



Matematikksenteret
Nasjonalt senter for matematikk i opplæringen

TEKNISK RAPPORT

NASJONAL PRØVE I REGNING 5. TRINN 2018

**Morten Svorkmo
Grethe Ravlo
Monica Rehaug
Ole Harald Johansen**

MATEMATIKKSENTERET, NTNU 2019

Nasjonal prøve i regning for 5.trinn 2018

Morten Svorkmo, Grethe Ravlo, Monica Rehaug og Ole Harald Johansen

NASJONALT SENTER FOR MATEMATIKK I OPPLÆRINGEN (MATEMATIKKSENTERET)
NTNU mars 2019

Innhold

INNLEDNING	4
KAPITTEL 1: PREPILOT	5
KAPITTEL 2: 1. PILOT	6
KAPITTEL 3: 2. PILOT	9
KAPITTEL 4: SAMMENSETNING AV KOHORTPRØVE.....	11
INNLEDNING	11
PSYKOMETRI	11
VALIDITET	14
FORKLARING TIL NOEN OPPGAVER I ENDELIG PRØVE	16
<i>Oppgaver med lav diskriminering</i>	<i>16</i>
<i>Oppgaver der distraktorer har positiv point-biserial</i>	<i>17</i>
<i>Oppgaver med bias</i>	<i>18</i>
ANKEROPPGAVER	19
GENERELLE VURDERINGER AV PRØVEN SOM HELHET.....	20
KAPITTEL 5: ANALYSE AV NASJONAL PRØVE I REGNING 2018.....	21
PSYKOMETRISKE VERDIER	21
<i>Oversikt over oppgaver NP5 2018</i>	<i>22</i>
<i>Testinformasjon og måleusikkerhet.....</i>	<i>23</i>
TOLKNINGER AV RESULTATENE	23
<i>Resultater innen områdene tall, måling og geometri, og statistikk.....</i>	<i>23</i>
<i>Resultater til enkeltoppgaver.....</i>	<i>24</i>
<i>Sammenligning av resultater over tid.....</i>	<i>24</i>
KJØNNSFORSKJELLER	25
<i>Kjønnsforskjeller 2014–2018.....</i>	<i>27</i>
<i>Ubesvarte oppgaver.....</i>	<i>27</i>
TEKNISKE DATA	29
OPPGAVEANALYSE FOR ALLE OPPGAVENE.....	29
KILDER:	31

Innledning

Denne rapporten beskriver prosessen for utvikling, gjennomføring og analyse av resultatene til nasjonal prøve i regning for 5.trinn 2018. Det tok to år å utvikle prøven, og oppgavene har vært gjennom tre piloteringsrunder før de ble godkjent til å være med i den nasjonale prøven. De tre første kapitlene gir en kortfattet oversikt over piloteringene av oppgavene, og sier litt om formålet med og utfordringer i de ulike piloteringsfasene.

I det fjerde kapitlet er prosessen med å sette sammen oppgaver til en endelig prøve beskrevet. Vi har forsøkt å begrunne de vurderinger vi har gjort før det foreligger et forslag til en endelig prøve. Siste del av rapporten, kapittel 5, er en analyse av resultatene etter at den nasjonale prøven for 2018 er gjennomført.

Gjennom hele utviklingsprosessen brukes klassisk test teori (CTT) og Item Response Theory (IRT) til å analysere resultatene. De utvalgte skolene som deltar i piloteringene, er trukket av Utdanningsdirektoratet og tilfredsstiller kravene til tilfeldige og representative utvalg.

For de ulike kapitlene er de viktigste analyseresultatene gjengitt i denne rapporten.

Kapittel 1: Prepilot

Prepiloter for 2018-prøven ble gjennomført i februar 2017. Det var sju prepiloter med 30 oppgaver i hver, i alt 210 oppgaver. Dette var i hovedsak åpne oppgaver der elevene fylte inn et svar uten at de hadde svaralternativer å velge blant. Prepilotene ble gjennomført med elever på 5. trinn, og det var i overkant av 200 elever som deltok på hver pilot. Svarfrekvenser ble gjennomgått i de åpne oppgavene for å se fordelingen av elevsvar. Elevsvar fungerer som et utgangspunkt for å lage flervalgsalternativer til 1. piloter. For å finne gode svaralternativer til flersvarsoppgaver, forsøker vi alltid å finne ut hvilken tankegang som ligger bak de svarene elevene har gitt. To personer fra prøveutviklingsgruppa overvar utprøving i to grupper på en skole. Det er nyttig for oss å se hvilke oppgaver elever stiller spørsmål til, og ikke minst se arbeidsomfanget i enkeltoppgaver. De fleste elevene ser ut til å forstå hva de skal gjøre i interaktive oppgaver uten at det er nødvendig med mye instruksjon. Vi fikk også mulighet til å se hvordan elevene hadde resonnert i enkeltoppgaver ved at vi fikk tilgang til anonymiserte kladdemark etter gjennomføringen.

Selv om vi har forsøkt å fordele oppgaver både etter vanskelighetsgrad og innhold i prepilotene, er vi i første rekke interessert i resultater til enkeltoppgaver. Vi vet at både vanskelighetsgraden i oppgavesettet som helhet og oppgavens plassering i settet vil innvirke på resultatet. For å kunne sammenligne resultater mellom prepilotene, har vi derfor de siste årene tatt med åtte oppgaver som vi kaller prepilotanker. En av oppgavene er felles for alle sju prepilotene, mens de sju andre oppgavene er fordelt med en oppgave i hver prepilot. Alle prepilot-ankeroppgavene er prøvd i tidligere prepiloter og kan gi gode signaler om forskjellene i resultater mellom de ulike oppgavesettene. Disse oppgavene plasseres på samme sted i oppgavesettene fra år til år for å kunne sammenligne verdiene så godt det lar seg gjøre. Det var 1546 elever som deltok i prepilot til 2018. Svarprosenten er høy, fra 95 til 97 prosent. Se tabell 1 for utvalg og resultater.

Prepiloter NP5 2018	Prepilot A	Prepilot B	Prepilot C	Prepilot D	Prepilot E	Prepilot F	Prepilot G
Antall gutter	108	111	113	102	125	108	105
Antall jenter	119	120	96	101	108	123	107
Totalt antall elever	227	231	209	203	233	231	212
Gj.snitt Theta gutter	0,154	0,120	0,157	0,165	0,027	0,130	0,213
Gj.snitt Theta jenter	-0,004	-0,016	-0,082	-0,052	0,046	-0,036	-0,082
Gj.snitt Theta totalt	0,071	0,049	0,047	0,057	0,036	0,042	0,064
Sign. forskjell J - G (Theta)	Nei	Nei	Nei	Nei	Nei	Nei	Ja
Standardavvik Theta	0,933	0,934	0,946	0,936	0,913	0,938	0,945
SE til gjennomsnitt Theta	0,062	0,061	0,065	0,066	0,060	0,062	0,065
Gj.snitt P-verdi gutter	47,8 %	44,3 %	51,0 %	49,6 %	52,1 %	46,2 %	48,0 %
Gj.snitt P-verdi jenter	43,6 %	41,5 %	46,0 %	45,1 %	51,6 %	42,3 %	42,0 %
Gj.snitt P-verdi totalt	45,6 %	42,9 %	48,7 %	47,4 %	51,9 %	44,1 %	44,9 %
Svarprosent	95,4 %	95,9 %	96,0 %	95,5 %	96,6 %	96,2 %	95,4 %
Cronbachs alpha	0,889	0,876	0,901	0,907	0,889	0,880	0,884
Antall items	30	30	30	30	30	30	30
Antall MC	3	6	7	7	9	8	5
Antall IA	3	5	7	7	6	7	7

Tabell 1: Resultater fra prepilot NP5 2018

Eksakte tall for alle oppgavene i prepiloten finnes i vedlegg *TR NP5-2018* under fanen Prepiloter NP5-2018.

Kapittel 2: 1. pilot

1. pilot ble gjennomført i mai 2017. Det var til sammen fire prøvesett med 37 oppgaver i hver, i alt 148 oppgaver. 1. pilot gjennomføres med elever på 4. trinn. Antall oppgaver ble redusert fra 40 til 37 for at utprøvingen ikke skulle bli for omfattende, siden elevene er fire måneder for unge, sammenlignet med elevgruppen prøven er ment for. De aller fleste oppgavene i 1. pilot er flervalgsoppgaver med høyfrekvente gale elevsvar som distraktorer, basert på resultatene og analysene da oppgavene var åpne i prepilotene. Et feilsvar betegnes som høyfrekvent når svarfrekvensen er 4–5 prosent eller høyere. For at en oppgave skal kunne endres fra åpen til flervalg må i tillegg rett svar og distraktorene som velges, utgjøre en stor andel av alle elevsvarene til sammen.

For å finne distraktorer til flervalgsoppgavene, undersøker vi feiltenkningene som ligger bak svarene til elevene da oppgaven var åpen i prepilotene. Svaralternativene skal være mulige elevsvar. Etter utprøvingen, forteller analysen av resultatene hvilke oppgaver som kan være med videre som flervalgsoppgaver. Oppgaver som får lave diskrimineringsverdier fordi svaralternativene ikke fungerer etter hensikten, kan ikke overleve som flervalgsoppgaver. Vi må også ta i betraktning at flervalgsoppgaver kan oppleves lettere å svare på enn åpne oppgaver, slik at oppgaver særlig mot slutten av prøven, kan få noen tilfeldige svar. I hvilken grad dette påvirker resultatet, avsløres imidlertid av analysene våre. I 1. pilot som i prepilot er det i første rekke enkeltoppgaver vi ønsker å finne ut om fungerer, men denne gang gjelder det i hovedsak flervalgsoppgavene.

Figur 1 og 3, og tabell 2 viser arbeidsgangen med en åpen oppgave fra prepilot til den er blitt en flervalgsoppgave som skal prøves i 1. pilot. Figur 1 viser oppgaven som den var i prepilot.

En familie skal plante hekk i hagen.
De ønsker at hekken skal bli 1,5 m høy.
Hekkeplantene de har kjøpt er 30 cm høye og vokser omtrent 20 cm i året.

Omtrent hvor mange år tar det før hekken er 1,5 m høy?

Svar: år



Figur 1: Item 5154506 Alperips som åpen i prepilot

Tabell 2 viser høyfrekvente svar på oppgaven og vår antagelse om hva slags tenkning som kan ligge bak feilsvarene. I noen tilfeller kan elevenes kladdeark fra prepilot hjelpe oss i denne fasen i likhet med de observasjoner vi kan gjøre ved å være tilstede under utprøvinger.

Elevar	Prosentandel	Mulig strategi
6 år	34,8	Riktig svar
4 år	12,0	Tenker at de vokser 30 cm i året og får dermed 4 år i tillegg til starthøyden
5 år	8,6	Legger sammen 20 cm + 30 cm og finner deretter at den må vokse med 20 cm per år i 5 år for at den skal bli 1,5 m høy
3 år	7,3	Legger sammen 20 cm + 30 cm = 50 cm, og tenker at det da tar 3 år til den er 1,5 m
7 år	6,4	
2 år	3,4	
Andre svar	21,5	
Ubesvart	6,0	

Tabell 2 Høyfrekvente svar fra prepilot item 5154506 Alperips

Figur 2 viser oppgaven som flervalg slik den ble prøvd i 1. pilot. Svaralternativer blir gjerne satt i kronologisk rekkefølge, enten stigende eller synkende hvis det gjelder tallverdier.

