Universitetet i Stavanger

DET TEKNISK-NATURVITENSKAPELIGE FAKULTET

MASTEROPPGAVE

Studieprogram/spesialisering: **Lektor realfag 8.-13.trinn**	Vårsemesteret, 2021 Åpen
Forfatter: **Håkon Berg Johannessen**	 Signatur forfatter
Fagansvarlig: Veileder(e): **Jan Terje Kvaløy**	
Tittel på masteroppgaven: **Statistiske metoder for analyse av data fra nasjonale prøver og undersøkelser i skoleverket.** Engelsk tittel: **Statistical analysis of data from the Norwegian national tests and research in the school system.**	
Studiepoeng: **30**	
Emneord: **Statistikk** **Psykometri** **Nasjonale prøver** **Item Response Theory**	Sidetall: **63** + vedlegg/annet: **5** Stavanger, **14.06.21**

Forord

Jeg vil takke min veileder Jan Terje Kvaløy for god støtte, nyttige tilbakemeldinger og interresante idéer gjennom hele masteroppgaven. Gjennom ukentlige møter har han hjulpet meg å holde fokus og motivasjon oppe. Jeg vil også rette en takk til Julius K. Bjørnsson ved Institutt for lærerutdanning og skoleforskning ved Universitetet i Oslo for tilgang på rådata fra nasjonale prøver, for hans entusiasme for prosjektet og at han stilte til rådighet med sin ekspertise.

Håkon Berg Johannessen

Sammendrag

Nasjonale prøver ble for første gang gjennomført våren 2004. En forskningsbasert evaluering av prøvene i 2005 anbefalte sterkt en kraftig kompetanseheving på testteori og psykometri for de som utvikler og for de som leder utviklingen av prøvene på nasjonalt nivå. Som konsekvens ble det fra og med 2014 introdusert Item Response Theory (IRT) som grunnlag for analyse av data fra nasjonale prøver. Denne oppgaven tar for seg grunnleggende IRT samt elementer av IRT som brukes i analyse av nasjonale prøver. Den tar også for seg DIF-analyse som brukes i piloteringen og IRT-metoder for lenking av prøvene som gjør det mulig å sammenligne resultater over tid. Deretter simuleres resultater fra nasjonale prøver for enkeltelever, klasser og skoler for å studere egenskapene til prøvene.

Det kommer fram at IRT-metodene som ble tatt i bruk i 2014 var en nødvendig og stor forbedring fra de tidligere brukte metodene. Det var før 2014 umulig å vite om endringene i resultater fra år til år skyldtes endring i prøvene eller endring i elevenes ferdighet. Med IRT-metoder for skalering av prøvene og ved bruk av ankeroppgaver ble det slik at samme tall beskriver samme ferdighet, til tross for at målingen er foretatt ved ulike prøver. Det ble fra 2014 også mulig å sammenligne resultater over tid.

Elevenes skåre på prøvene uttrykkes på en skala med nasjonalt gjennomsnitt på 50 skalapoeng og standardavvik 10. Simuleringene for enkeltelever viser at prøven er mest persis rundt gjennomsnittet på skalaen og det er noe høy usikkerhet ved antall skalapoeng eleven får rapportert. Simuleringene for klasser og skoler tydeliggjør at ved rapportering av gjennomsnittskåre for klassen eller skolen er det én usikkerhet i forhold til det prøven måler, og en annen usikkerhet knyttet til naturlig variasjon i populasjonen. Prøvene måler gjennomsnittet for en klasse mer presist jo større klassen er, og enda mer presist for skoler jo større de er. Den usikkerheten som rapporteres for gjennomsnittet til en skole på utdanningsdirektoratet sine sider er knyttet til naturlig variasjon i ferdigheten til elevgruppen på skolen, og blir med det høyere enn usikkerheten til ferdigheten prøven måler for akkurat den elevgruppen.

Innhold

1 Introduksjon

Nasjonale prøver ble for første gang gjennom ført i 2004. Formålet var å gi informasjon til lærere, skoleeiere og ledere om kvaliteten på utdanningstilbudet på lærestedet, samt å gi informasjon til den enkelte elev som grunnlag for videre utvikling. Prøvene høstet mye kritikk i media de første årene og aviser la ut rangeringslister over "beste" og "verste" skoler helt uten statistisk grunnlag. En rapport som i 2005 analyserte og vurderte kvaliteten på nasjonale prøver fant store mangler på den testteoretisk kompetansen som lå til grunn for utvikling og gjennomføring av prøvene. "Sammenliknet med andre land er det en forbløffene mangel på både grunnleggende og avansert psykometrisk kompetanse, og dette har gjenspeilet seg i forbindelse med årets prøver" (Bjørnsson, Caspersen & Lie, 2005). Det ble fra 2014 tatt i bruk mer moderne psykometriske metoder, Item Response Theory (IRT), som grunnlag for utvikling og rapportering av resultater for prøvene.

Denne oppgaven tar for seg grunnleggende momenter med Item Response Theory og de momenter som er relevante for nasjonale prøver. Deretter simuleres resultater på nasjonale prøver for enkeltelever, klasser og skoler for å få en bedre forståelse av resultatene og usikkerheten rundt disse. Til slutt diskuteres hva resultatene fra simuleringene forteller om bruk av IRT-metoder i nasjonale prøver.

Kapittel 2 tar for seg grunnleggende aspekter ved nasjonale prøver som kan være nyttig for forståelse for eksemplene som brukes i påfølgende kapitler, samt motivasjonen for simuleringene i kapittel 5. Kapittel 3 presenterer det grunnleggende ved IRT og aspekter som er nyttige for nasjonale prøver. Kapittel 4 tar for seg DIF-analyse, som brukes i piloteringen av nasjonale prøver. Kapittel 5 går gjennom enkelte metoder for lenking og ekvivalering av prøver, som er nødvendig for å kunne måle sammenheng over tid. Kapittel 6 viser hvordan en kan simulere resultater på nasjonale prøver for enkeltelever, klasser og skoler, og ved det studere egenskaper til nasjonale prøver.

2 Nasjonale prøver

Nasjonale prøver er prøver i lesing, engelsk og regning som gjennomføres på 5., 8. og 9. trinn ved barne og ungdomsskoler i Norge. Prøvene er et element i et nasjonalt kvalitetsvurderingssystem og ble først gjennomført våren 2004. Oppgavene utarbeides med egnede psykometriske metoder. Siden 2014 har Item Response Theory (IRT) vært basis for utvikling av oppgaver samt gjennomføring, bearbeiding og analyse av resultatene i de nasjonale prøvene i sin helhet.

Resultatene blir bearbeidet av utdanningsdirektoratet og publisert på en skala med gjennomsnitt på 50 skalapoeng og et standardavvik på 10. Elevene blir i tillegg delt inn i tre mestringsnivå på 5. trinn og fem mestringsnivå på 8. og 9. trinn. Hvor mange elever som er innenfor hvert mestringsnivå publiseres sammen med poengene. Disse resultatene er tilgjengelig for offentligheten på nasjonalt nivå, fylkesnivå, kommunalt nivå og skolenivå. Lærere, skoleledelse og utdanningsdirektoratet har tilgang til resultater på elevnivå (Udir, 2017d).

2.1 Prøvenes formål

Nasjonalt kvalitetsvurderingssystem (NKVS) ble etablert som en del av kunnskapsløftet i 2004. NKVS ligger til grunn når kunnskapsdeptartementet skal utvikle nye tiltak og ny skolepolitikk. I dag omfatter NKVS blant annet eksamensresultater, karakterstatistikk, elevundersøkelser, internasjonale studier og nasjonale prøver. Før 2004 hadde man data fra eksamener og standpunktkarakterer. Ulempen med å bruke tall fra eksamener er at prøvene kan være forskjellige fra år til år og sier lite om endring over tid. Standpunktkarakterer har den ulempen at det kan være underliggende forskjeller fra skole til skole. Samlet sett gir dette ikke gode nok statistiske grunnlag for å vurdere kvaliteten på utdanningen i Norge eller å sette nye tiltak. Det er her nasjonale prøver er satt inn som et tilskudd som kan gi relevante data for elevenes nivå over tid og på tvers av skoler (Sjøberg, 2019).

Nasjonale prøver kartlegger elevenes ferdigheter i regning og lesing på 5., 8. og 9. trinn, og engelsk på 5. og 8. trinn. De samme prøvene gis på alle skoler i hele landet, og prøvene er obligatorisk for alle elever. Det kan gis fritak til elever som har rett til spesialundervisning, tilrettelagt norskunder-

visning eller lignende tiltak. Andelen elever som får fritak har økt jevnt fra 1.1% i regning, 1.5% i engelsk og 1.9% i lesing i 2009, til 3.4% i regning 3.8% i engelsk og 3.7% i lesing i 2019 (Udir, 2019). Nasjonale prøver skal være et redskap for lærere og skoleledelsen som brukes til å vurdere og utvikle kvaliteten på utdanningstilbudet på skolen. Prøvene skal også være et element i vurderingen av grunnopplæringen i Norge på nasjonalt nivå.

2.2 Innhold

I kunnskapsløftet (LK06) definerer læreplanverket fem grunnleggende ferdigheter: digitale ferdigheter, muntlige ferdigheter, å kunne lese; å kunne regne og å kunne skrive. Disse ferdighetene gjelder på tvers av fagene. Rammeverket for de grunnleggende ferdighetene ble revidert av Kunnskapsdepartementet i november 2017 i forbindelse med fagfornyelsen. Ferdigheten å kunne regne gjelder ikke bare matematikkfaget, men viser til at eleven skal ha regneferdigheter som er tilstrekkelig til å nå kompetansemålene i alle fag. Det er denne regneferdigheten den nasjonale prøven i regning er ment å måle. Tilsvarende skal den nasjonale prøven i lesing måle lesing som ferdighet på tvers av fagene. Det er da viktig å understreke at den nasjonale prøven i regning ikke er en prøve i matematikkfaget, og den nasjonale prøven i lesing ikke er en prøve i norskfaget (Udir, 2017c).

2.2.1 Nasjonale prøver i regning

Den grunnleggende ferdigheten å regne er delt inn i fire ferdighetsområder som til sammen skal beskrive en problemløsningsprosess: gjenkjenne og beskrive, bruke og bearbeide, reflektere og vurdere, og kommunisere. Dette sammen med temaene tall, statistikk, og måling og geometri danner grunnlaget for utarbeiding av oppgaver til prøven for 5. trinn. Mer sentralt inneholder prøvene for 5. trinn oppgaver der de skal beskrive og gjenkjenne konkrete situasjoner fra virkeligheten der matematikk er involvert i både kjente og ukjente kontekster. Eksempelvis matlaging, kjøp og salg, reise, idrett eller kontekster knyttet til andre fag. For 8. og 9. trinn tar prøven utgangspunkt i kompetansemålene etter 7. trinn og inneholder prøvene algebra og sannsynlighet, i tillegg til temaene fra 5. trinn. Prøven for 9. trinn er den samme som for 8. trinn. (Udir, 2017b).

2.2.2 Nasjonale prøver i lesing

Å kunne lese handler om å kunne finne informasjon i, tolke og å reflektere over teksters form og innhold. Med tekster menes her alt som kan leses i ulike medier. Det vil i tillegg til ord inkludere illustrasjoner, symboler, grafiske framstillinger og andre uttrykksmåker. Tekstene som blir gitt som oppgaver i prøvene skal da representere de ulike tekstene elevene møter i de ulike fagene. Oppgavene inneholder både skjønnlitterære tekster og sakprosatekster. Dermed vil noen oppgaver bestå av å lete etter konkret informasjon i teksten, mens andre ber leseren tolke eller foreta en vurdering av teksten for å skape mening. Prøven inneholder flervalgsoppgaver, noen med fire alternativer, noen med to, samt åpne oppgaver der eleven svarer med egne ord (Udir, 2017d).

2.2.3 Nasjonale prøver i engelsk

Nasjonale prøver i engelsk måler i elevenes ferdigheter i engelesk, der ferdigheter i engelsk viser til kompetansemålene i læreplanen for engelesk: leseforståelse, ordforråd, begreper og grammatikk. Prøvene på 5. trinn ber elevene hente ut informasjon, forstå enkeltord og hovedinnhold i enkle tekster, forstå vanlige ord og uttrykk knyttet til dagligliv og fritid samt enkel grammatikk. Tekstene er fra et par setninger til rundt 250 ord. På 8. trinn utvides de samme kravene til mer omfattene tekster som inneholder historiske personer og engelskspråklig kultur. Elevene må i tillegg reflektere over innholdet i tekster. Tekstene på 8. trinn er fra et par setninger til rundt 400 ord (Udir, 2017a).

2.3 Gjennomføring

Det er kommunen som har det overordnede ansvaret for gjennomføringen av nasjonale prøver ved offentlige skoler. Dette innebærer blant annet å kontrollere at gjennomføringen av nasjonale prøver skjer korrekt og forsvarlig, følge opp resultatene på kommunenivå og se til at skolene følger opp og bruker resultatene. Det er skoleleder som er ansvarlig for gjennomføringen av nasjonale prøver på sin skole. Dette innebærer blant annet å sørge for at foreldre informeres om gjennomføring og resultat, legge til rette for at lærere følger opp resultatene i klassen og fatter enkeltvedtak for elever som fritas. Læreren er ansvarlig for å sørge for at elevene er kjent med oppgavetypene og bruk

av resultatene i tilbakemeldinger til elever og foreldre. Det er ikke meningen at elevene skal øve på prøvene. Kvaliteten på resultatene fra nasjonale prøver avhenger av at alle skolene følger retningslinjene for gjennomføringen av nasjonale prøver (Udir, 2018).

2.3.1 Fritak

I utgangspunktet er nasjonale prøver obligatoriske for alle elever, men skolene kan vurdere å innvilge fritak til elever som enten har vedtak om spesialundervisning eller særskilt språkopplæring. Det er rektor i samråd med elevens lærer som avgjør om eleven skal få fritak. Det skal vurderes individuelt for hver enkelt prøve om eleven skal fritas. Fritaket er et enkeltvedtak med tre ukers klagerett. Elever som mottar privat hjemmeundervisning har ikke krav om å delta på nasjonale prøver (Udir, 2018).

Internasjonale skoler eller skoler med alternative læreplaner kan søke om å legge de nasjonale prøvene på andre trinn eller om fritak fra prøvene. Dette er aktuelt dersom elevene ikke har fått tilstrekkelig opplæring i de kompetansemål som nasjonale prøver tester (Udir, 2018).

2.4 Rapportering av resultater

Det ble i 2014 utviklet en egen skala til nasjonale prøver hvor gjennomsnittet ble satt til 50 og standardavviket til 10. Alle resultatene fra påfølgende år blir kalibrert til den samme skalaen. Det medfører at dette gjennomsnittet kan endre seg over tid dersom elevenes ferdigheter endrer seg.
Resultatene fra de nasjonale prøvene skal følges opp på skolenivå og på kommunalt nivå. Prøveutviklerne skal utvikle analyserapporter som inneholder resultater for skolen, elevgruppene som har gjennomført prøven, og for hver enkelt elev. Skoler kan sammenlignes på nasjonalt nivå, og de kan se hvor mange elever som ligger innenfor hvert mestringsnivå. Lærere får beskrivelser for hver enkelt elev hvor på mestringsnivået eleven ligger, hvilke oppgaver som hører til hvert mestringsnivå og tema, og hvilke oppgaver eleven har svart rett eller galt på (Udir, 2017d).

2.4.1 Mestringsnivåer

Resultatene for nasjonale prøver blir delt inn i ulike mestringsnivåer. Det er tre mestringsnivåer for 5. trinn og fem mestringsnivåer for 8. trinn. Grensene for disse mestringsnivåene ble fastsatt etter en prosentil-fordeling etter gjennomføringen i 2014:

5. trinn: 25 - 50 - 25

8. trinn: 10 - 20 - 40 - 20 - 10

Dette ga følgende poenggrenser for hvert mestringsnivå i regning på 5. trinn:

Mestringsnivå 1: til og med 42

Mestringsnivå 2: 43 til 56

Mestringsnivå 3: 57 og høyere

Og følgende i regning på 8. trinn:

Mestringsnivå 1: til og med 36

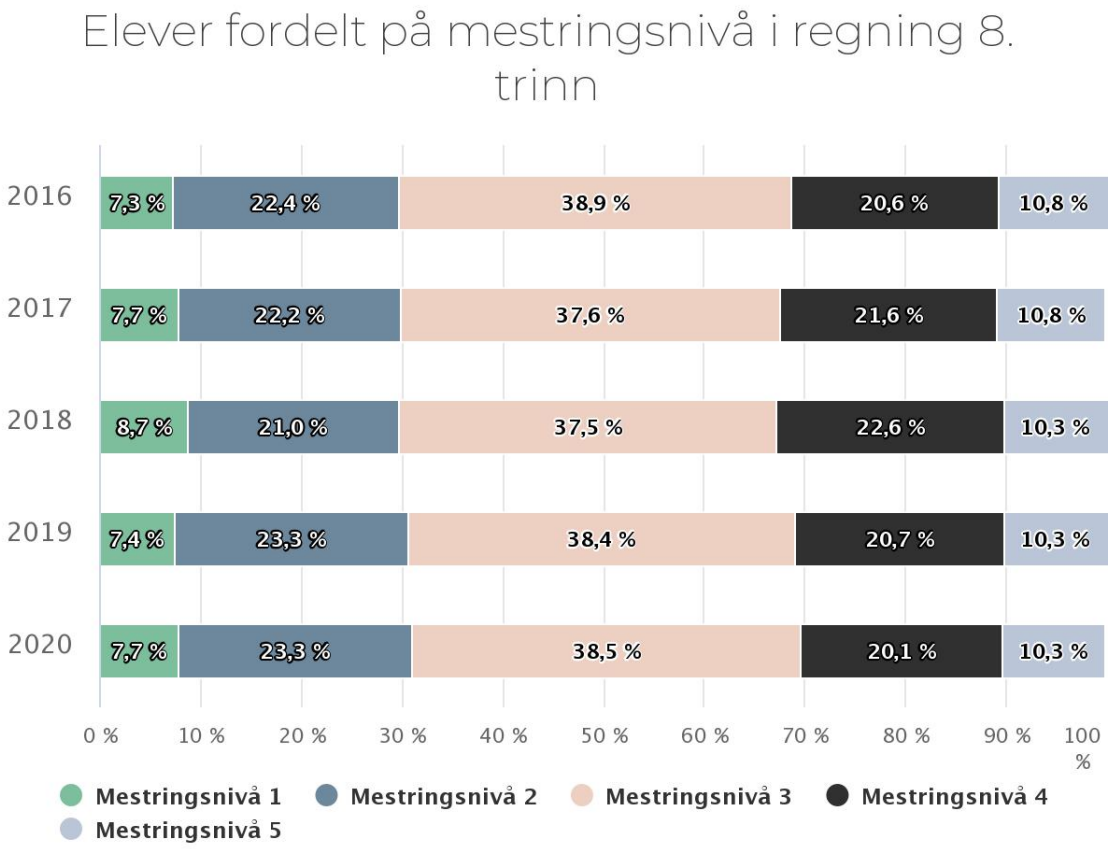
Mestringsnivå 2: 37 til 44

Mestringsnivå 3: 45 til 54

Mestringsnivå 4: 55 til 62

Mestringsnivå 5: 63 og høyere

På grunn av at poenggrensene ble satt etter en prosentil-fordeling vil de være noe ulike for regning, lesing og engelsk. Nøyaktig hvor poenggrensene går for hvert disiplin sammen med detaljerte beskrivelser av hva hvert mestringsnivå måler finnes på Udtanningsdirektoratet sine sider (Udir, 2017). Figur 1 viser et eksempel på hvordan Utdanningsdirektoratet rapporterer mestringsnivå.



Figur 1: Elever fordelt på mestringsnivå i regning 8. trinn. Nasjonalt nivå fra 2016 til 2020.

3 Item Response Theory

Før midten av 1980-tallet var psykometri og Educational Measurement i stor grad basert på klassisk test teori. Klassisk test teori ble utviklet på 1920-tallet og har en del svakheter sammenlignet med mer moderne teorier (Baker & Kim, 2017). En av de mest åpenbare svakhetene med klassisk test teori for måling av elevers ferdighet er at den ikke er i stand til å predikere hvordan en gruppe elever svarer på en spesifikk oppgave med mindre samme oppgave allerede har vært testet på en gruppe elever på samme ferdighetsnivå. Altså er vurdering av oppgavenes vanskegrad avhengig av elevenes ferdighetsnivå. Den klassiske test teorien tar ingen forutsetninger om parametere som er ute av kontroll for den som administrerer testen. Det er svakheter som dette som er bakgrunnen for at det ble utviklet en mer statistisk robust teori for kvantifisering av hvordan mennesker i den virkelige verden svarer på oppgaver og tester. Denne teorien, *Item Response Theory*, må kunne gjøre oss i stand til å estimere elevers ferdighetsnivå uavhengig av hvem som har tatt oppgaven før dem. Vi må også kunne vite noe om hvor godt de ulike oppgavene i testen skiller mellom elever på ulike ferdighetsnivå, altså oppgavens diskriminering (Lord, 1980).

