



UiO : Centre for Educational Measurement
Det utdanningsvitenskapelige fakultet

Utvikling av nasjonale prøver – rapport 1 og 2a

Rapportering av potensielle underdimensjoner og implementering av
vertikalt lenkedesign for de nasjonale prøvene i regning

Rapport til Utdanningsdirektoratet

28.2.2019

Henrik Galligani Ræder¹, Olav Dalsegg Tokle² og Rolf Vegar Olsen^{1,2}

¹ Centre for Educational Measurement at the University in Oslo

² Matematikksenteret, Nasjonalt senter for matematikk i opplæringen

Introduksjon

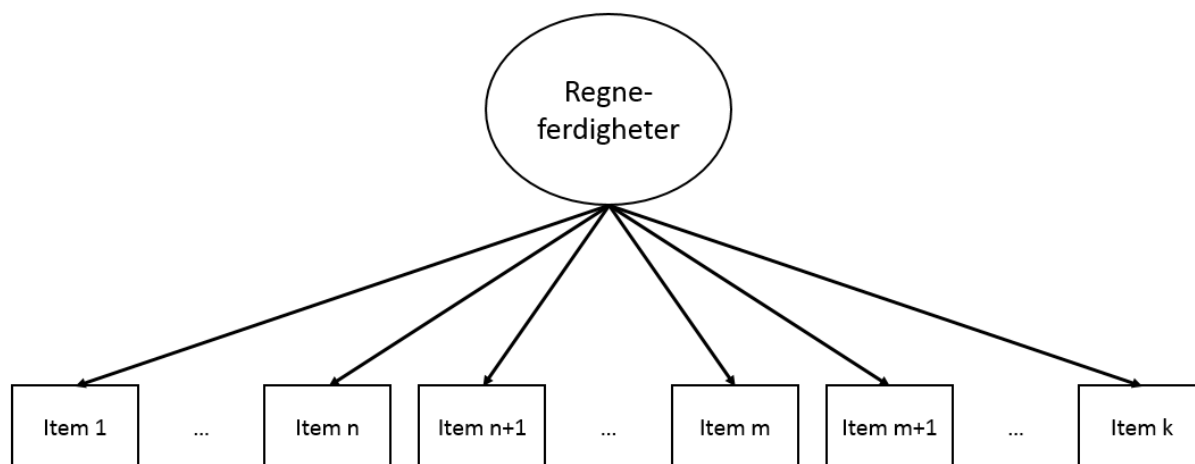
I mai 2018 ble Matematikksenteret tildelt et tilleggsoppdrag knyttet til utviklingen av nasjonale prøver i regning. I dette tilleggsoppdraget skulle det undersøkes om det er mulig, med et relativt enkelt design, å lenke sammen resultater på NP regning på 5. trinn med tilsvarende resultater for NP regning 8./9. trinn. Et slikt design har blitt utviklet og prøvd ut i forbindelse med gjennomføring av prøvene høsten 2018. En nærmere beskrivelse av det designet som er prøvd ut er gitt i figur 3. I forbindelse med oppdraget har Centre for Educational Measurement (CEMO) vært underleverandør for design og statistiske analyser. Dette er en første rapport som rapporterer fra de innledende analysene.

Rapportens første del gir resultater fra konfirmerende faktoranalyser for å etablere dimensjonaliteten i konstruktene som prøvene for 5. og 8. trinn måler. Dette er undersøkelser som foretas for 5. og 8. trinnsprøvene hver for seg, og analysene er gjort for de tre årene 2016, 2017 og 2018. Analyser av prøvene fra 2016 og 2017 ble gjennomført sommeren 2018 for å informere utviklingen av lenkedesignet. Formålet med analysene rapportert i denne delen er å avklare om lenkingen av de nasjonale prøvene i regning må gjøres flerdimensjonalt eller om det kan gjennomføres endimensjonalt, da sistnevnte vil være betraktelig enklere å implementere. Kombinert med at de teoretiske rammeverkene for de to prøvene er svært sammenfallende, vil en sammenfallende dimensjonalitetsstruktur på begge prøvene gi oss et plausibelt grunnlag for å anta at grunnleggende regneferdighet, slik dette er definert og målt, er ett enkelt konstrukt med en progresjon fra 5. til 9. trinn. I avtalen representerer denne delen av rapporten det som er benevnt Del 1.

Den andre delen rapporterer resultater fra utprøvingen av lenkedesign som ble implementert høsten 2018 med hensikt å koble sammen skalaene til de nasjonale prøvene i regning for 5. og 8. trinn. Analysen er basert på data fra gjennomføringen av de ordinære NP regning på 5. og 8. trinn. I tillegg ble det samlet data fra gjennomføring av lenkeprøvene gjennomført på 6. og 7. trinn. Denne delen består kun av deskriptive og innledende analyser. I avtalen representerer dette det som betegnes som Rapport 2a i Del 2

Del 1: Analyse av potensielle underdimensjoner i de nasjonale prøvene i regning

Med dagens fremgangsmåte benyttes en såkalt endimensjonal IRT-modell (se figur 1) hvor man antar at elevenes responser på oppgavene reflekterer et enkelt overordnet konstrukt, «regneferdigheter». Med utgangspunkt i dette konstruktet har det blitt operert med en skala definert i 2014, og gjennom det horisontale lenkedesignet³ som benyttes hvert år er det en konsistent skala som gjør det mulig å sammenligne på tvers av tid (Björnsson, 2018). Som en del av arbeidet vårt for å vertikalt lenke⁴ de nasjonale prøvene for 5. og 8. trinn ble det undersøkt om det er hensiktsmessig å fortsette med en slik endimensjonal tilnærming, eller om vi burde endre tilnærming til en flerdimensjonal modell.



