UiO : Centre for Educational Measurement
Det utdanningsvitenskapelige fakultet

Utvikling av nasjonale prøver – rapport 2b

Endelig rapport av vertikalt lenkedesign for de nasjonale prøvene i regning 5.-8. trinn.

Rapport til Utdanningsdirektoratet | 02.04.2020

Henrik Galligani Ræder og Rolf Vegar Olsen

Formål med rapporten

Denne rapporten tar for seg de endelige analysene av piloteringen av et vertikalt lenkedesign for de nasjonale prøvene i regning som ble gjennomført høsten 2018 av Nasjonalt senter for matematikk i opplæringen (Matematikksenteret) i samarbeid med Centre for Educational Measurement ved Universitetet i Oslo (CEMO).

Rapporten har til hensikt å svare på spørsmålet «Kan resultater fra de nasjonale prøvene i regning for 5. og 8. trinn plasseres på en felles skala?».

Analysene presentert i denne rapporten støtter oppunder å plassere de nasjonale prøvene i regning for 5. og 8. trinn på en felles skala.

Innhold

Formål med rapporten 2
Oppsummering av oppdraget så langt 4
Det vertikale lenkedesignet 5
Analytisk tilnærming 6
Flergruppe-modeller 6
Flerdimensjonale modeller 7
Resultater 9
Invariante oppgaver på tvers av trinn 9
Trinnspesifikke regneferdigheter 10
Revidert skalering 11
8. trinns skalapoeng på 5. trinns skalaen. 11
Sammenlikning av regneferdigheter fordelt over trinn 12
Konklusjon 14
Referanser 15

Oppsummering av oppdraget så langt

Våren 2018 ble Nasjonalt senter for matematikk i opplæringen (Matematikksenteret) tildelt et tilleggsoppdrag av Utdanningsdirektoratet knyttet til de nasjonale prøvene i regning (NP-R). Formålet med tilleggsoppdraget var å undersøke om NP-R egner seg til å fungere som «[…] robuste indikatorer for enkeltelevers utvikling av grunnleggende ferdigheter i regning over tid.»

Matematikksenteret, i samarbeid med Centre for Educational Measurement ved Universitetet i Oslo (CEMO), implementerte en vertikal lenke-pilot høsten 2018 for å undersøke om det er forsvarlig å koble sammen de nasjonale prøvene i regning. Hensikten med piloten var å undersøke om konstruktene som måles av NP-R i regning for 5. og 8. trinn er såpass like at det er forsvarlig å plassere skårene fra prøvene på en felles skala.

Data fra tidligere gjennomførte NP-R ble analysert før lenke-piloten. I tillegg ble det foretatt en gjennomgang av konstruktene som måles i NP-R. Disse analysene var lovende både i forhold til den underliggende psykometriske strukturen til testene, og i forhold til overlappet i det målte konstruktet fra bakgrunnsdokumentasjonen. Dette er tidligere rapportert i «Utvikling av nasjonale prøver - rapport 1».

I lenke-piloten deltok totalt 71 skoler med henholdsvis 1718 elever fra 6. trinn og 1522 elever fra 7. trinn. Omtrent halvparten av elevene på 6. trinn besvarte den ordinære 2018 versjonen av 5. trinn prøven og halvparten av elevene besvarte en egen variant av prøven som vi i det følgende referer til som lenkeprøven. Tilsvarende, i 7. trinn, tok omtrent halvparten av elevene 2018 versjonen av 8. trinns testen, og halvparten den samme lenkeprøven. Lenkeprøven som ble gitt til 6. og 7. trinns elever var laget fra oppgaver i de nasjonale prøvene i regning i 2018 for 5. og 8. trinn.

De første analysene av resultatene fra piloten viste et substansielt overlapp i elevers regneferdigheter som målt av de nasjonale prøvene i regning mellom trinnene, et lovende element med tanke på å etablere en robust skala som strekker seg på tvers av trinn. Det var også et substansielt overlapp i vanskegraden til oppgavene, et annet element som gjør det lovende med tanke på etableringen av en felles skala. Designet, utvalget og resultatene fra disse første analysene er tidligere rapportert i «Utvikling av nasjonale prøver – Rapport 2a».

I denne rapporten fokuseres det på om det er psykometrisk forsvarlig å koble sammen regneprøvene i 5. og 8. trinn. Modellene som benyttes er under det statistiske paradigmet Item Response Theory (IRT), spesifikt flergruppe-modeller og flerdimensjonale modeller.

Tilnærmingen kan best beskrives som en undersøkelse av følgende spørsmål:

1. Måler 5. trinn oppgavene det samme i 5., 6. og 7. trinn?
2. Måler 8. trinn oppgavene det samme i 6., 7. og 8. trinn?
3. For oppgavene i lenkeprøven: Måler 5. trinn oppgaver og 8. trinn oppgaver det samme i 6. og 7. trinn?