En familie skal plante hekk i hagen.
De ønsker at hekken skal bli 1,5 m høy.
Hekkeplantene de har kjøpt er 30 cm høye og vokser omtrent 20 cm i året.

Omtrent hvor mange år tar det før hekken er 1,5 m høy?

☐ 3 år

☐ 4 år

☐ 5 år

☐ 6 år



Figur 2: Item 5154506 Alperips som flervalg i 1. pilot

Nyproduksjon av oppgaver mellom prepilot og 1. pilot er alltid nødvendig for å erstatte oppgaver som for eksempel ikke har gode nok psykometriske verdier til å overleve etter prepiloten. Vi må hele tiden være sikre på at det finnes et bredt utvalg av oppgaver å velge mellom for alle mestringsnivåer, både når det gjelder relevans for de ulike prosessene gjenkjenne og beskrive, bruke og bearbeide og reflektere og vurdere, og når det gjelder matematiske områder og relevante kontekster. Primært er det oppgaver fra prepilot som er endret fra åpne til flervalg som blir pilotert i 1. pilotene, men også nyproduserte åpne oppgaver og oppgaver der tekst, tabeller/diagram og tall er blitt endret etter prepiloten, inngår i oppgavesettene.

Det var i alt 2246 elever som deltok i 1. pilot. Tabell 3 viser utvalg og resultater.

1.piloter NP5 2018	1.pilot A	1.pilot B	1.pilot C	1.pilot D
Antall gutter	275	295	291	273
Antall jenter	267	289	271	285
Totalt antall elever	542	584	562	558
Gj.snitt Theta gutter	0,037	0,038	0,089	0,173
Gj.snitt Theta jenter	-0,029	-0,044	-0,093	-0,182
Gj.snitt Theta totalt	0,005	-0,003	0,001	-0,008
Sign. forskjell J - G (Theta) $p < 0,05$	Nei	Nei	Ja	Ja
Standardavvik Theta	0,881	0,913	0,938	0,912
SE til gjennomsnitt Theta	0,038	0,038	0,040	0,039
Gj.snitt P-verdi gutter	40,6 %	46,3 %	43,7 %	50,8 %
Gj.snitt P-verdi jenter	39,5 %	45,0 %	40,6 %	44,0 %
Gj.snitt P-verdi totalt	40,1 %	45,7 %	42,2 %	47,3 %
Svarprosent	97,4 %	98,2 %	96,5 %	96,8 %
Cronbachs alpha	0,792	0,835	0,860	0,834
Antall items	37	37	37	37
Antall MC	33	33	32	32
Antall IA	4	4	4	3

Tabell 3: Resultater 1. piloter NP5 2018.

Ytterligere tall for alle oppgavene fra 1. piloter finnes i *TR NP5-2018* under fanene 1. pilot A-D

Kapittel 3: 2. pilot

Siste utprøving, 2. pilot, ble gjennomført i september 2017 med to fullskala oppgavesett, hver med 45 oppgaver. Begge oppgavesettene er konstruert som om de er en nasjonal prøve. Det betyr at oppgaverekkefølge, fordeling på mestringsnivå, fordeling på matematiske hovedtema og andel flervalgsoppgaver er så lik en endelig prøve (kohortprøven) som mulig.

Aktuelle oppgaver til 2. pilotene er oppgaver som fungerte som åpne oppgaver eller flervalgsoppgaver i 1. pilotene, og oppgaver som fungerte som åpne oppgaver i prepilotene. For at en flervalgsoppgave skal være aktuell for 2. pilotene, må den ha minimum to fungerende distraktorer i tillegg til det riktige svaret. Dette innebærer at distraktorene helst må være benyttet av minimum 4–5 prosent av elevene, og at gjennomsnittlig dyktighet til de som svarer riktig, skal være klart høyere enn for de som velger et annet alternativ.

Selv om vi i denne siste utprøvingen ønsker at begge oppgavesettene skal ha validitet og reliabilitet som en nasjonal prøve, er det fortsatt en utprøving. Vi prøver ut dobbelt så mange oppgaver som det skal være i den endelige prøven, og vi kan derfor ta med oppgaver som etter 1. pilot, har diskriminering lavere enn kravene i Rammeverket for nasjonale prøver (Utdanningsdirektoratet, 2017) i form av $a\text{-verdi} < 0,5$. Det kan for eksempel være oppgaver der vi har endret litt på oppgavetekst, spørsmålsformulering eller gjort en annen korrigering og derfor ønsker å prøve oppgavene på nytt, eller det kan være oppgaver som har pedagogisk verdi da den noe lave diskrimineringsverdien kan ha en faglig forklaring.

Opgavesettene til 2. pilotene må ta hensyn til ankeret til den fremtidige kohortprøven. Vi må finne oppgaver som kan fungere som skyggeoppgaver, det vil si oppgaver så like ankeret som mulig når det gjelder format, vanskelighetsgrad, innhold og psykometriske verdier. Slik legger de 20 oppgavene i ankeret føringer for hvilke oppgaver som skal plasseres på tilsvarende oppgavenummer når 2. pilotene settes sammen.

2. pilotene for 2018 ble gjennomført av elever på 5. trinn høsten 2017. Dette skjedde omtrent en måned før de samme elevene skulle gjennomføre årets nasjonale prøve. Det er viktig at oppgavene blir prøvd på riktig aldersgruppe, og gjennomføringstidspunktet for siste utprøving må derfor ligge så nært som mulig opp til tidspunktet for gjennomføring av endelig prøve. En måned forskjell i tid, kan ha betydning høsten på 5. trinn.

Tabell 4 viser resultatene etter gjennomføring av prøvene i 2. pilot. Det var 970 elever som gjennomførte 2. pilot A og 985 som gjennomførte 2. pilot B. Reliabiliteten til prøvene var henholdsvis 0,906 og 0,917 (tabell 4). Verdiene tilfredsstiller kravene i Rammeverket for nasjonale prøver (Utdanningsdirektoratet, 2017).

2.piloter NP5 2018	2.pilot A	2.pilot B
Antall gutter	483	491
Antall jenter	487	494
Totalt antall elever	970	985
Gj.snitt Theta gutter	0,168	0,154
Gj.snitt Theta jenter	-0,167	-0,157
Gj.snitt Theta totalt	0,000	-0,002
Sign.forskjell J - G (Theta) $p < 0,05$	Ja	Ja
Standardavvik Theta	0,963	0,971
SE til gjennomsnitt Theta	0,031	0,031
Gj.snitt P-verdi gutter	49,2 %	51,2 %
Gj.snitt P-verdi jenter	42,5 %	44,1 %
Gj.snitt P-verdi totalt	45,8 %	47,6 %
Svarprosent	93,4 %	93,6 %
Cronbachs alpha	0,906	0,917
Antall items	45	45
Antall MC	32	31
Antall IA	9	8

Tabell 4: Resultater 2. piloter NP5 2018

Ytterligere tall for alle oppgavene fra 2. piloter finnes i *TR NP5-2018* under fanene 2. pilot A og B.

Kapittel 4: Sammensetning av kohortprøve

Innledning

Høsten 2017 deltok til sammen 1955 elever fordelt på to piloter til 2018-prøven. 2. pilot er den siste utprøvingen av oppgavene før den endelige gjennomføringen, og det er svarmønstrene til disse elevene som danner grunnlag for tallene som blir løftet fram i dette kapitlet. I tillegg til at 2. pilotene er analysert, er det kjørt en IRT-analyse av kun de oppgavene som utgjør den endelige prøven (kohortprøven). Tabeller eller diagram som henviser til disse, er kommentert med «tall fra preanalyse». Det vil si at den er basert på elevbesvarelser fra 2. pilot, men at det er foretatt en analyse av de 45 oppgavene som utgjør forslag til endelig prøve og ankerprøve (tabell 5).

Nasjonal prøve for 5. trinn 2019 består av 45 oppgaver. Prøven foreligger i fire versjoner. Alle versjonene inneholder 45 oppgaver. Oppgavene i tre av versjonene er de samme, men rekkefølgen til de ti første oppgavene er ulik i de tre versjonene. Det er gjort for å redusere sannsynligheten for at to elever som sitter nær hverandre, arbeider med samme oppgave samtidig. Versjonen som er ulik de andre tre, er en ankerprøve som brukes til å måle utvikling over tid. Omtrent halvparten av oppgavene i ankerprøven har vært med i ankerprøven tidligere og utgjør det hemmelige ankeret. Ankeret skiftes ut i løpet av en fem-årsperiode. Nye oppgaver må derfor piloteres i ankerprøven som «preanker». Det gjør at mer enn halvparten av oppgavene i ankerprøven må hemmeligholdes. De øvrige oppgavene i ankerprøven er de samme som i de tre andre versjonene. Ankerprøven er kommentert senere i kapitlet.

Psykometri

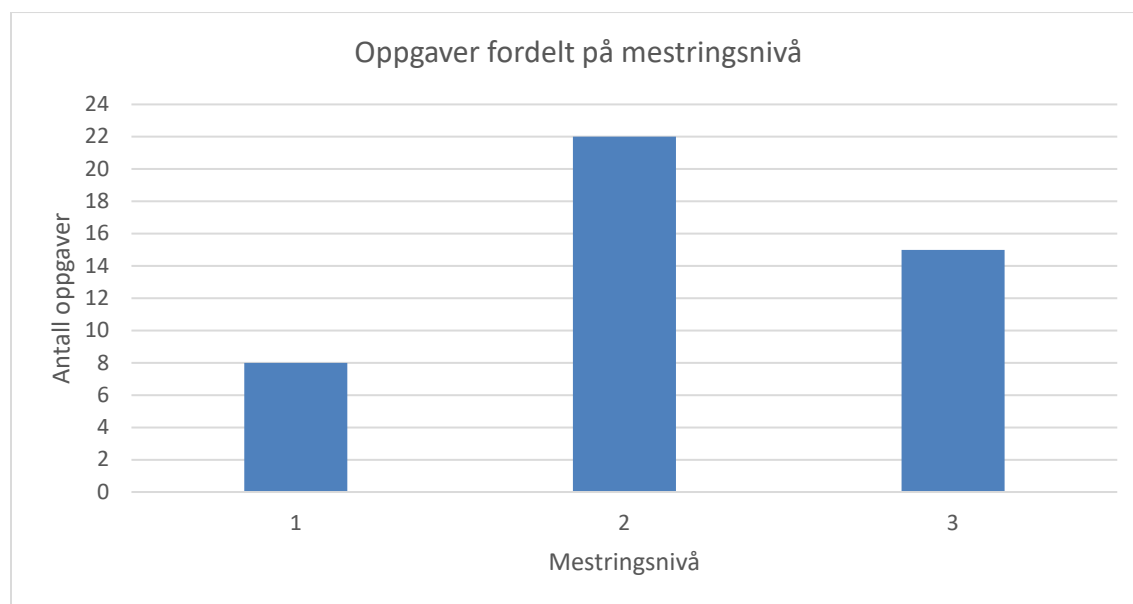
Tabell 5 viser sentrale verdier til prøven for 2018 ut fra en estimert preanalyse. Andelen flervalgsoppgaver i kohortprøven er 67 prosent, og det er totalt 20 prosent interaktive oppgaver. Gjennomsnittlig b-verdi til oppgavene er 0,130 og gjennomsnittlig a-verdi er 0,725. Gjennomsnittlig P-verdi er 47,6 prosent, og ved utprøving er P-verdien til guttene 6 prosentpoeng høyere enn jentenes. Det er åtte oppgaver i kohortprøven med bias mot kjønn, fem mot jenter og tre mot gutter, og det er som for tidligere år, grunn til å anta at det kan bli en signifikant forskjell mellom kjønnene i klassisk test-teori (CTT) ved endelig gjennomføring. Dette blir omtalt seinere i kapitlet.

