UiO : Universitetet i Oslo
Institutt for lærerutdanning og skoleforskning

Matematikkeksamen gjennom tre år

Psykometrisk analyse av eksamen i matematikk for 10. trinn

fra årene 2017, 2018 og 2019

Del av oppdraget:

Kvalitetssikring av eksamener og prøver

Julius K. Bjørnsson

Institutt for lærerutdanning og skoleforskning/EKVA

Januar 2020

Forord

Denne rapporten er en del av oppdraget Kvalitetssikring av prøver og eksamen, som ILS – EKVA gjennomfører på oppdrag fra Utdanningsdirektoratet. Det er mange personer som har bidratt til dette arbeidet, først og fremst Gregorios Brogstad som tok initiativet til denne studien, har koordinert datainnsamlingsarbeidet og sørget for at alt fungerte som det skulle. Eksamensnemden har bidratt med nye oppgaver til kalibreringsprøven og sørget for at de hadde samme kvalitet som de ordinære eksamensoppgavene. Hilde Olsen fra UDIR har bidratt i tekniske diskusjoner, lest gjennom resultatene og lagt fram mange gode forslag til endringer og forbedringer. I tillegg har Per Kristian Larsen Evjen og Marianne Elgen Bøhler lest igjennom rapporten, og bidratt med mange gode forslag til endringer og forbedringer. Til slutt må nevnes at Anna Eriksen på ILS/EKVA har foretatt en grundig språkvask av den endelige rapporten.

Alle disse menneskene får stor takk for sine bidrag, uten dem hadde dette ikke blitt gjort.

JKB

Innholdsfortegnelse

Figurer:

Tabeller:

Sammendrag

Denne rapporten beskriver resultatene fra et treårig prosjekt om eksamen i matematikk for 10. trinn. Hensikten har vært å undersøke om eksamen endret vanskegrad fra år til år, samt å se på egenskapene til enkeltoppgaver og eksamen i sin helhet. For å få til dette ble det utviklet en såkalt kalibreringsprøve (K-prøve), det vil si en digital øvelsesprøve i matematikk som alle elever fikk tilbud om å gjennomføre omtrent en måned før selve eksamen. Resultater fra elever som tok denne prøven og som i tillegg kom opp til eksamen, ble brukt i denne studien. Kalibreringsprøven ble holdt hemmelig og brukt i årene 2017, 2018 og 2019. Hensikten var å anvende den som ankerprøve for å lenke sammen eksamensresultater fra disse tre årene. Data fra de tre årene ble samlet inn fra alle elever som hadde tatt både K-prøven og eksamen i matematikk, og disse dataene ble samkalibrert ved hjelp av IRT slik at resultater fra de tre årene ble plassert på samme skala.

Resultatene tyder på at eksamen i matematikk for 10. trinn ikke har endret vanskegrad merkbart i sin helhet, men at de ulike delene av eksamen er ganske ustabile eller variable fra år til år. Del 1 av eksamen ble litt lettere i løpet av de tre årene, mens Del 2 ble litt vanskeligere. Ingenting tyder på at elevene i disse tre årene hadde forskjellig ferdighet. IRT-analysen viser i tillegg at eksamensdelene leverer forskjellig informasjon hver gang og at toppen av informasjonskurvene flytter seg ganske betydelig på ferdighetsskalaen fra år til år.

Eksamen viser seg å inneholde meget bra dikotome oppgaver, men av de oppgavene som hadde graderte svar, var det kun 34 % som fungerte på en adekvat måte. Hvordan vi arbeider med å utvikle denne typen oppgaver må derfor vurderes og forbedres.

Rapporten konkluderer med at den eneste måten å gjøre eksamen mer stabil på, vil være å pilotere oppgavene og ta i bruk moderne psykometriske metoder både i utviklingsfasen og i analysen av resultatene. I tillegg har erfaringen med en samkalibrering av kalibreringsprøven og eksamen vist at det er mulig å få til sammenliknbare målinger av kompetanse over tid.

Eksamen har mange gode egenskaper, med mange oppgaver som fungerer bra og lange tradisjoner for hvordan matematisk kompetanse kan måles. Disse fordelene går imidlertid tapt, hvis man ikke oppgraderer både utviklings-, analyse og rapporteringsmetodene.

Innledning

Eksamen i norsk skole bygger på lange tradisjoner om hvordan oppgaver blir laget, gjennomført og behandlet etter at elevene har utført sin del. Men eksamen kan kritiseres for at oppgavene ikke er pilotert, det vil si prøvd ut på forhånd, slik fleste andre store prøvesystemer gjør, og spesielt prøver hvor resultatene er viktige. Eksamen er viktig for norske elever, og bl.a. derfor er den usikkerheten som manglende pilotering fører med seg, en svakhet som det er viktig å peke på og forbedre.

Det er også påfallende at mange, og kanskje spesielt mediene, har tradisjon for å sammenlikne eksamensresultater mellom år, og til og med mellom kommuner og skoler. Dette er uforsvarlig, fordi sammenlikningen mellom kommuner og skoler oftest har blitt gjort uten å ta hensyn til den usikkerheten som ligger i gjennomsnittstall. Rangeringer som ikke sier noe om usikkerheten som ligger i dem, er ikke akseptable. Dette gjelder spesielt når sammenlikningene skjer mellom små grupper, fordi usikkerheten da er større enn ellers. Det å sammenlikne eksamensresultater mellom år er også uforsvarlig, fordi eksamen ikke er lenket eller ankret fra år til år. I tillegg har ikke eksamen det samme antallet oppgaver hvert år, og det foreligger ingen eller få opplysninger om oppgavene er på samme skala, eller om de blir bedømt på en konsistent måte fra år til år.

Ifølge Utdanningsdirektoratet er det blitt gjort en del for å forbedre denne situasjonen, bl.a. er retningslinjer for skåring av oppgavene blitt endret og forbedret. Men oppgavene er fortsatt ikke pilotert eller ankret fra år til år, noe som gjør sammenlikninger over tid umulig.

Hensikten med det prosjektet som denne rapporten omhandler, var først og fremst å undersøke hvor mye eksamen i matematikk for 10. trinn endrer seg i vanskegrad fra år til år. For å få dette til ble det anvendt en IRT-analyse («Item Response Theory») av alle eksamensoppgavene tre år på rad, og resultatene ble lenket til en såkalt kalibreringsprøve (K-prøve) som fungerte som et anker, eller lenke, mellom årene.

En IRT prøvekalibrering ble gjort med data fra eksamen i 2016. Denne analysen viste at mesteparten av oppgavene stort sett fungerte adekvat og at selv om eksamen bygger mest på tradisjoner og ikke er pilotert, så bygger den på et solid kunnskapsgrunnlag og benytter mange oppgaver som reflekterer læreplanen og måler på en bra måte.

Det er likevel noen uløste problemer, hvorav de viktigste er følgende:

- Sammenlikning av vanskegraden til eksamen fra år til år er per i dag ikke mulig.
- Eksamensoppgaver kan ikke gjentas fra år til år dersom de blir publisert etter gjennomføring, noe som gjør lenking og sammenlikninger mellom år umulig.
- Hvis resultatene endres fra et år til et annet, kan det være fordi:
  - *Oppgavene er lettere eller vanskeligere*
  - *Elevenes prestasjoner er blitt bedre eller dårligere*
  - *Sensureringen er endret: strengere/snillere*
  - ***Alle effekter samtidig – mest sannsynlig***

De fleste prøvesystemer som måler utvikling over tid, eller endringer fra et år til et annet, bruker en eller annen form for ankring og ekvivalering eller lenking av skårer fra et tidspunkt til et annet, og fra en prøve til en annen. Uten et slikt system er det ikke mulig å sammenlikne resultater over tid, eller fra eksamen et år til det neste.

Eksamensresultatet baserer seg på en lang kjede av faktorer som alle må være kvalitetssikret for at det skal bli reliabelt og valid. Det første som må sikres, er reliabiliteten til målingen. For eksamen sin del dreier dette seg om to typer reliabilitet: sensorreliabilitet og oppgavesettets (prøvens) reliabilitet. Først når disse to aspektene er på plass, kan validiteten til målingen i sin helhet sikres. Og det er verdt å understreke at ingen kjede er sterkere enn det svakeste ledd. Derfor må hele prosessen fra læreplan/oppgavespesifikasjon og formålsdefinisjon til det endelige resultatet, være av høyest mulig kvalitet.

Om sensorreliabilitet

Dersom mange av oppgavene i matematikkeksamen bygger på sensorbedømminger, er kvaliteten til det endelige resultatet sterkt avhengig av sensorenes reliabilitet. Det er et krav når sensorreliabiliteten bedømmes, at det skal være samsvar mellom ulike sensorer, og at de må bedømme konsistent over tid og på forskjellige steder. Derfor er sensorskolering og trening en sentral del av et system av denne typen, og uten

denne skoleringen kan ikke samsvaret mellom ulike sensorer sikres. Det burde derfor være et krav at alle som sensurerer eksamensbesvarelser, gjennomgår den samme treningen. Dette er ikke tilfellet i dagens system og bør endres. Og sensorene må bruke de samme bedømmingskriteriene, og disse må være så objektive som mulig, for å sikre at samme ferdighet ligger bak samme resultat.

Eksamen bør oppfylle følgende krav:

- Det må brukes en konsistent skala, hvor distansen mellom ulike karakterer er jevn.
- Det må brukes en invariant skala som betyr at målingen må kunne sammenliknes fra sted til sted og fra gang til gang, og være uavhengig av de gruppene som besvarer oppgavene.
- Bedømmingskriteriene og poengene bak karakteren må være så objektive og entydige som mulig.

Sensorreliabiliteten kan vurderes på mange måter. Utdanningsdirektoratet har gjennomført undersøkelser fra et utvalg av sensorer for matematikkeksamen. Disse undersøkelsene tyder på at sensorreliabiliteten er god.

Når sensorreliabiliteten viser seg å være tilfredsstillende, må også oppgavene og skalaen som brukes for rapportering av resultatene evalueres. Først når dette er på plass kan man begynne å tenke på hvordan resultater kan sammenliknes fra år til år, eller fra en eksamen til neste.

Hvis resultater fra to år i nåværende system viser seg å være forskjellige, er det ikke mulig å vite om det er fordi elevene presterer dårligere eller bedre, eller om det er fordi prøven er blitt lettere eller vanskeligere, eller begge deler samtidig.