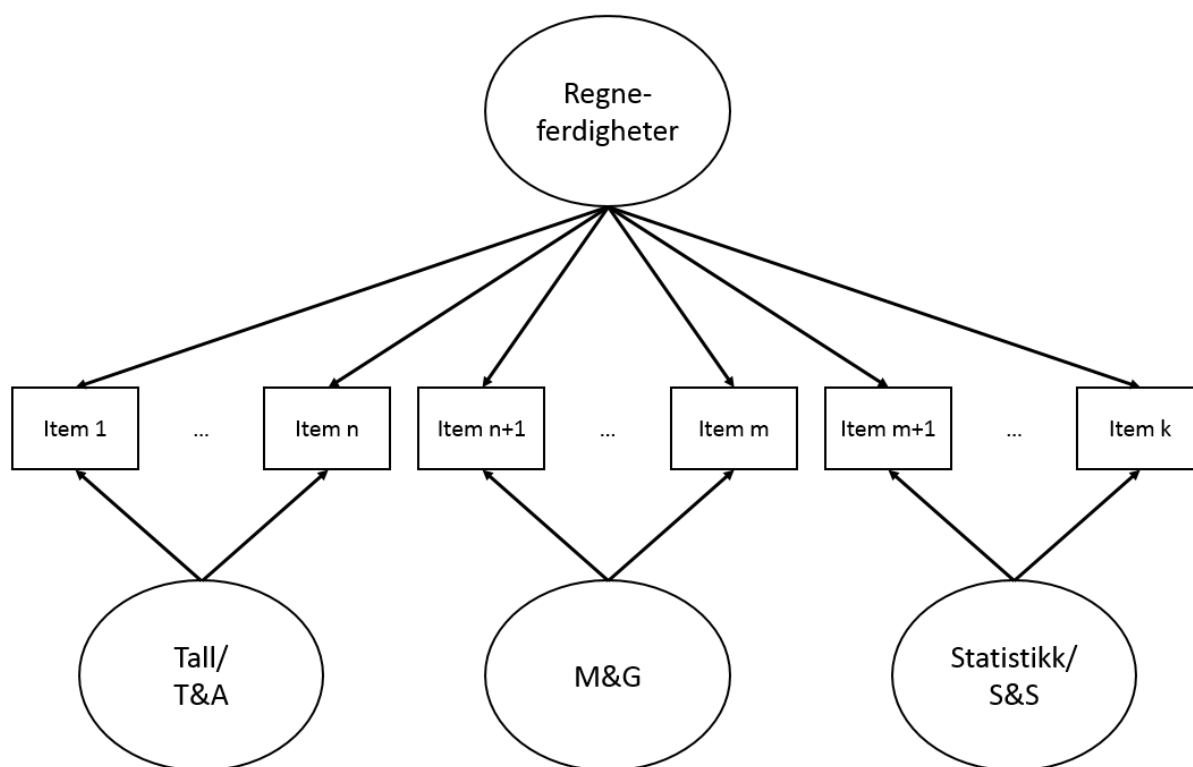
Figur 1: Dagens analysemetode antar at det er kun en faktor som måles, og alle oppgavene brukes til å måle denne ene faktoren. I figuren er det brukt grupper med oppgaver for å tilrettelegge for sammenligning med den påfølgende figuren.

For å teste om den endimensjonale tilnærmingen er forsvarlig benytter vi en modell som heller antar at regneferdighet er et flerdimensjonalt konstrukt (se figur 2), hvor underdimensjonene ble definert ut fra fagområdene til prøvene for 5. og 8. trinn. For 5. trinn er disse fagområdene henholdsvis 1) tall, 2) måling og geometri, og 3) statistikk. For 8. trinn er benyttet de delvis overlappende kategoriene 1) tall og algebra, 2) måling og geometri og 3) statistikk og sannsynlighet. I analysen har vi benyttet testutviklernes kategoriseringer av oppgavene inn i disse underområdene.

I modellen som figur 2 viser, er hver av oppgavene knyttet til to latente variabler: én overordnet regneferdighet og én underdimensjon knyttet til det fagområdet som oppgaven er klassifisert å tilhøre. I denne modellen er det altså en overordnet, generell faktor og til sammen tre faktorer som representerer de tre underdimensjonene. Slike modeller betegnes derfor også generelt som bi-faktor modeller (Reise, 2012).

³ Horisontal lenking, (ofte omtalt som *test equating*) innebærer at man kobler sammen forskjellige tester som er utviklet basert på det samme konstruktet for den samme målgruppen med tilnærmet lik vanskegrad.

⁴ Vertikal lenking innebærer at man kobler sammen forskjellige tester som er utviklet basert på sammenliknbare konstrukter for forskjellige målgrupper med forskjellige vanskegrad.



Figur 2: Bi-faktor modell med tre underdimensjoner. For 5. trinn er disse underdimensjonene tall, måling og geometri (M&G) og statistikk. For 8. trinn er disse underdimensjonene tall og algebra (T&A), måling og geometri (M&G) og statistikk og sannsynlighet (S&S).

En bi-faktor modell estimeres ved at man først beregne hvor mye av variansen til alle oppgavene som hovedfaktoren kan forklare, etterfulgt av at man beregner hvor mye hver av de enkelte underfaktorene kan forklare av variansen som er igjen, såkalt *residual variance*, for den gruppen av oppgaver som disse faktorene er definert fra. Denne fremgangsmåte gjør det mulig å tallfeste hvor sterk en underdimensjon representert ved en gruppe oppgaver er. Det er likevel viktig å være klar over at denne fremgangsmåten alltid vil gi en viss andel varians knyttet til underdimensjonene. Det kan derfor være utfordrende å definere et klart kriterium for når man kan slå fast at dataene reflekterer et rimelig endimensjonalt konstrukt, eller når det er mer hensiktsmessig å konkludere at dataene reflekterer et meningsfullt flerdimensjonalt konstrukt (Reise, Morizot, & Hays, 2007). En tommelfingerregel som ofte benyttes for å skille faktiske og tydelig underdimensjoner fra det som heller reflekterer støy eller neglisjerbare underdimensjoner, er å se på hvor stor del av variansen som hovedfaktoren forklarer sammenliknet med resten av modellen. Avhengig blant annet av antall oppgaver og antall underfaktorer er det anbefalt å gå for en endimensjonal tilnærming når hovedfaktoren står for over 70-80% av den forklart variansen (Rodriguez, Reise, & Haviland, 2016). Det er også verdt å påpeke at en flerdimensjonal tilnærming ofte vil redusere presisjonen på målingene, og derfor ikke bør benyttes med mindre det påvises en sterk flerdimensjonalitet.

Når vi i disse analysene benytter en bi-faktor modell for å evaluere en potensiell dimensjonalitetsstruktur er det samtidig viktig å påpeke at vi ikke anbefaler en slik modell ved potensiell implementering av en flerdimensjonal tilnærming til å analysere eller rapportere resultater fra de nasjonale prøvene. Bi-faktor modeller er et godt verktøy for vårt formål, men det er viktig å være klar over to svakheter med denne typen modeller. Den første svakheten er knyttet til at vi benytter en såkalt bekreftende (*confirmatory*) fremgangsmåte. Det vil si at vi tester en spesifikk struktur, men det kan godt hende det finnes andre strukturer som ville gitt tydeligere andre underdimensjoner. Grunnen til at vi tester akkurat denne strukturen er fordi fagområder er den mest

entydige og faglig relevante kategoriseringen i rammeverket til de nasjonale prøvene i regning. Andre potensielle underdimensjoner kan være f.eks. knyttet til oppgaveformat, kontekst, kognitive krav eller kompetansekategorier. Imidlertid er ikke dette dimensjoner som er like entydig definert som «inndelinger» av begrepet grunnleggende regneferdighet i rammeverket. Den andre grunnen er at det ofte er utfordrende å tolke resultatet fra en bi-faktor modell konseptuelt. I bi-faktor modellen som benyttes vil f.eks. underdimensjonen M&G representere en komponent i residualvariansen for en delmengde av oppgavene. Denne underdimensjonen vil derfor representere den unike restvariansen knyttet til nettopp disse oppgavene og en betegnelse av dette området vil derfor mer teknisk riktig være «ferdigheter i måling og geometri *som ikke er fanget opp av faktoren for generelle regneferdigheter*».

Resultater

I denne delen presenteres resultater fra analyser⁵ gjort av prøvene for 5. og 8. trinn gitt i 2016, 2017 og 2018.

Nasjonale prøver i regning for 5. trinn

For 5. trinn er de tre underdimensjonene definert ut fra oppgaver som dekker de tre fagområdene tall, måling og geometri (M&G) og statistikk. Tabell 1 gir en oversikt over hvordan disse fagområdene er representert i perioden 2016-2018.

Tabell 1: Fordeling av antall oppgaver klassifisert til å tilhøre de tre fagområdene i de nasjonale prøvene i regning for 5. trinn 2016-2018.