Oppgavene er plukket ut fra to 2. piloter, og samtlige oppgaver har vært gjennom minst to utprøvinger. De fleste oppgavene har vært prøvd ut tre ganger. Siden settet er satt sammen av oppgaver fra to 2. piloter, som er besvart av to uavhengige elevgrupper, er det ikke mulig å beregne reliabilitet for den endelige prøven. Hver av de to 2. pilotene har reliabilitetskoeffisient, Chronbach alpha $> 0,90$. Basert på tidligere erfaringer med å sette sammen prøver etter denne modellen, kan vi anta at reliabiliteten til den endelige prøven (kohortprøven) kommer til å bli i overkant av 0,90. Dette tilfredsstiller kravet i Rammeverket for nasjonale prøver (Utdanningsdirektoratet, 2017b).

Forslag til endelig prøve 2018	Kohortprøve	Ankerprøve ¹ NP5 - 2018
Gj.snitt P-verdi	47,6 %	49,2 %
Gj.snitt P-verdi jenter	44,4 %	46,1 %
Gj.snitt P-verdi gutter	50,7 %	52,3 %
Gj.snitt T-Rpbis	0,457	0,458
Gj.snitt a-verdi	0,725	0,716
Gj.snitt b-verdi	0,130	0,043
Antall item	45	45
Antall MC	30	29
Antall IA	9	10

Tabell 5: Psykometriske verdier i forslag til endelig prøve og ankerprøve. Tall fra preanalyse

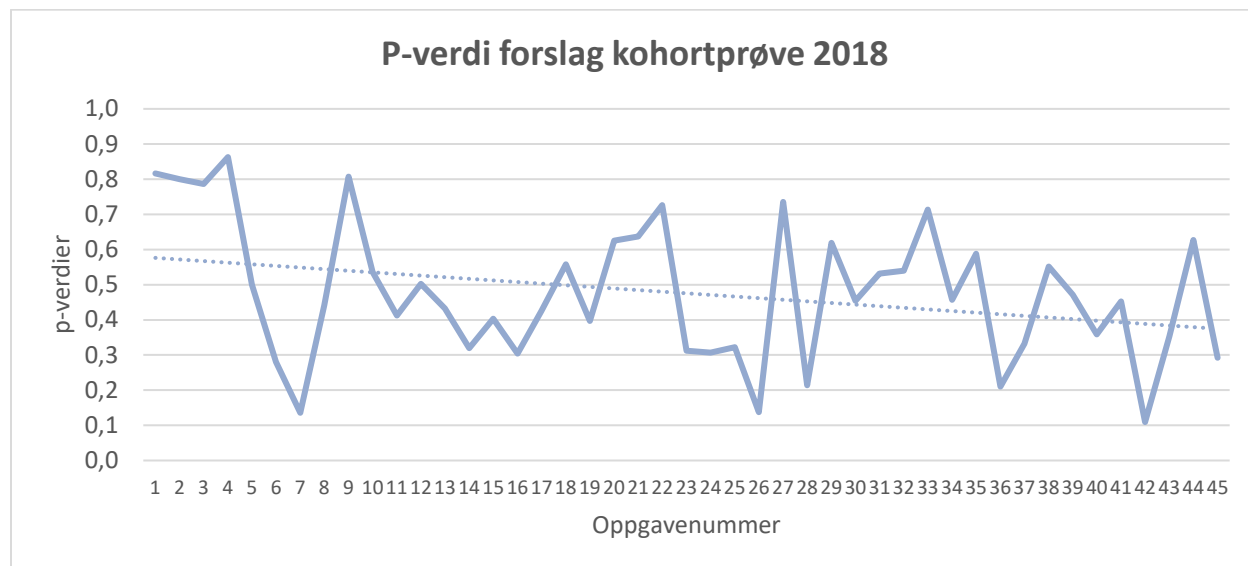
Som tidligere nevnt har vi gjennom flere år sett at gjennomsnittlig løsningsprosent øker, og at gjennomsnittlig b-verdi synker fra siste utprøving til endelig gjennomføring. Det kan muligens forklares med at en endelig prøve tas mer seriøst av både lærere og elever enn en utprøving, og at elevene er bedre forberedt til den endelige prøven enn ved en utprøving. Vi ser også at den gjennomsnittlige svarprosenten øker, noe som også vil påvirke resultatet. Vår erfaring er en økning på 4–6 prosentpoeng i gjennomsnittlig løsningsprosent og en tilsvarende endring i b-verdi, fra 2. pilot til endelig prøve. Det må vi ta hensyn til når vi ut fra vanskelighetsgrad og mestringsnivå skal velge oppgaver til endelig prøve basert på resultater fra 2. piloter.



Figur 3: Fordeling av oppgaver etter mestringsnivå, basert på b-verdier etter preanalyse

¹ Tall fra IRT-analysen i denne tabellen er kun estimerte, siden vi ikke har en lenking i datasettet (2. pilot og ankersettet) som gjør det mulig å beregne disse parameterne.

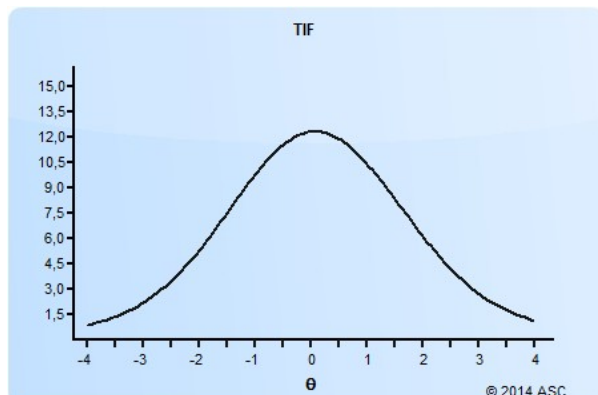
Figur 3 viser fordelingen av oppgaver på mestringsnivå for 2018-prøven, med verdier fra utprøving. Ut fra det vi vet om endringer fra siste utprøving til endelig prøve, vil vi anta at fordelingen etter endelig gjennomføring vil bli noe annerledes enn dette. Det er også en del oppgaver som ligger på grensen mellom to mestringsnivåer, og de oppgavene kan muligens ende opp på et lavere mestringsnivå enn ved pilotering.



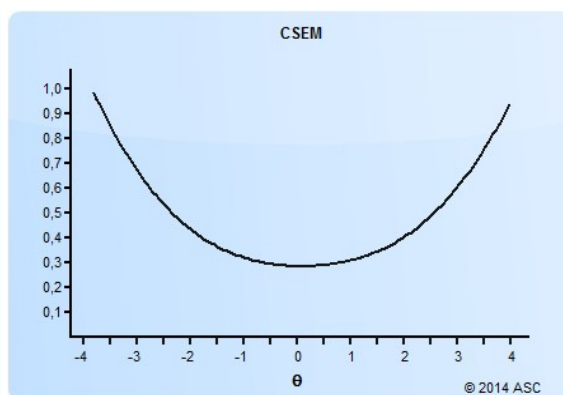
Figur 4: P-verdi fra pilotering til oppgaver i kohortprøve før gjennomføring. Oppgavene i rekkefølge som i prøven

Figur 4 viser oppgavenes P-verdi der vi med hensikt legger opp til en spredning i oppgaver med ulik vanskelighetsgrad gjennom hele prøven. Trendlinja viser en økende vanskelighetsgrad totalt utover i prøven, samtidig som mindre krevende oppgaver også er å finne langt ut i oppgavesettet. Vi vet imidlertid ikke om elevene alltid opplever oppgaver med lav løsningsprosent som vanskeligere enn oppgaver med høy løsningsprosent.

En preanalyse med IRT av de oppgavene som er tenkt i kohortprøven, viser at testinformasjonen (TIF) er høyest for elever med $\theta = 0,100$ (figur 5). Det betyr at prøven gir mest informasjon om elever som presterer midt på ferdighetsskalaen. Det er også der vi har flest elever. Med fastsettelse av ny skala i 2014 ble det bestemt at gjennomsnittet skulle være $\theta = 0$ (50 skalapoeng) og standardavviket 1 (10 skalapoeng). Det betyr at 68 % av elevene skal ligge mellom -1 og 1 på ferdighetsskalaen, og være i intervallet 40 til 60 skalapoeng. I tillegg ser vi at prøven gir god informasjon om elever som presterer lavt og høyt på prøven. Figur 6 viser at måleusikkerheten er lavest der det er flest elever, mellom -1 og 1 på ferdighetsskalaen. Kurvene for TIF og CESM oppfyller anbefalingene i Rammeverket for nasjonale prøver (Utdanningsdirektoratet, 2017b)



Figur 5: Testinformasjon før gjennomføring. Basert på tall fra preanalyse



Figur 6: Måleusikkerhet før gjennomføring. Basert på tall fra preanalyse

Validitet²

I Kunnskapsløftet (LK06) presiseres det at regning er en grunnleggende ferdighet i alle fag. Nasjonal prøve i regning skal avdekke i hvilken grad elevenes regneferdigheter er i samsvar med beskrivelsene av regning som grunnleggende ferdighet (Jf. Rammeverk for Nasjonale prøver, Utdanningsdirektoratet 2017b). Siden prøven tester regning i kompetansemålene for de ulike fagene etter 4. trinn i LK06, blir den gjennomført tidlig på 5. trinn.

Prøven for 2018 inneholder oppgaver innen områdene tall (T), måling og geometri (M&G), og statistikk (S). Oppgavene er knyttet til områdene i tråd med beskrivelsene i Rammeverket for nasjonale prøver (Utdanningsdirektoratet, 2017b). I tillegg oppfyller oppgavene kravene i Rammeverk for grunnleggende ferdigheter (Utdanningsdirektoratet, 2017a), ved at oppgavene utfordrer elevene i ulike prosesser i den helhetlige problemløsningsprosessen. Noen oppgaver har hovedutfordring i å formulere et regneuttrykk (gjenkjenne og beskrive), andre i å foreta beregningen (bruke og bearbeide) eller å avmatematisere innholdet i oppgaven (reflektere og vurdere).

Det å kunne bruke regneferdigheter i faglige og dagligdagse sammenhenger skal vektlegges, og prøven skal inneholde både åpne oppgaver og flervalgsoppgaver. Det er interaktive oppgaver innenfor begge oppgaveformatene. I åpne oppgaver skal elevene enten skrive et svar, merke antall ruter i et rutenett, tegne figurer, dra opp søyler eller stille klokka. I flervalgsoppgaver skal elevene enten velge ett av flere svaralternativer, flytte objekter som hører sammen eller koble sammen ulike representasjoner av samme tall. Oppgavene i prøven for 2018 knyttes til kompetansemål i fag som vist i tabell 6. Hver oppgave kan knyttes til kompetansemål i flere fag.

² Validitet betyr at prøven måler det den er ment å skulle måle, - grunnleggende ferdigheter i regning innenfor områdene tall, måling og geometri og statistikk i alle fag, relatert til LK06, og rammeverk for grunnleggende ferdigheter (Utdanningsdirektoratet 2012).