Den underliggende matematikken bak Item Response Theory (IRT) er mer avansert enn i den klassiske test teorien. Det kreves flere utregninger for hver oppgave i hver prøve. Som følge av dette kreves det datakraft for storskala gjennomføring av analyser basert på IRT, og det var da slik datakraft ble tilgjengelig for folk flest at IRT ble standarden for bruk i psykometri. Den mest vanlige bruken av IRT er innen Educational Measurement. IRT brukes i dag til å utvikle prøver og eksamener, spesielt for å se på endringer av elevers ferdigheter over lengre tid (Hambleton, Swaminathan & Rogers, 1991). Fra og med 2014 ble IRT bakgrunn for analyse og rapportering av resultater fra nasjonale prøver (Bjørnsson, 2018).

IRT er en samling statistiske modeller brukt for å analysere data i enkeltoppgaver samt tester og prøver i sin helhet. IRT estimerer ulike parametere for hver oppgave og for prøvetakeren og hvordan disse parameterne henger sammen med beskrivelser av enkeltoppgaver og prøvetakerens resultat. Brukt riktig vil IRT kunne gi et mer nøyaktig bilde av det testen er ment å måle. IRT er bakgrunnen for oppgavekonstruksjon, prøvekonstruksjon, skalering, sammensetning av prøver over tid, skåring og rapportering av skår. Bak-

grunnen for en hver IRT-modell er en beskrivelse av sannsynligheten for hva en prøvetaker med visse egenskaper svarer på en oppgave med visse parametere. En av de store fordelene med IRT framfor klassisk test teori er at oppgaveparameterne er uavhengige av ferdighetsnivå til prøvetakerne som svarer på oppgaven. Dette illustreres i delkapittel 3.5 etter at oppgaveparameterne er godt definert.

De ulike IRT modellene kan variere i hvilke egenskaper de beskriver oppgaver og prøvetaker med, og egenskapene i den matematiske funksjonen som beskriver forholdet mellom de. Modellene kan være dikotome eller polytome. En dikotom modell beskriver prøvetakerens svar på en oppgave binært (0 for galt svar, 1 for rett svar). Mens en polytom modell vil ha en gradert skalering av prøvetakerens svar på hver oppgave. Videre kan modellene være udimensjonelle eller flerdimensjonelle. En udimensjonell modell beregner sannsynligheten for rett svar i for eksempel regning, basert på prøvetakerens ferdighet i kun regning. Mens en todimensjonell modell vil beregne sannsynligheten for rett svar i regning basert både på prøvetakerens ferdighet i regning i tillegg til for eksempel i lesing. Denne oppgaven tar i hovedsak for seg dikotome, udimensjonelle modeller. Funksjonene som beskriver forholdet mellom prøvetakerens ferdighet og sannsynligheten for rett svar i disse modellene kalles ofte item response functions eller item characteristic curve (Brennan, 2006).

3.1 Item Characteristic Curve

En av de underliggende egenskapene hos et menneske en ønsker å måle i psykometri og Educational Measurement er elevers ferdigheter i ulike emner. For eksempel er nasjonale prøver ment å kartlegge hvilket nivå elever ligger på i engelsk, lesing og regning. I denne oppgaven brukes begrepet ferdighet generelt om det en ønsker å måle hos elever. Videre antas det at denne ferdigheten kan settes på en skala hvor 0 er midtpunktet og teoretisk ferdighet kan være fra $-\infty$ til $+\infty$. Men av praktiske årsaker, og på grunn av hvordan vi definerer parameterne til oppgavene vil denne ferdigheten som oftest ligge innenfor intervallet -3 til $+3$ (Baker & Kim, 2017). Ferdigheten som måles gis symbolet θ, og en tenker seg at θ ofte er fordelt tilnærmet lik en standard normalfordeling. For hvert ferdighetsnivå tilhører det en sannsynlighet for at eleven svarer rett på oppgaven, denne sannsynligheten blir da $P(\theta)$.

$$P_i(\theta) = P(X_i = x_i \mid \theta,\ \delta_i) \tag{1}$$

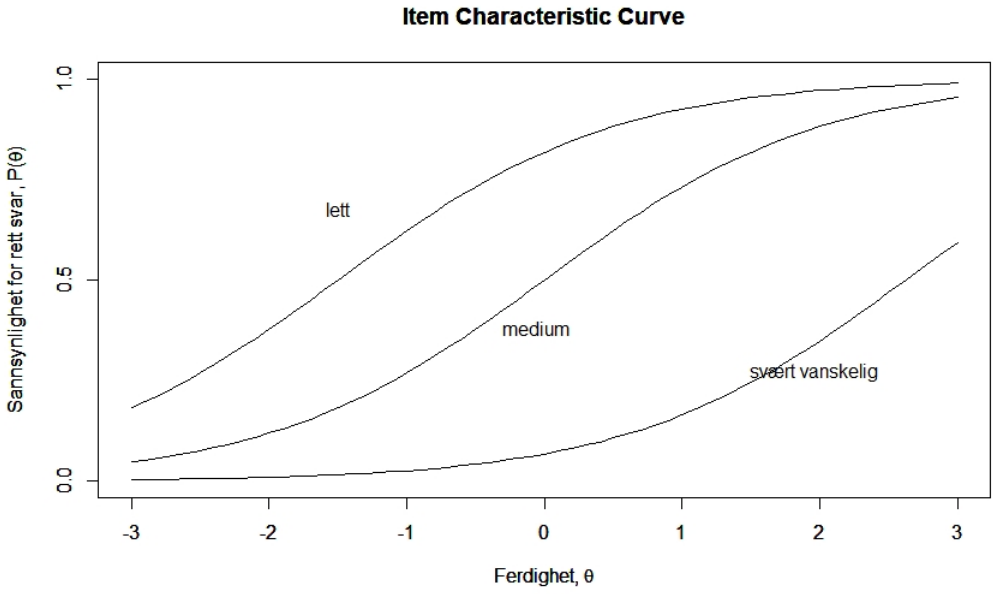
Ligning 1 kalles *item response funksjonen* (IRF) og gir oss sannsynligheten for at prøvetakeren gir svar x_i på den i-ende oppgaven gitt prøvetakerens ferdighetsnivå θ og oppgavens parametere δ_i. Videre gir *item characteristic function* (ICF) og dens grafiske representasjon *item characteristic curve* (ICC) forventet skår for en oppgave(item) som funksjon av ferdighet. For dikotome oppgaver er forventet skår 0 ganger sannsynligheten for galt svar pluss 1 ganger sannsynligheten for rett svar. Dermed er IRF lik ICF for dikotome modeller og uttrykkene kan brukes om hverandre (Brennan, 2006). Den grafiske representasjonen av forventet skår på en oppgave som funksjon av ferdighet vil i denne oppgaven kalles ICC. Denne sammenhengen presenteres i ulike modeller.

3.1.1 Rasch-modellen (1PL)

Rasch-modellen eller *en-parameter modellen* er den enkleste og en av de mest brukte IRT-modellene og beskrives av følgende ligning:

$$P_i(\theta) = \frac{1}{1 + e^{-(\theta - b_i)}} \tag{2}$$

I ligning 2 er θ som tidligere nevnt definert som prøvetakerens ferdighet. Parameteren b beskriver oppgavens vanskegrad og har høyere verdi ettersom vanskegraden på oppgavene i testen øker. I Rasch-modellen er b definert som punktet hvor prøvetakeren har 0.50 sannsynlighet for å svare rett dersom ferdighetsnivået matcher vanskegraden. Det er også det punktet hvor stigningstallet til ICC-kurven er på sitt høyeste (0.25). Eller mer presist: når $\theta = b$ er $P(\theta) = 0.5$ og $P'(b) = maxP'(\theta) = 0.25$. En av de store fordelene med IRT er at ferdighet og vanskegrad måles på samme skala, en annen stor fordel er at vanskegradsparameteren b er teoretisk uavhengig av ferdighetsnivået til prøvetakerne oppgavene er testet på (Brennan, 2006). For å gi en intuitiv forståelse av b parameteren listes eksempler på b verdier med tilhørende verbale uttrykk (Baker & Kim, 2017):



Figur 2: Rasch-ICC for tre oppgaver med ulik vanskegrad.

Vanskegrad: b
svært lett: -2.625
lett: -1.5
medium: 0
vanskelig: 1.5
svært vanskelig: 2.625

Figur 2 plotter Rasch-modell ICC for tre oppgaver med ulik b parameter (ulik vanskegrad). Vi kan se at jo vanskeligere oppgaven er jo høyere ferdighet må prøvetakeren ha for å ha 50% sannsynlighet for rett svar. Det er verdt å merke seg at ved det høyeste ferdighetsnivået vist ($\theta = 3$) er $P(\theta) = 0.6$ for en svært vanskelig oppgave. Rasch-modellen er en effektiv modell i den forstand at den trenger kun én parameter for å beskrive store mengder data. Andre IRT-modeller inkluderer flere parametere men felles for alle er at de beskriver store mengder data svært økonomisk (Brennan, 2006).

3.1.2 To-parameter modellen (2PL)

To-parameter modellen er en utvidelse av Rasch-modellen ved at den inkluderer en ekstra parameter for å beskrive oppgavene. Det er denne modellen som brukes av utdanningsdirektoratet i analysen av nasjonale prøver (Bjørnsson, 2018). Modellen beskrives av følgende ligning:

$$P_i(\theta) = \frac{1}{1 + e^{-Da_i(\theta - b_i)}} \tag{3}$$

I ligning 3 er θ prøvetakerens ferdighet, a og b parametere som beskriver oppgaven og D er en konstant som brukes for å approksimere egenskapene til en kumulativ normalfordeling (Lord, 1980). D-konstantens verdi avhenger av hvilken modell som brukes og det er personlig preferanse hos den som gjennomfører IRT-analysen som avgjør om D brukes eller ikke. Det blir teknisk sett rett å si at i Rasch-modellen er $D = 1.0$ og $a_i = 1$ for alle oppgaver, men i praksis blir de utelatt i beskrivelser av modellen. For to- og tre-parameter modellene er det blitt standard å bruke $D = 1.7$ eller $D = 1.702$. Det er viktig når man gjennomfører en IRT-analyse å presisere om man bruker D eller ikke, om ikke D defineres presist vil det kunne føre til unøyaktigheter (Brennan, 2006).

Parameter b er oppgavens vanskegrad og er i likhet med Rasch-modellen definert som den ferdigheten som gir $P(\theta) = 0.5$ men der er nå en parameter a som har effekt på stigningstallet til kurven:

$$P'(\theta) = \frac{-(-Da_i)e^{-Da_i(\theta - b_i)}}{(1 + e^{-Da_i(\theta - b_i)})^2} \tag{4}$$

For å vise hvor kurven er brattest setter vi:

$$x = e^{-Da_i(\theta - b_i)},$$

$$P'(\theta) = Da_i \frac{x}{(1+x)^2} \tag{5}$$

$$\text{setter } f(x) = \frac{x}{(1+x)^2},$$

$$f'(x) = \frac{1-x^2}{(1+x)^4} \tag{6}$$

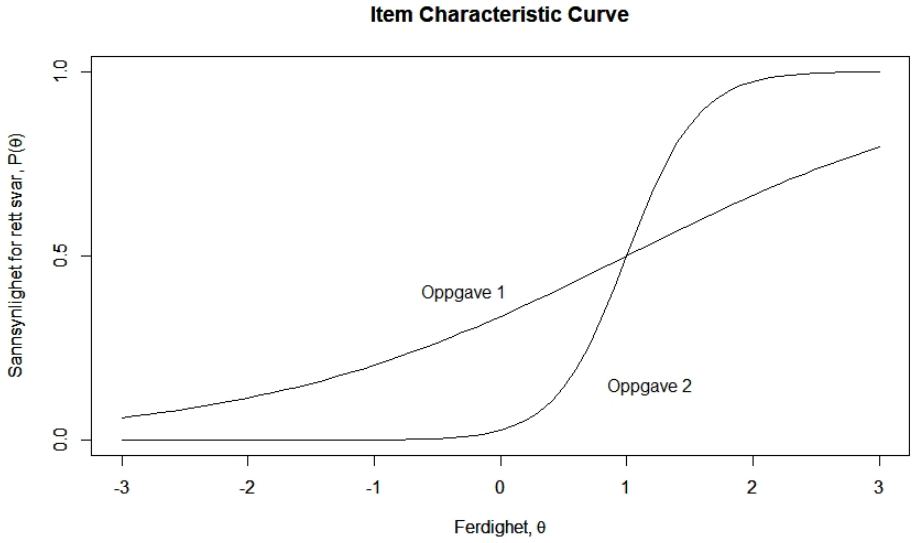
$$f('x) = 0 \text{ når } x = 1 \Leftrightarrow \theta = b;$$

$$\Rightarrow f(1) = \frac{1}{4}, \Rightarrow max\ P'(\theta) = \frac{Da_i}{4} \tag{7}$$

Parameter a er oppgavens diskriminering, altså hvor godt oppgaven skiller mellom lavere og høyere ferdighet. Vi ser fra utledningen over at stigningstallet til ICC-kurven i vendepunktet er proporsjonal med a-parameteren, hvor $P'(\theta) = (\frac{Da_i}{4})$. En ICC-kurve til en oppgave med høy diskriminering, det vil si høy a, vil da være bratt der $P(\theta) = 0.5$. I likhet med vanskegradsparameteren listes noen eksempler på a-verdier med numeriske verdier og tilhørende verbale uttrykk der $D = 1$ (Baker & Kim, 2017):

Diskriminering: a
ingen: 0
lav: 0.4
moderat: 1
høy: 2.1
perfekt: 999

Figur 3 plotter to oppgaver med samme vanskegrad ($b = 1$) og ulik diskriminering ($a_1 = 0.4, a_2 = 2.1$ og $D = 1.702$). Det kan være verd å merke seg at ved IRT-modeller som inkluderer diskrimineringsparameteren kan ICC-kurvene skjære hverandre. Det vil si at relativ vanskegrad på oppgaven kan være ulik for prøvetakere med ulik ferdighet (Brennan, 2006). I figur 3 vil en prøvetaker med $\theta = 0$ oppleve at oppgave 1 er lettere enn oppgave 2, mens en prøvetaker med ferdighet $\theta = 2$ vil oppleve oppgave 2 lettere enn oppgave 1 selv om oppgavene er av samme vanskegrad ($b = 1$).



Figur 3: 2PL-ICC for to oppgaver med samme vanskegrad og ulik diskriminering.$D = 1.702$.
Oppgave 1: $a = 0.4,\ b = 1$
Oppgave 2: $a = 2.1,\ b = 1$

Har diskrimineringsparameteren verdi $a = 0$ tilsvarer dette ingen diskriminering. I praksis betyr det at oppgaven ikke er i stand til å skille mellom ferdighetsnivå i det hele tatt og ICC-kurven blir følgelig en horisontal linje ved $P(\theta) = 0.5$ uansett vanskegrad. Perfekt diskriminering ved for eksempel $b = 2$ betyr i praksis at alle prøvetakere med $\theta < 2$ vil ha 0% sannsynlighet for rett svar, mens alle prøvetakere med $\theta > 2$ vil ha 100% sannsynlighet for korrekt svar, og ICC-kurven blir vertikal ved $\theta = 2$. Når a går mot uendelig går $P'(\theta)$ mot uendelig. Diskrimineringsparameteren kan i teorien ha verdier fra $-\infty$ til $+\infty$, men en $a < 0$ betyr i praksis at når θ synker, øker $P(\theta)$, altså vil en prøvetaker med lavere ferdighet ha høyere sannsynlighet for rett svar. Oppdager man at en oppgave i en test har negativ diskriminering bør dette undersøkes, det kan ofte bety at oppgaven bør forkastes (De Ayala, 2009).

3.1.3 Tre-parameter modellen (3PL)

En- og to-parameter modellene antar at sannsynligheten for rett svar nærmer seg 0 etter hvert som prøvetakerens ferdighet synker. I for eksempel prøver med svaralternativer er det mulig å gjette seg fram til rett svar, til tross for at prøvetaker ikke har tilstrekkelig ferdighet til å svare på oppgaven. Tre-parameter modellen tar hensyn til denne gjettingen. Tre-parameter modellen beskrives av ligningen:

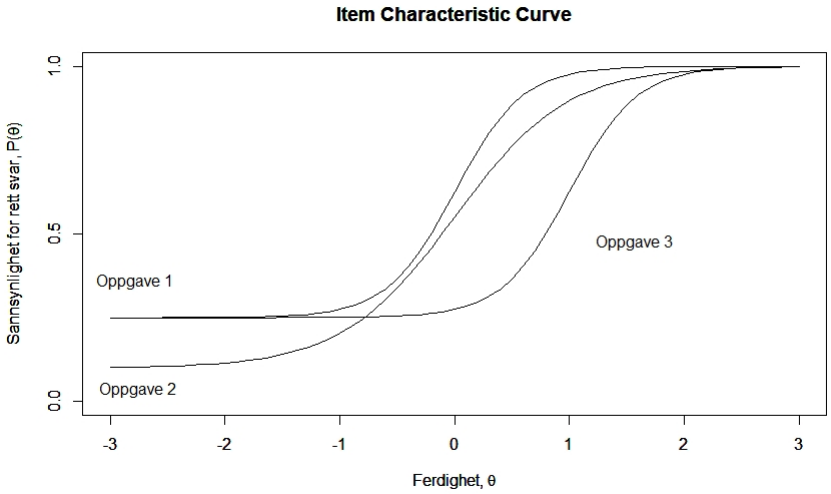
$$P_i(\theta) = c_i + \frac{1 - c_i}{1 + e^{-Da_i(\theta - b_i)}} \tag{8}$$

I ligning 8 er θ forstatt prøvetakerens ferdighet, a, b og c er parametere som beskriver oppgaven og D er konstanten for å approksimere kumulativ normalfordeling. Parameteren c er definert som sjansen for at en prøvetaker med $\theta = -\infty$ vil avgi korrekt svar på en oppgave. En kan si at to-parameter modellen er lik tre-parameter modellen med $c = 0$ (Brennan, 2006).

Parameteren c er ofte kalt gjetteparameteren eller psuedo-gjetteparameteren. Hvor psuedo kommer av at i en flervalgsoppgave er det ofte ikke ren gjetting. Teoretiske verdier for parameteren c er $0 \leq c \leq 1$, men i praksis aksepteres ikke verdier over 0.35 (Baker & Kim, 2017).

Figur 4 viser 3PL-ICC kurver for tre oppgaver med ulike parametere. Sammenligner man oppgave 1 og oppgave 3 ser man at å endre b parameteren har effekten å forskyve kurven horisontalt. En definisjon av parameteren c i 3PL-modellen, som kommer fram i figur 4, er at den er lik kurvens nedre asymptote (Brennan, 2006).

En konsekvens av å inkludere c i modellen er at definisjonen på b parameteren endres. Tidligere var b definert som det punktet på ferdighetsskalaen som tilsvarer $P(\theta) = 0.5$. Nå som nedre grense for $P(\theta) = c$, blir b definert som punktet på ferdighetssakalaen hvor $P(\theta) = (1 + c)/2$. Som følge av dette blir nå kurvens bratteste punkt $\frac{D}{4}a(1 - c)$. Selvom introduksjonen av c parameteren gir relativt små endringer i definisjonen av de andre parameterne kan det ha stor betydning for tolking av resultatene av IRT-analyser (Baker & Kim, 2017).



