# Nasjonale prøver på prøve

## *Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2004*

Svein Lie og Marion Caspersen, ILS, Universitetet i Oslo
Julius K. Björnsson, Námsmatsstofnun, Educational Testing Institute, Reykjavik

# Innhold

# 1 Innledning

I forbindelse med at de nasjonale prøvene for første gang gjennomføres denne våren er det framkommet et behov for å gjennomføre en utvalgsundersøkelse for å studere prøvenes kvalitet ut fra testteoretiske og pedagogiske kriterier. De konkrete kravspesifikasjoner knyttet til en slik utvalgsundersøkelse er angitt i neste kapittel, og det har styrt vårt arbeid. Vi har også tatt utgangspunkt i formålet med de nasjonale prøvene, slik de er beskrevet i prosjektbeskrivelsene 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 også slik at **prøvenes primære mål 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 primære målet. Vår forståelse av dette bygger her i stor grad på kravspesifikasjonen for denne undersøkelsen (gjengitt i kapittel 2).

Den nylig framlagte rapporten ”Rektorers og læreres erfaringer med de nasjonale prøvene 2004” fra TNS Gallup (heretter referert til som ”Gallup-rapporten”) diskuterer funn fra en spørreundersøkelse ute i skolene. Flere steder bruker vi funn i denne rapporten i diskusjoner av egne resultater.

Innholdet i denne rapporten er skrevet og senere revidert etter diskusjoner med representanter fra hver av faggruppene samt Anette Qvam fra Utdanningsdirektoratet. Konklusjoner og anbefalinger står likevel forfatterne helt ansvarlig for.

# 2 Kravspesifikasjon

I det følgende gjengir vi kravspesifikasjonen for undersøkelsen, slik den ble formulert fra Læringssenteret/ Utdanningsdirektoratet.

> ***Representativ utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver***
>
> *Formål med undersøkelsen*
>
> *I forbindelse med at de nasjonale prøvene for første gang gjennomføres denne våren skal det gjennomføres en utvalgsundersøkelse for å studere hvordan prøvene har*

*fungert dette året. Denne utvalgsundersøkelsen skal suppleres med en spørreundersøkelse til lærere og skoleledere om selve gjennomføringen av prøvene og hvordan den pedagogiske bruken av prøvene har fungert på skolene. Det utvikles en egen kravspesifikasjon for denne spørreundersøkelsen.*

*Formålet med utvalgsundersøkelsen er todelt. Undersøkelsen skal gi et grunnlag for å kunne vurdere hvilke rapporteringskategorier som man kan forsvare å publisere i Skoleporten. Det bør dokumenteres at alle publiserte data er av en tilstrekkelig høy kvalitet og at det er rapporteringskategorier som kommuniserer godt med brukergruppene. Den andre målsetting er å få kunnskap som gjør at man kan forbedre oppgaveutviklingen, prosedyrer for gjennomføring og vurdering for neste års prøver.*

*Utvalgsundersøkelsen skal gjennomføres på et representativt utvalg.*

*Dette skal utvalgsundersøkelsen belyse:*

*1. Hvilke metoder ble brukt i utviklingen av prøvene? (pilotering/item analyser/skalering)*

*2. Grunnleggende item-analyser*

*3. Reliabilitet:*
- *Prøvens reliabilitet (indre konsistens i prøven som helhet)*
- *Vurderings/skårings reliabilitet (på alle oppgavene som lærerne vurderer og gir en kode eller skåre på)*

*4. Validitet*
- *Om prøven prøver det den skal*
- *Lærerne blir (i spørreskjema) bedt om å vurdere om oppgavene i prøvene og de sammenfatningskategorier som brukes er relevante for deres elever og undervisning.*
- *Fungerer inndelingen i subskalaer på en god måte. Er de foreslåtte rapporteringskategoriene tilstrekkelig forskjellig fra hverandre til at de virkelig kan sies å måle forskjellig kompetanse?*
- *Har prøven den faktorstruktur som samsvarer med de delprøver/skalaer som ble brukt?*

*5.Sammenlikninger*
- *Kan de foreslåtte kompetansenivåer fungere som absolutte skalaer for kompetanse?*
- *Kan kompetansenivåene fungere som grunnlag for sammenligninger med resultater et annet år eller må sammenlikningene baseres på andre indikatorer fra prøvene?*

# 3 Utvalg og gjennomføring

### 3.1.1 Utvalg

Det var allerede lagt opp en strategi for å ”sjekke” hvor pålitelig rettingen av åpne oppgaver foregikk. Alle, eller de fleste, skoler sendte originaler eller kopier av noen elevhefter til ”ekspertvurderere” (ofte bare kalt ”eksperter”). Disse sendte så sine data tilbake til skolen for at skolen skulle få en slags vurdering av kvaliteten på egen retting. Det var imidlertid ikke lagt opp til at denne informasjonen skulle nå Utdanningsdirektoratet eller faggruppene.

Utvalgene av elevbesvarelser besto av alle besvarelser vurdert av et tilfeldig uttrukket utvalg av eksperter. Utdanningsdirektoratet har hatt ansvar for uttrekking av disse

ekspertene. Utvalget av eksperter er trukket 2 fra hver region (2x4=8). For matematikk i 10. klasse ble det valgt 9 for å nå ca. 1000 elevbesvarelser. Utvalgenes størrelse framgår av tabell 3.1.

Det var enighet om at for lesing på 4.trinn var 500 besvarelser et stort nok utvalg, dvs at fire eksperter skulle sende inn sine vurderinger. Dette var på grunn av at det bare var flervalgsoppgaver i prøven, så sammenlikning mellom sensorene var uaktuell. Ved en misforståelse trodde vi at åtte ekspertvurderere skulle delta, derfor ble de siste fire (i det store utvalget) purret opp uten at de hadde blitt varslet. Vi fikk kontakt med kun to av disse, som også sendte inn sine vurderinger. Det aktuelle utvalget ble derfor litt større enn forutsatt, men vi valgte å beholde alle. De aktuelle tallene står i parentes i tabell 3.1.

*Tabell 3.1: Utvalgsstørrelse for hver prøve*

| | Fag | Antall ekspertvurderere | Antall skoler | Antall elever |
|---|---|---|---|---|
| 10. trinn | Matematikk | 9 | 59 | 953 |
| | Lesing | 8 | 55 | 927 |
| | Engelsk | 8 | 60 | 929 |
| 4. trinn | Matematikk | 8 | 105 | 880 |
| | Lesing | 4 (6) | 62 (93) | 498 (750) |

Strategien gikk for 10. klasse ut på å hente inn data fra de samme elevene både fra ekspertene og fra skolene, slik at vi kunne sammenlikne vurderingene gjort av de to sensorene. Imidlertid ble det bestemt at heller ikke i matematikk ønsket man å be om data for 4. klasse fra skolene, for ikke å skape for mye turbulens ute i skolene. Man antok at informasjon om sensorreliabilitet for 10. klasse ville være nokså representativ også for 4. klasse, siden de to prøvene og vurderingene av dem er så like.

### 3.1.2 Deltakelse og gjennomføring

All informasjon om utvalgsundersøkelsen til de uttrukne eksperter og tilhørende skoler ble sendt fra Utdanningsdirektoratet . Skoler og ekspertvurderere ble bedt om å sende sine vurderingsark (evt. filer) til ILS v/Marion Caspersen. Her ble dataene tastet inn av studenter etter hvert som de er mottatt. Vi har også mottatt en del excel-filer som har blitt lagt rett inn i våre filer.

*Tabell 3.2: Antall skoler og elever i våre datafiler*

| | Fag | Ekspertfil | | Skolefil | | Sensorfil: Uavhengige data fra de to kildene | |
|---|---|---|---|---|---|---|---|
| | | Elev-besvarelser | Skoler | Elev-besvarelser | Skoler | Elev-besvarelser | Skoler |
| 10. trinn | Matematikk | 858 | 58 | 556 | 35 | 459 | 33 |
| | Lesing | 789 | 54 | 486 | 38 | 392 | 30 |
| | Engelsk | 833 | 55 | 574 | 35 | 372 | 24 |
| 4. trinn | Matematikk | 865 | 105 | - | | - | |
| | Lesing | 664 | 85 | - | | - | |

Det aktuelle innholdet av våre datafiler framgår av tabell 3.2. Stort sett gikk det greit å få inn data fra alle ekspertene. Det viste seg imidlertid at noen skoler ennå ikke hadde

sendt besvarelsene til ekspertene, slik at det manglet data fra noen få skoler i "ekspertfilene". Videre var det et betydelig, men ikke foruroligende, frafall av enkeltelever i forhold til det antallet som var trukket ut på skolene (maksimalt 20). For 10. klasse hadde vi bedt skolene fylle ut et skjema for å kunne registrere eventuelt frafall og grunnene til dette (nekt/boikott, fritak, annet fravær). Denne informasjonen er summert opp i tabell 3.3. For 4. klasse har vi ikke tilsvarende informasjon, men ut fra oversikten over antall aktuelle og deltakende elever tyder alt på at deltakerprosenten var betydelig høyere i 4. klasse enn i 10. klasse. 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. Men det ser ikke ut til at et stort fravær har vært et gjennomgående problem. Imidlertid vil vi peke på at i matematikk har det vært en påfallende høy andel boikott/nekt. Slike problemer bør drøftes spesielt for å finne tiltak for å unngå altfor ulik grad av deltakelse fra skole til skole.

*Tabell 3.3: Oversikt over frafall av enkeltelever i 10. klasse*

| Fag | % som deltok på prøven | % boikott/nektet av foreldre | % som har fått innvilget fritak / syk |
|---|---|---|---|
| Matematikk | 88,2 | 5,7 | 6,1 |
| Lesing | 94,0 | 0 | 6,0 |
| Engelsk | 91,5 | 0,9 | 7,6 |

Et stort problem framkommer tydelig ved å sammenlikne antall skoler i "ekspertfiler" og i "skolefiler" i tabell 3.2. Så mye som omtrent 35 % av skolene klarte ikke å sende inn sine data innen skoleårets slutt på tross av purring. Det ser altså ut til at en betydelig del av skolene på dette tidspunkt ennå ikke hadde gjennomført vurdering av besvarelsene, noe vi finner svært betenkelig. Man kan forvente at tilbakemeldingen til elever om resultatene har liten funksjon hvis den ikke gjøres før skoleårets slutt. Vi vil i det hele peke på at den pedagogiske bruk av prøvene på skolene, også ifølge skolenes svar på Gallup-undersøkelsen om dette, synes å være uklar. Vi vil peke på at det er viktig med god informasjon til skolene om hvilke prosedyrer skolene skal følge etter at prøvene er gjennomført.

Som det også framgår av tabell 3.2, er antall skoler og elever i sensorfilene lavere enn det er i skolefilene. Dette kommer for det første av at på noen skoler har lærerne endret sine vurderinger etter å ha sammenliknet med ekspertenes, på tross av at vi uttrykkelig advarte mot dette (se kap. 4.4). For det andre er det for noen skoler umulig å matche elevdataene fordi identitetsnummereringen har vært forskjellig for de to datasettene. Vi vil peke på at det er viktig kommende år å bruke et entydig sett av identitetsvariable slik at det er enkelt å matche elev- og skoledata.

# 4 Strategi og dataanalyse for undersøkelsen

I dette kapitlet vil vi gi en oversikt og beskrivelse av de analysene som er gjennomført, og bakgrunnen for disse.

## *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 (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).*

## 4.3 Prøvens reliabilitet (indre konsistens)

Vi har beregnet prøvens indre konsistens reliabilitet (Chronbachs 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 stor 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 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 grupper av personer med høy presisjon, 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 mål for alfa. Hvilket krav vi skal sette til alfa i vår undersøkelse, kan ikke settes på forhånd, siden det vil avhenge av hvordan skåreverdiene skal lagres og brukes som mål på elevenes kompetanse. For hver prøve har vi derfor tatt dette spørsmålet opp til en nærmere diskusjon.*

***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 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 stor 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 slik: Prøven i engelsk skriftlig har en reliabilitet på 0,66 og et*

*standardavvik på 1,3 målt i nivåer (se kap. 5.6.3). Hver elevs måleresultat har derfor en feilmargin på ca 120 % av 1,3 og det utgjør litt over 1,5 målt i nivåer. Hvis Per har fått resultatet 5 (nivå B1), så burde vi egentlig si at Pers kompetanse er målt til en verdi som med 95 % sannsynlighet ligger mellom mellom 3,5 og 6,5, eller med andre ord ligger på nivå enten 4, 5 eller 6. Og det er jo en svært upresis og nokså lite verdifull informasjon.*

*Tabell 4.1: Hvordan feilmarginer avhenger av reliabiliteten*

| ***Reliabilitets-koeffisient*** | ***Feilmargin s uttrykt i prosent av ett standardavvik*** |
|---|---|
| *(Ingen prove)* | *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 (der lærerne vurderer og gir en kode)*

Vi har foretatt en analyse av om rettingen er foregått på tilstrekkelig lik måte fra retter til retter slik at vi kan stole på de angitte kodene. For hver enkelt åpen oppgave har vi beregnet overensstemmelsen mellom lærerens vurdering og "eksperten". Vi har ikke kunnet gå inn på i hvor stor grad hver lærer eller ekspert har rettet konsistent, men har konsentrert oss om å måle inter-sensor reliabilitet. Vi har i denne analysen ikke betraktet ekspertenes vurdering som mer "riktig" enn lærernes. 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 tendenser til konsistent for streng/for snill retting 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 angitt resultatene ut fra kriteriene < 85 %, og også < 75 %. Imidlertid er den såkalte Cohen's Kappa 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 (chance corrected) og dette er meget viktig når man vurderer overensstemmelse på oppgaver som har få svar For eksempel kan man se at hvis en oppgave har enten riktig eller galt som svar ville fullstendig tilfeldig skåring 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 vi har valgt å oppgi sensorreliabiliteten på to måter.*

Siden ekspertene i flere tilfeller har sendt sine vurderinger tilbake til skolene, er det et problem at lærerne kan ha endret sin vurdering. Dette kan være gjort fordi de har gjort en revidert vurdering og blitt "overbevist" av eksperten. Eller de har gjort en endring bare for å oppnå bedre overensstemmelse i vår undersøkelse, på tross av at vi uttrykkelig ba om at dette ikke ble gjort. I noen tilfeller kan vi lett se at vurderingene er endret i ettertid, og slike skoler har vi konsekvent fjernet fra datagrunnlaget før sensorreliabiliteten er beregnet. Vi vil understreke 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.

## 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.

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. Når det gjelder matematikk og lesing, og delvis også i engelsk, har forfatterne av denne rapporten gjennom arbeid med Europarådets engelskundersøkelse, PISA- og TIMSS-prosjektene god kjennskap til fagdidaktiske diskusjoner om mål og mening i fagene. Likevel har vi bare i begrenset grad gått inn på en analyse av i overensstemmelse mellom prøvene og innholdet i læreplanene. Vi har imidlertid referert lærernes syn på dette, slik det er kommet til uttrykk i Gallup-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 over å studere korrelasjonen mellom de to variablene i lys av hvor "nøyaktig" hver av variablene representerer en "sann kompetanse", 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.*

Vi har 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. Videre har vi analysert med en såkalt konfirmerende faktoranalyse i hvor stor grad den foreslåtte inndelingen etter delkompetanser støttes av dataene.

> ***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 didaktiske 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.

Imidlertid hjelper ikke de beste intensjoner 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 pine". 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 kompetansenivåer fungere som absolutte (kriteriebaserte) skalaer for kompetanse?**

**Kan kompetansenivåene fungere som grunnlag for sammenligninger med resultater et annet år eller må sammenlikningene baseres på andre indikatorer fra prøvene?**

Disse to spørsmålene henger delvis sammen og kommenteres her samlet. Dersom kompetanser er målt langs en skala med klart beskrevne nivåer, sier vi at vi har en **kriteriebasert 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 og matematikk er eksempler på at man har prøvd å innføre en slik kriteriebasert 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 **normbasert vurdering**, der vurderingen foregår ved sammenlikning mellom elevene. Det sier seg selv at det ikke er mulig å sammenlikne resultater år for år ved en relativ vurdering, for gjennomsnittet vil per definisjon alltid måtte settes det samme. Det vil heller ikke gå an å bruke endring i prosent riktige svar 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 kriteriebasert 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 er mulig.

## *4.7 Råd for publisering 2004 og for videreutvikling*

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 får eller bør få for eventuell publisering på Skoleporten og for utvikling av neste års prøver.

# 5 RESULTATER

## *5.1 Data fra elevbesvarelsene*

I denne delen vil vi beskrive resultatene av undersøkelsen i detalj. For hver prøve vil vi først gi resultatene oppgave for oppgave i tabellform. Disse dataene er beregnet og gjengitt i hver tabell:

- Svarfordeling i prosent. For flervalgsoppgaver oppgis prosent for hvert svaralternativ, samt for blanke svar. For åpne oppgaver oppgis svarfordeling på de ulike skåreverdier (lesing 10. klasse) eller nivåer (matematikk).

- Neste sett av data gjelder hvor gode elevene er som har gitt disse svarene. Elevenes dyktighet er her gitt som deres skåre på testen som helhet. Disse skåreverdiene er normert på ulike måter for de forskjellige prøvene.
- Neste informasjon gjelder oppgavens diskriminering, og det er korrelasjonen mellom skåre/nivå/poeng på den aktuelle oppgaven og for testen som helhet.
- For åpne oppgaver kommer det deretter en kolonne med data om hvor likt oppgaven er rettet av de to uavhengige retterne. Dette er gjort i form av prosent overensstemmelse og verdien på Kappa (se 4.4).
- I den siste kolonnen er eventuelle problemer "flagget" i form av en fotnote og/eller kommentar.
- Når det gjelder engelsk, der det ikke finnes data oppgave for oppgave, vil data i stedet oppgis for hver vurderingskategori for seg.

Videre oppgir vi for hver prøve disse opplysningene:

- Prøvens reliabilitet (Chronbach's alfa, se 4.3)
- Tilsvarende for hver av de foreslåtte rapporteringskategoriene
- Korrelasjon mellom hver av de foreslåtte rapporteringskategoriene
- Informasjon om struktur i dataene ut fra eventuell faktoranalyse

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.