Fag	Antall oppgaver
Engelsk (eng)	2
KRLE (krle)	1
Kroppsøving (kro)	11
Kunst og håndverk (khv)	4
Mat og helse (mhe)	8
Matematikk (mat)	45
Musikk (mus)	0
Naturfag (nat)	14
Norsk (nor)	11
Samfunnsfag (saf)	17

Tabell 6: Antall oppgaver som kan knyttes til kompetansemål i ulike fag etter LK06 4. trinn

Prøven for 2018 har ingen oppgaver som kan knyttes direkte til kompetansemål i faget musikk til tross for at det i utprøvingen ble prøvd ut flere oppgaver med relevans for musikk. Så lenge en setter krav om at kompetansemålene som er knyttet til en oppgave skal være regning på fagets premisser, kan det være oppgaver der dette ikke lar seg forene med psykometriske krav. Vi opplever det utfordrende å finne gode oppgaver til en nasjonal prøve som måler regneferdighet blant annet i faget musikk.

Nasjonal prøve i regning for 2018 består av 21 oppgaver i området tall, 16 innenfor måling og geometri og 8 statistikkoppgaver (tabell 7). Denne fordelingen med litt flere oppgaver innen tall enn innenfor måling og geometri, samt omtrent halvparten av dette antallet i statistikk, gjenspeiler etter vår vurdering hvordan regning er synliggjort i kompetansemålene i kunnskapsløftet. Det er imidlertid viktig å påpeke at mange av oppgavene krever en kompetanse fra flere kategorier. I alle oppgaver er det en form for tallbehandling. Vi har forsøkt å kategorisere oppgavene etter i hvilken kategori vi vurderer at hovedutfordringen ligger.

Oppgaver på mestringsnivå	Mestringsnivå 1	Mestringsnivå 2	Mestringsnivå 3	Totalt
Tall	6	10	5	21
Måling og geometri	0	9	7	16
Statistikk	2	3	3	8
Totalt	10	20	15	45

Tabell 7: Antall oppgaver på mestringsnivå i hver kategori. Tall fra 2. pilot

Det er ingen oppgaver innen måling og geometri på mestringsnivå 1 etter preanalyse basert på verdier fra siste utprøving, men vi har to oppgaver som er så tett opp til mestringsnivå 1 i b-verdi, at vi tror begge disse mest sannsynlig vil havne på laveste mestringsnivå etter endelig gjennomføring. Inndeling i mestringsnivåer er basert på de samme kuttene til b-verdier som brukes i den nasjonale prøven i regning for 5. trinn.

Tabell 8 viser en oversikt over fordeling av oppgaver knyttet til prosessene i den helhetlige problemløsningsprosessen, gjenkjenne og beskrive, bruke og bearbeide, og reflektere og vurdere. Oppgavene er plassert der hvor hovedutfordringen i oppgaven er ment å ligge. Det er imidlertid viktig å

være klar over at elever har ulike ferdighetsprofiler og at én elev kan ha utfordringer i å gjenkjenne og beskrive et problem, mens en annen elev kan ha utfordringer knyttet til bruke og bearbeide i samme problem.

Oppgaver fra ferdighetsområder	Mestringsnivå 1	Mestringsnivå 2	Mestringsnivå 3	Totalt
Gjenkjenne og beskrive	2	12	5	19
Bruke og bearbeide	5	8	6	19
Reflektere og vurdere	1	2	4	7
Totalt	8	22	15	45

Tabell 8: Fordeling av oppgaver innen ferdighetsområder. Tall fra 2. pilot

Som tabell 8 viser, er det færrest oppgaver der hovedutfordringen er reflektere og vurdere. Det betyr ikke at dette er en prosess som i liten grad måles i prøven. Elevene bør i alle oppgaver vurdere og reflektere om det matematiske svaret de har funnet er et svar som er realistisk til konteksten oppgaven.

Eksakte tall for oppgavene i forslaget til endelig prøve og til ankerprøven finnes i vedlegget *TR NP5-2018* under fanene *Forslag endelig prøve 2018* og *Forslag Ankerprøve 2018*.

Forklaring til noen oppgaver i endelig prøve

Av de 45 oppgavene i endelig prøve er det 41 oppgaver som tilfredsstiller anbefalingene om at oppgavene skal ha en point-biserial-verdi $> 0,3$ og en α -verdi $> 0,5$. De fire som ikke tilfredsstiller begge disse kravene, er likevel tatt med i settet ut fra pedagogiske og didaktiske valg.

I IRT-analysen av 2. pilotene får åtte oppgaver som er med i endelig prøve, bias mot kjønn, tre mot gutter (oppgave 8, 22 og 27) og fem mot jenter (oppgave 13, 24, 25, 37 og 43). I tabell 9–12 er det satt inn tall for vanskelighetsgrad og diskriminering fra både CTT- og IRT-analysene. I tillegg er gjennomsnittlig dyktighet (θ) for elevene som har valgt de ulike alternativene i flervalgsoppgavene, tatt med.

Når det gjelder tall knyttet til forskjeller mellom gutter og jenter, har vi tatt med Mantel-Haenszel koeffisienten (M-H). I IRT-analysen er gutter satt som gruppe 1 og jenter som gruppe 2, noe som betyr at dersom $M-H > 1$, har guttene i gjennomsnitt over hele ferdighetsskalaen større odds for å løse oppgaven enn jentene. Dersom $M-H < 1$, har guttene mindre odds til å løse oppgaven enn jentene. Vi har også med signifikansindikator (p) fra IRT-analysen som angir hvor sannsynlig det er at forskjellen mellom de to gruppene skyldes tilfeldigheter eller ikke. I Xcalibre er 0,05 satt som signifikansmål, og oppgaver som har verdi lavere enn dette, kommer ut med bias.

Oppgaver med lav diskriminering

Forslaget til prøve har fire oppgaver som har point-biserial verdier under 0,3. Kun en av disse oppgavene har α -verdi $< 0,50$, og vi velger her å redegjøre for hvorfor vi mener denne oppgaven bør være med i prøven.

Oppgaver der en må foreta målinger, vurdere eller sammenlikne tall med måleenheter kan være noe uvant for elevene og derfor gi lave diskrimineringsverdier under utprøvinger. Dette kan skyldes at elever som skårer høyt på hele prøven ikke nødvendigvis svarer riktig i slike oppgaver, eller at noen elever som ikke skårer høyt, løser oppgaven. Oppgaver av denne typen måler en viktig dimensjon i den grunnleggende ferdigheten å kunne regne. I siste utprøving ser vi at oppgave 11, har a-verdi $< 0,5$, men S-Rpbis $> 0,3$, og vi mener derfor at den kan være med i den endelige prøven (tabell 9).

Oppgave 11 (lav diskriminering)

Termometeret viser temperaturen både i celsius (°C) og fahrenheit (°F).

Hvor mange grader fahrenheit tilsvarer 0 °C ?

- ☐ 15 °F
☐ 32 °F
☐ 60 °F



Itemnr.	T-Rpbis	S-Rpbis	P-verdi	M-H	P	a-verdi	b-verdi	Gjennomsnittlig theta			
								1	2	3	4
5163292 Termometer	0,299	0,331	0,41	1,264	0,240	0,443	0,53	-0,13	0,39	-0,18	x

Tabell 9: Item 5163292 Termometer (lav diskriminering). Tall fra 2. pilot

Oppgaver der distraktorer har positiv point-biserial

I forslaget til endelig prøve er det fire flervalgsoppgaver som har en distraktor som korrelerer svakt positivt. Oppgaven med høyest positiv korrelasjon på et feilsvar, er en oppgavetype som har vært behandlet i veiledningsmateriell tidligere år og som vi håper kan være et eksempel på en oppgave som kan brukes til å reflektere over realistiske svaralternativer. Mange elever som skårer relativt godt på prøven i sin helhet, velger siste alternativ i denne oppgaven, altså 1500 kg. Det at elever velger feilsvar som ut fra litt refleksjon åpenbart må være feil, ser vi går igjen fra år til år. Derfor ønsker vi at denne oppgavetypen også skal være representert i årets prøve.

Tabell 10 viser sentrale tall i oppgave 43 der en også ser hvordan den gjennomsnittlige dyktigheten til elevene fordeler seg på de ulike svaralternativene. Gjennomsnittlig theta på svaralternativ 4 er 0,11.

Oppgave 43 (Positiv point-biserial)

Emma skal lage middag til seks personer.
Hun beregner 250 g laks per person.

Hvor mange kilogram laks trenger hun?

- ☐ 1,5 kg
☐ 15 kg
☐ 1250 kg
☐ 1500 kg



Itemnr.	T-Rpbis	S-Rpbis	P-verdi	M-H	P	a-verdi	b-verdi	Gjennomsnittlig theta			
								1	2	3	4
5029478 Lakse- middag	0,371	0,405	0,35	1,577	0,034	0,554	0,75	0,38	-0,24	-0,15	0,11

Tabell 10: Item 5029478 Laksemiddag (positiv point biserial). Tall fra 2. pilot

Oppgaver med bias

Forslaget til prøve inneholder åtte oppgaver som gir bias for kjønn etter siste pilotering. Det er fem oppgaver som har bias mot jenter (oppgave 13, 24, 25, 37 og 43) og tre oppgaver som har bias mot gutter (oppgave 8, 22 og 27). Fire av oppgavene med bias mot jenter er innen området måling og geometri og gjelder omgjøring mellom enheter. To av oppgavene med bias mot gutter er divisjonsoppgaver.

Oppgave 37 (bias mot jente)

Aurora har et smykkeskrin med spilledåse.
Når hun åpner smykkeskrinet, spilles en melodi i 2,5 min.

Hvor mange sekunder varer melodien?

- ☐ 125 s
☐ 130 s
☐ 150 s
☐ 170 s



Itemnr.	T-Rpbis	S-Rpbis	P-verdi	M-H	P	a-verdi	b-verdi	Gjennomsnittlig theta			
								1	2	3	4
5127399 Spilledåse	0,459	0,419	0,33	1,731	0,015	0,658	0,78	-0,28	-0,35	0,65	0,13

Tabell 11: Psykometriske verdier for oppgave 37, som har bias mot jente. Tall fra 2. pilot.

Omgjøring fra gram til kilogram, meter til kilometer og minutter til sekunder får ofte verdier som gir bias mot jenter. Oppgave 37 er en omgjøringsoppgave der utfordringen er å gjøre om fra minutter til sekunder. Dette er en omgjøring vi også har sett gå i guttefavør tidligere. Oppgave 37 har en kontekst vi mener ikke favoriserer gutter. Å kjenne sammenhengene mellom sentrale enheter mener vi tilhører grunnleggende ferdigheter i å kunne regne innenfor flere fag, og vi ønsker derfor å ha med slike oppgaver i prøven.

Oppgave 27 (bias mot gutt)

24 elever skal spille fotball.
De skal deles inn i fire like store lag.

Hvor mange elever blir det på hvert lag?

☐ 4

☐ 6

☐ 8

☐ 12



Itemnr.	T-Rpbis	S-Rpbis	P-verdi	M-H	P	a-verdi	b-verdi	Gjennomsnittlig theta			
								1	2	3	4
5161484 Benkefotball	0,448	0,407	0,74	0,613	0,023	0,755	-1,06	-0,94	0,28	-0,52	-0,54

Tabell 12: Psykometriske verdier for oppgave 37, som har bias mot gutt. Tall fra 2. pilot.

To av oppgavene som har bias mot gutter, handler om divisjon. Vi ser at oppgaver som krever nøyaktighet og der en kan bruke en standardalgoritme, ofte går i favør av jenter. Slike oppgaver måler også en ferdighet vi mener en nasjonal prøve i regning bør måle.