Hvis svarene på alle disse spørsmålene er «ja» kan vi konkludere at de nasjonale prøvene kan brukes til å lage vertikale skalaer.

Det vertikale lenkedesignet

Lenkedesignet som ble implementert høsten 2018 hadde som formål å undersøke om de nasjonale regneprøvene var egnet til å utvikle en felles skala for regneferdigheter fra 5. til 8. trinn. For å undersøke om prøvene, og oppgavene som inngår i dem, var egnet til å lage en vertikal skala var det nødvendig å undersøke hvordan prøvene og oppgavene fungerte i trinnene mellom 5. og 8. trinn. Videre ønsket vi å legge til rette for å kunne koble sammen skalaene for 5. og 8. trinn.

For å oppnå begge disse målene ble et representativt utvalg skoler i Norge invitert til å delta i studien. Skolene ble valgt ut gjennom en vektet stratifisert trekning av alle skoler i Norge som hadde mer enn 15 elever på de aktuelle trinnene. Totalt ble 226 skoler invitert, og av disse deltok 71 skoler spredt rundt om i Norge. De 71 skolene som deltok var i utgangspunktet invitert til å delta med 1-2 klasser som vi skulle trekke tilfeldig for å sikre representativitet. På grunn av praktiske hensyn ble det opp til skolene å bestemme hvilke og hvor mange klasser som deltok. Utvalget kan dermed ikke regnes som representativt, på grunn av både selvvalg og intern seleksjon. Fra aggregert informasjon om skolene som deltok (se rapport 2a) er det imidlertid ingen grunn til å mistenke at utvalget elever som deltok har store skjevheter.

Resultatene fra den utvidete testingen i 6. og 7. trinn ble kombinert med datamaterialet fra den ordinære gjennomføringen av de nasjonale prøvene i regning, og danner grunnlaget for analysene i denne rapporten. Antall elever i de forskjellige trinnene er oppgitt i tabell 1.

Tabell 1: Oversikt over antall elever som gjennomførte de tre ulike regneprøvene i de forskjellige trinnene.

	5. trinns prøven	*Lenke-prøven*	*8. trinns prøven*	*Totalt*
5. trinn	60591			60591
6. trinn	899	819		1718
7. trinn		758	764	1522
8. trinn			59125	59125
Totalt	**61490**	**1577**	**59889**	**122956**

For å koble sammen 5. og 8. trinn prøvene ble det i tillegg utviklet en lenkeprøve. Denne prøven besto av oppgaver fra de ordinære prøvene for 5. og 8. trinn det samme året (for mer informasjon, se rapport 2a). I 6. trinn tok elevene enten 5. trinn prøven, eller lenkeprøven. Tilsvarende, i 7. trinn tok elevene enten 8. trinn prøven eller lenkeprøven. Antall elever som tok de forskjellige prøvene kan ses i tabell 1.

Siden prøver og oppgaver ble brukt i flere trinn kan vi undersøke om de er invariante på tvers av trinn, et viktig kriterium for at vi skal kunne anta at vi måler det samme i de forskjellige trinnene.

Lenkeprøven er hovedkoblingen mellom 5. og 8. trinn prøvene, og gjennom denne prøven får vi mulighet til å undersøke i hvilken grade oppgavene fra 5. trinn og oppgavene fra 8. trinn måler det samme konstruktet.

Analytisk tilnærming

For å undersøke de tre spørsmålene stilt innledningsvis bruker vi to typer analyser. For å svare på spørsmålene 1 og 2 («Måler 5./8. trinns oppgavene det samme på tvers av trinn?») benytter vi såkalte *flergruppe-modeller* (Muthén & Lehman, 1985). Denne typen modeller kan brukes til å evaluere om et måleinstrument er invariant på tvers av gruppene. I vårt tilfelle vil det si om oppgavene oppfører seg på samme måte på tvers av forskjellige trinn.

Vi bruker også disse slike flergruppe-modeller i den endelige skaleringen, siden tilnærmingen vi benytter i denne rapporten er en såkalt *samtidig kalibrering* (concurrent calibration). Dette er tilnærmingen vi benyttet i rapport 2a, og den vi anbefaler å bruke i en operativ skalering.

For å svare på spørsmålet «Måler 5. og 8. trinn oppgaver det samme i 6. og 7. trinn?» benytter vi såkalte *flerdimensjonale modeller* (Reckase, 2009; Reise, Morizot, and Hays 2007). Disse modellene gjør det mulig å undersøke hvor mange identifiserbare underliggende egenskaper/dyktigheter et instrument måler. I vårt tilfelle vil det si om prøvene måler én overordnet regneferdighet, eller om de måler to litt ulike ferdigheter knyttet til hvert av trinnene 5. og 8.