## *5.2 Matematikk 10. klasse*

### 5.2.1 Prøvens struktur, validitet og kriterier ved retting

Det er ingen flervalgsoppgaver i dette settet. Alle oppgavene er altså åpne, og de er rettet etter et spesielt designet system for å bestemme elevenes nivå langs en skala fra 0 til 5. Oppgavene er opprinnelig klassifisert i henhold til disse kategoriene:

- K: Kommunikasjon
- R: Representasjon, symbolbruk og formalisme
- T: Matematisk resonnement og tankegang
- M: Matematisk modellering og anvendelse
- P: Problembehandling
- H: Bruk av hjelpemidler (Denne kategorien måtte sløyfes på grunn av manglende og forvirrende data, se nedenfor)

Noen av oppgavene er klassifisert etter to (eller til og med tre) kategorier samtidig. Disse oppgavene er altså ment å skulle bidra til to forskjellige skalaer, og ifølge rettemanualen er kriteriene noen ganger forskjellig. Slike "dobbeltkategoriserte" oppgaver vil derfor måtte analyseres to eller tre ganger, en for hver kategori. Som det framgår av resultatene, var kriteriene så like at det i praksis har dreid seg om å gi samme vurdering to eller tre ganger. I alt har vi analysert data fra 63 oppgaver, eller så mange som 82 hvis vi teller med de dobbeltkategoriserte "oppgavene".

Det viste seg at den siste kategorien (H) var lite egnet. De forskjellige ekspertene hadde brukt helt forskjellige koder for disse oppgavene. Noen hadde skrevet nesten bare 0, andre hadde stort sett 5-ere, mens andre igjen ikke hadde gitt noen verdier i det hele tatt. Vi kan bare slå fast at det har hersket full forvirring rundt denne kategorien,

og vi har derfor fjernet den helt. Det har medført at tre oppgaver (oppgave 20a, b, og c) overhode ikke er analysert.

De foreslåtte kategoriene tar utgangspunkt i et internasjonalt velkjent teoretisk arbeid av den danske fagdidaktiker Mogens Niss. Dette systemet danner også et teoretisk utgangspunkt for rammeverket til matematikk i PISA-undersøkelsen. Men i PISA er det uttrykkelig unngått å klassifisere oppgavene etter dette systemet direkte, nettopp fordi det er betraktet som uegnet så lenge hver oppgave må klassifiseres i mer enn én kategori.

De som rettet, skulle først bedømme elevbesvarelsen og gi en kode (se prøvens KODEBOK). Deretter skulle disse kodene overføres til "nivåer" fra 0 (ubesvart) til 5 (høyeste nivå) i henhold til dokumentet "Innplassering av koder i kompetanseprofilen 10. klasse". Spesielt gjelder at svar som "vet ikke" eller liknende ikke regnes som noe svar, men fører automatisk til nivå 0.

Vår analyse tok utgangspunkt i nivåene fra 0 til 5. I henhold til prosedyrene for beregning av kompetanse skal da en elevs kompetansenivå innen en kategori kunne beregnes som gjennomsnittet av nivåene for de oppgavene som hører inn under kategorien. Vi har i vår analyse tatt utgangspunkt i denne framgangsmåten.

Et vesentlig poeng ved de foreslåtte kompetansenivåene er at de er ment å være kriterierelaterte (se 4.6), altså at de referer seg til nivåer som er beskrevet i ord ut fra hva elever på dette nivået faktisk kan (og ikke kan). Ideelt sett skulle bruk av et slikt system kunne gi den store fordelen at vi ut fra resultatene kan beskrive i detalj hvor mange elever som kan hva. Et stort problem er imidlertid at disse fem nivåene for hver kompetansekategori er hentet fra et rent teoretisk perspektiv. Det har ikke vært noen utprøving av hvorvidt elever som svarer riktig på en bestemt oppgave, *faktisk* befinner seg på nivå 5. Slik utformingen av nivåer og kriterier er for denne prøven, blir det for oss et viktig empirisk spørsmål å etterprøve dette: I hvor stor grad er det dekning i resultatene for de angitte nivåene? Vår oppgave blir her å sammenholde teori og empiri. I utgangspunktet fortoner det seg imidlertid for oss uforståelig at et riktig svar kvalifiserer til nivå 5 uavhengig av hvor vanskelig eller lett oppgaven er. Vi kan ikke gjøre annet i vår analyse enn å betrakte de gitte nivåene som poeng som angir hvor godt hver oppgave er besvart. Tilsvarende betrakter vi gjennomsnittet av disse poengene for en elev som et uttrykk for elevens dokumenterte dyktighet innenfor området.

Isolert sett vurderer vi prøvens innholdsvaliditet som høy. Lærerne rapporterer i Gallup-undersøkelsen at oppgavenes form "i stor grad" er velkjent for elevene, og oppgavene er etter vår vurdering meget gode når det gjelder å måle god forståelse i faget. Vi mener oppgavesettet reflekterer læreplanen på en meget god måte, noe lærerne er enig med oss i: Bare rundt 15 % av dem vurderte at oppgavene reflekterer sentrale mål i Læreplanen i "liten" eller "svært liten" grad. På spørsmålet "I hvilken grad mener du elevene fikk vist sine ferdigheter gjennom prøven?" svarte bare rundt 10 % av lærerne "i liten grad" eller "i svært liten grad". Vi mener at på bakgrunn av god validitet i forhold til læreplanens innhold og intensjoner, vil prøven på mange måter kunne gi et godt signal tilbake til skolene om hva som er viktig å vektlegge i undervisningen. Vi mener også at prøven inneholder en rekke gode oppgaver for å avdekke forståelse av grunnleggende begrepsforståelse i matematikk. Likevel vil vi

peke på at det er et åpent spørsmål, og derfor bør diskuteres, i hvilken grad utvalget av oppgaver vektlegger den delen av matematikkompetansen som Kvalitetsutvalget kalte ”basiskompetanse” og som St. meld. Nr. 30 kalte ”grunnleggende ferdigheter”.

### 5.2.2 Item-analyse

Resultatene fra item-analysen er gitt i tabell 5.1. Hver oppgave er angitt med oppgavens nummer samt en bokstav som angir hvilken kategori den tilhører. På grunn av dobbeltkategorisering vil noen oppgaver forekomme på to linjer. For mange oppgaver er svarene fordelt bare på noen få nivåer, eller noen nivåer er nesten ikke brukt.

Som mål for elevenes dyktighet har vi brukt gjennomsnittlig nivå for oppgavene som hører inn under den aktuelle kategorien. Dette innebærer altså at vi foreløpig bare bruker nivåene som en poengskala.

*Tabell 5.1: Item-analyse for matematikk i 10. klasse. Flere av oppgavene forekommer to eller tre ganger, en gang for hver kompetanse som er vurdert. Prosentfordelingen er avrundet til hele tall, og dyktigheten (gjennomsnittlig ”nivå” for de som har svart slik) til én desimal. D står for oppgavens diskriminering. Som et uttrykk for sensorreliabilitet står R for prosentandelen der de to sensorene har vurdert likt (første tall) eller der avviket er lavere enn 2 poeng (det andre tallet). K står for Kappa (se 4.4). I kolonnen for kommentarer (Kom) er det henvist til ulike fotnoter under tabellen. Et utropstegn i denne kolonnen betyr at vi har å gjøre med et betydelig problem.*

| Oppg nr | Svarfordeling i % | | | | | | Dyktighet | | | | | | D | R | K | Kom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | | | | |
| 1a R | 3 | 4 | | | | 94 | 1,1 | 1,7 | | | | 2,6 | ,28 | 99-99 | ,91 | b |
| 1b R | 1 | 49 | | | | 49 | 0,9 | 2,3 | | | | 2,9 | ,29 | 95-98 | - | b |
| 2a R | 4 | 13 | 0 | 0 | 0 | 83 | 0,9 | 1,8 | - | - | - | 2,8 | ,38 | 93-95 | ,74 | |
| 2b R | 7 | 34 | 2 | 1 | 0 | 57 | 1,2 | 2,1 | 2,5 | 2,0 | - | 3,0 | ,48 | 80-90 | ,67 | a, c |
| 2c R | 7 | 27 | 1 | 0 | | 65 | 1,1 | 2,1 | 2,3 | - | | 2,9 | ,44 | 93-97 | - | |
| 2d R | 17 | 42 | | | | 41 | 1,6 | 2,4 | | | | 3,2 | ,49 | 92-95 | - | |
| 2e R | 27 | 33 | | 0 | | 39 | 1,7 | 2,5 | | - | | 3,3 | ,55 | 92-97 | - | |
| 2f R | 20 | 39 | 2 | 2 | 1 | 37 | 1,6 | 2,4 | 2,5 | 3,1 | 2,8 | 3,2 | ,48 | 84-94 | ,76 | a, c |
| 2g R | 44 | 23 | 0 | 0 | 0 | 33 | 1,9 | 2,6 | - | - | - | 3,3 | ,51 | 93-97 | - | |
| 3a R | 3 | 10 | 0 | | 0 | 87 | 0,9 | 1,9 | - | | - | 2,7 | ,33 | 97-98 | - | |
| 3b R | 11 | 32 | 1 | 0 | 0 | 55 | 1,4 | 2,0 | 2,9 | - | - | 3,1 | ,59 | 91-96 | ,84 | |
| 3c R | 21 | 31 | 1 | 1 | 1 | 46 | 1,7 | 2,4 | 2,7 | 3,2 | 3,2 | 3,0 | ,36 | 89-96 | ,83 | a |
| 3d R | 45 | 32 | 1 | 0 | 1 | 22 | 2,0 | 2,6 | 3,2 | - | - | 3,7 | ,58 | 88-96 | ,81 | |
| 4a R | 36 | 41 | 0 | 0 | 0 | 22 | 2,0 | 2,5 | - | - | - | 3,7 | ,64 | 89-95 | - | |
| 4b R | 41 | 40 | 0 | 0 | 1 | 18 | 2,0 | 2,6 | - | - | 3,5 | 3,7 | ,54 | 92-97 | ,87 | |
| 4c R | 57 | 27 | 1 | 1 | 0 | 15 | 2,1 | 2,8 | 3,0 | - | - | 3,9 | ,56 | 91-98 | - | |
| 5a M | 4 | 18 | 4 | 0 | 1 | 73 | 0,9 | 1,6 | 2,0 | - | 2,0 | 3,1 | ,49 | 91-95 | - | |
| 5b M | 17 | 31 | 0 | 1 | 0 | 51 | 1,4 | 2,4 | - | - | - | 3,2 | ,39 | 89-92 | - | |
| 5c M | 27 | 22 | 28 | 0 | 0 | 21 | 1,9 | 2,3 | 2,9 | - | - | 3,7 | ,43 | 88-96 | - | |
| 6a K | 28 | 51 | 1 | 1 | 1 | 19 | 1,5 | 2,3 | 2,5 | 3,3 | 3,4 | 3,4 | ,46 | 85-92 | ,77 | c |
| 6a T | 28 | 48 | 3 | 1 | 1 | 19 | 2,2 | 2,9 | 3,1 | - | 3,5 | 3,7 | ,47 | 84-91 | ,75 | c |
| 6b K | 34 | 28 | 8 | 2 | 1 | 28 | 1,5 | 1,9 | 2,6 | 3,0 | - | 3,4 | ,58 | 79-88 | ,71 | c |
| 6b T | 34 | 28 | 7 | 2 | 1 | 28 | 2,2 | 2,7 | 3,2 | 3,3 | 2,9 | 3,7 | ,58 | 80-89 | ,72 | a, c |
| 7a T | 3 | 2 | | | | 95 | 1,1 | 1,9 | | | | 2,9 | ,21 | 98-98 | - | b |
| 7b T | 5 | 17 | 4 | | | 75 | 1,4 | 2,3 | 2,5 | | | 3,1 | ,37 | 95-97 | - | |
| 7c T | 5 | 11 | 2 | | | 82 | 1,4 | 2,2 | 2,3 | | | 3,0 | ,32 | 96-97 | - | |
| 8a R | 10 | 4 | 0 | | 0 | 86 | 1,6 | 1,9 | - | | - | 2,7 | ,35 | 97-98 | - | |
| 8a T | 10 | 4 | 0 | | 0 | 86 | 1,7 | 2,1 | - | | - | 3,0 | ,34 | 96-98 | - | |

| | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 8b R | 21 | 42 | 13 | 0 | 1 | 23 | 1,9 | 2,4 | 2,7 | - | - | 3,4 | ,52 | 90-96 | ,86 | |
| 8b T | 21 | 42 | 13 | 0 | 1 | 23 | 2,1 | 2,8 | 3,0 | - | - | 3,7 | ,52 | 90-96 | ,86 | |
| 9a R | 29 | 36 | 6 | 0 | 0 | 29 | 2,1 | 2,3 | 2,9 | - | - | 3,4 | ,58 | 91-97 | ,88 | |
| 9a T | 29 | 35 | 1 | 0 | 6 | 29 | 2,3 | 2,6 | 3,1 | - | 3,2 | 3,7 | ,58 | 89-96 | ,84 | |
| 9b R | 51 | 28 | 7 | 1 | 0 | 14 | 2,2 | 2,6 | 3,1 | 3,4 | - | 3,8 | ,56 | 89-96 | - | |
| 9b T | 51 | 28 | 3 | 1 | 4 | 14 | 2,5 | 2,9 | 3,7 | 3,6 | 3,4 | 4,0 | ,55 | 87-94 | ,80 | a |
| 9c R | 34 | 42 | | 0 | | 24 | 2,0 | 2,5 | | - | | 3,6 | ,59 | 94-98 | - | |
| 9c T | 34 | 42 | | 0 | | 24 | 2,3 | 2,8 | | - | | 3,8 | ,59 | 95-98 | - | |
| 10a T | 2 | 9 | 7 | 0 | | 82 | 1,4 | 2,1 | 2,0 | - | | 3,1 | ,36 | 95-98 | ,81 | a! |
| 10b T | 2 | 39 | 13 | 15 | 24 | 7 | 1,2 | 2,5 | 2,6 | 2,9 | 3,4 | 4,0 | ,57 | 76-93 | ,69 | c |
| 10c K | 18 | 12 | 7 | 1 | 35 | 27 | 1,0 | 1,6 | 2,0 | 2,7 | 2,7 | 2,9 | ,55 | 66-92 | ,54 | c! |
| 10c T | 17 | 13 | 7 | 1 | 34 | 27 | 1,8 | 2,4 | 2,5 | 3,4 | 3,2 | 3,4 | ,56 | 68-93 | ,58 | a,c! |
| 10d M | 41 | 30 | 3 | 5 | 4 | 16 | 1,9 | 2,6 | 3,1 | 3,5 | 3,5 | 4,0 | ,62 | 76-91 | ,67 | c |
| 11 T | 27 | 19 | 0 | 0 | 1 | 53 | 2,1 | 2,7 | - | - | 2,9 | 3,3 | ,50 | 88-94 | ,80 | |
| 12a P | 24 | 18 | 9 | 13 | 1 | 35 | 1,0 | 1,7 | 1,8 | 2,5 | 3,0 | 3,1 | ,50 | 73-85 | ,65 | c! |
| 12b K | 34 | 20 | 4 | 22 | 2 | 18 | 1,4 | 2,2 | 2,5 | 2,7 | 3,2 | 3,4 | ,54 | 65-75 | ,55 | c! |
| 13 R | 3 | 23 | | | | 74 | 1,3 | 1,9 | | | | 2,8 | ,40 | 95-97 | - | |
| 14 R | 2 | 61 | | | | 36 | 1,3 | 2,3 | | | | 3,1 | ,43 | 96-97 | ,92 | |
| 15a R | 46 | 23 | 2 | 1 | 1 | 29 | 1,9 | 2,6 | 3,1 | 3,4 | 3,1 | 3,5 | ,66 | 87-96 | ,81 | a |
| 15a P | 46 | 23 | 2 | 1 | 1 | 29 | 1,2 | 2,2 | 3,1 | - | 2,5 | 3,5 | ,65 | 86-96 | ,79 | a |
| 15b R | 32 | 32 | 2 | 5 | 3 | 26 | 1,8 | 2,5 | 2,7 | 3,1 | 3,0 | 3,5 | ,64 | 80-90 | ,72 | a,c |
| 15b P | 31 | 32 | 2 | 6 | 3 | 26 | 1,0 | 1,9 | 2,6 | 3,0 | 3,2 | 3,5 | ,64 | 80-90 | ,73 | c |
| 16 K | 48 | 38 | 6 | 1 | 4 | 4 | 1,8 | 2,4 | 3,3 | 3,2 | 3,6 | 4,2 | ,54 | 74-93 | ,60 | a,c! |
| 16 P | 46 | 39 | 6 | 2 | 4 | 4 | 1,6 | 2,2 | 3,3 | 3,5 | 3,8 | 4,5 | ,53 | 73-94 | ,59 | c! |
| 17 K | 15 | 22 | 7 | 4 | 2 | 49 | 1,0 | 1,6 | 2,3 | 2,3 | 2,5 | 3,0 | ,60 | 85-94 | ,78 | c |
| 17 P | 14 | 15 | 15 | 5 | 2 | 50 | 0,6 | 1,2 | 1,8 | 2,3 | 2,4 | 3,0 | ,62 | 82-93 | ,74 | c |
| 18 R | 3 | 8 | 18 | 1 | 1 | 70 | 1,2 | 1,8 | 2,2 | - | - | 2,8 | ,41 | 90-94 | ,79 | |
| 19 K | 10 | 29 | 2 | 7 | 11 | 41 | 0,9 | 1,7 | 2,1 | 2,4 | 2,0 | 3,1 | ,56 | 73-84 | ,61 | a!,c! |
| 19 T | 10 | 7 | 7 | 3 | 22 | 51 | 1,8 | 2,4 | 2,9 | 2,7 | 2,5 | 3,3 | ,42 | 76-91 | ,64 | a!,c |
| 19 M | 9 | 13 | 9 | 15 | 12 | 43 | 1,3 | 2,3 | 2,0 | 2,2 | 2,4 | 3,4 | ,50 | 73-87 | ,62 | a!,c! |
| 21 P | 14 | 44 | 1 | 0 | 1 | 40 | 0,8 | 1,5 | 1,5 | - | 3,5 | 3,3 | ,66 | 92-98 | - | a |
| 22a R | 6 | 27 | 0 | | | 67 | 1,1 | 2,1 | - | | | 2,9 | ,39 | 96-97 | - | |
| 22b R | 14 | 46 | 0 | 0 | 0 | 40 | 1,6 | 2,3 | - | - | - | 3,2 | ,48 | 91-95 | - | |
| 22c R | 15 | 60 | 0 | | | 25 | 1,5 | 2,5 | - | | | 3,5 | ,52 | 95-97 | - | |
| 22d R | 27 | 34 | 7 | | 0 | 31 | 1,7 | 2,6 | 2,5 | | - | 3,3 | 53 | 86-98 | - | a! |
| 22e R | 10 | 59 | 0 | | | 31 | 1,4 | 2,3 | - | | | 3,4 | ,54 | 96-98 | - | |
| 22f R | 14 | 51 | 0 | 0 | | 35 | 1,4 | 2,4 | - | - | | 3,3 | ,56 | 94-98 | - | |
| 23a M | 30 | 22 | 2 | 5 | 29 | 12 | 1,4 | 2,3 | 2,9 | 3,3 | 3,4 | 4,1 | ,70 | 71-89 | ,63 | c! |
| 23a P | 30 | 21 | 3 | 33 | 1 | 12 | 0,8 | 1,8 | 2,3 | 2,9 | 3,4 | 4,0 | ,72 | 74-85 | ,66 | c! |
| 23b M | 36 | 14 | 1 | 2 | 1 | 46 | 1,6 | 2,2 | 2,8 | 2,8 | 3,7 | 3,6 | ,68 | 88-94 | ,81 | |
| 23b P | 36 | 14 | 1 | 3 | 1 | 46 | 0,9 | 1,6 | 2,1 | 2,4 | 2,8 | 3,2 | ,68 | 87-93 | ,79 | |
| 24a T | 14 | 4 | 0 | 14 | | 68 | 1,6 | 2,2 | - | 2,5 | | 3,2 | ,46 | 82-84 | - | c! |
| 24b T | 23 | 12 | 3 | 19 | | 43 | 2,0 | 2,4 | 2,5 | 2,8 | | 3,5 | ,52 | 73-80 | - | c! |
| 24c T | 19 | 3 | 1 | 16 | 0 | 61 | 1,8 | 2,0 | 2,7 | 2,7 | - | 3,3 | ,47 | 80-83 | - | c! |
| 24d T | 29 | 28 | 3 | 14 | | 26 | 2,2 | 2,7 | 2,9 | 3,2 | | 3,7 | ,50 | 72-83 | - | c! |
| 25a R | 7 | 62 | | | | 32 | 1,7 | 2,4 | | | | 3,1 | ,36 | 96-97 | - | |
| 25b T | 25 | 51 | 4 | 4 | 6 | 11 | 2,1 | 2,9 | 3,1 | 3,1 | 3,7 | 4,0 | ,60 | 72-91 | ,58 | c! |
| 26 T | 21 | 38 | 1 | 1 | 1 | 38 | 2,0 | 2,7 | 2,6 | 3,2 | 3,5 | 3,5 | ,51 | 83-92 | ,76 | a,c |
| 27 K | 29 | 22 | 1 | 16 | 1 | 31 | 1,3 | 1,9 | 2,2 | 2,7 | 2,8 | 3,1 | ,55 | 76-86 | ,68 | c |
| 27 M | 28 | 17 | 4 | 1 | 0 | 48 | 1,6 | 2,2 | 2,3 | 2,9 | 3,4 | 3,4 | ,54 | 84-93 | ,76 | c |
| 28a K | 18 | 27 | 3 | 5 | 7 | 41 | 1,1 | 1,7 | 2,1 | 2,7 | 2,3 | 3,1 | ,54 | 68-84 | ,55 | a!,c! |
| 28a M | 15 | 26 | 3 | 5 | 7 | 44 | 1,3 | 2,0 | 2,6 | 3,0 | 2,7 | 3,5 | ,52 | 68-86 | ,55 | a!,c! |
| 28b K | 21 | 36 | 4 | 3 | 2 | 34 | 1,2 | 1,9 | 2,3 | 2,5 | 2,6 | 3,3 | ,59 | 73-85 | ,61 | c! |
| 28b M | 18 | 34 | 4 | 2 | 2 | 40 | 1,4 | 2,2 | 2,5 | 2,9 | 3,2 | 3,6 | ,56 | 73-85 | ,60 | c! |