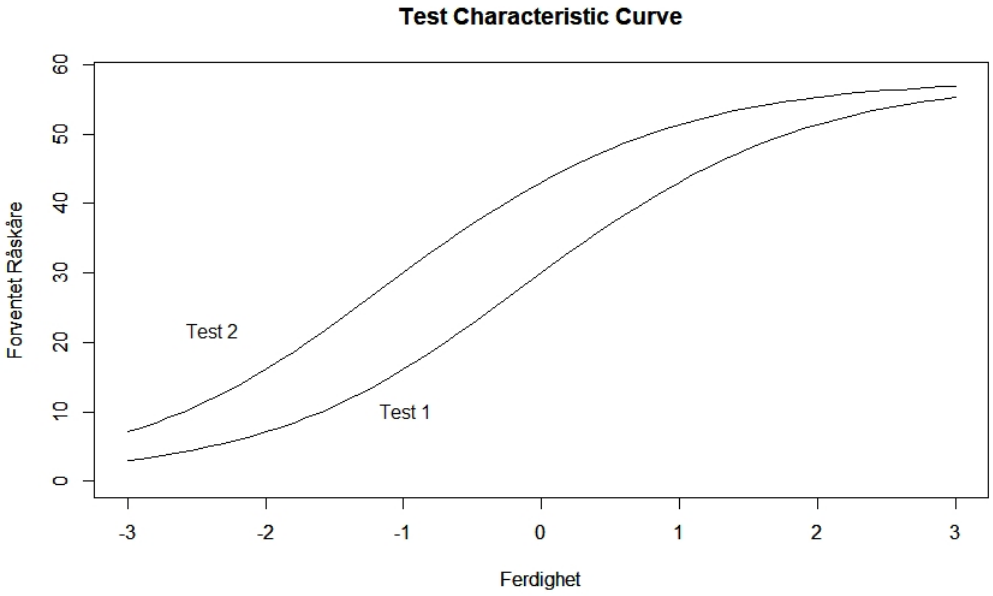
Figur 4: 3PL-ICC for tre oppgaver med ulike parametere. $D = 1.702$.
Oppgave 1: $a = 2.0,\ b = 0,\ c = 0.25$
Oppgave 2: $a = 1.2,\ b = 0,\ c = 0.10$
Oppgave 3: $a = 2.0,\ b = 1,\ c = 0.25$

3.2 Test Characteristic Curve

IRT-analyse er i stor grad basert på de individuelle oppgavene i en test, men når vi skal analysere tester i sin helhet må vi ha en metode for å legge sammen skåren fra de individuelle oppgavene for å gi et helhetlig bilde av det vi ønsker å måle prøvetakeren på. Når tester skåres dikotomt gis rett svar 1 poeng og galt svar gis 0 poeng, disse poengene summeres og vi får prøvetakerens *råskåre*. Råskåren vil alltid være et heltall mellom 0 og n, der n er antall oppgaver i testen. Om en prøvetaker tok en identisk prøve om igjen, antatt at den ikke husket noe fra forrige test, er det sannsynlig at prøvetakeren ville fått en annen råskåre. Tar prøvetakeren en slik test mange nok ganger vil råskårene samles rundt en gjennomsnittsverdi (Baker & Kim, 2017). Denne verdien kalles *true score* og er definert som *forventet råskåre* $= \xi$ (Lord, 1980).

$$\xi = \sum_{i=1}^{n} P_i(\theta) \tag{9}$$

Det følger av ligning 9 at hver prøvetaker med samme θ har samme ξ, og



Figur 5: 2PL TCC for to ulike tester, $a_1 = a_2$, $b_2 = b_1 - 1$, D = 1.702.

siden $P_i(\theta)$ er en økende funksjon av θ er forventet råskåre også en økende funksjon av θ. Forventet råskåre ξ og ferdighet θ er altså det samme, bare presentert på ulike skalaer, men med én vesentlig forskjell. Skalaen for ξ er avhengig av oppgavene i testen, mens skalaen for θ er uavhengig av oppgavene i testen. Dette gjør θ til mer nyttig parameter enn ξ når det kommer til å sammenligne ulike tester for samme ferdighetsnivå (Lord, 1980).

For hver $\theta \in \mathbb{R}$ kan det regnes ut en forventet råskåre ξ. Da vil altså enhver θ ha en korresponderende ξ. Plottes disse mot hverandre får vi det som kalles *Test Characteristic Cruve* (TCC). For 1PL- og 2PL-modellene nærmer venstre hale seg 0 når θ nærmer seg $-\infty$ og høyre hale nærmer seg n når θ nærmer seg $+\infty$. for 3PL-modellen nærmer venstre hale seg summen av c parameterne i stedet for 0. Dette reflekterer at prøvetakere med svært lav θ vil ha en forventer råskåre høyere enn 0 ved ren gjetting (Baker & Kim, 2017).

Figur 5 tar for seg TCC for to ulike tester der begge inneholder 58 oppgaver ($n = 58$). Test 1 består av b-verdier hentet fra nasjonale prøver i regning 2014

for 8. trinn (Bjørnsson, 2018). I Test 2 ser vi for oss en tenkt situasjon der alle oppgavene er lettere ved å sette: $\mathbf{b} = \mathbf{b} - 1$, der $\mathbf{b}$ er vektoren med b-verdiene fra nasjonale prøver 2014 . Begge testene har samme sett med a-verdier henter fra nasjonale prøver i regning 2014. En kan lese av grafen at forskjellen i forventet råskåre ξ mellom de to testene er større for $\theta \in \{-2, 1\}$ enn utenfor intervallet. Kurven til test 1 er brattest når $\theta \in \{-1, 1\}$, og kurven til test 2 er brattest når $\theta \in \{-2, 0\}$. Dette reflekterer for hvilke ferdighetsnivå testene best beskriver ξ. Utenfor disse intervallene blir ξ likere og det vil være naturlig å tro at usikkerheten også er større utenfor dette intervallet.

3.3 Estimering av ferdighet og oppgaveparametere

Innenfor IRT er hovedformålet med å administrere en test å finne ut hvor prøvetakeren ligger på ferdighetsskalaen, å estimere ferdigheten til prøvetakeren. De to mest brukte metodene er sannsylighetsmaksimering (MLE) og bayesianske metoder (Brennan, 2006). Når θ estimeres i neste delkapittel antas det at oppgaveparameterne er kjent, og oppgavene skåres dikotomt. Rett svar gis skåre 1 og galt svar skåre 0, svarene listes og lagres i en vektor $\mathbf{u}$. Denne vektoren kalles prøvetakerens responsvektor. Deretter estimeres oppgaveparameterne under forutsetningen at θ er kjent. I praksis skjer denne prosedyren samlet ved at N prøvetakere gir m svar på n oppgaver og vi får en svarmatrise. Dataen i denne svarmatrisen gjennomgår sannsynlighetsmaksimeringsprosesser helt til vi får stabile oppgaveparametere og θ-verdier slik at de kan tolkes innenfor rammene av IRT.

3.3.1 Sannsynlighetsmaksimering

Sannsynlighetsmaksimering innenfor IRT finner den $\hat{\theta}$ og de $\hat{\delta}$ ($\delta = (a, b, c)$, oppgaveparameterene i en prøve) som gir størst sannsynlighet for å oppnå prøvetakerens faktiske testdata (svarmønster), gitt valg av modell. Fordelen med sannsynlighetsmaksimering er at sannsynlighetsmaksimeringsestimatorene forblir forventningsrett etter hvert som antall oppgaver i testen og antall prøvetakere som tar testen øker, og de har minimale standardfeil (Brennan, 2006). I praksis finnes sannsynlighetsmaksimeringsestimatorene ved at logaritmen til rimelighetsfunksjonen deriveres og settes lik 0 (Lord, 1980). Den generelle rimelighetsfunksjonen for at prøvetaker k ($k = 1, .., N$) gir respons

j ($j = 1, ..., m$) på oppgave i ($i = 1, ..., n$) er

$$L = P(\mathbf{U} = u \mid \theta,\ \mathbf{a},\ \mathbf{b},\ \mathbf{c}) \tag{10}$$

$$= \prod_{k=1}^{N} \prod_{i=1}^{n} \prod_{j=1}^{m} P_{ij}(\theta_k)^{u_{ijk}} \tag{11}$$

Der $\mathbf{U}$ er prøvetakernes responsvektorer, $P_{ij}(\theta_k)$ er sannsynligheten for svar j på oppgave i gitt prøvetakerens ferdighet (θ_k), og der $u_{ijk} = 1$ dersom prøvetaker k svarte j på oppgave i, og 0 ellers.

3.3.2 Ferdighetsestimering

Når vi nå skal estimere ferdigheten θ_k til én enkelt prøvetaker der oppgavene skåres dikotomt ($m = 2$ og $u_{ik} = 0$ eller $u_{ik} = 1$) og vi antar at oppgaveparameterne er kjent blir rimelighetsfunksjonen:

$$L_k = \prod_{i=1}^{n} P_i(\theta_k)^{u_{ik}} (Q_i(\theta_k))^{1-u_{ik}} \tag{12}$$

hvor $Q_i(\theta_k) = 1 - P_i(\theta_k)$. Logaritmen til rimelighetsfunksjonen blir da:

$$ln\ L_k = \sum_{i=1}^{n} u_{ik}\ ln\ P_i\ (\theta_k) + (1 - u_{ik})\ ln\ Q_i(\theta_k) \tag{13}$$

Deriverer vi ligning 13 med hensyn på θ_k får vi da ligningene som løses for dikotome oppgaver når $\hat{\theta}$ skal estimeres:

$$\begin{aligned}
\frac{\partial ln\ L_k}{\partial \theta_k} &= \sum_{i=1}^{n} \frac{u_{ik}}{P_i(\theta_k)} \frac{\partial ln\ L_k}{\partial \theta_k} + \frac{1 - u_{ik}}{Q_i(\theta_k)} - \frac{\partial ln\ L_k}{\partial \theta_k} \\
&= \sum_{i=1}^{n} \frac{u_{ik} Q_i(\theta_k)}{P_i(\theta_k) Q_i(\theta_k)} \frac{\partial P_i(\theta_k)}{\partial \theta_k} - \frac{P_i(\theta_k) - u_{ik} P_i(\theta_k)}{P_i(\theta_k) Q_i(\theta_k)} \frac{\partial P_i(\theta_k)}{\partial \theta_k} \\
&= \sum_{i=1}^{n} \frac{u_{ik} - u_{ik} P_i(\theta_k) - P_i(\theta_k) + u_{ik} P_i(\theta_k)}{P_i(\theta_k) Q_i(\theta_k)} \frac{\partial P_i(\theta_k)}{\partial \theta_k} \\
&= \sum_{i=1}^{n} \frac{u_{ik} - P_i(\theta_k)}{P_i(\theta_k) Q_i(\theta_k)} \frac{\partial P_i(\theta_k)}{\partial \theta_k} = 0
\end{aligned} \tag{14}$$

Henholdvis for 1PL, 2PL og 3PL modellene blir ligningene:
For 1PL,

$$\sum_{i=1}^{n}[u_{ik} - P_i(\theta_k)] = 0 \tag{15}$$

For 2PL,

$$\sum_{i=1}^{n} a_i[u_{ik} - P_i(\theta_k)] = 0 \tag{16}$$

For 3PL,

$$\sum_{i=1}^{n} a_i \frac{P_i(\theta_k) - c_i}{(1 - c_i)P_i(\theta_k)}[u_{ik} - P_i(\theta_k)] = 0 \tag{17}$$

De ulike ligningene for estimering av θ_k løses iterativt med numeriske metoder. For eksempel vil ligning 16 løst med Newton-Rapson bli:

$$\hat{\theta}_{k,s+1} = \hat{\theta}_{k,s} - \frac{\sum_{i=1}^{n} a_i[u_i - P_i(\hat{\theta}_{k,s})]}{-\sum_{i=1}^{n} \alpha_i^2 P_i(\hat{\theta}_{k,s})Q_i(\hat{\theta}_{k,s})} \tag{18}$$

Der $\hat{\theta}_{k,s}$ er den estimerte ferdigheten til prøvetakeren i iterasjon s og $\hat{\theta}_{k,s+1}$ er neste iterasjon. Prosessen gjentas helt til $\hat{\theta}_{k,s+1} - \hat{\theta}_{k,s} \approx 0$, og den siste verdien av $\hat{\theta}_{k,s+1}$ brukes som prøvetakerens estimerte ferdighet.

3.3.3 Estimering av oppgaveparametere

I likhet med at det ved estimering av ferdighet i forrige delkapittel ble antatt at oppgaveparameterne var kjent, antas det ved estimering av oppgaveparametere i dette delkapittelet at ferdigheten er kjent. Videre brukes toparameter modellen i illustrasjonene i dette delkapittelet, og vi beholder også antakelsen om at prøvetaker k gir responsen $u = 1$ eller $u = 0$ på oppgave i. Når vi nå skal estimere oppgaveparametere bruker vi data fra N prøvetakere og ser på svar j på oppgave i. Hvis vi med disse antakelsene tar utgangspunkt i ligning 11 får vi ligningen:

$$\frac{\partial ln\ L}{\partial \delta_i} = \sum_{k=1}^{N} \frac{[u_{ik} - P_i(\theta_k)]}{P_i(\theta_k)Q_i(\theta_k)} \frac{\partial P_i(\theta_k)}{\partial \delta_i} = 0, \tag{19}$$

hvor δ er den oppgaveparameteren som estimeres. Ligning 14 blir lik 19 men de ulike antakelsene bestemmer hva som summeres og nå deriveres det med hensyn på oppgaveparameterne. Ligningene for hver parameter ved 3PL modellen blir da (Brennan, 2006):
For a_i

$$\frac{1}{1-c_i} \sum_{k=1}^{N} \frac{[\theta_k - b_i][P_i(\theta_k) - c_i]}{P_i(\theta_k)} [u_i k - P_i(\theta_k)] = 0 \tag{20}$$

For b_i

$$\frac{a_i}{1-c_i} \sum_{k=1}^{N} \frac{[P_i(\theta_k) - c_i]}{P_i(\theta_k)} [u_i k - P_i(\theta_k)] = 0 \tag{21}$$

For c_i

$$\frac{1}{1-c_i} \sum_{k=1}^{N} \frac{1}{P_i(\theta_k)} [u_i k - P_i(\theta_k)] = 0 \tag{22}$$

Når ligningene løses er initialverdier for oppgaveparameterne satt på forhånd (a priori) og verdiene brukes for å estimere $P(\theta_k)$ for hvert ferdighetsnivå med valgt modell. Ligningene løses da simultant og iterativt til vi får estimerte verdier for oppgaveparameterne som settes inn i uttrykket for modellen og en ICC-kurve kan plottes.

3.4 Kalibreringsprosessen

I de forrige delkapitlene har vi vist sannsynlighetsmaksimeringsprosesser for estimering av ferdighet og estimering av oppgaveparameterne separat. Når vi estimerte ferdighet antok vi at oppgaveparameterne var kjent, og når vi estimerte oppgaveparameterne antok vi at ferdigheten var kjent. I realiteten

kan vi ikke vite oppgaveparameterne til hver enkelt oppgave når vi konstruerer en test, vi kan heller ikke vite ferdigheten til prøvetakerne før de tar prøven. Dermed blir oppgaven å bestemme oppgaveparameterne og ferdigheten ved hjelp av svardata fra prøvetakerne, for så å få disse til å passe inn på en ferdighetsskala som passer til det spesifikke settet med oppgaver og prøvetakere, slik at resultatene kan tolkes ved hjelp av IRT. Denne prosessen kalles kalibreringsprosessen.

En mye brukt metode for test-kalibrering er Birnbaum paradigmet (Baker & Kim, 2017). Birnbaum paradigmet kombinerer de to sannsnylighetsmaksimeringsprosessene beskrevet tidligere i oppgaven til en samlet prosess (Joint maximum likelihood). I to steg estimeres oppgaveparameterne til n oppgaver og θ-verdiene til N prøvetakere. I det første steget antas de estimerte θ-verdiene å være sanne og oppgaveparameterne estimeres i samsvar med ligning 19. Oppgaveparameterne estimeres én oppgave om gangen under antakelsen at de er uavhengige av hverandre og vi får et sett med estimerte oppgaveparametere for hver oppgave i testen. I det andre steget antas oppgaveparameterne fra steg en å være sanne og ferdigheten til hver prøvetaker estimeres i samsvar med ligning 13. Disse to stegene gjentas etter hverandre iterativt til et gitt konvergenskriterium er satt. Det vi sitter igjen med er at oppgaveparameterne og ferdigheten har blitt estimert i en samlet prosess (Baker & Kim, 2017).

Alternativer til Birnbaum paradigmet som ofte brukes er bayesianske metoder (Brennan, 2006). En tidligere kjent fordeling av parameterne antas i estimeringsprosessen. Jo mer sikker brukeren er på informasjonen fra den tidligere fordelingen, jo mindre blir standardavviket i den tidligere fordelingen. Disse metodene baserer seg på Bayes' Teorem:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{23}$$

Der $P(X)$ er den tidligere fordelingen og informasjon fra den tidligere fordelingen kombineres med informasjon fra testen som kalibreres (Y) for å gi en posterior fordeling av sannsynligheten for en viss testskåre gitt prøvetakerens

ferdighet. Den generelle sammenhengen for 3PL-modellen blir da:

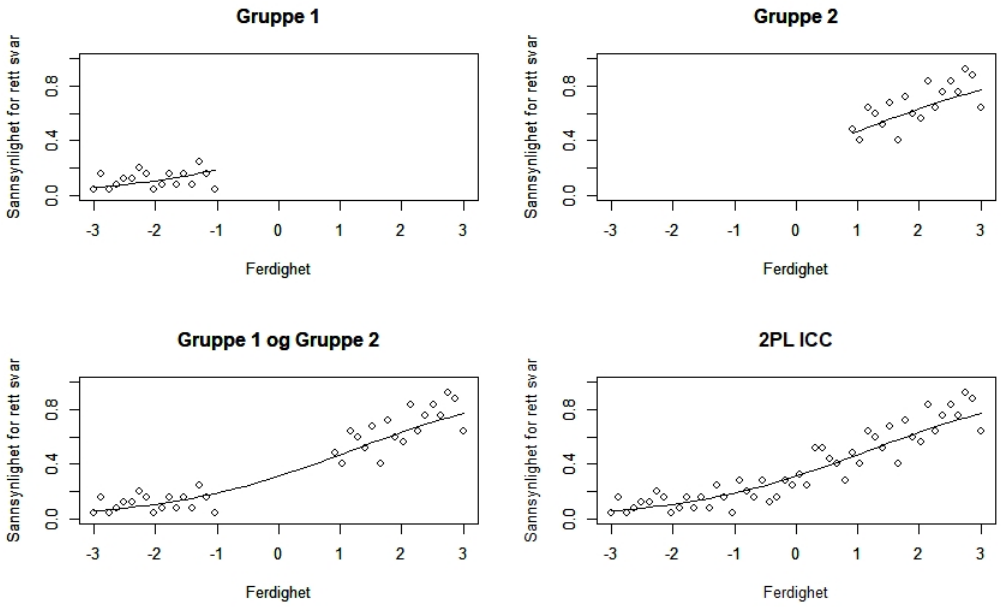
$$P(\theta|\mathbf{u_i}, \delta_\mathbf{i}) = \frac{P(\mathbf{u_i}|\theta, \delta_\mathbf{i})P(\theta)}{P(\mathbf{u_i})} \tag{24}$$

$$\text{der } P(\mathbf{u_i}) = \int P(\mathbf{u_i}|\theta, \delta_\mathbf{i})P(\theta)d\theta \tag{25}$$

Og der $P(\mathbf{u_i}|\theta, \delta_\mathbf{i})P(\theta)$ er sannsynligheten for poengskåre gitt valg av modell. $P(\theta)$ er den tidligere fordelingen av θ. Når gjennomsnittet fra posteriorfordelingen velges som ferdighetsestimat kalles det *expected a posteriori*(EAP) estimat. Det er EAP som i dag brukes i kalibreringen av nasjonale prøver. Fordelen med EAP framfor vanlige sannsynlighetsmaksimeringsmetoder er man unngår ekstreme målinger i ytterpunktene og da får lavere standardavvik i ytterpunktene. Ulempen med EAP er at bayesianske estimatorer kan ha systematisk skjevhet med at de trekker den estimerte ferdigheten mot gjennomsnittet til den tidligere fordelingen (Brennan, 2006). Denne ulempen blir mindre jo mer data vi har fra tidligere og jo sikrere vi er på dataen fra de tidligere fordelingene.

3.5 Gruppeinvarians

En stor fordel med IRT er at oppgaveparameterne er uavhengig av ferdighetsnivået til prøvetakerne som svarer på oppgavene. La oss si at vi har to grupper med prøvetakere fra samme populasjon der gruppe 1 har ferdighetsnivå fra -1 til 1 med gjennomsnitt -2 og gruppe 2 har ferdighetsnivå fra +1 til +3 med gjennomsnitt +2. Vi kan så ta for oss observert andel korrekte svar for gruppe 1 å estimere oppgaveparameterne a_1 og b_1. Estimerer vi så oppgaveparameterne a_2 og b_2 for gruppe 2 ser vi at $a_1 = a_2$ og $b_1 = b_2$. Dette illustreres i figur 6. Altså er oppgaveparameterne uavhengig av hvilken gruppe som svarte på oppgaven. Oppgaveparameterne er altså en egenskap som tilhører oppgaven, ikke prøvetakeren som svarte på oppgaven (Baker & Kim, 2017).



Figur 6: Illustrasjon av gruppeinvarians.

