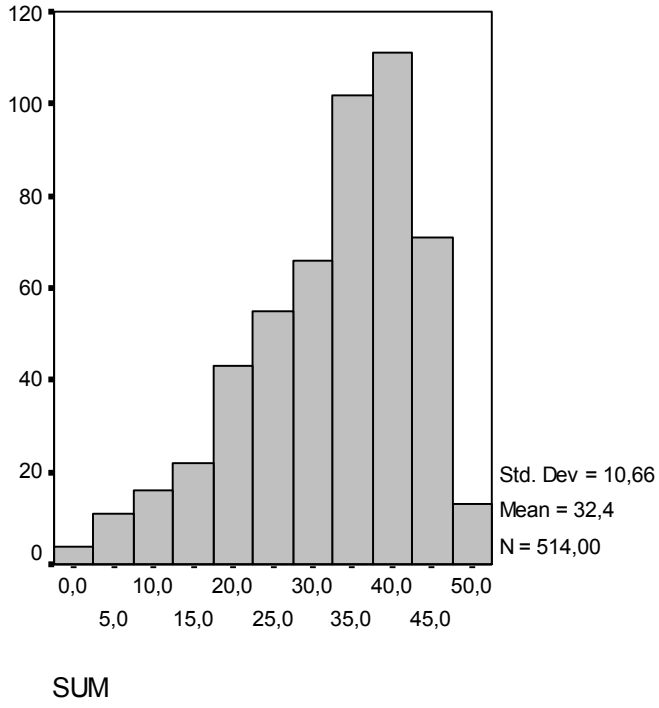
Tabell 5.3: Resultater for hver av kategoriene (N = 514)

Kategori	Antall oppgaver	Antall max. poeng	Gjennomsn. korrelasjon	Reliabilitet	Gjennomsn. andel av "fullt hus"
"Finne"	14	16	0,19	0,76	66 %
"Tolke"	16	19	0,19	0,78	63 %
"Reflektere"	14	17	0,25	0,81	59 %
Totalt	44	52	0,20	0,91	62%

5.1.5 Prøven som helhet for 10. trinn

På figur 5.1 er vist hvordan fordelingen av poeng er for prøven som helhet. Som det framgår der, har prøven en tydelig skjevhet, noe som gjenspeiler at den har vært litt lett. Prøven har imidlertid ingen tydelig "takeffekt" (at veldig mange har fullt hus). Totalt sett fungerer prøven tilfredsstillende for å måle lesekompetanse for enkeltelever og skalaer på det aktuelle trinntrinnet.

Figur 5.1: Fordeling av poeng for lesing på 10. trinn



5.1.6 Item-analyse av oppgavene til grunnkurs

Resultatene av item-analysene for grunnkurs er gitt i tabell 5.4 sortert etter delkompetanse. Oppgavene er som tidligere beskrevet, nøyaktig de samme som for 10. trinn. Ikke på noen viktige punkter er dataene vesentlig forskjellige for de to klassetrinnene. Kommentarene ut fra tabellen er derfor ikke gjentatt her. Det er i seg selv et viktig resultat at analysene gir nøyaktig de samme konklusjonene hva gjelder vanskelighetsgrad, diskriminering og sensorreliabilitet for enkeltoppgavene. Dermed styrkes tilliten til begge datasettene som er lagt til grunn for analysene her.

Tabell 5.4 Item-analyse for oppgavene i lesing for grunnkurs (N =528).
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering, og R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm) er det henvist til ulike fotnoter.

Oppg nr	Åpen eller Fler-valg	Svarfordeling i %				Dyktighet			D	R	Komm
		Blank	0 poeng	1 poeng	2 poeng	Blank/0 poeng	1 poeng	2 poeng			
Finne											
1	Å	1	22	77		29	34		,18	97	a
3	Å	2	20	78		22	36		,48	98	
5	Å	2	27	71		25	36		,44	96	
7	F	6	17	78		23	36		,44		
8	F	3	21	76		23	36		,49		
12	Å	2	26	72		26	36		,38	97	
14	Å	3	58	40		28	40		,46	92	
15	Å	6	6	88		18	35		,47	98	
25	F	3	9	88		17	35		,48		
27	Å	4	37	58		25	39		,58	94	
32	Å	18	42	24	17	28	38	42	,48	77	b
34	Å	21	11	14	54	20	34	40	,75	80	b
38	Å	12	17	71		21	38		,65	93	
39	F	9	21	71		21	38		,64		
Tolke											
2	F	3	41	57		27	37		,44		
9	Å	9	9	18	64	18	29	38	,67	74	b!
10	F	4	35	61		27	36		,37		
13	Å	8	30	62		25	38		,52	93	
17	Å	7	11	11	72	17	34	37	,59	85	
18	F	4	13	84		21	35		,46		
19	F	3	37	60		28	36		,36		
20	Å	11	44	44		28	40		,50	94	
21	F	4	23	73		23	36		,48		
26	F	3	9	89		15	35		,53		
30	F	3	8	89		16	35		,50		
31	F	6	17	78		21	36		,56		
35	F	9	33	58		25	39		,59		
40	Å	6	49	29	17	29	36	40	,36	78	b
41	F	10	40	50		28	37		,38		
43	F	9	21	70		25	36		,45		
Refl.											
4	Å	6	16	24	54	22	31	38	,56	72	b!
6	Å	5	18	77		22	36		,49	89	
11	Å	13	16	21	51	21	33	40	,66	71	b!
16	Å	7	21	72		22	37		,55	90	
22	Å	9	29	62		24	38		,56	90	
23	Å	12	59	29		30	41		,42	86	
24	Å	17	26	18	38	25	35	41	,60	66	b!
28	Å	10	23	67		23	38		,62	83	b
29	Å	14	33	52		26	39		,55	89	
33	Å	22	39	39		28	41		,54	84	b
36	Å	17	16	67		22	39		,67	90	
37	Å	14	28	58		24	39		,61	84	b
42	Å	12	11	77		21	36		,55	93	
44	Å	19	30	51		26	39		,56	81	b

a) Svak diskriminering (<0,30)
b) Dårlig overensstemmelse mellom rettere (< 85 %, b! betyr < 75 %)

5.1.7 Oversikt over de foreslåtte kategoriene for grunnkurs

Tabell 5.5 viser hvordan hver av de foreslåtte rapporteringskategoriene har fungert når det gjelder vanskelighetsgrad og reliabilitet. Angående vanskelighetsgraden er det svært liten forskjell fra 10. trinn. Det er litt høyere gjennomsnittlige prestasjoner på grunnkurs, men forskjellen er ikke signifikant. Det er ut fra elevutvalget på grunnkurs ingen grunn til å spekulere nærmere over hvorfor forskjellen på de to klassetrinnene er så liten. Både den høye boikottprosenten og den usikre representativiteten angående elevenes studieretning gjør at vi avstår fra nærmere sammenlikninger.

Alle tre skalaene har høyere reliabilitet for grunnkurs enn for 10. klassetrinn, siden oppgavene gjennomgående korrelerer litt høyere innbyrdes på det høyeste trinnet. Men likevel har i hvert fall to av skalaene for lav reliabilitet til at rapportering anbefales. Nå viser det seg imidlertid at også korrelasjonene mellom de tre kategoriene innbyrdes er større for det øverste trinnet (0,80-0,83) og altså også her omtrent like store som reliabilitetskoeffisientene (se tabell 5.5). Dette svarer til at de *latente* korrelasjonene (se kap. 4.5) er på nesten 1,0. Det er derfor igjen vanskelig ut fra empiri å se at de foreslåtte kategoriene virkelig representerer *forskjellige* kompetanser, og dette taler imot å rapportere dem hver for seg.

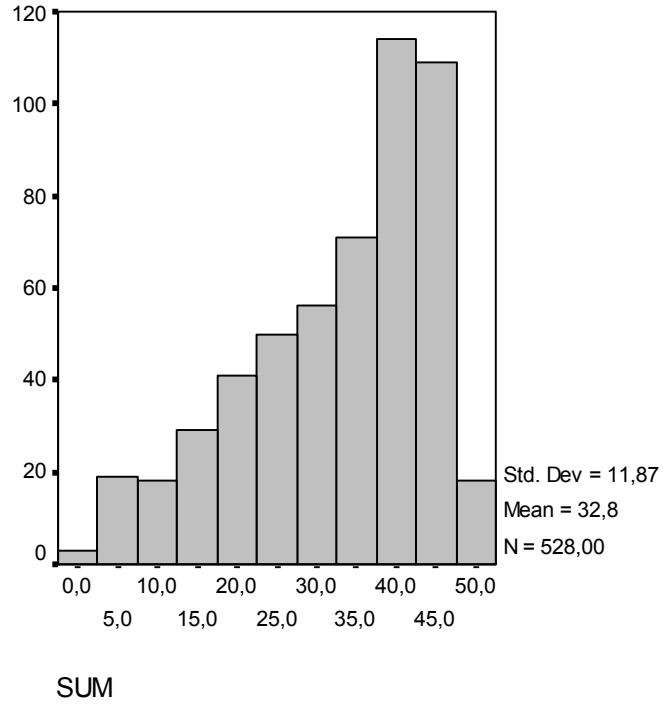
Tabell 5.5: Resultater for hver av kategoriene for grunnkurs

Kategori	Antall oppgaver	Antall max. poeng	Gjennomsn. korrelasjon	Reliabilitet	Gjennomsn. andel av ”fullt hus”
”Finne”	14	16	0,24	0,80	66 %
”Tolke”	16	19	0,23	0,81	65%
”Reflektere”	14	17	0,31	0,85	59 %
Totalt	44	52	0,25	0,93	63%

5.1.8 Prøven som helhet for grunnkurs

På figur 5.2 er vist hvordan fordelingen av poeng er for prøven som helhet. Fordelingen har en tydelig skjevhet, og som det framgår på figuren, har prøven en noe større, men ikke urovekkende ”takeffekt” for grunnkurs enn for 10. trinn.

Figur 5.2: Fordeling av poeng for lesing på grunnkurs



5.1.9 Konklusjon

Denne leseprøven har fungert bra på begge klassetrinn, men det frarådes å rapportere resultatene etter mer enn én skala. Vi anbefaler følgelig at man, hvis resultatene skal offentliggjøres, rapporterer etter én overordnet skala i lesing. De angitte poengene framstår som en poengskala som verken er norm- eller kriterierelatert. Dette gjelder enten man regner i poeng eller i prosent riktige svar. Man kan etter vår mening vurdere om ikke det vil være en fordel heller å standardisere resultatene så de blir normrelaterte, enten på elev- eller skolenivå. På den annen side vil det med rapportering etter poeng eller prosent riktige svar være meningsfullt å sammenlikne data mellom de to klassetrinnene som fikk den samme prøven.

Prøven har ideelt sett vært litt for lett, noe som medfører en viss takeffekt. Vi oppfatter imidlertid ikke dette som noe stort problem. Derimot har vi pekt på at det har vært en uforklarlig økning siden i fjor av andelen *åpne* oppgaver, stikk i strid med hva som ble anbefalt i fjorårets rapport. Dette har etter vår mening ført til unødig mye vurderingsarbeid for lærerne, og i tillegg har innslaget av oppgaver med problematisk sensorreliabilitet økt. Hvis dette er et resultat av ønsket om høyere vanskelighetsgrad, anbefaler vi heller at det nedlegges et konsentrert arbeid for å lage noen flere litt vanskelige flervalgsoppgaver neste år.

Vi vil igjen peke på betydningen av at det utarbeides et grundig rammeverk for de nasjonale prøvene i lesing på de aktuelle trinnene. Dette er særlig viktig fordi det formelt

sett ikke finnes noen læreplan i ”lesing”, men at ”faget” likevel er løftet sterkt fram av myndighetene som et viktig satsingsområde. Et slikt rammeverk bør inneholde definisjoner, kategorier og rasjonale for struktur og format av leseprøvene. Det vil også være viktig å drøfte forholdet til begrepet ”grunnleggende ferdigheter” i St.meld. nr. 30 og læreplanene i norsk og andre fag i Kunnskapsløftet. Den store forskjellen i format mellom leseprøvene på de ulike klassetrinn er ikke noe sted begrunnet og er for oss vanskelig å forstå. Det synes for oss å være et behov for overordnete diskusjoner om dette som en del av rammeverket.

5.2 Lesing i 7. trinn

5.2.1 Struktur og vurdering

Denne prøven inneholder tre lange tekster, den første (heretter referert til som A) består av en liten introduksjon til og et utdrag fra Anne Franks kjente dagbok. Neste tekst (B) handler om Benjamin og ”Holmlia-drapet”, mens den siste (T) er talen Kristin Clemet holdt i 2002 ved avdukingen av Benjamins minnesmerke. Til hver av de tre tekstene er det knyttet 13 oppgaver. Av de i alt 39 oppgavene er 11 åpne, mens de øvrige 28 er flervalgsoppgaver.

Som en innledning til prøven er det en ordkjedeprøve der elevene i løpet av 5 minutter skulle dele inntil 100 lange, meningsløse ord i fire vanlige ord.

Alle flervalgsoppgavene har fire graderte svaralternativer, slik at det ”beste” svaret gir 3 poeng, det nest beste 2 poeng, deretter 1 poeng, mens det antatt dårligste alternativet gir 0 poeng. Elevene fikk informasjon og instruksjon før prøven begynte om at de for hver oppgave skulle finne fram til det beste eller det mest *presise* svaret på spørsmålet.

Dette er en problematisk oppgavetype av flere grunner. For det første er det vanskelig å lage svaralternativer der graderingen av poeng faktisk støttes empirisk, altså at det beste alternativet gjennomgående velges av de beste elevene, det nest beste av de nest beste elevene osv. For det andre, og viktigere, er det krevende å lage alternativene slik at det faktisk er ulik *leseferdighet* som utfordres ved elevenes valg av alternativ, altså at oppgaven tifredsstiller kravet om høy validitet. Vår analyse vil vise i hvilken grad oppgavene og hele prøven fungerer etter forutsetningene.

De åtte åpne oppgavene er vurdert etter en detaljert vurderingsmal, som for alle oppgavenes vedkommende gir fra 0 til 3 poeng.

De tre tekstene er av forskjellig sjanger. Den første er hovedsakelig skjønnlitterær, men har en sakprosapreget innledning og en kronologisk oversikt i tillegg. Den andre teksten er informativ sakprosa, mens den siste er en tale som kjennetegnes ved at den appellerer til elevenes følelser og holdninger. Begge de sistnevnte tekstene er hentet fra Internettet. På ulike måter handler alle tekstene om rasisme, så tematisk er det liten spennvidde i prøven.

På vurderingsskjema er skolene bedt om i tillegg til poeng for hver oppgave å rapportere samlet antall poeng for hver tekst for seg, samt samlet poengsum. Også skåre på ordkjedeprøven skulle rapporteres. I vår kvantitative analyse gir vi resultater for enkeltoppgaver sortert etter hver tekst for seg før vi ser på prøven som helhet.

Som nevnt i forbindelse med leseprøven på 10. trinn, savner vi en begrunnelse for hvorfor det er så stor forskjell mellom prøvene på 7. og 10. trinn. Særlig undrer vi oss over de graderte flervalgsoppgavene som er brukt her, men også at det bare er tre lange tekster, mens det er mange flere, mer varierte og kortere tekster på 10. trinn. Vi kan ikke forstå annet enn at en mer enhetlig tilnærming til måling av lesekompetanse ville ha vært en fordel, eller i hvert fall at forskjellene var utførlig begrunnet. Vi vil stille et spørsmål om det kanskje burde vært flere og kortere tekster på 7. trinn. Siden leseforståelse alltid påvirkes av tidligere kunnskap, ville det etter vårt syn også ha vært en fordel med mer varierte emner.

5.2.2 Validitet

Som beskrevet inngående for leseprøven i 10. trinn, er det vanskelig å diskutere validitet så lenge det ikke finnes noen egentlig læreplan i "faget". Likevel tar vi her utgangspunkt i at prøven er ment å måle leseferdighet ut over ren avkoding. Det spesielle formatet som er brukt på de fleste oppgavene, graderte flervalgsoppgaver, innbyr til en nærmere undersøkelse av hvilke konkrete ferdigheter som i hvert tilfelle ligger til grunn for poenggraderingen av svaralternativene. Vi vil derfor bruke noe plass i det følgende på å diskutere noen utvalgte oppgaver i detalj.

Den første oppgaven til teksten om Anne Frank (A1) lyder slik: "*Hvem er Kitty*?". Det er fire svaralterantiver:

- Kitty er Anne Franks bestevenninne.
- Kitty er Anne Franks søster.
- Kitty er dagboka som Anne Frank bruker som venninne.
- Kitty er Anne Franks brevvenninne

Dette spørsmålet er enkelt (96 % får 3 poeng). Likevel kan det illustrere det problematiske med graderte svaralternativ. Hvis man har lest teksten på en slik måte at man tror at Kitty er Anne Franks søster eller brevvenninne, kan man ikke sies å ha forstått teksten. Likevel gir ett av disse svarene mer poeng enn det andre. I teksten kan man hente rett svar ut to steder. Det heter: "Kitty" er navnet på dagboka, og "…jeg vil la selve dagboka være venninnen, og denne venninnen skal hete Kitty". Det tredje alternativet peker seg følgelig ut som det svaret som er i overensstemmelse med teksten. Så har man i tillegg laget et "nesten riktig svar", nemlig det første. Hvorvidt en elev som velger det, bare delvis har forstått teksten, eller velger svaret av andre grunner (det står først, og det er på én måte riktig) er imidlertid høyst usikkert.

Spørsmål A6 lyder slik: "*Hvorfor føler Anne Frank seg ensom*?". Dette svaret gjenfinnes ikke i en enkelt setning, men forklares over et større avsnitt i teksten. Det er imidlertid ikke umiddelbart enkelt å avgjøre hva som regnes som det mest riktige svaralternativet her, altså det som gir tre poeng. Det første og det siste alternativet (se nedenfor) framstår

som nokså likeverdige, for begge er i overensstemmelse med det vi oppfatter som tekstens hovedpoeng, nemlig at Anne ikke har noen bestevenninne. Hva som nevnes i disse svaralternativene i tillegg til hovedpoenget, oppleves som mindre vesentlig å skille mellom. Går vi til item-analysen i tabell 5.6, ser vi at dette er eksempel på en oppgave der elevenes valg av svaralternativ ikke stemmer godt overens med deres allmenne lesedyktighet. Svaralternativene er:

- Fordi hun har mange slektninger og kjenninger, men ingen bestevenninne.
- Fordi ingen tar henne alvorlig
- Fordi hun må leve innestengt i noen hemmelige rom
- Fordi hun ikke har noen bestevenninne å dele tankene sine med.

I oppgave B1 spørres det ganske enkelt om hvor gammel Benjamin var da han ble drept. I teksten står det at han var 15 år, så dette alternativet gir da 3 poeng. De andre alternativene (16, 17 og 18 år) gir poeng (henholdsvis 2,1 og 0 poeng) etter hvor langt unna tallet er 15 matematisk sett. Men vi kan vanskelig forstå hva en slik gradering har med leseferdighet å gjøre. Dette er etter vår mening en oppgave som opplagt burde gitt 0 eller 1 poeng, 16 år kan i forhold til leseprosessen ikke anses "riktigere" enn 17 år. Oppgaven har derfor et alvorlig validitetsproblem. En annen sak er at nesten alle elevene svarer riktig, så dette problemet får ikke stor betydning for elevenes sluttresultat.

Det er relativt mange eksempler på oppgaver av denne typen der det rette svaret kan hentes rett ut av teksten. De dreier seg følgelig om spørsmål der det finnes ett rett svar. I slike tilfeller virker det oftest urimelig å gi poeng for alle de gale svaralternativene. I den tredje teksten, Kristin Clemets tale ved avduking av minnesmerket for Benjamin Hermansen, er det for eksempel opplyst to steder på den første siden at det var Clemet som holdt talen og ett sted hvilken dato talen ble holdt. I svaralternativene på det første spørsmålet (C1), om når talen ble holdt (1. november 2002), har man blant annet brukt andre datoer som nevnes i teksten, men som da forteller om helt andre begivenheter. ("*Benjamin Hermansen ble drept natten til 27. januar*" og "*Den 6. februar i år markerte 70 000 Oslo-elever Benjamin Hermansens død*"). Vi kan ikke se på hvilke måter svaralternativene, 27. januar 2001, 6. februar 2002 og 6. februar 2001, kan sies å representere gradvis leseforståelse av den datoen Clemet holdt talen. Elever som velger ett av de gale svarene kan rett og slett ikke ha lest overskriften til teksten. Det er vel det som er deres "leseproblem", og ikke at de velger en av de andre datoene. Tilsvarende gjelder for spørsmålet om hvem som holdt talen (C2), men her er navnet på taleren altså gitt to ganger på den første siden. Valg av ett av de andre alternativene kan vanskelig sies å representere nest best forståelse osv. (Alternativene var: *Martin Luther King, Hedda Himle Skandsen, En student med pakistansk bakgrunn*).

Spørsmål 12 til den samme teksten lyder slik: "*Hvor mange i familien Sachnowitz ble sendt til konsentrasjonsleiren Auschwitz?*" (C12). Dette er et detaljorientert spørsmål til noe Clemet bare nevner som et eksempel i sin tale. Det kan derfor stilles spørsmål ved hvor relevant spørsmålet i de hele tatt er i vurderingen av kompetansen i å lese en tale. Det aktuelle avsnittet i teksten lyder slik: "*Den norske jøden Herman Sacnowitz ble sammen med faren, fire brødre og en søster sendt med slaveskipet "Donau" til Stettin og videre til konsentrasjonsleiren Auschwitz i Polen*." Svaralternativene er:

- Faren, fire brødre og en søster
- Herman, faren hans og fem søsken
- Herman og foreldrene hans
- Hele familien på farssiden

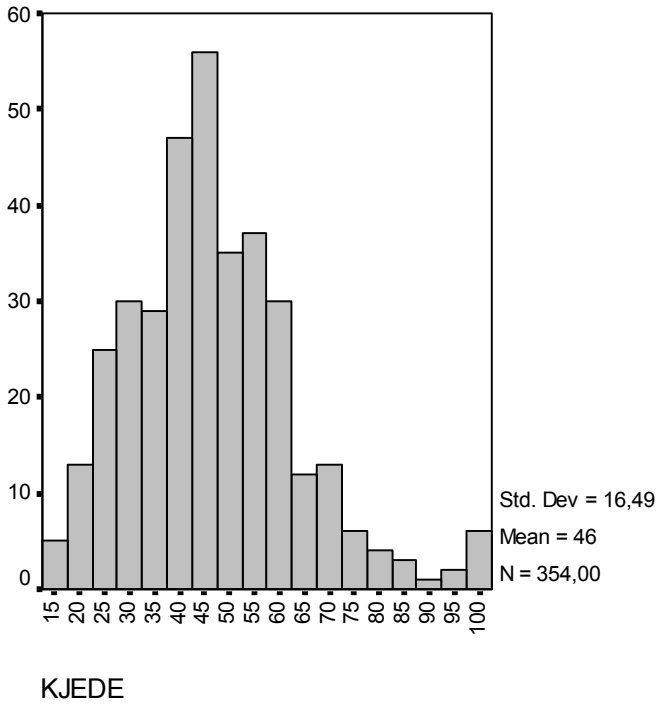
Igjen er det vanskelig å se hvordan de tre gale alternativene skal rangeres og gi poeng i forhold til forståelse av det leste.

Siden de graderte flervalgsoppgavene også er brukt i 4. trinn oppsummerer vi bruken av disse etter å ha sett på oppgavene på dette trinnet.

5.2.3 Resultater for ordkjedeprøven

Figur 5.3 viser fordelingen av antall riktige ord. Det synes å være et problem at noen kanskje har gitt for lang tid, i og med at det er noen ekstremt høye skåreverdier. Det er trolig ikke mulig å skåre fullt hus uten at elevene har fått bruke for lang tid. Men bortsett fra dette er det en grei fordeling. Påfallende mange elever har ikke data fra denne ordkjedeprøven, trolig fordi skolene ikke har sendt dette inn.

Figur 5.3: Fordeling av poeng (riktige ord) på ordkjedeprøven i 7. trinn (N=354)



Det er åpenbart at ordkjedeprøven kan gi viktig diagnostisk informasjon om enkeltelever, informasjon som kan være viktig som bakgrunn til å forstå hva som ligger bak eventuelle lave resultater på den egentlige leseprøven. Vi understreker imidlertid at ordkjedeprøven i seg selv tydeligvis ikke er ment å inngå i måling av leseforståelse, men å være en viktig *premiss* for tolkning av dårlige resultater. Vi vil likevel peke på at en slik prøve nødvendigvis opptar noe tid til fortrengsel for den "egentlige" prøven i lesing. Og spesielt

er vi bekymret for at den trolig tar mye tid å vurdere. Også det faktum at noen skoler synes å gi elevene for lang tid til rådighet, svekker verdien av ordkjedeprøven.

5.2.4 Resultater for leseprøven

Resultatene av item-analysene er gitt i tabell 5.6. Som mål på elevenes dyktighet har vi i vår analyse brukt antall oppnådde poeng. For hver av de 39 oppgavene har beste svar gitt 3 poeng. ”Fullt hus” er altså 117 poeng. I denne tabellen har vi i tillegg til prosentfordeling etter oppnådd poeng tatt med en egen kolonne, markert ”V”, som står for oppgavens vanskelighetsgrad. Med dette menes hvor mange prosent av ”fullt hus” elevene i gjennomsnitt har oppnådd. Et høyt tall betyr altså en lett oppgave.

Tabell 5.6: Item-analyse for leseoppgavene på 7. trinn (N=450).
Svarfordelingen i % og dyktigheten (gjennomsnittlig oppnådd poeng på hele prøven for de som har svart slik) er avrundet til hele tall. V angir vanskelighetsgrad, gjennomsnittlig poengsum i prosent av 3 poeng. D står for oppgavens diskriminering. For åpne oppgaver står R for henholdsvis % full overensstemmelse i vurderingen og % der avviket er 1 poeng. I kolonnen for kommentarer (Komm) er det henvist til ulike fotnoter.

Oppg.	For-mat	V	Svarfordeling i %					Dyktighet				D	R	Komm
			Blank	0p	1p	2p	3p	Blank / 0p	1p	2p	3p			
A1	F	98	0	1	1	2	96	-	-	70	89	,22		b
A2	F	86	0	6	7	12	76	74	76	82	91	,35		
A3	Å	67	1	10	18	29	42	70	83	87	95	,50	74 - 21	c!
A4	F	66	1	13	18	23	45	79	82	83	96	,41		
A5	Å	50	10	23	10	32	25	77	89	95	92	,44	85 - 12	a
A6	F	77	0	5	4	49	43	64	80	90	89	,26		a, b
A7	F	76	0	8	2	41	48	67	81	88	92	,41		
A8	F	74	0	9	6	42	44	74	85	89	91	,27		b
A9	F	78	0	12	13	5	70	69	77	80	94	,57		
A10	F	78	0	6	3	43	48	64	72	88	92	,41		
A11	F	52	1	14	43	11	30	82	86	87	95	,29		b
A12	Å	66	6	20	8	4	61	78	83	91	93	,41	94 - 4	
A13	F	91	0	6	5	3	87	71	68	71	91	,41		a
B1	F	99	0	0	1	2	97	-	-	70	89	,22		b
B2	F	93	0	4	2	4	90	70	59	74	90	,37		a
B3	F	74	0	2	17	39	42	69	79	85	95	,41		
B4	F	82	0	5	9	21	65	71	80	83	92	,36		
B5	Å	73	2	9	22	2	64	70	83	80	93	,49	93 - 6	a
B6	F	84	0	6	11	6	76	69	76	79	92	,47		
B7	Å	68	11	13	2	16	57	73	82	88	94	,56	92 - 6	
B8	F	80	0	6	6	26	61	72	73	85	93	,43		
B9	Å	59	3	8	25	39	25	65	82	90	101	,65	87 - 11	
B10	Å	48	11	10	32	29	18	72	86	95	100	,60	74 - 23	c!
B11	F	81	2	10	4	14	70	69	73	83	93	,52		
B12	F	79	2	5	18	7	68	76	83	77	92	,32		a
B13	Å	79	8	10	0	8	74	71	-	83	93	,53	97 - 2	
T1	F	82	1	12	5	9	74	74	76	79	92	,44		
T2	F	83	0	1	18	14	68	-	74	83	93	,50		
T3	F	77	1	5	13	29	53	68	77	86	93	,46		
T4	Å	86	3	7	5	2	83	69	87	79	91	,39	97 - 2	
T5	F	93	0	2	3	9	86	59	57	70	91	,56		a
T6	F	76	0	11	9	23	58	81	76	80	94	,40		a
T7	Å	70	9	7	10	22	52	70	87	89	93	,49	84 -15	c
T8	F	70	3	21	2	13	61	75	75	85	94	,51		a
T9	F	61	2	17	25	9	47	81	87	89	91	,22		b
T10	Å	34	20	22	36	2	21	79	91	93	99	,46	81 - 14	c
T11	F	87	2	2	5	17	74	67	80	77	92	,42		a
T12	F	71	2	2	4	66	26	61	67	89	93	,41		
T13	F	85	2	6	8	4	80	71	78	78	91	,40		a

a. Svaralternativene ikke ordnet etter dyktighet
b. Svak diskriminering (D<0,30)
c. Dårlig overensstemmelse mellom vurderingene (< 85 %, c! betyr < 75%))

Fra resultatene i tabellen vil vi peke på noen få trekk:

- Det er mange bemerkninger (”flagg”) i høyre kolonne i tabellen. For halvparten av oppgavene er det knyttet kommentarer som går på ett eller flere

kvalitetsproblemer. Selv om flere av disse problemene isolert sett kanskje kan sies å være små, blir summen av disse urovekkende.

- Diskrimineringen er gjennomgående god. Seks av oppgavene diskriminerer lavere enn 0,30, men i de fleste tilfellene dreier det seg om oppgaver med svært høy andel riktige svar, og da er dette vanskelig å unngå. De fleste svaralternativene synes ut fra empiri å være riktig gradert etter økende dyktighet, men i 11 av oppgavene fungerer ikke dette bra. Avvikene er imidlertid i de fleste tilfellene nokså små og gjelder for et lite antall elever.
- Med de gjeldende poengregler har oppgavene gjennomgående lav vanskelighetsgrad, idet gjennomsnittet ligger på omtrent 75 % av "fullt hus". Det er en tendens til inflasjon i poeng, men dette utgjør i seg selv ikke noen trussel mot prøvens kvalitet.
- Det er få blanke svar, og spesielt er det ingen sterk tendens til økende antall blanke svar på slutten av hver tekst. Dette tyder på at elevene har hatt rimelig tid til å svare på alt.
- Det er stort sett rimelig god overensstemmelse mellom de to uavhengige vurderingene. For fire av de 11 åpne oppgavene er det lavere enn 85 % overensstemmelse, og i to tilfeller er overensstemmelsen lavere enn 75 %. Totalt sett utgjør ikke sensorreliabilitet et sterkt bidrag til målefeilene. Det er en liten, men ikke problematisk tendens til at lærerne vurderer sine egne elever høyere enn i den eksterne vurderingen.

5.2.5 Resultater for hver del og samlet

Tabell 5.7: Data for hver av delene av leseprøven i 7. trinn (N=450)

Del	Antall oppgaver	Gjennomsn. korrelasjon mellom oppgavene	Reliabilitet	Gjennomsn. skåre
Ordkjede	1	-	-	47 %
A (Tekst 1)	13	0,14	0,68	74 %
B (Tekst 2)	13	0,20	0,77	77 %
T (Tekst 3)	13	0,18	0,71	75 %
Totalt	39	0,16	0,88	75 %

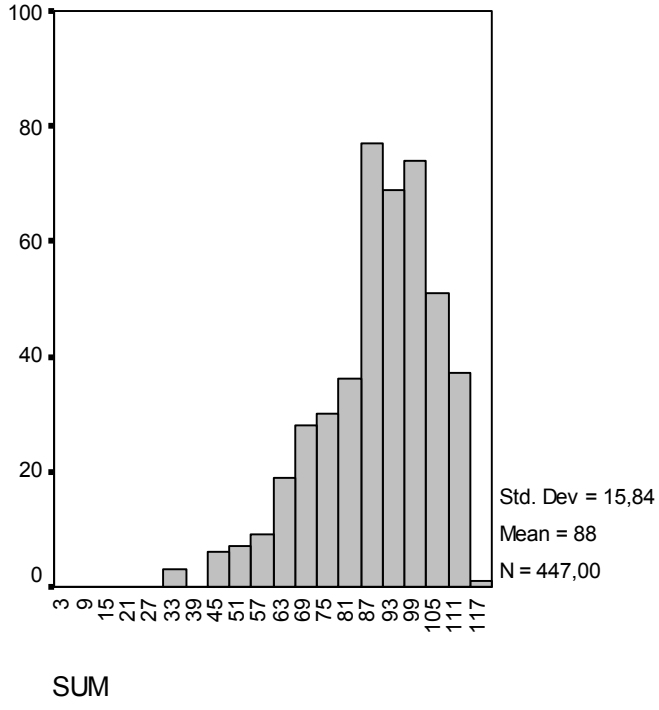
Tabell 5.7 inneholder data om de to delene av prøven samt for prøven som helhet. Det framgår tydelig at reliabiliteten for hver av de tre delene (tekstene) er for lav til å innby til rapportering. En annen sak er at det er lite meningsfullt å knytte skåreverdier til ulike tekster hvis det ikke framgår tydelig hva slags kompetanse som er målt i den aktuelle teksten. Det er videre lite som tyder på at oppgaver innen hver tekst er spesielt likere enn oppgaver på tvers av tekstene. Fra tabell 5.7 finner vi at korrelasjonene internt innen hver tekst er gjennomsnittlig på 0,17, nesten nøyaktig det samme som gjennomsnittet for *alle* oppgavene (0,16). Det er derfor ikke noe i dataene som tilsier at det dreier seg om *ulike* kompetanser i de tre delene.

For prøven som helhet er reliabiliteten tilfredsstillende, men ikke spesielt høy sett i lys av at bruken av graderte flervalgsoppgaver trolig har hatt høyere reliabilitet som mål. Gjennomsnittlig korrelasjon oppgavene imellom (0,16) er for eksempel betydelig lavere enn den er på 10. trinn (0,20) og på grunnkurset (0,25).

Det er noe uklart for oss hvordan ordkjedeprøven er tenkt brukt av lærerne. Den korrelerer middels høyt (0,52) med skåre for prøven for øvrig, så den kan ikke forklare mye (ca. ¼ av variansen) av den målte leseferdigheten. Det er heller ingen del av prøven som korrelerer spesielt høyt med ordkjeden.

Fordeling av poeng for prøven som helhet er vist som et histogram i figur 5.4. Ser vi på hele poengområdet fra 0 til "fullt hus" (117 poeng), synes det som prøven har en svært skjev fordeling. Men i lys av at oppgaveformatet har produsert en "inflasjon" av poeng, er det egentlig området fra 40 poeng og oppgaver som er viktig. En "blind" gjetting på alle de 28 flervalgsoppgavene vil automatisk honoreres med omtrent 42 poeng (gjennomsnittlig 1,5 per oppgave), slik at her ligger det vi kan tolke som det "egentlige" nullpunktet.

Figur 5.4: Histogram som viser fordelingen av poeng på prøven som helhet (N=447)



5.2.6 Konklusjon

Vår vurdering er at denne prøven ikke har fungert særlig bra. For det første har vi påvist at mange av de graderte flervalgsoppgavene har nokså tvilsomme svaralternativer i lys av

de poengene som er tilknyttet dem. Dette har betydd en alvorlig svekkelse av validiteten for prøven som måling av *leseferdighet.* Siden fjerdeklasseprøven er laget over tilsvarende mal, må kommentarene til 7. trinn sees i sammenheng med analysen av 4. trinn som følger nedenfor.

Vi er også skeptiske til den pedagogiske verdien av å fokusere på og innrapportere skåreverdier separat for hver tekst, så lenge det er helt uklart hva slags spesiell kompetanse som derved kommer til syne. Vi kan ikke se at elever og lærere gjennom slike resultater får fruktbare "profiler", altså informasjon som kan fortelle hvor elevene har sine sterke og svake sider. Videre mener vi man bør revurdere om variasjonen i emner og sjanger ikke bør bli større enn det er i foreliggende prøve.

5.3 Lesing 4. trinn

5.3.1 Struktur og vurdering

Leseprøven for 4. trinn inneholder to lange tekster som introduseres for elevene under en felles overskrift: "Fabeldyr". Den første teksten er skjønnlitterær, og det er en fantastisk fortelling skrevet av Astrid Lindgren, "Dragen med de røde øynene", heretter kalt D. Den andre er en sakprosatekst om ulike fabeldyr, "Enhjørningen og andre fabeldyr", heretter referert til som E. Flere fargeillustrasjoner er integrert i tekstene. Dette kan gjøre dem innbydende å lese, men det kan også virke distraherende i lesingen. De omtalte fabeldyrene er illustrert, men det er ikke lett å se hvilke bilder som skal illustrere hvilke dyr, særlig fordi bildene på side 36 og 37 synes å være forbyttet. Bildene passer derfor dårlig til teksten på samme side.

Til hver av tekstene er det 16 flervalgsoppgaver. Vi konstaterer at det altså bare er to lange tekster i prøvesettet, og at hver av dem er fulgt av mange oppgaver, alle av samme format (flervalg). I faggruppas omtale av prøven, *Nasjonal prøve i lesing for 4. trinn. Informasjons- og eksempelmateriale*, heter det at prøven er "modellert over PIRLS-designet".