a) Dyktigheten er ikke "ordnet" etter kompetanse.
b) Svak diskriminering (< 0,30)
c) Dårlig overensstemmelse mellom rettere

Med utgangspunkt i resultatene i tabell 5.1 har vi disse kommentarene:

- Disse oppgavene har mange gode egenskaper, særlig er det påfallende at nesten alle oppgavene diskriminerer godt. For de få oppgavene der

diskrimineringen er lavere enn 0,30 (b i høyre kolonne), er det i de fleste tilfellene fordi det for svært lette oppgaver er vanskelig å oppnå høye verdier.

- I tråd med dette ser vi at for de fleste oppgavene er dyktigheten ordnet med økende gjennomsnitt for økende nivå 0-5. Siden det er en så detaljert (seksdelt) poengskala, er det rimelig at dette ikke alltid er oppfylt (a i høyre kolonne). **Det er altså mange indikasjoner i dataene på at en så detaljert poenggiving på enkeltoppgaver er uegnet.**
- Den gode diskrimineringen tyder på at nesten alle oppgavene er "gode", i den forstand at de krever, og derfor måler, god forståelse og gode ferdigheter i matematikk. Svake elever faller ubønnhørlig igjennom på de fleste av disse oppgavene.
- Det er så mye som 20 % blanke svar (kategori 0) i gjennomsnitt for alle oppgavene, og for mange oppgaver ligger andelen høyere enn 40 %..
- De to (eller tre) vurderingene av samme oppgave etter forskjellig kompetanse har gitt nesten identiske vurderinger, siden kriteriene har vært så godt som identiske.
- Gjennomsnittlig poeng oppnådd for alle de 63 oppgavene (hver oppgave bare en gang) er 2,67, noe som tilsvarer 53 % av "fullt hus". Dette er svært lavt, sett i lys av at de får ett poeng "gratis" for et verdiløst svar. Hvis vi i stedet hadde gitt 0 poeng for verdiløst svar og for øvrig senket alle poengene med 1, ville gjennomsnittlig oppnådde poeng bare blitt 47 % av "fullt hus". Vi konkluderer dermed at prøven har vært noe for vanskelig, dette ut fra en pedagogisk og ikke en testteoretisk vurdering. Vi konstaterer fra Gallup-undersøkelsen at 24 % av lærerne er enig med oss i dette, mens ca 6 % mener at den var noe for lett.
- Det er dårlig samsvar mellom påstått kompetansenivå for et bestemt svar og det aktuelle gjennomsnittlige nivået (dyktigheten). Dette innebærer at de påståtte nivåene (0-5) ikke kan oppfattes som annet enn poeng for hver oppgave, og man kan regne gjennomsnitt av disse poengene. **Men disse gjennomsnittene kan ikke med rimelighet sies å tilsvare de kompetansebeskrivelsene som er angitt i veiledningen.**
- Det er svært vanskelig å begrunne en seksdelt skala for disse oppgavene. For de aller fleste av dem ville det vært bedre – og MYE enklere – å gi ett eller to poeng (evt. unntaksvis tre). Det ville da dreid seg om riktig/galt (1 – 0 poeng) eller Riktig/ delvis riktig/ galt (2 – 1 – 0 poeng).
- Spesielt vil vi peke på at det riktignok kan begrunnes empirisk å gi ett poeng for et galt svar og 0 for ikke noe svar, siden de som gir et svar, skårer høyere totalt (se tabellen). Likevel er det lett å skjønne at en slik politikk vil være umulig å gjennomføre i praksis. Når dette blir kjent, vil alle elever bli oppfordret til å gi i hvert fall et tåpelig svar på alle oppgavene. Det vil være god teststrategi, men dårlig pedagogikk. **Å gi poeng for verdiløse svar er uheldig!**
- På grunn av den fingraderte vurderingen er det mange oppgaver der de to vurderingene spriker. De alvorligste uoverensstemmelsene gjelder 12 av de 63 oppgavene og er angitt med "c!" i høyre kolonne. Dette er oppgaver der samsvaret er 75 % eller lavere og/eller der samsvaret er 85 % eller lavere hvis vi godtar et avvik på ett poeng som uvesentlig. Totalt sett vurderer vi likevel ikke uoverensstemmelsene mellom sensorene som spesielt store, snarere at ambisjonene har vært for høye.

- **Det store problemet med vurderingen gjelder at den er så komplisert og dermed tidkrevende**. Det er etter vår mening uforståelig at det lages vurderingskriterier som det tar lærerne så mye som i gjennomsnitt 41 minutter per besvarelse å rette (kfr Gallup-undersøkelsen).

### 5.2.3 Analyse av de foreslåtte kategoriene

Informasjon om hvordan hver av de forslåtte kategoriene har fungert, er gitt i tabell 5.2. For hver kategori har vi gitt antall oppgaver, kategoriens reliabilitet (Cronbachs alfa) og gjennomsnittlig kompetansenivå for alle elevene. Som det framgår av tabellen, er det svært ujevn fordeling av oppgaver, noe som er medvirkende til at reliabiliteten varierer. Imidlertid er det slik at der det er færrest oppgaver, korrelerer de sterkest innbyrdes, noe som gjør at reliabiliteten likevel ikke varierer altfor mye, men holder seg mellom 0,82 og 0,92.

*Tabell 5.2: Data for hver av de foreslåtte kompetansene*

| Kategori | Antall oppgaver | Gjennomsn. Korrelasjon | Reliabilitet | Gjennomsn. kompetansenivå |
|---|---|---|---|---|
| Kommunikasjon (K) | 10 | 0,32 | 0,82 | 2,27 |
| Representasjon, symbolbruk og formalisme (R) | 33 | 0,24 | 0,92 | 2,57 |
| Matematisk resonnement og tankegang (T) | 21 | 0,24 | 0,87 | 2,87 |
| Matematisk modellering og anvendelse (M) | 10 | 0,31 | 0,82 | 2,65 |
| Problembehandling (P) | 8 | 0,41 | 0,84 | 2,15 |

Vi ser også at hver av skalaene har et gjennomsnittlig ”nivå” som ikke avviker mye fra midtpunktet mellom 0 og 5. Fordelingen av skåreverdier viser seg å være rimelig symmetrisk for alle kategoriene.

Som tidligere nevnt, teller mange oppgaver inn under to eller tre kategorier samtidig. Dette har som naturlig konsekvens at de forslåtte skalaene blir likere og derfor korrelerer høyere med hverandre enn de ellers ville gjort. Informasjonsverdien av hver kategori blir da selvsagt lavere enn om de var mer forskjellige.

Problemer med de foreslåtte kategoriene:

- Noen av kategoriene har litt lav reliabilitet, særlig fordi det er for få oppgaver. Det dreier seg særlig om kategoriene K, M og P. På den annen side har R så høy reliabilitet at det synes å være unødvendig mange oppgaver i denne kategorien.
- Det er i tillegg umulig å se at kategoriene står for forskjellige ting. Korrelasjonene mellom de tre skalaene er av størrelse (0,86 for K-M, 0,83 for K-P, og 0,84 for M-P) like stor som reliabiliteten til hver av dem (se tabell 5.2), altså er de ”latente” korrelasjonene tilnærmet lik 1,0. Når foreslåtte skalaer ikke kan påvises å være forskjellige, må vi konstatere at de **dermed har lav (såkalt diskriminerende) validitet**.
- Vi ser også at hver av skalaene har et gjennomsnittlig ”nivå” som ligger omkring midtpunktet mellom 0 og 5, men det er vanskelig å innse at elevene skal være så mye (omtrent 0,7 poeng) ”bedre” i kategorien T enn i P. Vi

frykter at med slik informasjon man lett komme i skade for å gi feilaktig inntrykk av elevers (relativt sett) sterke og svake sider.
- Både betegnelser og beskrivelser av de foreslåtte kategoriene kommuniserer dårlig til lærere og ledere hva kompetansene egentlig går ut på, og særlig hva som er forskjellen mellom dem.
- Vi vil derfor advare både mot å publisere data for disse skalaene og mot å betrakte slike data som mål for enkeltelevers kompetanse.

### 5.2.4 Konklusjon

I tillegg til de validitets- og reliabilitetsproblemene som er diskutert ovenfor, vil vi understreke kommentarer fra skolene som tydelig indikerer at det har tatt altfor lang tid å rette besvarelsene. Mange lærere har vært frustrerte, og flere skoler hadde ennå ikke gjennomført vurderingen ved utgangen av juni . Med en så detaljert vurdering og så ambisiøse kategorier er det naturlig at vurderingen må bli svært komplisert. Vår analyse viser at en slik strategi vanskelig kan forsvares.

Vi stiller oss videre uforstående til at det ikke er brukt en eneste flervalgsoppgave i denne prøven. Innslag av slike oppgaver ville automatisk ha gitt mindre vurderingsarbeid og trolig også høyere reliabilitet (siden man da kan inkludere flere oppgaver) og lavere andel blanke svar. Vi foreslår at det neste år blir et betydelig innslag av slike oppgaver.

Vi må konkludere at på tross av at prøven inneholder mange ”gode” oppgaver, fungerer den dårlig som en nasjonal prøve som skal måle elevers kompetanse. Det anvendte rammeverket for prøven har ikke fungert etter hensikten, og det er urimelig komplisert. Vi finner at de fem foreslåtte kategoriene har dårlig validitet, rett og slett fordi de er for like empirisk sett og også begrepsmessig er vanskelig å skille. **Vi vil derfor sterkt fraråde at resultater blir publisert som opprinnelig planlagt.**

Med så mange oppgaver som er brukt her, kunne prøven gitt grunnlag for to (eller kanskje til og med tre) skalaer. Men da måtte prøven vært bygget opp for å gjøre nettopp dette, med jevn fordeling av oppgaver etter gjensidig utelukkende kategorier. Det kan ikke gjøres i ettertid.

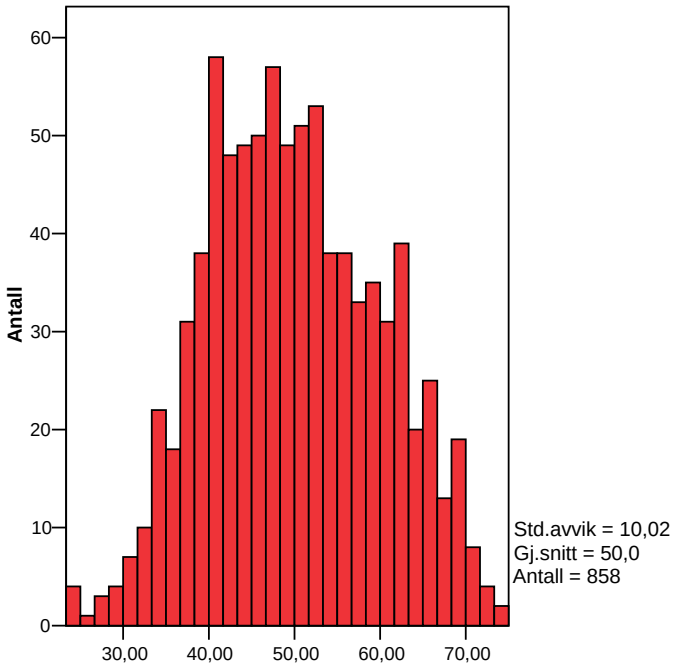
Derimot kan de foreliggende data slås sammen til én overordnet skala som måler matematikkompetanse generelt. En slik skala har høy både validitet og reliabilitet. For å få til dette i praksis er man nødt til å bygge på de data som skolene allerede har frambrakt, og som de skal rapportere til Utdanningsdirektoratet. Det dreier seg her om hver elevs ”nivå” (langs en skala fra 0 til 5) innen hver av de fem kategoriene K, R, T, M og P (se kap 5.2.1). Vi kan ikke ut fra disse tallene konstruere en samlet poengsum, der hver oppgave teller likt. Men vårt forslag til beste strategi er følgende:

- Man tar utgangspunkt i antall oppgaver innenfor hver kategori, på en slik måte at oppgaver som er dobbeltkategorisert, teller 0,5, mens de som er tredobbelt kategorisert teller 0,33. Antall oppgaver innenfor de fem kategoriene blir da henholdsvis 5, 33; 29,5; 16,33; 6,83 og 5,0. Ved å vekte de fem kategoriene etter disse ”antallene” kommer man fram til en gjennomsnittlig skåre på de 63 oppgavene langs en skala fra 0,0 (ikke svart på noe) til 5,0 (alt riktig).
  Dette betyr at man bruker denne formelen for elevenes skåre:
  **S = (5,33*K + 29,5*R + 16,33*T + 6,83*M + 5*P)/63**

- Siden skalaen fra 0 til 5, som påvist ovenfor, ikke representerer noen kriteriebasert skala, så er det imidlertid bedre å overføre denne skåren til en normrelatert skala med utgangspunkt i gjennomsnitt (2,60) og standardavvik (0,98) for alle elevene i utvalget.
- Hvis man bestemmer seg for dette, betyr det at man transformerer elevenes skåre S til en standardisert skåre, T, for eksempel en vanlig skala som har gjennomsnitt 50 poeng og standardavvik 10 poeng. Dette skjer ved å bruke formelen:
  **T = (S – 2,60) * 10/0,98 + 50**
- Med denne standardiserte skåreverdien betyr for eksempel en skåre på 60 poeng at eleven skårer ett standardavvik høyere enn gjennomsnittet, mens 45 betyr et halvt standardavvik lavere enn gjennomsnittet. Fordelen med en slik normrelatert, standardisert skala er at man uten videre kan sammenlikne elevenes relative prestasjoner i ulike fag, ut fra helt forskjellige prøver. Som det framgår av figur 5.1, framstår den standardiserte skåreverdien med en god og symmetrisk fordeling.

Vi vil videre sterkt anbefale at gruppa som arbeider med de nasjonale prøvene i matematikk, for årene framover legger helt om sitt arbeide slik at det er bedre i tråd med grunnleggende psykometriske prinsipper. Spesielt vil vi anbefale at man ved god utprøving sikrer høy reliabilitet til de planlagte rapporteringskategoriene og at det ikke er for mange av dem, rimeligvis to, eller i hvert fall ikke over tre. Vi vil anbefale å utvikle et nytt rammeverk med to eller i høyden tre kategorier, og der det innarbeides en rimelig jevn fordeling av oppgaver etter kategoriene. Vi mener å ha dokumentert at det rammeverket som er benyttet, ikke egner seg for dette formålet. Endelig vil vi anbefale å sørge for at prøven som helhet blir lettere i årene framover.

*Figur 5.1: Fordeling av standardiserte skåreverdier for matematikkprøven i 10. klasse*



## *5.3 Matematikk 4. klasse*

### 5.3.1 Prøvens struktur, validitet og kriterier ved retting

Vi vil henvise til 5.2.1 når det gjelder struktur og kriterier for vurdering av denne prøven, idet prinsippene for de to prøvene er nøyaktig de samme. Prøven inneholder 48 åpne oppgaver (oppgave 15 er ikke analysert, siden kategori H ikke fungerte etter hensikten, se 5.2.1), og med dobbeltretting var det i alt 69 ”oppgaver” som ble vurdert.