Et viktig utgangspunkt for valg av analytisk tilnærming er et ønske om å kunne gjøre en skalering som gjør det mulig å rapportere én enkelt dimensjon for regneferdigheter på tvers av trinn. Analysene er gjort i den statistiske plattformen R[1].

Flergruppe-modeller

Såkalte flergruppe-modeller brukes for å kunne evaluere i hvilken grad en prøve og/eller oppgavene i prøven er stabil på tvers av grupper. Det er eksempelvis vanlig å bruke slike analytiske tilnærminger i analyser av data fra de internasjonale undersøkelsene for å avgjøre om skalaer kan brukes for å gjøre sammenlikninger på tvers av land. Et annet eksempel er i de ordinære tekniske rapportene fra de nasjonale prøvene som gjerne inneholder en analyse av såkalt *differential item functioning* (DIF) på tvers av kjønn. Gitt at DIF analysene her er implementert på en ganske annen måte enn i de analysene som rapporteres i de ordinære tekniske rapportene fra prøvene, vil vi i det følgende forsøke å forklare hvordan den metoden vi har benyttet er annerledes, og vi begrunner hvorfor dette er en bedre tilnærming til å identifisere oppgaver med DIF.

DIF angir i hvilken grad oppgaver fungerer likt på tvers av grupper. Sagt enkelt: En oppgaves vanskegrad og diskriminering skal være uavhengig av hvem du er. Det eneste som skal påvirke hvor sannsynlig det er at en elev besvarer en oppgave korrekt, er forholdet mellom elevens overordnete dyktighet for det konstruktet som prøven måler og oppgavens parametre (her vanskegrad og diskriminering). I et eksempel hvor man undersøker kjønns-DIF: Hvis gutter og jenter med samme dyktighet har ulik sannsynlighet for å besvare en oppgave riktig, er det opplagt at oppgaven ikke fungerer slik den skal. Utfordringen med DIF er hvordan man skal operasjonalisere dette i en analyse.

I utgangspunktet har man en prøve med mange oppgaver, og man vet ikke hvor mange eller hvilke oppgaver det er som eventuelt har DIF. Den mest vanlige framgangsmåten – og dette er den som benyttes eksempelvis for å undersøke kjønns-DIF i nasjonale prøver – er å gjennomføre en prosedyre for hver enkelt oppgave ved å benytte en såkalt Mantel-Haenszel metode (M-H). Dette er metoden som er implementert i XCalibre (og i mye annen programvare for å gjøre IRT analyser). Denne metoden bruker en litt annen tilnærming for å finne oppgaver med DIF enn den vi har implementert her. Det som er viktig å påpeke her er at M-H (og andre beslektede prosedyrer for å undersøke DIF)

[1] Pakken «mirt» er brukt for estimering av alle IRT-modeller i denne rapporten (Chalmers, 2012).

tar utgangspunkt i en skala for den underliggende målte ferdigheten basert på hele testen som analyseres.

Hovedproblemet med en slik metode er at man behandler alle de andre oppgavene som om de er en fast referanse som fungerer likt for begge gruppene. Valg av hva man benytter som referanse, har en avgjørende rolle i prosedyren for å avdekke hvilke oppgaver som har DIF. I realiteten er situasjonen som regel den at det er flere oppgaver i en prøve som har DIF. Gjennom å bruke hele prøven som referanse for å avgjøre DIF, slik dette eksempelvis er implementert i M-H, har man egentlig ikke full kontroll over situasjonen: Det som skal være likt på tvers av gruppene er ikke nødvendigvis det. Man har dermed det som fritt oversatt kalles for en «forurenset referanse». Dette har blant annet blitt vist å medføre en høy grad av falske positive DIF klassifiseringer. Altså, at DIF blir rapportert for oppgaver som faktisk ikke har DIF (Kopf, Zeileis, & Strobl, 2015; Lopez Rivas, Stark, and Chernyshenko 2009; Meade & Wright, 2012).

Vi har derfor her valgt en annen tilnærming, spesifikt en tilnærming foreslått av Lopez Rivas et al. (2009). For å gjøre en sammenlikning på tvers av grupper trenger vi fortsatt en referanse. I stedet for å bruke «resten-av-prøven» som referanse i den endelige analyse, gjør vi i stedet en trinnvis tilnærming. Det første steget tilsvarer en typisk DIF-analyse, hvor vi identifiserer oppgaver som med stor sannsynlighet fungerer invariant på tvers av gruppene. En gruppe av disse oppgavene blir så valgt ut til å være vår referanse, eller om man vil, anker, i det neste analysetrinnet. Det vil si at parameterne til oppgavene i dette ankeret tvinges til å være like på tvers av gruppene. Antall oppgaver som blir inkludert i dette ankeret er en avveining mellom risikoen for å inkludere en invariant oppgave i ankeret (hvis ankeret er stort øker denne risikoen), og ønsket om størst mulig statistisk styrke (hvis ankeret er stort øker den statistiske styrken). I analysene i denne rapporten brukes et anker med fem oppgaver. Publiserte simuleringsstudier har vist at dette gir en god balanse mellom risiko og statistisk styrke (Lopez Rivas et al., 2009).