Ankeroppgaver

Utgangspunktet for ankerdesignet i nasjonale prøver i regning for 5. trinn, var 20 utvalgte oppgaver fra den ordinære prøven for 2012. I 2014 ble disse, sammen med 25 oppgaver fra kohortprøven for 2014, satt sammen til en egen ankerprøve. I etterkant av gjennomføringen i 2014, ble ankeroppgavene kalibrert sammen med kohortprøven. Ankeroppgavene i 2014 ble også brukt i 2015, men da var vanskelighetsgraden til ankeroppgavene fastsatt ut fra gjennomføringen i 2014. I etterkant av gjennomføringen i 2015 ble det oppdaget et behov for å ha en modell som gjør det mulig å skifte ut ankeroppgaver. Et viktig poeng var da å utvikle et design som gjør det mulig å endre ankeret, uten at det går på bekostning av antall oppgaver som er ferdig kalibrert.

I 2016 ble det innført en modell der ankerprøven består av 20 ankeroppgaver, 4 framtidige ankeroppgaver og 21 oppgaver fra kohortprøven. De framtidige ankeroppgavene, kalt preankeroppgaver, er med i ankerprøven, men fungerer ikke som ankeroppgaver og behandles som oppgaver fra kohortprøven. I etterkant av gjennomføringen blir preankeroppgavene kalibrert sammen

med kohortprøven, og ankeret, og går inn i ankeret som ankeroppgaver året etter. Det er derfor viktig at disse hemmeligholdes etter gjennomføring.

Modellen gjør det mulig å skifte ut opptil fire ankeroppgaver hvert år, som igjen betyr at hele ankeret kan fornyes i løpet av en femårsperiode. Ved hjelp av modellen kan dette skje uten å endre antall oppgaver som brukes til å kalibrere resultatene og måle utvikling over tid.

Ankeret i 2018 vil bestå av 20 oppgaver som er blitt kalibrert et tidligere år. 13 av disse oppgavene ble kalibrert i ankerprøven i 2014, to i 2015, én i 2016 og fire i 2017. I tillegg inneholder ankerprøven fire preankeroppgaver. Preankeroppgavene er med i ankerprøven for å bli kalibrert i 2018, og skal etter planen gå inn i ankeret i 2019.

Da vi valgte preankeroppgaver, var dette oppgaver fra 2. pilot med gode verdier og oppgaver som vi mente kunne fungere i noen år framover. For å kunne velge hvilke ankeroppgaver som skal skiftes ut, har vi analysert og sett på hvilke ankeroppgaver som har hatt størst variasjon i verdier fra år til år. Dette sammen med en vurdering av om det rett og slett er oppgaver som bør skiftes ut fordi de er «utgått på dato», har vært hovedkriteriene i forhold til hvilke ankeroppgaver vi har byttet ut. Samtidig er valg av innhold, format og område noe vi også må ta i betraktning slik at ankeret ikke endres vesentlig.

De nye ankeroppgavene bør stå på samme plass som da de var preanker, noe som gjør at vi må være bevisste på hvor i settet disse oppgavene plasseres. På den måten kan vi også justere plasseringene til ankeroppgavene så lenge vi skifter ut fire oppgaver hvert år. Det gjør at ankersettet for 2018 har flere ankeroppgaver som følger etter hverandre, noe som kan være en fordel for å kunne sammenligne resultater fra år til år.

Generelle vurderinger av prøven som helhet

Gjennom utprøvinger har hver oppgave i endelig prøve vært prøvd ut minimum to ganger. 26 oppgaver har vært med i tre utprøvinger, mens 19 har vært med i to. De fleste oppgavene har små endringer i verdier fra 1. pilot til 2. pilot og holder seg nokså stabile på mestringsnivå. Dette tyder på at oppgavene er relativt stabile.

Opgavene i prøven har stor bredde både når det gjelder vanskelighetsgrad, matematiske tema og oppgavenes kontekster. Selv om ikke regning i alle fag er representert i prøven, har prøven relevans for mange fag. Det er en god blanding av oppgaver hvor elevene skal fylle inn svar og velge svaralternativer.

Kapittel 5: Analyse av nasjonal prøve i regning 2018

Dette kapitlet er en analyse av resultatene til elevene som gjennomførte nasjonal prøve i regning på bokmål eller nynorsk for 5. trinn høsten 2018. Tallene det henvises til er basert på en analyse av resultatene fra kohortprøven og ankerprøven, der parameterne til ankeroppgavene er brukt for å kalibrere oppgavene i kohortprøven. Det var 60 591 elever på 5. trinn som gjennomførte prøven. Oppgavene i prøven er plassert på mestringsnivå ut fra vanskelighetsgrad, bestemt ved IRT-analyse. En oppgave med høy vanskelighetsgrad gir flere skalapoeng enn en oppgave med lav vanskelighetsgrad. Skalapoengene avgjør mestringsnivået til elevene.

Det er tre mestringsnivå for prøven på 5. trinn, der mestringsnivå 3 er det høyeste nivået. Grensene til mestringsnivåene er faste fra år til år, og følgende grenser gjelder for nasjonal prøve i regning for 5.trinn:

Mestringsnivå 1:	42 eller færre skalapoeng
Mestringsnivå 2:	43–56 skalapoeng
Mestringsnivå 3:	57 eller flere skalapoeng

Psykometriske verdier

Tabell 15 viser de tekniske verdiene til kohortprøven og ankerprøven.

NP5 2018	Endelig prøve ³	Ankerprøve
Antall gutter	30785	1864
Antall jenter	29806	1808
Totalt antall elever	60591	3672
Gj.snitt Theta gutter	0,124	0,101
Gj.snitt Theta jenter	-0,132	-0,111
Gj.snitt Theta totalt	-0,002	-0,003
Sign. forskjell J - G (Theta) $p < 0,05$	Ja	Ja
Standardavvik Theta	0,960	0,915
SE til gjennomsnitt Theta	0,004	0,015
Gj.snitt P-verdi gutter	58,0 %	55,2 %
Gj.snitt P-verdi jenter	52,4 %	49,3 %
Gj.snitt P-verdi totalt	55,2 %	52,3 %
Svarprosent	97,4 %	96,7 %
Cronbachs alpha	0,921	0,920
Antall item	45	45
Antall MC	30	29
Antall IA	9	10

Tabell 13: Psykometriske verdier til kohortprøven og ankerprøven 2018

³ Analysen er basert på alle elever som gjennomførte NPREG05, også de som tok ankerprøven som ikke tok alle oppgavene i kohortprøven

Oversikt over oppgaver NP5 2018

Oppgaver	Innhold	Område	Format	Relevans til fag	p-verdi ⁴ alle	p-verdi jenter	p-verdi gutter	Prosent- poeng forskjell J - G	b- verdi	Mestrings- nivå (IRT)
1	Subtraksjon, hele tall	T	Å	nat	85	81	89	-8	-1,53	1
2	Divisjon, hele tall	T	Å	mat	83	79	87	-7	-1,31	1
3	Multiplikasjon, hele tall	T	Å	mat	81	80	81	-1	-1,58	1
4	Lage diagram	S	Å	mat, nat, nor, saf	92	92	92	0	-3,79	1
5	Lengde	M&G	FV	mat, kro	59	58	61	-3	-0,45	2
6	Kjøp og salg	M&G	Å	mat, saf	35	32	38	-5	0,62	2
7	Velge regnearter	T	Å	kro, mat, nor	20	17	23	-7	1,57	3
8	Divisjon, hele tall	T	FV	khv, mat	67	68	67	1	-0,73	1
9	Addisjon, hele tall	T	FV	kro, mat, nor	86	86	86	0	-1,98	1
10	Addisjon, desimaltall	T	FV	mat, kro	61	55	67	-11	-0,58	2
11	Tolke illustrasjon	M&G	FV	mat	48	47	50	-3	0,10	2
12	Tolke og lese diagram	S	FV	mat, nat, nor, saf	66	67	65	2	-0,91	1
13	Forståelse av enheter (g –	M&G	FV	mhe, mat	39	30	47	-16	0,49	2
14	Sammensatt med	M&G	FV	kro, mat	45	42	48	-7	0,27	2
15	Sammenhengen brøk og	T	FV	kro, mhe, mat	49	44	53	-8	0,05	2
16	Multiplikasjon, hele tall	T	Å	mat	44	36	51	-14	0,23	2
17	Regne med forhold	M&G	FV	mat, nat	52	49	54	-6	-0,07	2
18	Multiplikasjon, hele tall	T	FV	mat, nor, saf	67	65	68	-3	-0,86	1
19	Stille analog klokke	M&G	Å	mhe, mat, nat, saf	49	48	50	-3	0,04	2
20	Velge regnearter	T	FV	kro, mat	62	58	67	-9	-0,46	2
21	Omgjøring mellom	M&G	FV	mhe, mat, nat	81	78	84	-6	-1,51	1
22	Lage diagram	S	Å	mat, nat, nor, saf	82	82	81	1	-1,31	1
23	Multiplikasjon, desimaltall	T	FV	eng, mhe, mat	43	41	45	-4	0,28	2
24	Forhold, sammenligne	M&G	FV	kro, mat, saf	34	30	38	-8	0,84	3
25	Sammensatt med	M&G	FV	mat, nat	41	34	49	-15	0,33	2
26	Tolke tabell, dobling	S	FV	mat, nat, nor, saf	20	18	23	-5	1,19	3
27	Divisjon, hele tall	T	FV	kro, mat	78	79	77	2	-1,39	1
28	Sammenligne brøker	T	FV	mhe, mat, saf	29	27	31	-4	0,83	3
29	Lese av tabell	S	FV	mhe, mat	72	72	72	1	-0,99	1
30	Egenskaper til mønstre	T	FV	khv, mat	53	53	54	0	-0,14	2
31	Divisjon, hele tall	T	FV	kro, mat	65	60	70	-10	-0,68	2
32	Divisjon, hele tall	T	Å	khv, mat	63	59	67	-7	-0,49	2
33	Multiplikasjon, hele tall	T	Å	mat	77	77	76	1	-0,99	1
34	Tolke tabell, beregne tid	M&G	FV	mat, nat, nor, saf	62	59	65	-5	-0,46	2
35	Lese av diagram	S	Å	mat, nat, nor, saf	64	61	67	-6	-0,55	2
36	Tolke tabell	S	FV	mat, nat, nor, saf	29	24	33	-9	0,91	3
37	Omgjøring (min – s)	M&G	FV	mat, saf	42	34	49	-15	0,35	2
38	Overslag	T	FV	eng, mat, saf	56	53	58	-5	-0,38	2
39	Temperatur, negative tall	M&G	FV	mat, nat	61	57	64	-7	-0,47	2
40	Sammensatt	T	FV	krle, mat, nor, saf	36	34	37	-4	0,87	3
41	Halvering	T	Å	mat, saf	50	44	57	-13	-0,02	2
42	Kvadrat, omkrets	M&G	Å	khv, mat	13	13	13	-1	2,17	3
43	Sammensatt med	M&G	FV	mhe, mat	45	36	55	-19	0,22	2
44	Koordinatsystem	M&G	FV	kro, mat	72	72	72	1	-1,12	1
45	Tolke diagram	S	Å	mat, nat, saf	35	33	38	-5	0,66	2
Gjennomsnitt					55,4	52,5	58,1	-0,06	-0,29	

Tabell 14: Oppgaverekkefølge i versjon 1 av Nasjonal prøve i regning, NP5 2018, N = 60 591
(Ng = 30 785 og Nj = 29 806)

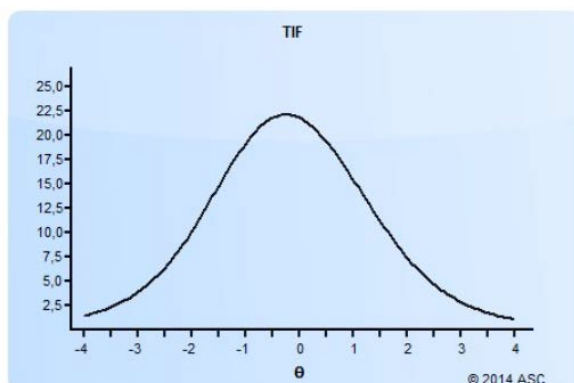
⁴ P-verdi betyr prosentandel riktig svar på oppgaven

Tabell 16 viser rekkefølgen til oppgavene i versjon 1 av kohortprøven, hva oppgavene måler, fagområde, oppgaveformat, oppgavens tilknytning til fag, løsningsprosenter for alle og jenter og gutter hver for seg, kjønnsforskjeller, oppgavens vanskelighetsgrad (b-verdi ut fra IRT-analyse) og mestringsnivå.