5. trinn	Tall	M&G	Statistikk	Antall
2016	19	16	10	45
2017	22	14	9	45
2018	21	15	9	45

For 5. trinn er andel varians forklart av de forskjellige faktorene vist i tabell 2. Her kommer det tydelig fram at den generelle faktoren er svært dominerende i modellen, noe som er en indikasjon på at prøvene har en høy grad av endimensjonalitet. Videre observeres en jevn, men liten, styrking av endimensjonaliteten til prøvene. For å avgjøre om disse endringene er statistisk signifikante vil det være nødvendig å gjennomføre en såkalt *bootstrap*-analyse. En slik analyse har vi foreløpig ikke hatt kapasitet til å gjennomføre. Vi begrenser oss derfor til å si at det er en observert tendens til at de nasjonale regneprøvene for 5. trinn har blitt mer endimensjonale i treårs-perioden. Dette vil imidlertid ikke ha innvirkning på de analysene og tolkningene som denne rapporten presenterer.

Tabell 2: Andel varians forklart av hovedfaktoren og de tre underfaktorene for de nasjonale prøvene i regning for 5. trinn 2016-2018.

G5	Regne-ferdigheter	Tall	M&G	Statistikk	Andel varians forklart av modellen	Derav forklart av hovedfaktoren «Regneferdigheter»
2016	31,2 %	1,1 %	1,3 %	0,8 %	34,4 %	90,7 %
2017	33,4 %	0,9 %	0,9 %	0,6 %	35,8 %	93,3 %
2018	34,7 %	0,7 %	0,5 %	0,5 %	36,4 %	95,3 %

⁵ I analysene har vi brukt R-pakken «mirt», og da spesifikt «bfactor» funksjonen (Chalmers, 2012).

Nasjonale prøver i regning for 8. trinn

For 8. trinn er det tre underdimensjonene definert ut fra oppgaver som dekker de tre fagområdene tall og algebra (T&A), måling og geometri (M&G) og statistikk og sannsynlighet (S&S). Tabell 3 gir en oversikt over hvordan disse fagområdene er representert i perioden 2016-2018.

Tabell 3: Fordeling av antall oppgaver klassifisert til å tilhøre de tre fagområdene i de nasjonale prøvene i regning for 8. trinn 2016-2018

8. trinn	T&A	M&G	S&S	Antall
2016	22	21	7	50
2017	25	18	7	50
2018	22	19	9	50

For 8. trinn prøvene ser vi fra tabell 4 at den generelle faktoren er enda mer dominerende enn for 5. trinn, og vi får en tilsvarende indikasjon på en høy grad av endimensjonalitet. Videre ser det også her ut som prøvene har blitt mer endimensjonale i perioden, men som for analysen for 5. trinns-prøven, har vi ikke undersøkt nærmere om dette er en statistisk signifikant tendens.

Tabell 4: Andel varians forklart av hovedfaktoren og de tre underfaktorene for de nasjonale prøvene i regning for 8. trinn 2016-2018.

G8	Regneferdigheter	T&A	M&G	S&S	Andel varians forklart av modellen	Derav forklart av hovedfaktoren «Regneferdigheter»
2016	32,8 %	0,9 %	0,9 %	0,7 %	35,3 %	92,9 %
2017	36,4 %	0,8 %	0,6 %	0,3 %	38,1 %	95,5 %
2018	36,1 %	0,9 %	1,1 %	0,4 %	38,5 %	93,8 %

Konklusjon

Fra analysene som er gjort i denne rapporten anbefales det å fortsette med en endimensjonal tilnærming når resultatene fra de nasjonale prøvene analyseres og rapporteres. For det spesifikke formålet som denne analysen er gjort for å belyse, lenking av NP regning 5. og 8. trinn, gir analysene støtte for å fortsette utviklingsarbeidet med et forholdsvis enkelt design basert på en antakelse om at prøvene måler det samme konstruktet på forskjellige vanskegradsnivåer.

Vi noterer også at prøvene ser ut til å få en økende grad av endimensjonalitet i perioden (spesielt prøven på 5. trinn). Gitt at man også i framtiden vil fortsette å rapportere resultater for de nasjonale prøvene ved hjelp av en enkelt skår, tyder dette på at prøvekonstruksjonen er «strammet opp» for å optimalisere dette formålet. En mindre optimistisk tolkning av dette funnet kan være at prøvene er snevret inn på en måte som gjør at konstruktet ikke måles i sin bredde i samme grad. Våre analyser gir imidlertid ikke informasjon som gjør det mulig å skille mellom disse to tolkningene.

Til tross for at denne analysen gir veldig tydelige resultater er det viktig å være bevisst på at vi her kun har testet én enkelt spesifisering av eventuell flerdimensjonalitet. Vi har ovenfor argumentert for at det fra rammeverket ikke finnes teoretisk grunnlag for å undersøke alternative spesifikasjoner av dimensjonaliteten.

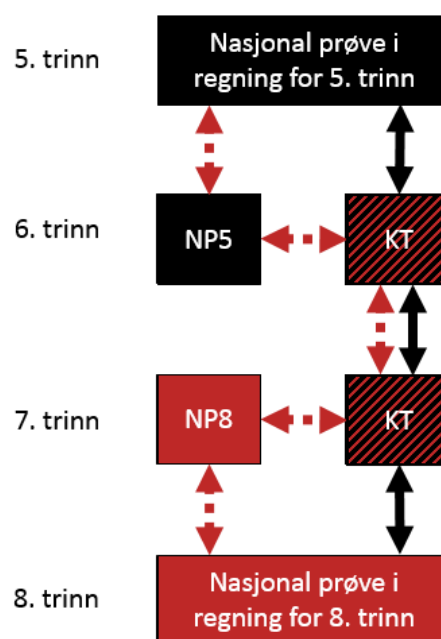
Her kan det være relevant å nevne at oppgaver som skal teste praktisk regneferdighet ofte vil involvere flere fagområder samtidig. Eksempelvis vil en oppgave som er relatert til beregning av temperaturforskjeller, og da spesielt i tilfeller hvor temperaturen skifter fra varmegrader til minusgrader, involvere både fagområdene *måling* og *tall*. Hvordan oppgaven til slutt blir klassifisert vil være avhengig av hva som anses å være hovedutfordringen gitt konteksten og representasjonsformene som er involvert i oppgaven (for eksempel digitalt vs. analogt termometer). Rammeverket opererer med en inndeling etter fagområder for å sikre at prøvene totalt sett skal ha en stabil representasjon av det relevante faglige innholdet. Dersom dette hadde vært et rammeverk som hadde spesifisert at fagområdene skal utvikles for å muliggjøre rapporteringer av underskalaer, ville dette ha ført til utvikling av oppgaver som mer entydig er knyttet til kun ett enkelt fagområde. Undersøkelsen av dimensjonalitet kan derfor kun tolkes i konteksten av de faktiske prøvene, og kan ikke tolkes som at grunnleggende ferdighet i regning som sådan, uansett operasjonalisering, vil framstå som et endimensjonalt begrep.

