En vurdering av vurderingen

En validitetsundersøkelse av matematikkeksamen i grunnskolen våren 2019

Erling Olbekk

Masteroppgave i matematikkdidaktikk

Institutt for lærerutdanning og skoleforskning

Det utdanningsvitenskapelig fakultet

Universitetet i Oslo

1. desember 2020

En vurdering av vurderingen

En validitetsundersøkelse av matematikkeksamen i grunnskolen våren 2019

Masteroppgave ved ILS, Institutt for lærerutdanning og skoleforskning

Erling Olbekk

© Erling Olbekk

2020

En vurdering av vurderingen: En validitetsundersøkelse av matematikkeksamen i grunnskolen våren 2019

Erling Olbekk

http://www.duo.uio.no

Trykk: Reprosentralen, Universitetet i Oslo

Forord

Det virker rart å kalle dette forord. Det er jo faktisk det aller siste som skrives, og er mer et lettelsessukk enn det er et bidrag til oppgaven. Det tok litt lenger tid enn planlagt, og betydelig mer frustrasjon enn jeg regnet med, men det ser ut som om det kanskje endelig er over.

Nå som den skal leveres inn er det viktig for meg å anerkjenne tre personer som har vært uvurderlige som støtte gjennom prosessen. Jeg vil derfor takke Guri Nortvedt ved ILS for inspirasjon og hjelp med datainnsamling og ideer. Jeg vil også takke min veileder Nils Fredrik Buchholtz for mange flere timer med hjelp enn jeg hadde krav på, og hans uendelige kilde til godt humør og gode tilbakemeldinger under prosessen. Sist, men ikke minst, vil jeg takke Henrik Ræder ved CEMO for uvurderlig hjelp med programmering for analysen. Uten hjelpen og støtten hadde aldri denne oppgaven blitt fullført, og jeg vil være evig takknemlig. Jeg vil til slutt også takke Utdanningsdirektoratet for tilgang til datamaterialet for min analyse.

Til alle mine medstudenter på lektorprogrammet vil jeg si takk for fem gode år, og ønske dere alle lykke til videre. Jeg føler meg trygg på at skolene dere drar til vil være en ressurssterk person rikere.

Sammendrag

Skriftlig eksamen i grunnskolen gis til omtrent femti tusen elever hvert år, på videregående tas det også tusenvis av eksamener. Og selv om mange klager på karakterer, og mener de har fått en urettferdig vurdering, er det sjelden man hører kommentarer om hvor gyldig selve eksamensoppgaven var. Skriftlige eksamener i Norge, og hvordan de relaterer seg til læreplanen, har det vært generelt lite forskning rundt (Utdanningsdirektoratet, 2019b). I denne oppgaven vil leser bli presentert for begrepet validitet i en prøvekontekst. Validitet er et mål på om slutninger som tas på grunnlag av en prøves resultater har tilstrekkelig grunnlag i teori og forskning (AERA, APA, & NCME, 2014). Deretter vil oppgaven gå gjennom en valideringsprosess av skriftlig eksamen gitt i faget MAT0010, våren 2019. Eksamen i MAT0010 er eksamen gitt ved utgang av grunnskolen.

Valideringsprosessen vil undersøke flere aspekter ved eksamen. Vi vil se på om oppgavenes innhold er representative for matematikkdidaktisk forskning rundt matematisk kompetanse, dette for å se om eksamen tester et rikt innhold av matematikk. Denne undersøkelsen er gjort ved hjelp av vurderingsskjemaet for PISA-undersøkelsen (Turner et al., 2015), og er en kvalitativ analyse av oppgavene gitt ved eksamen. Det vil også undersøkes kvantitativt om det er mulig tilstedeværelse av kjønnsbias for å se om deler av eksamen favoriserer enten gutter eller jenter. Resultater av analysene vil bli presentert mot slutten av oppgaven. Funnene tyder på at eksamen tester flere aspekter av hva forskning mener er viktig av grunnferdigheter i matematikk, og at oppgavene hovedsakelig er rettferdige. Det ble også funnet at det for enkelte oppgaver eksisterer betydelig kjønnsbias, og at den totale effekten av dette er betydelig for elever på lavere ferdighetsnivå.

Innholdsfortegnelse

1 Innledning

Hvor mange mennesker i Norge får livet sitt påvirket hvert eneste år på grunn av resultatet fra en test eller prøve? Stryk på teoriprøven til førerkortet kan føre til at du ikke får den jobben som budbilsjåfør. En lav karakter på eksamen kan føre til at du ikke kommer inn på sykepleierutdanningen du drømte om. Alle kan komme på en gang de måtte ta en prøve som ville påvirke mulighetene deres videre i livet. Vi har generelt ganske høy tiltro til resultatene disse prøvene gir, vi tror på at eksamen er en objektiv og rettferdig vurdering av våre evner og ferdigheter. Er det sant?

> A test is only a measuring instrument-an instrument far less precise than most practitioners believe. Such instruments should be used to arrive at inferences about examinees' status with respect to the domain of knowledge, skills, or affect represented by the test. And, because educational tests do not represent with unflawed perfection those domains, the resultant score based inference will often be less than completely accurate. But because educational tests typically yield numerical results, and because human beings usually ascribe excessive accuracy to numbers, many educators regard the results of educational tests with unwarranted deference. If educators think that the measuring instrument is valid, they'll also tend to regard its numerical results as «valid»-that is, as accurate. (Popham, 1997, s. 10)

I skoleåret 2018-2019 gikk nesten 60 000 elever ut av grunnskolen (Utdanningsdirektoratet, 2019a). Alle disse elevene har gjennomført en avsluttende skriftlig eksamen i ett av tre mulige fag, og omtrent en tredjedel har gjennomført eksamen i matematikk. I tillegg til grunnskoleelevene har det blitt gjennomført tusener av eksamener på videregående skole og i privatistsystemet. Karakterene som alle disse prøvene resulterer i har stor påvirkning for eksaminandenes fremtid. De avgjør hvilke videregående skoler de kan velge seg inn på, om de får godkjent fagbrev, og hvilke høyere utdanningsløp som ligger tilgjengelig. Det burde dermed være i alles interesse at resultatene fra disse eksamenene er til å stole på, og at de gir et pålitelig bilde av eksaminandens faktiske ferdighet i faget som testes.

Budskapet i det innledende sitatet fra Popham er verdt å tenke over. En eksamen er et instrument, et instrument designet for å måle en persons kompetansenivå. En vekt, en linjal, en klokke, og et litermål har usikkerhetsmarginer knyttet til sine målinger. Det er rimelig å anta at også en eksamen vil gi et usikkert bilde av kompetansenivået til eksaminanden.