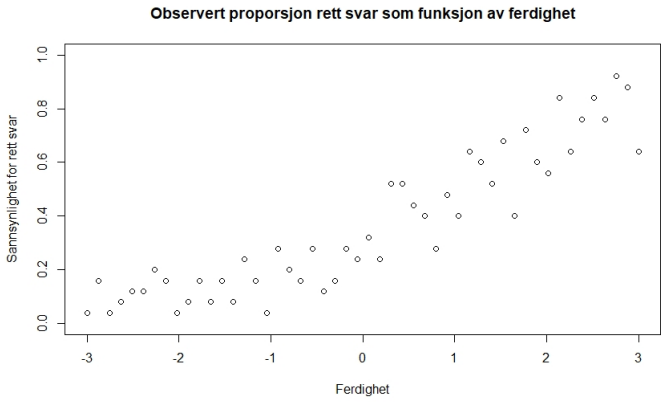
3.6 Modellsjekk

Et viktig element innenfor IRT-analyser er å måle hvor godt ICC-kurven passer til svardataene for den spesifikke oppgaven. For å sjekke om dataene passer overens med modellen ved ulike ferdighetsnivå deler vi inn prøvetakerne i G grupper langs ferdighetsskalaen slik at alle prøvetakerne innenfor hver gruppe har samme ferdighetsnivå θ_g. Innenfor hver gruppe g vil det være f_g prøvetakere og r_g prøvetakere som gir respons $u = 1$ (rett svar på oppgaven). Da får vi at på et gitt ferdighetsnivå θ_g er det observerte antallet rett svar $p(\theta_g) = r_g/f_g$ et estimat på sannsynligheten for rett svar for det ferdighetsnivået, $P(\theta)$. Hvis vi da får verdigen av r_g ved å administrere tester på ulike grupper prøvetakere kan vi regne ut p_g for hver g langs ferdighetsskalaen (Baker & Kim, 2017). Figur 7 viser observert andel rett svar som funksjon av ferdighet for en oppgave med gitte oppgaveparametere.

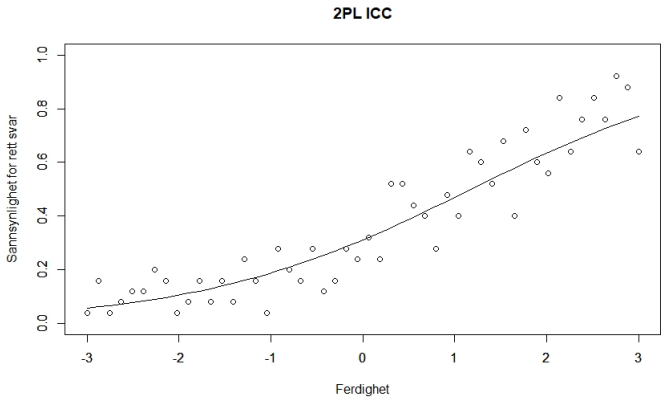
Hvor godt den observerte andelen av rette svar stemmer overens med sannsynligheten for rett svar gitt ved den estimerte ICC-kurven måles ofte med en kjikvadrat-observator (Baker & Kim, 2017). Kjikvadrat-observatoren er definert som:

$$\chi^2 = \sum_{g=1}^{G} f_g \frac{[p(\theta_g) - P(\theta_g)]^2}{P(\theta_g)Q(\theta_g)} \tag{26}$$

Hvis verdien av observatoren er større enn øvre kvartil for kjikvadratfordelingen ved gitt nivå (f.eks 5%) betyr det at ICC-kurven ikke passer til datasettet. Dette kan komme av man har anvendt feil modell i tilpassingen av kurven eller at andelen rett svar er for spredt til at en kan finne en god tilpassing, uavhengig av modell. I figur 8 er verdien av kjikvadrat-observatoren for de estimerte verdiene 46.18, mens den kritiske verdien ved 5% og 49 frihetsgrader er 66.34. Altså passer kurven til dataene. I en større test vil ofte enkelte oppgaver gi dårlige tilpassinger av kurven på grunn av for stor spredning i andel rette svar, men om flere av oppgavene gir dårlig tilpassede ICC-kurver er det grunn til å undersøke om man heller burde bruke en annen modell i analysen (Baker & Kim, 2017). Når en slik test brukes for å måle modelltilpasningen er det viktig å følge opp med videre analyser.



Figur 7: Observert andel rett svar som funksjon av ferdighet. $a = 0.67, b = 1.17, N = 50$.



Figur 8: Estimert 2PL-ICC kurve for data i figur 7

3.7 Informasjonsfunksjonen

Innenfor IRT ønsker vi å estimere ferdighetsparameteren θ til en prøvetaker. Estimerer vi ferdigheten med stor presisjon vet vi mer om ferdigheten enn hvis den er estimert med lavere presisjon. Verdien av presisjonen til ferdighetsestimatoren $\hat{\theta}$ er relatert til variasjonen i estimatoren rundt verdien av θ (Baker & Kim, 2017). Når vi skårer en enkeltoppgave dikotomt i en test gir vi skåre 1 for rett svar og 0 for galt svar. Vi lar så X være definert som råskåren, som er en funksjon av enkeltskårene i en test. Jo mer råskåren X i en test forteller oss noe om ferdigheten θ jo mer informasjon har vi. Informasjonen testskåren X gir oss om θ kan defineres som (Brennan, 2006):

$$I(X|\,\theta) = \frac{[\mu'(X|\,\theta)]^2}{\sigma^2(X|\,\theta)} \tag{27}$$

der $\mu' = dE(X|\,\theta)/d\theta$. Det kan vises at når antallet oppgaver i en prøve øker nærmer fordelingen av sannsynlighetsmaksimeringsestimatoren $\hat{\theta}$ seg en normal fordeling med gjennomsnitt θ og varians $\dfrac{1}{I(\hat{\theta}|\,\theta)}$, hvor $I(\hat{\theta}|\,\theta) = E\left[\dfrac{\partial lnL}{\partial\theta}\right]^2_{\theta=\hat{\theta}}$. Altså kan standardavviket fra testinformasjonen brukes til å lage konfidensintervall for θ, $\hat{\theta} \pm Z_{\alpha/2}SE(\hat{\theta})$ (Brennan, 2006).

$$SE(\hat{\theta}|\,\theta) = \frac{1}{\sqrt{I(\hat{\theta}|\,\theta)}} = \frac{\sigma(X|\,\theta)}{\mu'(X|\,\theta)} \tag{28}$$

Siden oppgaveinformasjonsfunksjonen gir informasjon basert kun på én enkelt oppgave vil mengden informasjon den gir være liten. En enkelt oppgave estimerer ferdighet best ved det ferdighetsnivået som korresponderer med oppgavens vanskegrad. Mengden informasjon nærmer seg 0 når avstanden fra ferdighetsnivået som korresponderer med oppgavens vanskegrad blir større. En prøve består av et sett med oppgaver og testinformasjonen for hele ferdighetsskalaen blir da summen av av oppgaveinformasjonen for hver oppgave (Baker & Kim, 2017). Testinformasjonsfunksjonen for tester som består av dikotome oppgaver med blir da summen av oppgaveinformasjonen:

$$I(\theta) = \sum_{i=1}^{n} I(X_i|\theta) \tag{29}$$

$$= \sum_{i=1}^{n} \frac{[P_i'(\theta)]^2}{P_i(\theta)[1 - P_i(\theta)]} \tag{30}$$

der $P_i'(\theta) = dP_i(\theta)/d\theta$ og da blir ulik for de ulike modellene. Oppgave-informasjonsfunksjonene for de mest vanlige modellene blir da (Brennan, 2006):

For 1PL:

$$I(X_i|\,\theta) = Q_i(\theta)P_i(\theta) \tag{31}$$

For 2PL:

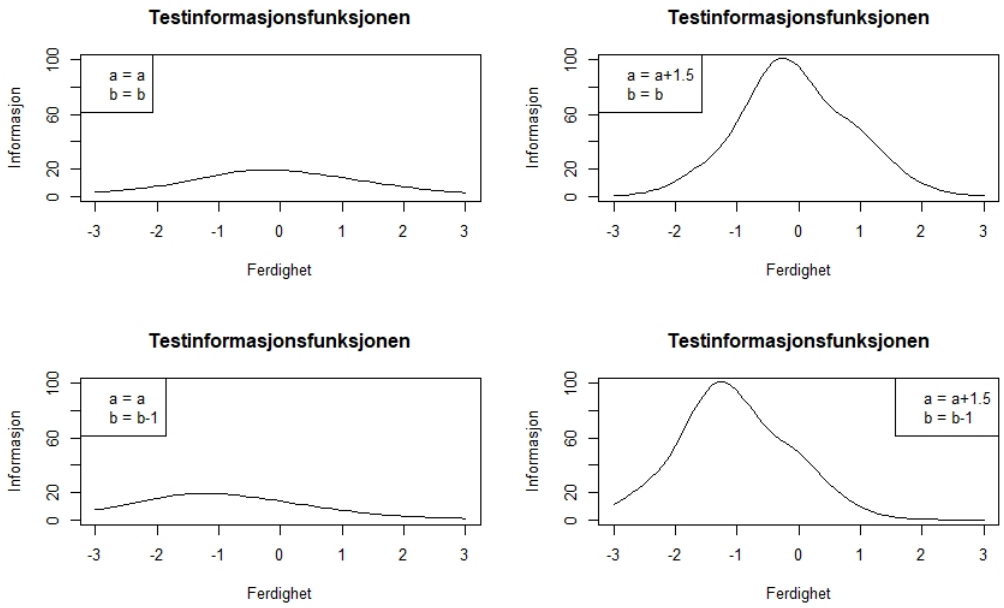
$$I(X_i|\,\theta) = D^2 a_i^2 Q_i(\theta)P_i(\theta) \tag{32}$$

For 3PL:

$$I(X_i|\,\theta) = D^2 a_i^2 \frac{Q_i(\theta)}{P_i(\theta)} \left[\frac{P_i(\theta) - c_i}{1 - c_i}\right]^2 \tag{33}$$

Plottet øverst til venstre i Figur 9 viser testinformasjonsfunksjonen for 58 oppgaver gitt ved nasjonale prøver i regning 2014. Leser vi av b-verdiene i tabell 1 er det naturlig å forvente mer informasjon ved $\theta = 0$ da de fleste av oppgavene har b i intervallet $(-1, 1)$, og det er kun 6 oppgaver som har $b > 1$ og 6 oppgaver med $b < -1$.

Ved å øke antall oppgaver med vanskegrad $b \approx -1$ øker vi informasjonen testen gir oss om prøvetakere med $\theta \approx -1$, det er dette som skjer i plottet nede til venstre. Men som vi kan se av plottene kommer dette på bekostning av informasjon om prøvetakere med θ lengre unna $b = \theta = -1$ på skala-en. Et universitet med strenge opptakskrav vil kanskje ønske å konsturere en test med toppunktet i testinformasonskurven lengre til høyre på ferdig-hetsskalaen, mens en skole som vil identifisere elever som trenger tilrettelagt undervisning i et fag ønsker å konstruere en test med toppunkt til venstre på θ skalaen, som vist i de to nederste plottene i figur 9. Nasjonale prøver i regning måler regneferdighetene til om lag 60 000 elever i hele landet, og



Figur 9: Testinformasjon for ulike a og b verdier, $n = 58$.

siden de fleste elevene ligger rundt gjennomsnittet er det konstruert en test med flest oppgaver med b parameter rundt 0 for å kunne gi best informasjon om der de fleste elevene ligger.

En annen metode for å oppnå mer informasjon er å konstruere oppgaver med brattere ICC-kurve ved det θ nivået en ønsker å måle. Ulempen med dette er at å øke diskrimineringen vil gi økt informasjon kun ved vanskegraden til oppgaven, på bekostning av informasjon på resten av skalaen. Plottene til høyre i figur 9 har blitt gitt en unaturlig høy diskriminering for å illustrere at ved å øke a-parameteren vinner man informasjon der $\theta = b$ på bekostning av informasjon i ytterkantene.

Tabell 1: a og b verdier, nasjonale prøver i regning 2014.

Opg	a	b	Opg	a	b	Opg	a	b	Opg	a	b
1	0.62	-0.96	16	0.65	-0.61	31	0.78	-0.55	46	0.89	-0.32
2	0.70	-1.76	17	0.44	0.19	32	0.84	0.13	47	0.61	0.01
3	0.82	-0.55	18	0.64	0.08	33	0.77	0.18	48	0.56	0.69
4	0.86	-0.59	19	0.63	0.03	34	0.56	0.40	49	0.46	0.46
5	0.46	-0.82	20	0.65	-1.47	35	0.46	-0.14	50	0.51	0.91
6	1.03	-0.97	21	0.65	-0.81	36	0.86	1.53	51	0.56	-1.79
7	1.07	-0.73	22	0.45	0.59	37	0.71	-0.37	52	0.90	-0.02
8	0.92	-0.25	23	0.50	-1.38	38	0.87	1.08	53	0.68	1.15
9	0.95	-0.30	24	1.33	-0.44	39	0.92	-0.12	54	0.39	0.64
10	0.44	-0.38	25	0.78	-0.15	40	0.61	0.76	55	1.00	0.99
11	0.38	-1.21	26	0.83	1.12	41	1.03	-0.64	56	0.66	-0.67
12	0.64	0.92	27	0.57	0.03	42	0.91	0.86	57	0.48	1.94
13	0.73	-1.23	28	0.74	-0.20	43	1.02	-0.12	58	0.90	1.52
14	0.79	-0.04	29	0.69	0.16	44	0.72	-0.37			
15	1.02	0.49	30	0.62	0.39	45	0.83	-0.43			

4 DIF Analyse

DIF-analyse (Differential Item Functioning) handler om å finne ut om enkelte oppgaver har ulike egenskaper basert på hvilken gruppe som svarer på oppgaven gitt at de som svarer har samme ferdighetsnivå. Et eksempel på en slik oppgave kan være følgende oppgave som ble gitt ved nasjonale prøver i regning til 8. trinn i 2018 (Tokle, Ravlo, Johansen, Myhre & Thoresen, 2019):

> Aleksander spiller Pokemon Go. Han har et egg som klekkes hvis han går 5km. Aleksander har gått 4,82km.
>
> Hvor mange meter har Aleksander igjen å gå før egget klekkes?

På denne oppgaven hadde gutter med lik ferdighet som jenter i gjennomsnitt 2.096 større odds enn jenter på å svare rett på oppgaven. Dette kan komme av at Pokemon Go i større grad fenger gutter enn jenter, eller at gutter generelt har gjort det bedre på oppgaver som handler om omgjøring av enheter (Tokle et al., 2019). Når DIF-analysen oppdager slik skjevhet i en oppgave bør det vurderes om oppgaven skal fjernes fra oppgavesettet. De finnes flere ulike metoder for DIF-analyser som brukes i ulike sammenhenger. De to neste delkapitlene ser nærmere på to metoder som er vanlig å bruke i IRT (Brennan, 2006).

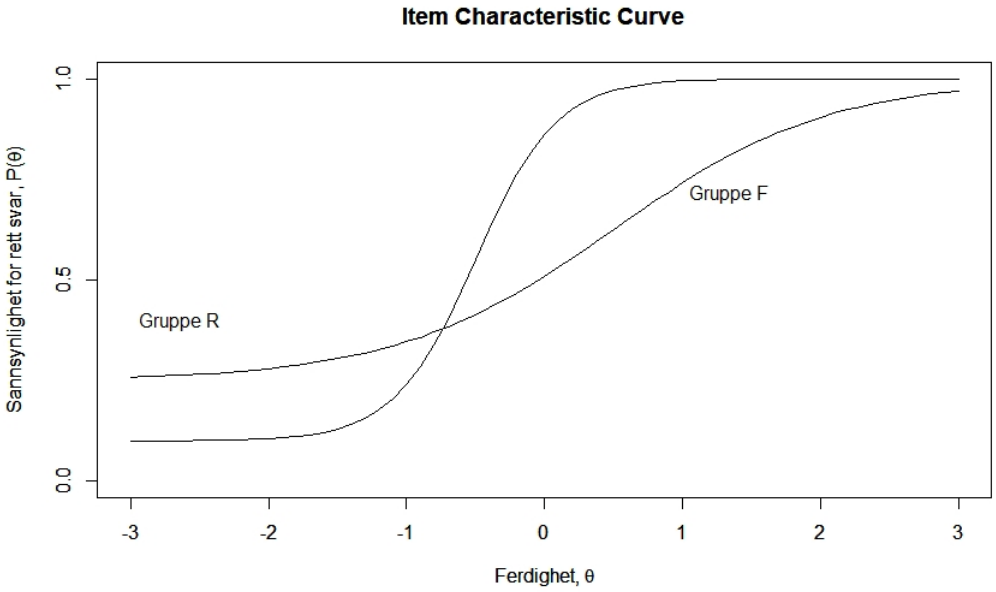
4.1 Forskjeller i ICC-kurvrer

En metode for DIF-analyse som brukes innenfor IRT er å se på selve ICC-kurven til de aktuelle oppgavene, mer spesifikt se på forskjellen i arealet under de respektive ICC-kurvene estimert separat for hver gruppe. Hvis vi ser på forskjeller i areal under kurven til en referansegruppe $P_R(\theta)$ og en fokusgruppe $P_F(\theta)$ får vi integralet:

$$A = \int_{-\infty}^{\infty} [P_R(\theta) - P_F(\theta)] d\theta \tag{34}$$

Har dette arealet positivt fortegn har vi skjevhet til fordel for referansegruppen. Om vi har kryssende ICC-kurver som i figur 10 kan araelene kansellere hverandre og vi får behov for å se på absoluttverdien i integralet:

$$AA = \int_{-\infty}^{\infty} |[P_R(\theta) - P_F(\theta)]| d\theta \tag{35}$$



Figur 10: Kryssende ICC-kurver, referansegruppe og fokusgruppe.

En ulempe med denne metoden er at arealmålene i ytterkantene av θ skalaen kan gi mye større effekt enn den egentlig har, særlig ved store forskjeller i c-parameteren (Brennan, 2006).

4.2 Mantel-Haenszel

En annen metode for å måle oppgaveskjevhet er Mantel-Haenszel testen, det er denne testen som i dag brukes som DIF-analyse i nasjonale prøver. Mantel-Haenszel testen brukes til å bestemme om to parametere er uavhengig av hverandre samtidig som de begge er avhengig av en tredje variabel. Et oppsett for en slik analyse kan være en 2 x 2 x S-1 krysstabell av oppgaver der S er høyeste oppnåelig råskåre på testen. Har vi en test med 5 oppgaver som skåres dikotomt vil vi få 4 tabeller, en for hver oppnåelig råskåre ($s = 1$ til $S-1$), ettersom tabellene for totalskåre 0 og 5 kollapser til en enkelt søyle. For hver oppnåelige råskåre blir prøvetakernes svar analysert og talt opp og vi får en krysstabell som i tabell 2, der v_{Rs} og v_{Fs} er antall prøvetakere i referanse- og fokus-gruppen som oppnåde råskåre s, u_{1s} og u_{0s} er antall rette og gale svar og T_s blir det totale antall prøvetakere tilhørende krysstabell s. En kan da regne ut hvor mye den ene gruppen favoriseres ved en Mantel-Haenszel koeffisient (Brennan, 2006):

$$\hat{\alpha}_{MH} = \frac{\sum_{s=1}^{S-1} \frac{A_s D_s}{T_s}}{\sum_{s=1}^{S-1} \frac{B_s C_s}{T_s}} \tag{36}$$

Om $\hat{\alpha}_{MH} = 1$ betyr det at fokusgruppen i gjennomsnitt har like stor sannsynlighet for å klare oppgaven som referansegruppen og vi har altså ikke oppdaget noen DIF i oppgaven. Om $\hat{\alpha}_{MH} = 2$ vil referansegruppen ha dobbelt så stor sannsynlighet som fokusgruppen til å svare rett på oppgaven, og er $\hat{\alpha}_{MH}$ lavere enn 1 vil referansegruppen gjøre det dårligere enn fokusgruppen. Ulempen med denne odds-skalaen er at den er asymmetrisk med øvre grense ∞ og nedre grense 0. Derfor konverteres odds-skalen ofte til en log-odds skala $\hat{\delta}_{MH} = ln(\hat{\alpha}_{MH})$ med tilhørende tilnærmet varians (Brennan, 2006):

$$SE(\hat{\delta}_{MH}) = \left[\frac{1}{2U^2} \sum_{s}^{S-1} T_s^{-2} (A_s D_s + \hat{\alpha}_{MH} B_s C_s) \cdot (A_s + D_s + \hat{\alpha}_{MH}(B_s + C_s)) \right]^{1/2} \tag{37}$$

Oppgaveskåre s

Gruppe	1	0	Total
Referanse	A_s	B_s	v_{Rs}
Fokus	C_s	D_s	v_{Fs}
Total	u_{1s}	u_{0s}	T_s

Tabell 2: Mantel-Haenszel krysstabell

$$\text{hvor } U = \sum_{s=1}^{S-1} \frac{A_s D_s}{T_s} \tag{38}$$

Log-odds skalaen og den tilhørende standardfeilen kan igjen konverteres videre til en mye brukt delta skala der:

$$\Delta_{MH} = -2.35 \cdot \hat{\delta}_{MH} \tag{39}$$

$$SE(\Delta_{MH}) = -2.35 \cdot SE(\hat{\delta}_{MH}) \tag{40}$$

Både i log-odds skalen og Δ skalaen vil en verdi på 0 tilsi ingen DIF i oppgaven.