Når ankeroppgavene er identifisert blir parameterne til alle de resterende oppgavene undersøkt på nytt. Hvis parameterne til den enkelte oppgaven er forskjellige nok til at det ikke er forsvarlig å holde dem konstante på tvers av gruppene, konkluderer vi med at oppgaven har statistisk signifikant DIF[2]. Oppgaver som identifiseres som å ha DIF på ulike trinn kan i den påfølgende skalering behandles som om de er ulike oppgaver for ulike trinn.

Det er også viktig å påpeke at denne tilnærmingen endrer tolkningen av hva DIF er. I den typiske fremgangsmåten for å undersøke DIF, vil man som sagt identifisere oppgaver med DIF i forhold til resten av *hele* testen som analyseres. Med den valgte tilnærmingen i denne rapporten ser vi derimot på DIF i forhold til et sett med oppgaver som vi kan være ganske sikre på at er invariante på tvers av gruppene. Vi kan dermed snakke om en reel og absolutt DIF i forhold til det underliggende konstruktet.

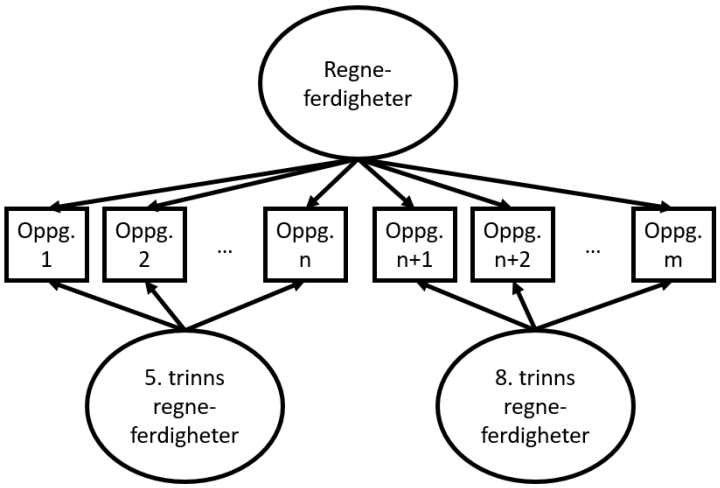
Det er verdt å nevne at selv om det er problematisk å benytte hele testen som en referanse når man undersøker DIF, er dette en tilnærming som er implementert i de fleste IRT-software. Dette gjelder også R-pakken brukt for analysene i denne rapporten, men på grunn av fleksibiliteten knyttet til R som en statistisk plattform er det mulig å implementere metoden beskrevet over manuelt.

Flerdimensjonale modeller

For å undersøke om det er en eller flere underliggende konstrukter som måles av en test er det nødvendig å benytte modeller som tillater flere dimensjoner. Ved å bruke slike modeller kan vi

[2] Evalueringskriteriet for å vurdere om en oppgave viser DIF er en såkalt *Likelihood-ratio test*, hvor det også i signifikanstestingen korrigeres for at man gjør mange sammenlikninger, noe som medfører en predikerbar forhøyet sannsynlighet for å gjøre type I feil. I vårt tilfelle benyttet vi en såkalt Bonferroni korreksjon.

undersøke hvor like dimensjonene vi undersøker er. Det kan likevel være utfordrende å finne en riktig underliggende faktorstruktur for å undersøke dette, og det er ofte avhengig av en rimelig og begrunnet hypotese som man på forhånd kan ha om den underliggende strukturen. I vårt tilfelle, hvor vi ønsker å undersøke om vi har en endimensjonal struktur, har en bi-faktor struktur vist seg å være godt egnet (Holzinger and Swineford, 1937; Reise, 2012). En tilsvarende struktur ble benyttet for å undersøke om fagområder var unike nok til at de trengte å bli modellert som egne dimensjoner i rapport 2a, og denne strukturen er beskrevet mer i detalj der. Den spesifikke strukturen som blir undersøkt i denne rapporten er vist i figur 1, hvor den generelle faktoren representerer dimensjonen *Regneferdigheter*, og de to sub-faktorene representerer henholdsvis dimensjonene *5. trinn regneferdigheter* og *8. trinn regneferdigheter*.



Figur 1: Bi-faktor struktur med to faktorer.

Tilsvarende som i rapport 2a bruker vi evalueringskriteriet *forklart felles varians* (Explained Common Variance – ECV). For å unngå at en endimensjonal skala blir forvrengt av underdimensjoner er det nødvendig at den generelle faktoren forklarer over 80% av ECV (Reise, 2012).