Målet på hvor pålitelig slutningene som trekkes fra testresultater kalles validitet; et ord de fleste har hørt, men som er vanskelig å definere. Sagt med en setning kan det beskrives som om resultatene

fra prøven er gyldige for prøvens hensikt, og er ifølge utdanningsdirektoratet selv det viktigste vurderingsteoretisk begrepet ved eksamen (Utdanningsdirektoratet, 2019b). Denne oppgaven undersøker validiteten til den skriftlige eksamen gitt i faget MAT0010 (matematikk på grunnskolen) våren 2019. Denne prosessen kalles validering, og det kan utforskes fra flere forskjellige utgangspunkt. Tester oppgavene på eksamen de egenskapene og ferdighetene eksamen er ment til å teste? Har alle elevene som tok eksamen hatt god mulighet til å få demonstrert hva de kan? Er noen av oppgavene vanskeligere for enkelte elever enn andre, har de for eksempel vanskelig tekst som noen ikke vil forstå? Alle disse spørsmålene handler om validering, og de fortjener alle et forsøk på besvarelse. Popham (1997), som er sitert over, skriver også at validiteten ved prøver i hovedsak handler om hvor presise slutningene som trekkes av disse prøveresultatene er. For å undersøke dette må vi gå utenfor matematikkdidaktisk teori og se på forskning rundt validering og konsekvenser av prøver. Vi må også se på hva forskning sier om hva matematisk kompetanse er. Forskningsspørsmålene for oppgaven vil presenteres etter teorikapittelet når vi har et godt overblikk over relevant kunnskap.

2 Teori

Relevant teoretisk bakgrunn for oppgaven vil presenteres i de følgende seksjonene under dette kapittelet. Denne teorien kan grovt deles inn i tre hoveddeler. Først vil en gjennomgang av oppgavers rolle i matematikk, og matematisk kompetanse, presenteres for å legge et grunnlag for egenskapen en eksamen i matematikk skal teste. Matematisk kompetanse har vært mye forsket på de siste tiårene, og mange av funnene fra feltet kan kobles opp mot læreplan i matematikk. Deretter følger en seksjon om validitet med utgangspunkt i måling knyttet opp mot tester og prøver. Validitet handler om gyldigheten av slutningene som trekkes av en målings resultater (. Disse slutningene kan ramme forskjellige grupper til ulik grad, og prøver kan dermed ha uheldige konsekvenser med tanke på likeverdig behandling av forskjellige grupper av befolkningen. Dette er også noe oppgaven forsøker å undersøke, og teorikapittelet inneholder derfor også seksjoner om bias i prøver og equity. Bias oppstår når en gruppe har lavere forutsetninger enn andre for å lykkes på en prøve, og dette har dermed konsekvenser for validiteten av prøveresultatene. Equity, som ofte kalles likeverd i norsk litteratur, handler om at alle har lik mulighet til å lære og til å nå så langt de kan.

2.1 Oppgaver i matematikk

Oppgaver benyttes i alle fag og av alle undervisere i skolen. Jeg påstår det kan gjøres et godt argument for at de har en særskilt viktig rolle i matematikkfaget. Oppgaver er på mange måter den viktigste veien til å kunne delta i faglig fordypelse og diskusjon for å lære matematikk. Hvordan oppgaver er dannet og formulert er derfor av interesse for matematikkundervisning. Shimizu, Kaur, Huang, & Clarke (2010) skriver at oppgaver kan avgjøre hvordan elever forstår materialet som undervises, og fungerer som en kontekst for elevenes tankeprosesser. Analyse av oppgaver kan dermed si mye om hvilke kognitive krav som stilles av en elev, og si noe om hvilke egenskaper som kreves for å løse oppgaven. I sin doktorgradsavhandling skriver Pettersen (2019) om matematikkoppgaver, og skiller mellom to kategorier av egenskaper man kan finne i dem. Den første av disse egenskapene er *oppgavens kjennetegn*, den andre egenskapen er *oppgavens krav*. Han skriver at kjennetegn ved en oppgave relaterer til hvordan oppgaven presenteres og formuleres. Med dette menes elementer som oppgavetekst, tilhørende figurer og grafikk, men også hvilket matematisk innhold man finner i oppgaven, slik som geometri, algebra, funksjonslære osv. Oppgavens krav henviser til de egenskapene som er nødvendige for å finne en løsning. Dette inkluderer matematisk kunnskap, men også hvilke mentale egenskaper som kreves. Oppgavens krav er altså ikke like overfladiske som oppgavens kjennetegn, og krever dypere analyse.

Det finnes et velde av litteratur om matematikkdidaktisk forskning rundt oppgaver, og det har tradisjonelt vært et skille mellom oppgaver som brukes i undervisning og oppgaver som brukes i

vurderinger (Pettersen, 2019). Da det tidligere ble sett mest på hvordan oppgaver ble brukt i undervisning har forskning i senere tid begynt å anerkjenne viktigheten av vurderingssituasjoner. Suurtamm et al. (2016) skriver om viktigheten av gode oppgaver i større vurderingstilfeller, slik som *eksamener* og *storskalaundersøkelser* som PISA og TIMSS. Dette kommer av en anerkjennelse av at store vurderinger, som nasjonale prøver og eksamener og internasjonale undersøkelser, har en påvirkende kraft på undervisning og hva som prioriteres i klasserommene. Disse storskalavurderingene bør derfor gjenspeile hva som menes med matematisk ferdighet ifølge forskningsfeltet. Når det nå kommer nye læreplaner i skolen som vektlegger egenskaper ved matematisk kompetanse i de nye kjerneelementene blir det altså enda viktigere at oppgavene på eksamen speiler dette i oppgavenes krav.

2.2 Matematisk kompetanse

En sentral del av analysen i denne studien omhandler komponenter av matematisk kompetanse i oppgavene ved eksamen. I seksjon 2.4.1 vil vi nevne hvordan Blum (2006) forklarer matematisk ferdighet som en egenskap med to dimensjoner; de mer tekniske ferdighetene og de mer grunnleggende ferdighetene som kreves for å mestre matematikk. Blum er ikke alene i dette synet på matematikk, og siden oppgaven undersøker hva som testes av oppgavene ved eksamen er det derfor viktig å se på hva matematisk kompetanse er, og hvordan begrepet brukes i denne oppgaven. Matematikk blir nå ofte sett på som en sammensatt ferdighet med et sammenflettet spekter av underbyggende elementer. På verdensbasis virker det som om de fleste skoleeiere ønsker at elever skal kunne demonstrere bred kunnskap, dyp forståelse og evne til praktisk bruk av matematikken som undervises (Burkhardt, 2014; Pettersen & Nortvedt, 2018). Dette synet på matematikkundervisning er også representert i den norske læreplanen for matematikk fellesfag, og i både kompetansemålene og de grunnleggende ferdighetene som beskrives kommer det tydelig frem (Kunnskapsdepartementet, 2013). Hvis dette skal være et mål for matematikkutdanningen i Norge er det viktig at elever både trenes i denne formen for matematisk arbeid, og at det også testes på eksamen. Hvis dette ikke testes på eksamen risikerer man at det nedprioriteres i undervisningen, og det har også implikasjoner for informasjonen man får fra eksamen og andre vurderingssituasjoner. Burkhardt (2014, s.29) sier:

« Yet it is common to ignore the effect of high-stakes tests on the implemented curriculum, seeing them as «just measurement», and to underfund key elements, notably professional development. »