Som en innledning til prøven er det en ordkjedeprøve der elevene i løpet av 5 minutter skulle dele 75 lange, meningsløse ord i tre vanlige ord. Denne er, som for 7. trinn, ikke ment å inngå som et *ledd* i måling av leseferdighet, men snarere som informasjon til *tolkning* av resultatene på selve prøven. Som for de eldre elevene, forekommer det oss nokså opplagt at noen elever har fått for lang tid på denne delen. Ellers hadde det trolig ikke vært mulig å oppnå så høyt antall riktige ord som noen elever har klart.

I informasjonsheftet heter det at hovedvekten i prøven er lagt på leseforståelse. Om oppgavene elevene skal løse, heter det "*Tekstoppgavene sier noe om hvor mye informasjon elevene er i stand til å finne i ulike teksttyper, hvordan de tolker og sammenholder informasjon, og i hvilken grad de klarer å granske og vurdere form og innhold i en tekst*". Resultatene innrapporteres på hver tekst og på ordkjedeteksten.

Vi mener prøvedesignet med to lange tekster fra samme emneområde kan diskuteres. Det er allment kjent at forståelsen av tekster i høy grad er avhengig av leserens

bakgrunnskunnskaper og ordforråd: Vi forstår tekster om emner vi kan noe om og er fortrolig med språket innenfor, bedre enn tekster om emner vi kan lite om. Når man i en nasjonal prøve velger å prøve forståelse av tekster innenfor bare ett emneområde, kan det hevdes at man ikke i tilstrekkelig grad tar hensyn til dette. Etter vår mening er det også uheldig at man i prøven for 4. trinn velger to *lange* tekster. Mange av oppgavene som stilles til tekstene, tar først og fremst sikte på å måle en nøyaktig ord- og setningsforståelse og ikke den mer helhetlige tekstforståelsen. I informasjonsheftet gis det eksempler på "representative spørsmål". Her kommer det tydelig fram at det først og fremst er en slik nøyaktig lesing man tilstreber:

> "Alle disse spørsmålene vil kreve nøyaktig lesing. Det vil si at elevene aktivt blar tilbake i teksten og finner det rette svaret, gjerne med utgangspunkt i de fire svaralternativene til hvert spørsmål. De vil oppdage at det ofte er små nyanser som skiller alternativene fra hverandre, og de beste leserne vil velge svar etter kritisk gjennomgang av hvilket alternativ som i størst grad er i pakt med det teksten sier. I dette ligger det innebygget en forutsetning om nøyaktig lesing, at elevene leser alle alternativene, og ikke bare velger det første og beste." (Nasjonal prøve i lesing for 4. trinn. Informasjons- og eksempelmateriale, s.6)

Den nøyaktige lesingen omfatter altså både svaralternativene i spørsmålene og teksten. Spørsmålene er av en slik art at man forventer at elevene "aktivt" skal bla tilbake og lete i teksten. Vi kan vanskelig se at denne typen lesing og oppgaver krever at oppgavesettet bare består av lange tekster. Med et slikt formål måtte det faktisk vært en fordel med variasjon i tekstlengden. Et annet ennå uavklart spørsmål er hvor stor plass denne typen spørsmål skal ha.

5.3.2 Validitet

Alle oppgavene er av typen graderte flervalgsoppgaver. Gradering er nytt av året og er i Utdanningdirektoratets informasjon på Internettet framstilt som ett av tiltakene som på kort sikt skal bidra til å forbedre prøven fordi elevene kan krediteres for "delvis forståelse", og fordi dette vil gi "*enda bedre opplysninger til pedagogisk bruk av resultatene*" (jf. *Nasjonale prøver publiseres og videreutvikles* hentet fra www.utdanningsdirektoratet.no). Elevene skal, som for sjuende trinn, velge ett av fire svaralternativer, og i poengsettingen gir de ulike alternativene også her enten tre, to, ett eller null poeng. Etter vår mening har dette formatet i betydelig grad svekket prøvens validitet. Dette vil bli nærmere begrunnet ved gjennomgang av noen eksempler fra oppgavene til den første teksten.

Ett problem er oppgaver som bare har ett rett svar. I spørsmål 15 til den første teksten (D15) spør man: "*Hva betyr pilutta*?". Dette er eksempel på et detaljorientert spørsmål som krever en nøyaktig lesing av en setning og at elevene kan finne tilbake til steder der dette stod i teksten, eller huske hva ordet betyr fra lesingen. Det aktuelle stedet i teksten er formulert slik: *"- Hører du det, din stabukk, sa bror min til ham en gang han var sånn. - Du skal aldri mer få så mye som en lysestubb, pilutta, pilutta. ("Pilutta" sa vi den gangen, det betydde omtrent det samme som "Haha")*." Alternativene elevene skal velge mellom når de svarer på spørsmålet, er disse:

- Lysestubb

- Haha
- Stabukk
- Dumming

Leser man teksten, er det bare ett av alternativene som er rett. Vi kan ikke se på hvilken måte man skal kunne si at noen av de andre alternativene representerer en gradvis forståelse, og etter hvilke kriterier som har sammenheng med leseforståelse, man skal kunne si at de tre øvrige skal tildeles to, ett eller null poeng.

Tilsvarende problem har man i andre oppgaver der ett alternativ er rett ifølge teksten, men der et annet svaralterantiv er konstruert slik at også det kan være "nesten rett". Et slikt eksempel har vi i det første spørsmålet, D1, som lyder slik: "*Hva har grisemoren fått om natten*?" Svaralternativene er:

- Grisunger
- Rampete unger, såkalte dragebarn
- Ti grisunger og en drage
- Ny halm til grisungene

Den innledende setningen i Lindgrens tekst er "*Jeg husker dragen vår*". Videre forteller jeg-personen om det overraskende som hendte en morgen da hun og broren skal i grishuset for å se på grisungene som purka har fått i løpet av natta, men oppdager at det ikke bare er ti grisunger der, men også en drage. *"-Hva er det der? sa bror min, og han var så forbauset at han nesten ikke kunne snakke. – Jeg tror det er en drage, sa jeg. – Purka har fått ti grisunger og en drage."* Som vi ser, kan svaret "*ti grisunger og en drage*" hentes rett ut av teksten. Setningen understøttes dessuten av det foregående innholdet i avsnittet. Dette svaralternativet gir altså 3 poeng. Når det gjelder de andre alternativene, er det kanskje forståelig at det første av disse regnes som bedre enn de to andre, i og med at det kan sies at det også er noe rett i det. Det er rett på den måten at det stemmer med hva vi vet om griser fra før. Men i teksten er hele poenget at grisemora så overraskende har fått ikke bare grisunger, men altså også en drage. Det er derfor ikke opplagt på hvilken måte dette alternativet kan sies å representere en gradvis forståelse av teksten og skal belønnes med poeng. De to siste alternativene er begge definitivt feil, i den forstand at det ikke refererer til noe som er gitt i teksten. Å gradere disse to alternativene mener vi er umotivert som mål på lesekompetanse.

Det er også eksempler på oppgaver der elevene inviteres til å tolke. I spørsmål D5 er det én av tekstens setninger som skal tolkes. Spørsmålet lyder slik: "*Hva betyr det at: "Dragen ville sikkert ha sultet i hjel om ikke bror min og jeg så iherdig hadde gått til grisehuset med den vesle korga vår*"? Svaralternativene er de følgende:

- At de to barna hadde mat til dragen i korga.
- At dragen ikke vil spise den maten han får av grisemoren.
- At dragen spiser korger.
- At dragen får den maten han trenger av de to barna.

Her er det ikke lett å avgjøre hvilket alternativ som er mest rett, nest mest rett og tredje mest rett. Dermed er det vanskelig å forstå hvordan det å sette ulike poeng på disse svarene skal kunne sies å reflektere fjerdeklassingenes ulike grader av leseforståelse.

Hva slags forskjell i lesekompetanse måles for eksempel gjennom at en elev krysser av for det første alternativet og en annen for det siste? Hvem av disse har lest og forstått teksten best? Det viser seg i dataene at denne oppgaven diskriminerer dårlig. Nesten alle elevene svarer ett av disse to alternativene, og de er omtrent like gode på prøven som helhet. Her mener vi "leseproblemet" rett og slett kan forklares med at man konstruerer svaralternativer som er så like.

Spørsmål D16 dreier seg om avlutningen på Lindgrens tekst. Spørsmålet er hvorfor dragen reiser. Her er et av de få eksemplene i prøven på at man ikke kan finne svaret direkte igjen i teksten. Dette er imidlertid ikke signalisert på noen måte i spørsmålet slik at fjerdeklassingene kan få hjelp til å vite at de nå må bruke teksten annerledes. Svaralternativene er de følgende:

- Han er lei av å spise lysestubber.
- Han er blitt større og vil klare seg selv.
- Han vil finne et sted der han kan være som de andre.
- Han vil prøve å finne drakemoren sin.

Igjen er det ikke lett å finne ut hvilket svar som her er mest rett osv. Bortsett fra det første alternativet, er alle de tre andre sannsynlige. Vi kan ikke se hvordan man måler ulike aspekter eller grader av leseforståelse gjennom å få elevene til å velge ett av disse alternativene. Dette er et eksempel på en oppgave hvor et åpent spørsmål trolig hadde vært bedre egnet til å fange elevenes forståelse.

Den andre teksten er på en måte en sakprosatekst om fabeldyr. Den beskriver de fysiske og mentale kjennetegnene til henholdsvis enhjørningen, basilisken, griffen og dragen. Elevene blir gjennom oppgavene invitert til å gå inn i fiksjonen og svare på spørsmålene som om disse dyrene er virkelige. I noen sammenhenger fører dette etter vår mening galt av sted. Spesielt har spørsmål E15 fått offentlig kritikk, en kritikk vi fullt ut deler. Spørsmålet lyder: "*Hvem av enhjørningen, basilisken, griffen og dragen bør du frykte mest?"* Svaralternativene er disse:

- Enhjørningen: fordi den stikker med hornet.
- Basilisken: fordi den kan drepe med bare et blikk.
- Griffen: fordi den er fryktinngytende.
- Dragen: fordi den kan sprute ild.

Det åpenbart eneste "riktige" svaret her at man ikke "bør" frykte noen av dem, siden de ikke finnes. Men dette svaret "finnes" altså ikke, noe som viser at elevene inviteres til en blanding av fiksjon og sakprosa, uten at denne dobbeltheten signaliseres i spørsmålsstillingen. Etter vår mening burde det ha framgått tydeligere at elevene skulle svare ved ikke å stille seg kritisk til premissene, altså ved å godta de fiktive beskrivelsene som saklig informasjon.

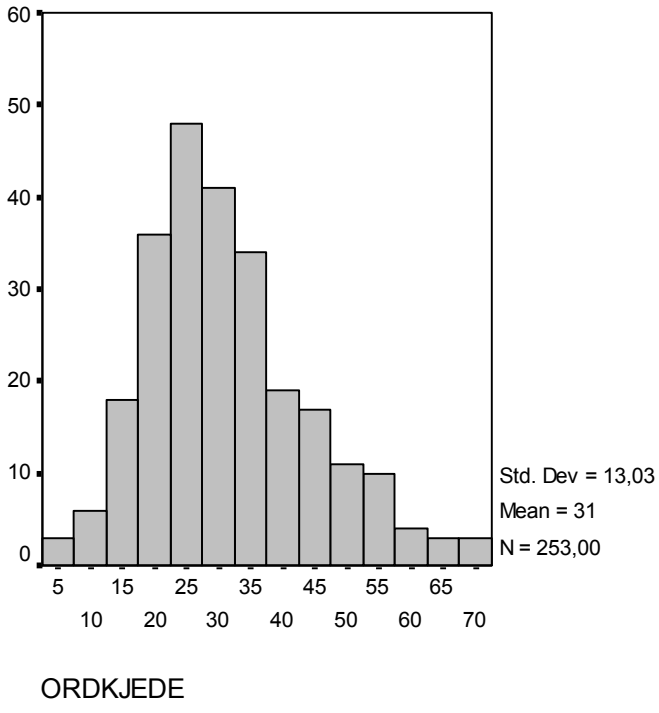
Oppsummerende kan vi slå fast at den uttalte intensjonen med å innføre slike graderte oppgaver blant annet har vært å imøtekomme kritikk av fjorårets prøve der det ble opplevd som urimelig at det kunne være "*små nyanser som skiller det riktige svaret fra de andre*". I informasjonsskrivet om prøven som faggruppen har utarbeidet, heter det videre at det er "*et helt klart poeng at leseforståelse ikke bare er 0 eller 100. Graderte*

svaralternativ fanger opp at utviklingen av leseforståelse er en prosess". Vi mener både dette prøvesettet og prøven for 7. trinn illustrerer hvor vanskelig det er å lage valide prøver med bruk av slike graderte svaralternativer. Det må også tilføyes at selv om det i mange situasjoner kan virke fornuftig å snakke om delvis forståelse, så er det samtidig aspekter ved leseforståelsen hvor dette ikke er rimelig. Prøveformatet med graderte svaralternativ fører til, som eksemplene ovenfor viser, at man kan komme til å belønne elever for gale svar samt å konstruere nesten riktige svar som kan forvirre elever som har forstått det leste, men som velger et svaralternativ som gir dem mindre poeng enn tekstforståelsen deres skulle tilsi. Vi mener vi har illustrert hvor vanskelig det er å lage rimelige mål for elevers leseforståelse på denne måten. Eksemplene vi har trukket fram er heller ikke enestående. Vi vil videre peke på at disse vurderingene langt på vei støttes av de empiriske analysene, se tabell 5.8.

5.3.3 Resultater for ordkjedeprøven

Figur 5.5 viser fordelingen av antall riktige ord. Det synes å være et problem at noen kanskje har gitt for lang tid, i og med at det er noen ekstremt høye skåreverdier. Men bortsett fra dette er det en grei fordeling.

Figur 5.5: Fordeling av poeng (riktige ord) på ordkjedeprøven på 4. trinn



Enda mer enn for 7. trinn er det åpenbart at ordkjedeprøven gir viktig diagnostisk informasjon om enkeltelever, informasjon som kan være viktig som bakgrunn til å forstå hva som ligger bak eventuelle lave resultater på den egentlige leseprøven. Vi

understreker imidlertid at ordkjedeprøven i seg selv tydeligvis ikke er ment å inngå i måling av leseforståelse, men kan gi en viktig *premiss* for tolkning av dårlige resultater. Som for 7. trinn kan denne delprøven forklare omtrent ¼ av variansen i den egentlige leseprøven. Og som på den prøven er vi bekymret for tiden det har tatt å vurdere den.

5.3.4 Resultater for leseprøven

Resultatene av item-analysene er gitt i tabell 5.8. Som mål på elevenes dyktighet har vi i vår analyse brukt antall oppnådde poeng. For hver av de 32 oppgavene har beste svar gitt 3 poeng. "Fullt hus" er altså 96 poeng. Av resultatene i tabellen ser vi at det er noen problemer knyttet til hvordan oppgavene har fungert. Vi vil peke på noen problematiske og noen positive trekk:

- Det er få blanke besvarelser. Elevene har svart på det aller meste og er tilsynelatende blitt motivert av oppgavene. Dette tyder også på at elevene har fått tilstrekkelig tid for prøven. Det kan selvsagt også ha vært en god del gjetting, noe som åpenbart er en god strategi ved tidsnød.
- I alt 13 av de 32 oppgavene er "flagget" for ett eller begge problemene a og b.
- For 11 av oppgavene er svaralternativene ikke ordnet etter forventet dyktighet. Og for flere av de andre oppgavene mangler det markerte forskjeller som skal til for at de forhåndsbestemte poengene skal få solid støtte i dataene. Vår bekymring når det gjelder svak validitet for slike "graderte" flervalgsoppgaver får her en støtte i dataene.
- Diskrimineringen er gjennomgående god, men for fem av oppgavene er den under 0,30, og spesielt ille er det for oppgavene D5 og D10.
- Oppgavene har gjennomgående svært lav vanskelighetsgrad, vurdert etter prosent av fullt hus (V i tabellen). Bare for to oppgaver er prosentandelen under 60%. Dette må imidlertid ses i lys av inflasjonen av poeng for dette oppgaveformatet.

Tabell 5.8: Item-analyse for leseoppgavene på 4. trinn (N=253).
Svarfordelingen i % og dyktigheten (gjennomsnittlig oppnådd poeng på hele prøven for de som har svart slik) er avrundet til hele tall. V angir vanskelighetsgrad, gjennomsnittlig poengsum i prosent av 3 poeng. D står for oppgavens diskriminering. I kolonnen for kommentarer (Komm) er det henvist til ulike fotnoter.

Oppg.	V	Svarfordeling i %					Dyktighet				D	Komm
		Blank	0p	1p	2p	3p	Blank / 0p	1p	2p	3p		
D1	96	0,4	0,4	1	7	91	64	30	65	75	,29	a, b
D2	89	0,4	1	8	15	76	64	68	64	76	,30	a
D3	74	0,4	6	4	53	37	56	61	73	79	,45	
D4	89	0	3	6	11	80	55	63	71	76	,35	
D5	79	1	4	7	35	53	64	69	73	76	,22	b
D6	89	1	1	6	17	76	50	58	68	77	,49	
D7	81	1	11	4	12	72	61	66	66	78	,44	a
D8	69	2	23	5	8	62	66	67	73	78	,40	
D9	79	2	10	6	17	66	60	64	71	78	,45	
D10	60	3	22	19	6	50	70	75	75	75	,15	a, b
D11	87	2	3	9	8	79	51	61	65	77	,54	
D12	81	2	2	13	16	66	49	64	70	78	,55	
D13	77	3	6	2	36	52	59	54	74	77	,40	a
D14	79	3	8	12	7	70	59	61	71	79	,58	
D15	87	3	2	8	4	82	51	64	58	77	,52	
D16	60	3	2	30	42	22	57	72	75	78	,29	b
E1	89	0,4	2	3	22	73	56	66	65	77	,40	a
E2	66	1	15	13	27	44	64	71	73	79	,39	
E3	84	0	9	6	12	74	60	60	65	78	,50	a
E4	66	1	13	21	17	48	66	69	71	79	,38	
E5	82	1	8	8	10	73	55	63	67	78	,58	
E6	56	3	17	32	7	41	66	71	72	80	,41	
E7	88	2	1	12	3	82	55	62	59	77	,48	a
E8	41	2	26	35	22	15	67	75	78	77	,29	a, b
E9	69	4	14	12	16	54	65	68	69	79	,45	
E10	78	2	13	6	9	70	61	60	63	79	,57	a
E11	66	2	10	24	19	45	60	70	73	80	,48	
E12	76	4	13	4	14	65	59	62	67	80	,62	
E13	80	4	6	9	10	70	62	65	72	77	,41	
E14	73	4	13	6	22	56	60	59	73	79	,56	a
E15	86	4	4	6	9	78	53	62	67	78	,56	
E16	86	4	5	5	8	79	54	66	71	77	,51	

a. Svaralternativene ikke ordnet etter dyktighet
b. Svak diskriminering (D<0,30)

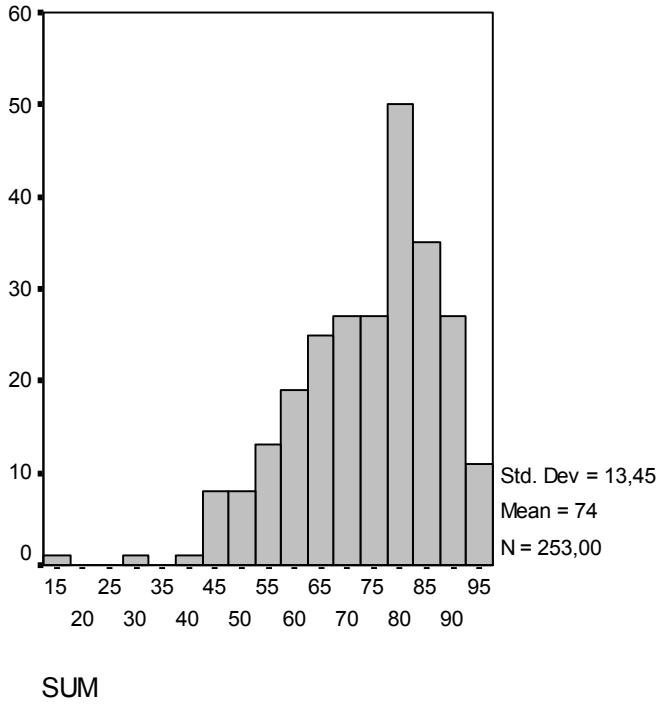
5.3.5 Resultater for hver del og samlet

Tabell 5.9 inneholder data om de to delene av prøven samt for prøven som helhet. Histogram for poeng for prøven som helhet er vist i figur 5.6.

Tabell 5.9: Data for hver av de foreslåtte delene av prøven i 4. trinn (N=253)

Kategori	Antall oppgaver	Gjennomsn. korrelasjon mellom oppgavene	Reliabilitet	Gjennomsn. skåre
Ordkjede	1	-	-	41 %
D (Tekst 1)	16	0,16	0,74	80 %
E (Tekst 2)	16	0,22	0,81	74 %
Totalt	32	0,17	0,86	77 %

Figur 5.6: Histogram som viser fordelingen av poeng på prøven som helhet (N=253)



Våre kommentarer til disse resultatene er:

- Prøven har vært lett, målt etter skåre i forhold til fullt oppnåelig poengsum (96 poeng). Særlig framstår den første delen som svært lett. Imidlertid vil en tilfeldig avkryssing gi gjennomsnittlig 1,5 poeng per oppgave, altså omtrent 48 poeng, som vi kan oppfatte som et slags "egentlig" nullpunkt. Sett i et slikt perspektiv har prøven rimelig vanskelighetsgrad. Problemet med et slikt resonnement er at noen elever som har skåret svært lavt, kunne med fordel ha brukt en slik strategi og tjent mye på det. Dette illustrerer nettopp hva vi er bekymret for, at vilkårlige gjettestrategier gjør sitt inntog i norsk skole.
- Sammenliknet med fjorårets prøve for 4. trinn er reliabiliteten denne gangen lavere (0,86 mot 0,88 forrige gang) på tross av at høyere reliabilitet rimeligvis har vært en viktig grunn for valget av graderte svaralternativer.

- Det er for lav reliabilitet for å rapportere for hver tekst for seg, og vi konstaterer derfor at man bare kan ha én skala. Innbyrdes korrelasjon mellom oppgavene på den første delen er lav, noe som har gitt altfor lav reliabilitet for denne delen.
- En annen sak er at korrelasjonen mellom de to delene (0,59) er betydelig lavere enn den kan bli, gitt reliabiliteten til hver del. Ut fra en latent korrelasjon på 0,77 (se kap. 4.5) kan vi konstatere at de to delene framstår som at de måler to ulike kompetanser. Men det er ikke lett å beskrive hva denne forskjellen går ut på, ut over at tekstene altså er forskjellige.
- Ordkjedeprøven står i et uklart forhold til resten av prøven, og den korrelerer middels (0,55) med skåre på den egentlige leseprøven. Mye tyder på at denne delen for noen elever kan fungere som en slags forklaringsvariabel for den målte leseforståelsen, og derved gi en nøkkel til å diagnostisere svake avkodere. Men gitt at det tar betydelig tid å vurdere og å telle opp antall riktige ord, kan det reises spørsmål ved om denne tiden er riktig anvendelse av tid i denne nasjonale prøvesammenhengen.

5.4 Konklusjon for lesing på 4. og 7. trinn

Det anvendte oppgaveformatet synes ikke å være egnet, idet både reliabiliteten og validiteten til prøven er påviselig en god del dårligere enn i 2004. Indre konsistens (gjennomsnittlig korrelasjon) for oppgavene er påtakelig lavere enn for prøven på 10. trinn og grunnkurs. I tillegg til dette har vi ved finlesing av oppgavene påvist betydelige validitetsproblemer, særlig knyttet til at det gis poeng for det vi oppfatter som definitivt gale svar.

I fjorårets vurderingsrapport (*Nasjonale prøver på prøve*) sto det:

> *"*Det har vært en del kritikk i media når det gjelder uklarheter for årets prøve. Særlig har det vært påpekt at flere distraktorer inneholder for mye som er delvis riktig. Vi støtter en slik kritikk, og vil framholde at dette bør unngås neste år. Faggruppas svar på dette er å foreslå å bruke gradert poenggiving for flervalgsoppgaver i år, altså at de ulike (ikke riktige) distraktorene kan belønnes med poeng som delvis riktig. Vi vil her advare mot en slik strategi, da den er veldig krevende å forsvare psykometrisk, idet man for hver eneste oppgave må påvise at det er empirisk dekning for den gitte poengsettingen. Med så strenge krav til hva som må fungere tilfredsstillende vil det trolig være altfor mange oppgaver som må kasseres etter utprøving. Et annet stort problem med det foreslåtte oppgaveformatet er at ren gjetting vil lønne seg i en helt urimelig grad. Dersom tre av fire alternativer gir "gevinst", er det jo opplagt at det lønner seg å gjette blindt selv om man ikke har noen som helst idé om hva som er det beste svaret. Det ligger altså gode begrunnelser bak den sterke internasjonale tradisjonen med at det bare er ett alternativ som er "riktig" og som derfor gir poeng." (s. 36)

Vi må bare konstatere at det innstendige rådet på dette punktet ikke har vært fulgt. Med lavere validitet, har vi også vanskelig for å forstå hvordan årets endringer i oppgaveformatet skal kunne gi prøvene "økt pedagogisk verdi" slik den gode ambisjonen har vært.

Det er bare aktuelt å rapportere resultatene etter én skala. Under noe tvil konkluderer vi med at denne kan betraktes som et generelt mål på elevenes lesekompetanse på det aktuelle trinnet. Men vi stiller oss altså undrende til mange av oppgavene, som etter vår vurdering ikke har høy validitet i forhold til en slik målsetting.

Vi mener man ikke bør fortsette å lage graderte flervalgsoppgaver. Videre mener vi at man bør tenke gjennom på nytt hva man vinner med å velge to eller tre lange tekster fra samme emneområdet i slike nasjonale prøver. Vi tror prøvene vil gi et bedre mål på lesekompetanse om man erstatter i hvert fall én av de lange tekstene med noen korte. Slik kunne også emnene bli flere.

Vi foreslår at det ved målrettet utprøving systematisk velges ut oppgaver med høy diskriminering for å legges større vekt på å få høy reliabilitet. Dersom dette blir gjennomført, kan det hende at det blir forsvarlig å rapportere etter to skalaer, selv om det kan være vanskelig med et så lavt antall oppgaver. Det må i så fall vurderes grundig hvilke kategorier det skal være. Å dele inn etter de to tekstsjangerne vil være svært tvilsomt, da det vil være vanskelig å generalisere til en ”tekstsjangerkompetanse” på basis av bare én tekst.

Som beskrevet for leseprøven på 10. trinn vil vi foreslå at det utarbeides et grundig rammeverk for nasjonale prøvene i lesing på alle de aktuelle trinnene som samtidig ivaretar kompetansen som skal oppnås på de ulike trinnene (se omtalen i 5.1). I et slikt arbeid vil det blant annet være naturlig å diskutere hvilken plass en ordkjedeprøve eller andre mål for ordavkoding eller lesehastighet skal ha på. Per i dag har ordkjedeprøven en noe uavklart status og pedagogisk nytte. Det kan synes som den tar forholdsvis mye tid å vurdere. Derfor må det vurderes om en slik oppgave er fornuftig bruk av den hardt tiltrengte tiden for elevene under prøven på 4. og 7. trinn og for lærerne ved vurdering. Det er mye som tyder på at noen elever har fått for lang tid på denne delen av prøven, noe som i så fall svekker dens mulige pedagogiske verdi som forklaring for resultater på den egentlige leseprøven..

6 Skriving

Det var skriveprøver på 4., 7. og 10. trinn samt grunnkurs. Prøven for grunnkurs var frivillig og ble bare besvart av omtrent 4 % av elevene, så vi ser bort fra denne prøven her. Skriving i 10. trinn

6.1.1 Beskrivelse av prøven

Prøven består av to oppgaver, som hver går ut på at elevene skal skrive en tekst om astronomiske og mer filosofiske spørsmål. Det faglige området for oppgavene dette første året er valgt til å være naturfag, eller mer presist astronomi, og spesielt solsystemet. Begge oppgavene er tilpasset et informasjonshefte knyttet til planeten Mars, som elever og lærere har fått til bruk ved en grundig forberedelse til prøven. Det har altså vært vesentlig at elevene har hatt et visst minimum av felles faglig bakgrunn for å besvare oppgavene.

Faggruppa har for øvrig gjort grundig rede for de faglige og praktiske avveiningene som har ledet utviklingen fram mot den nasjonale prøven i skriving. (Kjell Lars Berge har på vegne av faggruppa gitt en sammenhengende framstilling av dette i sitt innlegg i konferanserapporten "*Nasjonale prøver – veivalg og utviklingsmuligheter*", Høgskolen i Lillehammer.) Det er også viktig å peke på at faggruppa ser prøven, som de kaller "den sentralt gitte prøva", som del 1 av en helhetlig vurdering av elevenes skrivekompetanse. "Elevens tekstsamling" av tekster i ulike sjangrer og emner skal utgjøre del 2 i denne helheten. Vi understreker at vi her bare vurderer hvordan "den sentralt gitte prøva" har fungert.

Den første oppgaven ber om en saklig utredning om hva som taler for og hva imot at det er eller har vært liv på Mars:

> "Utforskingen av planeten Mars reiser på nytt spørsmål om det kan ha vært liv på planeten, og om det fortsatt kan være liv der.
>
> Skriv en fagtekst der du gjør greie for hva som taler for, og hva som taler imot at det er eller kan ha vært liv på planeten. Gi din vurdering av hva som virker mest sannsynlig. Lag en tittel til teksten din."

Den andre oppgaven ber om intet mindre enn en refleksjon rundt ett av de "store" spørsmål, meningen med livet, etikk og estetikk, hvem er jeg, osv.

> "Til alle tider har stjernene, planetene og verdensrommet inspirert mennesker til å tenke omkring "de store spørsmålene" og det å være menneske:
>
> > Hvem er jeg, og hvorfor er jeg her? Hva er meningen med livet? Hva kan vi vite sikkert egentlig? Hva er rett, og hva er galt? Hva er vakkert, og hva er stygt? Moder jord – hva er det vi gjør med vår blå-grønne planet?
>
> Dette er bare noen av de evige spørsmål som himmelhvelvingen kan få en til å fundere over. Kanskje har du fundert over slike spørsmål en gang i blant- kanskje i forlengelsen av tanker og ideer du er blitt kjent med gjennom KRL-faget?

> Kanskje har du egne erfaringer som gjør at disse eller andre spørsmål er blitt viktige for deg? Skriv en reflekterende tekst der du deler med andre mennesker dine tanker omkring ett av "de store spørsmålene". Gi teksten din en tittel."

De to oppgavene skal vurderes etter en spesifisert skala for mange ulike delkompetanser. Felles er at skalaen har tre nivåer etter hva som kan regnes som normen på 10. trinn:

1 = høyere enn forventet
2 = omtrent som forventet (dette nivået er ment å favne de aller fleste)
3 = lavere enn forventet

Det dreier seg altså ikke om en holistisk vurdering, men gjennom en analytisk tilnærming å vurdere hver elevs kompetanse på hvert område (hver "dimensjon"). De ulike områdene som skal vurderes, er disse:

1. Kommunikasjon
2. Innhold
3. Tekstoppbygging
4. Språkbruk
5. Rettskriving og tegnsetting
6. Bruk av skriftmediet

Hver av disse kompetansene vurderes for de to oppgavene hver for seg, og noen av kompetansene deles igjen opp i to underkompetanser (mer om dette senere). Berge (op cit.) sier det slik:

> "For vurdererne vil det være uvant og krevende å følge denne metoden. Hver tekst må leses flere ganger, og tekstene må studeres og vurderes for hver dimensjon."

Det er altså åpenbart at vurderingen av elevenes besvarelser innebærer mye arbeid for lærernes del. Det er like åpenbart at forberedelsene til prøven innebærer en betydelig allokering av oppmerksomhet mot skriveprosessen og ikke minst det aktuelle faglige emnet i tiden før prøven holdes.

6.1.2 Validitet

Faggruppa som har arbeidet med skriveprøvene, har utarbeidet en meget grundig rammeverk der bakgrunnen for prøven er gjort rede for. Vi vil anbefale Berges ovenfor nevnte innlegg som en orientering om dette.

Etter vårt skjønn representerer faggruppas arbeid med forberedelsene til prøvene en betydelig og viktig teoretisk bakgrunn for fremme av skriveopplæring. Den teoretiske avklaringen av hva skrivekompetanse er og hvordan den kan måles, utgjør en svært god forutsetning for å fremme bedre skriveopplæring i norsk skole. Imidlertid er det vår oppgave her å vurdere hvordan oppgavene har fungert i praksis, herunder særlig hvordan besvarelsene faktisk er blitt vurdert, og hvordan resultatet av vurderingen gjenspeiles i pålitelige skalaer.

Vi finner det ikke naturlig å overprøve faggruppas utredning, men vi vil bruke deres egne betraktninger i vår analyse av hvordan *resultatene* for skriveprøven forholder seg angående kvalitetsmessige krav.

Vårt første punkt er å peke på det faggruppa selv sier (Berge, op. cit.) om at skrivekompetanse bare kan vurderes på en meningsfull måte hvis:

> ”Alle elever må skrive flere tekster innenfor flere ulike skrivemåter, genrer og skrivesituasjoner. Dessuten må hver enkelt av elevens tekster vurderes flere ganger av forskjellige og uavhengige bedømmere. Bare på den måten kan vi danne oss et bilde av skriveferdighetene til en elev som er gyldig når elevens skriveutvikling skal stimuleres i skolen.”

Vi må konstatere at den nasjonale prøven i skriving består av en skrivesituasjon og to tekster i forskjellige sjangrer. Disse er i all hovedsak vurdert av én sensor, nemlig læreren selv. De elevbesvarelsene som er sendt inn til (dobbelt) uavhengig vurdering, danner et unntak, men det er et svært lite mindretall. Våre data er imidlertid hentet fra disse besvarelsene. Målt med faggruppas egne kriterier må vi derfor konstatere at resultatene fra skriveprøva, slik disse er rapportert inn fra hver skole til Utdanningsdirektoratet, isolert sett ikke kan anses som et valid uttrykk for elevenes skrivekompetanse. De ideelle kravene til gjennomføringen av prøver som faggruppa har forutsatt, har ikke vært forsøkt realisert.

6.1.3 Analyse av resultater for 10. trinn

For hver av kompetansene er det i tabell 6.1 gjengitt hvordan elevenes prestasjoner fra 3 (lavest) til 1 (høyest) fordeler seg prosentvis. Det framgår av tabellen at det (svært) store flertall av elevene vurderes til å ligge på det midterste nivået. Dette er kanskje omtrent slik man måtte forvente ut fra retningslinjene til vurderingen, men likevel blir det med en slik fordeling vanskelig å få god spredning blant elevene.

Det bør bemerkes at fordelingen ovenfor gjelder vurdering nr. 1. Vurdering nr. 2 er omtrent lik, men den er av en eller annen grunn systematisk litt mer samlet om verdien 2. Vi legger ikke noe avgjørende vekt på denne forskjellen her. Derimot vil vi nevne at vi ikke har brukt skolenes egne vurderinger i denne analysen.

Forskjellen mellom de to vurderingene står imidlertid i fokus i annen del av tabell 6.1 når det gjelder graden av overensstemmelse. Dette er gitt både i form av prosent av elever der de to sensorene er enig (R), og også som koeffisienten *kappa* (K, se om dette i kap. 4.4). Det framgår av resultatene at mange av kompetansene gir dårlig overensstemmelse. Selv om R er rimelig høy, så er altså K svært lav. Dette kommer av at det høye prosentvise samsvaret er en litt kunstig effekt av at de aller fleste havner på ”middels” nivå. Vi må derfor konstatere at det nokså klart ikke er etablert et godt nok tolkningsfellesskap for vurdering av skriving etter disse kriteriene.

Tabell 6.1: Resultater for skriveprøven på 10. trinn. Prosentvis fordeling av prestasjoner (1 er best) samt to ulike mål for overensstemmelse i vurderingen. (N=512)

Kategori	**Prosentvis fordeling**			**Overensstemmelse**	
Oppgave 1	**1 (høy)**	**2**	**3 (lav)**	**R**	**K**
0 Førsteinntrykk	11	79	11	76	,28
1 Kommunikasjon	13	80	8	76	,32
2 Innhold	12	75	14	74	**,37**
3 Tekstoppbygging	13	76	11	76	,35
3 Tekstbinding	10	80	10	79	,33
4 Språkbruk	7	87	7	82	,28
5 Rettskriving	6	84	10	82	,21
5 Tegnsetting	5	86	9	83	,21
6 Grafisk utforming	2	93	4	91	*
6 Håndskrift	4	87	9	84	**,11**
Oppgave 2					
0 Førsteinntrykk	11	77	12	77	,34
1 Kommunikasjon	12	81	8	74	,22
2 Innhold, relevans	8	81	11	76	,31
2 Innhold, konkret	10	79	11	76	,31
3 Tekstoppbygging	8	81	11	79	,29
3 Tekstbinding	7	83	10	81	,30
4 Språkbruk	4	88	8	86	,29
5 Rettskriving	5	85	9	83	,27
5 Tegnsetting	4	89	7	86	,21
6 Grafisk utforming	3	92	5	89	,11
6 Håndskrift	5	87	9	84	,16

*Kappa kan ikke beregnes, siden bare den ene sensoren har gitt (til i alt 11 elever) "karakteren" 1

Som en illustrasjon på hva slags overensstemmelse som ligger bak disse tallene, har vi vist to eksempler på krysstabeller for de to vurderingene, for henholdsvis oppgave 1, vurdering av innhold og av håndskrift (begge to uthevet i tabell 6.1).