Endring over tid

Et ankerdesign likt det som er brukt på nasjonale prøver, er av flere grunner ikke mulig fordi:

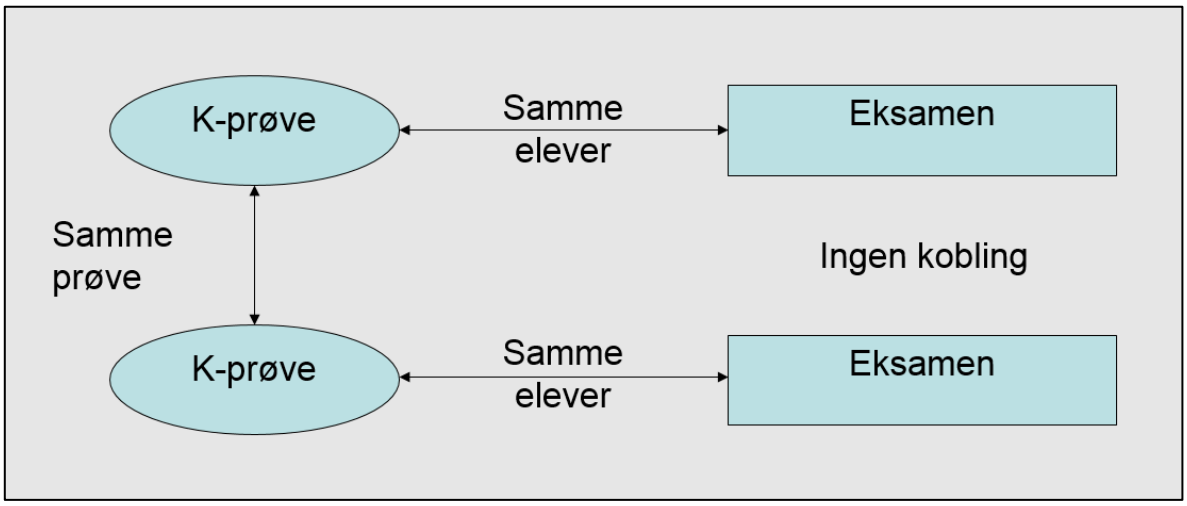
- alle oppgaver blir publisert etter gjennomføring av eksamen
- eksamensoppgavene blir ikke pilotert og kalibrert med moderne metoder (bruker poeng som ikke er sammenliknbare fra år til år, oversatt til tradisjonell karakter 1-6)
- elevbesvarelsene er ikke skåret med en moderne ferdighetsestimering, eller rapportert med en skalert skåre som er sammenliknbar over tid (se f.eks. Tan & Michel, 2011)

Men det er mulig å gjennomføre en prøve med uendret innhold ved siden av eksamen hvert år. Denne prøven kan gi indikasjoner på om elevenes prestasjon har endret seg eller ikke. Da er det overordnet viktig at de elevene som besvarer oppgavene er sammenliknbare fra år til år, men det er ikke nødvendig at de er representative for hele populasjonen.

Dette ble gjort i dette prosjektet ved å tilby alle elever i årene 2017, 2018 og 2019 å ta en såkalt kalibreringsprøve ca. en måned før eksamen. Kalibreringsprøven (heretter kalt K-prøven) er en digital prøve som inneholder oppgaver fra alle hovedområder i læreplanen, men hvor oppgavenes format ligner mest på Del 1 av eksamen. K-prøven ble ansett som en god elevøvelse for selve eksamen. Resultater fra de elevene som avla eksamen i matematikk og som også hadde tatt K-prøven ble analysert sammen.

Design

Det designet som ble brukt for dette formålet, er vist i figur 1. Dette er et skjematisk bilde av to år på rad med dette designet.



Figur 1. Elementer i prosjektdesignet.

En samkalibrering («Concurrent Calibration»)[1] av denne K-prøven og hvert års eksamen setter begge deler på samme skala. Når alle årene i dette prosjektet (2017, 2018 og 2019) er samkalibrert, kommer resultater fra alle årene på samme skala, og oppgavene i K-prøven får samme parametere (vanskegrad og diskriminering) for hvert år. Da kan forskjeller i prestasjon på eksamen sammenliknes mellom år. I tillegg ble det gjort en separat kalibrering av hvert år sammen med K-prøven for å se om K-prøven har endret seg fra år til år og for å gjøre det mulig å estimere lenkefeil mellom år.

En av fordelene med en IRT-analyse er at både oppgaver og elever kommer på samme skala, og dette vil gi informasjon om eksamensdelene endrer seg fra år til år.

Dette designet krever:

- At gruppen som tar kalibreringsprøven, må være sammenliknbar hvert år (utvalget må helst være trukket på samme måte fra samme populasjon)
- At K-prøven holdes hemmelig fra gang til gang
- At K-prøven og eksamen måler den samme kompetansen
- At tiden mellom K-prøven og eksamen er så kort som mulig
- At alle prøvesvar fra både K-prøven og eksamen for disse elevene samles inn paret, slik at analysen blir mulig

[1] For en beskrivelse av denne kalibreringsmetoden anbefales Kolen & Brennan, 2014.

Metoder

Kalibreringsmetoden

En IRT-analyse av eksamen for 10. trinn ble foretatt for årene 2017, 2018 og 2019, både hvert år for seg og for alle tre årene sammen. Samkalibreringen for alle de tre årene ble gjort på to måter, først for å estimere den totale kompetansen til hver elev (fra begge deler av eksamen sammen), og deretter separat for Del 1 og Del 2 av eksamen. IRT-programmet Xcalibre (Guyer & Thompson, 2014) ble brukt for alle IRT-analysene. Dette programmet bruker «marginal maximum likelihood» estimering av oppgaveparameterne som skal sikre «unbiased» resultater. Alle elevsvar ble skåret med en EAP-prosedyre («Expected a Posteriori») som leverte en såkalt theta-skår for hver elev. Denne prosedyren bygger på en evaluering av hele svarmønstret til hver elev som gjennomførte prøvene. (se f.eks. Embretson & Reise, 2013).

Alle dikotome (0,1) oppgaver ble kalibrert med en 2-parameter (2PL) logistisk IRT-modell:

$$P(X_{is} = 1|\theta_s, b_i, a_i) = \frac{e^{a(\theta-b)}}{1 + e^{a(\theta-b)}}$$

Her er P sannsynligheten for riktig svar, gitt en viss ferdighet, en vanskegrad og en diskriminering. a er oppgavens diskriminering, b oppgavens vanskegrad og θ (theta) er ferdigheten til elevene (Birnbaum, 1968).

Polytome oppgaver, dvs. oppgaver med graderte svar (f.eks. 0,1,2 eller flere), ble kalibrert med en såkalt GPCM-modell («Generalized Partial Credit Model»), som er en av de vanligste IRT-modellene for graderte svar (Muraki, 1982).

$$P_{i_{g|g-1.g}}(\theta_j) = \frac{\exp\left[\sum_{g=0}^{g} A_i(\theta_j - b_{i_g})\right]}{\sum_{h=0}^{m} \exp\left[\sum_{g=0}^{h} A_i(\theta_j - b_{i_g})\right]},$$

En samkalibrering («Concurrent Calibration») ble brukt, hvor alle årene er kalibrert sammen, noe som resulterer i at alle resultater kommer på samme skala. Eksamen fra alle årene var lenket sammen med en kalibreringsprøve som var en digital prøve i matematikk, ment å måle samme kompetanse som eksamen (figur 1.).

Det finnes naturligvis mange metoder for å lenke sammen prøver fra et år til et annet, men denne samkalibreringen har vist seg å være en av de beste (Martin et al., 2012). Til sammenlikning kan det nevnes at på de norske nasjonale prøvene, som har ankeroppgaver og er lenket fra år til år, er det brukt en annen metode, en såkalt FCIP metode («Fixed Common Item Parameters») (Bjørnsson, 2018). Da blir parameterne på ankeroppgavene fra året før brukt på inneværende år, og slik sikrer man at hvert år havner på samme skala. Denne metoden er litt enklere i bruk enn en samkalibrering, men den krever nøye sjekking av drift i alle ankeroppgaver, og kalibreringen er da bare basert på årets oppgaver, hvor parametere fra ankeroppgavene fra det første året eller året før, setter resultatene på den originale skalaen.

Samkalibreringen derimot, er basert på alle data samtidig, dvs. data fra alle år og alle oppgaver og gir derfor en vesentlig mindre lenkefeil fra år til år, dvs. mer presise resultater. Når den brukes til å måle trend, krever den også en lineær transformasjon ($\theta_i = \alpha\theta_j + \beta$) av eldre data over på den originale skalaen, for å sikre at alle resultatene havner på samme skala. Dette siste kompliserer metoden litt, men er nødvendig når et nytt år blir lagt til i en tidsrekke. En slik transformasjon var imidlertid ikke nødvendig i dette tilfellet, ettersom alle oppgaver ble kalibrert samtidig.

Samkalibreringen av oppgavene ble i tillegg foretatt på to forskjellige måter, først for hvert år samlet, slik at en helhetlig estimering av ferdigheten til hver av elevene kunne etableres. Da ble Del 1 og Del 2 av eksamen analysert sammen slik at individestimeringen av ferdighet ble basert på alle besvarte oppgaver. Dette gir opplysninger om eksamen i sin helhet hvert år. Den andre kalibreringen ble foretatt for Del 1 og Del 2 separat for å se nærmere på egenskapene til hver del og for å se endringer over tid for hver av dem.

Kalibreringsprøven og valg av elever

En digital prøve i matematikk ble laget spesielt for dette prosjektet. Prøven inneholdt 39 oppgaver, som først og fremst samsvarer med Del 1 på eksamen. Denne prøven ble laget av eksamensnemden for 10. trinn, som også er ansvarlig for de ordinære eksamensoppgavene. Alle elever fikk tilbud om å ta denne prøven som en eksamensforberedende prøve ca. en måned før eksamen, dvs. før de fikk vite om de skulle ta eksamen i matematikk eller ikke. På denne måten gikk resultatene fra alle de som tok K-prøven, men ikke matematikkeksamen tapt. Resultatene fra analysen bygger derfor på de elevene som tok K-prøven, og som også tok matematikkeksamen senere.

Denne utvalgsmetoden kan være litt tvilsom, ettersom man ikke sikrer på forhånd at utvalgene blir sammenliknbare. Metoden inkluderer en grad av selvvalg som kan ha påvirket resultatene, men uten sammenlikningsgrunnlag fra de som valgte å ikke ta K-prøven er det vanskelig å si noe om effekten av dette. Dette ville kreve et tilsvarende utvalg elever som tok eksamen, men ikke K-prøven, og en analyse av K-prøve-resultatene fra dem som ikke kom opp i matematikk.

Disse forholdene ved valg av elever var uheldig, men ikke til å unngå med dagens system. Men det er viktig å ha dette i tankene når resultatene av denne analysen blir vurdert, spesielt når det gjelder sammenlikning mellom årene.

Deltakere og antall oppgaver

Tabell 1 viser antall elever som deltok på begge prøvene, og antall oppgaver de besvarte hvert år.