Resultater

I denne delen presenteres først resultatene av analysene som undersøker om oppgavene er invariante på tvers av trinn. I den andre delen presenteres resultatene for analysene som undersøker hvor like konstruktene som måles av 5. og 8. trinns oppgavene er.

Invariante oppgaver på tvers av trinn

Som skissert tidligere definerer vi invariante oppgaver som oppgaver som ikke viser DIF. Slike oppgaver kan vi anta at måler det samme på tvers av trinnene, og dermed egner seg til å bruke til å bygge en vertikal skala. Under, i tabell 2, vises hvor mange oppgaver som viste DIF i 6. og 7. trinn, i forhold til trinnet de i utgangspunktet ble utviklet for.

Tabell 2: Antall oppgaver med signifikant DIF/totalt antall oppgaver brukt i det relevante trinnet, og andel av oppgaver som er invariante i prosent i de forskjellige trinnene oppgavene ble administrert.

	5. trinn	*6. trinn*	*7. trinn*	*8. trinn*
5. trinn oppgaver	**45 oppgaver**	7/45 16 %	5/23 22 %	
8. trinn oppgaver		11/25 44 %	1/50 2 %	**50 oppgaver**

Som det kommer fram fra tabellen var oppgavene i stor grad stabile når de ble brukt i et trinn over og et trinn under for henholdsvis 5. og 8. trinn oppgaver. Spesielt stabile er 8. trinn oppgavene når det ble gitt til 7. trinn elever, hvor kun en av oppgavene viste DIF. 5. trinn oppgavene som ble gitt til 6. trinn var også relativt stabile, om enn ikke på samme nivå.

Når det gjelder oppgaver administrert to trinn unna det intenderte trinnet, er situasjonen motsatt. Her er oppgaver brukt oppover, det vil si 5. trinn oppgavene i 7. trinn fortsatt relativt stabile. Oppgavene som ble brukt nedover er derimot veldig ustabile, og den store andelen med 8. trinn oppgaver med signifikant DIF i 6. trinn er problematisk med tanke på hvor sikre vi kan være på at resultatene stemmer. Siden andel oppgaver med DIF er så pass stor, kan vi ikke utelukke at vi har identifisert feil gruppe med oppgaver som er invariante.

Vår anbefaling er at slike oppgaver likevel skal være inkludert når vi beregner en endelig skala, men at de ikke begrenses til å ha like parametere på tvers av trinnene. I praksis betyr dette at slike oppgaver behandles som om de er to forskjellige oppgaver nå vi kobler sammen prøvene. Dette gjør at vi kan bruke informasjonen som oppgaven gir på det enkelte trinnet, uten at oppgaven forvrenger skalaen[3].

[3] Alternativet som ofte sees i slike sammenhenger, er at oppgaver med DIF tas ut fra den endelige skaleringen. Dette er uheldig fordi det reduserer informasjon (dvs at måleusikkerheten øker), og man vil ha risiko for å «trimme» konstruktet med den konsekvensen at enkelte viktige deler av rammeverket blir underrepresentert

Trinnspesifikke regneferdigheter

For å undersøke om regneferdigheter, slik dette er operasjonalisert i de nasjonale regneprøvene for 5. og 8. trinn, er én felles ferdighet brukte vi en flerdimensjonal IRT modell med en såkalt bi-faktor struktur. Resultatene av disse analysene kan ses i tabell 3. I tillegg til å analysere lenkeprøven, gjorde vi en tilsvarende analyse av 5. og 8. trinn prøvene for å gi et relativt tolkningsgrunnlag.

Tabell 3: Andel varians forklart av bi-faktor modellene dekomponert ved den generelle faktoren og sub-faktorene. Sub-faktorene for prøvene er definert på følgende måte: Oppgaver som ikke var med i lenkeprøven (NP5 unik og NP8 unik), og oppgaver som var med i lenkeprøven (NP5 LP og NP8 LP).

		Andel variansforklart av de de ulike faktorene					*Total varians forklart*	*ECV av generell faktor*
Trinn	*Prøve-form*	*Generell faktor*	*NP5 unik*	*NP5 LP*	*NP8 LT*	*NP8 unik*		
5. trinn	NP5	0,346	0,018	0,009	N/A	N/A	0,373	**0,928**
6. trinn	NP5	0,377	0,017	0,014	N/A	N/A	0,408	**0,924**
	LP	0,338	N/A	0,019	0,018	N/A	0,375	**0,901**
7. trinn	LP	0,350	N/A	0,026	0,017	N/A	0,393	**0,891**
	NP8	0,312	N/A	N/A	0,019	0,014	0,345	**0,904**
8. trinn	NP8	0,360	N/A	N/A	0,009	0,011	0,380	**0,947**

Fra tabellen ser vi at den generelle faktoren i hver av testene forklarer godt over 80 % av den totale forklarte variansen (ECV av generell faktor). I motsetning til resultatene fra DIF-analysen rapportert over ser det ut som at det er 5. trinn oppgavene administrert i 7. trinn som er mest unike. Vi har i disse analysene ikke undersøkt om disse sub-faktorene representerer statistisk signifikante underdimensjoner, da dette krever å implementere en såkalt *bootstrap-analyse*. Vi har ikke prioritert å dette, da det uansett vil være snakk om underdimensjoner som ikke har en substansielle størrelse.