5 Lenking og ekvivalering

For to tester er en lenking mellom skårene deres en overføring av skåren fra den ene testen til en skåre på skalaen til den andre testen. Det finnes flere forskjellige former for lenking av to tester X og Y, blant annet predikering, skalajustering og ekvivalering (Brennan, 2006). Dette kapittelet tar for seg ekvivalering av test X med test Y. I ekvivalering av to tester blir det satt opp en direkte lenking mellom skårene i de respektive testene. Formålet med å ekvivalere to tester er å kunne bruke skårene i de to testene om hverandre som om de hadde kommet fra samme test, for eksempel skal en skåre på 50 i regning fra nasjonale prøver i 2016 være det samme som en skåre på 50 i 2017. For at dette skal være mulig må det settes strenge krav til testene og metodene for ekvivalering. De må kunne måle samme egenskap med samme presisjon. Funksjonen som ekvivalerer test X med test Y må være symmetrisk, altså at å ekvivalere Y med X blir inversen av å ekvivalere X med Y. Det skal ikke ha noe å si for prøvetakeren om den tok test X eller test Y. Testene skal være populasjonsuavhengige, altså skal det ikke ha noe å si hvilken gruppe som brukes til å estimere ekvivaleringsfunksjonen mellom testene (Brennan, 2006).

5.1 Ikke-ekvivalente grupper med ankeroppgaver (NEAT)

Når tester ekvivaleres ved hjelp av NEAT metoden (Non-equivalent groups with anchor test) blir ofte oppgave og ferdighetsparametrene i test Y først estimert når test Y administreres, og parametrene i test X blir estimert når test X administreres. Ettersom prøvetakerne som tok test X ikke er ekvivalente med prøvetakerne som tok test Y havner de estimeterte parametrene på ulike skalaer. For å kunne lenke disse to skalaene lages et sett med oppgaver, ankeroppgavene, som er like de i to testene. De estimerte parametrene fra disse ankeroppgavene brukes til å kaliberere parametrene i test Y slik at test X og test Y havner på samme skala (Kolen & Brennan, 2014).

5.1.1 Parameterbaserte transformasjoner

Om en IRT-modell kan beskrives av et datasett kan en hvilken som helst lineær transformasjon av ferdighetsskalaen også beskrives av samme datasett, gitt at oppgaveparametrene også er transformert. Dermed kan en lineær sammenheng brukes til å kaliberere parameterestimatene fra de ulike testene til

samme skala. Disse kalibererte estimatene kan brukes til å ekvivalere ferdighetsskalaen fra test X til test Y. Hvis ferdighetsskalaen fra test X og ferdighetsskalaen Y kan sammenlignes lineært, kan sammenhengen mellom θ-verdiene for prøvetaker k beskrives ved:

$$\theta_{Xk} = A\,\theta_{Yk} + B \tag{41}$$

og sammenhengen mellom oppgaveparametrene for oppgave i beskrives ved:

$$a_{Xi} = \frac{a_{Yi}}{A} \tag{42}$$

$$b_{Xi} = A\,b_{Yi} + B \tag{43}$$

$$c_{Xi} = c_{Yi} \tag{44}$$

Videre kan A og B konstantene som brukes til å ekvivalere to oppgaver tatt av to prøvetakere beskrives ved:

$$A = \frac{\theta_{Xk_1} - \theta_{Xk_2}}{\theta_{Yk_1} - \theta_{Yk_2}} = \frac{b_{Xi_1} - b_{Xi_2}}{b_{Yi_1} - b_{Yi_2}} = \frac{a_{Yi}}{a_{Xi}} \tag{45}$$

$$B = b_{Xi} - Ab_{Yi} = \theta_{Xk} - A\,\theta_{Yk} \tag{46}$$

Generaliserer man ligningene 45 og 46 til å gjelde hele oppgavesett og større grupper med prøvetakere får vi sammenhengene:

$$A = \frac{\sigma(b_X)}{\sigma(b_Y)} = \frac{\mu(a_X)}{\mu(a_Y)} = \frac{\sigma(\theta_X)}{\sigma(\theta_Y)} \tag{47}$$

$$B = \mu(b_X) - A\,\mu(b_Y) = \mu(\theta_X) - A\,\mu(\theta_Y) \tag{48}$$

Der μ er gjennomsnitt og σ standardavvik. Dette betyr at gjennomsnittene og standardavvikene for oppgaveparametrene og for ferdigheten er definert for flere prøvetakere og oppgaver fra både test X og test Y. Hvilket betyr at om vi ønsker å ekvivalere ferdighetsskalaen fra test X med ferdighetsskalaen fra test Y er vi avhengig av *ankeroppgaver*, oppgaver som er både i test X og test Y.

En metode for ekvivalering kalt *mean/sigma metoden* bruker gjennomsnittene og standardavvikene til de estimerte b-parametrene fra ankeroppgavene i

ligning 47 og 48 til å estimere A- og B-konstantene. En annen metode, *mean/mean metoden* bruker gjennomsnittet til de estimerte a-parametrene fra ankeroppgavene til å estimere A-konstanten og gjennomsnittet til b-parametrene fra ankeroppgavene til å estimere B-konstanten. Når estimater for a- og b-parametrene er brukt er det gitt at likhetene i ligningene 47 og 48 holder, dermed kan de to metodene gi ulike resultater. En grunn til at mean/sigma metoden noen ganger er forretrukket er at estimatene for b-parametere ofte er mer stabile en parametere for a-parametere, mens noen ganger er mean/mean metoden foretrukket da gjennomsnitt ofte er mer stabile en standardavvik. Ettersom begge metodene kan gi ulike resultater avhengig av stabiliteten til elementene i metodene bør resultater fra begge metodene tas hensyn til når en gjennomfører ekvivalering (Kolen & Brennan, 2014).

5.1.2 Funksjonsbaserte transformasjoner

En potensiell ulempe med metodene beskrevet i delkapittel 5.1.1 er at ulike kombinasjoner av a- og b-parametere kan produsere like ICC-kurver. For eksempel kan to oppgaver som har to veldig ulike b-parametere ha like ICC-kurver avhengig av a- og c-parametrene. For å ta tak i dette problemet ble de utviklet metoder som tar hensyn til alle oppgaveparametrene samtidig. Disse metodene kalles *characteristic curve metoder* og er basert på funksjonen som beskriver sannsynligheten for rett svar i de ulike IRT-modellene, og særlig den egenskapen at sannsynligheten for rett svar gitt prøvetakerens ferdighet er den samme uavhengig av hvilken skala resultatene blir presentert i:

$$P_i(\theta_{Xk};\ a_{Xi}, b_{Xi}, c_{Xi}) = P_i(A\theta_{Yk} + B; \frac{a_{Yi}}{A}, Ab_{Yi} + B, c_{Yi}) \qquad (49)$$

Hvis det er brukt estimatorer for parametrene i ligning 49 er det ikke gitt at likheten holder for alle prøvetakere og oppgaver, for alle A- og B-konstanter. Det er forskjellen i ICC-kurvene som utnyttes i de funksjonsbaserte metodene.

Haebara-metoden (Kolen & Brennan, 2014) beskriver forskjellene i ICC-kurvene som en funksjon av summen av den kvadrerte forskjellen i ICC-kurvene for hver oppgave for prøvetakere ved en gitt θ. Forskjellen beskrives ved:

$$Hdiff(\theta_k) = \sum_{i=1}^{n} \left[P_i(\theta_{Xk};\ \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - P_i(A\theta_{Yk} + B; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi}) \right]^2 \tag{50}$$

Der det summeres over ankeroppgavene. Videre akkumuleres $Hdiff$ over flere prøvetakere. Estimeringen fortsetter ved å finne A og B-verdiene som minimerer følgende kriterium:

$$Hcrit = \sum_{k=1}^{N} Hdiff(\theta_k) \tag{51}$$

Stocking & Lord-metoden (Kolen & Brennan, 2014) bruker i motsetning til Haebara den kvadrerte forskjellene av summene over ankeroppgavene:

$$SLdiff(\theta_k) = \left[\sum_{i=1}^{n}\ P_i(\theta_{Xk};\ \hat{a}_{Xi}, \hat{b}_{Xi}, \hat{c}_{Xi}) - \sum_{i=1}^{n} P_i(A\theta_{Yk} + B; \frac{\hat{a}_{Yi}}{A}, A\hat{b}_{Yi} + B, \hat{c}_{Yi}) \right]^2 \tag{52}$$

Forskjellen på de to metodene er at $Hdiff$ er kvadratsummen av forskjellen i ICC-kurver, mens $SLdiff$ er kvadratsummen av forskjellen i TCC-kurver. I likhet med Haebara blir $SLdiff$ summert over gruppen med prøvetakere og estimeringen fortsetter ved å finne A- og B-verdiene som minimerer $SLcrit$:

$$SLcrit = \sum_{i=1}^{N} SLdiff(\theta_k) \tag{53}$$

5.1.3 Fixed item parameter calibration

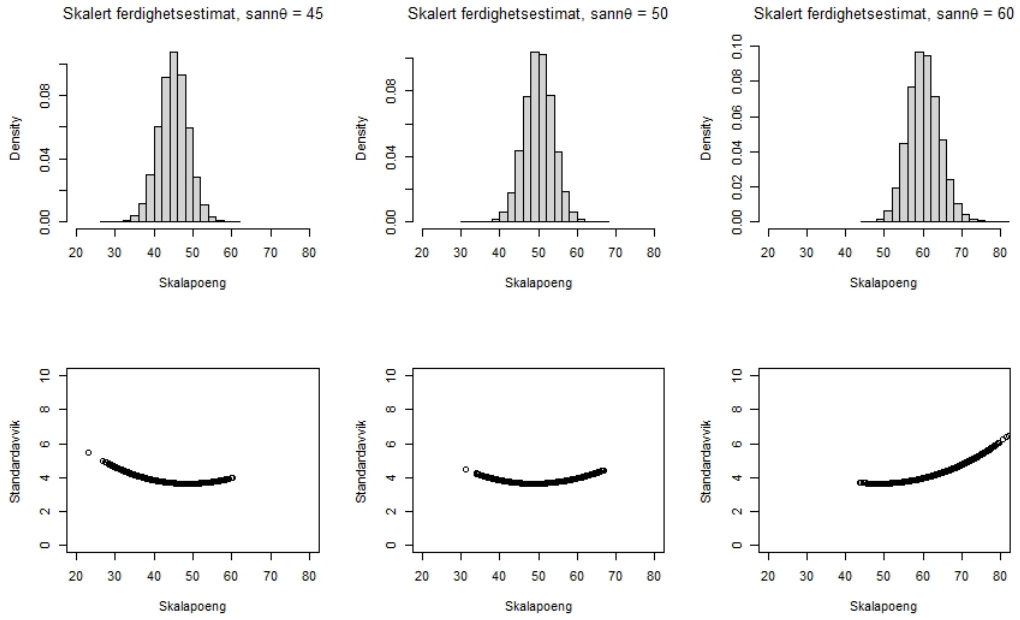
Fixed item paramterer calibration (FIPC) er kalibreringsmetoden som brukes i nasjonale prøver idag. Målet med FICP er å få estimert oppgaveparametrene til de nye oppgavene i test $X + 1$ ved hjelp av informasjon fra ankeroppgavene. Når FICP anvendes er oppgaveparametrene fra ankeroppgavene fiksert i kalibreringsprosessen og det er kun oppgaveparametrene til de nye oppgavene som estimeres. Denne typen lenking er spesielt nyttig når nye oppgaver skal flettes inn i en allerede eksisterende skala uten at parametrene fra de tidligere oppgavene endres (Chen, 2019), slik som i nasjonale prøver. For regning og engelsk består ankeret av 20 oppgaver. Ankeret vil også endre seg gradvis over tid ettersom oppgaver blir utdatert av ulike grunner. Dette skjer vet at det settes inn 2-4 nye ankeroppgaver hvert år og 2-4 oppgaver tas ut.

6 Simuleringer

Nasjonale prøver rapporterer resultater på blant annet elev-, klasse- og skolenivå. I dette kapittelet brukes metoder beskrevet i de forrige kapittelene på data hvor det simuleres ulike elever og klassesammensetninger, for å gi et bedre bilde av hva tallene som nasjonale prøver rapporterer kan fortelles oss. Bakgrunn for simuleringene er blant annet å kunne gi svar på hvor mye en elev eller en klasse må forbedre seg før det innenfor en rimelig sikkerhet er en faktisk forbedring, eller hvor mye som kan forklares ved naturlig variasjon. Kapittelet tar for seg spredningen i resultater for ulike enkeltelever, ulike klasser og ulike skoler. Det gjøres også simuleringen på differansen i estimert ferdighet fra et år til neste på bakrunn av at kullet i 9. klasse testes på de samme ferdighetene de ble testet på i 8. klasse.

For å simulere utfallet, u, av en dikotom oppgave med gitte oppgaveparametere, δ, for én elev med ferdighet θ, velges først en av modellene beskrevet i delkapittel 3.1. Deretter finner en ICC-kurven for oppgaven gitt oppgaveparameterne ($P(\theta, \delta)$). Så simuleres en $s \sim Uniform[0, 1]$ og en setter $u = 1$ dersom $s < P(\theta, \delta)$ og $u = 0$ hvis ikke.

Når vi simulerer den estimerte ferdigheten til en enkeltelev som tar en prøve trenger vi først å simulere svarvektoren $\mathbf{u}$ for hele prøven. Dette gjøres ved at vi først finner ICC-kurven for hver oppgave gitt oppgaveparameterne. I eksemplene i dette kapittelet brukes oppgaveparametere fra nasjonale prøver 2014. Da får vi en sannsynlighet for rett svar på hver oppgave, som funksjon av sann θ. Deretter simuleres svarvektoren. For hvert element i svarvektoren blir da $u_i = 1$ hvis $P(\theta, \delta_i) > s_i$ og $u_i = 0$ hvis ikke. Denne svarvektoren $\mathbf{u}$ samt a- og b-parameterne blir input i en funksjon som estimerer $\hat{\theta}$ basert på metodene beskrevet i kapittel 3.3.2. Disse stegene kan så settes i en for-løkke og gjentas N_{sim} ganger. Det vi får ut blir da gjentak av elevens estimerte $\hat{\theta}$ som er det de nasjonale prøvene rapporterer, basert på elevens sanne θ som er det faktiske ferdighetsnivået til eleven. I simuleringskapittelet her vil $\hat{\theta}$ rapporteres med samme lineære transformasjon som ved utdanningsdirektoratets rapportering av resultatene fra nasjonale prøver der skalapoeng, $\theta_S = \theta * 10 + 50$, som da tilsier et nasjonalt gjennomsnitt på 50 med et standardavvik på 10 dersom $\theta \sim N(0, 1)$.



Figur 11: Spredning i estimert θ_S basert på ulike sann θ_S med tilhørende standardavvik der standardavviket er det teoretiske standardavviket ved hver iterasjon, $N_{sim} = 10000$

6.1 Enkeltelever

Dette delkapittelet viser simulering av enkeltelever med ulik ferdighet for å se på spredningen i resultat når én enkeltelev, hypotetisk, gjennomfører nasjonal prøve i regning gjentatte ganger. Videre tar simuleringene i dette delkapittelet utgangspunkt i 2PL-modellen som gitt i ligning (3). Når simuleringene gjennomføres er a- og b-parameterne fiksert, og det er parameterne fra nasjonale prøver i regning 2014 som brukes i simuleringene. Det at parameterne er fiksert betyr at det er metodene beskrevet i delkapittel 3.3.2 som brukes i disse simuleringene og ikke de i delkapittel 3.4. Hadde vi brukt metodene i delkapittel 3.4 ville a- og b-parameterne blitt kalibrert etter svarmønsteret til eleven og vi hadde fått et estimert skalapoeng $\hat{\theta}_S = 50$, uavhengig av elevens sanne θ_S men med varierende a- og b- parametere.

Figur 11 viser spredningen og tilhørende standardavvik for $\hat{\theta}_S$ ved $\theta_S = 45, \theta_S = 50$ og $\theta_S = 60$. Standardavviket er her standardavviket for hver

Tabell 3: $\hat{\theta}_S$-estimat fra ulike sanne θ_S-verdier. Føreste kolonne: ulike θ_S, andre og tredje kolonne: gjennomsnittlig $\hat{\theta}_S$ og standardavviket til gjennomsnittet. De neste kolonnene viser kvantilfordeling av $\hat{\theta}_S$. De to siste kolonnene viser hvor ofte $\hat{\theta}_S$ havner innenfor ± 2.5 og ± 5 av θ_S. $N_{sim} = 10000$

θ_S	Gj.sn.	SD	2.5%	25%	50%	75%	97.5%	$\theta_S \pm 2.5$	$\theta_S \pm 5$
30	29.4	4.85	19.0	26.5	29.9	32.9	37.9	40.3%	71.8%
35	34.6	4.28	25.4	31.9	34.9	37.6	42.4	43.9%	75.7%
40	39.8	3.96	31.9	37.4	40.0	42.5	47.1	48.0%	79.8%
45	45.0	3.80	37.5	42.6	45.0	47.7	52.2	49.8%	81.9%
50	50.0	3.73	42.7	47.6	50.0	52.5	57.3	50.6%	82.4%
55	55.1	3.78	47.8	52.5	55.0	57.6	62.8	49.4%	81.4%
60	60.2	4.07	52.7	57.4	60.1	62.9	68.7	47.3%	78.8%
70	70.6	5.09	61.9	67.1	70.2	73.7	81.7	39.9%	70.2%

enkelt $\hat{\theta}_S$, som regnes ut via testinformasjonen (ligning 28). Tabell 3 viser $\hat{\theta}_S$ fordelt på kvantiler ved $N_{sim} = 10000$. Vi ser at kvartilbredden ved $\theta_S = 50$ er ca 5, og at den øker til $32.85 - 26.52 = 6.33$ ved $\theta_S = 30$ (ned to standardavvik) og den øker til $73.67 - 67.10 = 6.57$ ved $\theta_S = 70$ (opp to standardavvik). Dette gjenspeiler det vi finner ved å lese av informasjonsfunksjonen til prøven (plottet øverst til venstre i figur 9). Den estimerer mest presist på midten av skalaen og blir mindre presis mot ytterkantene. Den er noe asymmetrisk med det at den er litt mer presis på lavere vanskegrad enn høyere (se figur 9.

Vi kan med dette si noe om hvor ofte en elev blir plassert innenfor det mestringsnivået eleven egentlig hører til. En elev med sann $\theta_S = 50$, altså midten av mestringsnivå 3, vil i 82% av tilfellene bli plassert innenfor sitt mestringsområde ($\theta_S = (44.50, 54.49)$). Mens en elev med sann $\theta_S = 40.49$, altså midten av mestringsnivå 2, vil kun i 69.80% av tilfellene bli plassert innenfor mestringsområde 2 ($\theta_S = (36.50, 44.49)$), 16.12% av tilfellene i mestringsnivå 1 og 14.08% i av tilfellene mestringsnivå 3 eller høyere. I mostatt ende vil en elev på midten av mestringsnivå 4 i 69.15% av tilfellene bli plassert innenfor sitt eget mestringsnivå. Det her her verdt å merke seg at mestringsnivå 2 og 4 er smalere enn mestringsnivå 3 (bredde 7.99 mot bredde 9.99). En elev som ligger midt i mellom mestringsnivå 2 og 3, $\theta_S = 44.50$, vil i 97.57% av tilfellene bli plassert i enten mestringsnivå 2 eller mestringsnivå 3, i 0.33%

av tilfellene i mestringsnivå 4 og i 2.10% av tilfellene i mestringsnivå 1.

De to siste kolonnene i tabell 3 viser hvor ofte en elev får skalapoeng innenfor ± 2.5 og ± 5 av sin sanne θ_S. Vi ser at en elev med sann $\theta_S = 50$ i $\approx 50\%$ av tilfellene får skalapoeng innenfor ± 2.5 av sin sanne θ_S. Denne verdien synker ned til $\approx 40\%$ mot ytterpunktene ($\pm 2SD$) i begge ender av ferdighetsskalaen.