Isolert sett vurderer vi også denne matematikkprøvens innholdsvaliditet som høy. Lærerne rapporterer i Gallup-undersøkelsen at oppgavenes form ”i stor grad” er velkjent for elevene, og oppgavene er etter vår vurdering meget gode når det gjelder å måle god forståelse i faget. Vi mener oppgavesettet reflekterer læreplanen på en meget god måte, noe lærerne tydeligvis er enig med oss i: Bare rundt 10 % av dem vurderte at oppgavene reflekterer sentrale mål i Læreplanen i liten eller svært liten grad.

På spørsmålet ”I hvilken grad mener du elevene fikk vist sine ferdigheter gjennom prøven?” svarte bare rundt 13 % av lærerne ”i liten grad” eller ”i svært liten grad”. Vi mener at på bakgrunn av god validitet i forhold til læreplanens innhold og intensjoner, vil prøven på mange måter kunne gi et godt signal tilbake til skolene om hva som er viktig å vektlegge i undervisningen. Likevel vil vi peke på at som for 10. klasse er det et åpent spørsmål, som derfor bør nøye diskuteres, i hvilken grad prøven vektlegger den delen av matematikkompetansen som Kvalitetsutvalget kalte ”basiskompetanse”, og som St. meld. Nr. 30 kalte ”grunnleggende ferdigheter”.

### 5.3.2 Item-analyse

Resultatene fra item-analysen er gitt i tabell 5.3. Hver oppgave er angitt med oppgavens nummer samt en bokstav som angir hvilken kategori den tilhører. På grunn av dobbeltkategorisering vil noen oppgaver forekomme på to linjer. For mange oppgaver er svarene fordelt bare på noen få nivåer, eller noen nivåer er nesten ikke brukt.

Som mål for elevenes dyktighet har vi brukt gjennomsnittlig nivå for oppgavene som hører inn under den aktuelle kategorien. Dette innebærer altså at vi foreløpig bare bruker nivåene som en poengskala.

*Tabell 5.3: Item-analyse for matematikk i 4. klasse. Flere av oppgavene forekommer to eller tre ganger, en gang for hver kompetanse som er vurdert. Prosentfordelingen er avrundet til hele tall, og dyktigheten (gjennomsnittlig ”nivå” for de som har svart slik) til en desimal. D står for oppgavens diskriminering. I kolonnen for kommentarer (Kom) er det henvist til ulike fotnoter under tabellen. Et utropstegn i denne kolonnen betyr at vi har å gjøre med et betydelig problem.*

| Oppg nr | Svarfordeling i % | | | | | | Dyktighet | | | | | | D | Kom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 0 | 1 | 2 | 3 | 4 | 5 | 0 | 1 | 2 | 3 | 4 | 5 | | |
| 1a R | 1 | 1 | 1 | | | 98 | 2,1 | 3,0 | 2,6 | | | 3,6 | ,16 | a, b |
| 1b R | 2 | 82 | 1 | 0 | | 16 | 3,4 | 3,5 | - | - | | 4,1 | ,29 | b |
| 2 R | 2 | 22 | 0 | 0 | 0 | 75 | 2,8 | 3,1 | - | - | - | 3,8 | ,40 | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3 K | 2 | 20 | 5 | 27 | 2 | 44 | 1,6 | 2,1 | 2,4 | 2,6 | 2,7 | 3,4 | ,47 | |
| 3 R | 2 | 23 | 16 | 4 | 12 | 44 | 2,4 | 3,2 | 3,4 | 3,4 | 3,7 | 4,0 | ,50 | |
| 3 M | 2 | 8 | 7 | 16 | 13 | 55 | 2,2 | 2,5 | 3,0 | 3,4 | 3,7 | 4,3 | ,48 | |
| 4 K | 10 | 52 | 10 | 1 | 14 | 13 | 1,9 | 2,6 | 2,6 | 3,0 | 3,3 | 4,0 | ,49 | |
| 4 T | 10 | 29 | 18 | 17 | 13 | 14 | 2,3 | 2,8 | 3,2 | 3,5 | 3,5 | 3,9 | ,60 | |
| 4 P | 10 | 30 | 17 | 16 | 13 | 14 | 2,1 | 2,5 | 3,1 | 3,4 | 3,4 | 3,8 | ,59 | |
| 5a R | | 4 | 0 | 1 | 0 | 95 | | 3,3 | - | - | - | 3,6 | ,07 | b |
| 5b R | 1 | 4 | 6 | 2 | 1 | 88 | 1,7 | 2,6 | 3,3 | 3,5 | - | 3,7 | ,28 | b |
| 5c R | 2 | 10 | 8 | 2 | 1 | 79 | 1,8 | 3,0 | 3,4 | 3,6 | 3,5 | 3,7 | ,33 | a |
| 5d R | 2 | 24 | 4 | 2 | 1 | 68 | 1,9 | 3,1 | 3,6 | 3,9 | 3,3 | 3,9 | ,41 | a |
| 5e R | 2 | 9 | 1 | 2 | 1 | 86 | 1,9 | 3,0 | 3,3 | 3,3 | 3,4 | 3,7 | ,33 | |
| 5f R | 4 | 28 | 11 | 3 | 1 | 53 | 2,4 | 3,2 | 3,4 | 3,8 | 3,7 | 4,0 | ,46 | a |
| 5g R | 7 | 25 | 4 | 3 | 1 | 60 | 2,6 | 3,4 | 3,4 | 3,4 | 3,5 | 3,9 | ,37 | |
| 6a T | 0 | 0 | 3 | | 8 | 89 | - | - | 2,4 | | 2,7 | 3,3 | ,25 | b |
| 6a P | 0 | 0 | 3 | | 7 | 89 | - | - | 2,2 | | 2,4 | 3,1 | ,26 | b |
| 6b T | 5 | 8 | 4 | 1 | 44 | 39 | 2,3 | 2,6 | 2,8 | 2,8 | 3,2 | 3,4 | ,32 | |
| 6b P | 5 | 8 | 4 | 1 | 44 | 40 | 1,9 | 2,3 | 2,6 | 2,6 | 3,0 | 3,3 | ,32 | |
| 6c T | 5 | 26 | 31 | 1 | 1 | 36 | 2,5 | 2,7 | 3,0 | 2,9 | 3,4 | 3,7 | ,51 | a |
| 6c P | 5 | 26 | 31 | 1 | 1 | 36 | 2,0 | 2,5 | 2,8 | 2,9 | 3,5 | 3,7 | ,51 | |
| 6d T | 7 | 20 | 4 | 0 | 0 | 68 | 2,4 | 2,7 | 2,6 | - | - | 3,4 | ,42 | a |
| 6d P | 7 | 23 | 0 | 0 | 0 | 68 | 1,9 | 2,3 | - | - | - | 3,3 | ,42 | |
| 7 R | 9 | 16 | 2 | 1 | 11 | 61 | 2,8 | 3,1 | 3,5 | 3,5 | 3,5 | 3,9 | ,48 | |
| 8 R | 1 | 2 | 69 | 2 | 0 | 26 | 1,8 | 3,2 | 3,6 | 3,4 | - | 3,8 | ,22 | a, b ! |
| 9 K | 2 | 24 | 6 | 2 | 1 | 66 | 1,3 | 2,1 | 1,9 | 2,2 | 2,5 | 3,3 | ,46 | a ! |
| 9 M | 2 | 9 | 1 | 1 | 0 | 86 | 1,3 | 2,3 | 3,2 | 3,0 | - | 4,1 | ,46 | a |
| 10 R | 5 | 12 | 0 | 6 | 10 | 67 | 2,6 | 3,1 | - | 3,3 | 3,4 | 3,8 | ,43 | |
| 11a R | 1 | 4 | 1 | 12 | 10 | 72 | 2,2 | 2,4 | 2,9 | 3,3 | 3,5 | 3,8 | ,38 | |
| 11b R | 7 | 12 | 3 | 19 | 3 | 57 | 2,6 | 3,2 | 3,3 | 3,4 | 3,9 | 3,9 | ,45 | |
| 11c R | 8 | 6 | 10 | 6 | 0 | 71 | 2,6 | 3,1 | 3,2 | 3,4 | 3,8 | 3,8 | ,43 | |
| 11d R | 15 | 13 | 5 | 6 | 1 | 61 | 2,8 | 3,1 | 3,3 | 3,7 | 3,9 | 3,9 | ,53 | |
| 11e R | 9 | 29 | 6 | 15 | 7 | 33 | 2,6 | 3,4 | 3,7 | 3,8 | 3,9 | 3,9 | ,39 | |
| 12a R | 4 | 20 | 0 | | 36 | 40 | 2,7 | 3,1 | 3,2 | | 3,7 | 3,9 | ,48 | |
| 12a T | 4 | 20 | 0 | 0 | 35 | 40 | 1,9 | 2,6 | - | - | 3,2 | 3,6 | ,48 | |
| 12b R | 5 | 37 | 19 | 1 | 17 | 21 | 2,7 | 3,3 | 3,7 | 3,6 | 3,9 | 4,1 | ,49 | a |
| 12b T | 5 | 37 | 20 | 1 | 17 | 21 | 2,0 | 2,8 | 3,3 | 3,2 | 3,5 | 3,8 | ,49 | a |
| 12c R | 6 | 27 | 1 | 0 | 36 | 31 | 2,8 | 3,2 | 3,9 | - | 3,7 | 4,0 | ,49 | a |
| 12c T | 6 | 27 | 1 | 0 | 36 | 31 | 2,1 | 2,6 | 3,3 | - | 3,3 | 3,7 | ,49 | |
| 12d R | 6 | 35 | 42 | 0 | 13 | 3 | 2,8 | 3,3 | 3,8 | - | 4,0 | 4,2 | ,45 | |
| 12d T | 6 | 35 | 42 | 0 | 13 | 3 | 2,1 | 2,8 | 3,4 | - | 3,7 | 4,0 | ,45 | |
| 13a T | 5 | 9 | 2 | 1 | 1 | 83 | 1,8 | 2,4 | 2,5 | - | 2,8 | 3,4 | ,44 | |
| 13b T | 15 | 38 | 2 | 13 | 0 | 32 | 2,3 | 3,0 | 3,1 | 3,4 | - | 3,7 | ,46 | |
| 13c T | 16 | 11 | 8 | 1 | 0 | 64 | 2,3 | 2,7 | 3,0 | 3,3 | - | 3,5 | ,48 | |
| 13d T | 13 | 29 | 43 | 0 | | 15 | 2,1 | 3,0 | 3,4 | - | | 3,7 | ,37 | |
| 14 K | 11 | 40 | 12 | 2 | 1 | 34 | 1,8 | 2,5 | 2,8 | 2,9 | 2,8 | 3,6 | ,51 | a |
| 14 R | 11 | 41 | 3 | 11 | 1 | 34 | 2,9 | 3,4 | 3,5 | 3,8 | 3,8 | 4,0 | ,49 | |
| 16 K | 16 | 44 | 2 | 2 | 19 | 18 | 1,8 | 2,6 | 3,0 | 3,1 | 3,3 | 3,7 | ,52 | |
| 16 P | 16 | 23 | 23 | 14 | 10 | 13 | 2,1 | 2,7 | 3,0 | 3,4 | 3,6 | 3,8 | ,54 | |
| 17a K | 5 | 9 | 26 | 20 | 7 | 35 | 1,2 | 2,1 | 2,5 | 2,9 | 2,8 | 3,5 | ,41 | a ! |
| 17a T | 5 | 9 | 23 | 9 | 15 | 40 | 2,1 | 2,4 | 3,0 | 3,1 | 3,4 | 3,5 | ,47 | |
| 17a M | 5 | 11 | 6 | 25 | 3 | 51 | 1,7 | 2,6 | 2,9 | 3,7 | 3,9 | 4,4 | ,47 | |
| 17b K | 9 | 10 | 25 | 23 | 9 | 25 | 1,5 | 2,1 | 2,5 | 3,0 | 3,2 | 3,6 | ,52 | |
| 17b T | 9 | 15 | 4 | 16 | 29 | 28 | 2,3 | 2,7 | 2,8 | 2,9 | 3,4 | 3,7 | ,52 | |
| 17b M | 9 | 16 | 15 | 10 | 4 | 46 | 2,0 | 3,0 | 3,4 | 3,9 | 3,9 | 4,6 | ,52 | |
| 18 K | 17 | 25 | 2 | 26 | 14 | 18 | 1,9 | 2,4 | 2,6 | 3,0 | 3,3 | 3,7 | ,52 | |
| 18 P | 17 | 25 | 27 | 14 | 17 | 2 | 2,1 | 2,7 | 3,0 | 3,5 | 3,8 | 4,1 | ,52 | |
| 19a R | 10 | 2 | 0 | 2 | 0 | 86 | 2,3 | 2,5 | - | 3,2 | - | 3,8 | ,46 | |
| 19b R | 16 | 7 | 2 | 13 | 1 | 61 | 2,6 | 3,1 | 3,4 | 3,6 | - | 4,0 | ,56 | |
| 19c R | 12 | 3 | 1 | 4 | 1 | 79 | 2,4 | 3,1 | 3,1 | 3,3 | 3,8 | 3,9 | ,51 | |
| 19d R | 16 | 7 | 1 | 12 | 1 | 63 | 2,6 | 3,0 | 3,6 | 3,6 | - | 4,0 | ,54 | |
| 19e R | 11 | 1 | 1 | 0 | | 87 | 2,3 | 2,4 | 3,7 | - | | 3,8 | ,44 | |
| 19f R | 13 | 6 | 5 | 2 | 0 | 75 | 2,5 | 3,1 | 3,5 | 3,6 | - | 3,9 | ,51 | |
| 20 R | 21 | 16 | 12 | 3 | 8 | 40 | 2,8 | 3,3 | 3,8 | 3,8 | 3,9 | 4,0 | ,55 | |
| 21a P | 20 | 27 | 7 | 2 | 14 | 30 | 2,0 | 2,7 | 2,7 | 3,2 | 3,4 | 3,8 | ,49 | |
| 21b P | 18 | 27 | 3 | 2 | 1 | 49 | 2,1 | 2,4 | 2,8 | 3,1 | 3,4 | 3,6 | ,55 | |

| | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 22 K | 17 | 6 | 7 | 8 | 31 | 32 | 1,9 | 2,0 | 2,4 | 2,7 | 2,9 | 3,5 | ,52 | |
| 22 T | 17 | 5 | 10 | 3 | 33 | 33 | 2,5 | 2,5 | 2,8 | 2,8 | 3,2 | 3,7 | ,52 | |

a) Dyktigheten er ikke "ordnet" etter kompetanse.
b) Svak diskriminering (< 0,30)

Med utgangspunkt i resultatene i tabell 5.3 kan vi slå fast at denne prøven har veldig mange av de samme egenskapene som prøven for 10. klasse. Vi kommenterer derfor her bare det som er forskjellig for de to prøvene:

- Det er ca 8 % blanke svar (kategori 0) i gjennomsnitt for alle oppgavene, og bare for to oppgaver ligger andelen så vidt høyere enn 20 %. Dette er en mye bedre situasjon enn for prøven for 10. klasse.
- Gjennomsnittlig poeng oppnådd for alle oppgavene (hver oppgave bare en gang) er 3,48, noe som tilsvarer 69 % av "fullt hus". Dette er høyt, men vi må se det i lys av at de får ett poeng "gratis", altså for et verdiløst svar. Hvis vi i stedet gir 0 poeng for verdiløst svar og for øvrig senker alle poengene med 1, blir gjennomsnittlig oppnådde poeng 64 % av "fullt hus". Prøven har altså vært lett, men etter vår mening ikke for lett, ut fra hensyn til elevenes unge alder. Det er imidlertid vanskelig å forstå hvorfor det er lagt opp til så stor forskjell på vanskelighetsgraden til prøvene for de to klassetrinnene.
- Det er ut fra våre vurderinger vanskelig å forstå at så mye som 65 % av matematikklærerne i Gallup-undersøkelsen har vurdert prøven som "noe" eller "altfor" vanskelig. Her er vi tydeligvis ved et spørsmål som lærerne trolig tar stilling til mer ut fra tradisjoner og hensyn til svake elever enn hva som er god vurdering pedagogisk og testteoretisk sett.
- Vi har ikke kunnet studere hvordan vurderingen har fungert, siden vi ikke har data fra skolene. Vi har imidlertid indikasjoner på at som for 10. klasse har det vært mye uro på grunn av at prøven har vært så tidkrevende å rette. I gjennomsnitt har lærerne brukt så mye som 45 minutter per besvarelse (Gallup-undersøkelsen).

### 5.3.3 Analyse av de foreslåtte kategoriene

Informasjon om hvordan hver av de forslåtte kategoriene har fungert, er gitt i tabell 5.4. For hver kategori har vi gitt antall oppgaver, kategoriens reliabilitet (Cronbachs alfa) og gjennomsnittlig kompetansenivå for alle elevene. Som det framgår av tabellen, er det svært ujevn fordeling av oppgaver, noe som er medvirkende til at reliabiliteten varierer sterkt, fra 0,60 til 0, 88. Det er bare for kategorien R at reliabiliteten er høy nok til at publisering av resultatene kan være aktuell. Som tidligere nevnt, teller mange oppgaver inn under to eller tre kategorier på en gang. Dette har som naturlig konsekvens at de forslåtte skalaene blir likere og derfor korrelerer unormalt høyt med hverandre. Informasjonsverdien av hver kategori blir da selvsagt lavere enn om de var mer forskjellige.

Vi ser også at hver av skalaene har et gjennomsnittlig "nivå" som ligger godt over midtpunktet mellom 0 og 5, men det er vanskelig å innse at elevene skal være så mye (et helt poeng) "bedre" i kategorien M enn i K. Det er i det hele vanskelig å se hva kategorien "Kommunikasjon" egentlig dekker.