Tabell 6.2: Krysstabell mellom de to sensorene for oppgave 1, Innhold (N=496)

		Sensor 1			
		1 (høy)	**2**	**3 (lav)**	**SUM**
Sensor 2	**1 (høy)**	**26**	34		60
	2	31	**306**	34	371
	3 (lav)		30	**35**	65
	SUM	57	370	69	496

Tabell 6.3: Krysstabell mellom de to sensorene for oppgave 1, Håndskrift (N=496)

		Sensor 1			
		1 (høy)	**2**	**3 (lav)**	**SUM**
Sensor 2	**1 (høy)**		18		18
	2	12	**397**	11	420
	3 (lav)		36	**8**	44
	SUM				

Tabell 6.2 illustrerer den beste overensstemmelsen (K = 0,37). Det er en tydelig konsentrasjon ev besvarelser langs diagonalen (samme vurdering, uthevet i tabellen), men likevel er det en betydelig uenighet om hvem som fortjener vurderingene 1 og 3. Bare omtrent 1/3 av de som har fått slik vurdering, har fått den av begge sensorene.

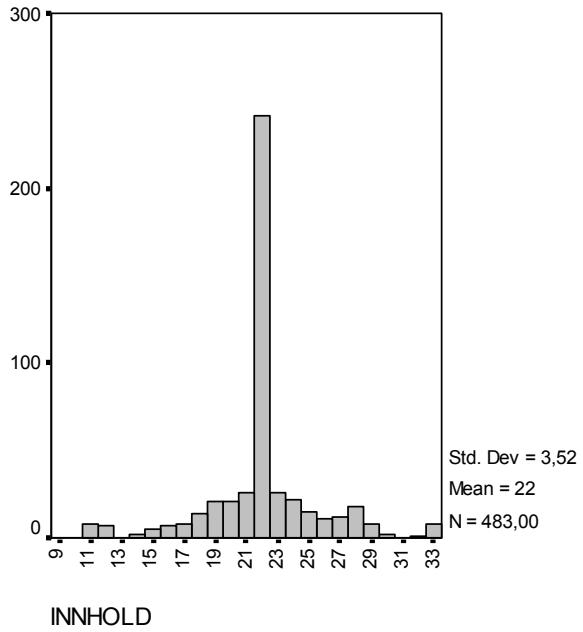
Den andre ytterligheten er representert i tabell 6.3 (K=0,11), som gjelder vurdering av håndskrift. Her ser vi at de to sensorene er nesten ikke enige om noen av de som er gitt vurderingene 1 eller 3.

6.1.4 Hvor forskjellige er kompetansene?

En faktoranalyse (se kap. 4.5) med to faktorer gir en tydelig og forståelig tendens i dataene. I den ene faktoren samles kompetansene 1-4 (i alt 11 variable, se tabell 6.1), mens kompetansene 5 og 6 (i alt åtte variable, se tabell 6.1) samler seg i den andre faktoren. Det er altså en struktur i dataene som separerer innholdsdimensjonen (kommunikasjon, innhold og tekstoppbygging) fra de formelle ferdighetene (rettskriving, tegnsetting og skriftforming). Selv om en slik todeling ikke er foreslått av faggruppa, kan vi likevel se hvordan dataene er i forhold til dette perspektivet. Hvis vi i stedet bruker tre faktorer, splittes den første faktoren ovenfor i to faktorer, en for hver av de to oppgavene.

Det er ut fra dette interessant å se hvordan førsteinntrykket korrelerer med hver av de spesifikke aspektene. Det viser seg at denne korrelasjonen ligger mellom 0,58 og 0,71 for kompetansene 1-4, mens den er så lav som mellom 0,27 og 0,42 for kompetanse 5 og 6. Av dette framgår at førsteinntrykket tydeligvis i mye større grad formes av innholdsdimensjonen enn av de formelle skriveferdighetene. Denne forskjellen er interessant, men ikke uventet.

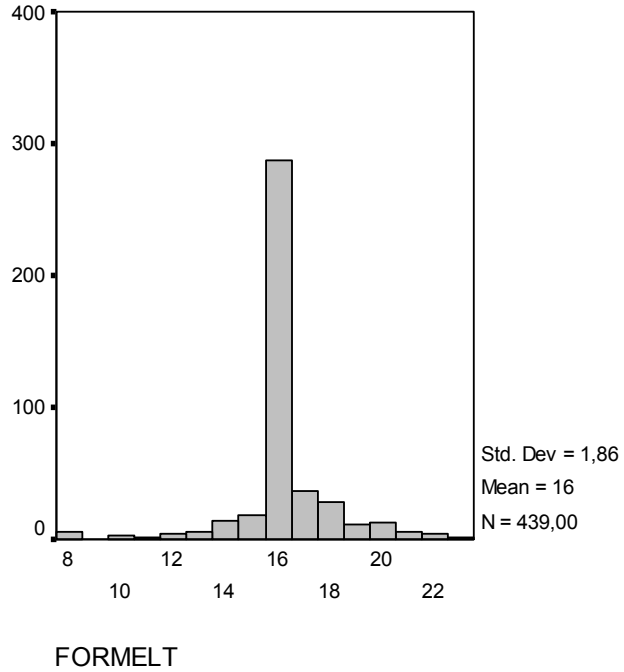
Figur 6.1: Fordeling av poeng etter en tenkt "summering" av delferdigheter innen innholdsdimensjonen (N=512)



Vi tenker oss at vi oppfatter kodene (1, 2 eller 3) som poenger og summerer disse for hver av de to faktorene. Fordeling av "poenger" blir da som vist på figurene 6.1 og 6.2.

Her må vi huske på at lave tall representerer høy kompetanse. Det framgår tydelig at diagrammet domineres av det store flertallet som har fått 2 (”Omtrent som forventet”) på *alle* variablene. Å lage en overordnet sum for hele prøven lar seg selvsagt gjøre, men blir ikke særlig meningsfullt på basis av de resultatene vi har sett hittil, og er heller ikke i tråd men prøvens intensjoner.

Figur 6.2 Fordeling av poeng etter en tenkt ”summering” av delferdigheter innen den formelle dimensjonen (N=512)



La oss nå se bort fra de store uoverensstemmelsene mellom ulike sensorer og studere hvor forskjellige disse to faktorene er. Vi ser da bare på data fra den ene sensoren og regner ut en reliabilitetskoeffisient (alfa) for de to faktorene på henholdsvis 0,91 og 0,83. Korrelasjonen mellom de to faktorene er 0,60, og fra dette kan vi regne ut den latente korrelasjonen (se kap. 4.5) til å bli 0,69. Siden dette tallet er langt under 1,0, er det et tydelig tegn på at de to faktorene virkelig framstår som forskjellige, og at en rapportering av disse to faktorene i og for seg ville være meningsfull hvis ikke uoverensstemmelsene mellom vurdererne hadde vært så altfor stor.

6.2 Skriving på 7. trinn

Prøven for 7. trinn er laget over samme lest som prøven for 10. trinn. Også faglig sett er det store likheter, idet et hefte om planeten Mars også her var det tematiske utgangspunktet for skrivingen.

Oppgavene 1 lød slik:

> *Beskriv planeten Mars* for en som ikke vet noe særlig om Mars fra før, slik at de kan lære om planeten. Om du vil, kan du lage en illustrasjon til det du har skrevet. Tittel lager du også selv.
>
> Bruk gjerne stoff fra notatarket ditt når du skriver, men husk at du skal velge ut den informasjonen som trengs for å beskrive planeten, og at du skal skrive din egen tekst nå i dag. Det som er viktig i denne oppgaven, er at du viser at du kan skrive en beskrivende fagtekst med sammenheng. Det vil blant annet si at du velger ut viktig informasjon, og ordner den på en god måte. Kan hende står det en del på notatarket ditt som du ikke har bruk for i teksten din.

Vi merker oss de klare og støttende anvisningene angående premissene for skrivingen, og videre de konkrete henvisningene til notatene fra det innledende arbeidet med temaheftet om Mars.

Oppgave 2 inviterte til å lage en fortelling fra Mars. Igjen merker vi oss hvordan elevene får hjelp i å tolke hva oppgaven går ut på:

> Tenk deg mange år framover i tid. Menneskene bor på Mars, og alt er svært annerledes enn på jorda. Hva kan hende?
>
> *Skriv en fortelling som foregår på Mars.* Du kan gjerne bruke både fantasien din og notatarket ditt når du skriver. Lag tittel selv. Om du vil, kan du lage en illustrasjon til det du har skrevet.

Som det framgår av de to oppgavene, er strukturen for prøven omtrent identisk med prøven for 10. trinn. Men oppgave 2 har her ikke noe av det filosofiske utgangspunktet som oppgave 2 hadde for de eldre elevene (se kap. 6.1.1).

Tabell 6.3 sammenfatter resultatene fra 7. trinn. I hovedsak er kategoriene de samme som for 10. trinn, men det er noen små forskjeller. Det viktige for oss her er det første tallet (1-6), som indikerer hvilken av de seks kategoriene det dreier seg om (se kap. 6.1.1). Uten å gå i detalj her, kan vi si fra tabellen at konklusjonene blir omtrent de samme som på 10. trinn, selv om det er noen små forskjeller.

Hvis vi tar gjennomsnittet av alle kategoriene, finner vi denne fordelingen mellom nivå 1 (høyest), 2 og 3 (lavest), der vi i parentes har vist de tilsvarende tallene for 10. trinn:

> Nivå 1: 6 % (8 %), Nivå 2: 73 % (83 %), Nivå 3: 20 % (9 %)

Det er altså en betydelig forskyvning mot svakere resultater for de yngre elevene. Eller vi kan kanskje like gjerne si at det er en forskyvning i retning mot ”lavere enn forventet”. Dette kan tolkes på flere måter, men det er ikke vesentlig for oss her. Viktigere er det at færre elever havner i midtsjiktet, og følgelig inneholder dataene mer informasjon om elevenes kompetanse, fordi spredningen er større.

Tabell 6.3 Resultater for skriveprøven på 7. trinn. Prosentvis fordeling av prestasjoner (1 er best) samt to ulike mål for overensstemmelse i vurderingen. (N=603)

Kategori	Prosentvis fordeling			Overensstemmelse	
Oppgave 1	**1 (høy)**	**2**	**3 (lav)**	**R**	**K**
0 Førsteinntrykk	8	73	19	72	,39
1 Kommunikasjon	11	71	19	64	,22
2 Innhold, relevans	11	80	10	72	,26
2 Innhold, utdyping	11	66	23	61	,25
3 Tekstoppbygging	8	72	20	65	,26
3 Tekstbinding	9	66	25	66	,30
4 Språkbruk	5	75	20	71	,25
5 Rettskriving	8	77	15	69	,19
5 Tegnsetting	5	77	15	71	,25
6 Grafisk utforming	10	78	12	70	,27
6 Håndskrift	2	83	15	79	,34
Oppgave 2					
0 Førsteinntrykk	5	69	26	75	,46
1 Kommunikasjon	6	67	27	63	,24
2 Innhold, fiksjon	4	74	22	67	,23
2 Innhold, utdyping	4	70	26	67	,26
3 Tekstoppbygging	4	73	22	64	,18
3 Tekstbinding	5	73	22	68	,27
4 Språkbruk	4	74	22	69	,24
5 Rettskriving	6	73	21	67	,20
5 Tegnsetting	6	64	30	70	,37
6 Grafisk utforming	5	78	16	74	,20
6 Håndskrift	1	79	20	76	,16

Hvis vi ser på overensstemmelsen mellom de to vurderingene, er det gjennomgående en betydelig lavere prosentandel av besvarelsene der det er full enighet: Gjennomsnittlig 69% på 7. trinn, mot 81% på 10. trinn. Men denne lavere overensstemmelsen skyldes bare at det er færre elever på det midterste nivået, for den gjennomsnittlige verdien for K (kappa, "coefficient of agreement") er den samme, nemlig 0,26 på begge klassetrinnene. Dette tallet er lavt og fører til at vi må fraråde enhver publikasjon av elev- og skoleresultater fra denne prøven.

6.3 Skriving på 4. trinn

Uten å diskutere innholdet nøyer vi oss for denne prøven å konstatere at resultatene ifølge tabell 6.4 er omtrent som for de andre trinnene. Overensstemmelsen mellom de to vurderingene er like lav her. Vi gjør oppmerksom på at antall elever (og dermed antall sensorer) her er lavt. Og med de store forskjellene det er fra sensor til sensor, er prosentfordelingen på nivåer beheftet med nokså store feilmarginer.

Tabell 6.4 Resultater for skriveprøven på 4. trinn. Prosentvis fordeling av prestasjoner ("karakteren" 1 er best) samt to ulike mål for overensstemmelse i vurderingen. (N=158)

Kategori	Prosentvis fordeling			Overensstemmelse	
Oppgave 1	**1 (høy)**	**2**	**3 (lav)**	**R**	**K**
0 Førsteinntrykk	7	66	26	66	,31
1 Kommunikasjon	6	75	19	63	,11
2 Innhold, relevans	8	82	10	73	,14
2 Innhold, utdyping	7	64	29	52	,10
3 Tekstoppbygging	8	54	38	60	,29
3 Tekstbinding	7	53	40	64	,34
4 Språkbruk	8	60	33	54	,16
5 Rettskriving	12	74	15	70	,20
5 Tegnsetting	6	80	14	74	,16
6 Grafisk utforming	5	90	5	80	,05
6 Håndskrift	4	62	34	59	,16
Oppgave 2					
0 Førsteinntrykk	3	63	35	66	,31
1 Kommunikasjon	4	75	21	60	,11
2 Innhold, passende	6	75	19	59	,03
2 Innhold, utdyping	5	63	31	60	,27
3 Tekstoppbygging	5	68	29	59	,19
3 Tekstbinding	8	54	39	59	,25
4 Språkbruk	8	62	30	52	,16
5 Rettskriving	7	73	20	70	,20
5 Tegnsetting	5	71	24	69	,21
6 Grafisk utforming	0	93	7	77	*
6 Håndskrift	1	61	38	65	,28

*Kappa kan ikke beregnes, siden bare den ene sensoren har gitt (til i alt 20 elever) "karakteren" 1

6.4 Konklusjon

Fra analysene av dataene, og faktisk også fra faggruppas egen tenkning, ser vi at skriveprøven egner seg lite til å lage pålitelige skåreverdier. I tråd med den teoretiske utredningen til faggruppa (Berge) slutter vi oss til at skriveferdigheter er mangfoldige, og å rapportere dette i form av én skåreverdi er ikke meningsfullt. Likevel vil vi peke på at det ut fra dataene ser ut til å være et tydelig skille mellom to aspekter av skrivekompetanse, en for innholdsdimensjonen og en for formelle ferdigheter. Men vi registrerer også at med den dårlige overensstemmelsen det er mellom de to uavhengige vurderingene, er enhver rapportering av skoleresultater i utgangspunktet uaktuell. Her vil vi igjen sitere faggruppa selv (Berge, op. cit.):

> "Store forskjeller i bedømmelsesmønsteret til ekspertvurdererne vil føre til at prøvene vil ha liten troverdighet, De vil ikke kunne fungere som vurderingsredskaper som på en pålitelig måte kan fortelle oss noe om elevens kompetanse i skriving."

Med den skalaen som er brukt for hver av variablene, havner de aller fleste elevene i den midterste kategorien, og skolers gjennomsnittsverdier blir da bestemt av de meget få ekstreme verdiene (1 og 3). Men som vi har sett, det er liten grad av enighet om *hvem* som skal få disse verdiene. Et naturlig spørsmål å stille er om situasjonen ville bli bedre

dersom den midtre kategorien ble gjort noe smalere. Da ville skalaen skilt bedre og det ville bli en bedre diskriminering.

Den store styrken til skriveprøven er at den representerer et pedagogisk verktøy, utviklet ved hjelp av en grundig teoretisk analyse av hva skrivekompetanse består av. Dette arbeidet ligger riktignok på et såpass høyt akademisk nivå at det vil være en viktig oppgave å formidle dette på en enklere måte til norsklærere i skolen. Gjennom dette utviklingsarbeidet har man fått et begrepsapparat og et viktig verktøy i arbeidet med å heve vurderingskompetansen i skriving i norsk skole.

Likevel er vi kritiske til at slike skriveprøver inngår som en del av de nasjonale prøvene. Resultatene egner seg ikke til å rapportere, og den diagnostiske verdien overfor elevene er begrenset idet de fleste elevene framstår som ”omtrent som forventet” på hver eneste av de mange variablene.

7 Engelsk

7.1 Engelsk skriving

7.1.1 Struktur og vurderingskriterier

Skriveprøvene, for 7. og 10. trinn samt grunnkurs, består av tre oppgaver av litt forskjellig karakter, men har det til felles at det dreier seg om "fri" skriving av en avgrenset tekst ut fra en lengdeangivelse i form av ett, et par eller flere avsnitt. Tidsrammen økes for hvert trinn: 60 minutter for 7. trinn, 90 minutter for 10. trinn og 120 minutter for grunnkurs. For 7. trinn er det flere fargeglade bilder som utgangspunkt for fri skriving, mens oppgavene for 10. trinn og grunnkurs er tekstbaserte. På de to øverste trinnene består prøven av tre ulike sjangere; personlig skriving, argumenterende tekst og fortelling/beskrivelse. Felles for alle trinn er at omtrent halvparten av tiden skal brukes på den siste oppgaven med beskrivelse/fortelling.

Elevbesvarelsene vurderes etter to vurderingsskjemaer, ett for *Formidling* og ett for *Språk*. De tre oppgavene vurderes hver for seg etter et oppgavespesifikt vurderingsskjema for aspektet *Formidling* og får i tillegg en "samlekarakter" for dette aspektet. Vurderingsskjemaet for *Språk* er ikke oppgavespesifikt, men dette aspektet vurderes på tvers av oppgaver i forhold til fire delferdigheter: *Tekststruktur*, *Grammatikk, Ord og uttrykk*, samt *Ortografi og tegnsetting*. Også aspektet *Språk* gis en samlet vurdering. I tillegg til de to overordnete vurderingene av henholdsvis *Formidling* og *Språk* får elevene til slutt en vurdering for den samlete prestasjonen, heretter kalt *Totalt*. Ved vurderingen av denne skal, i henhold til vurderingsveiledningen, de språklige ferdigheter telle mest.

Faggruppa i engelsk har tatt utgangspunkt i kompetansenivåene som er beskrevet i *Common European Framework of Reference for Languages: Learning, teaching and assessment* (heretter forkortet til *Rammeverket* og CEF-nivåer). I *Rammeverket* for ferdigheter i fremmedspråk finnes tre hovednivåer:

- Det laveste nivået *A- Basic user,* med oppdeling i *A1-Breakthrough* og *A2 Waystage.*
- Mellomnivået *B- Independent user,* med oppdeling i *B1-Threshold* og *B2-Vantage.*
- Det øverste nivået *C- Proficient user,* med oppdeling i *C1-Effective Operational Proficiency* og *C2- Mastery*. Det høyeste nivået C2 anses imidlertid som uaktuelt for norske skoleelever.

Faggruppa i engelsk har i tillegg til nivåene beskrevet ovenfor laget mellomnivåer for å få en mer differensiert vurdering av elevenes kompetanse. Med mellomnivåer består den aktuelle skalaen av nivåene A1, A1/A2, A2, A2/B1, B1, B1/B2, B2, B2/C1 og C1. Denne nivåskalaen er den samme som ble benyttet ved fjorårets prøver. B1-nivået, det såkalte terskelnivået, regnes som det nivået man bør oppnå for å kunne fungere sosialt og samfunnsmessig på et fremmedspråk. Det er utviklet terskelnivåbeskrivelser for de fleste europeiske språk.

Selv om skalaen altså er uendret, innebærer den ovenfor nevnte organiseringen med fire kategorier for *Språk* og oppgaverelaterte kriterier for *Formidling* for hver av de tre oppgavene, en vesentlig endring fra året før.

Et hovedpoeng med en slik skala, er at den er ment å representere en *absolutt* skala for kompetanse. Ved å anvende denne skalaen, kan vi oppnå at vurderingen i engelsk blir kriterierelatert. En åpenbar fordel er at skalaen kan gjøre det mulig på en enkel måte å studere enkeltelevers og landsgjennomsnittets læringskurve fra år til år, samt å studere om elever på et bestemt klassetrinn gjør fram- eller tilbakegang over tid. En forutsetning er imidlertid at beskrivelsen av nivåene er de samme. Ikke minst kreves det at vurderingen av besvarelsene kan gjøres pålitelig med en slik skala.

7.1.2 Vurdering av prøvenes validitet

Faggruppa i engelsk argumenterer i sin kommentar til fjorårets vurderingsrapport *Nasjonale prøver på prøve* sterkt for at prøven i skriftlig engelsk må ha positiv tilbakevirkningseffekt, og at siden fri skriftlig produksjon er sentral i en validitetstenkning, må man akseptere en lavere grad av reliabilitet. De ønsker videre at engelskprøvene også bør vurderes validitetsmessig i større grad enn det som ble gjort for 2004, og de lister opp sentrale områder som de gjerne ser vektlagt i en slik vurdering:

- Er det godt samsvar mellom oppgaver/vurderingskriterier og det vi ønsker prøven skal måle?
- Er vår tilpassing til CEF-skalaene holdbar?
- Legger sensorene vekt på de riktige trekkene når de vurderer?
- Er oppgavene passe vanskelige?
- Er oppgavene og vurderingskriteriene slik at man kan forvente en positiv tilbakevirkningseffekt på undervisningen?

Vi tar utgangspunkt i faggruppas egen definisjon av *Communicative Language Ability* og hva som er en god prøve (Vurderingsveiledning for 10. trinn 2005) og vil vurdere om prøvene virkelig måler den type språkferdighet som de intenderer å måle. *Communicative Language Ability* (CLA) defineres relativt likt med definisjonen av kommunikativ kompetanse som ligger til grunn i de eksisterende norske læreplanene (L97). CLA inneholder komponentene

- microlinguistic ability (lingvistisk kompetanse)
- textual ability (diskurskompetanse),
- pragmatic ability (pragmatisk eller sosiolingvistisk kompetanse)
- strategic ability (strategikompetanse).

Strategikompetanse testes ikke direkte i prøvene, og den pragmatiske kompetansen er noe uventet lagt inn under kategorien *Ord og uttrykk*. Den nasjonale prøven i engelsk er, som faggruppa understreker i veiledningsmaterialet, i hovedsak en ferdighetsprøve, og dette perspektivet gjenfinnes i vurderingsveiledningen der det framgår at kategorien *Språk* skal vektlegges mest for den totale vurderingen.

Et problem med vurderingen av engelsk skriftlig er knyttet til innføringen av CEF, som fortsatt er ukjent for de fleste norske engelsklærere. Da mange lærere følger elevene sine over flere år, er det grunn til å regne med at vurderingen av elevers skriftlige besvarelser i 2005 antakeligvis også er utført av andre lærere enn i 2004. Det vil med andre ord ta tid å få innført et tolkningsfelleskap. Begrunnelsen for å bruke nivåbeskrivelser fra CEF er ifølge faggruppas egne utsagn i vurderingsveiledningen for 7. klasse 2004/2005 som følger:

- Elevene får konkrete beskrivelser av hva de kan på alle trinn, og kan se hva de trenger å beherske for å utvikle seg videre.
- Eleven får en profil over ferdigheter, slik at deres sterke og svake sider synliggjøres. Elevene blir da i stand til å se hva de trenger å jobbe med.
- Fremskrittene elevene gjør når de beveger seg oppover i nivåene, er tydelige og motiverende både for sterke og svake elever.

Vurderingsskjemaet for *Språk* er ikke oppgavespesifikt. Idet CEF-nivåene er faste nivåbeskrivelser, dvs. en absolutt skala for språkferdigheter, var det forventet at de språkspesifikke kriteriene for de ulike trinn og nivåer var nøyaktig like. Men dette er ikke tilfelle. Snarere ser det ut til at beskrivelsene er trinndifferensiert. *Rammeverket* har følgende definisjon for C1-nivået for grammatisk korrekthet: ”*Consistently maintains a high degree of grammatical accuracy; errors are rare and difficult to spot*.”

I vurderingsskjemaet for 7. trinn er B2-nivået det høyeste nivået som er brukt, og nivåbeskrivelsen for *Grammatikk* for dette nivået er gitt slik:

> ”**Korrekthet:** Høy grad av grammatisk kontroll. Avvik kan forekomme i forbindelse med komplekse strukturer”.

Nivåbeskrivelsen for grammatisk korrekthet minner her mer om beskrivelsen for C1 enn for B2. I *Rammeverket* er det to nivåer for B2, grammatisk korrekthet, og det laveste B2-nivået burde vært brukt her, idet nivåbeskrivelsen for B2 også inkluderer delvis det lavere nivå B1/B2:*”Shows a relatively high degree of grammatical control. Does not make mistakes which lead to misunderstanding.”*

For 10 trinn på nivå C1 og over er nivåbeskrivelsen for *Grammatikk* slik:

> **”Korrekthet:** En meget høy grad av grammatisk korrekthet.”

For grunnkurset er imidlertid den tilsvarende beskrivelsen slik:

> **”Korrekthet:** Høy grad av grammatisk kontroll. Avvik er sjeldne og vanskelig å oppdage.”

”Meget høy grad” for 10. trinn og ”Høy grad” av grammatisk korrekthet for grunnkurset virker ulogisk, og det er vanskelig å forstå hvorfor trinnene ikke har identisk formulering. Denne type relativ nivåbeskrivelse går igjen i flere punkter og det virker unektelig problematisk at for eksempel både 7. trinn og grunnkurs skal avkreves høy grad av grammatisk korrekthet, mens høy grad av grammatisk korrekthet på 7. trinn honoreres med kun B2.

Det er vanskelig å forstå denne relativiseringen av det som gir seg ut for å være en absolutt ferdighetsskala. Etter vårt skjønn svekker dette også på en alvorlig måte vurderingskriterienes pålitelighet og særlig mulighet for å måle absolutt framgang.

Oppgavene for 10. trinn og grunnkurset.
De skriftlige oppgavene for 10. trinn og for grunnkurset på videregående skole er utformet over samme lest med tre oppgaver; oppgave 1: personlig skriving, oppgave 2: skriving av en argumenterende tekst, og oppgave 3: en friere kommenterende sjanger (beskrivende, resonnerende) eller fortelling.

> **Oppgave 1, 10. trinn:**
> "Some English/American friends are coming to visit you. Write a short letter to them telling them about a place they simply must see while in Norway. Tell them for example
> WHAT……WHERE….. and WHY."

> **Oppgave 1, Grunnkurs:**
> "You and some of your classmates have decided to throw a surprise party for a teacher who is retiring. Write a short message, 1 or 2 paragraphs, to your whole class, telling them about the party suggesting a few things you want people to do and bring."

Beskjeder/brev og liknende er praktisk anvendelige sjangere som alle bør beherske, og som sådan kan denne type oppgave ha positiv tilbakevirkning, men det er vanskelig å forsvare en slik oppgave når den ifølge veiledningen ikke kan oppnå mer enn B1. Faggruppa refererer selv til en CEF-definisjon for C1- nivået i sitt veiledningsmateriell:

> "I can express myself in clear, well-structured text, expressing points of view at some length. I can write about complex subjects in a letter, an essay or a report, underlining what I consider to be the salient issues. I can select style appropriate to the reader in mind."

Her stilles de samme krav til sjangrene "letter", "essay" og "report", og det er derfor vanskelig å forstå hvorfor brevet/beskjeden i oppgave 1 ikke er ment å kunne nå dette C1-nivået. På en måte kan vi derfor si at for de flinkeste elevene teller ikke oppgave 1, noe som må framstå som "urettferdig" for en som har gått alvorlig inn for å lage en god besvarelse.

> **Oppgave 2, 10. trinn:**
> **"Young people today care less about the environment than adults".** This headline appeared in your local newspaper recently. Write a response of 1 or 2 paragraphs to this editorial, providing arguments that will really convince the readers of your point of view."

Oppgave 2, Grunnkurs :
"There is a political proposal that people in Norway should no longer have to cover the cost of cosmetic treatment, to make them look better. This would make other medical services more expensive for people. Write a letter of 1 or 2 paragraphs to a national newspaper providing arguments that will really convince the readers of your point of view on this issue. Give your text a title."

Når det gjelder selve oppgaveformuleringen for 10. trinn, vil vi hevde at det er vanskelig å besvare et innlegg man ikke har lest. Det burde ikke være vanskelig å finne et autentisk innlegg som kan trykkes opp, og som kan fungere som utgangspunkt for et elevsvar. En slik styring ville kunne føre til mindre spredning i svaralternativer og klarere vurderingskriterier og, forhåpentligvis, dermed større sensorreliabilitet. Et lite grep som her er foreslått ville snarere kunne ha økt validiteten til oppgaven ut fra en kommunikativ tenkning.

I den oppgavespesifikke vurderingsbeskrivelsen for *Formidling* står det for oppgave 2 på C1-nivået for både 10. trinn og for grunnkurs:

"Kan fremme komplekse påstander på en kortfattet, velstrukturert og poengtert måte. Kan argumentere overbevisende for sitt syn."

Det virker vanskelig å skulle klare å beskrive komplekse tanker presist og kortfattet og ha hovedpåstander underbygget av overbevisende argumenter i ett eller to avsnitt med en tidsramme på rundt en halvtime. Nivået B1, som nesten 40 % på 10. trinn oppnådde (se kap. 7.1.4), har følgende formidlingskrav:

"Kan få fram sitt syn og gi noen grunner for dette, selv om det kan være noe vanskelig å se sammenhengen mellom påstander og argumenter".

For å oppnå selv dette gjennomsnittlige nivået trenger man rimeligvis flere enn et par avsnitt. Her vil vi hevde at oppgaveteksten, tidsramme foreslått i innledende instruks, samt vurderingskriteriene er lite i samsvar med hverandre. Rammene som settes, gjør at man ikke på en god måte får målt det faggruppa intenderer å få målt.

Oppgave 3, 10. trinn:
"EITHER.

a) A teen magazine is publishing a special edition on famous people. Write a text, 5-8 paragraphs, about a person you think has made a difference to the world. Give your text a title.

OR

b) Write a text, 5-8 paragraphs, with the title:
"The day everything went wrong"

Oppgave 3, Grunnkurs:
"EITHER
a) Imagine you are on a one-year exchange program in an English-speaking country. Write a text, 5-8 paragraphs, about your experiences at the school you

are attending. Feel free to use your imagination as well as drawing on what you actually know about schools in that country.

OR
b) Write a text, 5-8 paragraphs, illustrating the proverb "The grass is always green on the other side of the fence"."

Oppgavetekstene for 10. trinn virker gode med muligheter for å vise kunnskaper og fabulere fritt. For grunnkurset vil nok mange elever måtte velge oppgave b. Det å skulle beskrive sine opplevelser krever noe kunnskap, men definitivt helst egenopplevelse for å kunne skrive godt. Vurderingskriteriene for denne oppgaven er rimelig lik for begge trinn. Det kan være nyttig å spesifisere sjangerkrav og også et forventet innhold, slik at de oppgavespesifikke nivåbeskrivelsene for *Formidling* ikke likner for mye på underkategorien *Tekstoppbygging* under *Språk*.

B2, som bør være et nivå som mange elever når, har en nivåbeskrivelse for grunnkurset under *Formidling (O*ppgave 3*)*:

"Kan skrive en klar og detaljert tekst. Kan skape en tekst med en klar tråd og tematisk utvikling, og som stort sett er sammenhengende og flytende."

For samme nivå for *Tekstoppbygging* finner vi:

"**Organisering:** Stort sett velorganisert og med god bruk av avsnitt.
Logisk/tematisk utvikling: Det er en klar tråd og utvikling i teksten.
Sammenheng og flyt: Det er ganske god flyt og sammenheng i teksten, blant annet på grunn av variert bruk av småord som although, however, after all, og lignende.

Av eksemplet ovenfor kan det virke som omtrent det samme skal vurderes og vektes to ganger. Generelt kan det sies at de såkalte oppgavespesifikke kriteriene for *Formidling* er relativt uspesifikke og til dels språkrelaterte. De to hoveddimensjonene *Språk* og *Formidling* framstår derfor ikke som to klart ulike kompetanseprofiler.

Resultatene fra MMIs spørreundersøkelse i forbindelse med årets nasjonale prøver i engelsk skriftlig (elever på 10. trinn og på grunnkurset), viser at noe over halvparten av elevene mener at de kjenner oppgaveformen og at de har arbeidet med liknende oppgavetyper tidligere. Dette er en uventet lav andel ut fra lang eksamenstradisjon med fri skriftlig produksjon på 10. trinn og på grunnkurset. Likeledes er det en forbausende lav andel av elevene (38 % på 10.trinn/grunnkurset) som mener at de "helt eller delvis" fikk vist sine skriveferdigheter i engelsk på prøven, og hele 46 % av elevene er helt eller delvis uenige i utsagnet om at de fikk vist sine skriveferdigheter på prøven.

Oppgavene for 7.trinn.
Årets sett for 7. trinn består av tre ulike oppgaver. Ifølge vurderingsveiledningen for 7. trinn, skal oppgavene vurderes enkeltvis før en gjør en samlet vurdering av prøven, for øvrig i tråd med råd gitt i evalueringsrapporten i 2004. Det synes derfor hensiktsmessig å se på prøven oppgave for oppgave.

Oppgave 1 sier ganske enkelt: ”*Look at the picture. What do you see?”* Bildet viser en kvinne som holder et barn, samt en mann med topplue som sitter i en sofa og spiser en banan mens han ser på tv. Bildet rommer for øvrig detaljer som et kunstbilde, en fugl i bur, fisk i glassbolle, hund på teppe, leker, hyller med bøker og permer osv. Det er med andre ord nok for elevene å skrive om. Ifølge rettledningen skal denne oppgaven ikke vurderes over nivå A2. Oppgaven er med andre ord i seg selv rammet inn og satt tak på vurderingsmessig. I dette perspektivet kan det symbolrike bildet oppleves noe forvirrende for modne elever. Instruksjonen kan gjerne tolkes som hva elevene *forstår* eller assosierer ut fra dette bildet, og en kan forvente at noen elever reflekterer over innholdet som må sies å fremstå som noe spesielt i forhold til kjønnsrollene det reflekterer (stereotypiske framstillinger: mann med topplue/tv-titting, kvinne som passer barn).

Med den begrensede intensjonen som faggruppa tydeligvis har hatt med oppgaven, burde etter vårt syn en bedre instruksjon til elevene vært: ”*Write a list of what you can see in this picture”* eller ”*Describe what you can see in this picture”*. Slik ville oppgavens bestilling framgå tydeligere for elevene, og lærerne ville i større grad fått klarere kriterier å vurdere etter.

Oppgave 2 ber elevene skrive et postkort til en venn og fortelle hvor man er, hva man gjør og hva man liker eller ikke liker. Her har elevene tre ulike bilder å ta utgangspunkt i, slik at de kan skrive fra henholdsvis en by, en campingplass eller et sted fra kysten. Denne oppgaven kan oppnå et høyere nivå enn oppgave 1 i den forstand at rettledningen sier at elever kan vurderes inntil B1. Rettledningen gir ikke klare forklaringer på hvorfor et postkort skal være vanskeligere å skrive enn en beskrivelse av hva man ser på et bilde.

Oppgave 3 gir elevene valget mellom to oppgaver. De kan velge å ta utgangspunkt i setningen ”*One day when I came home from school, I found the front door wide open”*. Dette er en åpning som naturlig innbyr elevene til å skrive en spenningshistorie, og her kan elevene vise at de behersker fortellerteknikk, tekstoppbygning og formidling. Den andre oppgaven har igjen et bilde med følgende instruksjon: ”*Write your own text, 3 or 4 paragraphs using this picture”*. Oppgave 3 har klarere bestilling enn oppgave 1 og 2, noe som i seg selv bør gjøre den enklere å vurdere. I tillegg er elevene bedt om å bruke halvparten av tiden på denne oppgaven, og rettledningen sier at elever kan plasseres opp til B2 på denne oppgaven. Slik sett fremstår denne oppgaven som den klart beste hva oppgaveinstruksjon angår.

Resultatene fra MMIs spørreundersøkelse om engelsk skriftlig til engelsklærerne på 7. trinn, 10. trinn og på grunnkurset viser at et flertall av lærerne (57 %) mener at elevene bare i noen grad fikk vist bredden i sine skriveferdigheter. En tredel av lærerne mener likevel at den nasjonale prøven i skriftlig engelsk reflekterer læreplanen i ”ganske stor grad”.

Oppsummering av en validitetsvurdering av prøvene for engelsk skriftlig

Faggruppa i engelsk stilte konkrete spørsmål til fjorårets vurderingsrapport om vurdering av prøvens validitet. Når det gjelder godt samsvar mellom oppgaver/vurderingskriterier og det prøven skal måle, vil vi hovedsak si at det er rimelig samsvar mellom oppgaver og

vurderingskriterier, men dette gjelder ikke for den argumenterende teksten på de to øverste trinnene. Likeledes mener vi at Oppgave 1 på 7. trinn ikke hadde god nok instruks for å gi grunnlag for å vurdere ordforråd. Vi mener også at det er lite heldig at det settes tak på nivåoppnåelsen for visse oppgaver. Når det gjelder spørsmålet om tilpasningen til CEF-nivåene er gode nok, er hovedinnvendingen at nivåbeskrivelsene er relative, dvs. at CEF-nivåene har beskrivelser som ikke er identiske for alle trinnene.
Det europeiske rammeverkets hovedmål er primært knyttet til det enkelte individs språklæring, sett i et sosialt og kommunikativt perspektiv. Et klart siktemål er at *Rammeverket* med sine nivåbeskrivelser skal være åpne og tilgjengelige for den som lærer seg et språk slik at språkelever i alle aldre kan bli i stand til å vurdere sin egen språkkompetanse, sette seg realistiske mål og kunne dokumentere sin egen språkutvikling over tid ("*I can do*"- utsagn). En slik pedagogisk tenkning er viktig i en utdannings-sammenheng, og det at den nye læreplanen har et eget hovedområde som heter *Språklæring,* gjør at *Rammeverket* høyst sannsynlig vil bli brukt i norsk skole uansett nasjonale prøver. Men den nye læreplanen inneholder ingen fastsatte nivåbeskrivelser i tråd med *Rammeverket*, så den nye læreplanen er i seg selv ikke et argument for å bruke CEF-skalaen i de nasjonale prøver.