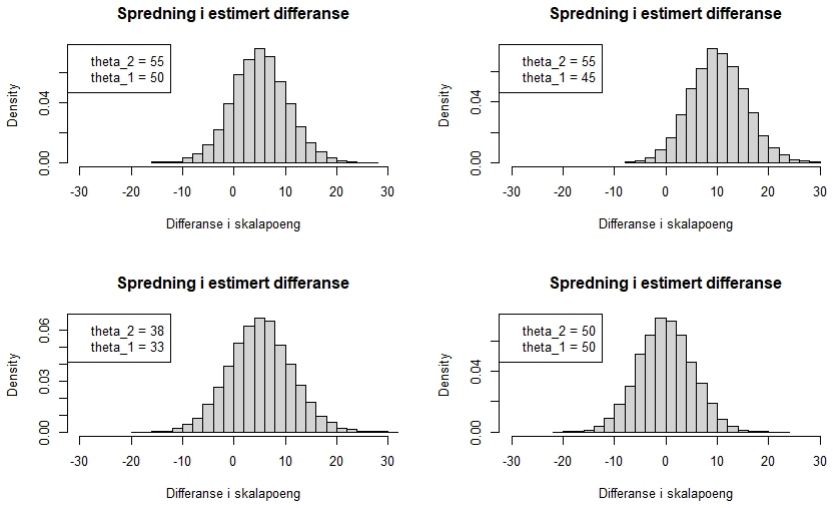
6.1.1 Endring i ferdighet

Vi har nå grunnlaget for å undersøke spredningen i den estimerte endringen i ferdigheten til en elev fra et år til et annet. I parksis gjennomføres nasjonale prøver for 8. trinn og for 9. trinn samme prøve, slik at en skal kunne måle endringen fra et år til neste. Dette kan simuleres ved å bygge videre på algoritmen i forrige delkapittel. La θ_1 betegne ferdighet i år 1 og θ_2 betegne ferdighet i år 2 å la $D = \theta_2 - \theta_1$ være endring i ferdighet fra år 1 til år 2. Hvis vi nå først estimerer θ_1, så estimerer θ_2 og setter $\hat{\theta}_2 - \hat{\theta}_1 = \hat{D}$ og så gjentar stegene N_{sim} ganger, vil vi få ut spredningen til endringen i ferdighet nasjonale prøver rapporterer.

Tabell 4: Estimert differanse fra år 1 til år 2. Andre kolonne viser hvor ofte det blir rapportert forbedring, de neste kolonnene viser hvor ofte den estimerte endringen er innenfor gitte intervaller. $N_{sim} = 10000$.

$\theta_{S,2} - \theta_{S,1}$	$\hat{D} < 0$	$\hat{D} > 0$	$\hat{D} \in [D \pm 2.5]$	$\hat{D} \in [D \pm 5]$	$\hat{D} \in [D \pm 10]$
$38 - 33$	19.5%	80.4%	32.6%	59.7%	90.1%
$45 - 35$	3.45%	96.6%	33.7%	61.9%	91.3%
$45 - 44$	42.4%	57.6%	36.4%	65.6%	94.0%
$50 - 50$	50.0%	50.0%	36.8%	66.1%	94.3%
$55 - 45$	2.73%	97.3%	35.0%	65.2%	93.7%
$55 - 50$	17.2%	82.8%	35.6%	65.0%	93.8%
$65 - 50$	0.26%	99.7%	33.9%	61.9%	91.6%
$65 - 65$	49.8%	50.2%	31.3%	57.9%	89.0%
$70 - 65$	21.8%	79.2%	29.9%	55.7%	86.8%

Figur 12 viser spredingen i den estimerte endring i ferdighet fra et år til neste for ulike sanne ferdigheter θ_1 og θ_2. Tabell 4 viser i første kolonne hvor



Figur 12: Spredning i estimert endring i skalapoeng. Fra øverst til venstre: $\theta_2 = 55, \theta_1 = 50$, $\theta_2 = 55, \theta_1 = 50$, $\theta_2 = 38, \theta_1 = 33$ og $\theta_2 = 50, \theta_1 = 50$, $N_{sim} = 10000$.

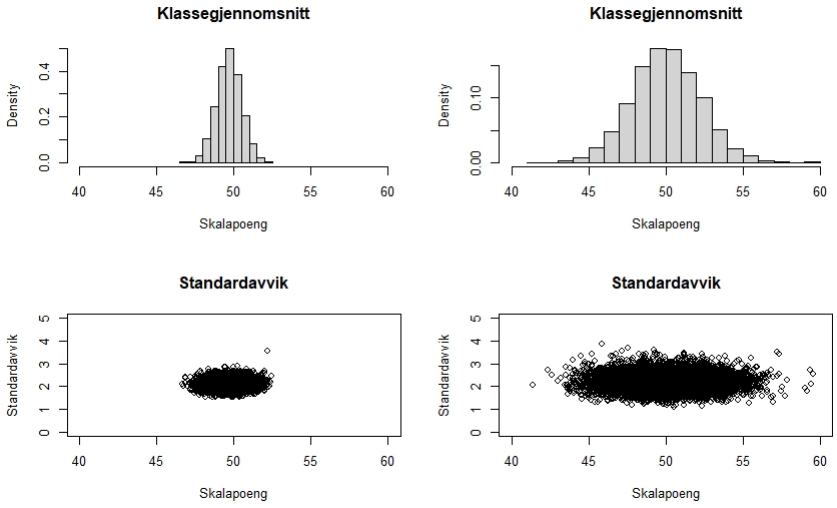
ofte det rapporteres ingen økning i θ og i neste kolonne hvor ofte det blir rapportert en økning i ferdigheten. De neste kolonnene viser hvor ofte den estimerte differansen er innenfor intervaller med størrelser på henholdsvis 5, 10 og 20 skalapoeng fra den faktiske differansen. Altså hvis en elev i realiteten har hatt en forbedring på D skalapoeng, hvor ofte elevens estimerte differanse havner innenfor intervallene $(D \pm 2.5)$, $(D \pm 5)$ og $(D \pm 10)$. I likhet med tidligere eksempler og simuleringer ser vi også ved estimert endring i ferdighet at den er noe mer nøyaktig ved gjennomsnittet ($\theta_S = 50$) enn ved ytterpunktene. En elev som har forbedret seg med 5 skalapoeng vil ved midten av skalaen få rapportert en forbedring i $\approx 82\%$ av tilefellene, denne andelen vil synke med litt over 2% når vi nærmer oss $\pm 2SD$, med større usikkerhet i den øvre delen av skalaen. En forbedring på fem skalapoeng vil i flere tilfeller være nok til å bli plassert i et høyere mestringsnivå, avhengig av verdien til θ_1. Ser vi på eksempelet i øverste rad i tabell 4 ser vi at en elev som har gått fra 33 skalapoeng i år 1 til 38 skalapoeng neste året vil i $\approx 80\%$ av tilfellene få rapportert en forbedring. Det ser ut til at usikkerheten i $\hat{D}$ avhenger mer av hvor på skalaen θ_2 og θ_1 ligger enn avstanden mellom de.

Ser vi på elever som har holdt seg på samme nivå begge årene ($\theta_2 = \theta_1$), kan vi se at elever på midten av skalaen ($\theta_S = 50$) i $\approx 35\%$ av tilfellene kan få rapporert enn differanse på over 5 skalapoeng påfølgene år, til tross for at ferdigheten i realiteten er uendret. Til sammenligning vil en elev med sann $\theta_S = 50$ som tar prøven et år i $\approx 50\%$ av tilfellene få skalapoeng innenfor ± 2.5. Dette kommer av at når vi sammenligner år for år vil det være usikkerhet knyttet til estimatene for begge år, mot bare det ene året i det andre tilfelle. I endene av skalaen øker usikkerheten, en elev med $\theta_1 = \theta_2 = 65$ vil i $\approx 42\%$ av tilfellene se en differanse på over 5, og i $\approx 11\%$ av tilfellene se en differanse på over 10 fra år 1 til år 2 til tross for ingen reell endring i ferdighet.

En lærer kan kanskje lure på hvor stor forbedring i resultatene fra nasjonale prøver man må se i en elev fra 8. klasse til 9. klasse før en kan være 90% sikker på at det er en reell forbedring. Siden usikkerheten er større i enden av skalaen vil dette avhenge av elevens sanne θ_1. Tar vi utgangspunkt i en sann $\theta_1 = 30$ vil en reell forbedring på ≈ 7.8 skalapoeng gi $\hat{D} > 0$ i $\approx 90\%$ av tilfellene. Tar vi derimot utgangspunkt i en sann $\theta_1 = 50$ vil en differanse $D \approx 6.8$ være tilstrekkelig for at $\hat{D} > 0$ i $\approx 90\%$ av tilfellene. Det betyr at svaret på dette på spørsmålet er at læreren må se en forbedring på mellom 6.8 og opp i mot 8 hvis elevene ligger helt i endene av ferdighetsskalaen, i følge simuleringene i dette kapittelet.

6.2 Skoleklasser

Når vi skal simulere resultatene for en hel klasse må vi på samme måte som for enkeltelever simulere en svarvektor for hver elev i klassen. Det betyr at vi må tildele en sann θ for hver elev i klassen. Dette kan gjøres ved for eksempel å trekke fra en normalfordeling, eller tildele hver elev en gitt θ. Deretter simuleres svarvektoren for hver elev, og algoritmen blir lignende det å simulere enkeltelever der klassestørrelse $= N_{klasse}$, men nå får hver elev i klassen sin egen θ som trekkes fra en gitt fordeling. Det lages en θ-vektor på størrelse N_{klasse}. Deretter kan vi finne gjennomsnittskåre til klassen og gjenta N_{sim} ganger for å finne spredningen i gjennomsnittsskåren til en klasse av størrelse N_{klasse}. Når vi gjentar algoritmen N_{sim} ganger kan vi enten fiksere klassens θ-vektor, eller trekke en ny for hver iterasjon. Fiksert θ-vektor representerer situasjonen der klassen med akkurat samme ferdighetsnivå tar den samme prøven N_{sim} ganger, og vi får en usikkerhet knyttet til selve prøven. Har vi ikke fiksert θ-vektoren vil det for eksempel kunne representere den naturlige



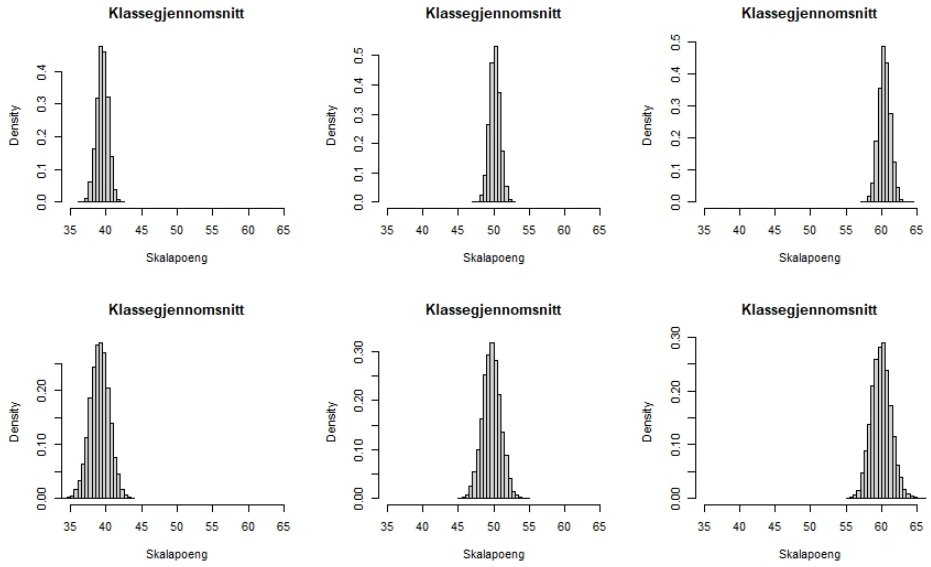
Figur 13: Klassesnitt med fiksert og ikke-fiksert θ, tilhørende teoretisk standardavvik regnet ut for hvert estimerte klassesnitt. $N_{klasse} = 25, N_{sim} = 10000$

spredningen blant ulike klasser på et kull, eller et årskull mot neste.

Tabell 5: Klassesnitt med fiksert og ikke-fiksert θ, SD = standardavviket til det estimerte klassesnittet fra variasjonen i estimatene. $N_{klasse} = 25, N_{sim} = 10000$

$\theta \sim N(0,1)$	Gj.snitt	SD	2.5%	25%	50%	75%	97.5%
Fiksert	49.7	0.80	48.1	49.2	49.7	50.2	51.3
Ikke-fiksert	50.0	2.20	45.7	48.5	50.0	51.5	54.3

Figur 13 viser forskjellen i spredningen når vi har fiksert klassens θ-vektor, mot når vi trekker en ny hver gang, standardavviket er standardavviket til hvert gjennomsnitt basert på spredningen i klassen ($SD = SD(\hat{\theta})/\sqrt{(N_{klasse})}$. Tabell 5 viser kvantilfordelingen til gjennomsnittsskåren til en klasse på 25 elever med henholdsvis fiksert og ikke-fiksert θ. At θ er trukket fra en standard dnormalfordeling tilsvarer et skalert snitt på 50 skalapoeng med standardavvik 10. For akkurat en klasse på 25 elever som gjennomfører nasjonale



Figur 14: Klassestørrelse 30 mot klassestørrelse 10, ulike, fikserte ferdighetsfordelinger, $N_{sim} = 10000$.

prøver er det estimerte gjennomsnittet til klassen i 50% av tilfellene innenfor $\approx \pm 0.5$ skalapoeng og i 95% av tilfellene innenfor $\approx \pm 1.6$ skalapoeng. Ser vi derimot på usikkerheten knyttet til variasjon i ferdighetsfordelingen, ser vi at vi i 95% av tilfellene er innenfor ± 4.3 av gjennomsnittlig skalapoeng for klassen.

Det er ikke gitt at alle klasser er like store og har elever med standardnormalfordelt ferdighet i den underliggende ferdigheten som testes. Det vil derfor være relevant å simulere klasser av ulik størrelse og ulike ferdighetsfordelinger.

Øverste rad i figur 14 viser spredningen i klassens estimerte gjennomsnittskåre for klasser på 30 elever med θ_S fordelt henholdsvis $\theta_S \sim N(40, 10)$, $\theta_S \sim N(50, 10)$ og $\theta_S \sim N(60, 10)$. Neste rad viser klasse på 10 elever med tilsvarende θ fordelinger. Tilhørende tabell, tabell 6 viser kvantilfordelingen for estimert gjennomsnittskåre for de ulike klassestørrelsene og fordelingene. Vi kan se at for en klasse på 30 elever med $\bar{\theta}_S = 50.1$ er det estimerte klassegjennomsnittet i 50% av tilfellene innenfor $\approx \bar{\theta}_S \pm 0.5$ og i 95% innenfor

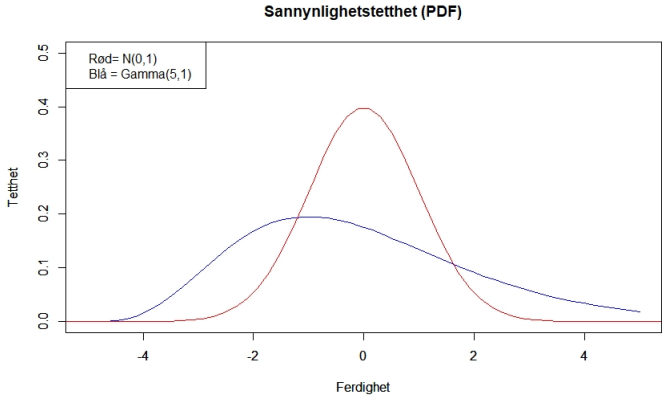
Tabell 6: Klassesnitt, store mot små klasser ved ulike normalfordelinger, fiksert θ_S, SD = standardavviket til det estimerte klassesnittet. $N_{sim} = 10000$, N_{klasse} = klassestørrelse.

$N_{klasse} = 30$	Gj.snitt	SD	2.5%	25%	50%	75%	97.5%
$\theta_S \sim N(40, 10)$	39.5	0.80	37.8	38.9	39.5	40.0	41.0
$\theta_S \sim N(50, 10)$	50.1	0.73	48.7	49.6	50.1	50.6	51.6
$\theta_S \sim N(60, 10)$	60.4	0.82	58.8	59.9	60.4	60.9	62.1
$N_{klasse} = 10$	Gj.snitt	SD	2.5%	25%	50%	75%	97.5%
$\theta_S \sim N(40, 10)$	39.1	1.34	36.4	38.2	39.1	40.0	41.7
$\theta_S \sim N(50, 10)$	49.7	1.26	47.2	48.8	49.7	50.5	52.1
$\theta_S \sim N(60, 10)$	59.9	1.39	57.3	59.0	59.9	60.8	62.8

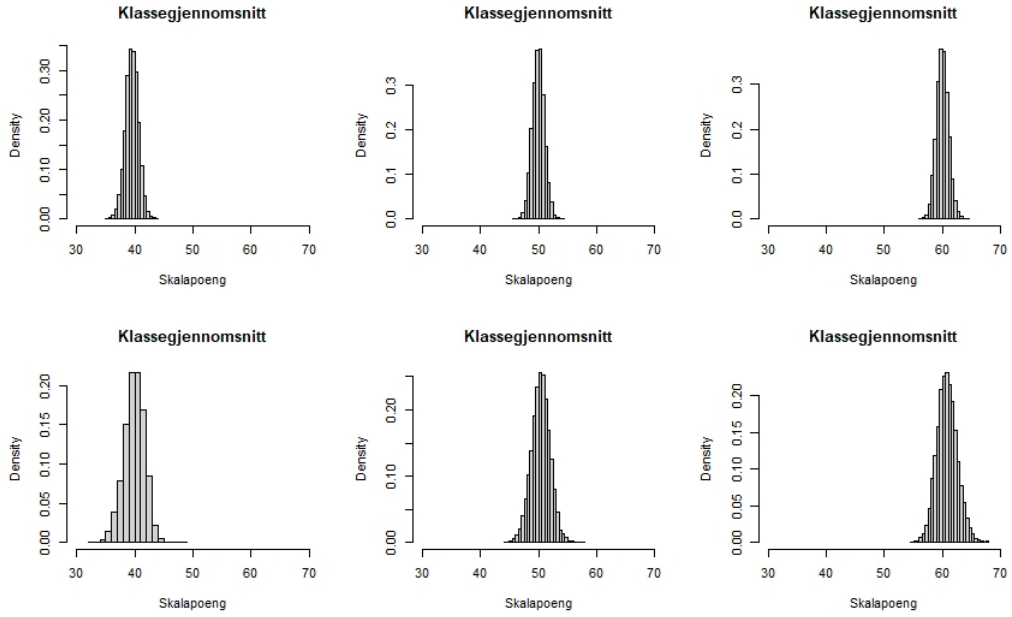
$\approx \bar{\theta}_S \pm 1.5$. For en klasse på 10 elever med $\bar{\theta}_S = 49.7$ er estimert klassesnitt i 50% av tilfellene innenfor $\approx \bar{\theta}_S \pm 0.9$ og i 95% av tilfellene $\approx \bar{\theta}_S \pm 2.5$. Spredningen i enden av skalaene øker ved begge klassestørrelsene opp til 95% innenfor $\bar{\theta}_S \approx \pm 2.9$ for en klasse på 10 elever med høy ferdighet. Dette betyr som forventet at selv ved små klasser er klassens estimerte gjennomsnittskåre mere presis enn for enkeltelever, gitt en normalfordelt ferdighet i klassen.

Det er blitt vist at underliggende ferdigheter av den typen som nasjonale prøver måler ofte ikke er normalfordelt. De kan være asymetriske og med flere observasjoner i endene av skalaen enn ved en standardnormalfordeling (Chen, 2019). På bakgrunn av dette velger Chen (2019) $\theta \sim \Gamma(5, 1) - \mu_\Gamma$, det vil si en gammafordeling med formparameter = 5, og skalaparameter = 1 og μ_Γ = gjennomsnittet i gammafordelingen, altså $form * skala$. Snittet trekkes fra for å få skiftet fordelingen slik at den får samme gjennomsnitt som normalfordelingen den sammenlignes med. I de neste simuleringene trekkes elevenes sanne θ-vektor fra tilsvarende fordeling: $\Gamma(5, 1)$, og fordelingen skiftes for å oppnå det klassesnittet vi ønsker som utgangspunkt for simuleringene.

Figur 15 viser tetthetsfunksjonen til normalfordelingen og gammafordelingen det trekkes fra. Gammafordelingen er forskjøvet for å ha samme gjennomsnitt som normalfordelingen.



Figur 15: Tetthetsfunksjoner for normalfordeling og gammafordeling.