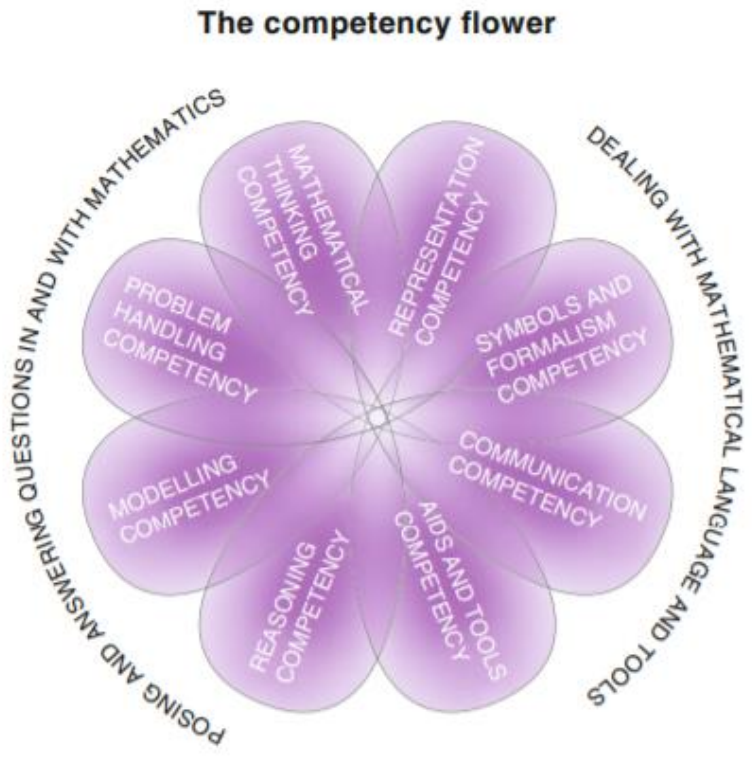
Implikasjonen er åpenbar, hva som testes på eksamen påvirker hva som undervises i klasserommene. Dersom undervisningen skal ruste elevene med et rikt arsenal av kompetanser er det derfor nødvendig at disse representeres i eksamensoppgavene slik at elever har et insentiv til å trene på dem, og at lærere har et insentiv til å tilrettelegge for dette i undervisningen.

Hittil har denne seksjonen bygget opp et bilde av nødvendigheten for at eksamen tester et bredt spekter av matematisk kompetanse, men et viktig spørsmål har ikke blitt besvart. Hva er egentlig matematisk kompetanse? Hvilke komponenter inngår i dette begrepet? Matematisk kompetanse er et konstrukt, en egenskap ved en person, men hvordan dette konstruktet er definert er et spørsmål uten en klar konsensus (Blömeke, Gustafsson, & Shavelson, 2015). På samme måte som for validitet, har mye arbeid har blitt gjort for å utvikle solide rammeverk som grunnlag for videre forskning på matematisk kompetanse (Kilpatrick et. al. 2001; Niss & Jensen, 2002). Opphavet til kompetanserammeverk fra et utdanningsperspektiv er Blooms taksonomi, som deler fag inn i de seks komponentene kunnskap, forståelse, anvendelse, analyse, syntese og evaluering (Kilpatrick,

2014). Denne inndelingen har blitt kritisert av mange, og en av grunnene er at det det ikke beskriver de forskjellige mentale prosessene som kreves for å løse oppgaver i matematikk. I matematikkfaget står oppgaver sentralt i undervisningen (Kilpatrick et al., 2001), og et rammeverk for matematisk kompetanse burde derfor ta mer hensyn til hvordan kompetanse er knyttet til oppgaver.

2.2.1 KOM-prosjektet: Matematikk som åtte kompetanser

KOM, eller «Kompetenceudvikling og Matematiklæring», er et prosjekt fra universitetet i Roskilde som startet sommeren 2000 for å se på muligheter for å utvikle en ny kompetansebasert læreplan for matematikkundervisning i Danmark (Niss & Jensen, 2002). Fruktene av denne arbeidsprosessen ble KOM-rammeverket, som deler konstruktet matematisk kompetanse inn i åtte delkompetanser fordelt mellom to grupper som illustrert i figur 1, og det var også en påvirkende faktor for Blum (2006).



Figur 1: Kompetanserammeverket fra KOM-prosjektet (Niss, 2015)

De åtte kompetansene er ikke illustrert som skarpt adskilte, og det er heller ikke rimelig å forvente at de skal være det; det er tross alt samspillet mellom dem som danner grunnlaget for konstruktet matematisk kompetanse, og de bør derfor ses på som deler av samme egenskap. Men selv om de beskrives som en del av samme kompetanse, og flyter over i hverandre, kan de allikevel beskrives

separat. Vi vil kort beskrive disse kompetansene her, da de er grunnlaget for rammeverket som benyttes senere i oppgaven. Videre i seksjonen bruker jeg min egen oversettelse av kompetansenavnene og beskrivelsene, alle er hentet fra Niss & Jensen (2002).

Tankegangskompetanse: Denne kompetansen beskriver i hovedsak evnen til å kjenne typen spørsmål og svar som er karakteristiske for og forventes i matematikk. Inkludert i denne kompetansen er også forståelsen for matematiske begrep og deres begrensninger.

Problemløsingskompetanse: Som navnet tilsier handler denne kompetansen i stor grad om å kunne stille opp, formulere og løse matematiske problem. Rutineoppgaver som kan løses ved anvendelse av en kjent algoritme faller ikke inn under denne kompetansen, som betyr at matematiske problemer er et subjektivt begrep, og avhenger av personen som løser dem.

Modelleringskompetanse: Kompetansen omhandler modelleringsprosessen, som enklest mulig kan beskrives som å anvende matematikk for å løse reelle problemstillinger. Det inkluderer både kompetanse til å lage egne modeller, og til å vurdere eksisterende modellers gyldighet og begrensninger.

Resonnementskompetanse: Evnen til å kunne følge og vurdere et matematisk resonnement, argumenter og bevis. Eksempler kan inkludere å finne motbevis til en påstand, eller påpeke hvilken del av at argument som er gyldig/ugyldig.

Representasjonskompetanse: Egenskapen til å forstå og gjøre mening av forskjellige former for matematiske representasjoner. Dette inkluderer blant annet grafer, figurer og tabeller. Kompetansen inkluderer også evnen til å se sammenhengen mellom forskjellige representasjonsformer og ferdighet til å lage sine egne representasjoner.

Symbol- og formalismekompetanse: Denne kompetansen dreier seg hovedsakelig om å forstå matematisk symbol- og formelspråk. Å kunne oversette mellom formler og normalt språk, og å kunne håndtere symbolske uttrykk. Et godt eksempel er å kunne jobbe med algebraiske uttrykk.

Kommunikasjonskompetanse: Den viktigste delen av denne kompetansen er å kunne tolke og sette seg inn i andres matematiske utsagn, både i form at tekst og muntlige utsagn. Det inkluderer også evnen til å uttrykke sine matematiske tanker og ideer på en slik måte at andre kan forstå dem.

Hjelpemiddelkompetanse: Den siste av kompetansene omhandler både å ha kjennskap til de forskjellige hjelpemidler og redskap som benyttes i matematikk, og evnen til å benytte dem i sitt

eget arbeid med matematikk. Alle former for hjelpemidler, om det er digitale verktøy eller passer og linjal, faller inn under denne kompetansen.

KOM-rammeverket har hatt stor innflytelse på matematikkundervisning i flere land, og presenteres her i noe detalj da det danner grunnlag for det senere rammeverket utviklet for analyse av oppgaver i PISA-prosjektet (Turner, Blum, & Niss, 2015), som vil bli viktig i analysen.