Det ser ut til at oppgavene representerer en positiv gjenkjennelse, men de representerer ikke noe spesielt nytt i prøvesammenheng. Det at tre sjangere brukes hvert år, kan føre til en positiv tilbakevirkningskraft på disse sjangrene, men potensielt på bekostning av andre sjangere. Dersom dette er intendert, bør det kanskje diskuteres om det er en ønskelig effekt. Det er vanskelig å vurdere om prøvene er passe vanskelige når det gjelder fri skriftlig produksjon. Mer relevant er det kanskje å diskutere om f. eks. 7.- klassinger klarer den tidsdisponering som prøven legger opp til. Her testes kanskje indirekte en form for generell teststrategisk kompetanse.

Når det gjelder prøvenes relasjon til gjeldende læreplaner, er denne uklar. Språkdefinisjonen samsvarer godt både med L97 og R94, men det å kunne skrive en argumenterende tekst er for eksempel ikke et krav i L97, men derimot i R94. En av tre lærere (på de tre trinnene som er testet) mener at prøvene reflekterer læreplanen i "ganske stor grad". Men når det gjelder prøvens validitet, er det lite positivt at både elever og lærere mener at det er vanskelig å få vist bredden i skriftlige engelskferdigheter gjennom de nasjonale prøvene (ifølge MMI-undersøkelsen).

7.1.3 Analyse av resultater for grunnkurs

Fordeling av prestasjoner
Tabell 7.2 viser svarfordelingen i prosent på de gitte nivåene. Dataene er gitt for ekstern vurdering. For enkelhets skyld har vi flere steder for hver delkompetanse brukt korte betegnelser på nivåene som framgår av tabell 7.1.

Tabell 7.1: Nivåer for vurdering av engelskprøvene

Nivå	Kort betegnelse
A1	1
A1/A2	2
A2	3
A2/B1	4
B1	5
B1/B2	6
B2	7
B2/C1	8
C1	9

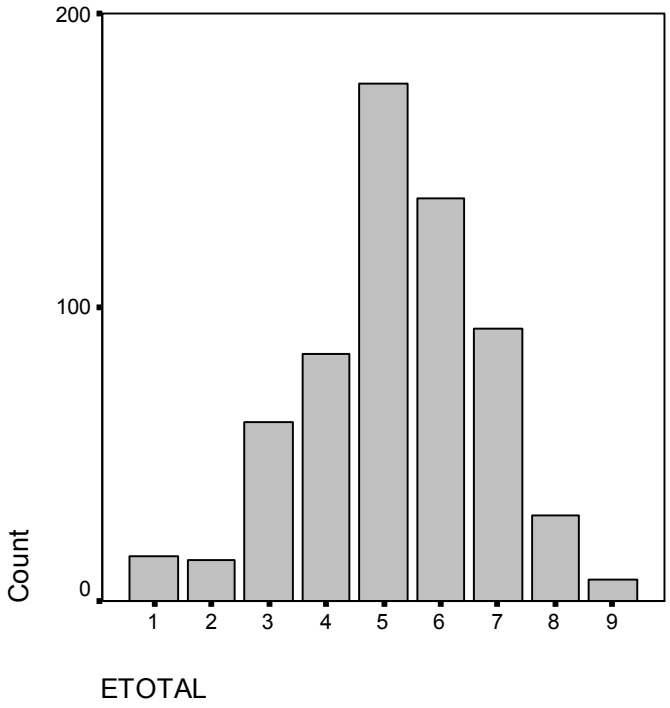
Tabell 7.2: Prosentvis svarfordeling på nivåer (avrundet til hele tall) for hver skala for engelsk skriftlig på grunnkurs. Ekstern vurdering (N = 616)

Kategori	**Svarfordeling i %**								
	A1 1	**A1/A2 2**	**A2 3**	**A2/B1 4**	**B1 5**	**B1/B2 6**	**B2 7**	**B2/C1 8**	**C1 9**
Formidling	3	4	12	13	28	21	14	6	1
Språk	2	3	10	15	30	22	13	5	1
Totalt	2	2	10	14	29	22	15	5	1

Fordelingen på nivåer viser en god normalfordeling, med et midtpunkt på omtrent B1 og de aller fleste elevene (86 %) på nivåer fra og med A2 til og med B2. Gjennomsnittet er 5,2, og standardavviket 1,6 for alle skalaene. Fordelingen er vist grafisk på figur 7.1 for *Totalt* nedenfor, men den er så godt som identisk for alle skalaene.

Det er som nevnt et påfallende trekk at fordelingene på nivåer er så like for de tre skalaene. Dette henger sammen med at elevene stort sett har fått samme nivå på hver av dem. De observerte korrelasjonene mellom de to skalaene er så høy som 0,91 mellom to og to skalaer. Og særlig sterk er korrelasjonen mellom kategorien *Språk* og den totale vurderingen, den er nesten perfekt (0,99). I de interne vurderingene finner vi de samme korrelasjonene som er nevnt her.

Figur 7.1: Fordeling på nivåer for den totale skalaen i engelsk. Grunnkurs, ekstern vurdering (N=616)



Samsvar i vurderingen

Tabell 7.3 viser samsvaret mellom den interne (lærernes) og eksterne uavhengige vurderingen. I tabellen har vi angitt antall elever for hver bestemte differanse mellom skolens og ekspertens vurdering. På grunn av dårlig tilgang på data fra skolene er det her et lavt antall besvarelser, bare 126. Dette gjør at vi må ta noe større forbehold om resultatenes pålitelighet.

Tabell 7.3: Prosentvis samsvar mellom skolenes og eksterne vurderinger for engelsk skriftlig på grunnkurs (N=126)

Differanse		Formidling	Språk	Totalt
Intern vurdering høyere	**5**	1	0	0
	4	0	0	1
	3	3	3	2
	2	15	14	14
	1	25	27	29
Lik vurdering	**0**	32	33	32
Intern vurdering lavere	**- 1**	15	15	14
	- 2	6	6	5
	- 3	1	3	2
	-4	2	0	2
	-5	1	0	0

Det framgår av tabell 7.3 at det bare er rundt 30 % av vurderingene som har vært identiske for de to sensorene, og nesten like stor andel har et avvik på to nivåer eller mer (utenfor skravert område i tabellen). Det er også et ikke helt ubetydelig antall besvarelser der avviket er på 3 eller til og med flere nivåer. En annen måte å angi overensstemmelse på er å beregne koeffisienten kappa (K, "coefficient of agreement", se kap. 4.4), og den ligger så lavt som 0,19. Siden det er så mange som ni kategorier (nivåer) å velge mellom, har vi også valgt å studere virkningen av å rekode den totale vurderingen til tre kategorier: 1 (Lav, under 4), 2 (Middels, 5-7) og 3 (Høy, over 7). Kappa mellom ekstern og intern vurdering øker da til 0,44, fortsatt et svært lavt tall.

Korrelasjonene mellom de to (totale) vurderingene er såpass lav som 0,69. Denne korrelasjonskoeffisienten sier imidlertid ikke så mye om *hva slags* uoverensstemmelse det er mellom vurderingene. Sensorene kan være relativt enige i hvor gode besvarelsene er i forhold til hverandre, men de kan allikevel være uenige i hvor nivået *stort sett* ligger i en gruppe (se kap. 4.4).

For noen av skolene har vi data fra to eksterne vurderinger. Her dreier det seg om mange forskjellige sensorer, og hvem som er betraktet som første og annen sensor, er i vårt datasett helt tilfeldig. For disse (269) elevene får vi en korrelasjon på 0,70 mellom de to vurderingene. Med samme inndeling i tre store kategorier som ovenfor (lav, middels, høy) blir kappa 0,34, altså lavere enn mellom ekstern og intern vurdering. En krysstabell mellom de to eksterne vurderingene, tabell 7.4, illustrerer avvikene mellom de to vurderingene. Det framgår av tabellen at det er stor uenighet om hvordan grensen mellom 1 og 2 og mellom 2 og 3 er å forstå i praksis.

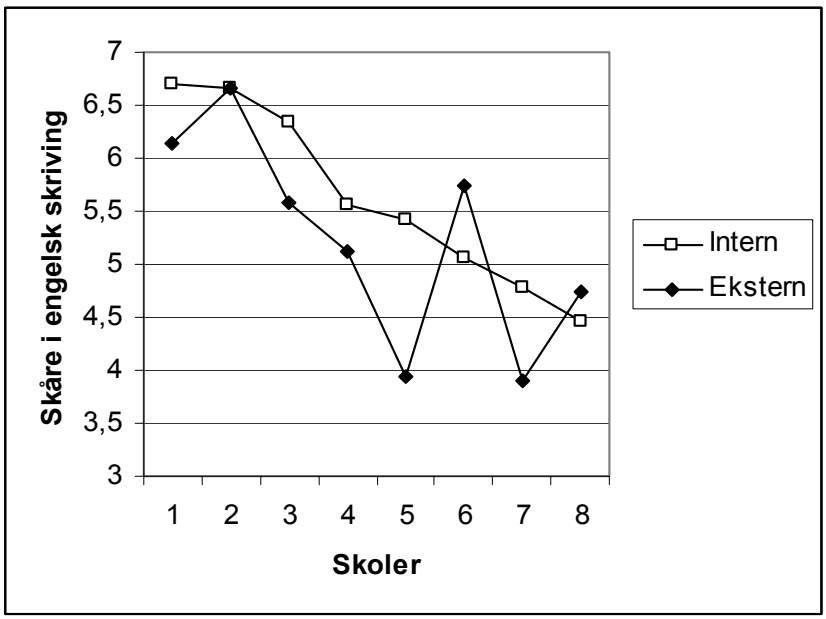
Vi må etter dette dessverre konstatere at sensorreliabiliteten for denne prøven rett og slett er for lav, langt under det som må kreves for en prøve som skal gi pålitelige mål for enkeltelevers (og skolers) kompetansenivå.

Tabell 7.4: Krysstabell mellom vurderingene til to eksterne sensorer. Engelsk skriftlig for grunnkurs (N=269)

		Ekstern vurdering 1			
		Lav 1	**Middels 2**	**Høy 3**	**Sum**
Ekstern vurdering 2	**Lav - 1**	13	13		26
	Middels - 2	8	124	56	188
	Høy - 3		16	39	55
	Sum	21	153	95	269

I tillegg til et stort sprik mellom vurderingene ser vi fra figur 7.2 at dette spriket i stor grad utgjør en *systematisk* forskyvning mot at skolene vurderer sine egne elever høyere enn ekspertene gjør. Denne effekten varierer imidlertid fra skole til skole, og det fører til at en sammenlikning av resultater mellom klasser og skoler, blir betydelig påvirket av hvor "streng" eller "mild" læreren er i sin vurdering. I figur 7.2 har vi demonstrert dette i detalj. Her er skoler med data fra minst 9 elever sammenliknet etter gjennomsnittskåre både ut fra den interne og den eksterne vurderingen.

Figur 7.2: Skolers gjennomsnittlige nivå sammenliknet mellom intern og ekstern vurdering. Skolene er sortert etter intern skåre. Skoler med færre enn 9 besvarelser er tatt ut. Engelsk skriftlig på grunnkurs



Som vi ser fra figur 7.2, er det svært dårlig samsvar mellom de to målene for skolens samlede kompetanse, og da har vi et dårlig grunnlag for en offentlig rapportering av slike resultater. Noen skoler har vurdert sine elever gjennomsnittlig mer enn et helt nivå høyere enn den eksterne vurderingen, mens det på andre skoler er en motsatt (men noe svakere) effekt.

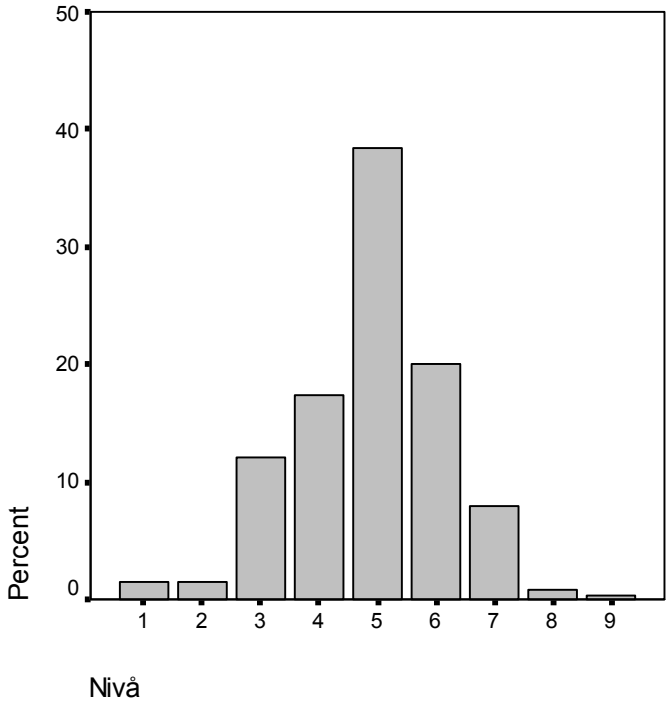
7.1.4 Analyse av resultater for 10. klasse

Fordeling av prestasjoner
Resultatene for 10. klasse er ikke mye forskjellig fra de for grunnkurs. Tabell 7.5 viser svarfordelingen i prosent på de gitte nivåene. Dataene er gitt for ekstern vurdering. Vi ser at som for grunnkurs ligger de aller fleste elevene på nivåene fra A2 til B2. Gjennomsnittet er omtrent 4,9, og standardavviket 1,3 for alle skalaene. Sammenliknet med grunnkurs (gjennomsnitt 5,2) ligger altså nivået litt lavere. Dette er i tråd med hva man vil forvente, men det er ikke naturlig å gå nærmere inn på denne forskjellen, siden utvalget, særlig for grunnkurs, på en ukjent måte kan være påvirket av boikott og annet fravær. For øvrig noterer vi at spredningen er betydelig lavere enn for grunnkurset. Fordelingen er vist grafisk på figur 7.3, og den er så godt som identisk for alle skalaene.

Tabell 7.5: Prosentvis svarfordeling på nivåer (avrundet til hele tall) for hver skala. Engelsk skriftlig for 10. klasse. Ekstern vurdering (N=899)

Kategori	Svarfordeling i %								
	A1 1	A1/A2 2	A2 3	A2/B1 4	B1 5	B1/B2 6	B2 7	B2/C1 8	C1 9
Formidling	2	3	13	20	36	20	6	1	0,2
Språk	2	2	11	17	39	21	8	1	0,3
Totalt	1	2	12	17	39	20	8	1	0,3

Figur 7.3: Fordeling på nivåer for den totale skalaen. Ekstern vurdering av engelsk skriftlig i 10. klasse (N=899)



Det er, som også diskutert for grunnkurselevene, et påfallende trekk at fordelingene på nivåer er så like for de tre skalaene. Korrelasjonen er så høy som 0,90 mellom de to aspektene (*Formidling* og *Språk*). Korrelasjonen mellom kategorien *Språk* og den totale vurderingen er, som for grunnkurselever, nesten perfekt (0,99). Hvis vi i stedet bruker de interne vurderingene, finner vi like store korrelasjoner. Den totale vurderingen framstår altså som tilnærmet identisk med språklig ferdighet, noe som i og for seg er i tråd med vurderingsveiledningen. Men når formidlingsaspektet i det hele tatt ikke ser ut til å influere på den totale vurderingen, blir det uklart hvilken rolle dette aspektet egentlig er tiltenkt. Den totale vurderingen framstår utelukkende som et uttrykk for elevenes rent *språklige* kompetanse.

Samsvar i vurderingen

Tabell 7.6 gir en oversikt over hvor godt samsvaret mellom intern og ekstern vurdering er på 10. trinn. Omtrent halvparten av besvarelsene vurderes likt av de to sensorene, og ser vi også på andelen med avvik på ett poeng, framstår samsvaret i alt som bedre enn på grunnkurs. Men likevel vil vi hevde at avviket er for stort til at sensorreliabiliteten kan anses som akseptabel. Riktignok er korrelasjonen mellom intern og ekstern vurdering overraskende høy (0,81), men igjen vil vi peke på at denne korrelasjonskoeffisienten isolert sett ikke sier så mye om hva som ligger bak uoverensstemmelse mellom vurderingene (se kap. 4.4 og 7.1.3). Korrelasjonen er mye lavere for to eksterne vurderinger seg imellom. For mange av elevene (omtrent 500) har vi dobbel ekstern vurdering. Disse to vurderingene har en korrelasjon seg imellom på 0,72 for total vurdering, omtrent den samme for alle tre skalaene. Dette tallet er altså betydelig lavere enn korrelasjonen mellom intern og (den ene) eksterne vurdering nevnt ovenfor. Og videre ligger dette tallet på samme nivå som for grunnkurs. En mulig årsak til denne uoverensstemmelsen når det gjelder grad av samsvar mellom to vurderere, er at noen lærere kan ha endret sine vurderinger av egne elever etter at de har sett resultatet av den eksterne vurderingen.

Siden korrelasjonskoeffisienten mellom intern og ekstern vurdering er såpass høy som 0,81, har vi studert hvordan dette forholder seg skole for skole. Det viser seg at korrelasjonene mellom to vurderinger på rundt 20 elever fra en og samme skole varierer fra 1,00 til -0,12. For den førstnevnte skolen er altså de to vurderingene helt identiske for alle elevene, og vi mener det er rimelig å tro at lærerne har endret sin vurdering etter å ha sett den eksterne vurderingen, noe som skjedde i ikke ubetydelig grad ved fjorårets prøver. For den sistnevnte skolen er det altså en tendens til at besvarelser som vurderes av den ene sensoren som blant de gode, bedømmes av den andre sensoren som dårlige.

Den systematiske forskyvningen i favør av egne elever er også problematisk, noe som framgår av tabell 7.6. Så mye som 32 % av elevene er vurdert høyere (total vurdering) internt enn eksternt, mens bare 13 % er vurdert høyest eksternt. Denne effekten er enda tydeligere for formidlingsaspektet, der 40 % av elevene vurderes høyere internt.

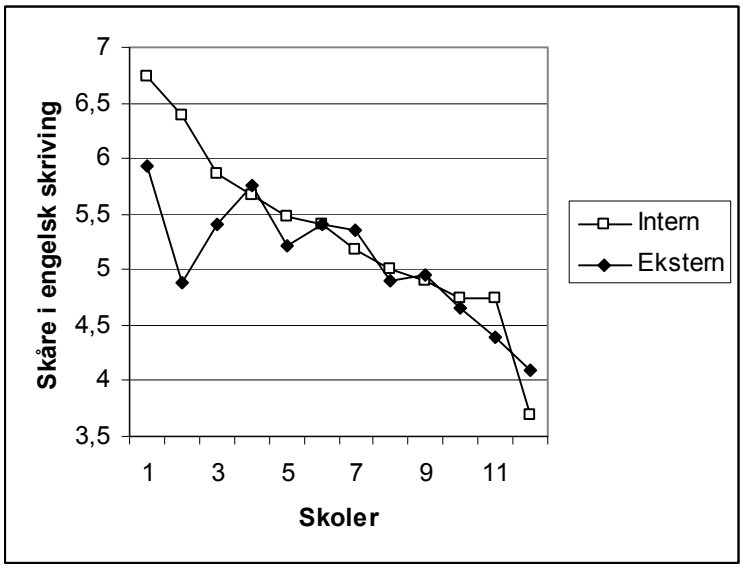
Tabell 7.6: Prosentvis samsvar mellom intern og ekstern vurdering. Engelsk skriftlig for 10. klasse (N=201)

Differanse		Formidling	Språk	Totalt
Intern vurdering høyere	**3**	1	2	1
	2	7	9	9
	1	32	22	22
Lik vurdering	**0**	46	55	56
Intern vurdering lavere	**- 1**	12	11	10
	- 2	2	2	3

Figur 7.4 illustrerer hvordan skolene vurderer sine egne elever sammenliknet med den eksterne vurderingen. Her er skoler med data fra minst 10 elever sammenliknet etter gjennomsnittskåre både ut fra den interne og den eksterne vurderingen. I gjennomsnitt gir skolen selv en vurdering som ligger 0,33 (tall framkommet ifølge omregning av nivå, se Tabell 7.1) høyere enn den eksterne vurderingen. Det framgår tydelig av figuren at denne

effekten varierer sterkt fra skole til skole, noe som fører til at en sammenlikning av resultater mellom klasser og skoler, blir betydelig påvirket av hvor "streng" eller "mild" læreren er i sin vurdering.

Figur 7.4: Skolers gjennomsnittlige nivå sammenliknet mellom intern og ekstern vurdering. Skolene er sortert etter intern skåre. Engelsk skriftlig i 10. klasse



Det er en tendens til at situasjonen i 10. klasse framstår som noe bedre enn den var for fjorårets prøve for dette klassetrinnet. Det er grunn til å tro at det i hovedsak ikke er de samme lærerne som var involvert i fjorårets prøve. I den grad dette er riktig, er det vanskelig å tolke høyere samsvar som et oppnådd tolkningsfellesskap blant lærerne.

7.1.5 Analyse av resultater for 7. klasse

Fordeling av prestasjoner

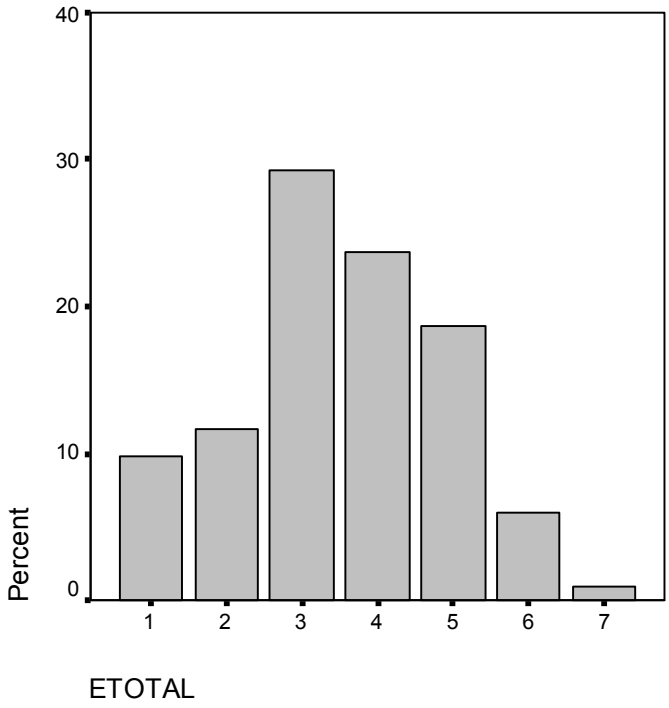
Tabell 7.7: Prosentvis svarfordeling på nivåer (avrundet til hele tall) for hver skala. Ekstern vurdering for 7. klasse (N=621)

Kategori	Svarfordeling i %								
	A1 1	A1/A2 2	A2 3	A2/B1 4	B1 5	B1/B2 6	B2 7	B2/C1 8	C1 9
Formidling	*9*	*18*	*20*	*25*	*19*	*6*	*3*	*-*	*-*
Språk	*10*	*11*	*29*	*23*	*20*	*6*	*1*	*-*	*-*
Totalt	*10*	*12*	*29*	*24*	*19*	*6*	*1*	*-*	*-*

Tabell 7.7 viser fordelingen for den eksterne vurderingen for elever i 7. klasse. Vi vil understreke at i henhold til veiledningen var de to øverste nivåene for de andre klassetrinnene (B2/C1 og C1) ikke aktuelle for bruk på 7. klassetrinn.

Disse elevene ligger naturlig nok betydelig lavere enn på 10. trinn og grunnkurs. Gjennomsnittet er omtrent 3,5, og standardavviket 1,4 for alle skalaene. Fordelingen er vist grafisk på figur 7.5, og den er så godt som identisk for alle skalaene.

Figur 7.5: Prosentvis fordeling på nivåer for den totale skalaen i engelsk skriving. Ekstern vurdering for 7. klasse (N=621)



Sammenhengene mellom de ulike skalaene er omtrent identisk med hva vi fant for de to andre klassetrinnene. Korrelasjonen er så høy som 0,89 mellom de to skalaene (*Formidling* og *Språk*). Det er enda sterkere korrelasjonen mellom kategorien *Språk* og den totale vurderingen, den er nesten perfekt (0,98). Hvis vi i stedet bruker de interne vurderingene, finner vi like store korrelasjoner.

Samsvar i vurderingen

Tabell 7.8 viser samsvaret mellom lærernes (skolenes) vurdering og ekstern vurdering. I tabellen har vi angitt prosentandel elever for hver bestemte differanse mellom intern og ekstern vurdering. Som for de andre klassetrinnene, er det en betydelig forskyvning i retning av at intern vurdering gir høyere skåre. Denne forskjellen i favør av skolens egen vurdering utgjør i gjennomsnitt omtrent et halvt nivå (0,45).

Tabell 7.8: Prosentvis samsvar mellom intern og ekstern vurdering. Engelsk skriftlig for 7. klasse (N=383)

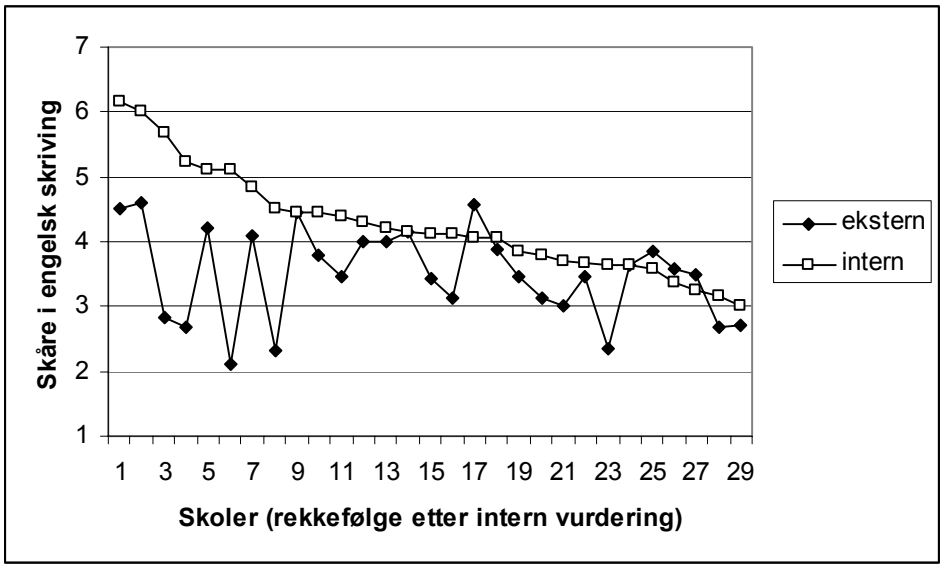
Differanse		Formidling	Språk	Totalt
Intern vurdering høyere	**5**	2	0,5	0,3
	4	3	4	4
	3	9	5	6
	2	11	17	13
	1	27	28	29
Lik vurdering	**0**	32	34	38
Intern vurdering lavere	**- 1**	12	12	10
	- 2	3	1	0,5
	- 3	0,5	0,3	0,3

Åpenbart er det svært stor forskjell mellom de to vurderingene, idet bare mellom 30 % og 40 % av "karakterene" er de samme for de to sensorene. Vi ser videre at den eksterne vurderingen gjennomgående ligger betydelig lavere. Mens over halvparten av besvarelsene vurderes høyere internt, så er det bare litt over 10 % der den interne vurderingen er lavest. Og videre det er generelt lite samsvar mellom de to.

Det er altså en tydelig tydelig tendens i retning av å bedømme egne elever mildere enn den eksterne sensoren gjør. Et annet viktig spørsmål for dette klassetrinnet er om denne tendensen er jevnt fordelt på skoler, eller om det varierer mye fra skole til skole. Figur 7.6 viser en sammenlikning mellom ekstern og intern vurdering for alle skoler med over 6 elever. Skolene er sortert etter avtakende skåre for intern vurdering. Det framgår av figuren at bare 4 av de 29 skolene har vurdert sine elever lavere enn de eksterne sensorene har gjort. De aller fleste skolene har høyere intern vurdering, og for noen skoler er det en ekstremt stor forskjell, helt opp til tre hele poeng! Dette betyr at noen lærere tydeligvis må ha oppfattet CEF-skalaen på en helt annen måte enn det som er intensjonen.

Som vi også ser av figuren, er det så dårlig samsvar mellom de to målene for skolens samlede kompetanse at de gradvis avtakende interne verdiene ikke gjenspeiles i en tilsvarende nedgang for eksterne verdier. Korrelasjonen mellom de to verdiene er så lav som 0,26. Med slike resultater sier det seg selv at offentlig rapportering blir nokså meningsløs. For å sette det på spissen: En skoles resultater sier ikke så mye om elevenes kompetanse, men er først og fremst bestemt av hvor strengt eller mildt lærerne har vurdert elevene! Åpenbart vil det føre galt avsted å publisere slike skoleresultater på Skoleporten, og dette kan få uheldige konsekvenser for enkeltskoler. Det er trolig vanlig at rektorer diskuterer skolens resultater med skoleeier. Naturlig nok blir egne resultater sammenliknet med andre skoler i kommunen, og det kan bli spørsmål om å begrunne resultatene. Det beste svaret på et dårlig skoleresultat er at skolens lærere trolig er strenge i sin bruk av CEF-skalaen.

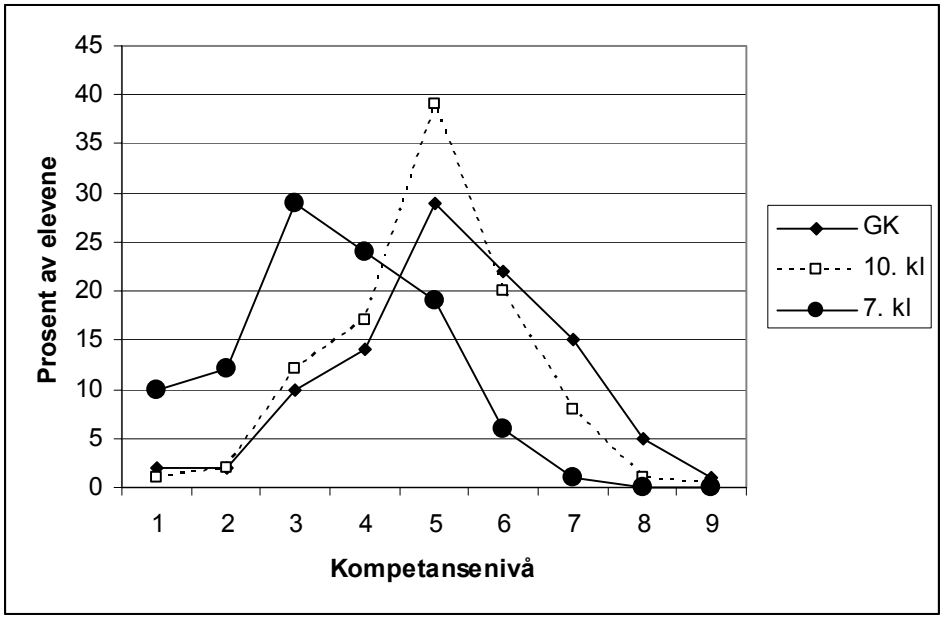
Figur 7.6: Sammenlikning mellom intern og ekstern vurdering når det gjelder gjennomsnitt for skoler. Engelsk skriving i 7. klasse



7.1.6 Oppsummering av den skriftlige prøven

Prøvene for de tre klassetrinnene er vurdert etter den samme skalaen, en tillempet versjon av den internasjonale CEF-skalaen (Common European Framework) med ni trinn. Fordelingen fra de tre prøvene er sammenliknet i figur 7.7. Det framgår her at endringene med klassetrinn virker fornuftig, altså at det gir mening å bruke dataene i en diskusjon om hvor mye "bedre" elevene blir i løpet av et skoleår. I så måte fungerer CEF-skalaen etter hensikten. Gjennomsnittsverdiene for de tre klassetrinnene er henholdsvis 3,5, 4,9 og 5,2.

Figur 7.7 Fordeling på CEF-nivåer (se tabell 7.1) for elever på tre ulike klassetrinn



Med dataene både fra de nasjonale prøvene og fra vårt vurderingseksperiment (se kap. 7.1.7) har vi tydelig påvist at det er svært vanskelig å oppnå en pålitelig vurdering av elevenes prestasjoner. Vi har sett at problemet for en sensor særlig ligger i å bestemme hvor nivået stort sett skal ligge, mens det er noe enklere å sammenlikne elevene imellom. Resultatet av det er blant annet at de skolevise resultatene (i form av gjennomsnittsverdier) i altfor liten grad reflekterer elevenes kompetanse, men snarere avhenger av hvordan lærerne har brukt skalaen; altså hvor strengt eller mildt de har vurdert besvarelsene.

Men problemet med tilfeldig vurdering er ikke "løst" ved å unnlate å rapportere skolevise resultater. Den pedagogiske tilbakemeldingen til skoler, klasser og elever vil også i stor grad være underlagt de samme tilfeldighetene, og verdien av denne synes derfor svært tvilsom.

7.1.7 Eksperiment for å vurdere sensorreliabilitet

Praktisk gjennomføring
I vår gjennomgang av sensorreliabiliteten har vi påvist et svært dårlig samsvar mellom intern og ekstern vurdering. For å studere nærmere hvilke faktorer som synes å påvirke sensorreliabiliteten mer generelt, gjennomførte vi et eksperiment med fire uavhengige vurderinger av et sett autentiske elevbesvarelser, ett for 10. klasse og ett for 7. klasse. Fire personer vurderte de samme besvarelsene uavhengig av hverandre og satte tre skåreverdier som beskrevet tidligere (*Formidling*, *Språk* og *Total*). Dataene ble så analysert på ulike måter, både når det gjaldt overensstemmelse mellom personene og når det gjaldt sammenheng mellom de tre ulike skåreverdiene for samme besvarelse.

De fire vurderingene ble gjennomført av studenter i engelsk og som nettopp hadde fullført sin praktisk-pedagogiske utdanning ved ILS. Disse personene representerte høy fagutdanning i engelsk, høyere enn det som er vanlig i norsk skole i 10. klasse og særlig i 7. klasse. Vi antok derfor at de ville ha bedre forutsetninger for å bruke vurderingskriteriene enn de fleste lærere. På den annen side manglet de lærernes erfaring med vurdering i praksis, men vi mente at dette trolig ikke ville bety mye så lenge kriteriene var helt nye og uvante også for de aller fleste lærerne. Slik sett stilte de likt med lærere i 7. klasse, som aldri tidligere hadde vurdert nasjonale prøver.

En lærer som hadde virket som ekstern vurderer, ble hentet inn for å gi de fire vordende lærerne en gjennomgang og trening i bruk av vurderingsveiledningen, dvs. omtrent den samme opplæringen som lærerne hadde fått før de gjennomførte vurderingen av egne elever.

Etter at alle besvarelsene var vurdert av alle fire, under kontrollerte forhold, ble dataene analysert og gjennomgått. Vi hadde deretter en diskusjon med dem om hvilke avveininger som førte til uoverensstemmelsene, og hvor lette kriteriene var å bruke i praksis.

Data fra 10. trinn

Tabell 7.9: Fordeling av enkeltresultater (totalskåre) for 44 besvarelser på 10. trinn vurdert av fem ulike personer A-E

Nivå	Sensorer				
	A	B	C	D	E
1 – A1	4	2	6	7	4
2 – A1/A2	3	1	2	3	2
3 – A2	3	5	5	5	5
4 – A2/B1	3	8	3	2	4
5 – B1	5	11	16	9	15
6 – B1/B2	9	5	5	4	10
7 – B2	7	10	4	9	3
8 – B2/C1	8	0	0	2	0
9 – C1	2	1	3	3	1
Gjennomsnitt	**5,48**	**5,02**	**4,59**	**4,80**	**4,64**
St. avvik	**2,35**	**1,75**	**2,15**	**2,48**	**1,79**

Tabell 7.9 viser fordeling på nivåer for de 44 besvarelsene slik de ble vurdert av de fire personene A-D samt den opprinnelige vurderingen (E). Tabellen viser også gjennomsnittsverdier og standardavvik for alle besvarelsene. Vi merker oss for øvrig at gjennomsnittet for alle besvarelsene og alle sensorene ligger på 4,9, som er det samme som for alle elevene på 10. trinn i våre andre data.

Fra tabellen ser vi at det er en markant forskjell mellom vurdererne når det gjelder det generelle nivået de har lagt seg på. Forskjellen mellom den mildeste (A) og strengeste (C) av de fem utgjør nesten 1 helt "poeng". B og D sine vurderinger framstår som nærmere gjennomsnittet av de fem vurderingene totalt sett. Vi ser spesielt at det er veldig store forskjeller når det gjelder hvor mye de to øverste nivåene er brukt. Vi må konkludere at kriteriene tydeligvis er oppfattet forskjellig når det gjelder hva kravene betyr i praksis. Et annet påfallende trekk er at noen (særlig C og E) har en sterk opphopning av resultatet 6, altså B1/B2. Et annet trekk er at det er betydelig større spreding (standardavvik) i vurderingene til A og D enn til B og E.