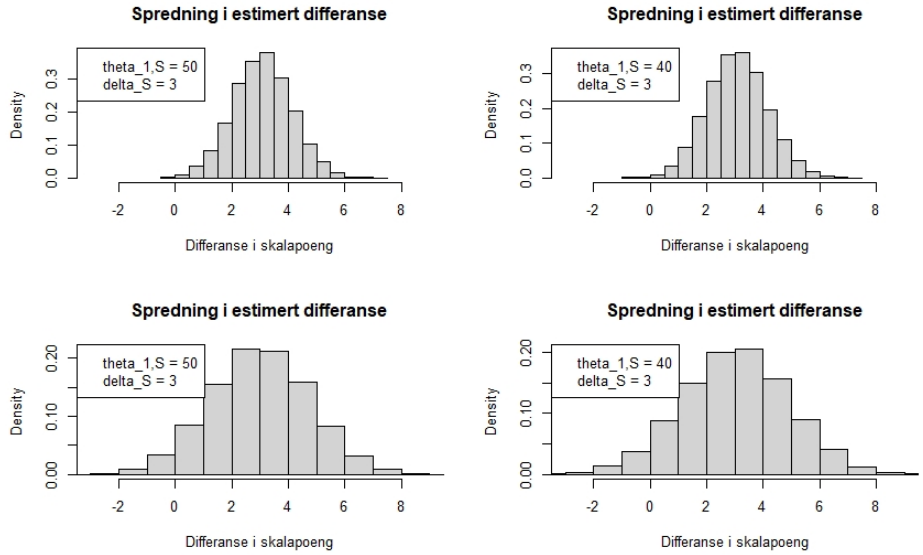


Figur 16: Spredning i estimert ferdighet for ulike klassestørrelser med gammafordelt ferdighet.

Tabell 7: Klassesnitt, store mot små klasser ved ulike gammafordelinger, fiksert θ, SD = standardavvik til det estimerte klassesnittet. $N_{sim} = 1000$, N_{klasse} = klassestørrelse.

$N_{klasse} = 30$	Gj.snitt	SD	2.5%	25%	50%	75%	97.5%
$\theta \sim \Gamma(5,1) - 6.0$	39.5	1.12	37.3	38.8	39.5	40.3	41.6
$\theta \sim \Gamma(5,1) - 5.0$	49.9	1.02	47.9	49.2	49.9	50.6	52.0
$\theta \sim \Gamma(5,1) - 4.0$	60.0	1.04	58.0	59.3	60.0	60.7	62.2
$N_{klasse} = 10$	Gj.snitt	SD	2.5%	25%	50%	75%	97.5%
$\theta \sim \Gamma(5,1) - 6.0$	40.0	1.76	36.3	38.8	40.0	41.2	43.1
$\theta \sim \Gamma(5,1) - 5.0$	50.3	1.60	47.1	49.3	50.3	51.4	53.4
$\theta \sim \Gamma(5,1) - 4.0$	60.8	1.76	57.5	59.6	60.7	61.9	64.4

Øverste rad i figur 16 viser spredningen i estimerte klassegjennomsnitt for klasser på 30 elever når vi antar en gammafordelt sann θ i klassen, med gjennomsnitt som skalert vil tilsvare henholdsvis 40, 50 og 60. Neste rad viser klasser på 10 elever med tilsvarende θ fordelinger. Tilhørende tabell, tabell 7 viser kvantilfordelinger for den estimerte gjennomsnittskåren for klassene av ulik størrelse og med gammafordelte θ med ulikt gjennomsnitt. For en klasse på 30 elever med $\bar{\theta}_S = 49.9$ er det estimerte klassegjennomsnittet i 50% av tilfellene innenfor $\approx \bar{\theta}_S \pm 0.7$ og i 95% innenfor $\approx \bar{\theta}_S \pm 2.1$. Reduserer vi klassestørrelsen øker vi usikkerheten ytterligere, for en klasse på 10 elever med $\bar{\theta}_S = 50.3$ er 50% innenfor $\approx \bar{\theta}_S \pm 1.1$ og 95% innenfor $\approx \bar{\theta}_S \pm 3.2$. Sammenligner vi de estimerte klassesnittene for klasser med antatt normalfordelt θ_S med klasser med antatt gammafordelt θ_S, rundt de samme snittene, ser vi at usikkerheten er større i klassene med gammafordelt θ_S. I de mest ekstreme tilfellene, der klassene er små og nivået er generelt høyt, er spredningen i det estimerte klassesnittet opp i mot 1.5 ganger høyere ved antatt gammafordeling mot normalfordeling. Dette er naturlig å forvente ettersom usikkerheten i klassesnittet øker jo færre elever det er i klassen og jo lengre ut i ytterkantene av skalaen vi kommer. I tillegg er det større spredning i klasser der ferdigheten antas å være gammafordelt, alle tre effektene vil sammen bidra til å øke usikkerheten i klassesnittet.



Figur 17: Spredning i estimert differanse i klassegjennomsnitt, øverste rad: $N_{klasse} = 30$, nederste rad: $N_{klasse} = 10$, $N_{sim} = 10000$.

6.2.1 Endring i ferdighet

På samme måte som vi kan simulere estimert endring i ferdighet hos en enkelelev fra 8. klasse til 9. klasse, kan vi også simulere estimert endring i klassegjennomsnittet fra 8. klasse til 9. klasse. Gjennom denne simuleringen kan vi estimere hvor ofte en økning i klassens snitt fra 8. til 9. klasse betyr en faktisk forbedring i klassens ferdighet. Algoritmen bygger på den som brukes ved simulering av enkeltelevers differanse. Det trekkes nå en $\theta_1 \sim N(\mu_\theta, 1)$, der μ_θ = klassens gjennomsnittlige ferdighet ved år 1. Deretter trekkes en $\Delta \sim N(\mu_\Delta, 0.1)$ der Δ = den reelle differansen i ferdighet per elev fra år 1 til år 2. Vi setter så $\theta_2 = \theta_1 + \Delta$. La så $\bar{\theta}_2 - \bar{\theta}_1 = D_k$ være endringen i klassens gjennomsnitt fra år 1 til år 2. Gjenta prosessen N_{sim} ganger og vi får den estimerte endringen i klassens gjennomsnitt fra det ene året til det neste, $\hat{\bar{\theta}}_2 - \hat{\bar{\theta}}_1 = \hat{D}_k$.

Figur 17 viser estimert endring i klassegjennomsnittet fra et år til neste basert på klassens estimerte gjennomsnittskåre i år 1 og år 2 for klasser med 30 elever og antatt normalfordelt ferdighet. Tabell 8 viser hvor ofte det ble

Tabell 8: Estimert differanse i klassesnitt, ulike klassestørrelser, $\bar{\theta}_{1,S}$ = gjennomsnittlig skalapoeng for klassen i år 1, Δ_S = reell differanse i skalapoeng, øvre halvdel: $N_{klasse} = 30$, nedre halvdel: $N_{klasse} = 10$, $N_{sim} = 1000$.

$\bar{\theta}_{1,S} + \Delta_S$	$\hat{D}_k > 0$	$\hat{D}_k \in (D_k \pm 1)$	$\hat{D}_k \in (D_k \pm 2)$
$50 + 0$	49.7%	66.6%	94.4%
$40 + 3$	99.9%	64.8%	93.9%
$50 + 3$	99.9%	67.5%	93.8%
$60 + 3$	99.9%	60.5%	89.3%
$\bar{\theta}_{1,S} + \Delta_S$	$\hat{D}_k > 0$	$\hat{D}_k \in (D_k \pm 1)$	$\hat{D}_k \in (D_k \pm 2)$
$50 + 0$	49.5%	41.2%	72.8%
$40 + 3$	94.8%	39.6%	69.7%
$50 + 3$	96.3%	44.2%	74.1%
$60 + 3$	93.3%	40.2%	69.0%

rapportert forbedring av snittskåre i klassen fra det ene året til det neste, og hvor ofte den rapporterte forbedringen var innenfor henholdvis ± 1 og ± 2 fra den faktiske forbedringen for klasser på størrelse 30, og klasser på størrelse 10. Det viser seg at selv ved små klasser ($N_{klasse} = 10$) og liten forbedring i snittskåre i klassen ($\Delta = 3$) vil en fortsatt få rapportert en forbedring i 90% av tilfellene. Dette betyr at om en lærer opplever at klassen sin fra 8. til 9. klasse går fra 50 til 53 i snittskåre, kan han være nokså sikker på klassen faktisk har blitt flinkere i de ferdighetene som prøven måler.

6.3 Skoler

Ved simulering av resultater på skolenivå blir algoritmen den samme som ved simulering av klasser. Vi simulerer en svarvektor $\mathbf{u}$ for hver elev på skolen basert på en gitt sann ferdighet. Når vi har svarvektoren til hver elev bruker vi sannsnylighetsmaksimering for å finne estimert $\hat{\theta}$ til hver elev på skolen. Denne prosessen gjentas N_{sim} ganger der vi enten simulerer et nytt sett med svarvektorer for hver iterasjon eller fikserer den sanne θ-fordelingen. Der den første metoden vil gi oss usikkerhet knyttet til spredningen i elevgruppen på skolen fra år til år, mens den andre gir oss usikkerheten til det prøven måler for akkurat den elevgruppen som tok prøven det året.

6.3.1 Ferdighetsfordeling

Ser vi på resultatene for nasjonale prøver på nasjonalt nivå der ($N = 60000$) ser det ut som en standard normalfordeling beskriver ferdigheten på landsbasis godt. Gjennomsnittlig skalapoeng i 2020 var 50 og spredningen i skalapoengene, målt fra 20. prosentil til 80. prosentil, $(\eta_{.20}, \eta_{.80}) = (41, 58)$ er symmetrisk. Vi kan også sjekke hvor godt normalfordelingen passer til hvor mange elever som ble plassert innenfor hvert mestringsnivå i 2020. Poenggrensene i skalapoeng for hvert mestringsnivå i regning ble i 2014 satt til:

$$\theta_S = (\leftarrow, 36), (37, 44), (45, 54), (55, 62), (63, \rightarrow)$$

Regner vi om til θ: $\theta = \theta_S - 50/10$ kan vi finne grensene mellom hvert mestringsnivå på ferdighetsskalaen. Lar vi M_l være nedre grense for mestringsnivået og M_u være øvre grense for mestringsnivået kan vi finne teoretisk andel av populasjonen innenfor hvert mestringsnivå gitt for eksempel en standard-normalfordelt ferdighet:

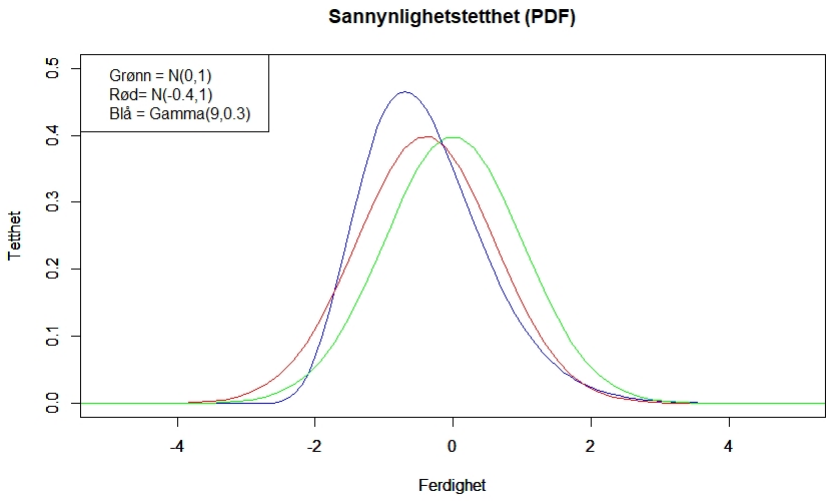
$$\text{Andel av populasjon i mestringsnivå } (M_l, M_u) = \int_{M_l}^{M_u} f(\theta) = \frac{1}{\sqrt{\sigma}} e^{-\frac{1}{2}\left(\frac{\theta}{1}\right)^2} d\theta \tag{54}$$

Det er nå mulig å gjøre en vurdering på hvor godt ulike fordelinger passer til de data som rapporteres av Utdanningsdirektoratet.

Tabell 9: Sammenligning av mestringsnivå i regning i 2020 med teorietisk fra en standard normalfordeling på nasjonalt nivå ($N = 60000$).

Mestringsnivå	1	2	3	4	5	Spredning
Udir 2020	7.7%	23.3%	38.5%	20.1%	10.3%	(41,58)
$\theta \sim N(0, 1)$	8.9%	22.0%	38.2%	22.1%	10.6%	(41,58)

Siden en standard normalfordeling ser ut til å passe godt med resultatene på nasjonalt nivå, er det naturlig å forvente at enkelte skoler vil ha elevgrupper der ferdighetsfordelingen er asymmetrisk. Vi har nå mulighet til å teste hvor godt ferdighetsfordelingen vi bruker til å estimere spredning i skalapoeng ved ulike skoler passer med de faktiske dataene rapportert for skolen.



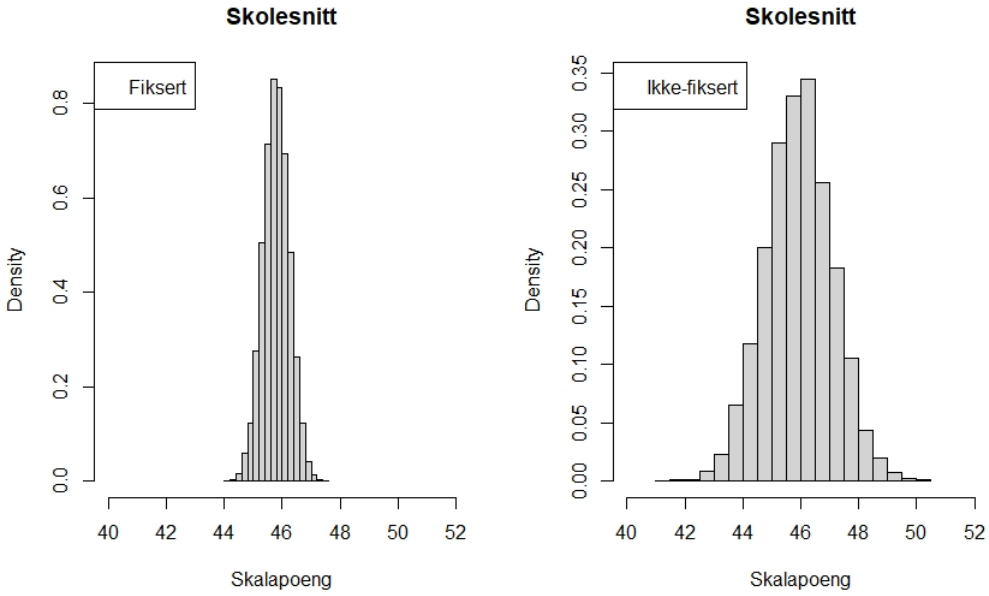
Figur 18: Sammenligning av $\Gamma(9, 0.3) - \mu_\Gamma$ og $N(-0.4, 1)$ fordelingene med standard normalfordeling.

6.3.2 Case: Vadsø kommune

Vadsø kommune har to ungdomsskoler der til sammen 75 elever på 8. trinn svarte på nasjonale prøver i regning i 2020. Gjennomsnittlig skalapoeng, $\bar{\theta}_S = 46$ og spredningen, $(\eta_{.20}.\eta_{.80}) = (39.52)$. La oss se hvor godt ulike sannsynlighetsfordelinger med et gjennomsnitt som tilsvarer 46 skalapoeng passer til de rapporterte dataene fra ungdomsskolene i Vadsø kommune i 2020.

Tabell 10: Sammenligning av rapporterte mestringsnivå i regning i 2020 ved ungdomsskolene i Vadsø kommune($N = 75$), med teoretisk andel av populasjon innenfor mestringsnivå ved gammafordeling og normalfordeling. $\bar{\theta}_S = 46$

Mestringsnivå	1	2	3	4	5	Spredning
Udir 2020	10.7%	36.0%	36.0%	12.0%	5.3%	(39,52)
$\theta \sim \Gamma(9, 0.3) - \mu_\Gamma$	13.6%	36.4%	35.6%	11.8%	4.8%	(39,53)
$\theta \sim N(-0.4, 1)$	17.1%	28.9%	36.2%	14.8%	4.9%	(38,54)



Figur 19: Spredning i estimert gjennomsnitt basert på $\theta \sim \Gamma(9, 0.3) - \mu_\Gamma$, fiksert og ikke-fiksert θ, $N = 75, N_{sim} = 10000$.

Fordelingen $\Gamma(9, 0.3) - \mu_\Gamma$ er en gammafordeling med formparameter $= 9$ og skalaparameter $= 0.3$, som er skiftet med $\mu_\Gamma = form * skala - \bar{\theta}$. Der $\bar{\theta}$ er gjennomsnittlig ferdighet regnet om fra skalapoeng rapportert i 2020. Det ble på bakgrunn av argumentene fra Chen (2019) valgt en gammafordeling som utgangspunkt, deretter ble det gjennom prøving og feiling funnet form og skalaparametere som fikk ligning 54 til å stemme best mulig overens med data fra Udir. Det kommer tydelig fram fra tabell 10 at den skiftede gammafordelingen gir et bilde av data fra Vadsø kommune som stemmer bedre overens med virkeligheten enn en normalfordeling rundt det samme gjennomsnittet. Det er på bakgrunn av dette at $\Gamma(9, 0.3) - \mu_\Gamma$ blir fordelingen som brukes videre i simuleringseksemplene for Vadsø kommune.

I tillegg til det vist i tabell 10 rapporterer Udir(2021) en usikkerhet til gjennomsnittet. Denne usikkerheten er et teoretisk 95% konfidensintervall rundt gjennomsnittet basert på spredningen i resultatene til alle elevene som tok testen det året. Denne usikkerheten vil synke jo flere elever som tar prøven. Usikkerheten som er rapportert for gjennomsnittlig skalapoeng for skolene i

Tabell 11: Spredning i estimert gjennomsnitt for en skole der elevgruppen har $\Gamma(9, 0.3) - \mu_\Gamma$-fordelt ferdighet. Første og andre kolonne er estimert gjennomsnittskåre og standardavviket til det estimerte snittet. De neste kolonnene viser kvantilfordeling til det estimerte snittet. Siste kolonne viser estimert spredning av skalapoeng i fordelingen der spredningen er $(\eta_{.20}, \eta_{.80})$, $N = 75, N_{sim} = 10000$.

Vadsø	$\hat{\bar{\theta}}_S$	SD	2.5%	25%	50%	75%	97.5%	Spredning
Fiksert	46	0.4	44.9	45.2	45.8	46.1	46.7	(38.1,53.6)
Ikke fiksert	46	1.2	43.7	45.2	45.9	46.7	48.2	(37.7,53.6)

Vadsø kommune 2020 er: $Usikkerhet = \pm 2.0$. Ser vi til tabell 11 kan vi se at når θ er fiksert blir tilsvarende usikkerhet: ± 0.9 og dersom θ ikke er fiksert blir tilsvarende usikkerhet ± 2.3. En person med ansvar for utdanning i Vadsø kommune kan lure på hvordan en skal tolke disse forskjellige usikkerhetene.

Svaret på det blir at Vadsø kommune kan være 95% sikker på at den ferdigheten som ble rapportert i regning i 2020 for akkurat den elevgruppen det året er innenfor ± 0.9 av den faktiske ferdigheten elevgruppen viste på prøven, så usikkerheten på ± 0.9 er altså knyttet til selve prøven. Usikkerheten til gjennomsnittet på ± 2.3 betyr at om de skal være 95% sikre på at neste års kull sitt gjennomsnitt skal være forskjellig fra årets må $\bar{\theta}_S \pm$ usikkerhet neste år, ikke overlappe med 46 ± 2.3. Altså om det neste år rapporteres en $\bar{\theta}_S = 48$ med usikkerhet 2.0, kan de ikke være sikre på at skolen har økt ferdigheten til elevgruppen sammenlignet med fjoråret. Denne usikkerheten er dermed knyttet til naturlig variasjon i elevgruppen. En mulig forklaring på at den estimerte usikkerheten (2.3) er høyere enn den rapporterte (2.0) kan være at selv om gammafordelingen passer godt med mestringsnivåene, vil det ved en gammafordeling kunne forekomme enkelte usannsynlig høye θ verdier som gir en kunstig økning i usikkerheten.

6.3.3 Case: Kannik skole

Kannik skole hadde i 2020 ved 8. trinn 182 elever som gjennomførte nasjonale prøver i regning. Gjennomsnittlig skalapoeng var 53. Siden ferdighetsfordelingen ved Kannik skole er asymmetrisk motsatt vei som skolene i Vadsø sjekkes det nå om en speilet versjon av gammafordelingen passer dataene bedre enn en normalfordeling.