2.2.2 PISA – En videreutvikling

Arbeidet med å utvikle et rammeverk for matematisk kompetanse ble gjenopptatt da PISA ville skape sin egen definisjon av hva «mathematical literacy», som jeg heretter vil kalle matematisk kompetanse, er da matematikk var hovedfokus for undersøkelsen for første gang i 2003 (Turner et al., 2015). For å kunne si noe om i hvilken grad oppgavene gitt i PISA-undersøkelsen samsvarte med definisjonen av matematisk kompetanse var det nødvendig å ha et rammeverk som kunne benyttes til å analysere oppgavene. Rammeverket er utviklet av PISAs Mathematics Expert Group, heretter kalt MEG.

Som nevnt tidligere var kompetansebeskrivelsene fra KOM-prosjektet en viktig grunnmur for arbeidet, og det deler derfor mange karakteristikker med PISA-rammeverket. Det bør nevnes at Mogens Niss, som ledet arbeidet med KOM-prosjektet også er med i MEG. Det er allikevel flere forskjeller mellom de to rammeverkene, som illustrerer at rammeverk må tilpasses formålet de lages for og at forskjellige institusjoner kan ha forskjellige definisjoner av konstrukter som testes. KOM-rammeverket forsøker kun å si noe om generelle aspekter ved matematisk kompetanse, mens PISA-rammeverket prøver også å være et verktøy for å analysere oppgaver (Turner, Dossey, Blum, & Niss, 2013). I tillegg til å beskrive de forskjellige komponentene som utgjør matematisk kompetanse inneholder rammeverket derfor også en nivåinndeling. Denne inndelingen beskriver i hvor stor grad en kompetanse aktiveres basert på kriterier ved oppgaven. Rammeverket har undergått endringer siden det først ble utviklet i 2003, men de seks kompetansene som beskrives i skjemaet denne oppgaven benytter er ifølge Turner et al. (2015), *Kommunikasjonskompetanse, Strategiutviklingskompetanse, Matematisering, Representasjonskompetanse, Symbol- operasjon og formalismekompetanse, og Resoneringskompetanse.*

Da dette rammeverket er sentralt for analysen vil disse seks kompetansene beskrives mer utdypende. Beskrivelsene er min tolkning og oversetting av rammeverket slik det presenteres i Turner et al. (2015).

Kommunikasjonskompetanse

Denne kompetansen beskriver egenskapen som kreves for å forstå hva som *sies* eller *vises* i en oppgave. Den kreves for å forstå hva oppgaven spør etter, hva det matematiske språket som benyttes betyr, hva slags svar som forventes og hvilken informasjon som er relevant for oppgaven. Den inneholder også en konstruktiv del, som kreves for å kunne formulere et svar slik at det tilfredsstiller oppgavens krav. Oppgaver som inneholder flere informasjonskilder, spesielt hvis man må gå gjennom dem flere ganger for å trekke ut informasjon som er relevant for steget i løsningsprosessen man jobber med i øyeblikket, aktiverer denne kompetansen i høy grad. Det

konstruktive aspektet ved en oppgave øker hvis den krever en detaljert løsningsprosess eller en beskrivelse av arbeidet. Den representerer dermed hovedsakelig språket i matematikkoppgaver, og ikke selve arbeidet med å løse den. Kompetansen inkluderer ikke å kunne tolke forskjellige matematiske representasjoner, som grafer og tabeller, da dette faller inn under representasjonskompetanse.

Strategiutviklingskompetanse

I rammeverket betyr ordet strategi et sett med steg som former en plan for å løse problemet, og strategiutviklingskompetanse er det aspektet ved matematikk som kreves for å *planlegge* hvordan man skal bruke informasjonen man har tilgjengelig for å finne en løsning på oppgaven. Kompetansen beskriver ikke ferdigheten til å lese data ut fra oppgaven, eller algebraferdigheter som kreves for å gjennomføre stegene i planen, disse ferdighetene faller inn under andre kompetanser. Kun kompetansen til å utforme en løsningsstrategi og egenskapen til å overvåke sin egen arbeidsprosess er en del av strategiutviklingskompetanse. Oppgaver som aktiverer denne kompetansen på lavt nivå er oppgaver som innebærer å løse ferdig oppstilte problemer; på høyere nivå kreves det gjerne å bruke forskjellige biter informasjon i flere steg for å lage en komplett løsningsstrategi.

Matematisering

Matematisering har sitt opphav i den gamle kompetansen modellering, som eksisterte i den opprinnelige formen av rammeverket. Matematisering er komponenten av modelleringsprosessen som beskriver handlingen/tankeprosessene som kreves for å oversette mellom et reelt problem og en abstrakt matematisk representasjon. Hvis man ønsker å lage en funksjon som gir summen man må betale for bensin som koster 15 kroner per liter kunne man skrevet dette på følgende måte.

$$Pris(x) = 15x$$

Det er egenskapen til å se koblingen mellom problemet og matematikken som er kjernen i matematiseringskompetansen. Kompetansen inkluderer i tillegg å kunne gjøre rimelige antagelser og å kunne se sammenheng mellom størrelser i det virkelige problemet og den matematiske tilnærmingen. Å kunne tolke de matematiske resultatene fra en oppgave for å kunne si noe om implikasjonen på den ekstra-matematiske situasjonen er også inkludert i denne kompetansen.

Representasjonskompetanse

Både når man jobber med matematikk, og i det daglige liv, kan man støte på mange forskjellige måter å representere størrelser og mengder. Grafer, tabeller, diagrammer, og andre figurer er forskjellige måter som benyttes for å representere tall og størrelser. Representasjonskompetanse innebærer egenskapen ved å kunne arbeide med disse forskjellige representasjonsformene. Å kunne lese verdier ut fra, og å tolke, grafer, å kunne sette opp en tabell for funksjonsverdier i gitte punkter, og å kunne omgjøre f.eks. et sektordiagram til et søylediagram er alle eksempler på representasjonskompetanse. Kompetansen inkluderer ikke å oversette en ekstra-matematisk representasjon til matematisk språk, som ville falt inn under matematisering, eller å tolke skriftlige representasjoner, som ville falt inn under kommunikasjon. Et eksempel på en oppgave som aktiverer denne kompetansen på lavt nivå ville vært å lese av isolerte verdier fra en graf, mens høyere nivåer kan inkludere mer komplekse representasjoner eller behov for å tolke innhold i flere forskjellige typer representasjoner i samme oppgave.

Symbol- operasjon og formalismekompetanse (SOF-kompetanse)

Denne kompetansen er kanskje den som ligger nærmest det de fleste forbinder med skolematematikk. Den beskriver kompetansen som kreves for å kunne anvende formler, definisjoner, algoritmer og prosedyrer. Divisjonsalgoritmen man lærer i grunnskolen er et godt eksempel på denne typen kompetanse; det samme er abc-formelen som generell metode for løsning av annengradsligninger. Å bruke kunnskapen om at vinklene i en trekant summerer til 180 grader, eller at arealet av en sirkel kan uttrykkes som $\pi \cdot radius^2$ er også eksempler på denne kompetansen. Kompetansen inkluderer ikke forståelsen bak hvorfor disse definisjonene og faktabitene er sanne, men ferdigheten som ligger i å kunne anvende dem som en del av løsningsprosessen. Den beskriver generelt all regning og algebra som brukes som steg på veien til svaret, men ikke den bakenforliggende forklaringen til hvorfor vi velger stegene vi bruker. Å sette opp en likning som beskriver en ekstra-matematisk situasjon tilhører matematisering, men å løse likningen krever SOF-kompetanse. Nivået av kompetansen en oppgave krever øker generelt med kompleksiteten til utregningene. Å manipulere brøker er eksempelvis mer krevende enn å jobbe med heltall, og å løse oppgaver med annengradslikninger krever gjerne denne kompetansen i høyere grad enn ligninger som kun inneholder førstegradsledd.