Del 2: Vertikal lenking av nasjonale prøver – innledende analyser

Høsten 2018 ble det gjennomført en studie av CEMO i samarbeid med Matematikksenteret for å koble sammen de nasjonale prøvene i regning for 5. og 8. trinn. Studien hadde to hovedmål. Det første var direkte knyttet til utviklingen av de nasjonale prøvene, i form av å videreutvikle de nasjonale prøvene til å være i stand til å måle progresjon. Det andre målet med studien var å etablere en skala for norske elevers regneferdigheter som strekker seg over flere trinn for å kunne gjøre det mulig å gjøre longitudinelle studier av samspill mellom ferdigheter og andre viktige faktorer i skolen som f.eks. motivasjon⁶.

Design

Utgangspunktet for studien var å etablere et enkelt og robust design som gjorde det mulig å etablere en slik lenking uten ressurskrevende utvikling av nye trinnspecifikke prøver. Det ble derfor valgt et design hvor lenkingen kunne studeres gjennom å inkludere utvalg på 6. og 7. trinn som besvarte prøver basert på oppgaver fra ordinær gjennomføring av de nasjonale regneprøvene. Figur 3 viser hovedprinsippene for dette designet. Det ble benyttet tre ulike prøver i lenkingen: De ordinære prøvene for 5. og 8. trinn (NP5 og NP8), samt en koblingstest (KT) konstruert fra en kombinasjon av oppgavene i 5. og 8. trinns prøvene. På 6. trinn fikk halvparten av elevene den prøven for 5. trinn og halvparten av elevene fikk koblingstesten. På 7. trinn fikk halvparten av elevene prøven for 8. trinn og halvparten av elevene fikk koblingstesten. Dette designet gjør det mulig å sammenlikne hvordan prøvene fungerer på forskjellige trinn (de vertikale røde pilene i figur 3), samtidig som at man kan undersøke hvor stabilt oppgavene oppfører seg når de kombineres med andre oppgaver designet for et annet trinn (de horisontale røde pilene i figur 3). Videre får vi gjennom designet også en sterk vertikal lenke-stamme (de svarte pilene i figur 3).



Figur 3 Lenkedesignet brukt i den vertikale skaleringen av de nasjonale prøvene i regning. Se tekst for ytterligere forklaring.

Koblingstesten ble konstruert fra en rekke kriterier, hvor de statistiske kriteriene var basert på resultatene fra den siste runden med pilotering av årets oppgaver. Hovedkriteriet for utvalg av oppgaver var at koblingstesten skulle ha en vanskegrad som var litt vanskeligere enn optimalt for 6. trinn og litt lettere enn optimalt for 7. trinn. For å kunne oppfylle dette kriteriet ble det tatt utgangspunkt i resultatene fra TIMSS 2015, hvor Norge deltok med 4. og 5. trinn i populasjon 1, og 8. og 9. trinn i populasjon 2. Selv om TIMSS ikke måler regneferdigheter var dette de eneste prøvene hvor vi hadde målinger av 4. og 5. trinn på sammenliknbare skalaer. I TIMSS 2015 var økningen i matematikk-ferdigheter fra 4. til 5. på 56 skalapoeng, og spredningen på trinnene (standardavviket) var på henholdsvis 71 og 72 skalapoeng. Dette ga oss en økning i ferdigheter på 0.78 SD fra 4. til 5. trinn. Tilsvarende var det en økning på 0.37 SD fra 8. til 9. trinn. En tilsvarende økning var det også fra 8. og 9. trinn på nasjonale prøver. Gitt en lineær avtakende vekst i ferdigheter fra 5. til 8. trinn ble det

⁶ Se f.eks. <https://www.uv.uio.no/ceмо/forskning/prosjekter/utvikling-av-motivasjon-og-prestasjon-i-matematikk/index.html>

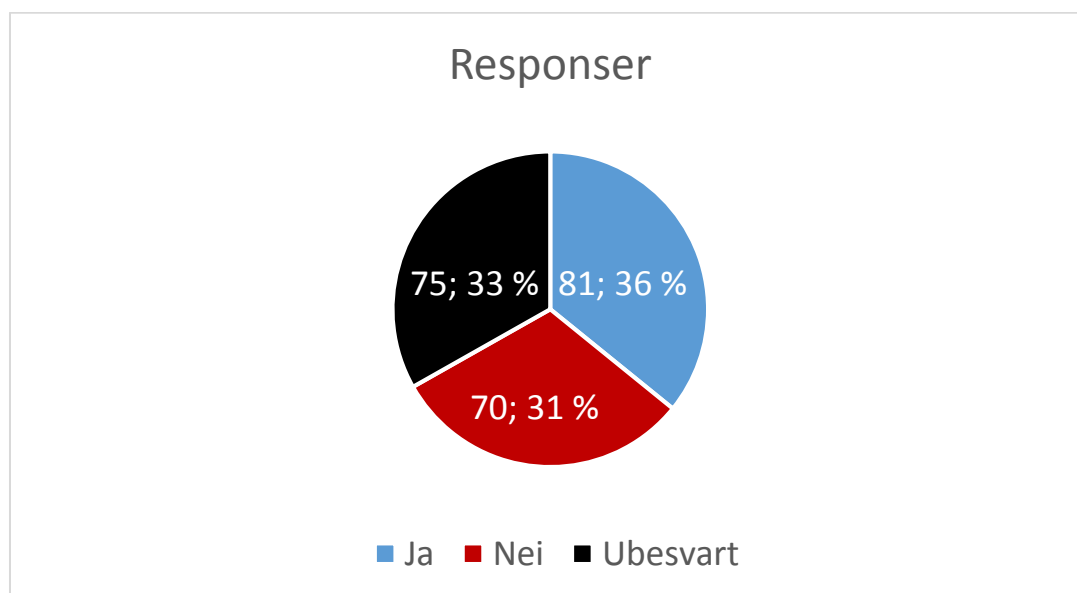
tatt utgangspunkt i en forskjell på 1.73 SD. På skalaen til de nasjonale prøvene vil det si at vi antok at 8. trinn skalaen lå omtrent 17 poeng over skalaen for 5. trinn.

Videre ble det lagt vekt på å inkludere oppgaver med høy diskrimineringssevne for å få maksimal informasjonen fra koblingstesten, samt en representativ fordeling av fagområdene for at koblingstesten skulle reflektere de nasjonale prøvene i størst mulig grad. Siden noen av oppgavene som var designet for 8. trinn kunne bli oppfattet som svært vanskelig eller uforståelig, og dermed demotiverende for elever på lavere trinn, ble oppgaver med høy opplevd vanskegrad (oppgaver som krever mye kognitivt arbeid, eksempelvis oppgaver som involverer dekodning av mye informasjon) brukt i minst mulig grad, og oppgaver med høy faktisk vanskegrad ble forsøkt etterfulgt av en eller flere lette oppgaver. For å unngå påvirkning fra en eventuell posisjonseffekt ble oppgavens posisjon bevart i så stor grad det lot seg gjøre, og ingen oppgaver hadde et avvik på mer enn +/-5 plasser fra posisjonen oppgaven hadde i de ordinære prøvene. For å minimere endring i påvirkningseffekten en oppgave kunne ha på den neste oppgaven, (*carry over effect*) ble det også benyttet klynger med oppgaver på 4-8 oppgaver. Totalt ble 23 oppgaver fra 5. trinns prøven og 25 oppgaver fra 8. trinns prøven benyttet, noe som ga en testlengde på 48 oppgaver i koblingstesten.