Tabell 1. Antall elever og oppgaver

	2017	2018	2019	Total
Antall elever	3120	1839	3035	7994
Oppgaver i K-prøve*	39	39	39	39
Oppgaver i Del 1 på eksamen	34	30	30	94
Oppgaver i Del 2 på eksamen	24	23	24	71
Samlet antall oppgaver på eksamen	58	53	54	165
Total				204

* Oppgavene i K-prøven var de samme hvert år, men alle eksamensoppgavene var nye hver gang

Tabellen viser at antall elever er likt i 2017 og 2019, men ganske lavt i 2018 sammenliknet med de andre to årene. Det er uvisst hvorfor det er slik, men kanskje var gruppen noe annerledes i 2018 enn de andre årene, ettersom det var frivillig deltakelse på prøven. Som vi skal se litt senere, så er resultatene i 2018 litt annerledes enn de andre to årene. Dette kan tyde på at deltakelsen har hatt en effekt som vi har liten eller ingen kontroll over.

Det er også verdt å nevne at det var fire oppgaver mer i Del 1 av eksamen i 2017 sammenliknet med de andre årene, og det totale antallet oppgaver er ikke det samme noen av årene. Dette fører til at det blir umulig å sammenlikne poengsummer. I tillegg dreier det seg naturligvis om forskjellige oppgaver hvert år, noe som gjør det enda mer umulig å sammenlikne resultatene med poengsummer eller karakter basert på dem.

Resultater

Resultatene fra IRT-analysen blir her presentert i tre hovedavsnitt. Det første handler om oppgavene og deres vanskegrad, det andre om elevprestasjonen hvert år og det tredje om analysen av enkeltoppgaver. De resultatene som blir presentert her, og som gjelder ferdighet og vanskegrad på oppgaver eller hele eksamen, er uttrykt som en såkalt theta-verdi når det gjelder elevenes ferdighet, men som en b-parameter når oppgavenes vanskegrad er omtalt. Begge disse har gjennomsnittet 0 og standardavviket 1. Dette er en konvensjon i IRT-analyse og er tilnærmet en normalfordeling. Men det må understrekes at dette ikke betyr at resultatene blir tvunget inn i en normalfordeling eller transformert til en slik fordeling.

De tallene som IRT-analysen leverer, kan naturligvis konverteres til hvilken som helst skala, som f.eks. på nasjonale prøver hvor de er multiplisert med 10 pluss 50, for å få en skala med gjennomsnittet 50 og standardavviket 10. De internasjonale storskalaundersøkelsene gjør nøyaktig det samme, men bruker gjennomsnittet 500 og standardavviket 100. Da er theta-verdien multiplisert med 100 pluss 500. Men her

blir theta-verdien brukt direkte, og ikke konvertert. Dette har den ulempen at tall under gjennomsnittet, blir minustall, og dette må man ha i tankene når tallene leses.

Reliabilitet

Før resultatene presenteres, er det nødvendig å se på hvor reliable disse målingene er.

Reliabiliteten til eksamen er uttrykt med en alfa-koeffisient og vises i tabell 2. Alfa-koeffisienten er det vanligste målet på reliabilitet til en prøve, men det bør understrekes at den er i realiteten prøvens indre konsistens[2]. Denne koeffisienten kan ha verdier fra 0 til 1, og et vanlig minimumskrav til en prøve er alfa over 0,8.

Tabell 2. Reliabilitet for tre år

År	Hele eksamen	Del 1	Del 2
2017	0,950	0,900	0,925
2018	0,949	0,911	0,907
2019	0,942	0,894	0,896

Alfa-koeffisienten viser seg å være meget bra i alle årene, både når man ser på eksamen som en helhet, og når man ser hver del for seg. Dette viser at den indre konsistensen i eksamen som en helhet, og i hver deleksamen, er meget bra, og dette støtter den antakelsen at de forskjellige oppgavene måler ulike aspekter av samme ferdighet. K-prøven hadde en alfa-koeffisient på 0,87.

Denne høye indre konsistensen (alfa) i begge deler av eksamen og i eksamen som en helhet, tyder på at oppgavene er meget homogene og måler stort sett den samme underliggende kompetansen. Ut ifra dette kunne man uten å tape målepresisjon, ha eksamen betydelig kortere.

Alfa er naturligvis høyere for eksamen som helhet, enn for de to delene, men hvor høy denne koeffisienten blir er en funksjon av antall oppgaver og hvorvidt oppgavene måler aspekter av samme ferdighet. Det er klart at det er godt mulig å måle en ferdighet med én eneste oppgave, men det ville ha gitt et svært upålitelig resultat. Gode prøver inneholder derfor et minimum antall oppgaver for å få en pålitelig måling av forskjellige elever.

[2] Andre metoder som test-retest metoden er bedre egnet til å estimere reell reliabilitet, dvs. om en prøve måler på samme måte hver gang den er brukt.

Oppgavenes vanskegrad

I det følgende presenteres vanskegrad og diskriminering for alle oppgaver fra alle år. Vanskegraden er representert med en b-parameter, og diskrimineringen, dvs. hvor bra oppgavene diskriminerer mellom individer med forskjellig ferdighet, med en a-parameter. Tabell 3 viser gjennomsnittlige verdier for hele eksamen hvert år.

Tabell 3. Gjennomsnittlige verdier for a- og b-parameterne. Eksamen som helhet.

År	Parameter	Antall oppgaver	Snitt	SD	Min	Max
2017	a	58	0,843	0,285	0,375	1,736
	b	58	-0,392	1,023	-4,000	1,532
2018	a	53	0,920	0,275	0,445	1,816
	b	53	-0,435	0,879	-2,274	1,417
2019	a	54	0,904	0,296	0,410	1,711
	b	54	-0,441	1,105	-2,713	2,310

Tabellen viser at den gjennomsnittlige vanskegraden på oppgavene er høyest i 2017, men går litt ned de andre to årene. Disse endringene er likevel meget små og ikke signifikante. Den gjennomsnittlige diskrimineringen er bra alle årene. Merk at her er de to delene av eksamen slått sammen. Tabell 4. viser gjennomsnittlig vanskegrad for de to delene av eksamen i de tre årene.

Tabell 4. Gjennomsnittlige verdier for b-parameteren. Deler av eksamen.

År	Antall oppgaver	Snitt	SD	Min	Maks
Del 1 2017	34	-0,667	1,256	-4,000	1,532
Del 1 2018	30	-0,774	0,780	-2,274	1,350
Del 1 2019	30	-0,837	1,008	-2,713	1,222
Del 2 2017	24	-0,002	0,251	-0,879	0,824
Del 2 2018	23	0,006	0,814	-1,687	1,417
Del 2 2019	24	0,054	1,034	-2,104	2,310

Her ser vi at vanskegraden for Del 1, ser ut til å bli marginalt lavere i løpet av de tre årene, ikke veldig mye, men likevel noe. Ingen av disse endringene mellom år er statistisk signifikante. Variasjonen (SD) for Del 1 er også ulik i disse årene: størst i 2017 og minst i 2018. Del 2 ser ut til å være mer stabil, men der går utviklingen i motsatt retning, og oppgavene blir gjennomsnittlig litt vanskeligere. Det er påfallende at variasjonen i Del 2 er meget lav i 2017. Dette vises med standardavviket og verdiene for maksimum og minimum det året. Alt i alt er endringen størst i Del 2 hvor variasjonen er blitt vesentlig større for hvert år. Her har det skjedd en betydelig endring, selv om endringen i gjennomsnitt ikke er signifikant.

I 2017 hadde Del 2 oppgaver som bare dekket en ganske snever vanskegrad (fra -0,8 til 0,8), men denne er blitt større for hvert år, selv om gjennomsnittet for hele eksamen ikke har endret seg så mye.

Del 1 hadde noen svært lette oppgaver i 2017, som det er blitt færre av de andre årene. I 2018 og i 2019 ser oppgavene ut til å dekke samsvarende ferdighetsnivåer, selv om variasjonen viser seg å være noe lavere i 2018 enn i de andre årene.

Informasjonsfunksjon for eksamen

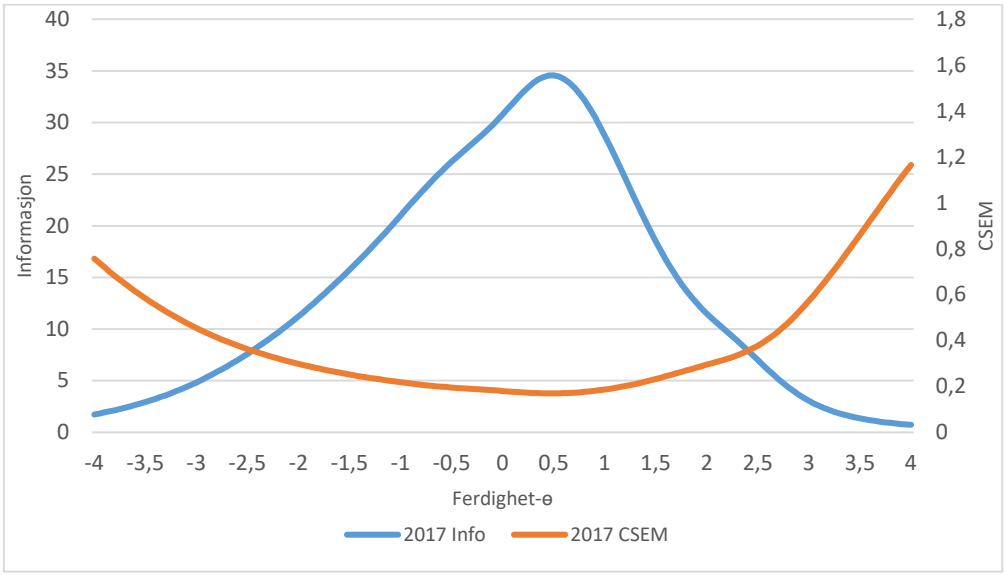
Hver prøve leverer informasjon ifølge følgende formel:

$$I_i = a_i^2 \, P_i(\theta) \, Q_i(\theta)$$

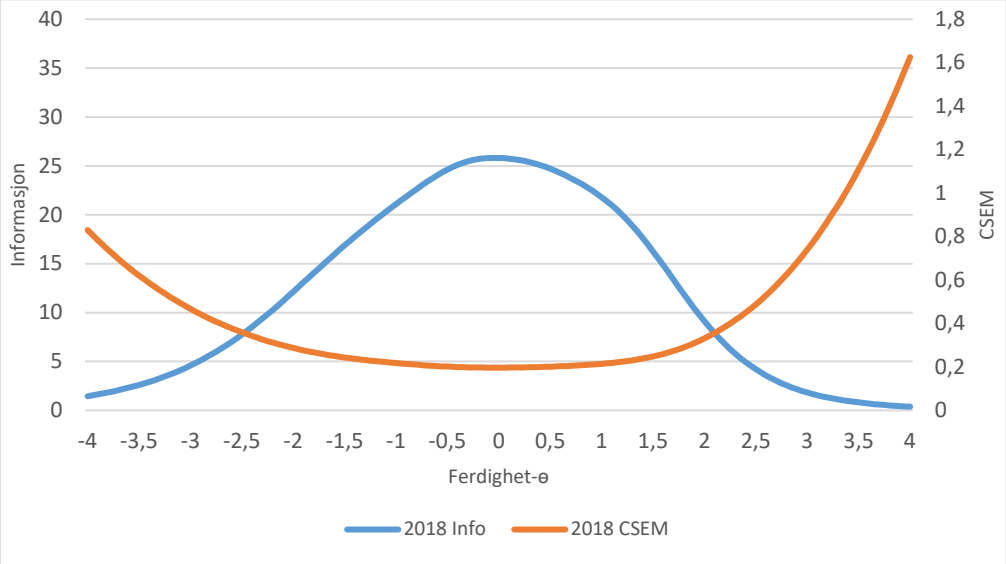
Her er I=informasjon, a er diskriminering, P og Q sannsynlighetene for riktig og galt svar, og θ er ferdigheten. Dette regnes i IRT-analysen ut for hver oppgave, men kan også summeres opp for en prøvedel eller en hel prøve og gir opplysninger om hvor på ferdighetsskalaen prøven leverer mest informasjon, og hvordan fordelingen av disse opplysningene er. Ut ifra informasjonsfunksjonen kan en regne ut prøvens kondisjonale standard målefeil («Conditional Standard Error of Measurement – CSEM») med:

$$SE=\frac{1}{\sqrt{I_i}}$$

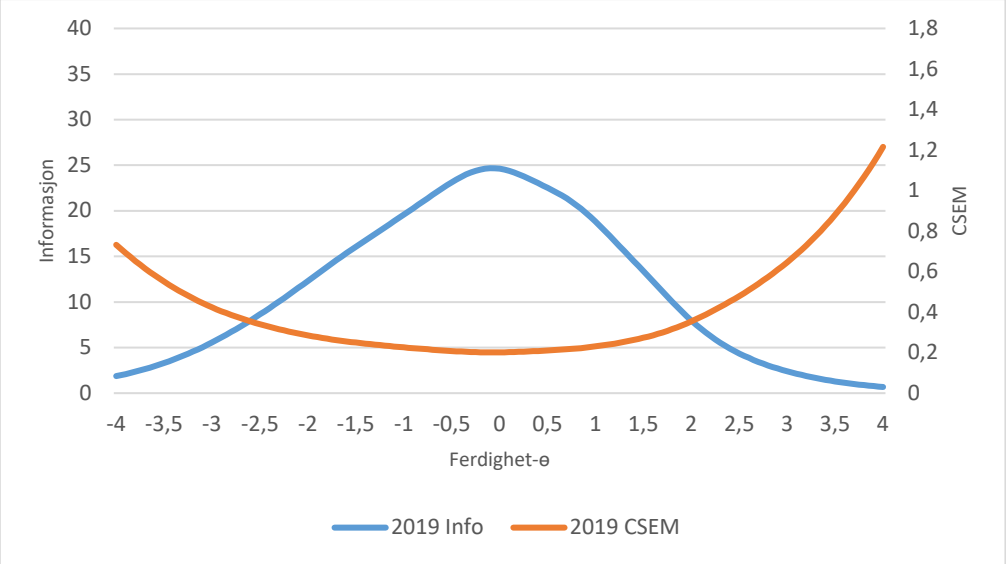
Prøvens målefeil er altså «resiprokalen» av informasjonen for forskjellige verdier av theta, kondisjonalt av Theta. Dette gir forskjellig målefeil forskjellige steder på ferdighetsskalaen. På de tre følgende figurene vises informasjon og målefeil for eksamen i sin helhet for de tre årene.



Figur 2. Informasjon for 2017 Maks=0,5

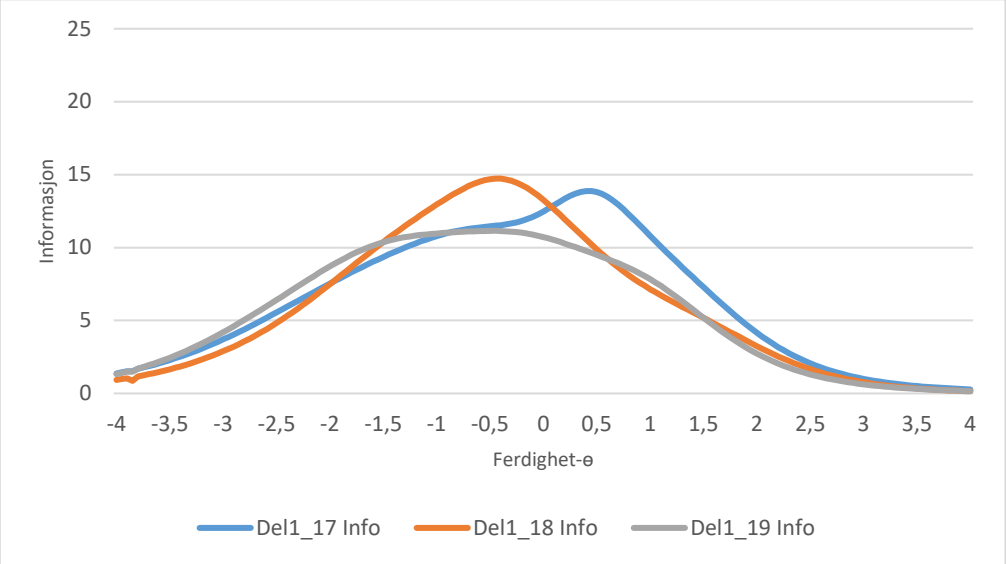


Figur 3. Informasjon for 2018. Maks=-0,15

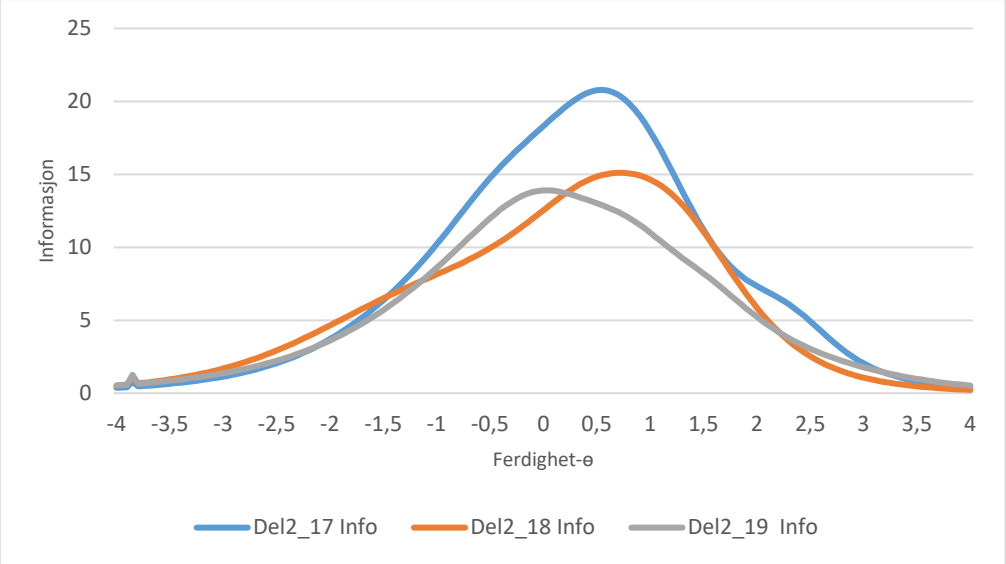


Figur 4. Informasjon for 2019. Maks=-0,1

Figurene viser at eksamen i 2017 måler vesentlig snevrere enn de to andre årene og at plasseringen på skalaen hvor eksamen leverer maksimal informasjon, flytter seg fra år til år (0,5 til -0,15 til -0,1). Figur 5 og 6 viser hvordan de to eksamensdelene måler hver for seg alle tre årene.



Figur 5. Del 1 alle årene.

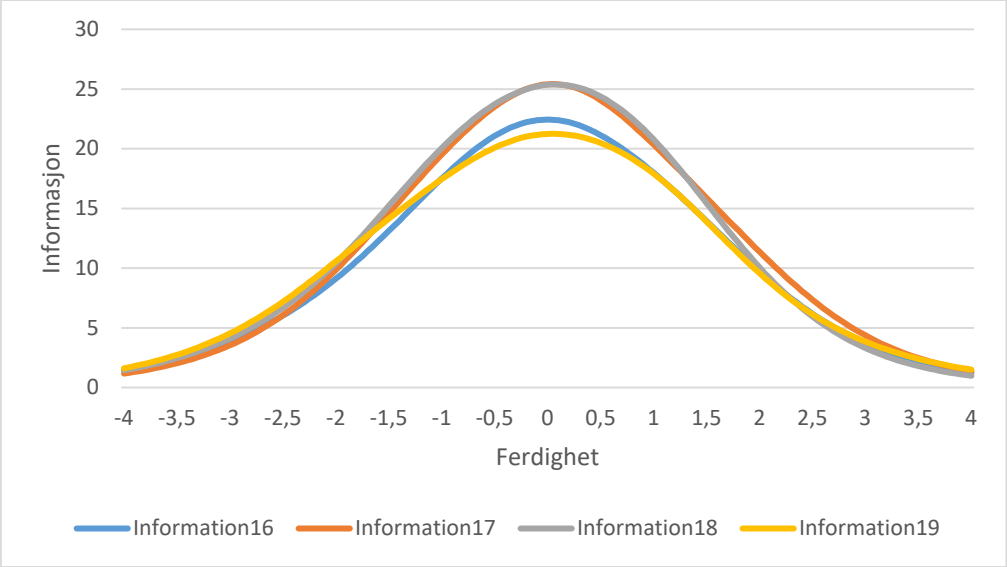


Figur 6. Del 2 alle årene.

Når en ser på disse figurene, så er det klart at eksamensdelene leverer forskjellig mengde informasjon hver gang (høyden på kurvene), og at maksimal informasjon er på forskjellig sted på ferdighetsskalaen hver gang. I 2017 er informasjonen maksimal høy, og det skyldes sannsynligvis tre forhold: 1) Det året hadde eksamen 58 oppgaver, men kun 53 og 54 de neste årene, samtidig med at 2) antallet polytome oppgaver var vesentlig høyere det året enn de andre, spesielt på Del 2. 3) I tillegg var oppgavene på Del 2 i 2017 veldig like og lå på omtrent samme sted på skalaen, noe som hever toppen på kurven akkurat der oppgavene ligger.