Resonneringskompetanse

Den siste kompetansen i rammeverket, og kanskje den vanskeligste å definere, står beskrevet innledningsvis som følgende sitat.

«This competency relates to drawing valid inferences based on the
internal mental processing of mathematical information needed to obtain well-founded results, and to assembling those inferences to justify or, more rigorously, prove a result.» (Turner et al., 2015, s. 114)

Resonering dreier seg altså om å trekke gyldige slutninger ut fra informasjonen man har tilgjengelig og ved å bruke matematiske gyldige argumenter. Det som gjør kompetansen vanskelig å identifisere i en oppgave er at mange komponenter av dette faller inn under de andre kompetansene, ofte matematisering eller strategiutvikling, og i eksemplene som gis i Turner et al. (2015) blir denne kompetansen ofte vurdert til å ikke være nødvendig av nettopp denne grunnen. Det virker derfor som om den ender opp med å bli en kompetanse som kun aktiveres når man ikke klarer å knytte en oppgave til noen av de andre kompetansene i rammeverket, og derfor er jeg usikker på hvor nyttig den er i en anvendelse av rammeverket.

Beskrivelsene av kompetansene i Turner et al. (2015) inkluderer også beskrivelser av nivåinndelinger for hver kompetanse, og kan derfor brukes for å beskrive i detalj hvilke kompetanser som er nødvendige, og i hvor stor grad disse må aktiveres, for å løse en oppgave. Studier som har forsøkt å benytte rammeverket for å predikere oppgavers vanskegrad har gitt gode resultater (Pettersen & Braeken, 2019; Turner et al., 2013), og rammeverket er derfor et lovende grunnlag for å kunne si noe om kompetansekravene til matematiske oppgaver.

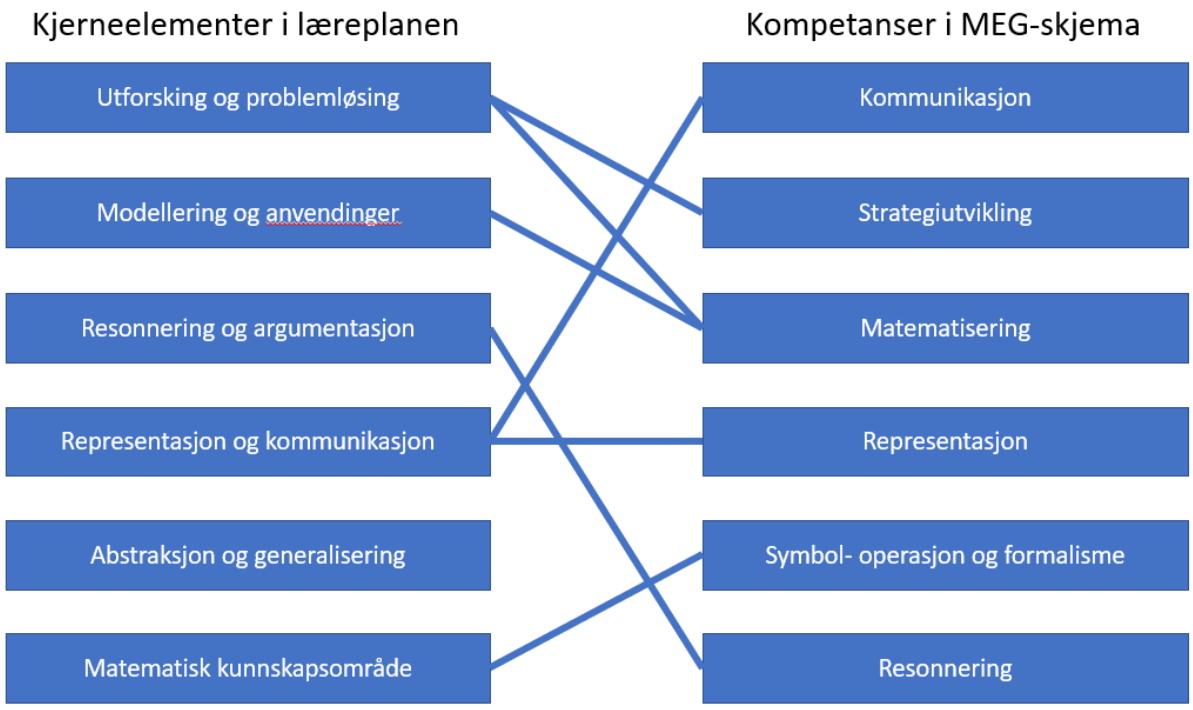
2.2.3 Læreplanens kjerneelementer

Hva som kjennetegner matematisk kompetanse har vært et viktig felt av matematikkdidaktisk forskning de siste tiårene, og det begynner å påvirke matematikkundervisningen i skolen. De nye norske læreplanene begynner også å anerkjenne at undervisningen skal trene og vurdere matematisk kompetanse som et sammenflettet spekter av ferdigheter. Dette synet deles av fagfeltet. Niss & Jensen (2002) poengterer dette i presentasjonen av sin kompetanserose (se Figur 1), og Kilpatrick et al. (2001) deler samme syn i deres rammeverk for matematisk kompetanse. De forskjellige kompetansene burde derfor ikke sees på som separate kompetanser. Matematisk kompetanse burde heller sees på som én overordnet egenskap, der rammeverkene presenterer forskjellige sider av helheten.

I skoleåret 20/21 begynner introduksjonen av de nye læreplanene i grunnskolen og videregående skole i Norge. I tillegg til nye kompetansemål har læreplanen i matematikk introdusert en ny del med navnet kjerneelementer (Kunnskapsdepartementet, 2019). I likhet med de grunnleggende ferdighetene presenterer kjerneelementene en del av matematikkfaget som går utover de mer spesifikke kompetansemålene, og beskriver overordnede kompetanser elevene skal mestre. Det presenteres seks kjerneelementer, som er felles for alle matematikkfagene, og de har følgende navn:

Utforsking og problemløsing, modellering og anvendinger, resonnering og argumentasjon, representasjon og kommunikasjon, abstraksjon og generalisering, og matematiske kunnskapsområder. (Kunnskapsdepartementet, 2019)

Likhetene mellom disse kjerneelementene og aspektene av matematisk kompetanse slik beskrevet i seksjon 2.2.2 er slående, og de tilhørende beskrivelsene av hvert enkelt kjerneelement gjør det mulig å etablere en kobling mellom kjerneelementene og aspektene av matematisk kompetanse i PISA-rammeverket.