Datamateriale

Denne rapporten er basert på datamateriale fra tre forskjellige kilder. To av datasettene er resultatene fra de ordinære gjennomføringene av nasjonale prøver i regning for 5. og 8. trinn. Resultatene fra elevene som tok ankerprøven er ikke inkludert. Disse to kildene er å se på som tilnærmet den fulle populasjonen. Vi angir derfor ikke utvalgsfeil for disse trinnene.

Den siste datakilden består av prøvene administrert til utvalgene av elever fra 6. og 7. trinn. Sommeren 2018 ble skoledata fra GSI benyttet for å trekke utvalg som ville gi oss omtrent 1000 elever på både 6. og 7. trinn. Før trekningen antok vi en positiv responsrate på 25% fra skolene.



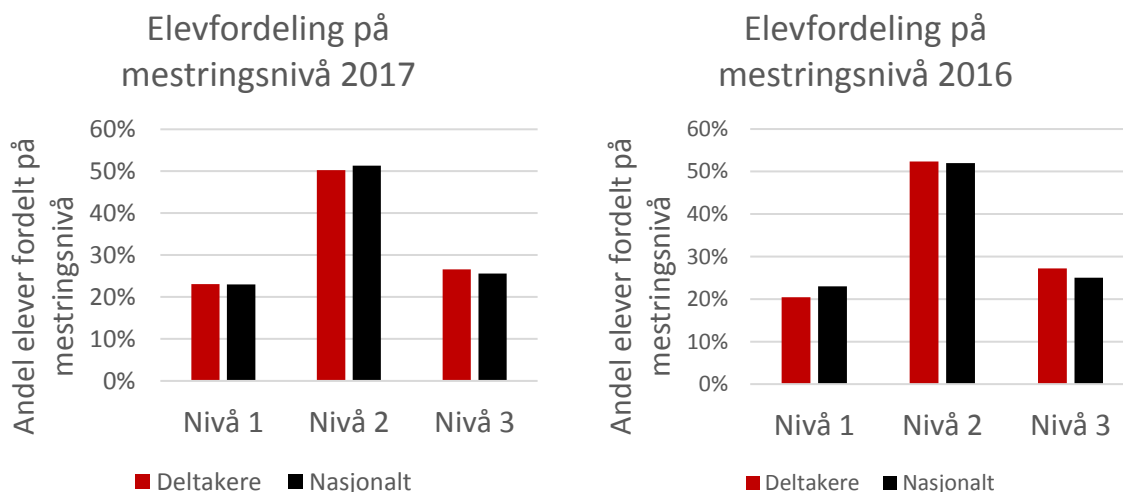
Figur 4: Responsrate fra inviterte skoler. 3 skoler meldte kort tid før gjennomføring at de ikke hadde mulighet til å gjennomføre prøven. 7 skoler som hadde meldt de skulle delta, deltok ikke.

I trekkingen ble skolene vektet etter antall elever på det aktuelle trinnet. For å redusere den administrative byrden ble kun skoler med mer enn 15 elever på det aktuelle trinnet invitert til å delta i studien. 116 skoler ble invitert til å delta på 6. trinn og 117 skoler ble invitert til å delta på 7. trinn. Av disse skolene ble 7 invitert til å delta med elever på både 6. og 7. trinn. Totalt ble det invitert 226

skoler, og responsraten kan ses over i figur 4. Av disse skolene svarte 81 skoler positivt på invitasjonen om å delta, og av disse deltok 71 skoler på gjennomføringen. Av deltakerskolene deltok 22 skoler med både 6. og 7. trinn⁷. På 6. trinn deltok 1718 elever på 47 forskjellige skoler. På 7. trinn deltok 1522 elever fra 46 skoler. Selv om utvalget vårt oversteg målet vårt var antall elever som deltok i nedre sjiktet av hva vi forventet gitt antall skoler og klasser som vi gjennom kommunikasjon med deltakerskolene antok at skulle delta.

Vi kan ikke slå fast at utvalget er representativt for populasjon. For det første ble skolene invitert til å delta med opptil 2 klasser for å spre belastningen av studien utover flere skoler. Vi hadde lagt opp til å trekke to tilfeldige klasserom fra skolene med tre eller flere parallellklasser, men det kom tydelig fram fra responsen fra noen av skolene at de selv ønsket å velge ut den eller de klassene som skulle delta. Noen skoler uttrykte også ønske om å delta med hele trinnet. For å sikre en så høy deltakelse som mulig, fant vi det nødvendig å være imøtekommende for skoler med slike ønsker. For det andre er det fravær innen den enkelte skole som er utenfor vår kontroll. Skolene ble instruert til å gjennomføre prøvene på samme måte som vanlige nasjonale prøver, både når det gjelder test-situasjonen og når det gjelder hvilke elever som burde fritas, men vi har ikke samlet inn informasjon som kan belyse avviket i antall elever eller avvik fra normal gjennomføring.

Videre var det også frivillig for skolene å delta i studien, noe som gjør det utfordrende å undersøke hvor godt utvalgene for 6. og 7. trinn representerer en nasjonal fordeling på skolenivå. Figur 5 viser fordelingen på de tre ferdighetsnivåene for deltakerskolene i tilleggsundersøkelsen den gangen disse elevgruppene deltok i den ordinære prøven på 5. trinn. I figuren sammenliknes disse fordelingene med de nasjonale fordelingene samme år, og vi ser at fordelingene er svært likes. Det viser at utvalget på skolenivå synes å være representative når det gjelder nivået i regneferdigheter.



Figur 5: Fordeling av elever på de forskjellige mestringsnivåene i 5. trinn for deltaker skolene kontra nasjonalt nivå.

Størrelsen på utvalget gjør det imidlertid mulig å få robuste estimater for oppgaveparametere, noe som ble ansett som en prioritet innenfor studiens rammer. På bakgrunn av dette vil vi fokusere på 5. og 8. trinn når vi rapporterer på populasjon.

⁷ At dette tallet er høyere enn de syv skolene som faktisk ble invitert til å delta med begge trinnene, skyldes at noen skoler meldte tilbake at de ønsket å delta med begge trinnene. I disse tilfellene vurderte vi det som mer hensiktsmessig å gi disse skolene en slik mulighet.

I tabell 5 vises det hvor mange elever som gjennomførte de forskjellige prøveformene på de forskjellige trinnene. Til sammenlikning blir det benyttet 6% (omtrent 3600) elever for den horisontale lenkingen av de nasjonale prøvene i regning.

Tabell 5: Antall elever inkludert i datasettet.

Trinn	Prøve	Antall	Totalt antall per trinn
5. trinn	Ordinær 5. trinns prøve	59125	59125
6. trinn	Ordinær 5. trinns prøve	899	1718
	Koblingstest	819	
7. trinn	Koblingstest	758	1522
	Ordinær 5. trinns prøve	764	
8. trinn	Ordinær 5. trinns prøve	59125	59125

Analysemetode

I resultatene blir det presentert to typer resultater. Den første delen er en enkel deskriptiv oversikt over hvordan de forskjellige trinnene gjorde det på de forskjellige oppgavegruppene. Den neste delen, og hovedformålet med denne rapporten, gir en innledende oversikt over skaleringen av de nasjonale prøvene i regning. Flere og mer detaljerte analyser vil være nødvendige for å validere lenkingen, og disse vil bli levert med rapport 2b i henhold til avtalen.