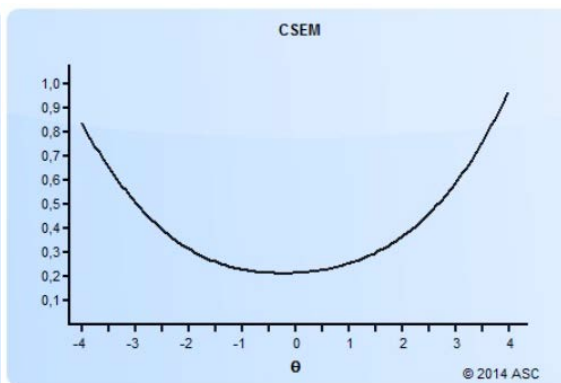
Nasjonal prøve i regning for 5. trinn måler elevenes kompetanse etter 4. trinn i grunnleggende ferdigheter i å kunne regne. Derfor må valideringen skje ut fra det en kan forvente av elever som nettopp har begynt i 5. trinn.

En lærergruppe med fire personer har gjennomført ekstern validering av prøvesettet. Det vil si at de har vurdert hver enkelt oppgave og prøvesettet som helhet opp mot kravene til validitet som står i rammeverket. Endelig godkjenning av prøven ble gjort av Utdanningsdirektoratet og deres egne interne og eksterne kvalitetssikrere.

Testinformasjon og måleusikkerhet



Figur 7: Testinformasjon NP5 2018



Figur 8: Måleusikkerhet NP5 2018

Prøven har stor informasjonsverdi med lav måleusikkerhet for elever som ligger midt på ferdighetsskalaen og innen ett standardavvik resultatmessig (figur 7). Maksimum testinformasjon var 22,1 ved $\theta = -0,250$. Måleusikkerheten er imidlertid stor for elever i begge ender av ferdighetsskalaen (figur 8). Dette er imidlertid en naturlig konsekvens av at det er flest oppgaver i prøven som ligger midt på ferdighetsskalaen, som igjen er en konsekvens av hvordan elevene fordeler seg utover ferdighetsskalaen. Minimum CSEM var 0,21 ved $\theta = -0,250$.

Tolkninger av resultatene

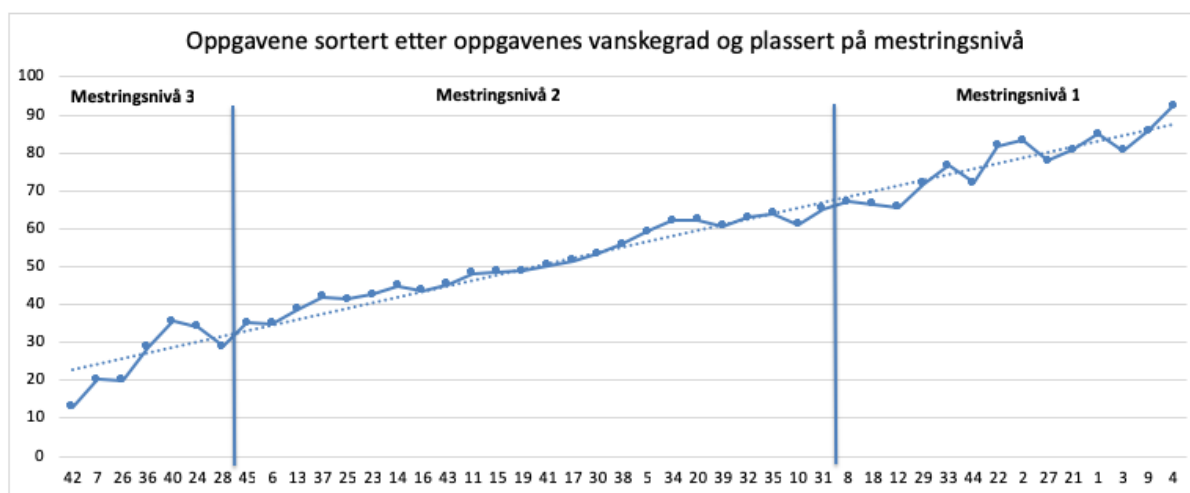
Resultater innen områdene tall, måling og geometri, og statistikk

En enkeltoppgave kan inneholde elementer fra flere områder, og det kan derfor være vanskelig å plassere oppgaver i ett bestemt område. I alle oppgavene behandler vi tall i en eller annen sammenheng. Det typiske for måling og geometri er imidlertid behandling av enheter for masse, lengde og volum, beregning av areal, kjøp og salg, egenskaper til geometriske figurer og regning med

temperatur og tid. Statistikk handler om å lage og lese tabeller og diagram og gjøre beregninger. Av den grunn vil både tallbehandling og riktig bruk av enheter kunne inngå i statistikkoppgaver.

Alle områdene har oppgaver med ulik vanskelighetsgrad og oppgaver på alle tre mestringsnivå. Vanskelighetsgraden i prøven for 2018 ble noe lavere enn vi hadde antatt ut fra verdiene i siste utprøving. I tidligere år har løsningsprosenten for prøven for 5. trinn gjerne endret seg i størrelsesorden 4–6 prosentpoeng fra siste utprøving til endelig prøve. Denne gang er endringen hele 7,8 prosentpoeng. Når vi samtidig ser at ankeret er nokså stabilt, kan vi ikke trekke en slutning om at denne endringen skyldes at elevene har blitt dyktigere. Som vi tidligere har vært inne på, er inntrykket at en endelig prøve bli behandlet mer seriøst enn en utprøving. Det viser blant annet svarprosenten som er omtrent 5 prosentpoeng høyere i den endelige gjennomføringen enn den var i den siste utprøvingen.

Resultater til enkeltoppgaver



Figur 9: P-verdi for hver oppgave. Oppgavene er sortert med avtagende vanskelighetsgrad og plassert på mestringsnivå ut fra IRT-analyse. Nivå 3 er høyeste nivå, og nivå 1 er det laveste mestringsnivået. N = 60 591

Den gjennomsnittlige P-verdien for hele prøven er 55,4 prosent. Det betyr at elevene i gjennomsnitt løste 55,4 prosent av alle oppgavene riktig. Oppgavene har god spredning i P-verdi. Den varierer fra 13 prosent til 92 prosent. Grafen i figur 9 viser også at det er stort samsvar mellom oppgavens vanskelighetsgrad og løsningsprosent.

Analysen viser at det er oppgave 42 som krevde høyest dyktighet og oppgave 4 som krevde lavest dyktighet for å løse oppgaven riktig. De samme oppgavene hadde også henholdsvis laveste og høyeste løsningsprosent.

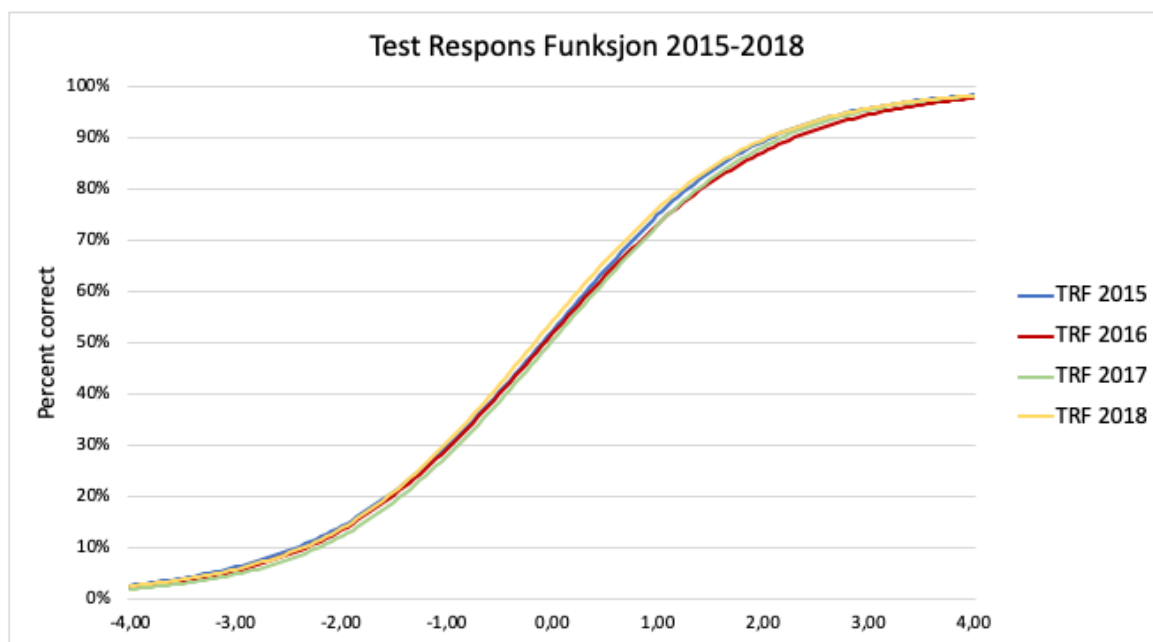
Sammenligning av resultater over tid

Innføring av ny vurderingsskala i 2014, gjør det mulig å måle utvikling over tid. Det er seks prosent av elevene som gjennomfører en ankerprøve hvert år. Omtrent halvparten av oppgavene i ankerprøven har

vært med i tidligere ankerprøver, og resten av oppgavene er fra årets ordinære prøve. Ankeroppgavene kan ikke offentliggjøres, og elevene som tar ankerprøven, får ikke vite hvordan oppgavene er vurdert.

2014 er år null for vurdering av utvikling over tid. Ved hjelp av ankeroppgavene er det mulig å sammenligne resultatene for 2018 med de fire foregående år og vurdere om elevenes dyktighet har endret seg. Fem år er fortsatt kort tid til å kunne trekke noen klare konklusjoner, og endringene er nokså små. Prøven for 5. trinn i 2018 har samme nasjonale gjennomsnitt som prøvene fra og med 2014, 50 skalapoeng.

Test respons funksjonen (TRF) viser sammenhengen mellom ferdigheten til elevene i form av skalapoeng, og antall riktige besvarte oppgaver på prøven. Diagrammet i figur 10 viser en sammenligning mellom de fire siste årene. Vi ser at prøvene er nokså like i vanskelighetsgrad, men enkelte steder ligger kurven for 2018 marginalt høyere enn de andre kurvene. Det betyr at prøven for 2018 har litt lavere vanskelighetsgrad enn i årene vi sammenligner med.