*Tabell 5.4: Data for hver av de foreslåtte kompetansene i 4. klasse*

| Kategori | Antall oppgaver | Gjennomsn. korrelasjon | Reliabilitet | Gjennomsn. kompetansenivå |
|---|---|---|---|---|
| Kommunikasjon (K) | 9 | 0,25 | 0,75 | 2,8 |
| Representasjon, symbolbruk og formalisme (R) | 31 | 0,18 | 0,88 | 3,6 |
| Matematisk resonnement og tankegang (T) | 16 | 0,20 | 0,80 | 3,2 |
| Matematisk modellering og anvendelse (M) | 4 | 0,26 | 0,60 | 3,8 |
| Problembehandling (P) | 9 | 0,22 | 0,73 | 3,0 |

Problemer med de foreslåtte kategoriene:

- Noen av dem har veldig lav reliabilitet, særlig fordi det er for få oppgaver. Som for 10. klasse dreier det seg særlig om kategoriene K, M og P.
- Det er i tillegg umulig å se at de ulike skalaene måler forskjellige ting. Korrelasjonene mellom skalaene er av samme størrelse som reliabiliteten til hver av dem, noe som innebærer at korrelasjonene er så store som de kan bli. Eller, om vi vil: de ”latente” korrelasjonene er tilnærmet lik 1,0 (eller faktisk høyere, fordi de samme oppgavene forekommer dobbelt). De foreslåtte kategoriene er altså beheftet med reliabilitetsproblemer samtidig som de på grunn av for høy innbyrdes korrelasjon har lav validitet.
- Både betegnelser og beskrivelser av de foreslåtte kategoriene kommuniserer dårlig til lærere og ledere hva kompetansene egentlig går ut på, og særlig hva som er forskjellen på dem. Dette er i 4. klasse et enda større problem enn i 10. klasse, siden lærerne er mye mindre skolert i matematikk, faglig og fagdidaktisk.
- Som for 10. klasse vil vi derfor advare både mot å publisere data for disse skalaene og mot å betrakte slike data som mål for enkeltelevers kompetanse..

### 5.3.4 Konklusjon

Denne prøven viser seg å ha nesten nøyaktig de samme egenskapene som prøven for 10. klasse, bortsett fra at vanskelighetsgraden synes å være lavere og derved mer hensiktsmessig. (Her er vi riktignok på kollisjonskurs med lærernes syn.) Vi konstaterer derfor at konklusjonen fra vår side om denne prøven stort sett blir den samme som for den andre. Spesielt betyr det at vi sterkt fraråder å rapportere som opprinnelig planlagt.

Derimot kan de foreliggende data som for 10. klasse slås sammen til én overordnet skala som måler matematikkompetanse generelt. Det dreier seg her om hver elevs ”nivå” (langs en skala fra 0 til 5) innen hver av de fem kategoriene K, R, T, M og P (se kap 5.2.1). Vi kan ikke ut fra disse tallene konstruere en samlet poengsum, der hver oppgave teller likt. Men vårt forslag til beste strategi er følgende, og den tilsvarer helt prosedyren for 10. klasse:

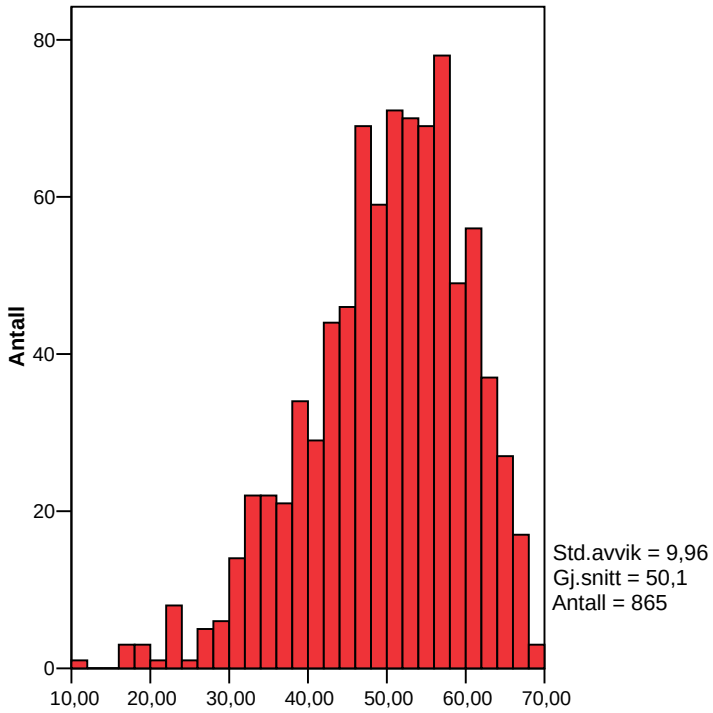
- Man tar utgangspunkt i antall oppgaver innenfor hver kategori, på en slik måte at oppgaver som er dobbeltkategorisert, teller 0,5, mens de som er tredobbelt kategorisert teller 0,33. Antall oppgaver innenfor de fem kategoriene blir da henholdsvis 3,83; 27,83; 9,5; 1,5 og 5,33. Ved å vekte de fem kategoriene etter disse ”antallene” kommer man fram til en gjennomsnittlig skåre på de 48 oppgavene langs en skala fra 0,0 (ikke svart på noe) til 5,0 (alt riktig).

Dette betyr at man bruker denne formelen for elevenes skåre:
**S = (3,83*K + 27,83*R + 9,5*T + 1,5*M + 5,33*P)/48**

- Siden skalaen fra 0 til 5, som påvist ovenfor, likevel ikke representerer noen kriteriebasert skala, så er det bedre å overføre denne skåren til en normrelatert skala med utgangspunkt i gjennomsnitt (3,40) og standardavvik (0,76) for alle elevene i utvalget.
- Hvis man bestemmer seg for dette, betyr det at man transformerer elevenes skåre S til en standardisert skåre, for eksempel en skala med gjennomsnitt 50 poeng og standardavvik 10 poeng. Dette skjer ved å bruke formelen:
  **T = (S – 3,40) * 10/0,76 + 50**
- Fordelen med en slik normrelatert, standardisert skala er, som nevnt for 10. klasse, at man uten videre kan sammenlikne elevenes relative prestasjoner i ulike fag, ut fra helt forskjellige prøver.

Figur 5.2 viser fordelingen av den endelige skåreverdien ved å følge den angitte transformasjonen. Som det framgår av figuren, er fordelingen litt skjev, siden prøven er litt lett og derfor har en liten ”takeffekt” og diskriminerer litt dårlig blant flinke elever.

*Figur 5.2: Fordeling av standardiserte skåreverdier for matematikkprøven i 4. klasse*




Våre videre anbefalinger til gruppa som arbeider med de nasjonale prøvene i matematikk, er for øvrig helt i tråd med det vi sa om prøven for 10. klasse. Men et enklere system for vurdering er enda viktigere på det lavere klassetrinnet.

## *5.4 Lesing 10 klasse*

### 5.4.1 Struktur, validitet og vurderingskriterier

Denne prøven framstår nesten som en kopi av 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 samlet i 10 oppgaveenheter med til sammen 50 oppgaver. Hver enhet og hver oppgave er klassifisert etter et sett av kriterier, og det er foreslått å rapportere etter de samme tre kategoriene som i PISA: Finne, Tolke og Reflektere (se mer om dette nedenfor).

En vurdering av prøvens validitet er ikke lett, så lenge det ikke finnes noen læreplan i "lesing". I Gallup-undersøkelsen sier bare 62 % av lærerne at prøven "i svært stor" eller "i noen grad" reflekterer sentrale mål i Læreplanen. Vi konstaterer imidlertid at målt med internasjonale mål, slik de er gjenspeilet i PISA-prosjektet, så representerer denne prøven nettopp det som internasjonalt forstås som "reading literacy", eller på norsk, leseforståelse. Vi mener derfor at prøven har rimelig høy validitet, ut fra hva man med rimelighet kan forlange. Imidlertid konstaterer vi at det åpenbart er et stort behov for at faggruppa lager et tydelig rammeverk for den nasjonale leseprøven (gjerne i samarbeid med faggruppa for lesing blant de yngre elevene). Et slikt rammeverk burde inneholde en tydelig definisjon av lesekompetanse og de eventuelle delkompetanser man tar mål av seg til å rapportere etter. Videre kunne det da framgå hvilken vekt (antall oppgaver) hver kategori skulle ha. Endelig kunne det gis et rasjonale for den foreslåtte struktur på prøven. Å kopiere PISA-prosjektets design har flere fordeler, men det bør begrunnes hvis dette som en selvfølge brukes år etter år.

Resultatene av item-analysene er i det følgende gitt for flervalgsoppgavene og de åpne oppgavene hver for seg. Som mål på elevenes dyktighet har vi i vår analyse brukt antall poeng oppnådd. For de 23 flervalgsoppgavene har riktig alternativ gitt ett poeng, mens galt eller blankt svar har gitt 0 poeng. De 27 åpne oppgavene er i de fleste tilfellene vurdert til 1 poeng (riktig) eller 0 poeng (galt). For noen få oppgaver er det brukt 2, 1 eller 0 poeng.

### 5.4.2 Item-analyse av flervalgsoppgaver

Resultatene for flervalgsoppgavene er vist oppgave for oppgave i tabell 5.5.

*Tabell 5.5: Item-analyse for flervalgsoppgaver i lesing for 10. klasse. Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. I kolonnen for kommentarer (Kom) er det henvist til ulike fotnoter under tabellen.*

| Oppg nr | Svarfordeling i % | | | | | Dyktighet | | | | | D | Kom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | Blank | A | B | C | D | Blank | | |
| 1 (B) | 1 | 98 | 1 | 0 | 0 | - | 35 | 20 | | - | ,19 | b |
| 3 (B) | 1 | 95 | 3 | 2 | 0 | 38 | 35 | 30 | 31 | - | ,11 | a, b |
| 5 (B) | 1 | 86 | 10 | 3 | 1 | 23 | 36 | 29 | 25 | 23 | ,31 | |
| 6 (C) | 0 | 3 | 96 | 0 | 1 | - | 26 | 35 | - | 19 | ,21 | b |
| 7 (A) | 93 | 5 | 0 | 1 | 1 | 36 | 28 | - | 22 | 15 | ,26 | b |
| 8 (B) | 7 | 64 | 27 | 2 | 1 | 32 | 38 | 30 | 24 | 10 | ,41 | |
| 15 (C) | 44 | 4 | 34 | 14 | 4 | 34 | 29 | 40 | 32 | 23 | ,35 | |
| 18 (B) | 2 | 66 | 4 | 26 | 3 | 21 | 37 | 27 | 33 | 20 | ,32 | |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 21 (C) | 9 | 10 | 74 | 3 | 5 | 26 | 32 | 38 | 27 | 20 | ,44 | |
| 22 (D) | 6 | 7 | 14 | 72 | 1 | 32 | 35 | 28 | 37 | 15 | ,34 | |
| 23 (B) | 3 | 65 | 26 | 4 | 2 | 33 | 38 | 32 | 28 | 16 | ,34 | |
| 27 (C) | 4 | 5 | 86 | 2 | 3 | 26 | 27 | 37 | 17 | 15 | ,46 | |
| 30 (A) | 83 | 1 | 2 | 10 | 4 | 37 | 17 | 18 | 33 | 16 | ,41 | |
| 31 (D) | 3 | 8 | 9 | 76 | 5 | 22 | 27 | 32 | 38 | 17 | ,51 | |
| 32 (D) | 1 | 22 | 1 | 71 | 5 | 25 | 33 | 21 | 37 | 19 | ,33 | |
| 36 (A) | 74 | 9 | 8 | 2 | 7 | 38 | 31 | 31 | 24 | 20 | ,45 | |
| 38 (B) | 8 | 72 | 4 | 8 | 9 | 30 | 39 | 21 | 30 | 22 | ,56 | |
| 40 (C) | 15 | 3 | 67 | 4 | 11 | 29 | 25 | 39 | 30 | 24 | ,54 | |
| 45 (B) | 3 | 70 | 7 | 6 | 15 | 33 | 38 | 30 | 26 | 27 | ,48 | |
| 46 (D) | 13 | 14 | 6 | 50 | 17 | 35 | 30 | 29 | 40 | 27 | ,46 | |
| 47 (C) | 20 | 20 | 20 | 21 | 19 | 31 | 35 | 42 | 38 | 28 | ,36 | a |
| 48 (D) | 2 | 9 | 3 | 66 | 20 | 27 | 30 | 33 | 38 | 28 | ,44 | |
| 49 (C) | 3 | 7 | 71 | 0 | 19 | 26 | 28 | 38 | - | 28 | ,49 | |

a) En distraktor velges av for flinke elever.
b) Svak diskriminering (<0,30)

Med utgangspunkt i resultatene i tabell 5.5 har vi disse kommentarene:

- Disse oppgavene har i store trekk fungert etter hensikten**.** Særlig er det påfallende at nesten alle oppgavene diskriminerer godt. For de få oppgavene der diskrimineringen er lav (b i høyre kolonne), er det fordi det for svært lette oppgaver er vanskelig å oppnå høye verdier.
- Vi ser også at de gale alternativene i så godt som alle tilfeller velges av elever som gjennomgående ligger under gjennomsnittet totalt på prøven (35 poeng).
- Det er noen distraktorer som er valgt av svært få elever, og som derfor med fordel kunne ha vært erstattet med andre (særlig for oppgave nr 5, 30, 32 og 49).
- Det er 6,7 % blanke svar i gjennomsnitt for alle flervalgsoppgavene. Det er ikke særlig høyt, men den sterke stigningen mot slutten av prøven er litt betenkelig, skjønt kanskje uunngåelig.
- Gjennomsnittlig poeng oppnådd for alle de 23 flervalgsoppgavene utgjør 72 % av ”fullt hus”. Dette er svært høyt, og det kunne med fordel vært noen vanskeligere oppgaver av dette formatet.

### 5.4.3 Item-analyse av åpne oppgaver

Item-analysen for de åpne oppgavene er vist i tabell 5.6.

*Tabell 5.6: Item-analyse for åpne oppgaver i lesing for 10. klasse. Svarfordelingen og dyktigheten (poeng oppnådd på prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. Når det gjelder sensorreliabilitet, står R for prosentandelen der de to sensorene har vurdert likt, mens K står for Kappa. I kolonnen for kommentarer (Kom) er det henvist til fotnoten under tabellen. Et utropstegn i denne kolonnen indikerer et betydelig problem.*

| Oppg nr | Svarfordeling i % | | | | Dyktighet | | | | D | R | K | Kom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Blank | 0 | 1 | 2 | Blank | 0 | 1 | 2 | | | | |
| 2 | 7 | 18 | 75 | | 26 | 30 | 37 | | ,38 | 97 | ,94 | |
| 4 | 1 | 11 | 88 | | 22 | 26 | 36 | | ,36 | 92 | ,61 | |
| 9 | 4 | 12 | 83 | | 17 | 26 | 37 | | ,50 | 92 | ,68 | |
| 10 | 6 | 26 | 68 | | 21 | 30 | 38 | | ,43 | 87 | ,69 | |
| 11 | 2 | 14 | 31 | 53 | 11 | 27 | 34 | 39 | ,48 | 79 | ,63 | c |
| 12 | 6 | 67 | 27 | | 19 | 34 | 41 | | ,35 | 94 | ,83 | |
| 13 | 13 | 38 | 49 | | 25 | 34 | 39 | | ,39 | 97 | ,95 | |

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 14 | 13 | 41 | 46 | | 25 | 33 | 39 | | ,38 | 96 | ,92 | |
| 16 | 4 | 26 | 70 | | 21 | 30 | 38 | | ,41 | 97 | ,92 | |
| 17 | 16 | 36 | 48 | | 26 | 33 | 39 | | ,39 | 83 | ,67 | c |
| 19 | 6 | 11 | 83 | | 17 | 27 | 37 | | ,52 | 89 | ,64 | |
| 20 | 9 | 35 | 57 | | 20 | 31 | 40 | | ,52 | 76 | ,50 | c! |
| 24 | 6 | 13 | 81 | | 16 | 28 | 38 | | ,51 | 95 | ,82 | |
| 25 | 8 | 8 | 84 | | 18 | 27 | 38 | | ,56 | 94 | ,73 | |
| 26A | 18 | 18 | 65 | | 25 | 29 | 39 | | ,59 | 90 | ,76 | |
| 26B | 19 | 34 | 47 | | 24 | 33 | 41 | | ,57 | 87 | ,74 | |
| 28 | 5 | 4 | 91 | | 14 | 20 | 37 | | ,59 | 99 | ,90 | |
| 29 | 8 | 14 | 79 | | 17 | 29 | 38 | | ,53 | 94 | ,79 | |
| 33 | 14 | 19 | 40 | 26 | 21 | 31 | 37 | 42 | ,58 | 65 | ,47 | c! |
| 34 | 11 | 30 | 59 | | 20 | 32 | 39 | | ,53 | 81 | ,60 | c |
| 35 | 11 | 8 | 81 | | 20 | 25 | 38 | | ,61 | 97 | ,90 | |
| 37 | 19 | 26 | 55 | | 24 | 32 | 41 | | ,60 | 82 | ,63 | c |
| 39 | 16 | 14 | 70 | | 23 | 29 | 39 | | ,62 | 90 | ,76 | |
| 41 | 12 | 16 | 72 | | 23 | 32 | 38 | | ,44 | 97 | ,91 | |
| 42 | 19 | 36 | 46 | | 25 | 35 | 40 | | ,41 | 77 | ,53 | c! |
| 43 | 28 | 22 | 38 | 13 | 26 | 34 | 39 | 44 | ,54 | 74 | ,57 | c! |
| 44 | 21 | 34 | 45 | | 26 | 32 | 42 | | ,58 | 87 | ,75 | |

c) Dårlig overensstemmelse mellom rettere (< 85 %, evt < 75 %, evt Kappa <0,60)

For de åpne oppgavene kan vi konkludere at disse i stor grad har fungert bra, både med hensyn til sensorreliabilitet og diskriminering. Dette tyder på et grundig arbeid med utvikling og utprøving av oppgaver med gode vurderingskriterier. Vi vil imidlertid påpeke at den rapporterte tiden til vurdering på 31 minutter per besvarelse (Gallup-undersøkelsen) er svært høy, etter vår mening urimelig høy. Denne kan, og etter vår mening bør, senkes neste år ved et noe lavere innslag av åpne oppgaver. Noen ytterligere kommentarer:

- For noen av oppgavene er det lavt samsvar mellom de to vurderingene. Dette gjelder særlig for ”2-poengs-oppgavene”. Kanskje burde man her ha nøyd seg med å gi ett poeng?
- Gjennomsnittlig er det 11 % blanke svar, og bare på de to siste oppgavene ligger andelen over 20 %. Dette er et gunstig resultat for dette formatet.
- Gjennomsnittlig poeng oppnådd for alle de 27 åpne oppgavene utgjør 62 % av ”fullt hus”. Dette er svært høyt til å være åpne oppgaver, og det kunne derfor med fordel vært noen vanskeligere oppgaver også av dette formatet.
- Det er for alle de 32 skolene samlet sett en liten tendens til at lærerne gir en litt høyere vurdering enn ekspertene. 10 av skolene har vurdert sine besvarelser til gjennomsnittlig minst 0,05 poeng per oppgave, eller totalt omtrent 1,3 poeng, bedre enn ekspertene. Tilsvarende er det bare én skole (med tre elever!) som ligger like mye lavere enn ekspertene. Denne tendensen til ”snill” retting er ikke urovekkende stor og vil ikke påvirke forskjellene mellom skoler i stor grad. Usikkerheter av denne størrelsesorden må man likevel regne med.