Fra bi-faktor analysene presentert over kan vi konkludere med at oppgavene fra 5. og 8. trinn ser ut til å måle det samme konstruktet, både i 6. og 7. trinn. Sammen med bi-faktor analysene presentert i rapport 1 kan vi videre konkludere med at regneferdigheter som målt ved de nasjonale prøvene i regning oppfører seg som et endimensjonalt konstrukt fra 5. til 8. trinn.

Revidert skalering

I rapport 2a presenterte vi resultater basert på innledende analyser - det vil si før vi hadde avklart om det var trinn-spesifikke dimensjoner og oppgaver med DIF som vi måtte ta hensyn til. Fra resultatene presentert over er det klart at vi ikke trenger å ta hensyn til førstnevnte, men det er nødvendig å ta hensyn til oppgaver med DIF. I denne delen presenterer vi derfor på nytt deskriptiv informasjon for en kombinert skala for regneferdigheter på fra 5. til 8. trinn. I disse analysene er oppgavene som viste DIF i analysene over ikke begrenset til å have de samme parameterne for trinnene de viste DIF i.

5. trinn 2018 er her benyttet som referansepopulasjon, med gjennomsnitt på 50 poeng og en spredning på 10 poeng[4]. Det vil si at resultatene for alle trinnene presenteres på en felles skala, og at denne skalaen refererer til 5. trinn skalaen som brukes i dag for de nasjonale regneprøvene. For komplettthetens skyld har vi inkludert datamaterialet fra 9. trinn 2018 i analysene.

8. trinns skalapoeng på 5. trinns skalaen.

Under vises populasjonsparameterne til de forskjellige trinnene plassert på 5. trinn skalaen. Sammenliknet med de innledende analysene ser vi at forskjellene er minimale for 8. trinn. Som man kan se fra tabellen blir spredningen noe større i de høyere trinnene.

Tabell 4: Populasjonsparametere for 5-9 trinn. Vi vil understreke at parameteren for 6. og 7. trinn ikke er basert på representative utvalg. Det er ingen feilmarginer for 5. trinns parameterne siden de ble definert, ikke estimert. For å beregne feil-estimatene har vi brukt en såkalt 'sandwich' tilnærming (Chalmers, 2012).

	5. trinn	*6. trinn*	*7. trinn*	*8. trinn*	*9. trinn*
Gjennomsnitt (SE)	50 (N/A)	53,65 (0,10)	58,55 (0,10)	61,58 (0,02)	65,99 (0,02)
Spredning (SE)	10 (N/A)	11,52 (0,16)	9,7 (0,15)	12,18 (0,04)	13,45 (0,04)

For å konvertere 8. trinn skalapoeng til 5. trinn skalapoeng, kan man gjøre en tilnærming ved å bruke følgende lineare transformasjon:

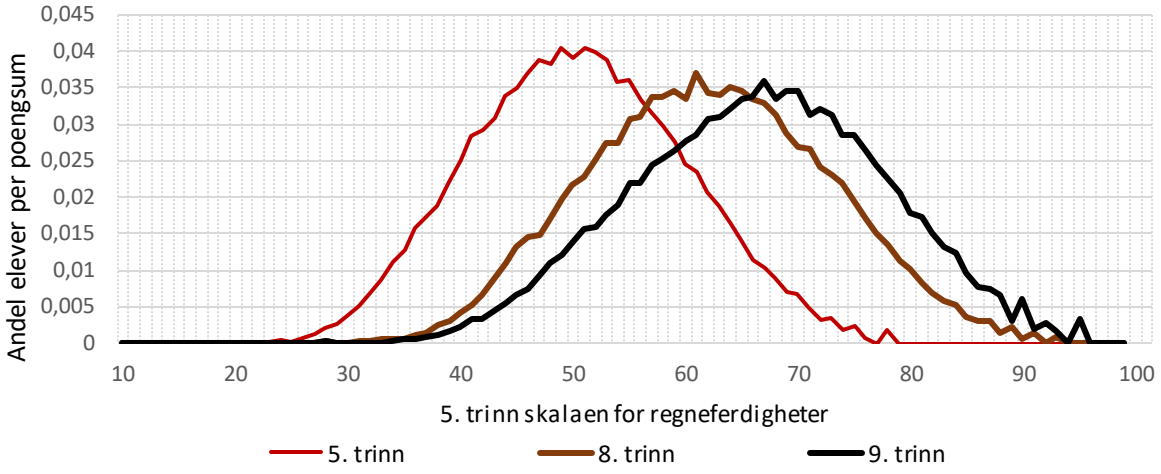
$$5.\,trinn\ skalapoeng = \frac{8.\,trinns\ skalapoeng - 50}{10} * 12{,}18 + 61{,}58$$

For en presis plassering er det nødvendig å gjøre en samtidig kalibrering, men med denne transformasjonen kan man på en enkel måte få en god tilnærming.