Figur 10: Test Response funksjonen for prøvene fra 2015, 2016, 2017 og 2018

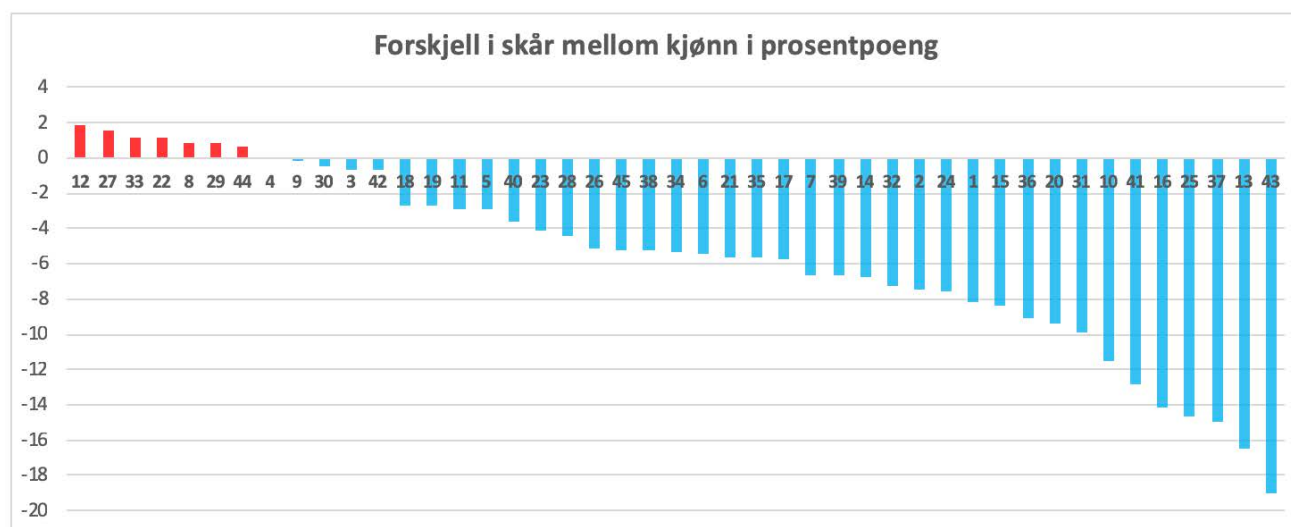
Kjønnsforskjeller

Guttene gjør det bedre enn jentene på prøven som helhet. Den gjennomsnittlige kjønnsforskjellen i guttenes favør er 5,6 prosentpoeng. Dette tilsvarer effektstørrelse 0,27 som betyr 27 prosent av standardavviket. Forskjellen mellom kjønnene er noe større enn den var i prøven for 2017 da den var 4,0 prosentpoeng og i 2016, da kjønnsforskjellen var 3,6 prosentpoeng.

	Gjennomsnittlig løsningsprosent	Effektstørrelse
Jenter	52,4	0,27
Gutter	58,0	

Tabell 15: Gjennomsnittlige P-verdier og effektstørrelse for jenter og gutter

Resultatene til de nasjonale prøvene i regning viser hvert år signifikante forskjeller i guttefavør. Vi ser også at det er de samme oppgavetyperne som hvert år viser forskjeller, og det ser ikke ut til at forskjellene blir mindre.



Figur 11: Forskjell i løsningsprosent for jenter og gutter for hver oppgave. Positive søyler (røde) for oppgaver hvor jentene gjør det bedre enn guttene, og negative søyler (blå) hvor guttene gjør det bedre enn jentene. Oppgavenummer på horisontal akse og prosentpoeng på vertikal akse

I likhet med tidligere år, ser vi en klar tendens til at guttene i gjennomsnitt presterer bedre enn jentene i oppgaver som krever omgjøring mellom enheter i masse og lengde (spesielt kg og g, km og m). Dette gjelder både rene omgjøringsoppgaver og oppgaver der konteksten kan være en hjelp til å vurdere størrelse eller mengde. Resultatene viser også at oppgaver som inneholder desimaltall uten at det nødvendigvis er omgjøring mellom enheter, er oppgaver som har en tendens til å gå i guttefavør. Oppgaven med størst kjønnsforskjell i årets prøve, 19 prosentpoeng i guttefavør, er oppgave 43 med omgjøring mellom kilogram og gram og der selve utregningen ikke bør være det som utfordrer mest. Oppgaven er omtalt tidligere i rapporten og er en oppgavetype som absolutt bør utfordre evne til å reflektere over valg av svar.

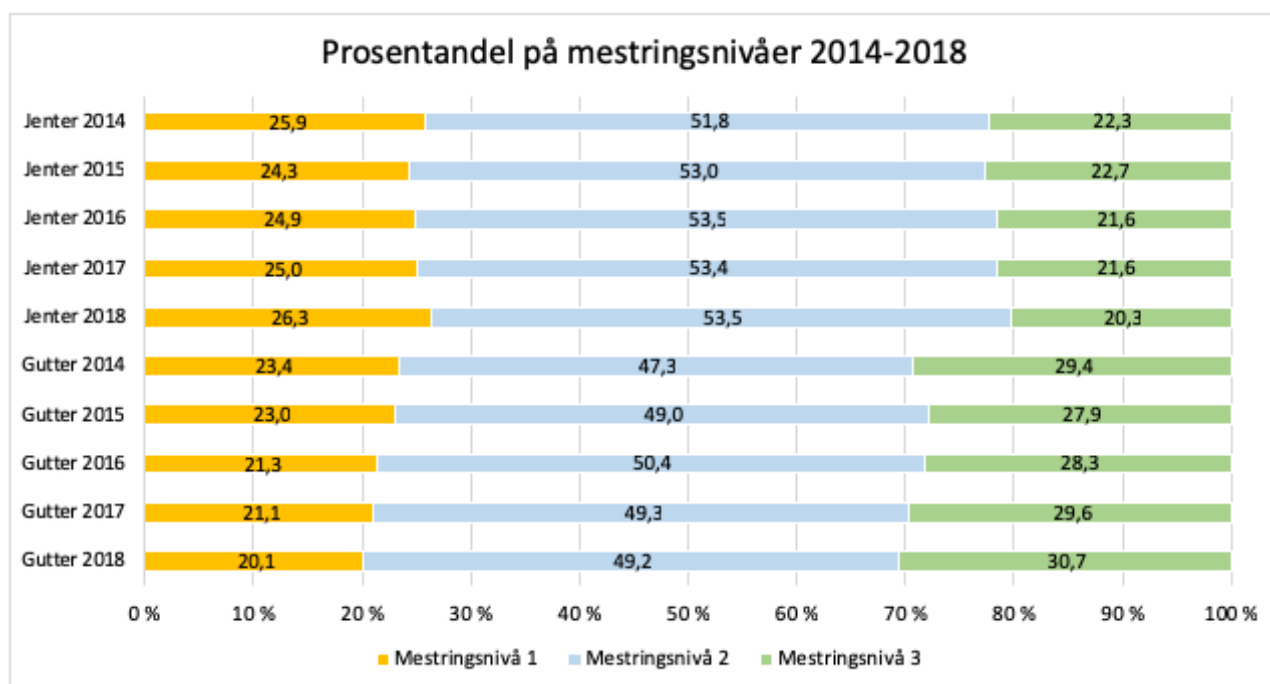
Når det gjelder oppgaver i jentefavør, ser vi av diagrammet i figur 11 at det kun er sju oppgaver der jentene løser oppgaven bedre enn guttene. Diagrammet viser også at forskjellene er mye større i oppgaver som går i guttefavør enn de som går i jentefavør. Oppgave 12 løses best av jentene sammenliknet med guttene, men forskjellen er kun på to prosentpoeng i jentefavør. Det er en oppgave

der du skal sammenlikne fire søylediagram med opplysningene i oppgaveteksten. Oppgaven krever nøyaktighet og evne til å sortere opplysninger.

Kjønnsforskjeller 2014–2018

Elevene er plassert på tre mestringsnivåer etter oppnådd skalapoeng. Diagrammet viser at det hvert år er prosentvis flere jenter enn gutter på mestringsnivå 1 og 2, hhv 6,2 og 4,3 prosentpoeng i 2018, mens det er omvendt på mestringsnivå 3 med 10,4 prosentpoeng flere gutter enn jenter i 2018. 30,7 prosent av guttene i utvalget oppnår en poengsum som tilsvarer mestringsnivå 3, mens 20,3 prosent av jentene oppnår det samme.

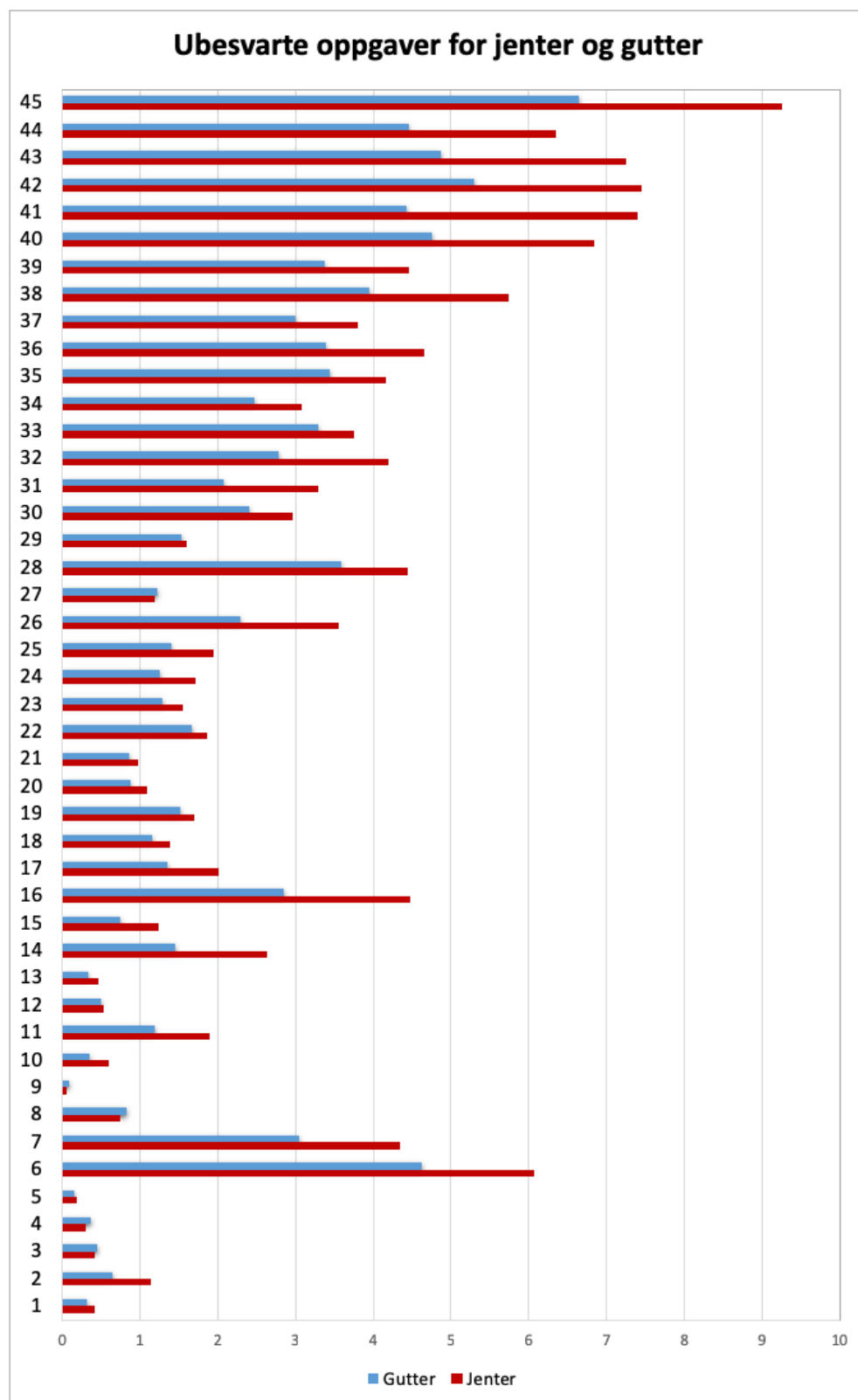
Tallene fra 2014 til 2018 viser at forskjellen mellom prosentandelen jenter og gutter på mestringsnivå 3 helt siden 2014 har vært i guttefavør. Andelen gutter på mestringsnivå 1 har gått jevnt ned siden 2014 mens andelen jenter på laveste mestringsnivå i 2018 har økt sammenlignet med årene før. Figur 12 viser utviklingen for begge kjønn gjennom de fem siste årene.