### 5.4.4 Oversikt over de foreslåtte kategoriene

*Tabell 5.7: Resultater for hver av kategoriene*

| Kategori | Antall oppgaver | Gjennomsn. korrelasjon | Reliabilitet | Gjennomsn. andel av ”fullt hus” |
|---|---|---|---|---|
| ”Finne” | 20 | 0,17 | 0,80 | 68 % |
| ”Tolke” | 18 | 0,16 | 0,77 | 71 % |
| ”Reflektere” | 12 | 0,26 | 0,80 | 57 % |

| | | | | |
|---|---|---|---|---|
| Totalt | 50 | 0,18 | 0,92 | 66 % |

Tabell 5.7 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 i de to første.

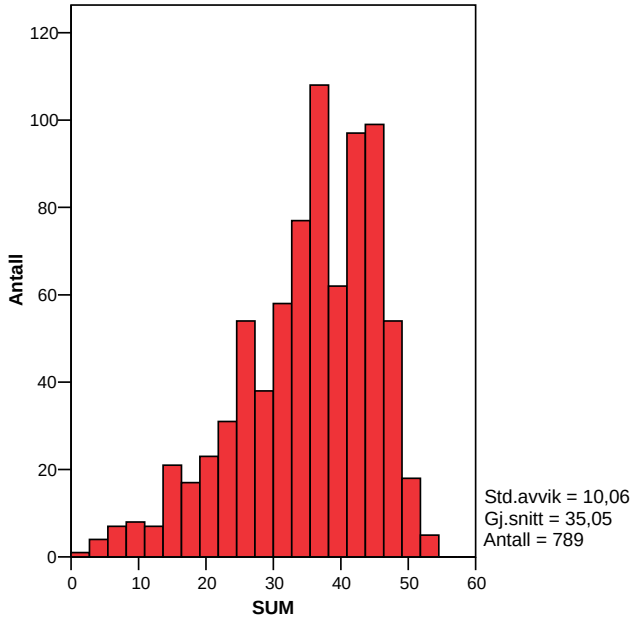
Dessverre viser det seg at alle tre skalaene har for lav reliabilitet til å kunne rapporteres. Videre er det slik at korrelasjonene mellom de tre kategoriene er alle omtrent like store (0,77) og omtrent like store som reliabilitetskoeffisientene (se tabell 5.7). Det er derfor veldig vanskelig ut fra empiri å se at de foreslåtte kategoriene virkelig representerer forskjellige kompetanser. Vi må derfor konstatere at de tre kategoriene hver for seg har litt tvilsom (diskriminerende) validitet. Slik dette ser ut, tyder det på at det beste ville være å rapportere bare én skala, generell kompetanse i lesing.

### 5.4.5 Prøven som helhet

På figur 5.3 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 for lett. Det er imidlertid noe forbausende at nesten ingen lærere (bare rundt 5 %) er enig med oss i dette, ifølge Gallup-undersøkelsen. Dette kan tolkes på mange måter, blant annet at lærerne har lave forventninger om elevenes lesekompetanse, eller at de kvier seg for å gi elevene noe å strekke seg etter. Utvilsomt ville innslag av noe vanskeligere oppgaver ha gitt bedre diskriminering blant de gode elevene og trolig også litt høyere reliabilitet. Prøven har altså en viss ”takeffekt”, men det hindrer ikke at den fungerer rimelig godt for å måle lesekompetanse for enkeltelever på det aktuelle klassetrinnet.

Men med en slik skjevhet blir det mer problematisk å beregne gjennomsnitt for skoler og klasser, siden de svake elevene vil ha en tendens til å påvirke dette gjennomsnittet i litt for stor grad, mens de gode elevene ikke riktig får vist sitt beste. I en slik situasjon vil det være særlig følsomt *hvilke* av de svakeste elevene i en klasse som faktisk deltok, og hvilke som ble holdt utenfor ved gjennomføringen av prøven.

*Figur 5.3: Fordeling av poeng for lesing i 10. klasse*




### 5.4.6 Konklusjon og anbefalinger for neste år

Prøven har fungert rimelig bra, men det ser vanskelig ut å rapportere resultatene etter mer enn én skala. Vi anbefaler følgelig at man rapporterer etter en overordnet skala i lesing. De angitte poengene framstår som en tilfeldig skala, verken norm- eller kriterierelatert. Og dette gjelder enten man regner i poeng eller i prosent riktige svar. Vi foreslår derfor at man regner om de rapporterte poengsummene (”Sum” av de tre skalaene) til en standardisert skåreverdi på samme måte som for matematikk. Dette gjøres ved å basere seg på gjennomsnitt (35,0) og standardavvik (10,06) til fordelingen. Formelen blir da:

**T = (Sum – 35)*10/10,06 + 50**

Med en slik transformasjon blir fordelingen i figur 5.3 uforandret, selv om verdien på x-aksen endres. Fordelen med en slik standardisert skåre er at resultatet i lesing kan sammenliknes direkte med matematikkresultatet, idet de to resultatene er målt med samme normrelaterte skala. Det gir da mening å si at en elev er ”et halvt standardavvik bedre” i lesing enn i matematikk. Det er det nærmeste vi kan komme en ”profil” for elevenes kompetanse for de nasjonale prøvene i 2004.

Med en målrettet utprøving og balansert fordeling av oppgaver er det trolig mulig å lage to gode skalaer i lesing, under forutsetning av at de to skalaene både gir pedagogisk mening og viser seg å være tilstrekkelig (i hvert fall *litt*) forskjellige. Vi vil ikke her spekulere over hvilke to kategorier dette kan være. For kommende år anbefaler vi at man konsentrerer seg om to kategorier, og det er da viktig for faggruppa å komme fram til hvilke to dette best kan være.

Det er også et problem at prøven har vært for lett. Dette har, som diskutert i 5.4.5, noen litt problematiske konsekvenser, særlig at svake elever ”teller” for mye for skoleresultatene. Vi anbefaler sterkt at det for neste år blir laget noen flere litt

vanskelige oppgaver. Vi har også kommentert ovenfor at vi for å redusere lærernes arbeidsbyrde foreslår et lavere innslag av åpne oppgaver neste år enn det var dette året.

Som beskrevet i 5.4.1, vil vi foreslå at det utarbeides et grundig rammeverk for nasjonale prøvene i lesing på de aktuelle trinnene. Dette er særlig viktig fordi det 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. Den store forskjellen i struktur mellom leseprøvene i 4. og 10. klasse 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.

## *5.5 Lesing 4.klasse*

### 5.5.1 Struktur, vurdering og validitet

Denne prøven inneholder to lange tekster, den første er et eventyr, mens den andre er en tekst som forklarer hva trolldom og magi dreier seg om. Til hver av tekstene er det 16 flervalgsoppgaver. Dette designet likner mye på det som var i PIRLS-undersøkelsen i 2002. De to tekstene er av forskjellig sjanger, den første er skjønnlitterær og den andre er sakprosa. I PIRLS dannet ulike sjangere basis for ulike rapporteringskategorier, men det er det ikke rimelig å gjøre her, siden det bare er én tekst i hver sjanger. Vi gir likevel resultater for hver tekst for seg før vi ser på prøven som helhet i 5.5.3.

Som nevnt i forbindelse med leseprøven i 10. klasse, savner vi en begrunnelse for hvorfor det er så stor forskjell mellom de to prøvene. Særlig undres vi over at det bare er to lange tekster her, mens det er mange flere i 10. klasse. 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 begrunnet. Uten at vi skal påberope oss å være fagdidaktiske eksperter på dette området, vil vi likevel hevde at det kanskje burde vært flere, kortere og noe enklere tekster enn det som er brukt her.

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.

Som beskrevet inngående for leseprøven i 10. klasse, er det veldig vanskelig å diskutere validitet så lenge det ikke finnes noen egentlig læreplan i "faget". Likevel vil vi peke på at lærerne synes å ha hatt problemer med å akseptere denne leseprøven for 4. klasse. Så mye som halvparten av lærerne mener at elevene "i liten grad" eller "i svært liten grad" fikk vist sine ferdigheter gjennom prøven. Tilsvarende mener omtrent halvparten at prøven "i liten grad" eller "i svært liten grad" reflekterer sentrale mål i Læreplanen. Det er altså tydelig at denne prøven har fått dårlig tilslutning ute i skolen.

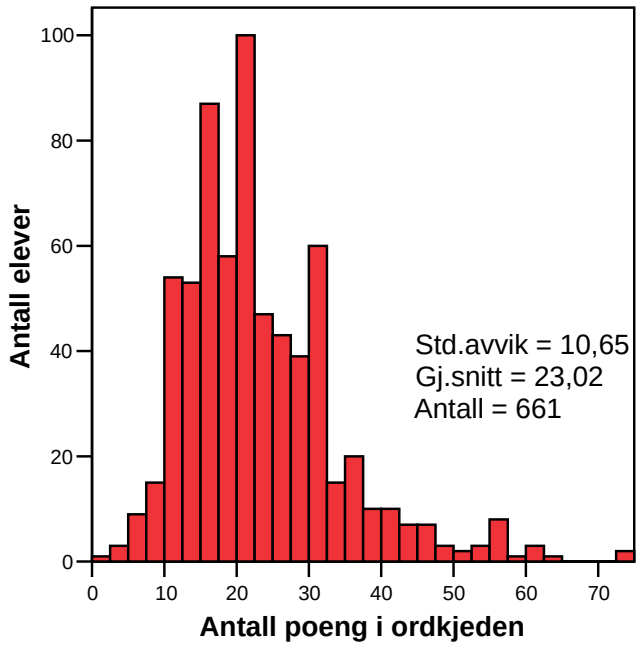
Prøven inneholder altså ingen åpne oppgaver, noe vi savner en begrunnelse for. På bakgrunn av dette er det forbausende at lærerne i Gallup-undersøkelsen rapporterer så mye tid som gjennomsnittlig 23 minutter brukt på en besvarelse. Det kan umulig ta så lang tid å registrere svar på flervalgsoppgavene. Dette må derfor trolig henge sammen

med at ordkjedeprøven har vært tidkrevende å telle opp, noe som i seg selv gir grunn til vurdering av om den forsvarer sin plass i en nasjonal prøve.

### 5.5.2 Resultater for ordkjedeprøven

Figur 5.4 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. Det har trolig vært svært uheldig at ordkjedeprøven som ble lagt ut på nettet, hadde en tidsmargin på 15 minutter, mens den virkelige prøven bare tillot 5 minutter. Dette kan ha vært årsaken til uklarheten om tiden til rådighet.

*Figur 5.4: Fordeling av poeng (riktige ord) på ordkjedeprøven*




Det er 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. Vi vil likevel peke på at en slik prøve nødvendigvis opptar noe tid til fortrengsel for den ”egentlige” prøven i lesing. Og ikke minst, en nøyaktig opptelling krever kanskje en urimelig bruk av lærernes tid. Hensiktsmessigheten av denne delen av prøven bør diskuteres spesielt.

### 5.5.3 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 riktig alternativ gitt ett poeng, mens galt eller blankt svar har gitt 0 poeng.

*Tabell 5.8: Item-analyse for oppgavene i lesing for 4. klasse. Svarfordelingen og dyktigheten (gjennomsnittlig oppnådd poeng på hele prøven for de som har svart slik) er avrundet til hele tall. D står for oppgavens diskriminering. I kolonnen for kommentarer (Kom) er det henvist til ulike fotnoter under tabellen.*

| Oppg nr | Svarfordeling i % | | | | | | Dyktighet | | | | | | D | Kom |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | A | B | C | D | Flere svar | Blank | A | B | C | D | Flere svar | Blank | | |
| S1 (A) | 86 | 2 | 3 | 3 | 3 | 4 | 18 | 9 | 10 | 12 | 10 | 5 | ,45 | |
| S2 (B) | 13 | 61 | 9 | 10 | 2 | 5 | 16 | 19 | 11 | 12 | 12 | 8 | ,46 | |
| S3 (C) | 9 | 13 | 57 | 15 | 2 | 4 | 11 | 14 | 19 | 14 | 10 | 3 | ,52 | |
| S4 (C) | 21 | 6 | 63 | 4 | 1 | 5 | 14 | 11 | 19 | 11 | 8 | 4 | ,53 | |
| S5 (D) | 3 | 8 | 5 | 77 | 1 | 6 | 12 | 12 | 14 | 18 | 11 | 5 | ,46 | |
| S6 (D) | 27 | 44 | 7 | 15 | 1 | 7 | 14 | 18 | 16 | 21 | 15 | 5 | ,27 | b |
| S7 (D) | 7 | 3 | 6 | 76 | 1 | 7 | 11 | 14 | 15 | 18 | - | 4 | ,51 | |
| S8 (C) | 31 | 16 | 37 | 6 | 1 | 10 | 17 | 16 | 20 | 13 | 14 | 6 | ,41 | |
| S9 (C) | 4 | 10 | 50 | 25 | 1 | 10 | 13 | 15 | 20 | 14 | 13 | 5 | ,56 | |
| S10 (B) | 12 | 61 | 10 | 4 | 1 | 13 | 14 | 20 | 14 | 12 | 7 | 7 | ,57 | |
| S11 (A) | 56 | 3 | 6 | 22 | 1 | 12 | 19 | 13 | 14 | 16 | 19 | 6 | ,49 | a |
| S12 (A) | 63 | 9 | 5 | 8 | 1 | 14 | 20 | 15 | 11 | 14 | - | 7 | ,61 | |
| S13 (B) | 12 | 62 | 4 | 7 | 1 | 14 | 15 | 20 | 11 | 14 | 11 | 7 | ,59 | |
| S14 (C) | 8 | 10 | 59 | 6 | 1 | 16 | 13 | 15 | 20 | 13 | 8 | 7 | ,64 | |
| S15 (B) | 40 | 33 | 5 | 6 | 1 | 16 | 16 | 21 | 15 | 17 | 15 | 7 | ,46 | |
| S16 (C) | 12 | 11 | 54 | 6 | 1 | 17 | 15 | 16 | 20 | 14 | 13 | 8 | ,53 | |
| T1 (B) | 5 | 85 | 6 | 2 | 1 | 1 | 11 | 18 | 10 | 6 | 7 | 2 | ,43 | |
| T2 (B) | 16 | 65 | 5 | 11 | 1 | 3 | 14 | 19 | 11 | 13 | 9 | 6 | ,43 | |
| T3 (B) | 13 | 62 | 8 | 14 | 0 | 4 | 12 | 19 | 15 | 14 | - | 5 | ,45 | |
| T4 (D) | 20 | 19 | 14 | 39 | 1 | 7 | 15 | 15 | 15 | 20 | 15 | 7 | ,39 | |
| T5 (C) | 21 | 23 | 38 | 1 | 1 | 8 | 15 | 16 | 20 | 15 | - | 7 | ,38 | |
| T6 (A) | 43 | 21 | 3 | 23 | 1 | 9 | 20 | 16 | 11 | 15 | - | 8 | ,42 | |
| T7 (C) | 9 | 27 | 30 | 26 | 0 | 1 | 16 | 18 | 20 | 15 | 13 | 5 | ,31 | a |
| T8 (B) | 19 | 63 | 3 | 7 | 1 | 8 | 12 | 20 | 15 | 13 | - | 5 | ,60 | |
| T9 (C) | 18 | 11 | 33 | 26 | 0 | 12 | 17 | 13 | 20 | 17 | - | 7 | ,39 | |
| T10 (B) | 10 | 40 | 26 | 12 | 1 | 12 | 14 | 19 | 18 | 15 | 14 | 7 | ,34 | a |
| T11 (A) | 38 | 10 | 20 | 16 | 1 | 15 | 20 | 14 | 17 | 16 | 7 | 8 | ,45 | |
| T12 (D) | 6 | 4 | 20 | 57 | 1 | 13 | 13 | 11 | 15 | 20 | 17 | 7 | ,53 | |
| T13 (A) | 44 | 17 | 3 | 22 | 1 | 12 | 19 | 17 | 16 | 17 | 15 | 7 | ,29 | b |
| T14 (B) | 8 | 37 | 18 | 18 | 4 | 16 | 14 | 19 | 19 | 17 | 18 | 8 | ,27 | a,b |
| T15 (B) | 20 | 32 | 19 | 13 | 1 | 15 | 15 | 21 | 17 | 17 | 15 | 8 | ,44 | |
| T16 (B) | 19 | 29 | 5 | 31 | 1 | 15 | 15 | 20 | 16 | 18 | 11 | 8 | ,34 | a |

a. Et svaralternativ velges av for flinke elever.
b. Svak diskriminering (<0,30)

Av resultatene i tabellen ser vi at det er få problemer knyttet til hvordan oppgavene har fungert. Vi vil likevel peke på noen få trekk:

- Diskrimineringen er gjennomgående god, men likevel ikke veldig god i siste del. Dette kan henge sammen med at noen oppgaver inneholder distraktorer med elementer av noe riktig, så det ikke alltid er enkelt å skille ”riktig” fra ”galt”.
- Oppgavene har gjennomgående litt høy vanskelighetsgrad. Særlig den siste delen har mange oppgaver med lav andel riktige svar.
- Det er en tendens til økende antall blanke svar på slutten av hver tekst. Dette kan tyde på at de har fått for liten tid, men denne effekten er ikke sterk.

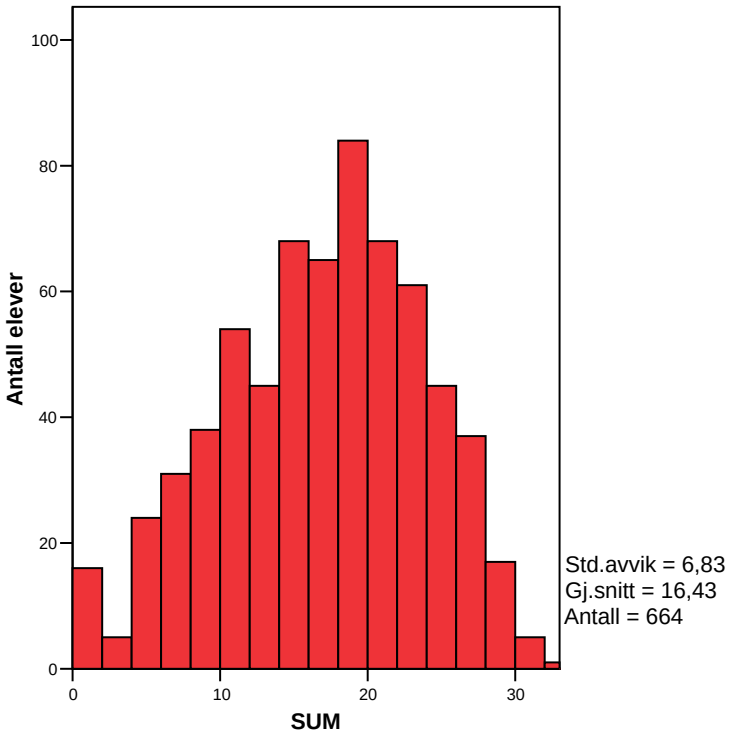
### 5.5.4 Resultater for hver del og samlet

*Tabell 5.9: Data for hver av de foreslåtte delene av prøven i 4. klasse*

| Kategori | Antall oppgaver | Gjennomsn. korrelasjon | Reliabilitet | Gjennomsn. skåre |
|---|---|---|---|---|
| Ordkjede | 1 | - | - | 23 |
| Tekst 1 | 16 | 0,24 | 0,84 | 56% |
| Tekst 2 | 16 | 0,14 | 0,73 | 46% |
| Totalt | 32 | 0,18 | 0,88 | 51% |

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 5.5.