En ting er at det generelle nivået varierer fra bedømmer til bedømmer. Men det kan likevel være en høy korrelasjon mellom vurderingene hvis svingningene fra besvarelse til besvarelse foregår "i takt". Tabell 7.10 viser noe om dette, her i form av korrelasjoner mellom vurderingene for to og to personer.

Av tabell 7.10 ser vi at alle korrelasjonene mellom vurderingene ligger i området 0,81 til 0,90. Den siste kolonnen i tabellen viser samsvaret mellom hver vurdering og gjennomsnittet av alle de andre, noe som kan oppfattes som det beste målet for samsvar med de andre fire vurderingene. På denne bakgrunnen kan vi si at A og E har gjort de mest "riktige" vurderingene av gode og mindre gode besvarelser. Men, som vi husker fra tabell 7.9, er det B og D som ligger på det "riktigste" (mest gjennomsnittlige) nivået totalt sett.

Tabell 7.10: Korrelasjoner mellom vurderingene til to og to av sensorene A-E for 10. klasse. Kolonnen lengst til høyre viser korrelasjonene mellom hver av vurderingene og gjennomsnittet av alle de andre.

	B	C	D	E	Gj snittet av de andre
A	,88	,83	,89	,90	,93
B		,81	,85	,90	,90
C			,86	,87	,88
D				,86	,91
E					,93

Hvordan er så prestasjonsvurderingene framkommet? For å svare på det har vi i tabell 7.11 vist sammenhengen mellom de to delkompetansene (*Formidling* og *Språk*) seg imellom og med den overordnete skåreverdien (totalskåre). Dette har vi vist for hver av de fire vurderingene i eksperimentet, mens vi ikke har tilsvarende data for delkompetansene for den eksterne vurderingen.

Tabell 7.11: Korrelasjoner mellom ulike kompetanser for hver av de fire personene

	A	B	C	D
Språk – Formidling	0,92	0,90	0,97	0,98
Språk – Total	0,99	0,99	1,00	1,00
Formidling – Total	0,95	0,93	0,97	0,98

Det er tydelig fra tabell 7.11 at alle korrelasjonene er høye, men det er også tydelig at korrelasjonene mellom *Språk* og totalskåre nesten er perfekte. Disse to framstår altså som tilnærmet identiske, hvilket får mening i lys av instruksjonen om at det er språkbruken som skal telle desidert mest. Men når det blir så høye korrelasjoner som her, så kan man med rette spørre om det har noen hensikt å operere med to så godt som identiske skalaer. Det er videre tydelig at C og D heller ikke ser ut til å skille noe særlig mellom *Språk* og *Formidling*, så for disse to er det en sterk tendens til at alle de tre kompetansene har samme verdi for nesten alle besvarelsene. De to andre personene, A og B, har tydeligvis i større grad skilt mellom kompetansene.

Kategorien *Språk* er på sin side framkommet av fire ulike aspekter: *Tekststruktur*, *Grammatikk, Ordbruk*, samt *Ortografi og tegnsetting*. Korrelasjonene mellom disse var svært høye, med unntak av vurderer B lå alle godt over 0,90. B, som den eneste, hadde for noen besvarelser noen markerte forskjeller mellom *Tekststruktur* og de andre tre kategoriene. Det ser ut til at denne kategorien er blitt oppfattet nokså forskjellig fra person til person. Som kategori nærmer dens innhold seg jo også formidlingsaspektet, og dette gjør trolig at den oppfattes substansielt forskjellig av ulike personer.

Vi tar ikke med flere resultater om dette her, men slike detaljerte data innbyr til en detaljstudie av hvordan vurderingsveiledningen har fungert i praksis.

Data fra 7. klasse

Tabell 7.12 viser hvordan de fire sensorene har vurdert de 64 besvarelsene fra 7. klasse når det gjelder total vurdering. Vi har ingen annen ekstern eller intern vurdering for disse

elevene. Vi ser av tabellen at det er variasjon fra sensor til sensor av omtrent samme størrelsesorden som det var for 10. klasse. Vi ser også at det er omtrent det samme mønsteret mellom de fire sensorene A-E. A er også her den mildeste i sin vurdering, mens C er den strengeste, og forskjellen er nesten ett "poeng".

Tabell 7.12: Fordeling av enkeltresultater (Total) for 64 besvarelser i 7. klasse vurdert av fire ulike personer A-D

	A	**B**	**C**	**D**
1 – A1	8	3	5	6
2 – A1/A2	2	7	15	7
3 – A2	14	26	22	22
4 – A2/B1	16	11	16	9
5 – B1	11	14	4	15
6 – B1/B2	9	3	2	0
7 – B2	4	0	0	5
Gjennomsnitt	**3,98**	**3,55**	**3,08**	**3,63**
St. avvik	**1,68**	**1,22**	**1,16**	**1,57**

Tabell 7.13 viser korrelasjoner mellom vurderingene for to og to personer. Av tabellen ser vi at korrelasjonene mellom enkeltvurderingene ligger i området 0,70 til 0,85, altså betydelig lavere enn for 10. klasse. Den siste kolonnen i tabellen viser samsvaret mellom hver vurdering og gjennomsnittet av alle de andre, noe som kan oppfattes som det beste målet for samsvar med de andre tre vurderingene. På denne bakgrunnen kan vi si at A helt klart har gjort de mest "riktige" vurderingene av gode og mindre gode besvarelser. Men, som vi husker fra tabell 7.12, er det B og D som ligger på det "riktigste" (mest gjennomsnittlige) nivået totalt sett. Konklusjonen for hver av sensorenes vurdering er igjen nokså lik den for 10. klasse.

Tabell 7.13: Korrelasjoner mellom vurderingene til to og to av sensorene A-D for 7. klasse. Kolonnen lengst til høyre viser korrelasjonene mellom hver av vurderingene og gjennomsnittet av alle de andre.

	B	**C**	**D**	**Gj snittet av de andre**
A	,78	,85	,83	,91
B		,70	,72	,79
C			,74	,83
D				,83

Tabell 7.14 viser sammenhengen mellom de to delkompetansene (*Formidling* og *Språk*) seg imellom og med den overordnete skåreverdien (*Total*). Dette har vi vist for hver av de fire vurderingene i eksperimentet.

Tabell 7.14: Korrelasjoner mellom ulike kompetanser for hver av de fire personene

	A	B	C	D
Språk – Formidling	0,93	0,89	0,96	0,93
Språk – Total	1,00	1,00	0,99	1,00
Formidling – Total	0,93	0,90	0,96	0,93

Det er tydelig fra tabell 7.14 at alle korrelasjonene er høye, men det er også tydelig at korrelasjonene mellom språk og totalskåre nesten er perfekt. Disse to framstår altså også for dette klassetrinnet som så godt som identiske. Det er videre tydelig at det heller ikke er noe stort skille mellom *Språk* og *Formidling*. Person B har i størst grad skilt mellom de to kompetansetypene.

Hva har vi lært av eksperimentet?
Vårt eksperiment har gått ut på å se systematisk på samsvar mellom ulike vurderinger av samme besvarelser, og videre å se hva som ligger bak ulikhetene. Våre data synes å innby til noen tydelige konklusjoner:

- Samsvaret mellom uavhengige vurderinger, i form av korrelasjoner, ligger mellom 0,80 og 0,90 for 10. klasse og mellom 0,70 og 0,85 for 7. klasse. Dette gjelder for personer med høy kompetanse og med gjennomført opplæring i å bruke vurderingskriteriene. Sensorene er altså i ganske stor grad enige om hvilke besvarelser som er gode og dårlige. Isolert sett oppfatter vi dette som et positivt funn.
- Derimot er det meget stor forskjell mellom personene når det gjelder hvilket gjennomsnittlig nivå de legger seg på. I så måte synes CEF-skalaen å fungere for dårlig. Sensorene synes å ha helt forskjellig oppfatning av hvor "gode" besvarelsene er i forhold til nivåene i skalaen.
- Det er mye større uenighet, og det synes derfor vanskeligere å vurdere de yngste elevenes besvarelser. Det samme viste seg for øvrig i dataene fra nasjonale prøver. Ifølge studentene henger dette ikke minst sammen med uklarhet i forhold til å vurdere innslag av norske ord og oppgaver med manglende svar.
- Også spredningen synes å være svært forskjellig. Noen sensorer bruker hele skalaen i større grad enn andre.
- En etterfølgende diskusjon med sensorene har overbevist oss om at det uten omfattende trening (med tilbakemelding) i bruk av CEF-skalaen i praksis, og derved et bedre tolkningsfellesskap, ikke er mulig å oppnå bedre overensstemmelse i vurderingen enn dette eksperimentet har vist. Dette betyr rett og slett at en skriftlig vurderingsveiledning, uansett hvor detaljert og god den eventuelt kan bli, i seg selv ikke er tilstrekkelig for god sensorreliabilitet. Heller ikke "kursing" i bruk av vurderingsveiledningen er tilstrekkelig.

7.2 Engelsk lesing

7.2.1 Innledning

De nasjonale prøvene i engelsk lesing er nettbasert. Vi vil berømme engelskgruppa for det arbeidet som er gjort ved utviklingen av prøvene, idet lignende prøver aldri har vært gjennomført i så stor skala i Norge. Ikke minst vil vi berømme den fyldige BITE-IT-rapporten utarbeidet av faggruppen i engelsk som dokumenterer i detalj utviklingen av prøvene.

BITE-IT-rapporten fra 2004 refererer til følgende grunnmodell for prøvene:

- Forprøve – et oppgavesett bestående av ca. 20 korte oppgaver som måler elevenes nivå for å finne riktig nivå på hovedprøven.
- Hovedprøve – et oppgavesett som inneholder ca. 35 lengre oppgaver som skal måle elevenes faktiske nivå relatert til CEF.

Prøvene er adaptive, og det er en utviklet en utvelgelsesalgoritme som velger ut en hovedprøve på riktig nivå basert på elevenes prestasjoner i forprøven.*"Leseprøvene i engelsk er laget for datamaskiner og er adaptive, det vil si at de tilpasser seg elevenes ferdighetsnivå automatisk" (...) " Alle elevene vil få oppgaver som er individuelt tilpasset deres ferdighetsnivå" (BITE-IT-rapporten 2004).* Det er utviklet tre hovedsett som elevene kan bli testet i. Til leseprøven på 10 trinn i 2004 gjennomførte 2 % av elevene prøvesett 1, 94 % av elevene fullførte prøvesett 2 og 4 % prøvesett 3. I etterkant ble "karakterene" utregnet etter et bestemt nivåfastsettingssystem. Resultatene fra 2004 viser at nivåfordelingen av de som gjennomførte prøven, 90 % av elevene befinner seg på B-nivåene (B1, B1/2, B2). For 2005 er innslaget over B2 en god del høyere.

Vi har i tabell 7.15 satt opp nivåfordelingen på leseprøven fra 2004 og 2005 sammen med resultatene fra skriveprøven i 2005 for å se i hvilken grad nivåene samsvarer. I utgangspunktet forventet vi noe høyere resultater for lesing enn for skriving, men ikke så stort sprik som det vi ser nedenfor. Noe uventet har svært få fått lavere enn B1 i lesing. Vi stiller oss i utgangspunktet noe tvilende til om dette kan være riktig måling av elevenes lesekompetanse i henhold til beskrivelsene for CEF-nivåene. Også det høye og økende prosentandelen på helt eller nesten oppnådd C1-nivå er for oss noe vanskelig å forstå.

Tabell 7.15: Fordeling av resultater fra engelsk skriving og lesing i 2005 og engelsk lesing i 2004. 10. trinn

	Prosentvis fordeling på nivåer								
	A1	A1/A2	A2	A2/B1	B1	B1/B2	B2	B2/C1	C1
Skriving 2005	1	2	12	17	39	20	8	1	0,3
Lesing 2004	0	0	0,2	0,9	40	29	20	4	5
Lesing 2005	0	0	0,1	1,0	38	28	19	4	11

7.2.2 Generelt om oppgavenes validitet.

Det å lage leseprøver på internett medfører en del utfordringer. I en validitetsdiskusjon handler det for det første om hvorvidt det er mulig å måle det man ønsker å måle. For det andre er det stor sannsynlighet for at elevenes erfaring med digitale medier kan spille en rolle. I hvert fall på lavere trinn er det store forskjeller mellom elevene når det gjelder det å kunne håndterer en PC.

Faggruppa i engelsk gir følgende korte definisjon av hva prøven skal måle:

"De nasjonale leseprøvene i engelsk skal måle elevenes grunnleggende ferdigheter i å lese engelsk tekst. Det vil si at prøvene måler en del av det elevene lærer i

engelskfaget, mens andre deler av faget må vurderes på andre måter."
(http://bite.intermedia.uib.no/tests/about_reading.page).
I tillegg er det tre leseferdigheter som hovedsakelig måles:

- Å trekke konklusjoner
- Forstå hovedinnholdet
- Forstå detaljer, hente ut informasjon

Oppgavene er i hovedsak bygget over følgende lest: en oppgaveinstruksjon, en tekst eller et bilde, en påstand eller et spørsmål og et felt for oppgavebesvarelser. Besvarelsen av en flervalgsoppgave består i at eleven klikker i en svarrute av flere mulige, eventuelt ved å klikke direkte på bilder eller ved å markere ordet i teksten. Oppgavene grupperer seg i følgende kategorier ifølge BITE-IT-rapporten:

- Multiple choice/flervalg
- Highlight/markér ord
- Drag and drop/dra og slipp
- True-False/rett-galt
- Colour/fargelegg
- Who could say-click name/ Hvem kan si - klikk på navnet

Ved å få passord, fikk vi begrenset tilgang til et sett prøver på hvert trinn, og vi i evalueringsgruppa har gått igjennom lesetester på alle trinn med ulik intensjon, fra "flink" til "svak"elev. Til tross for at vi konsekvent prøvde å svare feil på flere prøver, var det vanskelig å få lavt nivå. Uansett fikk vi opp omtrent det samme settet med oppgaver, om enn i noe ulik rekkefølge, og med rimelig bra resultat. Lesetesten er utformet slik at man er nødt til å svare på oppgaven for kunne gå videre i prøven. Idet det er flervalgsoppgaver av ulikt slag, kan eleven klare seg bra med gjetting.

På **11. trinn** viste gjennomprøving at det var litt over 30 oppgaver. I hovedsak var dette flervalgsoppgaver, med to til fem svaralternativer. Oppgavene baserte seg hovedsaklig på korte tekster. Det var kun en lang tekst og skjermbildet var til dels uleselige uten ytterligere justeringer på PC skjermen. Spørsmålene testet de ulike formene for leseforståelse som er beskrevet ovenfor, men prøveformatet hadde ingen variasjon som på de lavere trinnene. Det er uheldig at elever på dette trinnet blir presentert med primært detaljlesing av korte tekster. Dette kan få en uheldig tilbakevirkningseffekt dersom leseprøven blir tatt på alvor.

Gjennomgangen på **10. trinn** viste litt over 30 oppgaver, av noe ulike format, hovedsakelig flervalgsoppgaver. Elevene fikk lese korte tekster med svaralternativer som true/false, eller to til flere svaralternativer. Ifølge en rektor vi intervjuet likte elevene på ungdomstrinnet godt denne formen for testing: *"... elevene likte godt leseprøven. Dette ble som spill, oppfattet som moro og mange fikk gode resultater."* Leseprøven inneholdt også enkelte illustrasjoner slik at tekst og bilder skulle kombineres for å teste elevenes leseferdigheter.

Leseprøvene på **7. trinn** var basert på både tekst og bilder og hadde omtrent 27 oppgaver, i hovedsak flervalgoppgaver. I tillegg var det "click and drag"- oppgaver og en

fargeoppgave. Noen av oppgavene har tekstmessige og tekniske utfordringer som kan påvirke resultatene, og måler like gjerne elevenes tekniske og strategiske kompetanse som leseferdigheter.

Leseprøvene på **4. trinn** hadde cirka 9 oppgaver, i hovedsak flervalgsoppgaver, "matche bilder" og "click and drag". Ved gjennomgang av 4. trinn ser vi at det i noen oppgaver kan være utfordrende instruksjoner, til tider er instruksjonene vanskeligere enn selve oppgavene. Når elevene må dobbeltklikke på et ord for at det skal markeres, bør det ganske enkelt stå: "Double click on the correct word." Det vil da være mindre tvil om hva elevene skal gjøre. Man bør unngå ordet "mark", da dette er mer tvetydig (det gir assosiasjoner til å markere, setter ring rundt, krysse av osv.) Instruksjoner bør videre unngå ordet "right", som forveksles med det norske ordet "høyre". Vi noterer oss at 10. trinns nasjonale prøver og demo - testen bruker ordet mark, mens tilsvarende oppgaver for 4. trinn benytter seg av ordet "Click" i instruksjonen. Vi anbefaler altså det siste for denne type oppgaver.

Det kan se ut som om en del oppgaver er utfordrende pga. skrolling og layout, og noen av oppgavene ser ut til å kunne måle elevenes teststrategier likeså mye som leseforståelse. Det er for øvrig grunn til å tro at gjennomgang av demo - settet er svært viktig på dette trinnet, slik engelskgruppen selv skriver, da elevene er uerfarne med testsituasjoner.

Ved gjennomgang av både demo - testene og de nasjonale prøvene er dette et vesentlig punkt, da det ikke er åpenbart enkelt for elevene å forstå hva de faktisk skal gjøre. Flere av oppgaveinstruksjonene kunne med fordel vært forenklet, særlig for elever på 4. trinn.

7.2.3 Resultater

Fordeling av elevene etter hovedtester
Prøvene er som tidligere nevnt, adaptive. Det betyr at elevene ledes til ulike (hoved-) prøver etter hvordan de svarer på den innledende delen. Tabell 7.16 viser hvordan den innledende delen av prøven (20 oppgaver) har ledet elevene inn til ulike prøver etter de innledende spørsmålene. Det er stort sett tre alternativer, henholdsvis test 1 (den letteste), test 2 (den midterste, gjelder de fleste) og test 3 (den vanskeligste). Som det framgår av tabellen, er det noen tester som er besvart av veldig få elever, og som vi derfor ikke vil diskutere her. Vi vil særlig konsentrere oss om resultater fra test 2, men også se på noen av versjonene av test 1 (for 4. trinn) og test 3 (7. og 10. trinn).

Tabell 7.16: Fordeling av elever etter hvilken hovedtest de har gjennomført

Klassetrinn	Test 1 (lettest)	Test 2 (middels)	Test 3 (vanskeligst)
4	50	50	(ingen elever)
7	0,4	80	20
10	3	85	13
Grunnkurs	4	81	14

Når vi i det følgende diskuterer reliabiliteten til prøvene og diskriminering av enkeltoppgaver, vil vi peke på to viktige forhold. For det første er elevene allerede skilt

etter prestasjon i og med den innledende prøven, og det vil ikke være rimelig å stille like strenge krav til reliabilitet og diskriminering som når alle elevene tar samme prøve. For det andre er mange av oppgavene av typen riktig/galt, og det er da 50 % sjanse for å tippe riktig. I slike situasjoner er det vanskelig å oppnå høy diskriminering, noe vi ser tydelig utslag av for alle leseprøvene. Men også for slike oppgaver er det viktig å unngå svært lav diskriminering, for slike oppgaver svekker prøvens reliabilitet.

Resultater fra 4. trinn

Tabell 7.17: Item-analyse av oppgavene i test 2 for engelsk lesing på 4. trinn. N=280

Oppgave	Prosent riktig	Diskriminering	Kommentarer
1	96	,24	a
2	81	,43	
3	76	,41	
4	81	,36	
5	60	,36	
6	74	,38	
7	65	,41	
8	60	,31	
9	43	,39	
10	53	,19	a!
11	49	,30	
12	72	,35	
13	68	,40	
14	62	,43	
15	89	,28	a
16	95	,32	
17	51	,43	
18	64	,40	
19	33	,35	
20	23	,27	a
21	68	,45	
22	53	,44	
23	69	,44	
24	69	,39	
25	64	,34	
26	58	,39	
27	13	-,13	a!!
28	43	,12	a!
29	35	,38	
30	38	,08	a!!
31	45	,17	a!
32	57	,30	
33	80	,24	a
34	32	,04	a!!
35	75	,37	
36	57	,19	a!
37	60	,20	a
38	58	,26	

a Lav diskriminering, <0,30
a! Svært lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10

Tabell 7.17 viser resultatene for test 2, oppgave for oppgave. Dette gjelder altså prøven som ble gjennomført av omtrent halvparten (de flinkeste) av elevene. Det framgår av tabellen at de fleste oppgavene diskriminerer rimelig bra, men likevel er det et betydelig antall som diskriminerer dårlig. Gjennomsnittlig er det 60 % riktige svar på denne prøven, noe vi betrakter som ideelt. Og reliabiliteten er etter forholdene tilfredsstillende (se ovenfor), alfa = 0,83. Det er verdt å merke seg at hvis man rett og slett fjernet alle oppgavene markert med a! eller a!! i tabell 7.17, ville alfa økt til 0,85, og vi ville etter vårt syn ha fått en bedre prøve.

Tabell 7.18. Item-analyse av oppgavene i test 1 for engelsk lesing på 4. trinn. N=299

Oppgave	Prosent riktig	Diskriminering	Kommentar
1	71	,28	a
2	52	,29	a
3	52	,42	
4	58	,33	
5	58	,36	
6	72	,44	
7	91	,30	
8	89	,28	a
9	88	,33	
10	61	,23	a
11	71	,21	a
12	73	,00	a!!
13	75	,30	
14	83	,41	
15	94	,29	a
16	85	,43	
17	80	,46	
18	36	,23	a
19	82	,39	
20	87	,40	
21	52	,21	a
22	81	,43	
23	46	,11	a!
24	34	,03	a!!
25	30	-,01	a!!
26	54	,34	
27	45	,25	a
28	62	,28	a
29	59	,27	a
30	55	,19	a!
31	63	,37	
32	71	,01	a!!
33	63	,12	a!
34	56	,13	a!
35	65	,25	a
36	54	,18	a!
37	63	,40	
38	58	,31	

a Lav diskriminering, <0,30
a! Meget lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10

Vi konstaterer fra tabell 7.18 at det er litt større problemer med å få høy diskriminering på denne prøven enn på test 2. Det er noen oppgaver som definitivt ikke burde vært med, i og med at de faktisk diskriminerer negativt. Gjennomsnittlig korrelasjon oppgavene imellom er 0,09, og det er betenkelig lavt. Reliabiliteten er derfor litt lav (alfa= 0,78). Hvis vi hadde fjernet alle de ni oppgavene markert med a! eller a!! i tabell 7.18, ville alfa økt til 0,82. I alt er det 65 % riktige svar, noe vi oppfatter som fint for disse elevene.

Ut fra resultatene på de to prøvene for 4. trinn er ikke reliabiliteten en alvorlig hindring mot å publisere resultatene. Vi kan imidlertid ikke si noe sikkert om hvordan de to prøvene har fungert i forhold til hverandre. Dette henger sammen med vår usikkerhet når det gjelder overføring fra skåre på prøvene til CEF-nivå.

Resultater fra 7. trinn

Tabell 7.19 viser tilsvarende resultater fra test 2 for 7. trinn, som ble gjennomført av 80% av elevene. Vanskelighetsgraden viser stor variasjon, og gjennomsnittet på 57 % riktige svar er omtrent ideelt. Problemet med prøven er lav reliabilitet, alfa = 0,75. Det framgår av tabellen at det er mange oppgaver med lav diskriminering, markert med a, a! eller a!!. Gjennomsnittlig korrelasjon mellom oppgavene er så lav som 0,075, og da blir nødvendigvis reliabiliteten forholdsvis lav på tross av at det er mange oppgaver. Det kan være verdt å merke seg at uten de 12 oppgavene som i tabell 7.19 er merket med a! eller a!!, ville alfa ha økt til 0,79. Det er også grunn til å peke på at mange av de problematiske oppgavene er oppgaver som nesten alle (spesielt oppgave nr. 17, 32 og 33) eller nesten ingen (særlig oppgave nr. 5 og 19) elever har besvart riktig.

Tabell 7.20 viser tilsvarende de resultatene for test 3, som i alt ble gjennomført av de 20% presumptivt beste elevene. Denne prøven har en forholdsvis god reliabilitet, alfa= 0,82. Men en svakhet med prøven er at den framstår som for lett for disse beste elevene. Med så mye som gjennomsnittlig 80 % riktige svar har ikke det adaptive formålet fungert bra på dette trinnet. Det hadde åpenbart vært en fordel om denne prøven hadde inneholdt flere vanskelige oppgaver i stedet for alle de oppgavene som nesten alle elevene har fått riktig.

Tabell 7.19: Item-analyse av oppgavene i test 2 for engelsk lesing på 7. trinn. N=299

Oppgave	Prosent riktig	Diskriminering	Kommentar
1	71	,20	a
2	67	,35	
3	54	,13	a!
4	5	,33	
5	6	,22	a
6	58	,33	
7	61	,37	
8	81	,38	
9	82	,31	
10	77	,39	
11	84	,35	
12	59	,44	
13	57	,33	
14	63	,28	a
15	71	,40	
16	62	,38	
17	99	,10	a!
18	74	,47	
19	01	-,13	a!!
20	65	,30	
21	26	-,03	a!!
22	46	,23	a
23	39	,31	
24	31	,13	a!
25	24	-,08	a!!
26	72	,20	a
27	65	,11	a!
28	90	,34	
29	63	,16	a!
30	81	,18	a!
31	91	,30	
32	97	,14	a!
33	100	-	b
34	25	,37	
35	23	,05	a!!
36	31	,10	a!
37	39	,21	a

a Lav diskriminering, <0,30
a! Meget lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10
b Ingen informasjon, siden alle svarer riktig

Tabell 7.20: Item-analyse av oppgavene i test 3 for engelsk lesing på 7. trinn. N=297

Oppgave	Prosent riktig	Diskriminering	Kommentar
1	85	,35	
2	32	,10	a!
3	86	,28	a
4	87	,20	a
5	71	,14	a!
6	93	,38	
7	76	,33	
8	85	,40	
9	31	,28	a
10	66	,38	
11	84	,36	
12	73	,30	
13	93	,31	
14	98	,40	
15	95	,38	
16	98	,24	a
17	96	,27	a
18	46	,07	a!!
19	90	,25	a
20	99	,29	a
21	97	,28	a
22	99	,32	
23	99	,15	a!
24	92	,38	
25	84	,48	
26	76	,42	
27	93	,40	
28	70	,39	
29	88	,41	
30	92	,50	
31	67	,35	
32	67	,46	
33	67	,37	
34	51	,36	
35	87	,45	
36	78	,38	

a Lav diskriminering, <0,30
a! Meget lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10

Resultater fra 10. trinn

Resultater for test 2 på 10. trinn er vist i tabell 7.21. Denne prøven er gjennomført av en klar majoritet av elevene, 85 %. Riktignok er det noen oppgaver med svært dårlige tekniske egenskaper, som burde vært fjernet. Men likevel fungerer prøven rimelig bra, med en gjennomsnittlig vanskelighetsgrad på 69 % riktig (som er litt for høyt for å være ideelt) og med en reliabilitet på alfa=0,84.

Tabell 7.21: Item-analyse av oppgavene i test 2 for engelsk lesing på 10. trinn. N=288

Oppgave	Prosent riktig	Diskriminering	Kommentar
1	53	,52	
2	88	,23	a
3	93	,38	
4	89	,39	
5	88	,43	
6	71	,35	
7	79	,45	
8	74	,35	
9	68	,38	
10	74	,36	
11	72	,53	
12	69	,41	
13	33	,37	
14	88	,44	
15	82	,09	a!!
16	88	,27	a
17	54	,22	a
18	65	,21	a
19	64	,31	
20	75	,33	
21	11	,09	a!!
22	67	,42	
23	84	,51	
24	92	,36	
25	74	,43	
26	87	,44	
27	92	,44	
28	56	,36	
29	74	,25	a
30	50	,00	a!!
31	85	,38	
32	61	,27	a
33	30	-,01	a!!
34	57	,47	
35	40	,38	
36	41	,28	

a Lav diskriminering, <0,30
a! Meget lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10

De presumptivt litt over 10 % beste elevene har gjennomført test 3, og resultater fra denne prøven er vist i tabell 7.22. Fra denne tabellen ser vi at prøven har fungert svært dårlig. For det første har den vært altfor lett, idet det i gjennomsnitt er 94 % riktige svar, og det er en sterk takeffekt. Når nesten alle svarer riktig, blir diskrimineringen automatisk lav, og det gjelder så godt som alle oppgavene. Reliabiliteten blir derfor svært lav, alfa= 0,62. Vi kan bare konkludere at denne prøven ikke har fungert etter forutsetningen. Uten å vise resultatene i detalj vil vi her bare gi noen få data også om test 1 for de antatt svakeste 3 % av elevene. Også denne prøven har fungert dårlig. Den framstår som for vanskelig (35 % riktige svar), og reliabiliteten er lav, alfa= 0,68. Følgelig må vi

konstatere at det adaptive elementet ved prøven(e) på 10. trinn ikke har fungert etter forutsetningene.

Tabell 7.22: Item-analyse av oppgavene i test 3 for engelsk lesing på 10. trinn. N=288

Oppgave	Prosent riktig	Diskriminering	Kommentar
1	94	,18	a!
2	97	,16	a!
3	97	,23	a
4	85	,21	a
5	96	,00	a!!
6	99	,24	a
7	99	,01	a!!
8	99	,29	a
9	99	,21	a
10	99	,19	a!
11	84	,25	a
12	86	,21	a
13	94	,20	a
14	95	,20	a
15	99	-,05	a!!
16	99	,19	a!
17	97	,17	a!
18	98	,01	a!!
19	85	,16	a!
20	77	,15	a!
21	77	,27	a
22	96	,12	a!
23	99	,24	a
24	98	,26	a
25	99	,41	
26	99	,44	
27	97	,22	a
28	100	-	b
29	97	,15	a!
30	95	,31	
31	88	,13	a!
32	90	,22	a
33	99	,21	a
34	99	26	a
35	97	,11	a!

a Lav diskriminering, <0,30
a! Meget lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10
b Ingen informasjon, siden alle svarer riktig

Resultater for grunnkurs

For grunnkurs nøyer vi oss med å vise resultatene for test 2, som ble besvart av litt over 80 % av elevene. I tabell 7.23 har vi vist resultater for denne prøven. Det framgår av tabellen at det er svært få tekniske problemer med prøven. Det er gjennomsnittlig 65 % riktige svar, og alfa er så høy som 0,92. Denne svært høye reliabiliteten framstår som positiv. Vi mener at dette i hovedsak skyldes at det er brukt vanlige flervalgsoppgaver med fire eller flere alternativer som diskriminerer bedre enn riktig/galt-formatet. På den

annen side kan en så høy verdi for alfa gi grunn til å spørre om oppgavene er *for* like, at de i for stor grad måler det samme og derfor måler liten faglig bredde (se kap. 4.3).

Tabell 7.23: Item-analyse av oppgavene i test 2 for engelsk lesing på grunnkurs. N= 300

Oppgave	Prosent riktig	Diskriminering	Kommentar
1	63	,41	
2	75	,60	
3	61	,49	
4	59	,39	
5	71	,53	
6	80	,59	
7	85	,51	
8	74	,38	
9	90	,44	
10	68	,48	
11	81	,53	
12	73	,41	
13	80	,54	
14	80	,50	
15	73	,52	
16	58	,26	a
17	68	,48	
18	45	,45	
19	54	,60	
20	65	,54	
21	54	,54	
22	57	,55	
23	74	,48	
24	53	,38	
25	66	,67	
26	80	,63	
27	47	,42	
28	58	,28	a
29	55	,49	
30	48	,50	
31	52	,42	
32	59	,35	
33	65	,42	
34	57	,43	
35	65	,40	
36	54	,46	
37	58	,37	

a Lav diskriminering, <0,30
a! Meget lav diskriminering, <0,20
a!! Ekstremt lav diskriminering, <0,10

7.2.4 Resultater trinn for trinn

Ut fra resultatene for alle elevene som har gjennomført prøvene, har vi i tabell 7.24 vist hvordan fordelingen på nivåer er for hvert klassetrinn. Vi understreker at dette ikke er data fra vårt utvalg, men en landsoversikt som vi har fått tilsendt fra faggruppa.

Tabell 7.24: Fordeling etter nivåer for leseprøvene i engelsk. Hele elevkullet

Klassetrinn	Svarfordeling i %								
	A1 1	A1/A2 2	A2 3	A2/B1 4	B1 5	B1/B2 6	B2 7	B2/C1 8	C1 9
4. trinn	19	44	29	7	1	0	0	0	0
7. trinn	0	0	2	28	57	9	3	0	0
10. trinn	0	0	0	1	38	28	19	4	11
Grunnkurs	0	0	3	12	12	16	20	17	21

For å sammenlikne resultater for skriving og lesing på engelsk har vi i tabell 7.25 sammenliknet gjennomsnittskårene for hvert trinn. Tallene refererer her til skalaen 1-9 for CEF-nivåene.

Tabell 7.25: Gjennomsnittlig skåre i engelsk skriving og lesing

Klassetrinn	Skriving	Lesing
4. trinn	-	2,3
7. trinn	3,5	4,8
10. trinn	4,9	6,2
Grunnkurs	5,2	6,7

Vi vil ut fra disse resultatene peke på to forhold:

- For grunnskoleelevene ser det ut til å være en fornuftig progresjon mellom trinnene. Men prøven(e) for grunnkurs gir svært forskjellige resultater, med mange flere elever høyt oppe og lavt nede på skalaen enn det er for 10. klasse. Dette virker rett og slett urimelig.
- Det framgår at prestasjonene gjennomgående ligger godt over ett "nivå" høyere i lesing enn i skriving, og vi er rett og slett usikre på om dette virkelig er i overensstemmelse med beskrivelsene av CEF-nivåene.

7.2.5 Oppsummering av engelsk lesing

Basert på diskusjonene av resultatene vil vi her summere opp noen funn og synspunkter:

- Som faggruppa i engelsk selv understreker i sin egen BITE-IT- rapport fra 2004, er det noen layoutmessige og tekniske problemer knyttet til leseprøvene. Dette gjelder fortsatt, og det vil si at det er en fare for at man i en del tilfeller tester elevenes digitale ferdigheter og teststrategier mer enn selve leseferdigheten i engelsk.
- Oppgaveinstruksen er ikke alltid like klar, og for de lavere trinnene hender det at instruksjonene er vanskeligere enn selve oppgavene. Usikkerhet knyttet til

hvordan oppgavene skal løses, kan øke elevenes vilkårlige gjetting idet formatet innbyr til det. Elevene må gi et svar på hver oppgave for å komme videre.

- Vi kan ikke forstå at elevenes kompetanse virkelig er så høy som resultatene viser, vurdert etter hva nivåbeskrivelsene sier. Ut fra resultatene får for eksempel nesten alle 10. klassingene B1 eller bedre. Kun ca 1% skårer lavere enn det. Vi betviler om dette kan gjenspeile den faktiske lesekompetansen. På mange måter er det fint at elever får god tro på egne ferdigheter, og som en rektor sa ”*Det ble som spill, oppfattet som moro og mange fikk gode resultater*”(kap1.4). Vi vil påpeke at det å gi elevene en urealistisk tro på egne leseferdigheter også kan medføre problemer. Våre egne erfaringer ved å gjennomføre prøvene på 10.trinn ga oss følelsen at det var vanskelig, selv om vi prøvde bevisst, å komme på A2-nivået eller lavere.
- Idet oppgaver og resultater ikke kan studeres i etterkant, er det vanskelig for læreren å vite hva slags problemer elevene har og hva slags feil de gjør. Den pedagogiske verdien av leseprøven blir derfor begrenset. Et nivå alene gir ikke nok informasjon om den enkelte elev til at leseopplæring kan tilpasses den enkelte elev.
- På mange måter er det positivt med oppøving i digitale ferdigheter, og en testsituasjon kan gi skolene et incitament til å arbeide med dette. Men for mange elever kan en testsituasjon på PC bli unødig stressende. Det er viktig at et ønske om pedagogisk utviklingsarbeid ikke må forringe kvaliteten på de nasjonale prøvene.
- Når det gjelder prøvenes reliabilitet, må alle hovedtestene fungere tilfredsstillende for at prøven for trinnet kan kalles god. Selv om noen av hovedtestene har fungert rimelig bra, er det summen av prøvene som avgjør kvaliteten. Slik item-analysene viser, oppfyller ikke prøvene på noen av trinnene dette kriteriet.
- Fra item-analysene og egne erfaringer med testene, kan vi heller ikke se at prøvene fungerer tilfredsstillende adaptivt. Det er også store usikkerhetsmomenter knyttet til at elevene får ulike oppgaver, uansett om de differensierte prøvesettene hadde fungert bra psykometrisk sett.