Figur 2: Sammenheng mellom PISA-rammeverket og kjerneelementene i den nye læreplanen

Figur 2 viser en sammenheng mellom PISA-rammeverket for matematisk kompetanse og de nye kjerneelementene i læreplanen. Denne sammenhengen og figuren er min tolkning, basert på beskrivelsene i de respektive kildene. Henholdsvis Turner et al. (2015) for PISA og kunnskapsdepartementet (2019) for læreplanens kjerneelementer. Det eneste kjerneelementet i læreplanen som ikke er koblet til et av PISA-rammeverkets seks aspekter er *abstraksjon og generalisering*. Beskrivelsen av dette kjerneelementet er bred nok til at den kan kobles til veldig mange av aspektene ved matematisk kompetanse i MEG-skjemaet, og det blir vanskelig å velge en av dem som mer representativ enn de andre.

Grunnen til at kjerneelementene tas opp i denne oppgaven er for å danne en sammenheng mellom fagfornyelsen og eksamen slik den fungerer i dag. Eksamen påvirker hva som undervises i skolen, og hva elever bruker tid på å lære seg. Hamilton, Stecher, Marsh, McCombs, and Robyn (2007) fant at nesten halvparten av lærerne i deres studie brukte tid på å undervise prøvestrategier og hvordan man kunne løse spesielle typer eksamensoppgaver, og over halvparten fokuserte mer på temaer og emner som de visste ofte dukket opp på prøvene. Selv om deres funn også indikerte at lærerne i studien i større grad bruker læreplanen som retningslinje enn de brukte standardiserte tester, var det spesielt i matematikkundervisningen at standardiserte tester hadde innflytelse på undervisningen. At vurderingsformer kan påvirke undervisningen burde ikke komme som en overraskelse på noen, og Nortvedt & Buchholtz (2018) poengterer at innflytelsen vurderinger har på undervisning kan brukes av utdanningsmyndigheter ved å utnytte at endringer i vurderingsformer kan påvirke undervisning og utdanningspraksis. Hvis de nye kjerneelementene i fagfornyelsen skal være en del av de ferdigheter og kompetanser elever skal tilegne seg i skolen er det derfor viktig at de representeres på eksamen og andre vurderinger. I tillegg til dette kan vi argumentere fra et validitetssynspunkt. Hensikten bak eksamen er å teste kompetanse i faget som testes, og om det kan demonstreres at eksamen ikke tilstrekkelig dekker læreplanen har dette negative implikasjoner for testens innholdsvaliditet.

Rammeverket fra Turner et al. (2015) har nok felles trekk og tilsvarende beskrivelser til å kunne benyttes som en representasjon av de nye kjerneelementene i læreplanen. Det er derfor et hensiktsmessig rammeverk å benytte for å analysere om oppgavene fra eksamen er en god implementering av kjerneelementene med tanke på anbefalinger og diskusjoner rundt utvikling av eksamensoppgaver i fremtiden.

2.3 Validitet

Denne oppgaven er i hovedsak en undersøkelse av forskjellige aspekter ved validiteten til eksamen i MAT0010 gitt våren 2019. For å kunne tolke resultatene av analysen er det derfor viktig å ha et bilde av hva validitet er. Når jeg diskuterer validitet i dette kapittelet mener jeg validitet i forbindelse med måling i utdanningsforskning, og spesielt i kontekst av tester og prøver.

Validitet er et kritisk konsept i alt av måling, testing og undersøkelser, men samtidig et konsept som er ytterst komplekst og kan være vanskelig å måle (Blömeke, 2013). Definisjoner av begrepet validitet har endret seg gjennom historien, og dette reflekteres i debatten rundt validitet i nyere tid (Blömeke, 2013; Lind Pantzare, 2018). Generelt kan validitet defineres som «sannhetsverdien» til undersøkelsen eller målingen, og tidlige definisjoner av validitet i vår sammenheng (måling i utdanning) er knyttet til i hvilken grad en test måler det den forsøker å måle (Newton & Shaw, 2014). Dette er et godt grunnlag for hva begrepet validitet betyr, men det leder til et uunngåelig oppfølgende spørsmål. Hvordan vet man at en test måler det den forsøker å måle? I ekstreme tilfeller kan man lett se at validiteten til en test er lav. Hvis en norsk klasse fikk utlevert en matematikkprøve skrevet på gresk er det ikke vanskelig å forstå at resultatene ikke er et godt mål på deres faktiske matematikkferdigheter. Hvis vi ser bort fra slike urealistiske situasjoner er det derimot vanskeligere å si noe om en prøves validitet uten en grundigere undersøkelse. Burkhardt & Schoenfeld (2018) presiserer at tester i utdanning ikke er presisjonsinstrumenter, et syn som deles av Popham (1997), og sier at dette er noe som ignoreres av testkjøpere av politiske årsaker. At en eksamen i matematikk, eller et hvilket som helst annet fag, har høy validitet er altså ikke en antagelse vi kan gjøre uten nærmere undersøkelser.

Som nevnt over er validitet et komplekst begrep som lenge har blitt diskutert i akademia, men selv om det ikke er total enighet om begrepets definisjon finnes det nødvendige kriterier for at en test skal kunne sies å være valid. *Objektivitet* er et av disse kriteriene som har vært anerkjent lenge som viktig for testvaliditet. Newton & Shaw (2014) skriver om hvordan utdanningsforskeren Walter Monroe i 1923 noterte hvordan eksaminator var en viktig kilde til forskjell mellom testresultater og faktisk evne til testkandidaten. Objektivitet beskrives av Blömeke (2013) som uavhengighet mellom testresultatene og generelle forhold tilknyttet testen. Dette betyr at objektivitet er et mål på i hvor stor grad testresultatene påvirkes av den som vurderer prøven, på forhold i prøvelokalet eller andre omstendigheter. Hvis slike ytre forholds innflytelse er betydelige, er det lett å se hvordan resultatene fra en prøve blir mindre anvendelige.

Et annet nødvendig kriterium for validitet i måling er *reliabilitet*. Reliabilitet og validitet er begreper som er nært knyttet hverandre, men med en nyansert forskjell i betydning. Reliabilitet sier noe om

presisjonen, ikke sannhetsverdien, til målingen ifølge Newton and Shaw (2014). De presiserer at reliabilitet handler om i hvor stor grad en måling vil gi samme resultat hvor gang den gjennomføres. Hvis du veier samme gjenstand fem ganger og vekten viser fem svært forskjellige tall kan du ikke stole på måleinstrumentet ditt. Hvis vekten viser samme tall hver gang er den reliabel, men det er fortsatt mulig at den systematisk viser en annen verdi enn den faktiske vekten og dermed ikke er til å stole på. Reliabilitet er derfor en nødvendig forutsetning for validitet, men er ikke alene tilstrekkelig. I tillegg til kravene objektivitet og reliabilitet har validitet flere aspekter som kan, og bør, undersøkes for å kunne stole på resultatene en prøve gir. Disse aspektene ved validitet vil nå beskrives ytterligere.