*Figur 5.5: Histogram som viser fordelingen av poeng på prøven som helhet*




Våre kommentarer til disse resultatene er:

- Prøven har høy vanskelighetsgrad, og vi savner et sterkere innslag av lettere oppgaver. Del 1 har en gjennomsnittlig andel riktige svar på 57 % og del 2 bare 46 %. Totalt er det 51,4 % riktige svar. Det hadde vært en fordel med betydelig høyere andel, for eksempel omtrent 60 %, noe som kunne ha økt motivasjonen til å gi flere svar.
- Lærerne mener også at prøven er for vanskelig. Så mye som 83 % av dem mener at den er ”noe for” eller ”altfor” for vanskelig, og da er det et åpenbart problem.
- Vi konstaterer igjen at prøven ikke inneholder noen åpne oppgaver, et format som kanskje med hell kunne vært prøvd ut, for eksempel slik at elevene skriver ett eller noen veldig få ord som svar på spørsmålet.

- 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 siste delen er veldig lav, noe som har gitt altfor lav reliabilitet for denne delen. Dette kan henge sammen med de mange blanke svarene på slutten. Mens det for del 1 er 15 oppgaver som diskriminerer bedre enn 0,40, er det i del 2 bare åtte.
- Det har vært kritisert i pressen, og vi er enig i en slik kritikk, når det gjelder at flere oppgaver har distraktorer som et langt stykke på vei er "riktig". Særlig for så unge elever mener vi at distraktorene bør klart peke seg ut som uriktige.
- En annen sak er at korrelasjonen mellom de to delene (0,69) er nettopp så stor som den kan bli, gitt reliabiliteten til hver del. Ut fra en slik perfekt latent korrelasjon må vi konstatere at det er umulig å finne empirisk belegg for at det er to forskjellige kompetanser som måles. Det er altså ikke hold i å snakke om at de to tekstene måler ulike kompetanser.
- Ordkjedeprøven står i et uklart forhold til resten av prøven, og den korrelerer svakt (0,47 og 0,43) med de to tekstene. Det har etter sigende (fra noen eksperter) på noen skoler blitt gitt uklare beskjeder om hvor lang tid elevene har til disposisjon. Mye tyder på at denne delen kan fungere som en slags forklaringsvariabel for den målte leseforståelsen, men at den ikke vil fungere som en kompetanse som kan rapporteres på egne bein.
- Konklusjonen for 2004 gir seg selv: Det eneste fornuftige er å rapportere bare én generell kompetanse, leseforståelse. Fordelingen av poeng for en slik skala er ideell psykometrisk sett, men særlig av pedagogiske grunner burde den likevel vært noe lettere.

### 5.5.5 Konklusjon og anbefalinger for neste år

Prøven har fungert rimelig bra teknisk sett, men det er bare aktuelt å rapportere resultatene etter én skala. Som for 10. klasse foreslår vi å regne om de innrapporterte poengsummene ("Sum" av de to delene) til en standardisert skåreverdi. Dette gjøres ved å basere seg på gjennomsnitt (16,4) og standardavvik (6,83) til fordelingen. Formelen blir da:

**T = (Sum – 16,4)*10/6,83 + 50**

Med en slik transformasjon blir fordelingen i figur 5.5 uforandret, selv om verdien på x-aksen endres. Fordelen med en slik standardisert skåre er at resultatet i lesing kan sammenliknes direkte med matematikkresultatet, idet de to resultatene er målt med samme normrelaterte skala. Det gir da mening å si at en elev er "et halvt standardavvik bedre" i lesing enn i matematikk. Det er det nærmeste vi kan komme en "profil" for elevenes kompetanse for de nasjonale prøvene i 2004.

Det er et pedagogisk (men ikke et teknisk) problem at prøven har vært for vanskelig. Prøven har framkalt en del negative holdninger blant lærerne. Men vi må ikke underslå at dette også kan avspeile en tendens hos lærerne til å kreve forholdsvis lite av elevene. De nokså svake norske resultatene på internasjonale prøver gir en pekepinn om at det kan være viktig å beholde forholdsvis høye krav, særlig hvis ambisjonene for årene framover er en betydelig kompetanseheving på dette området. Vi konstaterer at dette spørsmålet bør løftes fram for diskusjon i et bredere forum.

Vi foreslår ut fra det foregående at det for årene framover lages en noe lettere prøve, i hvert fall med over 60 % riktige svar. Videre foreslår vi at det ved å systematisk velge ut oppgaver med høy diskriminering 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 "sjangerkompetanse" på basis av bare én tekst.

Som beskrevet for leseprøven i 10. klasse vil vi foreslå at det utarbeides et grundig rammeverk for nasjonale prøvene i lesing på alle de aktuelle trinnene.

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 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.

Det må vurderes om en ordkjedetest er fornuftig bruk av den hardt tiltrengte tiden for elevene under prøven og for lærerne ved vurdering.

## *5.6 Engelsk 10. klasse*

### 5.6.1 Struktur, validitet og vurderingskriterier

Når det gjelder engelsk, så har vi her bare vurdert skriveprøven. Det er også gjennomført en leseprøve på data, som vi altså ikke kan si noe om. Skriveprøven består av tre "oppgaver" av litt forskjellig karakter, men har det til felles at de krever at elevene skriver noen avsnitt ur fra en gitt kontekst. Imidlertid er sjangeren svært forskjellig for de tre oppgavene, noe som gjør at de stiller nokså ulike krav til elevene. Vi stiller oss generelt litt uforstående til at de tre oppgavene ikke er vurdert hver for seg, med kriterier tilpasset oppgaven. Spesielt vil vi peke på at oppgave 1 ber elevene skrive en e-mail til noen venner, noe som synes å være en sjanger som krever helt spesielle vurderingskriterier.

Den helhetlige tilnærmingen det er lagt opp til i vurderingen, tillater oss ikke å gjøre noen detaljerte analyser oppgave for oppgave. Vi kan bare studere fordelingen på nivåer og hvor godt samsvaret mellom retterne har vært.

Faggruppa har tatt utgangspunkt i nivåene som er beskrevet i Common European Framework of Reference for languages: Learning, teaching and assessment (CEF).

Disse nivåene gjelder samlet for lesing av engelsk tekst. Faggruppa har videre delt den samlete kompetansen inn i tre delkompetanser: Content, Text structure og Language. Denne tredelingen er innført for å stemme overens med kriteriene for vurdering til eksamen i 10 klasse. For hver av de tre delkompetansene har de så tilpasset en versjon av CEF med nivåer fra A1 til C2 for hver kompetanse, se de aktuelle nivåene det er referert til i veiledningen for prøven.

Ambisjonene med disse nivåene er altså å kunne vurdere elevenes kompetanse på hver av de tre områdene etter kriteriebaserte (absolutte) skalaer. Dataene vil kunne si noe om i hvor stor grad dette lykkes. Vi konstaterer at faggruppa har gjort et ambisiøst grep med å innføre et system av kriterier for vurdering som verken lærere eller "eksperter" kjenner fra før. I 10. klasse kunne man jo tenkt seg at man tok utgangspunkt i karakterskalaen, der det allerede langt på vei gjennom eksamen er etablert et tolkningsfellesskap for vurdering av elevbesvarelser.

Når det for øvrig gjelder prøvens innholdsvaliditet i forhold til læreplanen, tyder alt på at denne er høy. I Gallup-undersøkelsen har lærerne gitt overveiende god tilslutning til at prøven i "noen" eller "svært stor" grad reflekterer læreplanens mål og at elevene fikk vist sine ferdigheter. Også når det gjelder prøvens vanskelighetsgrad, er lærerne positive til prøven. En overveiende andel (84 %) mener at prøven er "passende" vanskelig. På mange måter har denne prøven fungert utmerket.

### 5.6.2 Resultater fra ekspertenes vurdering

Tabell 5.11 viser svarfordelingen i % når det gjelder de gitte nivåene. Vi har her for hver delkompetanse brukt betegnelsene på nivåene som framgår av tabell 5.10.

*Tabell 5.10: Nivåer for vurdering av engelskprøven*

A1 = 1
A1/A2 = 2
A2 = 3
A2/B1 = 4
B1 = 5
B1/B2 = 6
B2 = 7
B2/C1 = 8

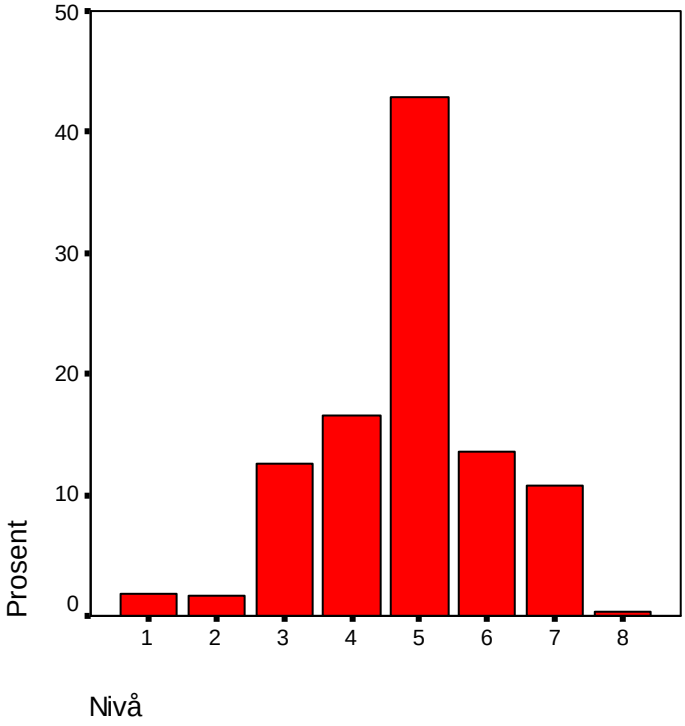
*Tabell 5.11: Svarfordeling på nivåer (avrundet til hele tall) for hver skala*

| Kategori | Svarfordeling i % | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** |
| Innhold | 2 | 2 | 12 | 16 | 42 | 13 | 11 | 0 |
| Tekststruktur | 2 | 2 | 12 | 18 | 44 | 11 | 9 | 0 |
| Språk | 2 | 2 | 12 | 18 | 41 | 16 | 7 | 1 |
| Totalt | 2 | 2 | 11 | 18 | 42 | 13 | 9 | 0 |

Fordelingen på nivåer viser som ventet en god normalfordeling, og de aller fleste elevene ligger som forutsagt på nivåer fra A2 til B2. Gjennomsnittet er 4,8, og standardavviket er 1,29 for alle skalaene. Fordelingen er vist grafisk på figur 5.4

nedenfor, og den er så godt som identisk for alle skalaene. Så langt ser altså resultatene meget bra ut.

*Figur 5.4: Fordeling på nivåer for den totale skalaen (omtrent den samme for alle)*




Det er imidlertid et påfallende trekk at fordelingene på nivåer er så forbløffende like for de tre skalaene. Dette henger sammen med at elevene stort sett har fått samme nivå på hver av dem. Utregning som er gitt i tabell 5.12 viser at korrelasjonen mellom dem er omtrent 0,90 mellom to og to skalaer. Foreløpig konstaterer vi at det synes å være påfallende liten forskjell mellom de tre delkompetansene, slik de er målt her.

*Tabell 5.12: Korrelasjon mellom de tre kriteriene*

| | Innhold | Tekststruktur |
|---|---|---|
| Tekststruktur | 0,88 | |
| Språk | 0,89 | 0,90 |

### 5.6.3 Samsvar i rettingen

Tabell 5.13 viser samsvaret mellom lærernes (skolenes) og ekspertenes uavhengige vurdering. I tabellen har vi angitt antall elever for hver bestemte differanse mellom skolens og ekspertens vurdering.

*Tabell 5.13: Prosentvis samsvar mellom skolenes og ekspertenes vurderinger*

| | Differanse | Innhold | Tekststruktur | Språk | Totalt |
|---|---|---|---|---|---|
| **Skolene** | 4 | 1 | 2 | 2 | 2 |

| | | | | | |
|---|---|---|---|---|---|
| **høyere** | 3 | 2 | 2 | 3 | 2 |
| | 2 | 16 | 13 | 12 | 13 |
| | 1 | 23 | 26 | 26 | 25 |
| | 0 | 42 | 41 | 39 | 43 |
| **Skolene lavere** | - 1 | 11 | 12 | 13 | 11 |
| | - 2 | 5 | 4 | 5 | 3 |
| | - 3 | 1 | 1 | 0 | 1 |

Det framgår av tabell 5.13 at det bare er rundt 40 % av vurderingene som har vært like for de to sensorene, og i rundt 20 % av tilfellene er avviket 2 nivåer eller mer. Vi har videre konstatert et ikke helt ubetydelig antall besvarelser der avviket er 3 eller til og med 4 nivåer. Vi må dessverre konstatere at sensorreliabiliteten rett og slett er for lav og langt under det som må kreves ved en nasjonal prøve som skal gi pålitelige mål for enkeltelevers (og skolers) kompetansenivå. Vi ser også at det er en betydelig forskyvning mot at skolene vurderer sine egne elever høyere enn ekspertene gjør.

Vi har også studert hvordan dette samsvaret har vært for hver skole, og dette har understreket problemene med vurderingene. **Noen skoler har systematisk vurdert sine elever mer enn et helt nivå høyere enn eksperten**, mens det på noen få skoler er en motsatt (og noe svakere) effekt. Så mye som 17 av de 24 skolene har vurdert sine elever høyere enn ekspertene har gjort. En måte å illustrere dette på har vi vist i tabell 5.14. I tabellen har vi gitt gjennomsnittlig kompetansenivå for skoler slik de ble vurdert av henholdsvis skolen selv og av eksperten. Vi har her bare tatt med skoler der vi har data fra minst 10 elever. For hver av de to vurderingene har vi også tatt med hvilket nummer skolen ville komme i en tenkt rangering av disse skolene etter prestasjoner. Som vi ser, 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.

*Tabell 5.14: Skolers gjennomsnittlige kompetansenivå og rangering sammenliknet mellom skolenes egen og ekspertenes vurdering. Hver linje representerer en skole, og skoler med færre enn 10 besvarelser er tatt ut.*

| **Kompetanse vurdert av skolen selv** | | **Kompetanse vurdert av ekspert** | |
|---|---|---|---|
| **Nivå** | **Rangering** | **Nivå** | **Rangering** |
| 6,50 | 1 | 6,00 | 1 |
| 6,30 | 2 | 4,70 | 12 |
| 6,11 | 3 | 4,84 | 10 |
| 5,94 | 4 | 5,06 | 5 |
| 5,70 | 5 | 4,85 | 9 |
| 5,67 | 6 | 4,72 | 11 |
| 5,63 | 7 | 5,50 | 2 |
| 5,62 | 8 | 5,08 | 4 |
| 5,45 | 9 | 5,05 | 6 |
| 5,40 | 10 | 5,10 | 3 |
| 5,37 | 11 | 3,89 | 19 |
| 4,86 | 12 | 4,50 | 14 |
| 4,85 | 13 | 3,80 | 20 |
| 4,84 | 14 | 3,95 | 18 |
| 4,80 | 15 | 4,70 | 13 |
| 4,75 | 16 | 4,95 | 8 |
| 4,25 | 17 | 4,95 | 7 |

| | | | |
|---|---|---|---|
| 4,10 | 18 | 4,30 | 16 |
| 4,06 | 19 | 4,44 | 15 |
| 4,00 | 20 | 4,11 | 17 |

Korrelasjonen mellom de to sensorene for hele prøven er 0,66 og ubetydelig lavere for de tre delkompetansene (0,65, 0,65, 0,64). Disse korrelasjonene kan vi oppfatte på samme måte som reliabiliteten til de andre prøvene. Og vi konstaterer også på denne måten at med så lav reliabilitet vil det være helt urimelig å publisere resultatene.

### 5.6.4 Konklusjon

Engelskprøven synes å ha høy validitet hva gjelder de kompetansene den tar mål av seg til å måle. Prøvens validitet synes også å være høy, vurdert fra lærernes side.

Sensorenes vurdering av besvarelsene til denne prøven tilfredsstiller dessverre langt fra de krav man må sette til en rimelig sensorreliabilitet. Den store forskjellen det synes å være mellom skolene når det gjelder bruk av vurderingskriteriene, er også foruroligende. Noen skoler ser ut til systematisk å vurdere sine elever høyere enn ekspartene har gjort. Vi vurderer dette som så alvorlig at vi innstendig vil fraråde å rapportere noe fra denne prøven. Å sammenlikne skoler med et slikt instrument vil være urimelig. Vi vil også fraråde skolene å anse resultatene som et objektivt mål for enkeltelevers eller skolers kompetanse når det gjelder skriving på engelsk.

Det er sterkt å beklage at resultatene er blitt slik, men langt på vei er dette som forventet. Mange undersøkelser har vist at det er svært vanskelig å oppnå høy sensorreliabilitet for fri skriving, uansett hvilket språk det gjelder. Dette er nokså rimelig, idet det naturlig nok er vanskelig å enes om hva en god tekst er. Problemet synes derfor å være en uklarhet når det gjelder oppdragets art. Faggruppa har gjort et stort og prisverdig utviklingsarbeid med å tilpasse CEF for norske forhold. Og de har hatt et svært ambisiøst utgangspunkt med å basere vurderingskriteriene på CEF, som er ganske ukjent i norsk skolesammenheng. Og når man så i tillegg har utvidet dette til nivåer for tre delkompetanser, er vanskelighetene kanskje blitt ytterligere forsterket. Sett på bakgrunn av dette er det kanskje heller grunn til å si at faggruppa har klart å få til en rimelig høy sensorreliabilitet. Problemet er altså at den likevel langt fra er høy nok.