8 Matematikk

Faggruppa for matematikk har laget prøver for alle elever på 4., 7. og 10. trinn. For grunnkurselever har de laget tre versjoner av prøven avhengig av elevenes studieretning og valg av matematikk. Prøvene har mye til felles, blant annet er det nesten bare åpne oppgaver, og hver av dem skal vurderes etter en nokså detaljert vurderingsveiledning som ivaretar både riktighet (hvor *godt* er svaret?) og det diagnostiske aspektet (*hvilket* svar er gitt?). Et annet fellestrekk er at hver oppgave er ment å måle ett av i alt tre kompetanseområder:

1. Representasjoner, symbolbruk og formalisme
2. Matematisk resonnement, tankegang og kommunikasjon
3. Matematisk anvendelse, problembehandling og modellering

Det er laget et bakgrunnsdokument for prøvene med tittelen ”Kompetanser i matematikk”. Her har man delt inn matematisk kompetanse i åtte delkompetanser som beskrives kort i dokumentet. Beskrivelsene av kompetansene bygger på rapporten ”Kompetencer og matematiklæring- ideer og inspirasjon til utvikling af matematikundervisning i Danmark” av Mogens Niss og Thomas Højgaard Jensen. For de nasjonale prøvene er disse åtte kompetansene slått sammen til fire kompetanseområder, nemlig de tre områdene presentert over, samt ”Hjelpemiddelkompetanse”. ”Hjelpemiddelkompetanse” blir imidlertid ikke vurdert i de skriftlige delene av nasjonale prøver.

Data til våre analyser er hentet fra både skolens egen vurdering og fra ekstern vurdering av de samme besvarelsene.

Vi har vurdert prøvenes validitet i forhold til de gjeldende læreplanene. Imidlertid vil vi peke på at det er et åpent spørsmål om intensjonen er at prøvene i særlig grad skal måle grunnleggende ferdigheter i faget. I så fall ville dette kunne ført til en annen utforming av prøvene. Særlig gjelder dette prøvene for grunnkurs.

8.1 Matematikk 4. trinn

8.1.1 Prøvens struktur og validitet

Prøven for 4. trinn består av en blanding av ferdig oppstilte oppgaver og tekstoppgaver. Oppgavesettet innledes med noen oppgaver som tester ferdigheter i regneartene med ferdig oppstilte oppgaver. Deretter følger noen oppgaver hvor elevene må lese matematikken ut av tekst, for så å beregne riktig svar. Oppgavesettet tester også eksplisitt elevenes forståelse av posisjonssystemet, som i oppgave 8. Kunnskap om enkle geometriske figurer undersøkes også, samt grunnleggende forståelse av brøkbegrepet. Videre testes kjennskap til forskjellen mellom partall og oddetall. Avslutningsvis i oppgavesettet inngår oppgaver knyttet til avlesing fra en kalender. Innholdsmessig vurderer vi prøven til å generelt ha høy validitet i forhold til den gjeldende læreplanen for matematikk generelt og for 4. trinn spesielt. Og for 4. trinn omhandler også prøven i hovedsak grunnleggende ferdigheter.

Oppgavesettet består av totalt 20 oppgaver, hvorav noen består av flere deloppgaver (heretter kalt oppgaver). Det er få flervalgsoppgaver i prøven. En umiddelbar anbefaling er å inkludere flere flervalgsoppgaver i framtidige prøver. Mange åpne oppgaver skaper en stor rettebyrde, og flere av oppgavene i prøven kunne med letthet vært gitt i flervalgsformatet.

Totalt antall skårepoeng på prøven er 59. Det er utviklet en vurderingsveiledning med et relativt detaljert kodesystem for hver oppgave. I følge kodeboka er oppgavene vektet med poeng fra 1 til 2, avhengig av arbeidsmengde og vanskelighetsgrad. Hver oppgave er definert å i hovedsak måle ett av tre *kompetanseområder*:

1. Representasjoner, symbolbruk og formalisme (heretter kalt RSF).
2. Matematisk resonnement, tankegang og kommunikasjon (heretter kalt RTK).
3. Matematisk anvendelse, problembehandling og modellering (heretter kalt APM).

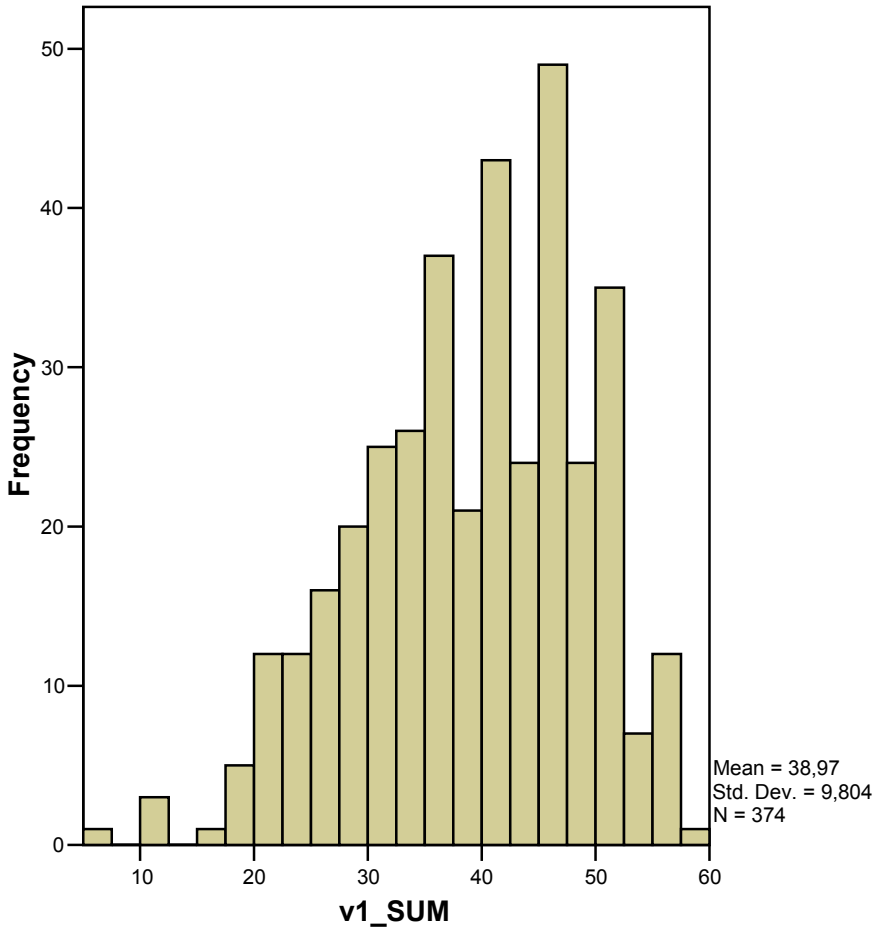
Maksimalt antall poeng innen hvert kompetanseområde er henholdsvis 22, 18 og 19. For hvert kompetanseområde er det definert 5 *kompetansenivåer* ut fra hvor mange poeng som oppnås på oppgavene knyttet til hvert område. Det er imidlertid uklart hvordan disse fem nivåene er tenkt brukt i praksis.

Vi er kritiske til systemet som er valgt for å kode oppgavene. I internasjonale studier som TIMSS og PISA anvender man kodesystemer hvor første siffer alltid angir antall poeng, mens det andre sifferet bare har en kategorisk funksjon. Veiledningen for de nasjonale prøvene følger ikke et slikt system, noe som blir ulogisk og derfor en klar ulempe i videre databehandling og analyse. For eksempel gir kode 1 noen ganger 1 poeng, andre ganger 2 poeng. Det er heller ikke noe prinsipielt skille mellom kodene 10-19 og 20-29. En bestemt tallkode gir noen ganger poeng, andre ganger ikke. Det er grunn til å tro at et mer logisk og enhetlig system vil være til hjelp for lærerne som skal anvende systemet, og ikke minst være tidsbesparende. Vår anbefaling er å endre kodesystemet mer i retning av det som anvendes i internasjonale komparative studier i matematikk.

8.1.2 Fordeling av skåreverdier og koder

Figur 8.1 viser hvordan elevenes totalskåre på matematikkprøven fordeler seg. På grunn av sen innsending av resultater, var det for mange elever kun enten intern eller ekstern vurdering som forelå på det tidspunktet analysene måtte gjennomføres. Fordelingen som er presentert på figur 8.1, er derfor basert på delvis ekstern og delvis intern vurdering. I analysene har totalt 374 elever inngått. Totalt antall oppnåelige poeng på prøven er som tidligere nevnt 59, og gjennomsnittskåre for prøven er 39 poeng, som tilsvarer 67 % av ”fullt hus”. Fordelingen er noe forskjøvet mot høye verdier, og den framstår derfor som nokså lett for de fleste. En ulempe er da at prøven skiller mye mellom de svakeste, men ikke så godt mellom flinke elever. Svake elever vil påvirke gjennomsnittsverdiene forholdsvis mye. Prøven som helhet har en relativt god spredning (et standardavvik på 9,8). Tilsvarende fordelinger finner vi også for hver av de tre delskalaene RSF, RTK og APM, men vi viser ikke disse her.

Figur 8.1: Fordeling av skåre på matematikkprøven for 4. trinn basert på delvis intern og delvis ekstern vurdering (N=374)



Kodesystemet som er anvendt i prøven, er som nevnt svært omfattende. Tanken bak dette er opplagt å søke å fange diagnostisk informasjon om elevenes tenkning. Som vist i tabell 8.1, fanger imidlertid nesten halvparten (45 %) av kodene mindre enn 5 prosent av elevene. Antallet koder på noen av oppgavene er også svært høyt. Hvis vi ser bort fra kodene som ikke har separat diagnostisk verdi (1 = riktig, 0 = blankt, 99 = andre svar), fanger over 60 % av kodene færre enn 5 % av elevene.

En slik inflasjon av koder vil føre til høyere tidsbruk enn nødvendig blant lærerne som skal vurdere prøven. Vi foreslår at antallet koder reduseres betraktelig i framtidige prøver. Ideelt sett bør man pilotere oppgaver og kodesystemer slik at man kan luke bort koder som er svært lite frekvente før prøven gjennomføres i stor skala (for eksempel med kriteriet <5 %).

Tabell 8.1: Prosentvis svarfordeling på kodenivå for 4. trinn. Koder som fanger mindre enn 5 % av elevene, er skravert. Poengene er ikke angitt, men 1 er alltid det beste, og 0 er blankt. N=374

Oppgave	0	1	11	12	13	21	22	23	24	25	26	99
1a	0	94				4	1					1
1b	2	83				2	1	3	1			8
1c	0	94				1						5
1d	1	84				2	4					9
1e	4	80				1	4					12
1f	1	89				2	1					7
1g	12	53				1	5	2	3			24
1h	5	76				2	0	3				14
2	0	82				3	6	3				5
3	3	41	6	23	3	4	4					17
4	2	74				4	20					0
5a	3	53	9	3		10						21
5b	9	10	2			1	1	1	1	45	5	25
6a	1	73	2	7		6	0					11
6b	12	35				12	17					25
6c	9	63	16	2		2						8
6d	12	40				2	1	12				35
7a	0	97				1	1					1
7b	1	82				9	6					2
7c	0	82				16	1					2
8	3	75				3	2	1	7			9
9	1	35	3	2	2	50						8
10a	1	86				5						8
10b	4	27				53	2					15
10c	16	61				5						18
10d	15	50				3	4	4				24
11a	1	89				6						5
11b	1	78				17						4
11c	1	85				2	2					11
11d	2	46				4	16	9	9			13
12a	1	58	39									3
12b	2	53	35			4	3					4
13	2	45	17	4	2	1						28
14	6	56				8	2	10	1	3		15
15a	6	75				6	1					11
15b	7	70				1	1	1	5			14
15c	8	70				1	1	1	5			14
15d	7	65				1	11	3				12
16	1	68	11	9		1	5	4				2
17	1	7	54	4	9							26
18a	3	81				2	5					9
18b	5	72				9	5					9
18c	8	45				1	10	5				31
19a	2	87				2						8
19b	3	75	4	3	7	2	1					5
20a	5	43				51						1
20b	5	45				48						2

8.1.3 Analyser av enkeltoppgaver

Tabell 8.2 presenterer en item-analyse av alle enkeltoppgavene i matematikkprøven for 4. trinn. Resultatene i tabellen viser at for alle oppgavene, med ett unntak, er poengene riktig ordnet etter elevenes dyktighet. Med dette menes at elevene som får 2 poeng i gjennomsnitt er dyktigere enn de som får 1 poeng osv. For oppgave 6a er de elevene som får 1 poeng i gjennomsnitt like gode som de som får 0 poeng. Skillet mellom 0 og 1 poeng er derfor ikke meningsfullt, og de to poengnivåene burde vært slått sammen. Det er imidlertid et klart skille i dyktighet mellom de som får 1 og de som får 2 poeng. Ut fra kriteriet vi har satt, er det om lag en firedel av oppgavene som ikke har god nok diskriminering, men flere av disse ligger kun marginalt under den valgte grensen på 0,30 (se kapittel 4.2). Tre av oppgavene (angitt ved a!) har imidlertid en diskriminering på godt under 0,20, og disse oppgavene burde etter vår mening vært luket ut etter utprøvingen.

Som vist i tabell 8.2, er det gjennomgående svært høy sensorreliabilitet (R) i prøven for 4. trinn. På det tidspunktet hvor analysene måtte gjennomføres, forelå ikke både ekstern og intern vurdering for alle elevene i utvalget. Analysene som er presentert i tabell 8.2, er basert på resultater for 198 elever. Resultatene viser at kun én av oppgavene har lavere sensorreliabilitet enn grensen vi har satt på 85 prosent (se kap 4.4). Dette gjelder oppgave 13 som har en enighet på 83 prosent. For de fleste av oppgavene ligger imidlertid sensorreliabiliteten tett opp mot 100 prosent.

Tabell 8.2: Item-analyse for matematikkoppgavene for 4. trinn. N=374.
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng				Gjennomsnittlig dyktighet etter poeng					
	Blank	0 poeng	1 poeng	2 poeng	Blank/ 0 poeng	1 poeng	2 poeng	D	R	Komm.
1a	0	6	94	-	33	39	-	0,16	100	a!
1b	2	15	83	-	31	41	-	0,35	98	
1c	0	6	94	-	33	39	-	0,14	100	a!
1d	1	15	84	-	30	41	-	0,38	97	
1e	4	16	80	-	31	41	-	0,41	99	
1f	1	10	89	-	29	40	-	0,35	99	
1g	12	35	53	-	33	44	-	0,57	97	
1h	5	19	76	-	29	42	-	0,55	98	
2	0	18	82	-	31	41	-	0,40	99	
3	3	25	32	41	31	39	44	0,55	90	
4	2	24	74	-	34	41	-	0,31	99	
5a	3	32	12	53	32	37	44	0,55	88	
5b	9	79	2	10	38	44	47	0,29	95	a
6a	1	17	9	73	32	31	42	0,42	94	b
6b	12	53	35	-	36	45	-	0,43	95	
6c	9	10	18	63	31	35	43	0,51	87	
6d	12	48	40	-	35	45	-	0,48	93	
7a	0	3	97	-	26	39	-	0,21	99	a
7b	1	17	82	-	31	41	-	0,37	100	
7c	0	18	82	-	30	41	-	0,45	99	
8	3	22	75	-	31	42	-	0,48	97	
9	1	57	7	35	35	38	45	0,45	95	
10a	1	13	86	-	32	40	-	0,27	98	a
10b	4	69	27	-	36	47	-	0,48	97	
10c	16	23	61	-	32	43	-	0,56	96	
10d	15	35	50	-	34	44	-	0,54	97	
11a	1	10	89	-	31	40	-	0,27	98	a
11b	1	22	78	-	35	40	-	0,24	97	a
11c	1	15	85	-	35	40	-	0,16	96	a!
11d	2	52	46	-	35	43	-	0,39	96	
12a	1	3	39	58	33	34	43	0,42	96	
12b	2	10	36	53	27	36	44	0,58	97	
13	2	30	24	45	32	39	44	0,53	83	c
14	6	38	56	-	33	44	-	0,56	95	
15a	6	19	75	-	31	42	-	0,45	96	
15b	7	12	81	-	31	41	-	0,39	99	
15c	8	23	70	-	32	42	-	0,48	90	
15d	7	28	65	-	33	42	-	0,47	96	
16	1	12	19	68	30	36	42	0,41	87	
17	1	26	66	7	34	40	48	0,35	88	
18a	3	16	81	-	34	40	-	0,24	98	a
18b	5	23	72	-	34	41	-	0,29	88	a
18c	8	47	45	-	36	43	-	0,34	93	
19a	2	11	87	-	34	40	-	0,20	99	a
19b	3	9	14	75	31	34	41	0,39	92	
20a	5	52	43	-	35	44	-	0,44	98	
20b	5	50	45	-	35	44	-	0,43	93	

a) Svak diskriminering (<0,30, a! betyr <0,20)
b) Poengene ikke ordnet etter dyktighet
c) Dårlig overensstemmelse mellom vurderingene (< 85 %)

8.1.4 Analyse av foreslåtte skalaer

Faggruppa har som nevnt delt oppgavene inn etter tre kategorier, forkortet som RSF, RTK og APM. Et viktig spørsmål er hvordan denne inndelingen fungerer empirisk. Tabell 8.3 viser reliabiliteten (Cronbachs alfa, se kapittel 4.3) til totalskalaen og de foreslåtte delskalaene i matematikk. Reliabiliteten til totalskalaen er høy nok til at det er naturlig å rapportere etter denne. På grunn av lav reliabiliteten til delskalaene må vi fraråde at disse legges til grunn for rapportering. Særlig er det tydelig at oppgavene i den siste kategorien i for liten grad innbyrdes måler det samme til at vi med under 20 oppgaver kan få høy nok reliabilitet.

Tabell 8.3: Reliabilitet (Cronbachs alfa) til totalskala og foreslåtte delskalaer for prøven i matematikk for 4. trinn.

Skala	**Reliabilitet**
Representasjonskompetanse og kompetanse i symbolbruk og formalisme (RSF)	0,77
Resonnements-, tankegangs- og kommunikasjonskompetanse (RTK)	0,77
Anvendelses-, problembehandlings- og modelleringskompetanse (APM)	0,68
Totalskala	0,88

Hvor forskjellig er så de foreslåtte skalaene? For at de hver for seg skal ha noen pedagogisk verdi, må de i hvert fall framstå som rimelig forskjellige. Resultatene i tabell 8.4 viser både de observerte og de latente korrelasjonene (se kapittel 4.5) mellom de tre delskalaene. Tabellen viser at delskalaene APM og RTK har den høyeste latente korrelasjonen seg imellom, mens RSF og RTK er de mest forskjellige. Ingen av skalaene framstår som tilnærmet identiske. Men uansett har delskalaene for lav reliabilitet til at de kan tillegges pålitelig informasjonsverdi.

Tabell 8.4: Observerte og latente korrelasjoner mellom delskalaer for 4. trinn. Latente korrelasjoner i parentes.

	RSF	RTK
RTK	0,64 (0,83)	
APM	0,65 (0,90)	0,68 (0,94)

8.2 Matematikk 7. trinn

8.2.1 Prøvens struktur og validitet

Prøven for 7. trinn er bygget over samme lest som den for 4. trinn. Oppgavesettet består av totalt 23 oppgaver, hvorav noen består av flere deloppgaver (heretter kalt oppgaver). Det finnes kun noen få flervalgsoppgaver (som oppgave 3 og 4). Mange av de åpne oppgavene har omfattende kodesystemer. Oppgave 18 har så mye som 12 koder. Man kan stille spørsmålstegn ved hvor hensiktsmessig det er å anvende så omfattende kodesystemer, noe vi vil komme nærmere tilbake til.

Totalt antall skårepoeng på prøven er 54. Det er utviklet en vurderingsveiledning med detaljert kodesystem for hver oppgave. Hver oppgave er definert å i hovedsak måle ett av de tre kompetansene som vi for korthets skyld har kalt RSF, RTK og APM. Maksimalt antall poeng innen hvert kompetanseområde er henholdsvis 13, 17 og 24. For hvert kompetanseområde er det definert 5 *kompetansenivåer* ut fra hvor mange poeng som oppnås på oppgavene knyttet til hvert område. Matematikkprøven for 7. trinn skal løses uten lommeregner.

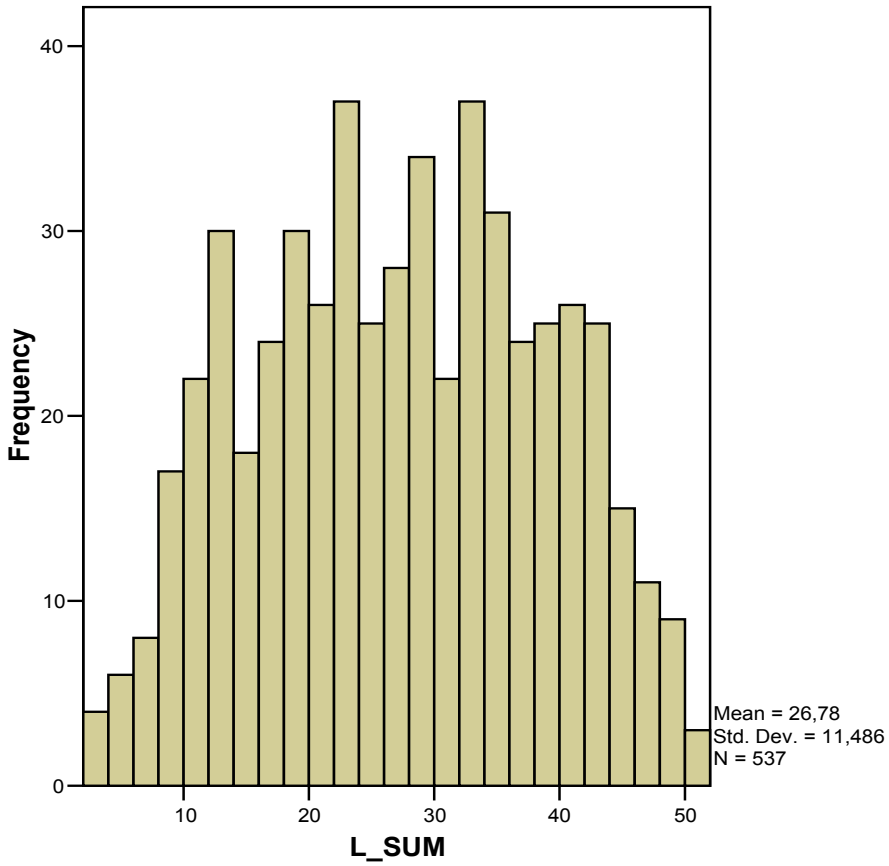
Innholdsmessig vurderer vi prøven til å ha høy validitet i forhold til norsk læreplan på det aktuelle nivået. Oppgavesettet innledes med oppgaver som tester grunnleggende tallforståelse. Litt lenger ut i prøven finner vi også en oppgave som tester elevenes forståelse av sammenhengen mellom prosent og brøk. Forståelse av negative tall berøres også mot slutten av prøven. Etter de innledende oppgavene i tallforståelse, følger noen oppgaver som tester grunnleggende regneferdigheter. Prøven innholder videre oppgaver som tar for seg areal og omkrets, og også enkle figurtall. I prøven finner vi også oppgaver med et mer virkelighetsnært utgangspunkt, blant annet et sett av oppgaver basert på en togtabell og avlesninger fra denne. På samme måte finner vi en oppgave som dreier seg om prosentregning i en praktisk kontekst. Høy innholdsvaliditet i forhold til L97 anser vi som en stor styrke ved denne prøven.

8.2.2 Fordeling av skåreverdier og koder

Figur 8.2 viser hvordan elevenes totalskåre på matematikksprøven fordeler seg. Totalt antall oppnåelige poeng er som tidligere nevnt 54, mens gjennomsnittskåre for prøven er 27 poeng, som utgjør nøyaktig 50 % av det totale antallet. Fordelingen ligger tett opp til en normalfordeling med standardavvik på 11,5. Prøven som helhet har en god spredning og skiller godt mellom sterke og svake elever. Tilsvarende fordelinger finner vi også for hver av de tre delskalaene RSF, RTK og APM, men vi viser ikke disse her. Fordelingen for RSF har imidlertid en svak forskyving mot høye skåreverdier, men tendensen er ubetydelig.

Denne prøven framstår altså som mye vanskeligere for elevene enn prøven for 4. trinn, og vi finner det umotivert at det er så stor forskjell mellom klassetrinnene i så henseende. Teknisk sett er den høye vanskelighetsgraden overhode ingen svakhet, men for elevene representerer dette sannsynligvis en prøve som er mye vanskeligere enn de er vant med. Etter vår mening er dette uheldig.

Figur 8.2: Fordeling av skåre på matematikkprøven for 7. trinn basert på intern vurdering (N=537)



Kodesystemet som er anvendt i prøven, er som nevnt til dels svært omfattende. Tanken bak dette er opplagt å søke å fange diagnostisk informasjon om elevenes tenkning. Som vist i tabell 8.5, er det imidlertid en stor andel av kodene som fanger mindre enn 5 prosent av elevene, og disse er skravert i tabellen. Antallet koder på noen av oppgavene er også svært høyt; i ett tilfelle (oppgave 18) finner vi så mange som 12 koder! Som for 4. trinn foreslår vi derfor en kraftig reduksjon ved at lite frekvente koder lukes ut før prøven gjennomføres i stor skala (for eksempel med kriteriet <5 %).

Tabell 8.5: Prosentvis svarfordeling på kodenivå for 7. trinn. Koder som fanger mindre enn 5 % av elevene, er skravert. Poengene er her ikke angitt, men 1 er beste svar og 0 er blankt. N=537.

Oppgave	0	1	11	12	13	14	15	16	17	21	22	23	24	25	99
1a	3	85								4	1	1			7
1b	2	86								7	0	1			5
2	1	79								6	7	1			6
3	0	50		0						25	2	22			0
4	0	25								60	4	11			0
5a	13	44								1	4	2			36
5b	15	46								3	9				28
6	8	50								3	4	1	1		33
7a	10	44								14	4	4			24
7b	4	67								6	5	2			16
7c	13	36								33	4				14
8a	2	90								4					4
8b	13	46								15	10	3			14
8c	13	48								21	2				17
9	12	52								9	7	2			19
10a	1	83								6	3				7
10b	4	26								55	2				14
11	6	47	4							4	7	1	3		27
12	1	54								2	31	6	4		1
13a	6	18	21	11	2					10					32
13b	37	43								5	1				14
14a	3	73								1	3				20
14b	5	65	9												21
14c	13	27	16	2											42
15	7	11	9	50	1	2									22
16a	6	65	0							10	7	1			11
16b	13	58	0							1	10				18
16c	15	14								3	10	5	2	3	46
17	12	31	3	8	4					13	3	0			27
18	9	31	16	3	4	2	3	5	2	0	5				21
19a	8	46								20	3	7	4		12
19b	19	46								16	2				18
19c	23	9								29	18	4			17
20a	24	59								3	1				13
20b	27	62								5		2			6
20c	15	76								3					6
20d	40	46								6					9
20e	35	37								8	13				6
21a	9	68								8	4				11
21b	11	41								10	14				24
21c	7	77								11	2				4
22a	13	21								16	3	7	17		24
22b	24	22								6	3	7	4		35
23	16	26	7	0	2					1	21	1	1		24

8.2.3 Analyser av enkeltoppgaver

Tabell 8.6 presenterer en analyse av alle enkeltoppgavene i matematikkprøven for 7. trinn. Resultatene i tabell 8.6 viser at for alle oppgavene er poengene riktig ordnet etter elevenes dyktighet. Med dette menes at elevene som får 3 poeng gjennomsnittlig er dyktigere enn de som får 2 poeng, som igjen er dyktigere enn de som får 1 poeng osv. De aller fleste av oppgavene har også god diskriminering. Dette tyder på meget grundig

utprøving av oppgavene. Ut fra kriteriet vi har satt er det bare seks oppgaver som ikke har god nok diskriminering, og flere av disse ligger kun marginalt under den valgte grensen på 0,30.

Den høye vanskelighetsgraden har satt et tydelig spor i den høye andelen ubesvarte oppgaver. For mange oppgaver dreier det seg om over 20 %, særlig i den siste halvdelen av prøven.

Den oppgaven som diskriminerer dårligst, er oppgave 10b. Her skal elevene fylle inn et manglende tall i ei rute slik at svaret blir korrekt: $3+3\cdot\square=24$. Årsaken til at oppgaven diskriminerer dårlig, er at elevene som svarer 4, er tilnærmet like gode som de elevene som gir det korrekte svaret 7. Dette er et svært interessant tilfelle. Diagnostisk sett er dette en glimrende oppgave, fordi den avslører elever som har problemer med operasjonenes rekkefølge når det ikke står parentes. Men psykometrisk fungerer den dårlig, fordi også ellers flinke elever viser seg å slite med de samme problemene. Og da måler denne oppgaven i for stor grad noe annet enn de andre oppgavene.

Som vist i tabell 8.6, er det gjennomgående høyt samsvar mellom de to vurdererne. Kun én av oppgavene har lavere enighet enn 85 % når det gjelder poengsetting. Mange av oppgavene har en enighet på tett opp mot 100 % på poengnivå. Dette er ikke overraskende, sett i lys av at vurderingen av mange av oppgavene kun består i å avgjøre om elevene har kommet fram til riktig(e) tall eller ikke. Resultatene viser for øvrig at de interne vurdererne gjennomgående er noe mildere i sin poenggiving enn de eksterne vurdererne, men forskjellen er ikke stor. Gjennomsnittlig er det snakk om 0,70 poeng på totalskåren i matematikk, og i forhold til et gjennomsnitt på 27 poeng er det nokså ubetydelig.

Tabell 8.6: Item-analyse for matematikkoppgavene for 7. trinn.
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. 1 kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng					Gjennomsnittlig dyktighet etter poeng						
	Blank	0 poeng	1 poeng	2 poeng	3 poeng	Blank/ 0 poeng	1 poeng	2 poeng	3 poeng	D	R	Komm
1a	3	12	85	-	-	16	28	-	-	0,40	98	
1b	2	13	85	-	-	21	27	-	-	0,21	99	a
2	1	20	79	-	-	17	29	-	-	0,44	98	
3	0	50	50	-	-	21	32	-	-	0,51	98	
4	0	75	25	-	-	24	32	-	-	0,29	98	a
5a	13	42	44	-	-	21	33	-	-	0,51	99	
5b	15	39	46	-	-	21	33	-	-	0,52	99	
6	8	42	50	-	-	22	31	-	-	0,41	85	
7a	10	46	44	-	-	21	33	-	-	0,56	99	
7b	4	29	67	-	-	19	30	-	-	0,45	98	
7c	13	52	35	-	-	22	34	-	-	0,47	98	
8a	2	8	90	-	-	14	28	-	-	0,36	99	
8b	13	41	46	-	-	21	33	-	-	0,54	97	
8c	13	39	48	-	-	23	30	-	-	0,33	95	
9	12	36	52	-	-	21	32	-	-	0,50	89	
10a	1	16	83	-	-	19	28	-	-	0,29	100	a
10b	4	71	26	-	-	25	30	-	-	0,19	99	a
11	6	43	4	47	-	21	35	-	-	0,48	98	
12	1	45	54	-	-	23	30	-	-	0,31	99	
13a	6	42	34	18	-	20	31	36	-	0,60	89	
13b	37	20	43	-	-	22	32	-	-	0,41	89	
14a	3	24	73	-	-	17	30	-	-	0,49	94	
14b	5	21	9	65	-	16	25	31	-	0,56	97	
14c	13	43	17	27	-	21	29	35	-	0,54	77	b
15	7	22	51	10	11	19	26	35	39	0,56	88	
16a	6	29	66	-	-	21	29	-	-	0,33	99	
16b	13	30	58	-	-	22	30	-	-	0,36	99	
16c	15	71	14	-	-	25	37	-	-	0,40	96	
17	12	43	14	31	-	21	25	36	-	0,60	92	
18	9	26	14	20	31	17	26	31	35	0,66	87	
19a	8	46	46	-	-	21	33	-	-	0,55	98	
19b	19	35	46	-	-	24	29	-	-	0,23	94	a
19c	23	68	9	-	-	26	35	-	-	0,25	98	a
20a	24	17	59	-	-	20	31	-	-	0,46	98	
20b	27	12	62	-	-	19	31	-	-	0,50	98	
20c	15	9	76	-	-	16	30	-	-	0,53	97	
20d	40	14	46	-	-	22	32	-	-	0,46	97	
20e	35	28	37	-	-	22	34	-	-	0,54	97	
21a	9	23	68	-	-	18	30	-	-	0,52	95	
21b	11	48	41	-	-	21	34	-	-	0,58	91	
21c	7	16	77	-	-	19	29	-	-	0,35	86	
22a	13	66	21	-	-	24	37	-	-	0,47	98	
22b	24	54	22	-	-	23	37	-	-	0,52	99	
23	16	48	10	26	-	22	28	36	-	0,56	94	

a) Svak diskriminering (<0,30)
b) Dårlig overensstemmelse mellom rettere (< 85 %)

8.2.4 Analyse av foreslåtte skalaer

Tabell 8.7 viser reliabiliteten til totalskalaen (Cronbachs alfa) og de foreslåtte delskalaene for denne prøven. Reliabiliteten til totalskalaen er høy, men for de to første delskalaene er den definitivt ikke høy nok for rapportering. Derimot har skalaen APM rimelig høy reliabilitet. Ut fra dette kunne det være aktuelt å slå sammen de to første for å få to aktuelle delskalaer. For å undersøke om en slik todeling vil være aktuell, har vi studert hvor forskjellige de tre delskalaene er ved å beregne de observerte og latente korrelasjonene (se kapittel 4.5) mellom dem. Resultatene i tabell 8.8 viser at delskalaene APM og RTK har en latent korrelasjon seg imellom på tett opp mot 1,00 og framstår derfor som identiske. Skalaen RSF er imidlertid noe mer forskjellig fra de to andre foreslåtte delskalaene. Dette gjør at en sammenslåing av RSF og RTK vil være meningsløs. Konklusjonen er at bare totalskalaen er aktuell for rapportering.

Tabell 8.7: Reliabilitet (Cronbachs alfa) til totalskala og delskalaer i matematikk for 7. trinn.

Skala	**Reliabilitet**
Representasjonskompetanse og kompetanse i symbolbruk og formalisme (RSF)	0,75
Resonnements-, tankegangs- og kommunikasjonskompetanse (RTK)	0,71
Anvendelses-, problembehandlings- og modelleringskompetanse (APM)	0,83
Totalskala	0,91

Tabell 8.8: Observerte og latente korrelasjoner mellom delskalaer for 7. trinn. Latente korrelasjoner i parentes.

	RSF	RTK
RTK	0,65 (0,89)	
APM	0,74 (0,89)	0,76 (0,99)

Men uavhengig av hvordan resultatene rapporteres er vi bekymret for hvordan disse delskalaene brukes som pedagogisk tilbakemelding. Skåreverdier etter disse delskalaene kan ikke brukes som informasjon om sterke og svake sider av en elevs matematikkompetanse så lenge de framstår med så tvilsom validitet og reliabilitet. For øvrig tror vi at kompetansebetegnelsene er for kompliserte til å formidle pedagogisk mening.

8.3 Matematikk 10. trinn

8.3.1 Prøvens struktur og validitet

Prøven i matematikk for 10. trinn består av to deler, en uten og en med lommeregner. Totalt består prøven av 26 oppgaver, hvorav mange består av flere deloppgaver (heretter kalt oppgaver). Heller ikke i dette oppgavesettet er det mange flervalgsoppgaver; vi finner kun noen få eksempler som oppgave 11 og 26a. Totalt antall oppnåelige poeng på prøven er hele 99.

Hver oppgave er som for de andre prøvene definert som i hovedsak å måle ett av de tre kompetanseområdene RSF, RTK og APM. Maksimalt antall oppnåelige poeng innen de tre kompetanseområdene er henholdsvis 39, 29 og 31 poeng. For hvert

kompetanseområde er det definert 5 *kompetansenivåer* ut fra hvor mange poeng som oppnås på oppgavene knyttet til hvert område.

I likhet med for prøvene for 4. og 7. trinn, kan man undre seg over motivasjonen for det kodesystemet som er valgt i kodeboka. Vi anbefaler å endre systemet i retning av det som brukes i PISA og TIMSS i framtida.

Første del av matematikkprøven for 10. trinn skal løses uten bruk av lommeregner, mens i andre del er lommeregner, passer og gradskive tillatt. Første del tester elevenes ferdigheter i tallregning, inklusive regning med brøk og desimaltall. Ferdigheter i regning med bokstavuttrykk testes også. I andre del av prøven testes elevene i temaer som enkel kombinatorikk, proporsjonalitet, forståelse av tallenes plassering på tallinja, geometriske beregninger, konstruksjon, likningsløsning, sannsynlighet, avlesing av grafiske framstillinger og koordinatsystemet. Etter vår vurdering gir de to delprøvene til sammen et balansert utvalg av grunnskolens matematikkpensum, og vi mener derfor at prøven har høy innholdsvaliditet i forhold til L97.