For å analysere resultatene vil det her bli benyttet en såkalt *multigroup IRT-model* (Muthén & Lehman, 1985)⁸. For å gjøre resultatene sammenliknbare med de nasjonale prøvene vil det bli benyttet en 2-parameter logistisk modell (2PL). Dette er den samme metoden som benyttes ved ordinære analyser⁹. Videre tas det utgangspunkt i populasjonen på 5. trinn når vi skalerer prøvene. Det vil si at 5. trinn (i 2018) definerer vår baseline skala. Vi gjennomfører dermed en forenkling ved å rapportere på en 50/10¹⁰ skala definert av årets 5. trinns populasjon, i motsetning til den ordinære skalaen som er definert fra 2014 populasjonen. Det er rimelig å anta at dette ikke påvirker analysene som presenteres her i nevneverdig grad siden det var liten endring i ferdighetsnivået og spredningen til 5. trinn i 2018 sammenliknet med basisåret 2014.

Ved en endelig analyse med formål å forene skalaene for de nasjonale prøvene i regning anbefaler vi å kalibrere resultatene fra koplingstesten gjennom en analyse som inkluderer data fra samtlige prøver gjennomført siden 2014 (såkalt *concurrent calibration*). Tilsvarende anbefaling for den horisontale lenkingen ser vi også fra Björnsson (2018).

Det vil i denne delen av rapporten bli referert til fire grupper med oppgaver, som vist i tabell 6.

Tabell 6: Oversikt over hvilke prøver de forskjellige gruppene med oppgaver var en del av. Rekkefølgen på nummereringen er valgt for å gi enklere visuell fremstilling i resten av rapporten.

	Inkludert i koblingstesten	Ikke inkludert i koblingstesten
Nasjonal prøve i regning for 5. trinn	Oppgavegruppe 2	Oppgavegruppe 1
Nasjonal prøve i regning for 8. trinn	Oppgavegruppe 3	Oppgavegruppe 4

⁸ R-pakken «mirt» er brukt for analysene i denne delen av rapporten, og da spesifikt «multipleGroup» funksjonen (Chalmers, 2012).

⁹ I analysene som ble rapportert til de deltakende skolene ble det anvendt en én-parameter modell.

¹⁰ 50/10 skala referer til måten dagens skalaer for de nasjonale prøvene er definert. Gjennomsnittet er i en slik skala definert som 50 poeng og et standardavvik er definert som 10 poeng.

Resultater

Som vist i tabell 6 ser vi at trinnene, som forventet, presterer gradvis bedre på alle gruppene med oppgaver. Fra fordelingen i tabellen under ser man en stor forskjell i vanskegraden på oppgavegruppene som ble gitt til de forskjellige elevene. For 6. trinn fikk den ene halvdel av elevene en prøve bestående av oppg.gr. 1 og 2, og den andre en prøve bestående av oppg.gr. 2 og 3. Som vist, er det over 20% forskjell i andel riktig på oppgavene i oppg.gr 1 og 3. Tilsvarende for 7. trinn fikk den ene halvdel av elevene en prøve bestående av oppg.gr. 2 og 3, og den andre en prøve bestående av 3 og 4. Her ser vi en enda større forskjell mellom prøvene, da forskjellen i andel riktig for 7. trinn på oppg.gr. 2 og 4 er over 25%. Heldigvis kan vi undersøke i hvilken grad forskjellen i prøvenes vanskegrad påvirker hvordan elevene gjør det på de enkelte oppgavene. Denne potensielle kilden til lenkefeil vil bli fulgt opp i den neste rapporten (rapport 2b i oppdraget).

Tabell 7: Oversikt over prosent riktig for elever på de forskjellige studiene på de forskjellige gruppene av oppgaver.

	Oppgavegruppe 1	Oppgavegruppe 2	Oppgavegruppe 3	Oppgavegruppe 4
5. trinn	57,6 %	53,3 %	N/A	N/A
6. trinn	63,5 %	60,8 %	41,4 %	N/A
7. trinn	N/A	69,5 %	51,2 %	42,9 %
8. trinn	N/A	N/A	56,8 %	49,2 %

Fra resultatene i tabell 7 ser vi en klar progresjon fra trinn til trinn. Det er derimot nødvendig med en annerledes statistisk tilnærming hvis vi skal kunne gjøre mer presise sammenlikninger av resultatene fra prøvene på 5. og 8. trinn, og av den grunn er resultatene som presenteres videre gjort i et IRT-rammeverk. Denne tilnærmingen gjør det mulig å tolke prøveresultatene på en såkalt intervall-basert skala¹¹, et nødvendig kriterium hvis vi ønsker å sammenlikne resultater fra prøver som måler ferdigheter på forskjellige nivåer.

¹¹ En intervall-basert skala har blant annet den egenskapen at en endring i f.eks. 3 poeng fra 40 til 43 betyr det samme som en endring fra 60 til 63. Fra råskår (antall riktig på en prøve) kan vi ikke si det samme av flere grunner. Stevens (1946) gir en leservennlig introduksjon til forskjellige typer psykologiske skalaer som er aktuelle i en utdanningssammenheng.

Skalapoeng fra 8. trinn til skalaen for 5. trinn

Ved å tilpasse en IRT-modell til alle resultatene på de nasjonale prøvene i regning i 2018, inkludert resultatene fra den utvidete testingen for 6. og 7. trinn får vi muligheten til å koble sammen skalaene til prøvene for både 5. og 8. trinn. Koblingen som er gjort i denne rapporten er som nevnt, basert på skalaen til 5. trinn, det vil si at vi definerer gjennomsnittet til 5. trinn til å være 50 poeng og et standardavvik til 10 poeng. Tabell 8 viser at gjennomsnittet til 8. trinn er omtrent 12 poeng over 5. trinn. Dette var en mindre forskjell enn vår forventning om en forskjell på rundt 17 poeng.

Tabell 8: Populasjonsparametere for 5-8 trinn. Vi vil understreke at parameteren for 6. og 7. trinn ikke er basert på representative utvalg.

	5. trinn	6. trinn	7. trinn	8. trinn
Gjennomsnitt	50	53,5	58,4	61,8
Spredning	10	10,5	10,8	12,2

Tabell 8 viser også at spredningen i skåren er større for 8. trinn enn for 5. trinn. Med utgangspunkt i årets 8. trinn populasjon viser tabell 9 konvertering av skillene mellom mestringsnivåene for 8. trinn til skalaen for 5. trinn.

Tabell 9: Konvertering av skalaen for 8 trinn til skalaen for 5. trinn

Gammel skala	Ny skala
36	44,7
44	54,5
56	69,2
63	77,7

Gitt konverteringen i tabell 9, viser tabell 10 samsvaret mellom mestringsnivåene til de nasjonale prøver for 5. og 8. trinn. Her ser vi at mestringsnivåene 1-3 for 5. trinn i stor grad er overlappende med de tre første mestringsnivåene som benyttes for å rapportere resultater for 8. trinn.

Tabell 10: Oversikt over skillene mellom mestringsnivåene til nasjonale prøver for 5. og 8. trinn plassert på felles skala. Avviket i skillet mellom mestringsnivå 2 og 3 for 8. trinn skyldes antall viste siffer i tabell 9.