Figur 12: Prosentandel jenter og gutter på mestringsnivåer 2014-2018

Ubesvarte oppgaver

I gjennomsnitt er 2,6 prosent av oppgavene i prøven ubesvart. Det betyr at gjennomsnittlig svarprosent på årets prøve er litt høyere enn i fjor da andelen ubesvart var 3,6 prosent. Ikke overraskende er det større andel ubesvart blant de åpne oppgavene enn blant flervalgsoppgavene. Diagrammet viser at andelen ubesvart øker utover i prøven, og at andelen ubesvart er høyere blant jentene enn blant guttene (figur 13).



Figur 13: Prosent jenter og gutter som ikke har svart på oppgavene. Rødt for jenter og blått for gutter. Oppgavenummer langs vertikal akse, prosent langs horisontal akse

Tekniske data

Antall oppgaver	Flervalgsoppgaver	Reliabilitet	Gjennomsnittlig p-verdi	Gjennomsnittlig poengsum	Std.avvik	Std.feil til gj.sn.
45	30	0,92	55,2	24,8	9,5	0,04

Tabell 16: Tekniske data for prøven

Antall elever	60 591
Antall gutter	30 785
Antall jenter	29 806
Gj.snitt gutter poeng og prosent	26,1 p (58,0 %)
Gj.snitt jenter poeng og prosent	23,6 p (52,4 %)
G.snitt alle poeng og prosent	24,8 p (55,2 %)

Tabell 17: Antall elever i utvalget. Gjennomsnittlige resultater i poeng og P-verdier

Oppgaveanalyse for alle oppgavene

Oppgaveanalyse for alle oppgavene er vist i tabell 20

- I kolonnen forskjell j-g prosentpoeng, betyr positive verdier at jentene har høyere løsningsprosent enn guttene, og negative verdier at guttene har høyere løsningsprosent enn jentene
- FV betyr flervalgsoppgave. A, B, C og D er svaralternativene
- Å betyr åpen oppgave, 0 betyr galt svar og 1 betyr rett svar
- To oppgaver har point biserial < 0,3
- Noen distraktorer velges av forholdsvis få elever, men vi har likevel valgt å beholde dem
- Distraktorene i flervalgsoppgavene blir i alle oppgavene valgt av elever som har lavere gjennomsnitt/dyktighet (theta) i antall riktige besvarte oppgaver på prøven enn de som har rett svar på oppgavene

5. trinn 2018		Svarfordeling %							Dyktighet (Theta means)							Point biserial	IRT-analyse		P-verdi %	Differanse j-g prosent-poeng	Signifikans-nivå 5 %
Nr	FV Å	A 0	B 1	C	D	E	F	Ubesv	A 0	B 1	C	D	E	F	Ubesv art		a-verdi	b-verdi			
1	Å	15	85					0	-0,99	0,23					-0,86	0,40	0,94	-1,53	85	-8	G
2	Å	16	83					1	-1,06	0,27					-0,86	0,46	1,12	-1,31	83	-7	G
3	Å	19	81					0	-0,72	0,24					-0,80	0,35	0,65	-1,58	81	-1	G
4	Å	7	92					0	-0,55	0,10					-0,88	0,17	0,42	-3,79	92	0	
5	FV	6	26	59	9			0	-0,68	-0,40	0,41	-0,55			-0,60	0,40	0,59	-0,45	59	-3	G
6	Å	60	35					5	-0,31	0,74					-0,44	0,46	0,76	0,62	35	-5	G
7	Å	76	20					4	-0,14	0,84					-0,41	0,34	0,62	1,57	20	-7	G
8	FV	67	17	8	8			1	0,40	-0,59	-1,06	-0,44			-0,41	0,46	0,79	-0,73	67	1	J
9	FV	7	5	86	2			0	-0,88	-0,77	0,19	-0,56			-0,47	0,31	0,65	-1,98	86	0	
10	FV	61	8	17	14			0	0,38	-0,57	-0,61	-0,22			-0,52	0,38	0,56	-0,58	61	-11	G
11	FV	17	48	33				2	-0,26	0,42	-0,31				-0,34	0,32	0,44	0,10	48	-3	G
12	FV	17	6	11	66			1	-0,32	-0,69	-0,49	0,31			-0,58	0,33	0,48	-0,91	66	2	J
13	FV	45	11	5	39			0	-0,25	-0,62	-0,40	0,65			-0,45	0,44	0,69	0,49	39	-16	G
14	FV	12	45	23	18			2	-0,43	0,46	-0,20	-0,33			-0,16	0,34	0,49	0,27	45	-7	G
15	FV	49	25	22	3			1	0,52	-0,52	-0,17	-0,96			-0,39	0,43	0,64	0,05	49	-8	G
16	Å	53	44					4	-0,44	0,68					-0,49	0,51	0,87	0,23	44	-14	G
17	FV	15	14	18	52			2	-0,28	-0,68	-0,59	0,58			-0,37	0,51	0,87	-0,07	52	-6	G
18	FV	14	14	67	4			1	-0,65	-0,38	0,33	-0,46			-0,46	0,37	0,56	-0,86	67	-3	G
19	Å	50	49					2	-0,43	0,55					-0,53	0,46	0,71	0,04	49	-3	G
20	FV	20	9	7	62			1	-0,76	-0,71	-0,57	0,50			-0,37	0,55	1,04	-0,46	62	-9	G
21	FV	18	81					1	-0,74	0,23					-0,47	0,37	0,71	-1,51	81	-6	G
22	Å	16	82					2	-0,95	0,27					-0,62	0,44	0,98	-1,31	82	1	J
23	FV	10	15	31	43			1	-0,89	-0,80	-0,07	0,67			-0,44	0,50	0,83	0,28	43	-4	G
24	FV	64	34					1	-0,23	0,58					-0,31	0,36	0,53	0,84	34	-8	G
25	FV	16	21	20	41			2	-0,59	-0,45	-0,19	0,67			-0,29	0,49	0,82	0,33	41	-15	G
26	FV	77	20					3	-0,21	1,09					-0,36	0,47	1,01	1,19	20	-5	G
27	FV	4	78	6	11			1	-0,90	0,25	-0,75	-0,54			-0,33	0,37	0,66	-1,39	78	2	J
28	FV	67	29					4	-0,28	0,88					-0,39	0,49	0,90	0,83	29	-4	G
29	FV	5	7	72	14			2	-0,56	-0,64	0,32	-0,71			-0,35	0,42	0,73	-0,99	72	1	J
30	FV	6	23	15	53			3	-0,84	-0,46	-0,56	0,55			-0,24	0,51	0,87	-0,14	53	0	
31	FV	5	65	20	7			3	-0,56	0,37	-0,51	-0,80			-0,31	0,44	0,72	-0,68	65	-10	G
32	Å	34	63					3	-0,68	0,46					-0,47	0,53	0,97	-0,49	63	-7	G
33	Å	20	77					4	-0,95	0,33					-0,45	0,51	1,14	-0,99	77	1	J
34	FV	19	4	13	62			3	-0,80	-0,79	-0,47	0,48			-0,28	0,54	0,99	-0,46	62	-5	G
35	Å	32	64					4	-0,65	0,43					-0,43	0,50	0,89	-0,55	64	-6	G
36	FV	22	29	5	3	24	13	4	-0,40	0,80	-0,55	-0,46	-0,02	-0,24	-0,24	0,44	0,78	0,91	29	-9	G
37	FV	34	10	42	11			3	-0,40	-0,51	0,58	0,03			-0,27	0,43	0,66	0,35	42	-15	G
38	FV	12	56	18	10			5	-0,47	0,31	-0,19	-0,28			-0,19	0,30	0,42	-0,38	56	-5	G
39	FV	61	8	8	19			4	0,42	-0,66	-0,92	-0,35			-0,25	0,46	0,74	-0,47	61	-7	G
40	FV	59	36					6	-0,19	0,49					-0,24	0,31	0,45	0,87	36	-4	G
41	Å	44	50					6	-0,53	0,60					-0,33	0,55	0,97	-0,02	50	-13	G
42	Å	81	13					6	-0,07	0,89					-0,20	0,30	0,62	2,17	13	-1	G
43	FV	45	16	16	16			6	0,48	-0,58	-0,51	0,11			-0,20	0,39	0,56	0,22	45	-19	G
44	FV	23	72					5	-0,59	0,27					-0,21	0,37	0,61	-1,12	72	1	
45	Å	57	35					8	-0,29	0,65					-0,19	0,43	0,68	0,66	35	-5	G
Gjennomsnitt ubesvart:								2,6	Gjennomsnitt:							0,42	0,74	-0,28	55,4	-5,6	G

Tabell 18: Oppgaveanalyse for alle oppgavene i kohortprøven. Svarfordeling angitt i prosent. Gjennomsnittlig dyktighet for hvert svaralternativ og for rett eller galt i de åpne oppgavene. Point- biserial betyr oppgavens korrelasjon med summen av poeng. a-verdi betyr oppgavens diskriminering i IRT-analyse, og b-verdien er oppgavens vanskelighetsgrad på en skala fra -4 til 4 ut fra IRT-analyse. Guttene har høyest løsningsprosent i de oppgavene hvor differansen (j-g) er negativ.

G) Kjønnsskjeller i gutters favør, signifikant

J) Kjønnsskjeller i jenters favør, signifikant

Kilder:

Bergem, O.K., Kaarstein, H. & Nilsen, T (2016). Vi kan lykkes i realfag: resultater og analyser fra TIMSS 2015. Oslo: Universitetsforlaget

Grønmo, L. S., Onstad, T., Nilsen, T., Hole, A., Aslaksen, H. & Borge, I.C. (2012). *Framgang, men langt fram. Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2011*. Akademika forlag.

Utdanningsdirektoratet. (2017a). Rammeverk for grunnleggende ferdigheter. Retrieved from https://www.udir.no/globalassets/upload/larerplaner/lareplangrupper/rammeverk_grf_2012.pdf

Utdanningsdirektoratet. (2017b). Rammeverk for nasjonale prøver. Retrieved from <https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover/>