Det er vår klare forståelse når det gjelder vurdering av elevtekster, enten de er på norsk eller engelsk, at en noenlunde pålitelig vurdering er helt avhengig av at det på forhånd er etablert et tydelig tolkningsfellesskap mellom sensorene. Til en viss grad er dette gjort i 10. klasse gjennom flere år i forbindelse med skriftlig eksamen, men da med den uttrykkelige forutsetning at vurderingen foregår med bruk av de tradisjonelle karakterene og at to uavhengige sensorer diskuterer eventuell uenighet seg imellom.

Den skriftlige engelskprøven, slik den framstår med sin tilknytning til CEF-nivåene, kan være et viktig bidrag til å knytte norske elevers kompetanse til internasjonale kompetansebeskrivelser. Og i så måte har den en funksjon gjennom den diskusjon som den kan reise blant lærere og fagdidaktikere. Men vi synes ikke det er forsvarlig å gjennomføre en nasjonal prøve i full skala når resultatene ikke med rimelighet kan brukes som objektivt mål på kompetanse. Et ufravikelig krav burde være at før en slik prøve blir sendt ut til hele elevmassen, bør det være påvist under utprøving at sensorreliabiliteten er god.

Uten at vi vil framstå med høy kompetanse i engelsk fagdidaktikk vil vi likevel peke på en alternativ strategi, som trolig vil gi høyere sensorreliabilitet, kan være å øke antall oppgaver og vurdere hver oppgave for seg, med vurderingskriterier spesielt tilpasset hver oppgave. Videre, ved å bruke en poengskala for hver oppgave kan hver elev tilordnes en samlet poengsum. Denne poengsummen kan så i sin tur tilordnes både en karakter og et CEF-nivå. Dette kunne være en mer forsiktig og kanskje noe enklere kopling av CEF til de nasjonale prøvene.

## *5.7 Faktoranalyse*

For å undersøke nærmere strukturen av de målte resultatene har vi foretatt en såkalt "principal component" eksplorerende faktor analyse hvor hensikten var å se om prøven naturlig delte seg i de skalaer som opprinnelig var konstruert. Hver analyse ble foretatt to ganger, første gang uten premisser på hvor mange skalaer prøven skulle inneholde, og den andre gangen med det antall skalaer som var foreslått. Her rapporteres begge forsøkene for lesing 4. klasse, lesing 10. klasse og matematikk i 10 klasse. Nokså strenge kriterier ble brukt for analysen, hvor det ble brukt en såkalt "direct oblimin" løsning, som tillater korrelasjon mellom faktorer, og bare egenverdier over 0,35 ble godtatt. Dette kan naturligvis gjøres på andre måter, men for denne analyses hensikt synes dette å være rimelig.

*Tabell 5.15: Resultater av faktoranalysene*

| | | 1.kjøring | | 2.kjøring | | |
|---|---|---|---|---|---|---|
| **Fag** | **Antall foreslåtte skalaer** | **% forklart** | **Antall faktorer** | **% forklart** | **Antall faktorer** | **% forklart 1. faktor** |
| Lesing – 4 | 2 | 47,1 | 8 | 26,3 | 2 | 21,6 |
| Lesing – 10 | 3 | 51,6 | 12 | 29,6 | 3 | 21,1 |
| Matematikk 10 | 5 | 69,5 | 22 | 40,6 | 5 | 27,8 |

Som man kan se i tabell 5.15, har ingen av prøvene klart god støtte for de skalaer som ble brukt. De ser alle sammen ut til å ha en overveiende faktor (henholdsvis i lesing og i matematikk), hvor alle andre delene er veldig høyt korrelert med den underliggende faktoren. Det er derfor neppe grunnlag for å rapportere mer enn én skala fra hver prøve.

Det er ikke lett å tolke dette resultatet, spesielt når man tar hensyn til at oppgavene har forskjellig vanskelighetsgrad. Det er mulig at en del av de faktorer som framkommer, skyldes rett og slett at det naturlig blir en ganske høy korrelasjon mellom oppgaver av samme vanskelighetsgrad. Man kan konstatere at i alle tilfeller forklarer den ubegrensede løsning ("1. kjøring") en stor del av totalvariansen, noe høyere for matematikk enn for de andre prøvene. Den første faktoren har i alle tilfeller en så stor andel av totalvariansen at også ifølge denne analysen synes det å være åpenbart at prøvene er mer "endimensjonale" enn opprinnelig forutsatt. Dette er et ytterligere argument for at man bør rapportere færre skalaer enn opprinnelig planlagt.

# 6 Oppsummering og konklusjoner

## *6.1 Oppsummering*

En oppsummering av de viktigste funnene i vår undersøkelse er gitt nedenfor.

**Begge matematikkprøvene** (10. og 4.klasse) er laget etter et meget ambisiøst opplegg, med svært kompliserte vurderingskriterier og med intensjoner om å rapportere mange delkompetanser. Prøvene inneholder mange svært gode oppgaver som hver for seg kan gi viktig diagnostisk informasjon om enkeltelevers forståelse av grunnleggende matematiske begreper. I det hele tatt har faggruppa lagt ned et stort arbeid med betydelig originalitet. Som forskningsprosjekt er dette uhyre interessant, men det har dessverre vist seg å være altfor vidløftig for det aktuelle formålet. Ikke minst har det ført til en etter vår mening uforholdsmessig stor byrde for lærerne. Dette er svært uheldig, ikke minst i lys av at våre analyser viser at det ikke er empirisk hold i de foreslåtte kategoriene. Alle disse kategoriene har ikke god nok reliabilitet til å kunne rapporteres, og de er heller ikke tilstrekkelig forskjellige til at de hver for seg innehar verdifull informasjon. Vi fraråder derfor sterkt å publisere resultater som planlagt. I stedet har vi anvist hvordan man for hvert av de to klassetrinnene kan kombinere delresultatene til et generelt resultat for matematikkompetanse. Denne skalaen har høy reliabilitet og validitet, og den kan derfor gi en meningsfull verdi både for enkeltelever og for skoler. Vi anbefaler videre at man vurderer en normbasert rapportering ved å standardisere resultatene til for eksempel et gjennomsnitt på 50 poeng og et standardavvik på 10 poeng (T-skåre).

Prøven i10. klasse har etter vår (og lærernes) mening vært noe for vanskelig. Vi stiller oss imidlertid nokså uforstående til at et flertall av lærere har kritisert prøven i 4. klasse for det samme. Snarere vil vi hevde at det kanskje forteller noe om svake forventninger blant lærerne. Vanskelighetsgraden i de nasjonale prøvene på alle trinn (og for 4. trinn spesielt) bør nøye diskuteres i lys av ambisjoner om høyere elevkompetanse, slik disse kommer til uttrykk i St. meld. nr 30.

Det er videre vår sterke anbefaling at man i det videre arbeid i matematikk forlater det ambisiøse systemet det hittil er lagt opp til. Vi mener det bør lages prøver som er mye enklere å vurdere, blant annet ved å inkludere et betydelig antall flervalgsoppgaver, og som i større grad ivaretar grunnleggende testteoretiske prinsipper. Vi anbefaler en grunnleggende nytenkning når det gjelder matematikkprøvens mål og mening.

**Leseprøven for 10. klasse** er laget med PISA-undersøkelsen som mal. Denne prøven har tydeligvis vært noe for lett, men ellers har den fungert stort sett bra. De aller fleste oppgavene har fungert godt teknisk sett, og det er overveiende god sensorreliabilitet. Imidlertid er det ikke et høyt nok antall oppgaver til å få høy reliabilitet for hver av de tre foreslåtte rapporteringskategoriene. Vi anbefaler derfor at de foreslåtte kategoriene slås sammen til én. Vi foreslår videre at denne skalaen standardiseres med samme gjennomsnitt og standardavvik som for matematikkprøven (f eks gjennomsnitt 50 og standardavvik 10).

Vi foreslår at faggruppa i det videre arbeidet prøver å klargjøre hvor mange og hvilke rapporteringskategorier man skal tilstrebe, og at oppgaver velges ut på en balansert

måte for å ivareta dette. Vi anbefaler også, av hensyn til lærernes arbeidsbyrde, å redusere noe på andelen av åpne oppgaver.

**Leseprøven for 4. klasse** er laget etter mønster av PIRLS-undersøkelsen, og den inneholder bare flervalgsoppgaver. Denne prøven er blitt sterkt kritisert av lærerne for å være for vanskelig og å stemme dårlig overens med kravene i læreplanen. Til det har vi anført at dette bør diskuteres nøye i lys av de aktuelle ambisjoner om å heve elevenes kompetanse i leseferdighet, og ikke minst i lys av at det er vanskelig å se at de nåværende læreplanene i det hele tatt definerer noe nivå for hva elevene skal kunne. Det står veldig lite i L97 om hva som menes med leseferdighet ut over det helt elementære grunnlaget. Vi er for øvrig helt enig i at neste års prøve bør være lettere enn årets.

Vi har påvist at det ikke er hold i å rapportere mer enn én skala for leseferdighet, enten ved å rapportere antall riktige svar direkte, eller helst slik vi anbefaler, at denne skalaen standardiseres med samme gjennomsnitt og standardavvik som for lesing i 10. klasse og matematikkprøven (f eks gjennomsnitt 50 og standardavvik 10).

Faggruppa har foreslått for neste år å lage flervalgsoppgaver med gradert riktighet, altså at de fire svaralternativene blir tillagt 0, 1, 2 og 3 poeng istedenfor som vanlig 0, 0, 0 og 1 poeng. Vi har advart mot en slik framgangsmåte, både fordi det stiller uhyre sterke tekniske krav til hver oppgave for at den skal fungere tilfredsstillende, men også fordi det med slike oppgaver ville bli en fornuftig strategi å gjette blindt der en elev ikke vet hva som er riktig svar.

Vi har foreslått at det utarbeides et felles rammeverk for de nasjonale prøvene i lesing, der design og fordeling av oppgaver og tekster begrunnes for de enkelte klassetrinn. Dette vil være særlig viktig i lesing, siden det ikke finnes noen egentlig læreplan i dette ”faget”.

**Engelskprøven for 10. klasse** består av tre ”essay-oppgaver” som skal vurderes ut fra en helhetlig vurdering, altså ikke oppgave for oppgave. Faggruppa har gjort et stort arbeid med å tillempe det internasjonale Common European Framework of Reference for languages: Learning, teaching and assessment (CEF) for bruk ved vurdering av denne prøven. Våre analyser av sensorreliabiliteten viste dessverre at denne er så lav som korrelasjon 0,66 mellom uavhengige vurdering av samme besvarelser. Det kan hevdes at dette er et lovende resultat ut fra at det tar flere år å oppnå et tolkningsfellesskap og at dette er det beste vi kan håpe på. Men vi har likevel innstendig advart mot å rapportere resultater med så dårlig reliabilitet. I tillegg kommer en tydelig tendens til at noen skoler i sterk grad vurderer sine elever systematisk høyere enn ekspertene gjør.

Vår konklusjon er altså at ingen resultater bør rapporteres fra denne prøven.

## *6.2 Konsekvenser for det videre arbeidet*

Som en avslutning på denne rapporten vil vi her gi noen anbefalinger om det videre arbeidet med de nasjonale prøvene, slik de framstår i et litt mer overordnet perspektiv. Kravspesifikasjonen til vår undersøkelse hadde disse punktene:

- *Hvilke konkrete indikasjoner finner vi på at oppgaveutviklingen neste år bør foregå grundigere og mer målrettet?*
- *Hva slags opplæring og kravspesifikasjoner overfor fagmiljøene er nødvendig for at dette skal kunne skje?*
- *Hva har vi lært om prosedyrer for gjennomføring, retting og kontroll?*

Vi har konstatert at årets nasjonale prøver på flere måter har fungert lite tilfredsstillende, og vi har av den grunn advart mot å rapportere resultatene slik det var planlagt. De nasjonale prøvene representerer noe nytt i norsk skole. Det er mye som skal til for at et slikt omfattende nasjonalt tiltak skal lykkes, og det er naturlig at det tar tid å bygge opp et system for dette. Når vi avslutningsvis vil peke på noen åpenbare områder for forbedring og videre utvikling, er det ikke fordi vi har noe ønske om å fordele skyld for det som ikke har lykkes så bra, men fordi vi håper å kunne bidra til utviklingen på en konstruktiv måte.

Vi har i våre analyser påvist at det særlig er på den testteoretiske siden det er behov for opprustning. Sammenliknet med andre land er det en forbløffende mangel på både grunnleggende og avansert psykometrisk kompetanse, og dette har gjenspeilet seg i forbindelse med årets prøver. Vi har en lang tradisjon på å lage gode oppgaver til eksamen og andre prøver som har høy validitet i forhold til læreplanene. Men vi har ingen tilsvarende tradisjon for å interessere oss for i hvilken grad prøvenes utforming og dens bestanddeler tillater oss å stole på resultatene. Eller kortere sagt: Validitet har vært i høysetet, reliabilitet har man stort sett ikke snakket om. Problemet med dette er at høy validitet er lite verdt i praksis hvis den ikke kombineres med høy reliabilitet. Høy reliabilitet for de nasjonale prøvene, og da mener vi BÅDE i form av indre konsistens og i form av sensorreliabilitet, er en nødvendig forutsetning for å kunne gi et nokså presist mål for enkeltelevers kompetanse og for skolenes faglige nivå. Det er åpenbart at hvis man vil publisere skoleresultater, så er det viktig at forskjellen mellom skolene kan påvises å representere reelle forskjeller og ikke bare tilfeldigheter ved prøvens utforming eller ved vurderingen av besvarelsene.

Vi vil altså sterkt anbefale en kraftig kompetanseheving på det testteoretiske feltet. Her har vi mye å lære av andre land, der vi finner både universitets- og høyskoleinstitutter samt selvstendige institusjoner for pedagogiske målinger. Økt slik kompetanse er viktig for alle som arbeider med å utvikle prøvene. Men særlig er dette avgjørende for de som leder utviklingen av prøvene på nasjonalt nivå, særlig for å kunne få til en felles forståelse og tilnærming blant de ulike faggruppene.Uten god forståelse av grunnleggende testteoretiske prinsipper er det ikke mulig å vurdere hva man bør legge vekt på for å kunne oppnå et bestemt formål med prøvene. Og uten slik forståelse er det heller ikke mulig på forhånd å sette konkrete krav til prøvenes utforming og innhold.

I forlengelse av dette forslaget om kompetanseheving foreslår vi videre at det mer generelt tas et nasjonalt grep for å heve den testteoretiske innsikten blant lærerutdannere og fagdidaktikere. Økt kunnskap om vurdering og utvikling av gode oppgaver er også viktig for lærere. Økt forståelse på dette feltet mener vi også er viktig blant skolepolitikere og ansatte i Utdanningsdirektoratet og i departementet.

Dette første året har det tydeligvis vært en bevisst politikk å ”la de tusen blomster blomstre”, faggruppene har i stor grad selv kunnet bestemme utformingen av prøvene

både med hensyn til design, oppgaveformater, vurderingskriterier og rapporteringsskaler. Vi anbefaler en mye sterkere styring og koordinering på alle disse områdene. Med en mer enhetlig tilnærming kan man bedre kommunisere med skoleeiere, skoler, lærere og elever om formål og betydning av resultater.

En mer enhetlig tilnærming er særlig viktig når det gjelder utprøving av oppgaver. Det synes å herske stor usikkerhet om hva en god utprøving egentlig innebærer. Vi foreslår at det settes konkrete krav til både hvordan, når og med hvor mange elever utprøving bør foregå. Videre foreslår vi at det settes konkrete krav til enkeltoppgaver og prøven som helhet knyttet til vanskelighetsgrad, diskriminering og reliabilitet. Gjennom god utprøving og analyse av data derfra bør det kunne påvises at slike krav med god sannsynlighet vil bli oppfylt ved selve gjennomføringen. Vi mener det er avgjørende for de nasjonale prøvenes funksjon og omdømme at prøvene PÅ FORHÅND er påvist å tilfredsstille grunnleggende krav. I dette ligger det blant annet at det ikke legges opp til å rapportere fagkompetanse i form av "profiler" uten at hver delkompetanse på forhånd kan påvises å ha god validitet og reliabilitet. I denne sammenheng vil vi peke på at det bør engasjeres folk utenfor faggruppa selv til uavhengig vurdering av et sett besvarelser for å undersøke sensorreliabiliteten der dette er aktuelt. Et ofte brukt sitat blant folk som arbeider med vurdering i stor skala er: *It is not enough to be good, you must prove you're good!* Altså: Det må påvises at vesentlige kriterier for prøvene er oppfylt før de sendes ut til hele elevkullet.

Den foreslåtte strategien vil innebære at faggruppene før en slik systematisk "generalprøve" allerede har prøvd ut grupper av enkeltoppgaver med noen få elever slik at helt ubrukbare oppgaver allerede er luket ut. Likevel må utprøvingen gjennomføres med mange flere (gjerne så mange som det dobbelte antall) enkeltoppgaver enn det som er aktuelt i den endelige prøven for at man systematisk kan sette sammen en god prøve.

Det bør etter vår mening legges opp til en enhetlig plan for hvordan lærere og skoler kan nyttiggjøre seg resultater fra prøvene. Etter at prøven er gjennomført, er det viktig at lærerne med en gang får tydelig beskjed om hva som videre skal skje med vurdering og innsending av resultater, og ikke minst hvordan de kan og bør bruke resultater i pedagogisk sammenheng. Med dette siste tenker vi her på både vurdering og diagnostisering av enkeltelever samt tilbakemelding om på hvilke felter det synes å være rom for forbedring for enkeltelever, klasser og skoler.

Vi har flere steder pekt på behovet for et grundig faglig rammeverk for hvert av fagområdene. Et slikt rammeverk kunne inneholde et rasjonale for prøvene på ulike klassetrinn, samt en oversikt over design, fordeling av oppgaver etter oppgavetype, samt hvilke rapporteringskategorier som er tilstrebet. Et slikt dokument vil kunne være av betydelig verdi i en oppbygningsfase av de nasjonale prøvene. Og videre ville et slikt offisielt dokument klargjøre hvilket forhold det er mellom læreplanen og hvilke kompetanser som måles. En skriftlig prøve kan selvsagt ikke dekke alle kompetansene. Og videre er det vel slik at prøvene særlig skal konsentrere seg om det som Kvalitetsutvalget kalte "basiskometanse" og som stortingsmeldingen kalte "grunnleggende ferdigheter".

Dersom man virkelig mener 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. Det er fullt mulig, men vil kreve en systematisk og målrettet utvikling, for eksempel ved at et utvalg av elevene prøver ut deler av neste års oppgaver samtidig med årets prøver (uten å vite hva som er hva).