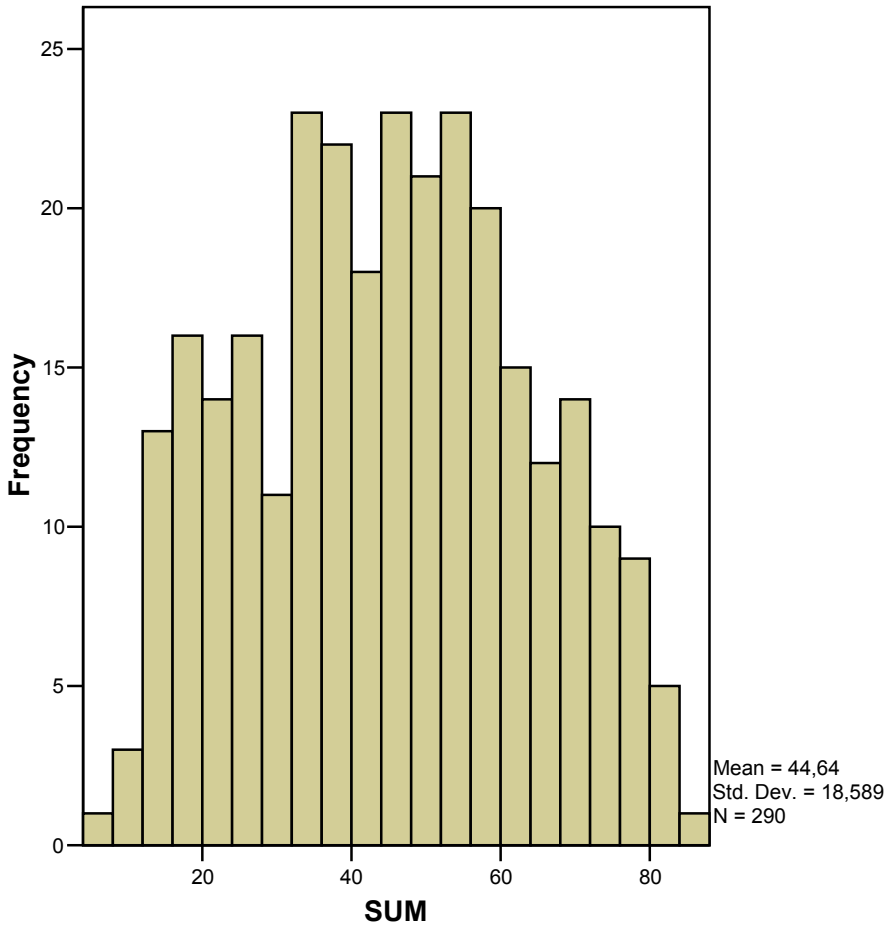
8.3.2 Fordeling av skåreverdier og koder

Figur 8.3 viser hvordan elevenes totalskåre på matematikksprøven fordeler seg. Totalt antall oppnåelige poeng er som tidligere nevnt 99, og gjennomsnittskåre for prøven er 45 poeng. Fordelingen ligger som vist på figuren tett opp til en normalfordeling med et standardavvik på 18,6. Også denne prøven har som helhet en god spredning og skiller godt mellom sterke og svake elever.

Med et gjennomsnitt på 45 % av maksimum framstår prøven som svært vanskelig. Dette er definitivt en mye vanskeligere prøve enn det elevene er vant til. Motivasjonsmessig er trolig dette et stort problem, men psykometrisk sett er det ingen særlig ulempe. Vi vil særlig studere virkningen når det gjelder blanke svar på oppgavene, og fra tabell 8.9 ser vi at av de 66 oppgavene er det for 24 av dem flere enn 20 % og for 13 av dem flere enn 30 % blanke svar. Åpenbart blir den diagnostiske verdien av en oppgave lav hvis mange elever ikke svarer på den.

Kodesystemet som er anvendt i prøven er også på dette trinnet svært omfattende. Som vist i tabell 8.9, er det imidlertid en stor andel av kodene som fanger mindre enn 5 prosent av elevene. Oppgave 26b har for eksempel så mange som 11 koder, men kun to av dem fanger 5 prosent eller flere av elevene. Vurderingsveiledningen bærer preg av at man har laget koder ut fra hva man teoretisk kan tenke seg det er mulig å svare. Empirien viser imidlertid at i svært mange tilfeller slår ikke disse antagelsene til. Samtidig viser det seg at for mange av oppgavene er prosentandelen for kode 99 ”Andre svar” høy. Vi foreslår at antallet koder reduseres betraktelig i framtidige prøver. Ideelt sett bør man pilotere oppgaver og kodesystemer slik at man kan luke bort koder som er svært lite frekvente før prøven gjennomføres i stor skala (for eksempel med kriteriet <5 %).

Figur 8.3: Fordeling av skåre på matematikkprøven på 10. trinn.



Tabell 8.9: Prosentvis svarfordeling på kodenivå for 10. trinn. Koder som fanger mindre enn 5 % av elevene, er skravert. Poengene er her ikke angitt. N=290

Oppgave	0	1	2	11	12	13	14	15	16	17	21	22	23	24	99
1a	1	91									2	0			6
1b	2	51									41	0			7
2a	2	86									5	1			7
2b	4	66									2	15	0	1	11
2c	3	82									5	2			8
2d	33	31									2	1	8		25
3a	6	57	3								18	2	2		14
3b	14	24		34							3	2	1		21
3c	33	16		11	2	1					6	1	0	0	29
3d	37	37									2	7			16
4a	6	48	18								6	15	1	0	7
4b	13	41	13								23	3	1	0	6
5a	2	77									12	2			7
5b	3	83									4	6	0		4
5c	2	62									25	2	2		7
5d	6	35									35	10	0		14
5e	22	34									24	3	0	4	13
6a	15	31		27	0										26
6b	30	23		18	1										28
7a	9	49	20								7	1	0		13
7b	16	35	16								6	1	11		14
7c	26	28									6	3	5	1	31
7d	36	26	7	7							3				22
8a	7	80									4	1			8
8b	16	69									3	5	0	0	6
8c	24	54									3	0	1	6	12
8d	34	19									6	2	10	2	29
9	0	93									1	1	5	0	0
10a	2	70		4							8	7			9
10b	2	84		5							3				5
10c	4	37		1	14						10	8			26
10d	5	2		9	12						7	8	8		49
10e	37	13	15	3							1				30
11	1	72									8	12	7		0
12	0	55									1	34	9		0
13	1	51									37	8	2		0
14	1	65		1	12						9	2			10
15a	2	57									19	6	6	9	1
15b	22	11		11	1						3	5	5		41
16a	9	75		2	1						1	1			11
16b	11	60		4	0						2	6			18
17a	10	56		2	10						3				18
17b	29	4		0	0	0	1	9	2	9	16				29
18a	10	25		13	2	0	22	6			7				14
18b	25	30		9	0						10	2			24
18c	36	20	1	2	3						3	10			24
19	24	31		11							18				17
20a	10	64									2	7	0		16
20b	22	32		20							0	5			21
21a	5	87													9
21b	7	72									8				13
21c	26	10		1							13	15	1		33
21d	36	2		1	2						7	8			45
21e	48	3	1	7							5				35
22a	3	74									8	3	6	1	6
22b	5	71									7	0			17
22c1	7	62									3	8	12	7	1
22c2	23	51	4								9				12
23a	9	69									3	2			17
23b	24	42									7	7	6		14
23c	34	18	1	2	2						3				40
24a	8	63	0								5	4	3		16
24b	12	61	3								8	1			15
25	30	23		6	2	14	2								24
26a	12	55									20	13			
26b	59	1		4	1	1	2	0	0	3	1				27

8.3.3 Analyser av enkeltoppgaver

Tabellene 8.10A og 8.10B presenterer en analyse av alle enkeltoppgavene i matematikkprøven for 10. trinn. Resultatene i tabell 8.10 viser at for de aller fleste oppgavene er poengene riktig ordnet etter elevenes dyktighet. Med dette menes at elevene som får 3 poeng gjennomsnittlig er dyktigere enn de som får 2 poeng, som igjen er dyktigere enn de som får 1 poeng osv. Vi finner faktisk bare ett unntak fra dette, og det gjelder for oppgave 17b. Her har de 2 prosent av elevene som har fått 2 poeng lavere dyktighet enn de elevene som får 1 poeng. De aller fleste av oppgavene har også god diskriminering. Ut fra kriteriet vi har satt, er det bare 7 av totalt 66 oppgaver som ikke har god nok diskriminering, og flere av disse ligger kun marginalt under den valgte grensen på 0,30. Når oppgavene er så lette som noen av disse, er det dessuten ofte i praksis nesten umulig å oppnå høy diskriminering når vi bruker et slikt mål.

Vi ser av tabell 8.10 at seks av oppgavene er ”3 poengs-oppgaver”. Vi kan ikke se fra dataene at dette er godt motivert. For fire av disse er det bare en marginal andel som får 3 poeng, og det ser ikke ut til at diskrimineringen er blitt spesielt høy av å tillegge så mange poeng.

For 154 av elevene forelå vurdering fra både intern og ekstern sensor på det tidspunktet analysene måtte gjennomføres. Sensorreliabiliteten i tabellene 8.10A og 8.10B er basert på analyser av data for disse elevene. Resultatene viser at sensorreliabiliteten gjennomgående er svært god for denne prøven. For mange av oppgavene ligger den på tett opp mot 100 %. Fem av oppgavene har lavere sensorreliabilitet enn den nedre grensen vi har satt på 85 prosent. Den første av disse er oppgave 10e. Spesielt er det her mange uenigheter mellom kode 2 med beskrivelsen ”alle forklaringer der eleven har begrunnet svaret sitt i d skikkelig, selv om svaret i d er feil” og kode 99 ”Andre gale svar”. Kode 2 gir 2 poeng, og dette fører til at en så stor andel som 9 prosent av sensorene har en poengdifferanse seg imellom på 2 poeng. Også oppgave 17a har noe lav sensorreliabilitet. Her skal elevene konstruere en normal til ei linje gjennom et punkt. Flest uenigheter er det her mellom kode 11 for de som har tegnet en normal nøyaktig og kode 99 ”Andre svar”. Trolig er det her problematisk å vite hva som skal regnes som nøyaktig nok. Videre har oppgave 18a den aller laveste sensorreliabiliteten i oppgavesettet. Her skal elevene konstruere en trekant ut fra oppgitt lengde og vinkler. For denne oppgave er det ingen typiske uenigheter som peker seg ut, men det framstår som generelt vanskelig å skille mellom kodene på en entydig måte. Oppgave 18b har også noe lav sensorreliabilitet. Her viser særlig skillet mellom kodene 1 og 11 seg uklart. Forskjellen mellom disse to kodene er om elevene har gitt riktig begrunnelse for svaret eller ”noe tynn begrunnelse eller delvis rett begrunnelse”. I oppgave 26b er det en rekke koder som gir 3, 2 eller 1 poeng, og mange av uenighetene er mellom koder som gir poeng og kode 99 ”Andre svar”.

Tabell 8.10A: Item-analyse for matematikkoppgavene for 10. trinn. N=290. Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng					Gjennomsnittlig dyktighet etter poeng						
	Blank	0 poeng	1 poeng	2 poeng	3 poeng	Blank/ 0 poeng	1 poeng	2 poeng	3 poeng	D	R	Komm.
1a	1	8	91	-	-	33	46	-	-	0,19	99	a
1b	2	47	51	-	-	40	49	-	-	0,26	99	a
2a	2	12	86	-	-	29	47	-	-	0,34	99	
2b	4	30	66	-	-	34	50	-	-	0,40	99	
2c	3	15	82	-	-	30	48	-	-	0,38	100	
2d	33	36	31	-	-	38	59	-	-	0,53	98	
3a	6	35	59	-	-	33	53	-	-	0,52	99	
3b	14	28	34	24	-	37	43	62	-	0,38	94	
3c	33	36	15	16	-	39	48	64	-	0,49	90	
3d	37	26	37	-	-	37	56	-	-	0,48	97	
4a	6	28	66	-	-	32	51	-	-	0,47	95	
4b	13	32	55	-	-	35	53	-	-	0,49	95	
5a	2	20	77	-	-	31	49	-	-	0,39	98	
5b	3	14	83	-	-	27	48	-	-	0,44	100	
5c	2	36	62	-	-	37	50	-	-	0,34	100	
5d	6	59	35	-	-	37	58	-	-	0,52	97	
5e	22	44	34	-	-	39	56	-	-	0,43	100	
6a	15	26	28	31	-	31	51	57	-	0,61	99	
6b	30	28	19	23	-	36	49	63	-	0,61	99	
7a	9	22	70	-	-	33	50	-	-	0,40	99	
7b	16	33	51	-	-	35	54	-	-	0,51	99	
7c	26	46	29	-	-	39	60	-	-	0,51	97	
7d	36	25	7	33	-	37	51	58	-	0,53	88	
8a	7	13	80	-	-	26	49	-	-	0,51	99	
8b	16	15	69	-	-	31	51	-	-	0,51	99	
8c	24	22	54	-	-	34	54	-	-	0,53	97	
8d	34	48	19	-	-	41	60	-	-	0,41	100	
9	0	7	93	-	-	29	46	-	-	0,23	100	a
10a	2	25	4	70		30	43	50	-	0,48	98	
10b	2	9	5	85	-	28	35	47	-	0,34	97	
10c	4	44	15	37	-	34	48	57	-	0,57	85	
10d	5	72	12	9	2	40	54	68	69	0,51	95	
10e	37	31	4	28	-	40	49	56	-	0,40	81	c
11	1	27	72	-	-	31	50	-	-	0,45	100	
12	0	45	55	-	-	35	53	-	-	0,49	100	
13	1	48	51	-	-	36	52	-	-	0,43	100	
14	1	21	13	65	-	30	34	52	-	0,50	91	
15a	2	41	57	-	-	41	48	-	-	0,18	99	a
15b	22	54	12	11	-	40	57	60	-	0,39	83	c
16a	9	13	3	75	-	26	39	51	-	0,56	89	
16b	11	26	4	60	-	29	49	54	-	0,64	91	
17a	10	22	12	56	-	30	40	54	-	0,58	80	c
17b	29	45	20	2	4	40	57	51	65	0,40	89	b
18a	10	21	28	15	25	32	41	56	58	0,56	69	c
18b	25	36	9	30	-	37	47	60	-	0,56	81	c
18c	36	36	6	21	-	38	56	64	-	0,58	89	
19	24	34	11	31	-	37	44	60	-	0,56	95	
20a	10	26	64	-	-	32	51	-	-	0,49	97	
20b	22	26	20	32	-	33	48	60	-	0,65	89	

a) Svak diskriminering (<0,30)
b) Poengene ikke ordnet etter dyktighet
c) Dårlig overensstemmelse mellom vurderingene (< 85 %)

Tabell 8.10B: Item-analyse for matematikkoppgavene for 10. trinn. N=290. Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng					Gjennomsnittlig dyktighet etter poeng						
	Blank	0 poeng	1 poeng	2 poeng	3 poeng	Blank	0 poeng	1 poeng	2 poeng	D	R	Komm.
21a	5	8	87	-	-	29	47	-	-	0,34	99	
21b	7	21	72	-	-	33	49	-	-	0,38	97	
21c	26	63	1	10	-	42	58	66	-	0,40	97	
21d	36	59	2	1	2	44	50	66	76	0,26	97	a
21e	48	41	7	4	-	42	65	67	-	0,39	93	
22a	3	23	74	-	-	37	47	-	-	0,25	97	a
22b	5	24	72	-	-	34	49	-	-	0,36	97	
22c1	7	31	62	-	-	33	52	-	-	0,49	99	
22c2	23	22	55	-	-	32	55	-	-	0,64	92	
23a	9	22	69	-	-	33	50	-	-	0,41	95	
23b	24	34	42	-	-	37	55	-	-	0,48	99	
23c	34	43	4	19	-	39	60	64	-	0,55	91	
24a	8	29	63	-	-	33	52	-	-	0,50	98	
24b	12	24	64	-	-	32	52	-	-	0,52	91	
25	30	23	16	8	23	38	42	50	60	0,50	86	
26a	12	33	55	-	-	41	47	-	-	0,16	96	a
26b	59	28	6	7	1	41	64	72	47	0,42	82	c

a) Svak diskriminering (<0,30)
b) Poengene ikke ordnet etter dyktighet
c) Dårlig overensstemmelse mellom vurderingene (< 85 %)

8.3.4 Analyse av foreslåtte skalaer

I tabell 8.11 har vi angitt noen karakteristiske data for å vurdere og sammenlikne de tre delskalaene. Vi ser for det første at det er svært stor forskjell mellom skalaene, og særlig at RSF-skalaen inneholder desidert mye lettere oppgaver enn de to andre.

Tabell 8.11: Vanskelighetsgrad og reliabilitet til totalskala og delskalaer i matematikk for 10. trinn. N=290

Skala	Max poeng	Gj.snitt poeng	Prosent av max	Reliabilitet (alfa)
Representasjonskompetanse og kompetanse i symbolbruk og formalisme (RSF)	39	22,7	58	0,89
Resonnements-, tankegangs- og kommunikasjonskompetanse (RTK)	29	11,8	40	0,78
Anvendelses-, problembehandlings- og modelleringskompetanse (APM)	31	10,2	33	0,81
Totalskala	99	44,6	45	0,94

Tabell 8.11 viser at reliabiliteten til totalskalaen i matematikk er 0,94 og med andre ord meget tilfredsstillende. Det kan fullt ut forsvares å rapportere etter denne totalskalaen. Av de tre delskalaene har RSF høyest reliabilitet med 0,89, mens de to andre er mye lavere. En nærliggende tanke kunne da være å slå sammen de to siste for å få to reliable delskalaer til rapportering. For å vurdere en slik mulighet vil vi studere hvor forskjellige de tre skalaene framstår ut fra dataene. For dette formålet har vi i tabell 8.12 angitt korrelasjonene mellom de tre, både de observerte og de latente (se kapittel 4.5).

Tabell 8.12: Observerte og latente korrelasjoner mellom delskalaer for matematikk på 10. trinn. Latente korrelasjoner i parentes.

	RSF	RTK
RTK	0,76 (0,92)	
APM	0,82 (0,90)	0,78 (0,99)

Resultatene i tabell 8.12 viser at delskalaene RTK og APM har en latent korrelasjon seg imellom som nesten er perfekt. Dette er for øvrig akkurat samme situasjon som vi fant for prøven for 7. trinn. For disse to prøvene framstår derfor de to skalaene som umulig å skille fra hverandre empirisk, og følgelig vil det være logisk å slå dem sammen. I motsetning til situasjonen på 7. trinn har RSF-skalen høy nok reliabilitet til å stå "på egne bein". Et alternativ for rapportering av resultater fra prøven på 10. trinn er altså i tillegg til den totale skalaen å rapportere etter to delskalaer: RSF og RTK/APM. Men forutsetningen for det er at innholdet i disse to kategoriene, og særlig forskjellen mellom dem, blir kommunisert på en forståelig måte.

8.4 Matematikk på grunnkurs

8.4.1 Prøvenes struktur og validitet

Prøvene for grunnkurs består av en felles del for alle studieretninger som delprøve 1. Del 2 i prøvene er spesifikke for de to variantene (1MX og 1MY) på studieretning for allmenne- og økonomisk-administrative fag, samt matematikk på de andre studieretningene (1M).

Det presiseres at del 1 av prøven skal løses uten elevbok, formelsamling og kalkulator. I denne delen testes elevene i de fire regningsartene anvendt på desimaltall og brøk, samt på bokstavuttrykk. Forenkling av bokstavuttrykk inngår også, samt forståelse av forholdstall. Elevene prøves videre i løsning av likninger, og i forståelse av brøk og potenser. Avslutningsvis inngår en oppgave om relasjonen mellom ulike volumenheter.

Inklusive oppgavene i del 1 består 1M-prøven av totalt 20 oppgaver, med flere deloppgaver (heretter kalt oppgaver). Maksimalt antall oppnåelige poeng på de tre kompetanseområdene RSF, RTK og APM er henholdsvis 23, 15 og 19, noe som gir 57 som totalt antall oppnåelige poeng. For hvert kompetanseområde er det definert kompetansenivåer fra 1 til 5, der 5 er det høyeste oppnåelige nivået. Nivåene er bestemt ut fra antallet oppnådde poeng.

I delprøve 2 for 1M er kalkulator tillatt hjelpemiddel, men ikke elevbok eller formelsamling. Prøven innholder oppgaver om avlesing av søylediagram, geometriske beregninger, beregning av volum, forenkling av bokstavuttrykk, beregning av reallønn, beregning av kostnader basert på faste og variable utgifter, formulering av enkle matematiske modeller og prosentberegninger.

Delprøve 2 for 1MY har noen av de samme oppgavene som prøven for 1M. De samme kompetanseområdene som for 1M-prøven er definert. Totalt i prøven for 1MY er det 22

poeng innen kompetanseområdet RSF, 23 innen RTK og 20 innen APM. Totalt antall poeng på denne prøven er følgelig 65. Prøven inneholder som nevnt noen av de samme oppgavene som for 1M. Her finnes i tillegg oppgaver om lineær regresjon, løsning av likninger og likningssystemer, formulering av funksjonsuttrykk og vinkelberegninger.

Delprøve 2 for 1MX har også felles oppgaver med prøvene for 1M og 1MY. I 1MX-prøven finner vi dessuten oppgaver om løsing av eksponentiallikninger, samt forståelse av grunnleggende trigonometriske begreper og trigonometriske beregninger. Vi finner også en oppgave hvor elevene skal skissere av grafen til en tredjegradsfunksjon og finne koordinatene til toppunkt, bunnpunkt og nullpunkt. Totalt antall oppnåelige poeng i prøven for 1MX er 25 for RSF, 24 for RTK og 19 for APM. Antall oppnåelige poeng for prøven total er følgelig 68.

Også prøvene for 11. trinn innholder mange gode oppgaver som innholdsmessig burde passe godt til de ulike læreplanene på dette trinnet. Vi skal imidlertid se at 1M- og 1MY-prøvene viser seg å falle svært vanskelig ut for elevene.

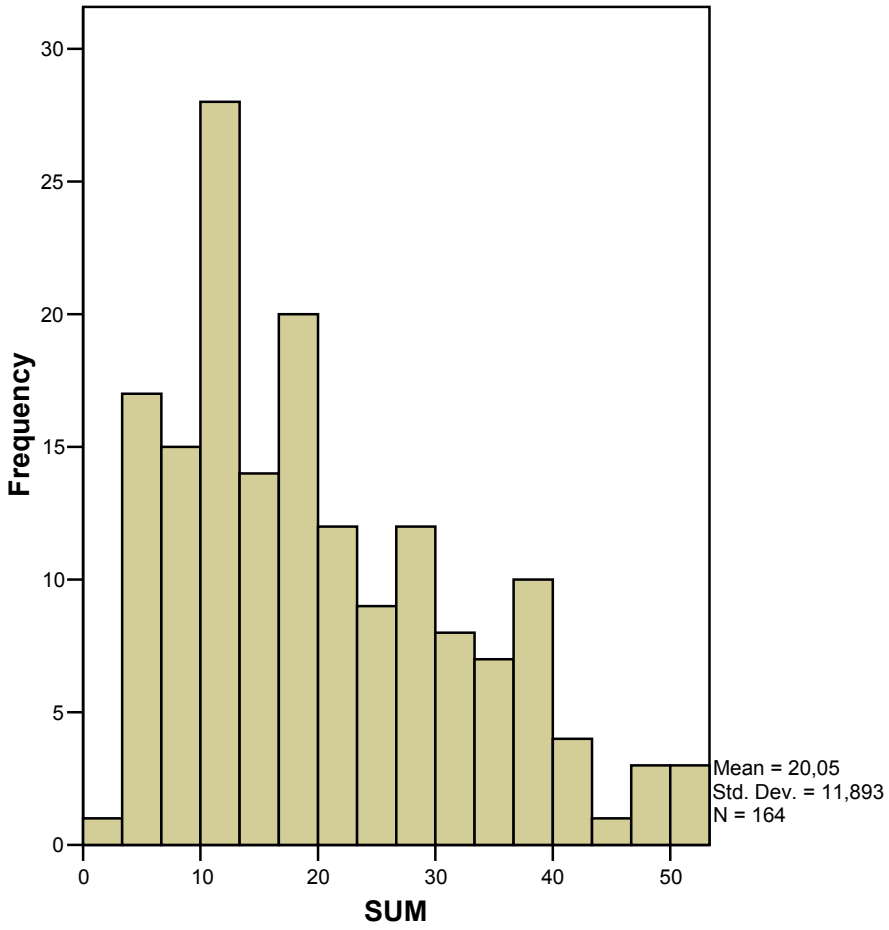
Også på 11. trinn er det opplagt at mange av oppgavene lett kunne vært gitt i flervalgsformatet, og vi vil på det sterkeste anbefale at flere flervalgsoppgaver inkluderes i eventuelle framtidige prøver.

8.4.2 Analyse av 1M-prøven

Fordeling av skåre

Figur 8.5 viser hvordan elevenes totalskåre på matematikkprøven for 1M fordeler seg. Boikott av prøven i kombinasjon med sen innsending av resultater førte til at kun data for 164 elever forelå på det tidspunktet analysene måtte gjennomføres. For mange elever var det også kun enten ekstern eller intern vurdering som var tilgjengelig. Figur 8.5 viser fordelingen av totalskåre på prøven for disse 164 elevene, basert på enten intern vurdering hvis denne forelå, eller på ekstern vurdering. Figuren viser at fordelingen er forskjøvet mot lave verdier og at gjennomnittet ligger på 20 poeng. Totalt antall oppnåelige poeng på prøven er som tidligere nevnt 57. Med bare 35 % av fullt oppnåelig framstår framstår prøven totalt sett som for vanskelig for den aktuelle elevgruppa. Tilsvarende fordelinger finner vi også for de tre delskalaene, men de er ikke vist her.

Figur 8.5: Fordeling av skåre på matematikkprøven for 1M basert på delvis intern og delvis ekstern vurdering.



Svarfordeling på kodenivå

I likhet med grunnskoleprøvene har også 1M-prøven et til dels svært omfattende kodesystem. Tabell 8.13 viser prosentvis fordeling av elevsvarene på de enkelte kodene. Som det går fram av tabellen, er det også her en rekke koder som fanger ubetydelige andeler av elevsvarene. Mange av kodene er tydelig laget ut fra hva man teoretisk kan tenke seg det er mulig svare. Oppgave 6d kan tjene som et eksempel. Her skal elevene trekke sammen uttrykket 4a- (2b + 3a). Kodeboka foreslår her en kode 22 for det å trekke sammen tallene for deretter å føye til bokstavene, noe som leder til 5ab som svar. Det er mulig en slik strategi er påvist i tidligere forskning, men den virker svært lite sannsynlig. I det foreliggende empiriske materialet forekommer den dessuten ikke. Vårt råd er å redusere antallet koder betydelig med utgangspunkt i grundig pilotering av instrumenter og kodesystem.

Tabell 8.13: Prosentvis svarfordeling på kodenivå for 1M-prøven. Koder som fanger mindre enn 5 % av elevene, er skravert. Poengene er her ikke angitt. N=164

Oppgave	0	1	2	3	11	12	13	14	21	22	23	24	99
1a	3	81							0	9			7
1b	6	62							10	2			20
1c	25	24	1						13	1			36
2a	27	41							1	1			31
2b	47	22							5	7			19
2c	58	18							5	0			18
2d	66	14							2	5	1		13
3a	11	78							4				8
3b	15	39							18				28
3c	22	22							5				51
4a	10	60							2	10			18
4b	20	50							1	14			16
5	17	20							31				32
6a	9	48							25				19
6b	11	35							31	7			16
6c	28	22							30				21
6d	23	37							14	0			26
6e	17	29							7				48
6f	43	17							9	1	8		22
7	5	54							8	17	2	12	1
8a	13	59							2	5			21
8b	32	19			0	7			3	0	0		38
9a	32	43											25
9b	38	12	8	7									34
10	7	59							31	2	1		0
11a	1	95											4
11b	1	83											16
11c	31	25			5	1	0		1	7			31
12	18	23			2				12				44
13	17	23			15	21	2		5				16
14a	28	24			2	0			6	1			38
14b	37	24	3		1				7				28
15a	58	1			1	1			7				31
15b	57	6	9										29
16a	31	32							9	13			16
16b	30	30							9	12			20
17a	63	15							0	1			21
17b	62	12			3	0			0	9			15
17c	69	7			1				1				22
18a	15	50			2	4	8	10	1				10
18b	31	37	12										20
18c	42	8			45								6
19a	43	19							8	7			22
19b	70	8							3				20
20a	25	57							1				18
20b	30	34							18	1			17

Analyser av enkeltoppgaver

Tabell 8.14 viser en analyse av alle enkeltoppgavene i 1M-prøven. Vi merker oss at mange av oppgavene har store andeler ikke svart, med så mye som 70 prosent på det meste. Resultatene i tabell 8.14 viser videre at i de aller fleste tilfeller er poengene riktig ordnet etter elevenes dyktighet. Som tidligere nevnt, betyr dette at de elevene som har fått 3 poeng gjennomgående er dyktigere enn de som har fått 2 poeng osv. Oppgavene 12, 14a, 14b og 18c avviker imidlertid fra dette. For disse oppgavene er elevene som får 1 poeng dyktigere enn de som får 2 poeng. For de tre første oppgavenes vedkommende er imidlertid andelen elever som får 1 poeng så liten, at estimatet for dyktigheten blir svært

lite presist. Men nettopp fordi bare 1-2 prosent av elevene får 1 poeng, bør man vurdere å innføre en todelt poenggradering. Det samme bør vurderes for oppgave 18c. Resultatene i tabell 8.14 viser også at så godt som alle oppgavene har tilfredsstillende diskriminering. Oppgaven 11b har lavest diskriminering. Dette henger sammen med at oppgaven er svært lett; 95 prosent av elevene får 1 poeng. Som tidligere nevnt, var ikke begge vurderingene tilgjengelige for alle elevene på det tidspunktet de foreliggende analysene måtte gjennomføres. For 118 elever var både intern og ekstern vurdering tilgjengelig, og sensorreliabiliteten (R) i tabellen er beregnet basert på disse elevene. Som det går fram, er sensorreliabiliteten gjennomgående svært høy, og kun i noen få tilfeller ligger den lavere enn den valgte grensen på 85 prosent. I de fleste tilfeller ligger faktisk reliabiliteten tett opp mot 100 prosent.

Tabell 8.14: Item-analyse for matematikkoppgavene i 1M-prøven.
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng					Gjennomsnittlig dyktighet etter poeng						
	Blank	0 poeng	1 poeng	2 poeng	3 poeng	Blank/0 poeng	1 poeng	2 poeng	3 poeng	D	R	Komm.
1a	3	16	81	-	-	11	22	-	-	0,38	92	
1b	6	32	62	-	-	13	24	-	-	0,43	97	
1c	25	49	26	-	-	17	28	-	-	0,43	94	
2a	27	33	40	-	-	16	26	-	-	0,44	96	
2b	47	32	21	-	-	17	31	-	-	0,50	97	
2c	58	24	18	-	-	17	34	-	-	0,56	100	
2d	66	20	14	-	-	18	35	-	-	0,50	99	
3a	11	10	79	-	-	16	21	-	-	0,19	98	a
3b	15	45	40	-	-	16	25	-	-	0,39	96	
3c	22	55	23	-	-	17	32	-	-	0,54	96	
4a	10	31	59	-	-	13	25	-	-	0,52	92	
4b	20	29	51	-	-	15	25	-	-	0,44	92	
5	17	62	21	-	-	17	33	-	-	0,55	83	b
6a	9	44	47	-	-	15	26	-	-	0,47	97	
6b	11	55	34	-	-	16	28	-	-	0,50	97	
6c	28	50	22	-	-	17	31	-	-	0,48	97	
6d	23	39	38	-	-	16	27	-	-	0,45	95	
6e	17	54	29	-	-	16	29	-	-	0,50	97	
6f	43	40	17	-	-	17	33	-	-	0,49	98	
7	5	41	54	-	-	16	24	-	-	0,32	97	
8a	13	29	59	-	-	13	25	-	-	0,50	98	
8b	32	41	7	20	-	16	28	34	-	0,63	91	
9a	32	35	43	-	-	17	27	-	-	0,58	96	
9b	38	34	28	-	-	17	27	-	-	0,36	84	b
10	7	35	59	-	-	16	23	-	-	0,26	98	a
11a	1	4	95	-	-	16	20	-	-	0,09	99	a!!
11b	1	17	82	-	-	13	21	-	-	0,26	94	a
11c	31	37	7	26	-	15	25	32	-	0,61	91	
12	18	56	2	24	-	16	36	32	-	0,61	96	
13	17	21	24	15	23	13	17	26	29	0,57	85	
14a	28	46	2	24	-	16	34	32	-	0,60	97	
14b	37	36	1	27	-	15	50	32	-	0,63	97	
15a	58	38	2	1	-	19	38	51	-	0,39	97	
15b	57	28	15	-	-	18	32	-	-	0,42	85	
16a	31	37	32	-	-	16	30	-	-	0,54	95	
16b	30	40	30	-	-	15	32	-	-	0,64	97	
17a	63	22	15	-	-	18	34	-	-	0,51	99	
17b	62	23	3	12	-	17	34	37	-	0,58	92	
17c	69	23	8	-	-	18	43	-	-	0,57	93	
18a	15	11	24	50	-	9	21	26	-	0,56	92	
18b	31	20	49	-	-	13	27	-	-	0,56	96	
18c	42	5	45	8	-	12	28	24	-	0,56	84	b
19a	43	38	20	-	-	16	36	-	-	0,67	97	
19b	70	22	8	-	-	18	39	-	-	0,48	97	
20a	25	18	57	-	-	12	26	-	-	0,57	93	
20b	30	35	35	-	-	14	31	-	-	0,66	97	

a) Svak diskriminering (<0,30)
b) Dårlig overensstemmelse mellom vurderingene (< 85 %)

Analyse av foreslåtte skalaer

Tabell 8.15: Reliabilitet til totalskala og delskalaer i 1M-prøven. N=164

Skala	Reliabilitet (Cronbachs alfa)
Representasjoner, symbolbruk og formalisme (RSF)	0,87
Resonnement, tankegang og kommunikasjon (RTK)	0,78
Anvendelse, problembehandling og modellering (APM)	0,81
Totalskala	0,92

Tabell 8.15 viser at reliabiliteten til totalskalaen i matematikk er 0,92, med andre ord meget tilfredsstillende. Det kan fullt ut forsvares å rapportere etter denne totalskalaen. Av de tre delskalaene har RSF høyest reliabilitet med 0,87, en verdi som isolert sett kan invitere til at denne delskalaen kan brukes som rapporteringskategori.

Tabell 8.16: Observerte og latente korrelasjoner mellom delskalaer i 1M-prøven. Latente korrelasjoner i parentes.

	RSF	RTK
RTK	0,67 (0,82)	
APM	0,75 (0,89)	0,76 (0,96)

Resultatene i tabell 8.16 viser at som for de andre prøvene har delskalaene RTK og APM en latent korrelasjon seg imellom på tett opp mot 1,00. Skalaen RSF er imidlertid noe mer forskjellig fra de to andre delskalaene. Dette betyr at det isolert sett kunne forsvares å rapportere etter to delskalaer for denne prøven, henholdsvis RSF og RTK/APM. Imidlertid er rapportering av resultater fra grunnkurs uaktuelt på grunn av høy andel boikott.

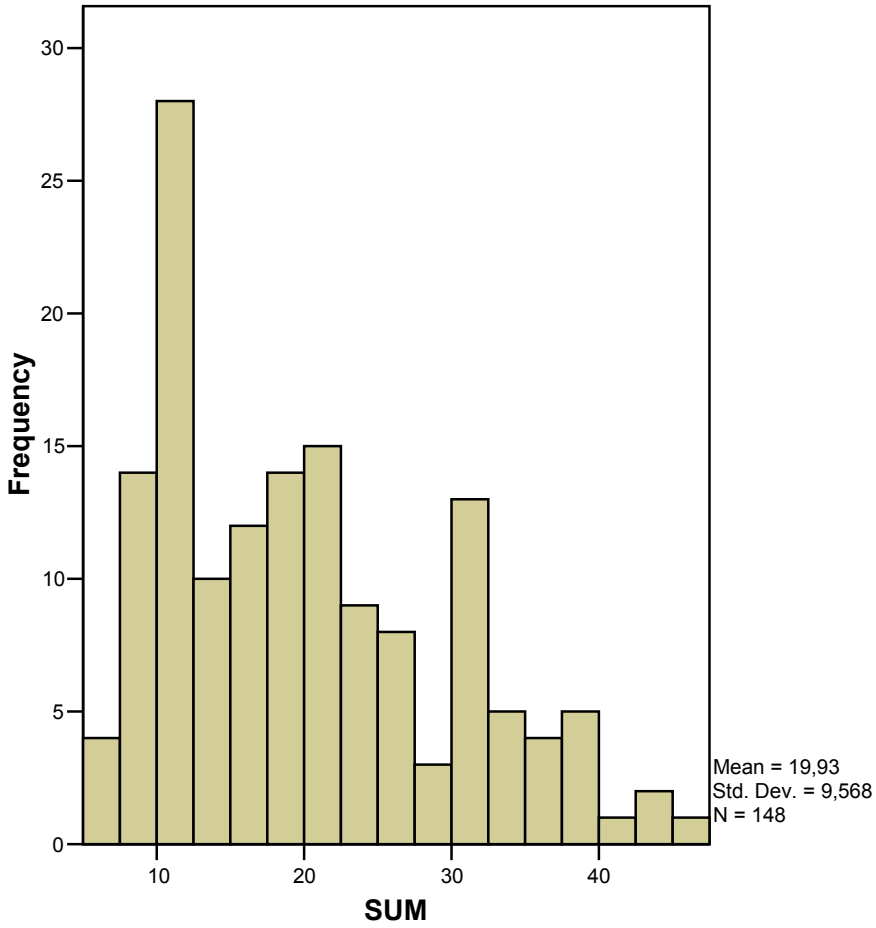
8.4.3 Analyse av 1MY-prøven

Fordeling av skåreverdier og koder

Figur 8.6 viser hvordan elevenes skåre på matematikkprøven for 1MY fordeler seg. På grunn av boikott og sen innsending av resultater forelå kun resultater for 148 elever på det tidspunktet analysene måtte foretas. For mange elever var det også kun enten ekstern eller intern vurdering som foreslå. Resultatene i figur 8.6 baserer seg på intern vurdering for de elevene hvor denne var tilgjengelig, og for øvrig på ekstern vurdering. Maksimalt antall oppnåelige poeng på 1MY-prøven var 65. Figur 8.6 viser at fordelingen er klart forskjøvet mot lave verdier og at gjennomsnittet ligger så lavt som på 20 poeng. Med bare 31 % av "fullt hus" framstår MY- prøven altså som svært vanskelig for den aktuelle elevgruppa. Videre er standardavviket til fordelingen 9,6. Hvis vi sammenlikner med 1M-prøven, ser vi at prøven for 1MY har mindre spredning. Som vi allerede har vært inne på, er det foreslått tre delskalaer basert på 1MY-prøven: "Representasjoner, symbolbruk og formalisme" (RSF), "Resonnement, tankegang og kommunikasjon" (RTK) og "Anvendelse, problembehandling og modellering" (APM). Studier av fordelingene til de enkelte skalaene avdekker interessante forskjeller. Totalt antall oppnåelige poeng på skalaen RSF er 22. Gjennomnittet for denne skalaen er 9 poeng, mens standardavviket er 4,5. Fordelingene for RTK og APM avviker markant fra fordelingen for RSF. Disse to

fordelingene, og særlig den siste, er sterkt forskjøvet mot lave verdier. Totalt antall oppnåelige poeng for RTK er 23, og her ligger gjennomsnittet på 7. Tilsvarende tall for APM er henholdsvis 20 og 4. Standardavviket for RTK er 3,3, mens det er 3,5 for APM.