Det er som figurene viser, store svingninger i hvordan disse prøvedelene oppfører seg fra år til år. Dette kan delvis forklares med at oppgavene diskriminerer forskjellig i de tre årene, men svaret på hvorfor 2017 leverer høyest informasjon, er de tre forklaringene over. I tillegg hadde eksamen 23 oppgaver med graderte svar i 2017, men kun henholdsvis 11 og 8 i 2018 og 2019. Graderte oppgaver leverer vanligvis mer informasjon enn dikotome, og det høye antallet av dem i 2017 pluss at det var 4 og 5 flere oppgaver enn de andre årene, forklarer høyden det året. De polytome oppgavene blir spesielt omtalt senere.

Til sammenlikning kan vi se på informasjonskurver fra Nasjonal Prøve i regning fra årene 2016 til 2019.



Figur 7. Informasjonskurver fra Nasjonal prøve i regning for 8. trinn.

Her er kurvene mye jevnere og likere mellom år. Toppen av kurvene er omtrent på samme sted på ferdighetsskalaen alle årene, men disse prøvene leverer litt forskjellig mengde informasjon hvert år på de samme stedene på grunn av varierende diskriminering av enkeltoppgaver. Spredningen på kurvene er den samme hver gang. Alle prøvene har like mange oppgaver. (NB: 58 i 2014, men 50 etter det). Oppgavene i disse prøvene er alle sammen pilotert minst to ganger og analysert med IRT, slik at denne sammenliknbarheten er blitt sikret på forhånd. Her er det også viktig at standardmålefeilen blir omtrent den samme for hvert ferdighetsnivå hvert år, og varierer ikke som den gjør på eksamen.

Hovedbekymringen, når en ser på informasjonskurvene fra eksamen, er imidlertid at toppen på informasjonskurvene aldri er på samme sted; den flytter seg fra år til år. Fordelingen av informasjonen er også forskjellig fra år til år med en snever fordeling i 2017 og litt bredere etter det. Konsekvensen av dette er også at målefeilen på eksamen varierer, sikkerheten i målingen er ikke den samme eller lik fra år til år. Dette er i grunnen en annen side av de opplysningene som kom frem i tabell 4, hvor rekkevidden i målingen på Del 2 viser seg å være meget liten i 2017, litt større i 2018 og størst i 2019.

Elevenes ferdighetsestimering

De små endringene i oppgavenes vanskegrad over tid må ses i sammenheng med hva som skjer med elevenes ferdighetsestimater disse tre årene, men merk at selv om disse er på samme skala så er IRT-modellen invariant. Dette betyr at oppgavenes vanskegrad skal være uavhengig av gruppen som besvarer dem, gitt at dataene er kalibrert på en gruppe der hele ferdigheten er representert. Dette er en meget viktig egenskap til IRT-modellene og godt dokumentert i litteraturen (f.eks. Lord, 1980). Her må det igjen nevnes at det er et ubesvart spørsmål om 2018 var sammenliknbart med de to andre årene, og ut ifra de forrige resultatene er det mulig at det året manglet noen elevgrupper.

Om trend og lenkefeil

For å vurdere om endringene i prestasjon fra et år til et annet er signifikante eller ikke, må selve størrelsen av endringen bestemmes. Det gjør man enkelt ved å trekke den ene verdien fra den andre: År1 – År2, osv.

I denne undersøkelsen ble det brukt en såkalt samkalibrering av alle årene samtidig. Det betyr at alle resultater er på samme skala fra starten av. Dataene fra alle år ble satt i en datafil, sammen med resultatene fra kalibreringsprøven fra alle år. Det er gjennom kalibreringsprøven at lenken mellom år opprettholdes. Men ingen ankerprøver er perfekte, og kalibreringsprøven som ble gjentatt hvert år, kan ha endret seg, eller «sklidd» i alle fall på noen oppgaver. Derfor er det viktig å etablere såkalt lenkefeil mellom de tre årene. Lenkefeilen legges da til standardfeilen for forskjellen mellom hvert år, på følgende måte:

SE for forskjellen må bestemmes, og det er gjort med følgende formel: Hvis dette var prestasjon to år på rad, År 1 og År 2, så er forskjellen, År2 minus År1 (eller omvendt):

$$SE(\text{År}2 - \text{År}1) = \sqrt{SE(\text{År}2)^2 + SE(\text{År}1)^2 + (LE)^2}$$

Når denne standardfeilen foreligger, kan det foretas en enkel t-test på følgende måte:

$$t = \frac{\text{År}2 - \text{År}1}{SE(\text{År}2 - \text{År}1)}$$

Hvis denne verdien er høyere enn 1,96 eller lavere enn -1,96, så er forskjellen signifikant ved $p \leq 0{,}05$.

For å beregne lenkefeilen må en separat kalibrering fra alle årene og data fra samkalibreringen for alle årene sammenliknes. Da har man en forskjell i ferdighetsestimering fra separat og samkalibrerte data, og gjennomsnittene av standardfeilen på disse forskjellene er lenkefeilen. For nærmere forklaring av metoden henvises til Martin et al. (2012).

Denne beregningen ble foretatt for alle tre årene under samkalibreringen, og lenkefeilen mellom alle årene viste seg å være veldig liten, f.eks. så er SE for forskjellen i separat og ankret kalibrering 0,00041 for 2017 0,00064 for 2018 og 0,0027 for 2019. De andre forskjellene viste seg også å være av samme størrelsesorden, og derfor blir lenkefeilen utelatt i signifikanstestingen her, ettersom den ikke er viktig. Dette viser også at lenkefeil under en samkalibrering hvor alle data er benyttet, blir veldig liten, sammenliknet med en beregning av lenkefeil hvor kun ankeroppgavene er benyttet. Med andre ord; samkalibreringen resulterer i en bedre lenking.

Endringer over tid

Tabell 5 viser utviklingen av ferdighetsestimatene for eksamen som helhet.

Tabell 5. Elevenes ferdighetsestimering alle år.

År	Antall elever	Snitt	SD	Skew	Min	Median	Maks
2017	3120	-0,034	0,989	-0,030	-3,352	-0,035	2,881
2018	1839	0,025	1,018	-0,135	-3,231	0,049	2,643
2019	3035	0,002	0,992	0,030	-3,264	0,003	2,803

En test av signifikans mellom alle årene vises i tabell 6.

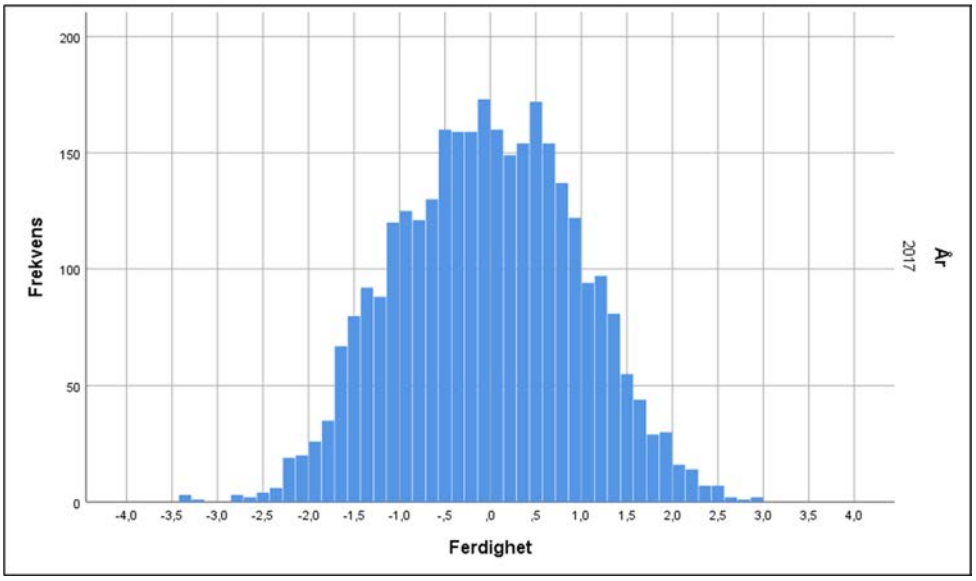
Tabell 6. Signifikans mellom år.

	17–18	18–19	17–19
Forskjell	0,059	-0,022	0,037
SE forskjell	0,030	0,030	0,025
t	**1,999**	-0,752	1,458

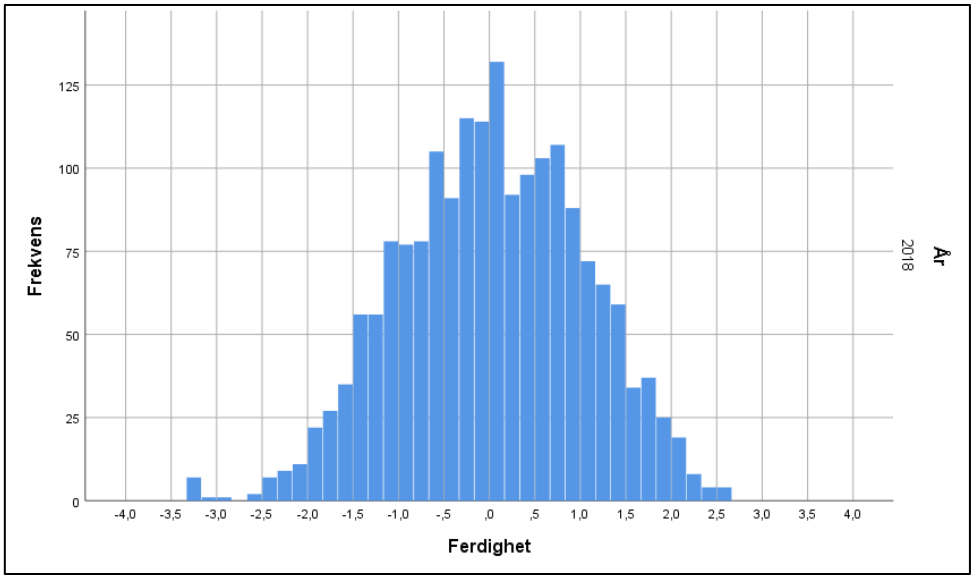
Her ser vi at gjennomsnittet for helthetsestimeringen endres svært lite i de tre årene. Likevel er endringen mellom 2017 og 2018 signifikant, (t=1,99), men her må man huske på at denne endringen er på mindre enn ett poeng i summen av poeng fra eksamen. Endringen mellom 2018 og 2019 er ikke signifikant. Her er det, som tidligere nevnt, en mulighet for at elevgruppen i 2018 var annerledes enn de andre årene; det er mulig at de hadde en høyere ferdighet, at de «flinkeste» elevene var de som tok K-prøven dette året, eller at antallet elever som presterte lavt var betydelig mindre da enn i de andre årene.