2.4 Aspekter av validitet

Vi har nå sett på to nødvendige forutsetninger for validitet, men validitetsbegrepet i seg selv er fortsatt ikke definert. I litteraturen rundt validitet har det vært vanskelig å finne en tydelig definisjon, og Newton and Shaw (2014) sier direkte at det har vært få forsøk på å lage en samlet oversikt, og at litteraturen i hovedsak har vært en samling av argumenter og innsyn i forskjellige aspekter av validitet. Akademikere snakket om flere forskjellige former for validitet, og *The Standards for Educational and Psychological Testing* (heretter kun *Standards*) nevnte tre former for validitet i sin versjon fra 1966, Innholdsvaliditet, Kriterievaliditet, og Konstruktvaliditet (Newton & Shaw, 2014). Disse begrepene brukes fortsatt, og to av disse, Innholdsvaliditet og konstruktvaliditet er nært knyttet målet for denne oppgaven, og vil bli presentert i mer detalj i seksjon 2.4.1 og 0.

2.4.1 Innholdsvaliditet

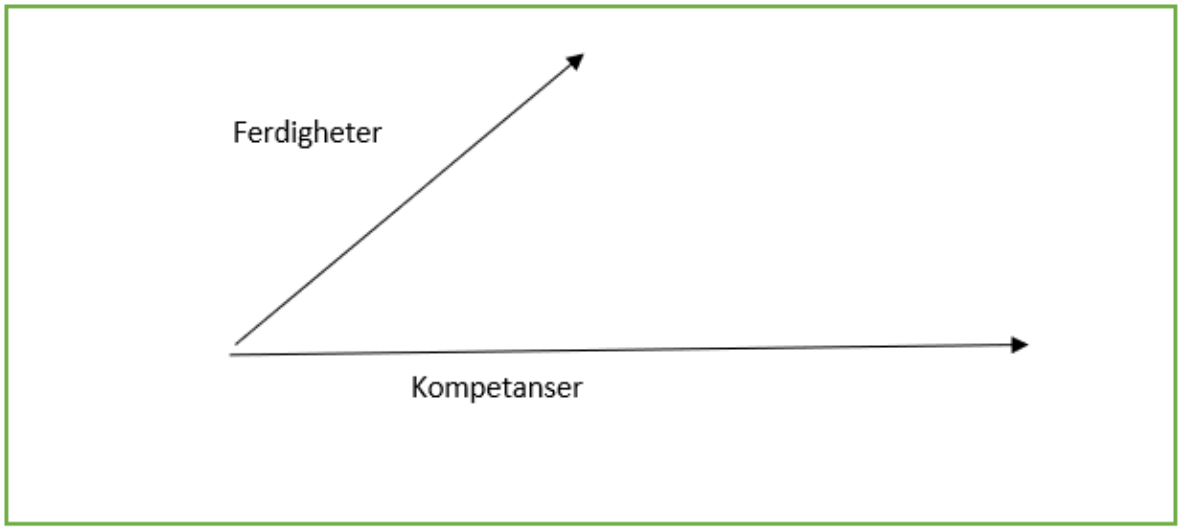
Content validity, eller innholdsvaliditet, er et mål på i hvilken grad innholdet i en test er representativt for konstruktet den forsøker å måle (Blömeke, 2013; Leder & Forgasz, 2018). Konstrukt i denne sammenhengen betyr omtrent det samme som ordet egenskap, og vil diskuteres nærmere i seksjon 0. Sagt med andre ord er innholdsvaliditet et mål på om prøvens innhold er relevant for prøvens formål. Innhold betyr i denne sammenhengen de kunnskapene og ferdighetene hos testtager prøven forsøker å avdekke (Embretson, 2019). Hvis alle oppgaver som potensielt kan dukke opp på en eksamen utgjør det totale universet av eksamensoppgaver, så kan én enkelt eksamen sees på som et utvalg fra dette universet. Innholdsvaliditeten til prøven blir da et mål på hvor godt prøven representerer dette universet av eksamensoppgaver.

Det finnes to hovedkilder til svekkelse av validiteten av innholdet i en test, og begge er viktige å unngå om innholdsvaliditet skal ivaretas. Disse kildene er konstruktunderrepresentasjon og

konstruktirrelevans ifølge Blömeke (2013) og Leder and Forgasz (2018).
Konstruktunderrepresentasjon oppstår når prøvens innhold tester et utvalg av konstruktets omfang som er for snevert; et eksempel ville vært en eksamen i matematikk som kun inneholdt oppgaver om geometri, eller hvis flere deler av læreplanen ikke testes. Resultatene fra en slik prøve mangler for mye informasjon om konstruktet den forsøker å teste til å gi et pålitelig bilde av hva elevene som tar testen faktisk kan. Konstruktirrelevans oppstår når prøvens innhold er irrelevant for konstruktet som testes. Dette kan være vanskeligere å oppdage, men et eksempel kan være matematikkoppgaver som inneholder språk og termer elever ikke forstår. Da er det ikke lenger matematikkompetansen disse oppgavene tester, men kunnskap om andre emner som ikke er en del av faget.

Innholdsaspektet ved validitet er det vanskelig å måle kvantitativt, og *Standards* sier det generelt ikke er mulig å måle ifølge Newton & Shaw (2014). De skriver at validering av innholdet i tester generelt innebærer at testens oppgaver må vurderes av eksperter som har erfaring med konstruktet som testes på prøven og erfaring med metoden som ble benyttet for utvalg av oppgavene. Sagt med litt andre ord må innholdet i matematikkeksamen vurderes av erfarne matematikklærere, matematikere og prøveprodusenter.

Innholdet som testes i eksamen i MAT0010 er ferdigheter i og kunnskaper om matematikk, og defineres av læreplanen for matematikk fellesfag (Kunnskapsdepartementet, 2013). Uten å dykke for dypt inn i læreplanen her kan det nevnes at den generelt deler inn faget i to hovedsegmenter, kompetansemål og grunnleggende ferdigheter. De grunnleggende ferdighetene beskriver generelle ferdigheter som er nødvendige for å arbeide med matematikk; de er, muntlige ferdigheter, å kunne skrive i matematikk, å kunne lese i matematikk, å kunne regne i matematikk, og digitale ferdigheter. Kompetansemålene beskriver mer spesifikke kompetanser elevene skal besitte etter endt utdanning. Et eksempel fra kompetansemålene fra 10. år er «rekne med brøk, utføre divisjon av brøkar og forenkle brøkuttrykk» (Kunnskapsdepartementet, 2013). Dette inndelingen korresponderer godt med beskrivelsen av oppgaver fra Pettersen (2019), som skiller mellom kjennetegn og krav ved matematikkoppgaver. De to dimensjonene er representert i Figur 3 (min figur) på neste side.



Figur 3: Matematikkferdighet i to dimensjoner

Blum (2006) beskriver den overhengende ideen bak matematisk kunnskap på en tilsvarende måte. Blum bruker ordet kompetenzen om de delene av matematikkunnskap som i læreplanen kalles grunnleggende ferdigheter, og veiledende ideer om matematiske områder som tall og måling. Ideen er at matematikkferdighet kan måles i to dimensjoner. Langs den ene dimensjonen finner vi matematiske prosedyrer og den mer instrumentelle delen av matematikk; regneferdigheter, algoritmer for å løse likninger, og formler for utregning av areal og volum er eksempler på denne dimensjonen. Langs den andre dimensjonen finner vi de mer generelle ferdighetene som kreves for å forstå matematikk; problemløsing, modellering, og bruk og håndtering av matematisk språk og formalismer er eksempler på dette. Denne andre dimensjonen av matematikkferdigheter sammenfaller med kompetansene beskrevet i seksjon 2.2.