[4] Skalaen som benyttes i dag er basert på en annen referansepopulasjon (2016). Kommunikasjon med Julius Björnsson, som er ansvarlig for den horisontale lenkingen, tilsier at vi likevel kan være trygge på at denne tilnærmingen vil gi veldig like resultater. Det har kun vært små og ubetydelige endringer i populasjonsparameterne siden lenkingen ble introdusert.

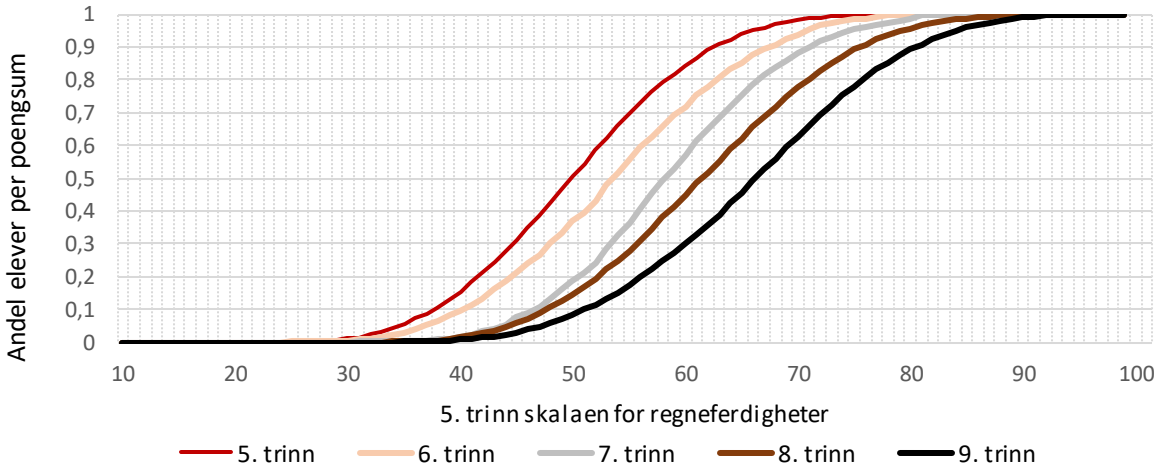
Sammenlikning av regneferdigheter fordelt over trinn

Som populasjonsparameterne viste i forrige avsnitt, er det 11,6 poeng forskjell i den gjennomsnittlige ferdigheten til elevene i 5. og 8. trinn. Videre er det større forskjell i ferdighetene til elevene på 8. trinn. Dette vises også grafisk under i figur 2.



Figur 2: Fordeling av oppnådd skalapoeng for elevene på 5. og 8. trinn på skalaen for 5. trinn.

I figur 3 er den samme fordelingen representert kumulativt. Her kan vi tydelig se at det er relativt liten forskjell mellom nære trinn, men det er også en markant overlapp mellom 5. og 8. trinn til tross for 3 år med skolegang.



Figur 3: 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 utvalget for 6. og 7. trinn ikke er representativt.

I figur 3 er den samme fordelingen vist kumulativt. Fordelen med denne fremstillingen er at det er enkelt å lese av andel av elever som lå under eller over en gitt poengsum. For eksempel ligger 15% av elevene på 8. trinn og 8% av elevene på 9. trinn under 50 poeng, og prester er dermed dårligere enn gjennomsnittet på 5. trinn. 13% av elevene på 5. trinn ligger over 62 poeng, og presterer derfor bedre enn gjennomsnittet på 8. trinn.

En annen måte å se på overlappet i regneferdigheter mellom trinnene er å sammenligne poengintervallene for de forskjellige mestringsnivåene når de plasseres på en felle skala. Mestringsnivåene for 5. og 8. trinn ble i referanseåret definert normativt. For 5. trinn var den

normative fordeling basert på prosentandelene 25/50/25, og for 8. trinn ble 10/20/40/20/10 brukt. I tabell 5 gis en oversikt over hvilke poengsummer disse mestringsnivåene vil tilsvare på en felles skala. Vi har tatt utgangspunkt i de samme prosentandelene, men med resultatene fra 2018 som referanse. Her ser vi at mestringsnivåene 1-3 for begge trinn beskriver omtrent det samme området på skalaen.

Tabell 5: Poengintervaller for de forskjellige mestringsnivåene for 5. og 8. trinn plassert på en felles skala.