Selv om forskjellen i prestasjoner mellom 2017 og 2018 er statistisk signifikant, så kan man spekulere på hvor mye det betyr ettersom forskjellen er så liten som 0,06. I skalapoeng på nasjonale prøver med gjennomsnitt 50 og standardavvik 10, ville dette vært forskjellen mellom 50,24 og 49,66, eller 0,6 poeng, altså veldig lite. I poeng på eksamen er dette en forskjell på rundt 1 poeng, men her er det umulig å være eksakt, ettersom eksamen i disse tre årene ikke har samme antall poeng.

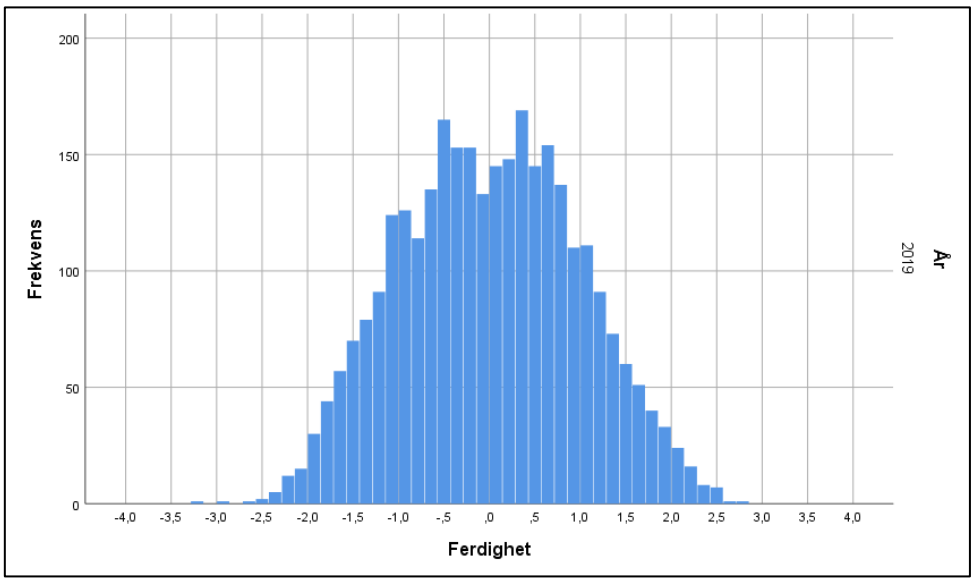
Ferdighetsfordelingen alle disse årene er vist i figur 8, 9 og 10.



Figur 8. Ferdighetsfordeling 2017.



Figur 9. Ferdighetsfordeling 2018



Figur 10. Ferdighetsfordeling 2019.

Dette er svært like fordelinger av ferdighet i disse tre årene, og ingenting tyder på at elevene presterer vesentlig dårligere eller bedre i løpet av perioden. Merk at det på midten av fordelingene er en antydning til en todeling («bimodal distribution») alle årene. Man kan lure på hva dette betyr, men det er i alle fall et skille mellom elevene der som kanskje må undersøkes nærmere og forklares. Ofte betyr en slik todeling at elevgruppen kan deles i to grupper som har forskjellige egenskaper. Dette må kanskje undersøkes nærmere, men her er denne bimodaliteten ikke stor.

Hvis man ser på de to delene av eksamen hver for seg, viser det seg at forskjellen mellom 2017 og 2018 kan knyttes til prestasjoner på Del 2, men merk at denne forskjellen er svært liten selv om den er signifikant. I forrige avsnitt ble det dessuten vist at det mellom 2017 og 2018 var en minimal vanskegradsendring i oppgavene som kunne ses over tid i Del 2.

Tabell 7. Ferdighetsestimering – deler av eksamen alle år.

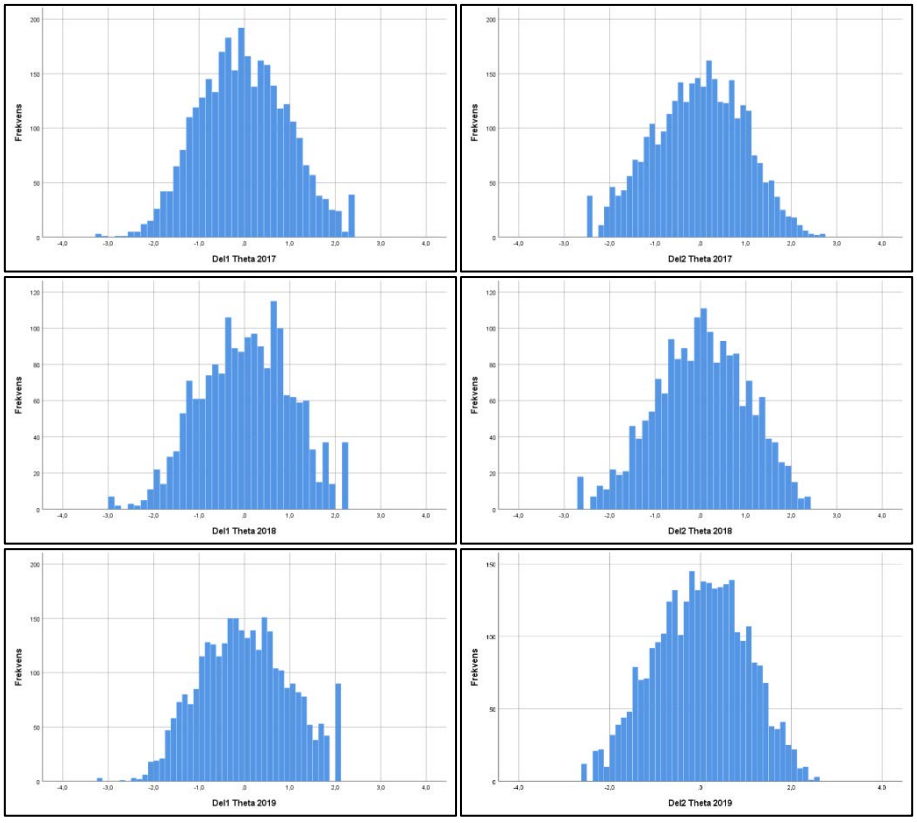
	n	Del1	SD	SE	Del2	SD	SE
2017	3120	-0,023	0,974	0,017	-0,048	1	0,018
2018	1839	0,028	0,993	0,023	0,018	1,017	0,024
2019	3035	0,01	0,98	0,018	-0,012	1,004	0,018

Tabell 8. Signifikanstesting mellom år-eksamensdeler.

	Forskjell Del 1	SE-DIFF	t	Forskjell Del 2	SE-DIFF	t
18-17	0,051	0,029	1,759	0,066	0,030	**2,221**
19-18	-0,018	0,029	-0,616	-0,03	0,030	-1,003

Her kan det fortsatt tenkes at utvalget i 2018 kan ha hatt en effekt, at det har hatt en overvekt av høyt presterende elever eller en mangel av svakt presterende. Som tidligere konstatert er det vanskelig å være sikker på dette, uten en grundig analyse av alle andre eksamens- og K-prøveresultater (fra de som bare gjennomførte én av delene).

En ferdighetsfordeling fra begge deler av eksamen for alle årene vises i figur 11.



Figur 11. Ferdighetsfordeling for alle deler av eksamen alle tre årene.

Det er ingen merkbar meningsfull endring i elevenes prestasjoner gjennom disse tre årene. Prestasjonen på eksamensdel 2 er visstnok signifikant forskjellig mellom 2017 og 2018, men en kvalitativ undersøkelse (granskning av forskjeller i ferdighet) av hva som ligger bak dette, ville sannsynligvis ikke ha kunnet påvist en tydelig forskjell, ettersom den er meget liten.

Kjønnsforskjeller

Det viste seg å være signifikante kjønnsforskjeller i nesten alle deler av eksamen i de tre årene. Tabell 9 viser disse og hvilke av dem som er signifikant. K-prøven derimot hadde ingen signifikante forskjeller mellom kjønnene.

Tabell 9. Kjønnsforskjeller

	Kjønn	N	Snitt Theta	SD Theta	Theta SE	Forskjell G-J	SE forskjell	T*
Del 1 2017	G	1625	-0,07	0,974	0,024	-0,098	0,035	**-2,811**
	J	1495	0,028	0,972	0,025			
2018	G	931	-0,01	0,995	0,033	-0,077	0,046	-1,663
	J	908	0,067	0,991	0,033			
2019	G	1573	-0,028	0,978	0,025	-0,079	0,036	**-2,220**
	J	1462	0,051	0,981	0,026			
Del 2 2017	G	1625	-0,142	1,003	0,025	-0,197	0,036	**-5,529**
	J	1495	0,055	0,986	0,026			
2018	G	931	-0,053	1,046	0,034	-0,144	0,047	**-3,046**
	J	908	0,091	0,981	0,033			
2019	G	1573	-0,078	1,04	0,026	-0,138	0,036	**-3,803**
	J	1462	0,06	0,959	0,025			
K-prøve	G	4129	0,022	1,024	0,016	0,027	0,022	1,209
	J	3865	-0,005	0,973	0,016			

*Signifikante forskjeller vises med fet skrift

Jenter presterer bedre på all deler av eksamen, unntatt på Del 1 i 2018. Dette er et særdeles interessant resultat sett i lys av at gutter presterer generelt litt bedre enn jenter på nasjonale prøver (8. trinn) og på PISA undersøkelsen (10. trinn), inntil PISA 2015, da jentene presterte litt bedre. Dette er noe som burde undersøkes nærmere, med en grundig gransking av de forskjellige oppgavene på disse ulike prøvene, slik at man fikk bedre forståelse av hva det er som forårsaker denne diskrepansen, og om det er en utvikling i jentenes favør på gang.

K-prøven hadde heller ingen signifikante forskjeller, men her var det en tendens til at guttene gjorde det litt bedre, i motsetning til på eksamen der jentene presterte best. Disse kjønnsforskjellene er ganske vanlige når en sammenlikner papirprøver og elektroniske prøver, og i litteraturen finnes det en del undersøkelser av dette, men denne forskjellen kan muligens også være på grunn av forskjellig prøvelengde hvor jentene holder ut lenger eller har høyere motivasjon for å løse oppgavene. Naturligvis kan andre faktorer ha en effekt her., Del 2 av eksamen er delvis elektronisk og der peker ikke forskjellene i samme retning som på k-prøven.

Dersom en skal konstatere hva som påvirker disse kjønnsforskjellene er det derfor behov for en videre undersøkelse.

Oppgavene og deres egenskaper

Tabell 10 viser en oppsummering av antall oppgaver på eksamen hvert år, samt antallet dikotome og polytome oppgaver.