Figur 8.6: Fordeling av skåre på matematikkprøven for 1MY basert på delvis intern og delvis ekstern vurdering.



Tabell 8.17 viser prosentvis fordeling av elevsvarene i de enkelte kodene. Som det går fram av tabellen, er det også her en rekke koder som fanger ubetydelige andeler av elevene. Igjen er det derfor på sin plass å etterlyse en grundig pilotering av oppgaver og kodesystem før anvendelse i stor skala. Mange av kodene er også her tydelig laget ut fra hva man teoretisk kan tenke seg det er mulig svare. Ofte slår imidlertid ikke disse antakelsene til. Samtidig er det store prosentandeler kode 99 ”Andre svar” for mange av oppgavene. Vårt råd er å redusere antallet koder betydelig i framtidige prøver med utgangspunkt i grundig pilotering av instrumentene.

Tabell 8.17: Prosentvis svarfordeling på kodenivå for 1MY-prøven. Koder som fanger mindre enn 5 % av elevene, er skravert. Poengene er her ikke angitt.

Oppgave	0	1	2	3	11	12	13	21	22	23	24	99
1a	0	88						0	7			6
1b	1	73						11	1			14
1c	8	39	2					16	1			33
2a	15	51						2	2			30
2b	35	17						1	7			39
2c	42	18						9	2			29
2d	48	18						1	1	1		31
3a	14	70						6				11
3b	18	29						28				25
3c	22	22						4				52
4a	9	68						0	14			9
4b	15	55						1	22			9
5	14	14						45				28
6a	3	70						17				11
6b	9	32						34	13			13
6c	16	43						24				18
6d	18	42						11	0			29
6e	15	36						5				44
6f	22	27						8	6	15		22
7	5	57						5	22	3	8	0
8a	10	71						2	5			13
8b	25	36			0	6		7	1	1		24
9a	20	44										36
9b	26	27	5	10								32
10	3	64						28	2	3		0
11a	0	97										3
11b	1	90										10
11c	20	29			6	1	0	2	3			40
12	16	19			7			11				47
13	14	34			11	17	3	6				15
14a	32	13			6	1		7	1			40
14b	43	14	6		1			22				15
15a	59	1			0	0		3				37
15b	60	9	5									25
16a	59	7						0				34
16b	72	7	1	4								17
16c	82	4			7							7
17a	66	15						0	1			19
17b	63	13			5	1		1	4			14
17c	73	2			3			1				21
18	37	1			4							57
19a	57	19			2							22
19b	77	4			3							16
20a	26	34						13	3	0		24
20b	45	4						9				42
21	27	8	3		24			13	21			5
22a	33	34						7				25
22b	38	9						7				46
22c	52	9						16				22
22d	64	5			11							21

Analyser av enkeltoppgaver

Tabell 8.18 presenterer en analyse av de enkelte oppgavene som inngår i prøven for 1MY. Tabellen viser at for alle oppgavene er poengene riktig ordnet etter elevenes dyktighet. Elevene som får 3 poeng er i gjennomsnitt dyktigere enn de som får 2 poeng osv. Resultatene viser videre at om lag en tredel av oppgavene har for lav diskriminering sett i forhold til det kriteriet vi har satt (D<0,30). Som allerede nevnt, var ikke alltid både ekstern og intern vurdering tilgjengelig for elevene i utvalget på det tidspunktet analysene måtte gjennomføres. Sensorreliabiliteten (R) i tabell 8.18 er beregnet på basis av ekstern og intern vurdering for bare 89 elever. Resultatene viser at sensorreliabiliteten gjennomgående er svært god, og for de fleste av oppgavene ligger enigheten opp mot 100 prosent. Kun to av oppgavene faller under den valgte kritiske grensen på 85 prosent. Lavest enighet finner vi for oppgave 15b. Her skal elevene sette et merke på en figur omtrent der hvor en løper i ytre bane starter. Korrekt svar er i kodeboka definert som ”rimelig god avmerking”, videre definert som cirka 60 grader eller bortimot en tredel av svingen. På bakgrunn av denne definisjonen er det ikke merkelig at sensorreliabiliteten blir noe lav.

Tabell 8.18: Item-analyse for matematikkoppgavene i 1MY-prøven.
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng					Gjennomsnittlig dyktighet etter poeng						
	Blank	0 poeng	1 poeng	2 poeng	3 poeng	Blank/0 poeng	1 poeng	2 poeng	3 poeng	D	R	Komm.
1a	0	12	88	-	-	17	20	-	-	0,13	93	a
1b	1	26	73	-	-	15	22	-	-	0,29	97	a
1c	8	51	41	-	-	18	23	-	-	0,24	99	a
2a	15	34	51	-	-	16	24	-	-	0,39	92	
2b	35	48	17	-	-	19	26	-	-	0,29	97	a
2c	42	40	18	-	-	18	30	-	-	0,49	100	
2d	48	34	18	-	-	18	28	-	-	0,40	98	
3a	14	16	70	-	-	16	22	-	-	0,27	99	a
3b	18	53	29	-	-	18	25	-	-	0,31	96	
3c	22	56	22	-	-	18	28	-	-	0,45	98	
4a	9	23	68	-	-	15	22	-	-	0,36	96	
4b	15	30	55	-	-	16	24	-	-	0,43	96	
5	14	72	14	-	-	18	30	-	-	0,41	96	
6a	3	27	70	-	-	15	22	-	-	0,31	98	
6b	9	59	32	-	-	17	26	-	-	0,43	97	
6c	16	41	43	-	-	16	25	-	-	0,46	94	
6d	18	40	42	-	-	17	25	-	-	0,41	96	
6e	15	49	36	-	-	16	26	-	-	0,48	99	
6f	22	51	27	-	-	18	26	-	-	0,41	97	
7	5	38	57	-	-	17	22	-	-	0,25	98	a
8a	10	19	71	-	-	14	22	-	-	0,38	98	
8b	25	34	6	36	-	15	24	26	-	0,56	87	
9a	20	36	44	-	-	16	25	-	-	0,47	99	
9b	26	32	42	-	-	19	22	-	-	0,17	96	
10	3	30	64	-	-	16	22	-	-	0,29	100	a
11a	0	3	97	-	-	8	20	-	-	0,23	100	a
11b	1	10	90	-	-	15	20	-	-	0,18	97	a
11c	20	45	7	29	-	16	23	28	-	0,56	92	
12	16	58	7	19	-	17	21	29	-	0,47	89	
13	14	19	20	11	34	14	19	23	26	0,53	81	b
14a	32	48	7	13	-	18	26	30	-	0,48	97	
14b	43	47	1	20	-	17	28	31	-	0,59	96	
15a	59	40	0	1	-	20	-	39		0,19	100	a
15b	60	25	15	-	-	19	25	-	-	0,24	77	a, b
16a	59	34	7	-	-	19	31	-	-	0,31	100	
16b	72	16	12	-	-	19	31	-	-	0,41	93	
16c	82	7	7	4	-	19	30	31	-	0,36	92	
17a	66	19	15	-	-	18	28	-	-	0,37	94	
17b	63	19	5	13	-	18	23	33	-	0,53	96	
17c	73	22	5	-	-	19	32	-	-	0,29	96	a
18	37	58	4	1	-	19	34	-	-	0,23	97	a
19a	57	22	2	19	-	17	22	30	-	0,53	98	
19b	77	16	3	4	-	19	21	34	-	0,30	93	
20a	26	40	34	-	-	16	28	-	-	0,60	96	
20b	45	51	4	-	-	19	31	-	-	0,24	91	a
21	27	38	24	11	-	18	20	30	-	0,32	88	
22a	33	33	34	-	-	17	25	-	-	0,41	98	
22b	38	53	9	-	-	19	29	-	-	0,29	99	a
22c	52	39	9	-	-	19	29	-	-	0,28	100	a
22d	64	21	11	5	-	19	26	29	-	0,31	92	a

a) Svak diskriminering (<0,30)
b) Dårlig overensstemmelse mellom vurderingene (< 85 %)

Analyse av foreslåtte skalaer

Tabell 8.19: Reliabilitet til totalskala og delskalaer i 1MY-prøven.

Skala	Reliabilitet (Cronbachs alfa)
Representasjoner, symbolbruk og formalisme (RSF)	0,79
Resonnement, tankegang og kommunikasjon (RTK)	0,69
Anvendelse, problembehandling og modellering (APM)	0,68
Totalskala	0,87

Tabell 8.19 viser reliabiliteten til totalskalaen i matematikk er 0,87, og med andre noe lavere enn for 1M-prøven. Av de tre delskalaene har RSF høyest reliabilitet med 0,79. Ingen av delskalaene har høy nok reliabilitet til at det kan forsvares å rapportere etter disse. Riktignok er disse verdiene basert på et relativt lavt antall besvarelser (N=148), men her er tendensene så tydelige at konklusjonen synes holdbar.

Tabell 8.20: Observerte og latente korrelasjoner mellom delskalaer i 1MY-prøven. Latente korrelasjoner i parentes.

	RSF	RTK
RTK	0,49 (0,66)	
APM	0,57 (0,78)	0,67 (0,99)

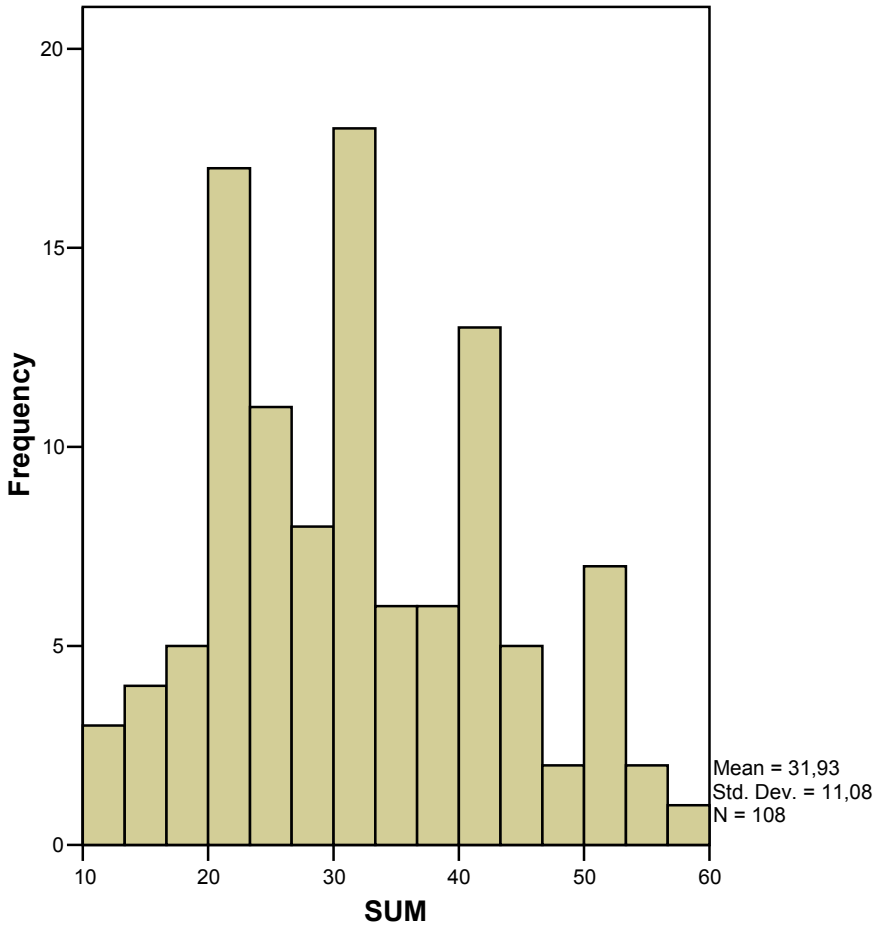
Resultatene i tabell 8.20 viser at delskalaene RTK og APM har en latent korrelasjon seg imellom på tett opp mot 1,00. Skalaen RSF framstår imidlertid som tydelig forskjellig fra de to andre delskalaene.

8.4.4 Analyse av prøven for 1MX

Fordeling av skåreverdier og koder

Figur 8.7 viser fordeling av skåre for 1MX-prøven. Også her la boikott og sen innsending av resultater begrensninger på antallet elever som kunne inkluderes i analysene. Fordelingen i figur 8.7 er basert på intern eller ekstern vurdering for bare 108 elever. I de tilfeller hvor ikke den interne vurderingen forelå, ble ekstern vurdering lagt til grunn. Figur 8.7 viser at fordelingen har et gjennomsnitt på 32 poeng og standardavvik på 11,1. Totalt antall oppnåelige poeng på prøven var 68, som tidligere nevnt. Det er interessant å observere at vi ikke finner den samme forskyvningen av fordelingen mot lave skåreverdier som vi fant for 1M- og 1MY-prøvene. Studier av fordelingene for delskalaene ”Representasjon, symbolbruk og formalisme (RSF)”, ”Resonnement, tankegang og kommunikasjon (RTK)” og ”Anvendelse, problembehandling og modellering (APM)” viser at fordelingen for RSF er noe forskjøvet mot høye skårverdier, mens de to andre fordelingene er noe forskjøvet mot lave skåreverdier.

Figur 8.7: Fordeling av skåre på matematikkprøven for 1MX basert på delvis intern og delvis ekstern vurdering.



Tabell 8.21 viser hvordan elevens svar fordeler seg i de ulike kodene som er definert i kodeboka. Igjen er et svært omfattende kodesystem anvendt, men mange av kodene forekommer nesten ikke i det empiriske materialet. Kodesystemet burde følgelig vært forenklet basert på grundig pilotering.

Tabell 8.21: Prosentvis svarfordeling på kodenivå for 1MX-prøven. Koder som fanger mindre enn 5 % av elevene, er skravert. Poengene er her ikke angitt.

Oppgave	0	1	2	3	11	12	13	21	22	23	24	99
1a	0	91						1	4			4
1b	0	85						5	2			7
1c	6	65	3					4	0			22
2a	2	80						4	2			13
2b	11	59						5	4			20
2c	15	55						3	1			26
2d	19	55						0	3	2		21
3a	2	87						4				6
3b	6	55						16				23
3c	10	58						2				31
4a	1	81						0	6			13
4b	5	74						2	12			6
5	3	33						37				26
6a	1	86						9				4
6b	2	60						23	7			8
6c	5	71						15				10
6d	6	76						4	0			15
6e	2	73						6				19
6f	3	53						8	6	14		17
7	0	76						4	11	2	6	1
8a	2	90						1	3			4
8b	4	62			0	8		0	6	0		20
9a	9	68										23
9b	16	30	11	11								32
10	2	79						16	1	1		0
11a	0	99										1
11b	0	93										7
11c	10	46			14	1	1	2	2			25
12	2	42			6			20				30
13	3	62			12	5	5	4				9
14a	9	48			10	0		7	4			23
14b	16	45	2		6	0		12	0			20
15a	35	8	2		2	7		13				35
15b	56	3	3		0	0		2				38
16a	36	41						0				23
16b	40	28	2	10								19
16c	47	11			28							15
17a	25	44						6				25
17b	50	28			2							20
18	15	12			20							52
19	56	11			9							24
20a	31	28						13				28
20b	56	10			5	4						24
20c	62	3			3							31
21a	20	16	7		30			16	11			0
21b	60	2	4		0	4		3	1	3		25
22a	24	50										26
22b_i	48	26	4	2				1	0			20
22b_ii	53	28	0					1				19
22b_iii	50	19	2	8								21

Analyser av enkeltoppgaver

Tabell 8.22 viser en analyse av de enkelte oppgavene i 1MX-prøven. Tabellen viser at, med kun ett unntak, er poengene riktig ordnet etter elevens gjennomsnittlige dyktighet på prøven totalt sett. Elevene som har fått 3 poeng er i gjennomsnitt dyktigere enn de som har fått 2 poeng osv. Det eneste unntaket er for oppgave 21a. Her er det ingen forskjell i dyktighet mellom elevene som har fått 1 poeng og 0 poeng. Dette skillet kan derfor

vanskelig forsvares. Om lag en firedel av oppgavene har lavere diskriminering enn grensen vi har satt på 0,30. Oppgave 11a er den letteste i oppgavesettet med en p-verdi 0,99. I psykometrisk forstand er denne oppgaven unødvendig, i og med at man i praksis "gir bort" et poeng til alle elevene. Oppgaven tester om elevene kan foreta en enkel avlesning fra et søylediagram, noe mange vil ha automatisert allerede på barnetrinnet.

Sensorreliabiliteten (R) i tabell 8.22 er beregnet på basis av de 63 elevene hvor både ekstern og intern vurdering forelå på det tidspunktet analysene måtte gjennomføres. Resultatene i tabellen viser at også for denne prøven ligger sensorreliabiliteten gjennomgående svært høyt, og i de fleste tilfeller tett opp mot 100 prosent. Det er bare noen få oppgaver som faller under vår valgte grense på 85 %. Oppgave 16c skiller seg imidlertid spesielt negativt ut. Her skal elevene drøfte en modells gyldighet. Kodeboka gir 2 poeng til elevene som "drøfter flere momenter, har gode begrunnelser". 1 poeng gis til elever som "har få momenter eller mangler begrunnelser". Andre svar får her 0 poeng. Sensorreliabiliteten for denne oppgaven er så lav som 60 prosent, og dataene viser at det har vært uklart for sensorene hvor skillet mellom 0 og 1 poeng skal gå.

Tabell 8.22: Item-analyse for matematikkoppgavene i 1MX-prøven.
Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. R for prosentandelen der de to sensorene har vurdert likt. I kolonnen for kommentarer (Komm.) er det henvist til ulike fotnoter under tabellen.

Opp-gave	Prosentvis svarfordeling etter poeng					Gjennomsnittlig dyktighet etter poeng						
	Blank	0 poeng	1 poeng	2 poeng	3 poeng	Blank/0 poeng	1 poeng	2 poeng	3 poeng	D	R	Komm.
1a	0	8	92	-	-	21	33	-	-	0,30	98	
1b	0	14	86	-	-	24	33	-	-	0,30	98	
1c	6	26	68	-	-	28	34	-	-	0,27	90	a
2a	2	18	80	-	-	26	33	-	-	0,26	94	a
2b	11	31	58	-	-	28	35	-	-	0,28	97	a
2c	15	32	53	-	-	26	37	-	-	0,47	100	
2d	19	30	51	-	-	27	37	-	-	0,47	93	
3a	2	8	90	-	-	26	33	-	-	0,18	97	a
3b	6	39	55	-	-	25	37	-	-	0,55	92	
3c	10	34	57	-	-	25	37	-	-	0,54	95	
4a	1	19	80	-	-	31	32	-	-	0,05	95	a!
4b	5	22	73		-	25	34	-	-	0,36	90	
5	3	62	35	-	-	26	42	-	-	0,68	92	
6a	1	14	85	-	-	25	33	-	-	0,27	95	a
6b	2	40	58	-	-	26	36	-	-	0,48	94	
6c	5	27	69	-	-	30	34	-	-	0,31	97	
6d	6	21	73	-	-	24	35	-	-	0,41	98	
6e	2	27	71	-	-	26	34	-	-	0,25	90	
6f	3	44	53	-	-	27	37	-	-	0,47	95	
7	0	22	78	-	-	26	34	-	-	0,31	97	
8a	2	9	89	-	-	27	33	-	-	0,16	98	a
8b	4	30	7	58	-	26	29	36	-	0,44	87	
9a	9	23	68	-	-	26	35	-	-	0,37	90	
9b	16	30	54	-	-	28	36	-	-	0,35	89	
10	2	17	81	-	-	23	34	-	-	0,40	100	
11a	0	1	99	-	-	-	32	-	-	-	100	
11b	0	7	93	-	-	23	33	-	-	0,24	97	a
11c	10	26	18	46	-	28	33	35	-	0,28	95	a
12	2	52	6	41	-	27	40	37	-	0,43	85	
13	3	13	8	13	63	25	27	30	35	0,34	76	b
14a	9	35	9	46	-	25	31	39	-	0,60	90	
14b	16	32	5	47	-	24	36	39	-	0,65	89	
15a	35	50	8	7	-	30	43	46	-	0,46	97	
15b	56	43	3	-	-	32	42	-	-	0,16	100	a
16a	36	25	39	-	-	27	40	-	-	0,59	93	
16b	40	25	35	-	-	27	41	-	-	0,60	93	
16c	47	17	26	10	-	27	40	43	-	0,60	60	b
17a	25	32	43	-	-	27	39	-	-	0,54	96	
17b	50	26	2	22	-	28	32	44	-	0,58	86	
18	15	56	17	12	-	29	38	43	-	0,48	92	
19	56	24	9	11	-	29	42	44	-	0,48	85	
20a	31	40	29	-	-	28	41	-	-	0,49	96	
20b	56	25	5	6	9	29	40	42	48	0,57	94	
20c	62	32	4	2	-	31	46	51	-	0,34	96	
21a	20	30	28	22	-	29	29	42	-	0,41	78	b
21b	60	32	3	6	-	31	39	46	-	0,32	97	
22a	24	31	45	-	-	29	35	-	-	0,24	86	a
22b_i	48	22	30	-	-	31	34	-	-	0,14	94	a
22b_ii	53	21	26	-	-	31	35	-	-	0,16	94	a
22b_iii	50	23	27	-	-	30	38	-	-	0,35	83	b

a) Svak diskriminering (<0,30)
b) Dårlig overensstemmelse mellom vurderingene (< 85 %)

Analyse av foreslåtte skalaer

Tabell 8.23: Reliabilitet til totalskala og delskalaer i 1MX-prøven.

Skala	Reliabilitet (Cronbachs alfa)
Representasjoner, symbolbruk og formalisme (RSF)	0,78
Resonnement, tankegang og kommunikasjon (RTK)	0,73
Anvendelse, problembehandling og modellering (APM)	0,71
Totalskala	0,88

Tabell 8.23 viser reliabiliteten til totalskalaen i matematikk er 0,88, med andre noe lavere enn for 1M-prøven og omtrent som for IMY-prøven. Av de tre delskalaene har RSF høyest reliabilitet med 0,78. Ingen av delskalaene har høy nok reliabilitet til at det kan forsvares å rapportere etter disse. Her må vi imidlertid ta et forbehold på grunn av det begrensende antallet elever som kunne inkluderes i analysene (N=108).

Tabell 8.24: Observerte og latente korrelasjoner mellom delskalaer i 1MX-prøven. Latente korrelasjoner i parentes.

	RSF	RTK
RTK	0,64 (0,85)	
APM	0,54 (0,73)	0,67 (0,93)

Resultatene i tabell 8.24 viser at delskalaene RTK og APM har en latent korrelasjon seg imellom på tett opp mot 1,00. Skalaen RSF er imidlertid mer forskjellig fra de to andre delskalaene.

8.5 Oppsummering

De seks prøvene som er analysert i dette kapitlet, varierer sterkt når det gjelder vanskelighetsgrad for de ulike elevgruppene. Dette er oppsummert i tabell 8.25. Tabellen viser gjennomsnittet for de enkelte prøvene i prosent av full skåre. Tabellen viser at prøvene for 4. trinn, 7. trinn og 11. trinn (1MX) har omtrent tilsvarende vanskelighetsgrader ved at gjennomsnittet ligger rundt 50 prosent av full skåre. Prøven for 4. trinn er imidlertid klart lettere, mens prøvene for variantene 1M og 1MY for grunnkurs har falt svært vanskelig ut.

Rapporten fra MMI om de nasjonale prøvene i 2005 viser at de nasjonale prøvene i matematikk er blitt oppfattet som vanskeligst av elevene. Over halvparten av de spurte elevene (10. trinn og grunnkurs) karakteriserer disse prøvene som meget eller ganske vanskelige. Elevene på grunnkurs synes at matematikkprøvene var noe vanskeligere enn elevene på 10. trinn, noe som er rimelig sett i lys av våre analyser. Videre mener 40 % av lærerne på 10. trinn og grunnkurs at prøvene er meget eller ganske vanskelige. Tilsvarende prosenttall for 4. og 7. trinn er 51.

På samme måte som for de andre trinnene, er det ikke lett å se at prøvene for 1MY og 1M innholdsmessig ligger på siden av det som læreplanene legger vekt på. Sammenliknet med prøvene i de andre fagene viser MMI-undersøkelsen at mange lærere mener oppgavene i matematikkprøvene generelt reflekterer læreplanen i faget. 63 % av lærerne mener prøvene i meget eller ganske stor grad reflekterer læreplanen. De støtter her med

andre ord vår vurdering av at prøvene har høy validitet i forhold til læreplanen. På den annen side svarer bare 31 % av lærerne at elevene i meget eller ganske store grad har arbeidet med liknende oppgaver på forhånd. 40 % mener at elevene i meget eller ganske stor grad er kjent med oppgaveformen som ble brukt i prøven. Ut fra dette kan uvante oppgavetyper være noe av forklaringen på den høye vanskelighetsgraden til de fleste av matematikkprøvene. Dette bør tas med i betraktningen ved utformingen av eventuelle framtidige prøver.

Tabell 8.25: En sammenlikning av vanskelighetsgraden til prøvene

Klassetrinn	**Gjennomsnitt på prøven i prosent av full skåre**
4. trinn	66 %
7. trinn	50 %
10. trinn	45 %
Grunnkurs (1MX)	47 %
Grunnkurs (1MY)	31 %
Grunnkurs (1M)	35 %

Alle seks prøvene har gjennomgående svært høy sensorreliabilitet. Dette er ikke overraskende, i og med at rettingen av mange av oppgavene kun består i å avgjøre om elevene har kommet fram til riktig tall eller ikke. Som vi har vært inne på flere ganger, er kodesystemet som anvendes i alle prøvene svært omfattende. For alle prøvene bør kodesystemet forenkles betydelig. Svært mange av kodene forekommer praktisk talt ikke i det empiriske materialet. Denne typen koder bør lukes bort gjennom pilotering før prøven gjennomføres i stor skala. Færre koder vil kunne bidra til å redusere tiden det tar for lærerne å sette seg inn i systemet og følgelig den totale tidsbruken på rettingen. Vi anbefaler også å innføre et mer enhetlig kodesystem etter mønster av det som er vanlig i internasjonale komparative studier i matematikk. Dette har klare fordeler for lærerne som skal anvende systemet, og ikke minst i videre dataanalyse. Det vil også være ressursbesparende å omformulere mange av oppgavene til tradisjonelle flervalgsoppgaver som ikke krever ekspertbedømmelse. For mange av oppgavene vil dette være svært enkelt å gjøre.

Generelt tilfredsstiller prøvene som helhet grunnleggende psykometriske krav. Det må likevel nevnes at særlig prøvene for 1MX og 1MY har mange oppgaver med lav diskriminering. Flere av disse oppgavene er enten så enkle at ”alle” får poeng eller så vanskelige at ”ingen” får poeng. Denne typen oppgaver har liten verdi både rent psykometrisk og pedagogisk.

Tabell 8.26: Reliabiliteten til prøvene i matematikk

Klassetrinn	Reliabilitet (Cronbachs alfa)
4. trinn	0,88
7. trinn	0,91
10. trinn	0,94
11. trinn (1MX)	0,88
11. trinn (1MY)	0,87
11. trinn (1M)	0,92

Reliabiliteten til prøvene ligger rundt 0,90, noe som danner et godt grunnlag for å rapportere totalskåren for den enkelte prøve, se tabell 8.26. For alle prøvene finner vi svak diskriminerende validitet mellom de foreslåtte delskalaene ”Resonnement, tankegang og kommunikajson” og ”Anvendelse, problembehandling og modellering”. De latente korrelasjonene ligge tett opp mot 1,00 for alle prøvene. Det er derfor ikke empirisk belegg for å hevde at disse skalaene måler ulike dimensjoner. I alle prøvene er imidlertid delskalaen ”Representasjon, symbolbruk og formalisme” noe mer forskjellig fra de to andre delskalaene. I de fleste tilfellene er likevel reliabiliteten til denne delskalaen lavere enn 0,80, noe som gjør det problematisk å rapportere etter skalaen. I to tilfeller ligger imidlertid reliabiliteten til delskalaen tett opp mot 0,90, nemlig på 10. trinn (0,89) og grunnkurs, 1M (0,87). Ut fra dette kan det være grunnlag for å rapportere to delskalaer for disse prøvene. Forutsetningen er imidlertid at innholdet i de to skalaene kan kommuniseres på en forståelig måte.

Tabell 3.2 i kapittel 3 viser at andelen boikott på grunnkurs har vært høy; 36 % for 1MX, 43 % for 1MY og 45 % for 1M. Generelt sett mener vi derfor at det ikke er grunnlag for å rapportere skoleresultater for grunnkurs på Skoleporten. Med så stort frafall er det grunn til å anta at det er store skjevheter hva gjelder hvilke elever som gjennomførte prøven. Sammenlikninger blir derfor lett helt misvisende.

Vedlegg: Uttalelser fra to rektorer

Uttalelse fra en rektor på en barneskole

Jeg må si med en gang, at jeg er blant de som har blitt begeistra for de nasjonale prøvene, selv om vi ikke akkurat gledet oss i forkant. Det jeg ser er at nasjonale prøver gir læring for de voksne. Denne læringen går på vurderingskompetanse. Lærere som bare har vært på barneskoler har ikke den samme vurderingskompetanse som lærere på ungdomstrinnet. Dette ser jeg fordi jeg også har vært rektor på en kombinert barne- og ungdomsskole. Jeg ser at barneskolelærerne har fått ord og begreper de ikke hadde før. I norsk skriftlig har jeg en lærer som har hatt en aha-opplevelse på 7. trinn, nettopp i forhold til ord og faglige begreper i vurderingsarbeidet. Som Kjell Lars Berge sa på kurs, får kanskje ikke lærerne vite noe nytt, fordi de vet det samme om elevene fra før av. På den annen side er dette lærernes personlige oppfatninger av elevene, og man kan stille spørsmål om det er på det nivået man skal være sammenlignet med andre skoler.

Både jeg og lærerne er enig om at dette har vært god etterutdanning for lærerne! Men vi har brukt mye tid. Når vi ser på helheten, mener vi likevel at dette har vært nyttig tid.
Som skoleleder har jeg lagt til rette for at teamene skulle planlegge perioden de skulle gjennomføre nasjonale prøver. Vi har sagt at de lærerne skulle frigjøre hverandre, og vi måtte ha beskjed når de trengte en ekstra lærer inn for å hjelpe elevene. Lærerne har frigjort hverandre slik at de har sittet halve dager av gangen og rettet på skolen. I utgangspunktet trodde vi at det var matematikk som ville ta mest tid, men de har korrigert noe på rettingen siden i fjor, så matematikken viste seg å ta minst tid. Engelsk skriftlig tok heller ikke så lang tid, det gjorde derimot norsken. Læreren som hadde både lesing og skriving på 7. trinn, brukte lang tid på rettingen.

Når det gjelder lærerne på 4. trinn har ikke de hatt samme aha-opplevelse i forhold til vurdering, fordi en av dem er ekspertvurderer og kunne mye fra før. Jeg har vært heldig, fordi jeg har tre ekspertvurderere på min skole. Jeg har derfor fått opplysninger om de nasjonale prøvene med en gang den har kommet og jeg spurt dem og fått raskt svar. De har også vært nyttige og inspirerende både for meg og resten av skolen. Vi føler sånn sett at vi er ”på banen”, og vi tenker at dette med nasjonale prøver ikke er så farlig.
Det har vært lite ytringer av typen ”Å nei, hvordan skal vi ordne dette?” I stedet har vi hatt fokus på hva det er konkret innenfor det enkelte faget vi nå skal sette fokus på og hva vi og elevene kan utvikle oss og bli bedre på.

Når det gjelder de nye elektroniske leseprøvene, har ikke dette vært noe problem så vidt jeg har hørt. Men alle elevene har fått øvd seg på demo-testene, slik at de skjønte hva det gikk ut på. Elevene var her positive og synes ikke dette var noe problem.

Vi har ikke fått noen veiledning på hvordan resultatene skal brukes, men vi har tilbakeført informasjon om resultatene til elev og foreldre i foreldresamtalen som vi har på våren. Dette kan vi sikkert bli bedre på, men vi har hatt noe rutine her i kommunen, vi har tross alt hatt kartleggingsprøver tidligere som vi også har snakket med elev og

foreldre om. Nå er det kommunal enighet om at profilene fra 7. trinn skal overføres til ungdomsskolene, slik at lærerne kan bruke dem videre.

Uttalelse fra en rektor på en ungdomsskole

Uansett hva som skjer, er det viktigste fremover å få orden på informasjonsstrømmen til skolene og i første rekke strømmen av brev og mailer til rektorene. Det er en formidabel flyt av informasjon gjennom brev og mail fra henholdsvis fagmiljøene og Utdanningsdirektoratet. Slik kan det ikke fortsette – for meg er det utmattende som skoleleder.

På tross av informasjonsstrømmen, er det så vidt jeg har kunnet sett, ingen informasjon om hvordan vi skal bruke resultatene. All informasjon dreier seg om hvordan prøvene har vært utviklet, forarbeidet, registrering av data, sikring av resultatene, men ingenting om hvordan vi skal ta disse resultatene inn i skolen for videre organisasjonslæring. For oss kom dessuten resultatene sent, slik at ikke alle 10. trinns elever fikk snakket med lærerne sine om profiler, før de sluttet skolen. Noen lærere har klart å snakke med foreldre og elever, men ikke alle. Vi har heller ikke snakket med videregående skole om overføring av resultater for den enkelte elev.
Derimot har rektorgruppa i vår kommune blitt enige om at resultatene fra 7. trinns - elevene skal følge dem til ungdomsskolene, slik at vi kan samarbeide om elevene i grunnskolen.

Jeg har som rektor adgang til våre elevers resultater, og kan printe ut disse å gi dem til lærerne. Lærerne har derimot hatt ulike følelser knyttet til nasjonale prøver, og særlig norsk lærerne har hatt mye frustrasjon knyttet til prøvene. Etter å ha brukt mye tid i 10. klasse på å lese informasjon om skrive prøvene i norsk, gjennomføre dem og rette dem, var lærerne svært skuffa da de fikk resultatene i form av ”som forventet”. To språklærere med 25 års erfaring følte det nærmest provoserende. Etter å ha lagt ned rimelig mye tid, opplevde de det som nedbrytende å få så lite informasjon tilbake. De uttalte: Hva tror de (fagmiljøet) om oss norsklærere og vår kompetanse? De mener selv at de kjenner sine egne elever svært godt etter tre års undervisning, og derfor ikke fikk noe igjen for denne skriveprøven. Skriveprøven i norsk har ikke fått fram det beste i mine lærere.

Når det gjaldt matematikk var lærerne mer fornøyd, da de så på dette som en diagnostisk prøve, som kunne gi dem nyttig informasjon om den enkelte. Her hadde også flere lærere kommentert at de oppdaget hva de kunne arbeide videre med for den enkelte elev. Lærerne var heller ikke så negative til engelskprøven, og elevene likte godt leseprøven. Dette ble som spill, oppfattet som moro og mange fikk gode resultater.

Elevene på 10. trinn var først innstilt på å boikotte, men rektor oppfordret dem til å ta prøven for å se om de kunne lære noe av den. Elevene gjennomførte prøven og de flinkeste uttalte at de var mest fornøyd. Rektor hadde ellers inntrykk av at elevene likte den engelske lesetesten best, fordi den var IKT- basert.

Elevene har ellers vært irritert på de nasjonale prøvene fordi de ikke skulle telle på karakteren. De mener selv at dersom de skal ta disse prøvene, må de telle med i en samlet vurdering.

I rektorgruppa er barneskolerektorene mest fornøyd, fordi de sier at lærerne deres har fått et vurderingsverktøy som de ikke hadde før. Ungdomsskolerektorene jeg kjenner, ser ikke hva de nasjonale prøvene har å tilby i vurderingsarbeidet slik de nå er.

Det har ikke vært noen kommunal strategi så langt for bruk av resultatene, men det vil være naturlig å trekke inn resultatene i skolenes årlige vurderingsmøte med skoleeier og i skolens årlige egenvurdering.

Acta Didactica **1/2005**

Institutt for lærerutdanning og skoleutvikling
Det utdanningsvitenskapelige fakultet
Universitetet i Oslo
Postboks 1099 Blindern
0317 OSLO

Dept. of Teacher Education and School Development
Faculty of Education
University of Oslo
P.O.Box 1099 Blindern
0317 Oslo
Norway

www.ils.uio.no

ISSN 1502-2013
ISBN 82-90904-81-9