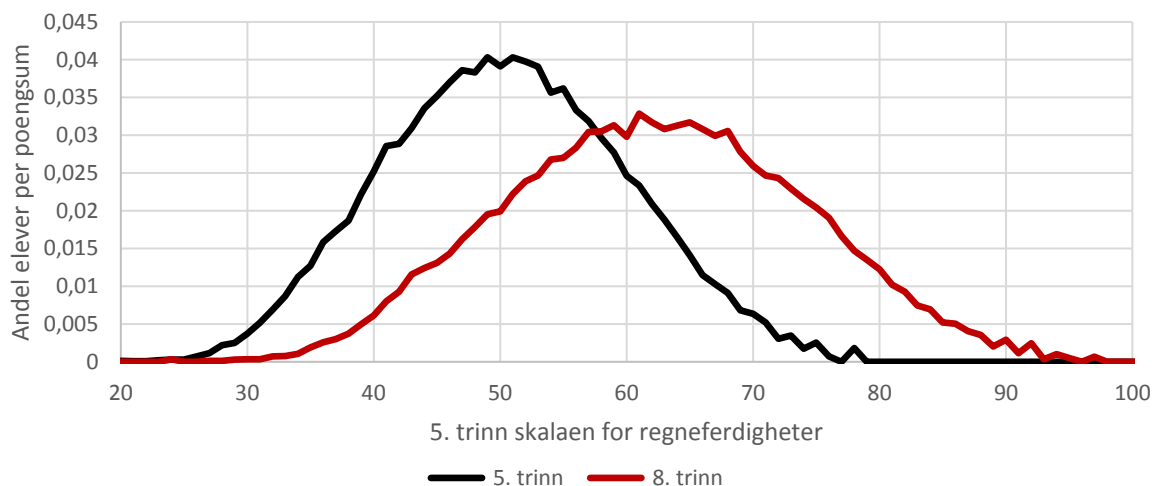
5. trinn		8. trinn	
Under 42	Mestringsnivå 1	Under 45	Mestringsnivå 1
43 til 56	Mestringsnivå 2	45 til 54	Mestringsnivå 2
Over 57	Mestringsnivå 3	54 til 69	Mestringsnivå 3
N/A	N/A	69 til 78	Mestringsnivå 4
N/A	N/A	Over 78	Mestringsnivå 5

For å konvertere 8.trinns skalapoeng til skalaen for 5. trinn benyttes følgende linear transformasjon:

$$5. \text{ trinn skalapoeng} = \frac{8. \text{ trinns skalapoeng} - 50}{10} * 12,2 + 61,8$$

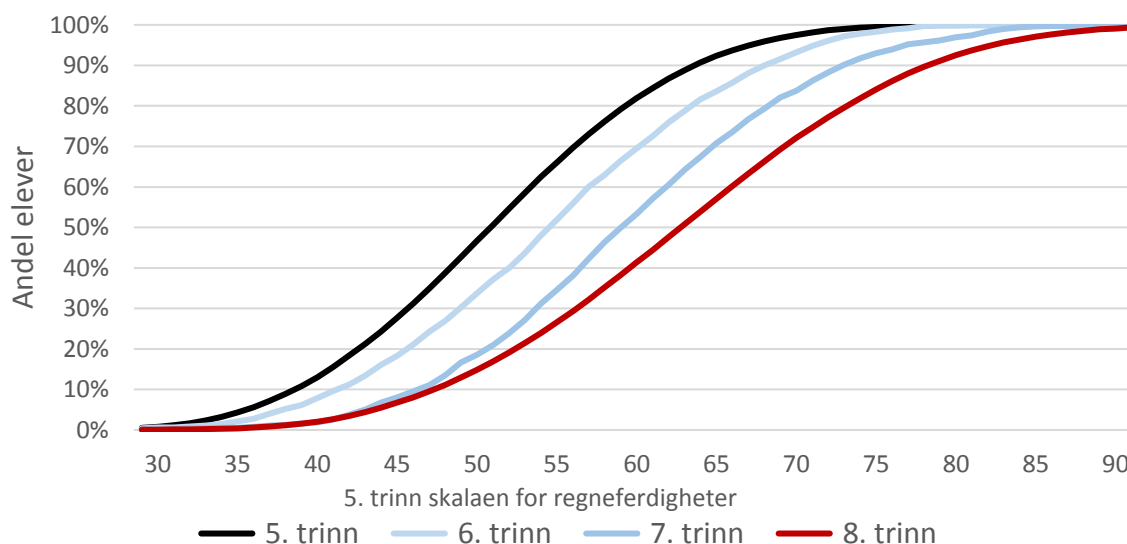
Sammenlikning av regneferdigheter på 5. og 8. trinn

Med utgangspunkt i skaleringen mulighet til å sammenlikne hvordan ferdighetene fordeler seg mellom trinnene på en felles skala. I figur 6 vises fordelingene av elever på 5. og 8. trinn. Som fastslått over at det er en forventet og tydelig høyere grad av regneferdighetene i 8. trinn sammenliknet med 5. trinn. I tillegg kan man i denne figuren tydelig se at spredningen også er større for 8. trinn i den lenkede skalaen.



Figur 6: Fordeling av faktiske skalapoeng for elevene på 5. og 8. trinn på skalaen for 5. trinn.

I figur 7 er den samme fordelingen representert kumulativt. Her kan vi se at det er en markant overlapp mellom 5. og 8. trinn til tross for 3 år med skolegang. Totalt ligger 15% av 8. trinn elevene under 50 poeng, og 13% av elevene på 5. trinn ligger over 62 poeng (som er gjennomsnittet for 8. trinn).