Tabell 10. Antall oppgaver etter år og oppgavetype.

	2017	2018	2019	Total
Oppgaver i del 1 på eksamen	34	30	30	94
Oppgaver i del 2 på eksamen	24	23	24	71
Samlet antall oppgaver på eksamen	58	53	54	165
Dikotome oppgaver	35	42	46	123
Polytome oppgaver	23	11	8	42

Dikotome oppgaver

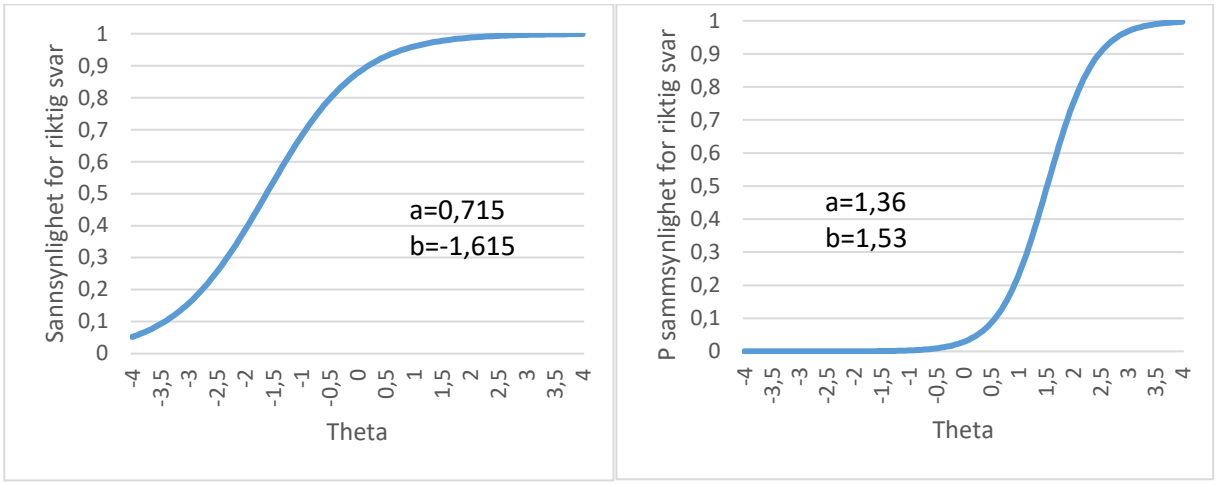
Tabell 11 beskriver egenskapene til alle de dikotome oppgavene for hvert år. Her ble det brukt en 2PL logistisk modell som gir diskriminering (a) og vanskegrad (b). Merk at modellen bruker konstanten D=1,7 som betyr at diskrimineringstallene er lavere enn hvis D=1. Dette betyr at alle a-parametere under 0,3–0,5 er svak diskriminering.

Tabell 11. Parameteropplysninger dikotome oppgaver.

	2017		2018		2019	
	a	b	a	b	a	b
Snitt	0,88	-0,65	0,99	-0,48	0,94	-0,39
Maks	1,7	1,53	1,82	1,42	1,71	2,31
Min	0,44	-2,16	0,47	-2,27	0,53	-2,71

Som tabellen viser, har ingen av disse oppgavene en dårlig/ubrukelig diskriminering, og det er et meget bra resultat. Høyeste vanskegrad derimot, ligger lavt i 2017 og 2018, men er betydelig bedre i 2019. Dette stemmer med de forrige resultatene som viste at oppgavene måler høyere opp på ferdighetsskalaen i 2019.

Bildet under viser to eksempler på dikotome oppgaver. Det første viser en oppgave som er relativt lett, og det andre en som er vanskelig.



Figur 12. Eksempler på dikotome oppgaver – ICC kurver

Nærmere opplysninger om alle oppgaver finnes i vedlegg 1, Xcalibre-rapporten fra samkalibreringen.

Om graderte svar – polytome oppgaver:

Graderte svar er de som har kode 0, 1, 2 eller høyere. Disse svarene ble alle sammen kalibrert med en GPCM («Generalized Partial Credit Model») IRT-modell (Muraki, 1992). Dette er den vanligste modellen å bruke på graderte oppgaver, men det finnes også andre varianter som f.eks. Samejima-modellen som er brukt i nasjonale prøver.

Noen av de polytome oppgavene i eksamen ser ut til å fungere bra, men det er dessverre også flere av dem som ikke fungerer så bra. Dette viser at det er helt nødvendig å pilotere oppgaver av denne typen for å sikre at de fungerer på riktig måte. Selv de mest erfarne oppgavekonstruktører klarer ikke å lage gode graderte oppgaver uten pilotering. I tillegg ligger det en sensorvurdering bak mange av disse resultatene, og det kan føre til reliabilitetsproblemer som kan ha en effekt her.

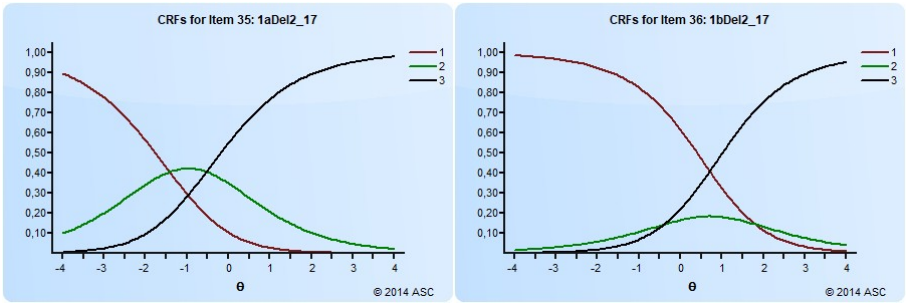
Tabell 12 viser hvor mange graderte oppgaver det er totalt i eksamen alle de tre årene, og hvor mange av dem som ikke fungerer som de skal.

Tabell 12. Antall graderte oppgaver som ikke fungerer optimalt.

	Antall graderte oppgaver	Antall som ikke virker optimalt
2017	23	17 (73%)
2018	11	5 (45%)
2019	8	6 (75%)
Samlet:	42	28 (66%)

Det er helt klart ut fra denne oppsummeringen at her må man ta grep og teste disse oppgavene på forhånd (pilotere). Alternativet ville være å endre skåringen til kun 0 og 1, dvs. riktig og galt, siden de dikotome oppgavene i eksamen fungerer meget bra. Det ville sannsynligvis ha gitt et bedre og riktigere resultat, men litt lavere informasjonsverdi for hele eksamen. Det er ikke akseptabelt at eksamen inneholder så mange oppgaver som ikke virker som de skal.

Dette betyr naturligvis ikke at oppgavene ikke måler den ferdigheten de skal i noen grad, det gjør de ganske sikkert. Men det peker på at bruken av poengene, spesielt de som ligger mellom 0 og øverste poeng, er nokså tilfeldig og kanskje usystematisk og må forbedres. Mellomkategoriene i disse oppgavene har i veldig mange tilfeller aldri en større sannsynlighet enn laveste (0) eller øverste kategori. Det er ikke akseptabelt at 66 prosent av disse oppgavene ikke virker, spesielt fordi de potensielt bidrar ganske mye til den totale informasjonen som eksamen leverer (Se vedlegg 1 for nærmere opplysninger om alle de polytome oppgavene).



Figur 13. Eksempler på ICC fra graderte oppgaver.

I eksemplet i figur 13, er det vist to graderte oppgaver. Den første virker bra, og mellomkategorien (oppgaven hadde 0, 1 og 2 poeng) har størst sannsynlighet for å beskrive ferdigheten midt på skalaen. Den andre oppgaven er et eksempel på en oppgave hvor mellomkategorien ikke virker; den har aldri den høyeste sannsynligheten for å beskrive ferdigheten. Her er mellomkategorien nytteløs, og oppgaven burde endres over til å bli dikotom (0, 1). Ofte kan små endringer i tekst eller visning av oppgaven hjelpe, men dette kan en ikke vite på forhånd uten å pilotere oppgavene og kjøre dem igjennom en IRT-analyse.

Kalibreringsprøven – en kort oppsummering

Som tidligere nevnt og beskrevet i innledningen, er kalibreringsprøven sentral i dette prosjektet. Det er derfor passende å beskrive den litt og se på de egenskapene den har. K-prøven lenker sammen resultatene, sentrerer dem og sikrer at alle tre årene faller på samme skala og muliggjør de analysene som er presentert her foran.

K-prøven inneholder 39 oppgaver som alle sammen er dikotome og presentert elektronisk. Prøven ble gjennomført som en frivillig eksamensøvelse for alle elever som ønsket det, omtrent én måned før eksamen og før elevene visste om de ville bli trukket ut til matematikkeksamen. Disse forholdene kan selvsagt ha hatt en innvirkning på hvordan prøven virket. Utvalget var naturligvis påvirket av denne frivilligheten og som nevnt tidligere, er det ikke sikkert at utvalgene disse tre årene er helt sammenliknbare. Spesielt råder det tvil om resultatene fra 2018, da ble utvalget betydelig mindre enn i de andre årene.

Tabell 13. Kalibreringsprøven.

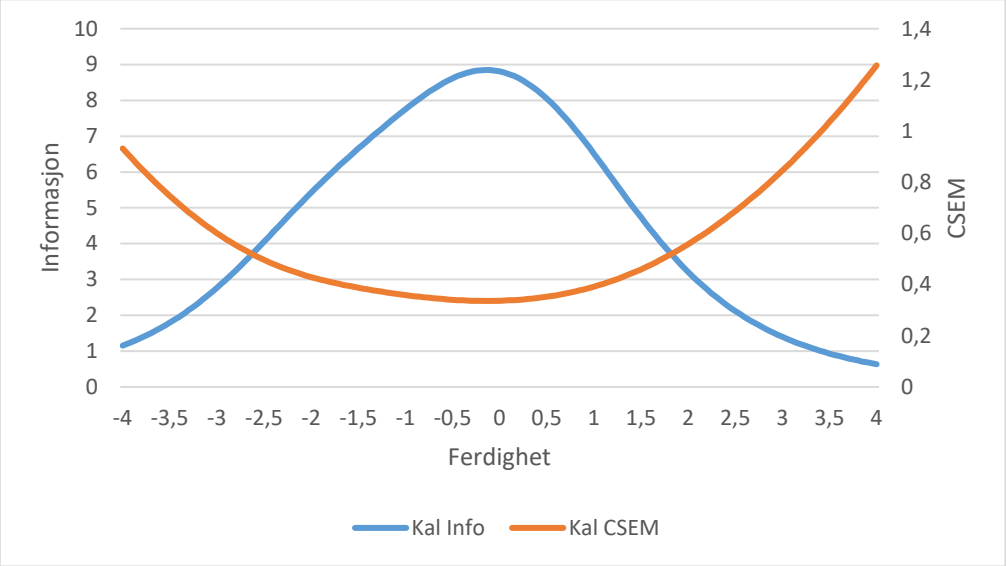
	N oppgaver	Snitt	SD	Min	Maks
a	39	0,596	0,249	0,146	1,094
b	39	-0,376	1,065	-3,279	1,386
Antall elever	7994				
Snitt Theta	0,009		1,00	-3,14	2,67

Gjennomsnittet på K-prøven viser bra hvordan samkalibreringen virket; det havner på 0 med standardavviket 1. Alle andre verdier i samkalibreringen må sees i relasjon til dette, og dette understreker rollen til K-prøven som en ankerprøve i en samkalibrering. Merk også at den er kalibrert en gang for alle elever fra alle tre årene.