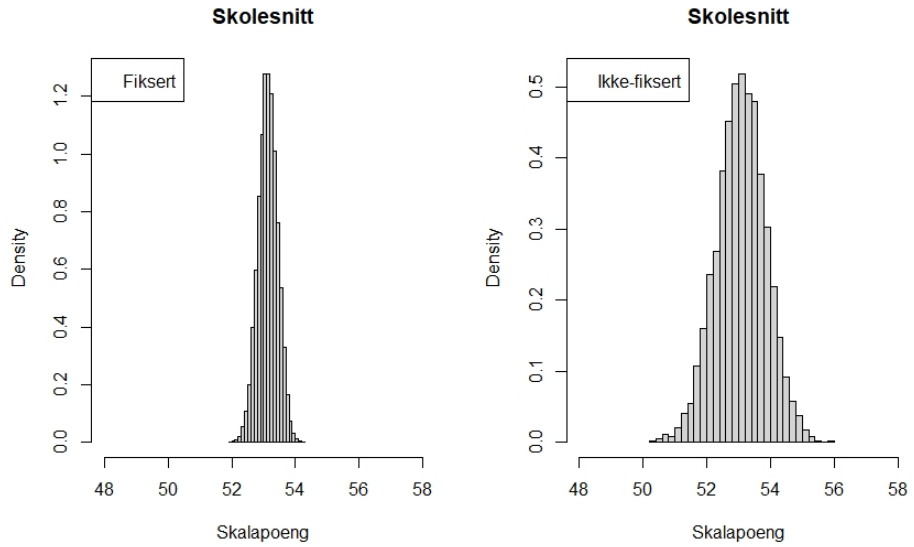
Tabell 12: Sammenligning av rapporterte mestringsnivå i regning i 2020 ved Kannik skole($N = 182$), med teoretisk andel av populasjon innenfor mestringsnivå ved gammafordeling og normalfordeling. $\bar{\theta}_S = 53$

Mestringsnivå	1	2	3	4	5	Spredning
Udir 2020	4.4%	15.4%	34.1%	27.5%	18.7%	(45,62)
$\theta \sim -\Gamma(9, 0.3) + \mu_\Gamma$	5.4%	13.3%	34.0%	32.5%	15.6%	(46,62)
$\theta \sim N(0.3, 1)$	4.9%	16.2%	36.2%	26.9%	17.1%	(45,61)

Det ser ut i fra tabell 12 ikke ut til at den speilede gammafordelingen gir en bedre beskrivelse av de virkelige dataene enn normalfordelingen. Dette kan trolig forklares ved at jo større elevgruppen blir jo mer vil ferdigheten til elevgruppen nærme seg en normalfordeling. På bakgrunn av data fra tabell 12 vil en normalfordeling med et snitt som tilsvarer 53 skalapoeng brukes som utgangspunkt for simuleringseksemplene for Kannik skole.

Tabell 13: Spredning i estimert gjennomsnitt for en skole der elevgruppen har $N(0.3, 1)$-fordelt ferdighet. Første og andre kolonne viser estimert gjennomsnittskåre og standardavviket til det estimerte snittet. De neste kolonnene viser kvantilfordeling til det estimerte snittet. Siste kolonne viser estimert spredning av skalapoeng i fordelingen der spredningen er $(\eta_{.20}, \eta_{.80})$.

Kannik	$\hat{\bar{\theta}}_S$	SD	2.5%	25%	50%	75%	97.5%	Spredning
Fiksert	53	0.3	52.5	52.9	53.1	53.3	53.7	(44.1,62.4)
Ikke fiksert	53	0.8	51.5	52.6	53.1	53.6	54.6	(44.8,61.8)



Figur 20: Spredning i estimert gjennomsnitt basert på $\theta \sim N(0.3, 1)$, fiksert og ikke-fiksert θ, $N = 182, N_{sim} = 10000$.

Utdanningsdirektoratet rapporterer en usikkerhet på ± 1.5 rundt gjennomsnittskåren i regning for elevene ved Kannik skole. Tabell 13 viser den estimerte usikkerheten til å være ± 0.6 knyttet til prøven og ± 1.6 knyttet til fordelingen i elevgruppen. Som forventet blir estimatene mer presise når størrelsen på elevgruppen øker. Usikkerhetene her kan tolkes slik at om neste års kull skulle få en gjennomsnittskåre på 55, kan en med rimelig sikkerhet si at det kullet hadde høyere regneferdighet enn fjorårets, men en kan ikke med rimelig sikkerhet konkludere med at det skyldes noe annet enn naturlig variasjon i ferdighetsfordelingen til elevgruppen. Om skolen har satt inn tiltak og neste års kull får $\bar{\theta}_S = 50$, eller $= 56$, kan det være et grunnlag for å tro at tiltakene har hatt en effekt.

7 Diskusjon

Denne oppgaven har tatt for seg deler av IRT som er relevante for bruk ved nasjonale prøver og lignende tester i undervisningssammenheng. Deretter har IRT-metoder beskrevet i kapittel 3 blitt tatt i bruk for å simulere tenkte situasjoner som lærere og andre med ansvar for kvalitetssikring av undervisning kan oppleve, for så å analysere og tolke data som kommer ut av disse simuleringene. Dette kapittelet vil diskutere bruk av IRT-metoder i nasjonale prøver samt diskutere resultatene av simuleringseksemplene.

7.1 IRT i nasjonale prøver

Når en ønsker å måle underliggende ferdigheter hos elever over tid, er det særdeles viktig at måleverktøyet er godt og at måleverktøyet brukes riktig og konsekvent. Med bruk av klassisk testteori, eller poengsummer fra prøver er det vanskelig å si om elevers endring i resultat fra år til år skyldes endring i prøvene eller endring i elevenes ferdighet. Konsekvent bruk av IRT-analyser og lenking og ekvivalering over tid vil kunne sørge for at resultatene fra prøvene er sammenlignbare over tid.

IRT-metoder definerer oppgavens vanskegrad og prøvetakerens underliggende ferdighet på samme skala. Oppgavenes vanskegrad er også uavhengig av ferdigheten til personen som svarer på oppgaven. I motsetning til klassisk testteori gjør dette det mer meningsfult å sammeligne ferdighet og vanskegrad på en prøve eller oppgave. Ved klassisk testteori eller en vanlig eksamen vil da resultatet til prøvetakeren være definert på grunnlag av akkurat den prøven. IRT-metoder gjør det mulig å gi mere presis og sammenlignbar informasjon om de underliggende ferdighetene til elever som tar ulike prøver på tvers av prøvene og over tid.

Et viktig element som introduseres i IRT er testens reliabilitet gjennom informasjonsfunksjonen. Utenfor IRT måles testers presisjon ofte med en endimensjonell usikkerhet. Det kommer tydelig fram av informasjonsfunksjonen at testens presisjon sjeldent er uniform over hele ferdighetsskalaen. Dette gjør IRT-metoder svært nyttig for de som skal konstruere prøver med spesifikke formål, da de kan kalibrere prøven til å ha høy presisjon på det område på ferdighetsskalaen som er relevant for formålet.

Det at IRT-metoder ble tatt i bruk fra 2014 for å konstruere, gjennomføre og analysere resultater fra nasjonale prøver gir oss presis informasjon som er sammenlignbar over tid, som ikke hadde vært mulig uten å innføre IRT. Når regjeringen nå bestiller endring i nasjonale prøver i samsvar med innføring av nye læreplaner er det viktig å være klar at om endringene i prøvene skjer for fort og ustrategisk mister vi en god mulighet til å presist måle effekten av de nyelæreplanene.

7.2 Resultater fra simuleringer

Det ble i simuleringene i Kapittel 6 tatt utgangspunkt i et oppgavesett på 58 oppgaver fra nasjonale prøver i regning for 8. trinn i 2014. Om oppgavenes egenskaper har hatt signifikante endringer fra 2014 vil tolkning av resultater i 2020 bære preg av dette. Det ble simulert resultater for elever på enkeltnivå, klassenivå og skolenivå. Dette delkapittelet diskuterer resultatene fra simuleringene i tråd med formålet for prøvene, som en del av NKVS er å gi informasjon om kvaliteten på utdanningstilbudet i landet. Det er også ment som veiledning til lærere for planlegging av undervisning og kartlegging av elevens kompetanse.

Simuleringene for enkeltelever viser stor usikkerhet rundt skalapoeng rapportert for elevene. En elev på midten av ferdighetsskalaen vil i $\approx 80\%$ av tilfellene få rapportert skalapoeng innenfor ± 5 av det som representerer elevens sanne ferdighet. For elevene som ligger to standardavvik unna midten er det i kun $\approx 70\%$ av tilfellene de vil få skalapoeng innenfor ± 5 det som representerer deres sanne ferdighet. Ser vi på elevers forbedring fra 8. trinn til 9. trinn er det $\approx 80\%$ sannsynlig å se en forbedring i skalapoeng hos en elev som i realiteten har hatt en forbedring som tilsvarer 5 skalapoeng. Altså er det $\approx 20\%$ sannsynlig at eleven opplever ingen forbedring eller en lavere skåre i 9. trinn til tross for en reell forbedring som tilsvarer 5 skalapoeng. Dette betyr at om en lærer mener at en elev i et enkelttilfelle er flinkere eller mindre flink enn hva skalapoengene på nasjonaleprøver forteller er det grunnlag for læreren å stole mer på egen kjennskap til eleven enn nøyaktig hvilket tall nasjonale prøver rapporterer. Nasjonale prøver gir for enkeltelever beskrivelser av hvilke fagområder elever skårer bra og dårlig på. Det blir mer naturlig for en lærer å se til disse beskrivelsene å bruke de som et supplement til allerede etablert kunnskap om elevens ferdighet.

Simuleringseksemplene for skoleklasser tok for seg to ulike situasjoner. Den ene simulerer fra akkurat den samme populasjonen hver gang, og data fra simuleringene fortelles oss noe om usikkerheten knyttet til prøven for akkurat den populasjonen. Den andre simulerer fra en ny antatt ferdighetsfordeling for hver iterasjon, og data herfra vil fortelle oss om hva vi kan forvente av naturlig variasjon fra år til år, eller mellom klasser på samme trinn. Det er viktig at de som tolker data fra nasjonale prøver er klar over om usikkerheten i gjennomsnittet er knyttet til populasjonen eller prøven. Prøven måler klassegjennomsnittet nokså godt, en lærer med en klasse på 30 elever som får gjennomsnittskåre på 50, kan være $\approx 95\%$ sikker på at innenfor ± 1.5 av det faktiske klassesnittet, dersom en antar at klassens ferdighet er normalfordelt. Det viser seg at det i tilfeller er fordelinger med større variasjon og mindre symmetri enn normalfordelingen som bedre beskriver virkeligheten, antar den samme læreren en $\Gamma(5.1)$ fordeling av ferdigheten i klassen rundt det samme gjennomsnittet vil usikkerheten øke til ± 2.5. Usikkerheten knyttet til fordelingen av elevene er større, ved antatt normalfordelt klasse og et gjennomsnitt på 50 skalapoeng vil det estimerte gjennomsnittet i 95% av tilfellene havne innnenfor ± 4.3 av det faktiske gjennomsnittet. For en lærer vil dette bety at om en klasse det ene året har 5 skalapoeng høyere snitt i regning enn forrige års klasse, kan han med rimelig sikkerhet si at klassen i år er bedre i regning enn den han hadde i fjor. Læreren kan fortsatt ikke med rimelig sikkerhet si at dette skyldes noe annet enn naturlig variasjon i ferdighetsfordelingen til elevgruppene.

I likhet med simuleringene for skoleklasser ble det i forsøkene med hele skoler gjort simuleringer for usikkerhet knyttet til prøven og usikkerhet til variasjon i elevgruppen. Som forventet blir resultatene mer presise jo større elevgruppen som tar prøven er. Selve prøven måler gjennomsnittet på skolenivå svært presist. Selv ved små skoler ($N = 75$) kan vi med 95% sikkerhet si at gjennomsnittet som prøvene rapporterer for skolen er innenfor ± 1 av det faktiske gjennomsnittet for det kullet. En bør likevel være forsiktig når en ser på gjennomsnittskåren til små skoler og skoleklasser isolert sett, ettersom usikkerheten knyttet til naturlig variasjon i elevers ferdighet vil ha mye å si for hvor godt akkurat den skolen akkurat det året gjør det. Det viser seg også at denne usikkerheten kan være større enn ved en antatt normalfordeling.

Referanser

Baker, F.B. & Kim, S.-H. (2017). *The basics of item response theory using r.* Springer. Hentet fra http://search.ebscohost.com/login.aspx?direct=true&db=nlebk&AN=2540801&scope=site

Bjørnsson, J.K. (2018). *Metodegrunnlag for nasjonale prøver.* (Rapport fra utdanningsdirektoratet). Hentet 15. juni 2021 fra https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover-august-2018.pdf

Bjørnsson, J.K., Caspersen, M. & Lie, S. (2005). *Nasjonale prøver på prøve: rapport fra en utvalgsunde rsøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2004.* (Rapport fra utdanningsdirektoratet). Hentet 15. juni 2021 fra https://www.udir.no/globalassets/upload/forskning/5/nasjonale_prover_pa_prove.pdf

Brennan, R.L. (red.). (2006). *Educational measurement* (4th ed. utg.). Westport, Conn: Praeger.

Chen, K. (2019). *A comparison of fixed item parameter calibration methods and reporting score scales in the development of an item pool.* (Doktorgradsavhandling). University of Iowa.

De Ayala, R. (2009). *The theory and practice of item response theory.* New York: Guilford press.

Hambleton, R.K., Swaminathan, H. & Rogers, H.J. (1991). *Fundamentals of item response theory* (vol. 2). Sage.

Kolen, M.J. & Brennan, R.L. (2014). *Test equating, scaling, and linking: Methods and practices.* Springer Science & Business Media.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems.* New York: Routledge.

Sjøberg, S. (2019). *Nasjonalt kvalitetsvurderingssystem for skolen.* i Store norske leksikon på snl.no. Hentet 15. juni 2021 fra https://snl.no/NKVS_Nasjonalt_Kvalitetsvurderingssystem_for_skolen

Tokle, O.D., Ravlo, G., Johansen, O.H., Myhre, S.A. & Thoresen, M. (2019). *Teknisk rapport for nasjonal prøve i regning, 8.trinn 2018* (Teknisk rapport). Matematikksenteret, NTNU.

Udir. (2017a). *Hva måler nasjonale prøver i engelsk?* Hentet 15. juni 2021 fra https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/mestringsbeskrivelser-og-hva-provene-maler/kva-maler-nasjonale-prove-i-engelsk/

Udir. (2017b). *Hva måler nasjonale prøver i regning?* Hentet 15. juni 2021 fra https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/mestringsbeskrivelser-og-hva-provene-maler/hva-maler-nasjonal-prove-i-regning/

Udir. (2017c). *Rammeverk for de grunnlegende ferdigheter.* Hentet 15. juni 2021 fra https://www.udir.no/laring-og-trivsel/rammeverk/rammeverk-for-grunnleggende-ferdigheter/

Udir. (2017d). Rammeverk for nasjonale prøver. Hentet 15. juni 2021 fra https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover/

Udir. (2018). *Administrere nasjonale prøver.* Hentet 15. juni 2021 fra https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/administrere-nasjonale-prover2/#ansvar-for-gjennomforingen

Udir. (2019). *Resultater på nasjonale prøver 2019 for 8. og 9. trinn - analyse.* Hentet 15. juni 2021 fra https://www.udir.no/tall-og-forskning/finn-forskning/tema/nasjonale-prover/nasjonale-prover-8.-og-9.-trinn-2019/

A R Kode

```
#Definerer funksjoner til bruk i simuleringen:

#Estimerer ferdighet med fikserte parametere:
ability <- function(mdl,u, b, a ,c){
  J <- length(b)
  if (mdl == 1 | mdl == 2 | missing(c)) {
    c <- rep(0,J)
  }
  if (mdl == 1 | missing(a)) { a<- rep(1, J)}
  x <- sum(u)
  if (x == 0) {
    th <- -log(2*J)
  }
  if (x == J) {
    th <- log(2*J)
  }
  if (x == 0 | x == J){
    sumdem <- 0.0
    for (j in 1:J) {
      pstar <- 1 / (1 + exp(-a[j] * (th - b[j])))
      phat <- c[j] + (1.0 - c[j]) * pstar
      sumdem <- sumdem - a[j]^2 * phat * (1.0 - phat) * (pstar/phat)^2
    }
    se <- 1 / sqrt(-sumdem)
  }
  if (x!=0 & x!=J) {
    th <- log(x/(J - x))
    S <-10
    ccrit <- 0.001
    for (s in 1:S) {
      sumnum <- 0.0
      sumdem <- 0.0
      for (j in 1:J) {
        pstar <- 1 / (1+exp(-a[j] * (th - b[j])))
        phat <- c[j] + (1.0 - c[j]) * pstar
```

```
        sumnum <- sumnum + a[j] * (u[j] - phat) * (pstar/phat)
        sumdem <- sumdem -a[j]^2 * phat * (1.0 -phat) * (pstar/phat)^2
      }
      delta <- sumnum /sumdem
      th <- th - delta
      if (abs(delta) < ccrit | s == S) {
        se <- 1/sqrt(-sumdem)
        break
      }
    }
  }
  #cat(paste("th=", th, "\n")); flush.console()
  #cat(paste("se=", se, "\n")); flush.console()
  thse <- c(th, se)
  return(thse)
}


#Funksjon for     f   svarvektor:
IRcurve <- function(a,b,theta)
  1/(1+exp(-a*(theta-b)))

SimScores <- function(a,b,theta){
  probs <- IRcurve(a,b,theta)
  scores <- as.numeric(probs>runif(length(b)))
  return(scores)
}


# Gjenta mange ganger for samme person/ferdighetsniv
#Ferdighetsniv  0
nsim <- 100000
esttheta <- numeric(nsim)
sdtheta <- numeric(nsim)
truetheta <- 0
for(i in 1:nsim){
  res <- ability(mdl=2,u=SimScores(a,b,truetheta),b=b, a=a)
  esttheta[i] <- res[1]
```

```
  sdtheta[i] <- res[2]
}



#Klassest rrelse 10, normalfordelt:
#Set.seed(187) for theta_S = 50
nrep <- 10000
meanvektor <- numeric(nrep)
nsim <- 10
esttheta <- numeric(nsim)
set.seed(187)
truetheta <- rnorm(nsim,0,1)
mean(truetheta)
for(k in 1:nrep){
  for(i in 1:nsim){
    res <- ability(mdl=2,u=SimScores(a,b,truetheta[i]),b=b, a=a)
    esttheta[i] <- res[1]
  }
  SESTTH <- esttheta*10+50
  meanvektor[k] <- mean(SESTTH)
}



#KLASSEDIFFERANSE

#53 - 50 | set.seed(186) <- N_Klasse 30
nrep <- 10000
meanvektor_1 <- numeric(nrep)
meanvektor_2 <- numeric(nrep)
meandiff <- numeric(nrep)
nsim <- 30 #Klassest rrelse 30
esttheta_1 <- numeric(nsim)
esttheta_2 <- numeric(nsim)
```

```
set.seed(186)
theta_1 <- rnorm(nsim,0,1) #Skalasnitt 50
set.seed(5)
differ <- rnorm(nsim,0.3,0.1)
for(k in 1:nrep){
  for(i in 1:nsim){
    res1 <- ability(mdl=2,u=SimScores(a,b,theta_1[i]),b=b, a=a)
    res2 <- ability(mdl=2,u=SimScores(a,b,theta_2[i]),b=b, a=a)
    esttheta_1[i] <- res1[1]
    esttheta_2[i] <- res2[1]
  }
  meanvektor_1[k] <- mean(esttheta_1)
  meanvektor_2[k] <- mean(esttheta_2)
  meandiff[k] <- meanvektor_2[k] - meanvektor_1[k]
}
Klassediff <- meandiff * 10 #Skalere differansen
```

```
#Vads kommune:

shape = 9
scale = 0.3

nrep=10000
meanvektor <- numeric(nrep)
usvek <- numeric(nrep)
Q20 <- numeric(nrep)
Q80 <- numeric(nrep)
nsim <- 75
esttheta <- numeric(nsim)

#Gjennomsnittlig skalapoeng: 46
set.seed(186)
truetheta <- (rgamma(nsim,shape=shape,scale=scale)-shape*scale-0.4)
```

```
mean(truetheta)
for(k in 1:nrep){
  for(i in 1:nsim){
    res <- ability(mdl=2,u=SimScores(a,b,truetheta[i]),b=b, a=a)
    esttheta[i] <- res[1]
  }
  SESTTH <- esttheta*10+50
  meanvektor[k] <- mean(SESTTH)
  Q20[k] <- quantile(SESTTH,0.2)
  Q80[k] <- quantile(SESTTH,0.8)
  usvek[k]<- 1.96*sd(SESTTH)/sqrt(nsim)
}
```