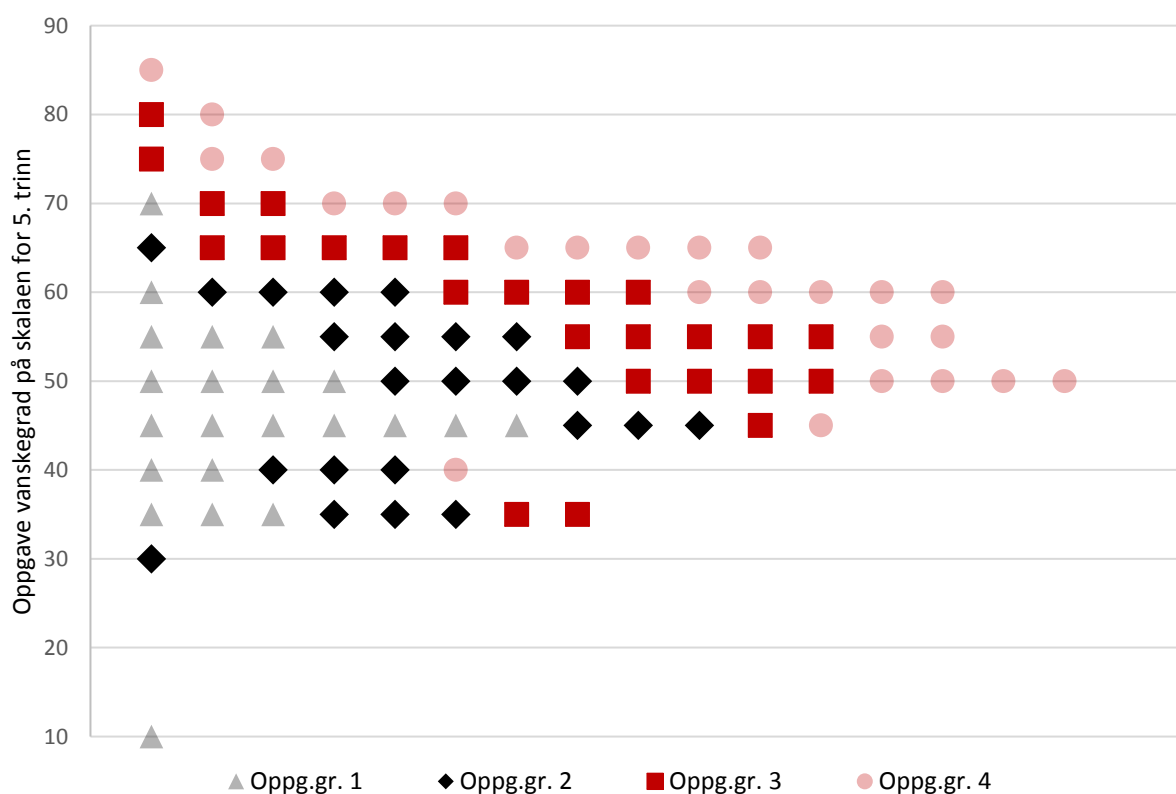


Figur 7: Kumulativ fordeling av regneferdigheter for elever på 5. og 8. trinn. 6. og 7. trinn er også tatt med for å illustrere overlappen mellom trinnene, men vi vil igjen understreke at det er usikkerheter knyttet til representativiteten til utvalgene for 6. og 7. trinn.

Sammenlikning av regneprøvene for 5. og 8. trinn

Ut fra enkle statistiske kriterier fastslår vi at koblingstesten (oppg.gr 2 og 3) er godt tilpasset ferdighetsnivået til 6. og 7. trinn. I figur 8 vises det hvordan oppgavene i de forskjellige gruppene fordelte seg i forhold til vanskegrad. Oppgavene er sortert etter gruppetilhørighet, og stablet bortover, hvor hver oppgave er representert med et punkt. Som vi kan se var det en god spredning i vanskelighetsgraden til koblingstesten (oppg.gr. 2 og 3), samtidig som hovedvekten ligger rundt det gjennomsnittlige ferdighetsnivået til 6. og 7. trinn. Vi ser også at oppg.gr. 2 har en markant høyere vanskegrad enn oppg.gr. 1, og at oppg.gr. 3 består av flere lettere og færre vanskeligere oppgaver enn oppg.gr. 4. Vi har ikke undersøkt i hvilken grad oppgaveparameterene fra piloteringen var stabile.

Videre viser figur 8 at det er en markant overlapp mellom prøvene som utvikles for 5. og 8. trinn, og oppgavene har en god fordeling med en overvekt av oppgaver rundt forventet ferdighetsnivå og med en god spredning av oppgavenes vanskegrad. Til sammen gir en slik fordeling av vanskegrader at man måler godt i midten av fordelingen hvor de fleste elevene befinner seg, samtidig som at man har oppgaver som også får informasjon om elevene i nedre og øvre ende av skalaen.



Figur 8: Oversikt over vanskegraden til oppgavene i de nasjonale prøvene i regning for 5. og 8. trinn i 2018. Oppgavegruppene er definert på samme måte som i tabell 6.

I tabell 11 på neste side er overlappet mellom prøvenes vanskegrad spesielt tydelig, og vi merker oss at det er et stort antall oppgaver i begge prøvene som har en vanskegrad mellom 50 og 60 skalapoeng. Denne typen oppgaver er det behov for i prøven for både 5. og 8. trinn. Vi anser det som sannsynlig at oppgaver i dette vanskegradsintervallet vil fungere godt i begge prøvene basert på disse innledende analysene.

Tabell 11: Antall oppgaver i prøvene for 5. og 8. trinn fordelt på vanskegradintervaller.

	NPR5	NPR8
>70	1	7
60-70	2	18
50-60	16	17
40-50	17	5
<40	9	3
Totalt	45	50

Konklusjon

Innledende analyser indikerer at den vertikale lenkingen av de nasjonale prøvene kan gjøres på et solid grunnlag uten utvikling av ytterligere test-materiale designet for å teste 6. og 7. trinn. Vi er optimistiske både når det gjelder et kosteffektivt design som kan koble de nasjonale regneprøvene og med tanke på senere vedlikehold av koblingen av mellom de prøvesystemene.

Fra skaleringen ser vi at overlappet mellom 5. og 8. trinn er noe større enn vi forventet. Dette kan muligens forklares med at vi brukte TIMSS-resultater som grunnlag for forventet forskjell mellom 5. og 8. trinn. I TIMSS tester man matematiske ferdigheter, og dette konstruktet skal reflektere læreplanene i matematikkfaget til deltakerlandene på de respektive trinnene som testes. Regneferdigheter, som testes i de nasjonale prøvene i regning, er derimot ikke et konstrukt som er direkte knyttet opp mot pensum på et aktuelt trinn. Dette konstruktet trekker i større grad på kunnskap og ferdigheter som både kommer fra flere fag og flere trinn. Denne forskjellen kan være en årsak til at våre prøver for 5. og 8. trinn ligger nærmere hverandre i antall poeng enn vi forventet.

At man i de fire populasjonene finner en betydelig andel elever med tilsvarende prestasjoner i regning tilsier at det både er relevant og nyttig å utvikle én felles og trinnuavhengig mestringsbeskrivelse. Slike mestringsbeskrivelser vil være et nyttig verktøy for lærere for å identifisere den nærmeste utviklingssonen til elever og grupper av elever. Overlappet mellom de fire populasjonene viser også at ambisjonen om å lage nye læreplaner med sterkere vektlegging av progresjon kan følges opp med vurderingsverktøy som direkte modellerer en slik progresjon som en målbar størrelse.

Analysene viser altså foreløpig at et enkelt og kosteffektivt design kan benyttes for å lenke resultatene for nasjonale prøver i regning for 5. og 8. trinn. Designet kan forsterkes ytterligere ved også å ha felles oppgaver i 5. og 8. trinnsprøvene. Vi vil også anbefale at det utvikles et felles helhetlig rammeverk for de to prøvene, inkludert tekst som beskriver et begrenset antall mestringsnivåer (5-8 nivåer). Denne teksten bør utformes spesielt med tanke på å vise hvordan elevens progresjon i grunnleggende ferdigheter typisk kan utvikle seg fra 5. til 8. trinn.

Referanser

- Björnsson, J. K. (2018). Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid. 12(4). doi:<http://dx.doi.org/10.5617/adno.6273>
- Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.
- Muthén, B., & Lehman, J. (1985). Multiple Group IRT Modeling: Applications to Item Bias Analysis. 10(2), 133-142. doi:10.3102/10769986010002133
- Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research*, 47(5), 667-696. doi:10.1080/00273171.2012.715555
- Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16(1), 19-31.
- Rodriguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying Bifactor Statistical Indices in the Evaluation of Psychological Measures. *Journal of Personality Assessment*, 98(3), 223-237. doi:10.1080/00223891.2015.1089249
- Stevens, S. S. (1946). On the Theory of Scales of Measurement. *Science*, 103(2684), 677-680.