K-prøven ligger i vanskegrad litt over Del 1 av eksamen alle årene og litt under Del 2. Den fikk en reliabilitet på 0,87, som er et bra resultat. En interessant egenskap ved K-prøven er at den ser ut til å fungere litt bedre for gutter enn jenter, og dette er merkbart ettersom jentene er litt bedre på alle deler av eksamen alle de tre årene. Dette kan ha å gjøre med at K-prøven er elektronisk og kanskje mer motiverende for gutter enn en papirprøve, og i tillegg var den kort (maks to timer), mens eksamen er 5 timer. Disse forholdene, en elektronisk prøve og kort tid, hever vanligvis guttenes prestasjon.

Oppgavene i K-prøven ble ikke pilotert, akkurat som eksamen, og det viser seg at det har slått ut på noen oppgaver som kunne ha vært luket ut i en piloteringsprosess. De 39 oppgavene fungerte ikke alle sammen like bra, noen hadde lav diskriminering, seks av dem hadde en a-parameter under 0,3. K-prøven ville også ha blitt bedre hvis det var noen flere vanskelige oppgaver i den.

K-prøven hadde en korrelasjon på 0,85 med eksamen de tre årene. Informasjonskurven for K-prøven er vist i figur 14.



Figur 14. Informasjon og CSEM for kalibreringsprøven. Maks ved -0,15

Denne informasjonskurven viser det som er nevnt tidligere; K-prøven måler litt svakt på den øvre delen av skalaen, og kunne forbedres med litt flere vanskeligere oppgaver. Da ville den høye målefeilen øverst på skalaen gå ned, kurven ville bli symmetrisk, og prøven ville levere mer informasjon øverst. Ellers kan det konstateres at K-prøven likner mye mer på «vanlige» piloterte prøver enn begge deler av eksamen, med unntak av de oppgavene som diskriminerer dårlig.

K-prøven ser ut til å ha virket bra som en ankerprøve i denne samkalibreringen, og man burde se nærmere på hva det er ved den som gjør at gutter presterer bedre enn jenter, i motsetning til det som skjer på eksamen, hvor jenter er generelt flinkere.

Diskusjon og konklusjoner

Det er ikke meningen å gjenta her de resultatene som allerede er blitt presentert, men sammenfatte dem og peke på noen konsekvenser av denne analysen av eksamen i matematikk for 10. trinn for årene 2017, 2018 og 2019.

Dette prosjektet må ansees å være vellykket, samkalibreringen virket utmerket og resultatene peker på både positive og negative sider ved nåværende eksamen i matematikk, og på hva man i framtiden kan gjøre for å forbedre resultatene. Det anbefales på det sterkeste at eksamen tar i bruk moderne metoder for prøveutvikling og rapportering. Hvis man i fremtiden vil måle utvikling over tid, blir det nødvendig å sette opp et system for det, kanskje med en samkalibrering av to eller flere år hver gang eksamen gjennomføres og utvikling av ankeroppgaver og et system for å gjennomføre dem. Et slikt system trenger ikke å gjennomføres for alle som kommer opp til eksamen, bruk av en representativ gruppe for ankring som på nasjonale prøver ville være mulig. Men en oppgradering av metodene brukt for eksamen er absolutt nødvendig.

Hovedkonklusjonene etter denne analysen kan oppsummeres i noen viktige punkter:

1. Eksamen i sin helhet har ikke endret vanskegrad i de tre årene.
2. Elevenes kompetanse i matematikk har ikke endret seg i de tre årene, med unntak av en liten forbedring fra 2017 til 2018. Denne forbedringen kan muligens være på grunn av utvalgsproblemer i 2018, men selv om denne forbedringen er statistisk signifikant, er den sannsynligvis veldig liten når en ser kvalitativt på den kompetansen elevene viser.
3. Målingen av matematikk-kompetanse er blitt bredere i løpet av de tre årene, og eksamen måler høyere på ferdighetsskalaen i 2019 enn de andre årene.
4. Eksamensdelene innbyrdes har endret seg litt; Del 1 er blitt litt lettere og Del 2 litt vanskeligere.
5. Eksamensdelene er ganske ustabile fra år til år og leverer verken informasjon på samme sted på ferdighetsskalaen hver gang, eller samme dekning av hele ferdigheten (spredning).
6. Dikotome oppgaver på eksamen fungerer meget bra alle årene; vanskegraden er blitt bedre fordelt på ferdighetsskalaen, og oppgavene diskriminerer bra. Det er ganske imponerende at ingen av de dikotome oppgavene hadde for lav diskriminering.
7. En stor del av de polytome oppgavene (66 %) fungerer ikke optimalt og må forbedres, enten ved å endre dem til dikotome oppgaver før poengene for hver elev blir bestemt, eller fjerne oppgavene.
8. En samkalibrering av eksamen over tid ser ut til å virke meget bra, men er avhengig av en god ankerprøve. K-prøven kan forbedres, men den virket etter hensikten. Diskrimineringen til noen oppgaver må forbedres, og prøven trenger noen litt vanskeligere oppgaver for å måle optimalt.
9. Eksamensoppgavene må piloteres, og dette gjelder spesielt de polytome oppgavene som ikke kan konstrueres på en bra måte uten pilotering. Å ikke pilotere disse oppgavene er som et lotteri; man vet aldri om man vinner eller ikke, og som i alle typer av gambling er sannsynligheten for å tape større enn for å vinne (66 %). En pilotering av alle oppgaver ville også bidra til å stabilisere målingen fra år til år, slik at den leverer resultater fra samme kompetansenivå hver gang, og analysen viser klart at dette trenges.

10. Eksamen må ta i bruk moderne metoder (IRT) for oppgaveutvikling og rapportering av resultater.
11. Hvis man velger å måle endring over tid, må det etableres et system for dette. Noe som likner på kalibreringsprøven er brukbart, men dette må anvendes for hele gruppen som tar eksamen, slik at alle får resultater som er så sikre som mulig. Dette vil gi sikre sammenlikninger over tid og sikker estimering av ferdigheten til elevene, og dermed opplysninger om hvordan kompetansen i matematikk endres over tid. Det siste er overordnet viktig nå som alle læreplaner er i endring.

Det er tydelig etter denne gjennomgangen at eksamen i matematikk bygger på gode kunnskaper om hva matematikk er, og hvordan disse kompetansene kan måles. Uten disse grunnleggende ferdighetene ville man ikke ha lykkes med å konstruere såpass bra eksamen uten bruk av psykometriske metoder og pilotering av oppgaver. Det er meget viktig å ta vare på disse kunnskapene og erfaringene for fremtiden. Men det er like viktig å ta i bruk moderne metoder for evaluering av oppgaver, for utvikling av nye oppgaver og for produksjonen av det ferdige resultatet, som er karakteren til hver elev. Nå som alle læreplaner oppdateres er det en opplagt mulighet for å få dette til, som en del av den endringen som er på gang i skolen.

Dette prosjektet og resultatene må evalueres i lys av svakheter ved datagrunnlaget. Det største problemet er utvalgene som ble brukt. Elevene valgte selv å ta k-prøven, og det er mye i disse resultatene som tyder på at utvalget i 2018, ikke var helt sammenliknbart med utvalget fra de to andre årene. Dette gjør resultatene litt mindre sikre, men burde likevel ikke endre helhetsbildet. Nåværende eksamen er heller ikke konstruert for å måle utvikling over tid, og en endret sammensetning av oppgaver og endret organisering av gjennomføringen er sikkert nødvendig for å få til optimale resultater. Det ville være mulig å inkludere nye oppgaver med gjennomføringen av eksamen og også ankeroppgaver for tilfeldig valgte elevgrupper. Dette ville man trenge å gjøre for omtrent 5– 10 % av elevene (6 % på nasjonale prøver). I et slikt system kunne derfor en liten gruppe (6 %) av elevene få ankeroppgaver, og en annen tilsvarende gruppe kunne få nye oppgaver som skulle prøves ut for neste omgang av eksamen.

Det er også et åpent spørsmål om man trenger en så lang prøvetid som nå er brukt, men det er et empirisk spørsmål hvor mye den kan forkortes.

Eksamen i matematikk har allerede mange, meget gode egenskaper som kan benyttes for å bygge videre et stabilt og moderne vurderingsverktøy, som kan produsere nyttige opplysninger om prestasjonen til hver elev. Dette må i fremtiden baseres på moderne psykometriske metoder sammen med solide kunnskaper om matematikk, samt den lange erfaringen i disse målingene, noe som er uvurderlig å ha med ved oppdatering av eksamenssystemet. Her begynner man ikke på bar bakke, men kan bygge på lange tradisjoner fra eksamen og brede og solide kunnskaper om matematikk.

Referanser:

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 397-479). London, UK: Addison-Wesley.

Björnsson, J. K. (2018). "Om lenkefeil og ekvivaleringsmetoder på nasjonale prøver: Evaluering av endring over tid." Acta didactica Norge [elektronisk ressurs] 12(4): 24-24.

Embretson, S. E. & Reise, S.P. (2013). Item Response Theory for Psychologists. Hoboken, Taylor and Francis.

Guyer, R., & Thompson, N. A. (2014). User's Manual for Xcalibre item response theory calibration software, version 4.2.2 and later. Woodbury MN: Assessment Systems Corporation.

Kolen, M. J. and R. L. Brennan (2014). Test Equating, Scaling, and Linking: Methods and Practices. New York, NY, Springer New York.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, N.J, Lawrence Erlbaum.

Martin, M.O., Mullis, I.V.S., Foy, P., Brossman, B. & Stanco, G.M. (2012). Estimating linking error in PIRLS. IERI Monograph, Vol. 5. Ch. 2. IERI Institute.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. Applied Psychological Measurement 16, 159–176.

Tan, X. & Michel, R. (2011). Why Do Standardized Testing Programs Report Scaled Scores? Why Not Just Report the Raw or Percent-Correct Scores? ETS: R&D Connections, No. 16, September.

Vedlegg: Rapporter fra samkalibreringen gjort med Xcalibre version 4,2.