5. trinn		*8. trinn*	
Nivå 1	Mindre enn 43	*Nivå 1*	Under 48
Nivå 2	43 til 57	*Nivå 2*	48 til 56
Nivå 3	Mer enn 57	*Nivå 3*	56 til 67
		Nivå 4	67 til 80
		Nivå 5	Over 80

Konklusjon

Innledningsvis stilte vi tre spørsmål som vi har forsøkt å besvare med denne rapporten. For det første spørsmålet, «Måler 5. trinns oppgavene det samme i 5., 6. og 7. trinn?», viser analysene at 5. trinns oppgavene i stor grad var invariante i både 6. og 7. trinn, og vi kan dermed anta at de måler det samme.

For det andre spørsmålet, «Måler 8. trinns oppgavene det samme i 6., 7. og 8. trinn?», viser analysene presentert her at 8. trinns oppgavene kun var invariante i 7. trinn. Dette betyr at vi ikke anbefaler en vertikal skalering basert på 8. trinns oppgavene i 6. trinn, men heller la dem bli estimert fritt for denne gruppen for å benytte oss av den informasjonen som disse oppgavene bidrar med lokalt på trinnet.

For det tredje spørsmålet, «For oppgavene i lenkeprøven: Måler 5. trinn oppgaver og 8. trinns oppgaver det samme i 6. og 7. trinn?», viser analysene presentert her at det var et felles underliggende konstrukt som ble målt av oppgavene fra 5. og 8. trinn i både 6. og 7. trinn. Vi kan dermed konkludere med at regneferdigheter, slik dette blir målt i de nasjonale prøvene, reflekterer et felles overordnet konstrukt, og ikke separate unike konstrukter for de to trinnene.

En del av oppgavene anbefales som sagt å bli estimert separat for ulike trinn (pga. DIF), og dette gjør at koplingen i lenkedesignet blir noe redusert. Til tross for dette er vår konklusjon at piloten har vist at det er empirisk meningsfullt å plassere 5. og 8. trinns prøvene på en felles skala.

For å sikre at en slik skala holder seg stabil gjennomfører flere test-systemer i USA separate kalibreringsstudier med 3-5 års mellomrom. Et annet alternativ er å innføre et sett med felles oppgaver for 5. og 8. trinns prøvene. Vi har ikke undersøkt hvor stabile 5. trinns oppgaver er i 8. trinn, men gitt resultatene for 8. trinns oppgaver administrert i 6. trinn, vil vi ikke anbefale å bruke 8. trinns oppgaver i 5. trinns testen.

Generelt ser vi at overlappet mellom trinnene, både med tanke på ferdigheter og det målte konstruktet åpner for en rekke muligheter. Blant annet kan det være gunstig å utvikle oppgavebanker som er delt på tvers av trinn, både for å redusere arbeidsomfanget knyttet til oppgaveutvikling, og for å ha oppgaver tilgjengelig for flere ferdighetsnivåer. En annen mulighet er å åpne for at lærere og skoler kan benytte oppgaver/prøver ved andre tidspunkter enn de obligatoriske måletidspunktene. En slik mulighet vil gjøre det mulig for skoler å måle progresjonen til grupper med elever over tid, en mulighet som skoler per dags dato ikke har tilgjengelig instrumenter for.

Referanser

Chalmers, R. P. (2012). mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software, 48*(6), 1-29.

Holzinger, K. J., & Swineford, F. (1937). The Bi-factor method. *Psychometrika, 2*(1), 41-54. doi:10.1007/bf02287965

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor Selection Strategies for DIF Analysis:Review, Assessment, and New Approaches. *Educational and Psychological Measurement, 75*(1), 22-56. doi:10.1177/0013164414529792

Lopez Rivas, G. E., Stark, S., & Chernyshenko, O. S. (2009). The Effects of Referent Item Parameters on Differential Item Functioning Detection Using the Free Baseline Likelihood Ratio Test. *Applied Psychological Measurement, 33*(4), 251-265. doi:10.1177/0146621608321760

Meade, A. W., & Wright, N. A. (2012). Solving the measurement invariance anchor item problem in item response theory. *Journal of Applied Psychology, 97*(5), 1016-1031. doi:10.1037/a0027934

Muthén, B., & Lehman, J. (1985). Multiple Group IRT Modeling: Applications to Item Bias Analysis. *10*(2), 133-142. doi:10.3102/10769986010002133

Reckase, M. D. (2009). Multidimensional Item Response Theory Models. In *Multidimensional Item Response Theory* (pp. 79-112). New York, NY: Springer New York.

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. doi:10.1007/s11136-007-9183-7

Reise, S. P. (2012). The Rediscovery of Bifactor Measurement Models. *Multivariate Behavioral Research, 47*(5), 667-696. doi:10.1080/00273171.2012.715555