Niss & Jensen (2002), som er beskrevet tidligere, og Kilpatrick, Swafford & Findell (2001) har begge utviklet teoretiske rammeverk for matematisk kompetanse for å beskrive denne andre dimensjonen til matematikkferdigheter, at det å kunne matematikk består av mer enn å kunne de mer instrumentelle oppskriftene faget benytter. Årsaken til å nevne dette her er for å illustrere at matematikkferdigheter, og dermed innholdet til en eksamen, kan vurderes langs mer enn en dimensjon; det er altså mulig at innholdsvaliditeten til en eksamen kan være svekket selv om den dekker alle kompetansemålene, da kompetansemålene kun svarer til én av dimensjonene som bør testes.

2.4.2 Konstruktvaliditet

Construct validity, eller konstruktvaliditet, er et mål på om en test måler det konstruktet den forsøker å måle (Leder & Forgasz, 2018). Dette er nært knyttet innholdsvaliditet, og denne seksjonen burde leses som en fortsettelse av forrige. Messick (1989) argumenterte for at all validitet bør sees på som aspekter ved konstruktvaliditet, at alle former for validering er med på å validere konstruktet målingen representerer. I kontekst av prøver og tester er et konstrukt et annet ord for en psykologisk egenskap eller kvalitet, og denne egenskapen antas å kunne måles med en test. I motsetning til innholdsvaliditet, som kan vurderes ved å se at en prøve eller oppgave samsvarer med innholdet den skal teste, kan man ikke anta at en prøve direkte representerer konstruktet den skal teste, men at den fungerer som et grunnlag for å antyde et underliggende konstrukt (Newton & Shaw, 2014). Sagt med andre ord, og her skriver jeg min tolkning av deres tekst, kan man ifølge dem ikke gjøre en kvalitativ analyse for å undersøke konstruktvaliditeten til en prøve.

Siden konstruktvaliditet ikke kan måles kvalitativt ved å vurdere en test må det avgjøres empirisk. Et konstrukt, i vårt tilfelle matematisk ferdighet, defineres, og en test konstrueres og antas å være representativ for konstruktet. Teorien bak konstruktet brukes for å danne hypoteser, og testen benyttes for å sette prøve på hypotesene. Resultatene fra testen benyttes til å endre på teorien bak konstruktet, og dermed også fremtidige tester. Validering av konstruktet innebærer dermed også å samtidig validere testen som representerer konstruktet (Newton & Shaw, 2014). Blömeke (2013) beskriver hvordan konstruktvaliditet testes gjennom å finne empiriske bevis for konstruktets eksistens. Hvis konstruktet som testes f. eks. antas å være endimensjonalt, med andre ord at alle oppgavene på testen representerer samme underliggende egenskap, burde dette reflekteres i en empirisk undersøkelse av testresultatene. En annen måte å validere konstruktet er å bruke beskrivelsene av konstruktet til å forutsi vanskegrad til testens oppgaver, slik som Turner et al. (2013) gjorde for oppgaver fra PISA 2012. Hvis konstruktet har beskrivelser av hva som kreves på høyt og lavt nivå bør dette reflekteres i en empirisk undersøkelse av vanskegraden til enkeltoppgaver.

Læreplanen har ikke en detaljert definisjon av konstruktet matematikkferdigheter. De grunnleggende ferdighetene «Å kunne regne i matematikk», «Å kunne lese i matematikk», «Å kunne skrive i matematikk», «Muntlige ferdigheter», og «Digitale ferdigheter» (Kunnskapsdepartementet, 2013) kan som nevnt i forrige seksjon benyttes sammen med kompetansemålene som grunnlag for et konstrukt, men de beskrives generelt og uten tydelige definisjoner av hva som kreves for forskjellige nivå. Det finnes rammeverk for matematisk kompetanse som inneholder nivådelte beskrivelser, og derfor er egnet for å undersøke innholdet i en matematikkeksamen. Et av disse er rammeverket utviklet av PISAs Mathematical Expert Group (MEG), som ble presentert i seksjon

2.2.2. Dette rammeverket er basert på arbeidet til Niss & Jensen (2002), og er en operasjonalisering av de grunnleggende ferdighetene, og spesielt de nye kjerneelementene, i læreplanen.

2.4.3 Validitet fra et konsekvensperspektiv

Validitetsdebatten endret seg mot slutten av 80-tallet med arbeidet til Samuel Messick. På denne tiden var det ifølge Newton and Shaw (2014) vanlig å skille mellom validitet for måling og validitet for prediksjon. De forteller at Messick argumenterte for at validering ikke kan splittes på denne måten, og at senter for validitetsfokuset er måling; uten en sikker forståelse av hva som måles kan ikke prediksjoner forsvares. Messick (1989) argumenterer for at all validitet er konstruktvaliditet, og at alle andre former for validitet bidrar til konstruktvalidering. Han brakte også opp et nytt perspektiv i validitetsdebatten som fokuserte mer på konsekvenser av, og formålet bak, tester.

De fleste prøver har en eller annen form for påvirkningskraft eller innflytelse på personen som tar testen sitt liv, og de har også gjerne et spesifikt formål. En eksamen forsøker å si noe om en kandidats kompetanse i faget som testes, og en god karakter på eksamen øker dine sjanser for å tas opp til et ønsket universitetsstudium eller annen høyere utdanning; hvis du stryker på vegvesenets teoriprøve får du ikke lov til å kjøre bil. Dette er testenes formål, og derfor er det naturlig at det er slik de fungerer. Her bør leser huske at tester ikke er perfekte måleinstrumenter, og at en prøve dermed kan ha negative konsekvenser for testtageren som ikke er fortjente og dermed uønskede. Messick satt søkelyset på nettopp denne konsekvensbiten av testvalidering, og mente at validitet handler om i hvor stor grad konsekvenser og bruk av testresultater kan forsvares. Flere begynte å snakke om et nytt aspekt ved validitet, «consequential validity», men dette ble raskt et omdiskutert begrep. Blömeke (2013) påpeker at konsekvenser er noe som vanskelig lar seg teste, og ikke passer inn i konstruktdefinisjoner. Et argument som benyttes av Popham (1997) er at hva en test benyttes til i prinsipp er uavhengig av testen. En myndighet kan teoretisk velge å bruke en synstest som inntakskrav for sitt medisinstudium. Dette betyr ikke at synstesten i seg selv er problematisk og ikke kan stoles på, bare at den benyttes til et upassende formål. At konsekvenser av en test skal spille inn i validiteten til testen er derfor noe mange har bestridet (Leder & Forgasz, 2018), og (Popham, 1997, s.13) skriver om de som forsøker å knytte konsekvens til validitet med følgende ord «Their mistake, I believe, is in trying to tie social consequences into a validity framework. Such a wedding of related but distinctive concepts will not be symbiotic, it will be septic. »

Allikevel kunne man ikke bestride at konsekvenser av slutninger som tas på bakgrunn av testresultater var nært knyttet testvaliditet, og at forholdet mellom validitet og konsekvenser ikke kunne skilles. Arbeidet til Messick ble videreført av Kane (2006) som argumenterte for at validitet

handlet om å ha tilstrekkelig empirisk evidens og teoretisk grunnlag for å forsvare slutninger testresultater fører til. Konsekvenser av testresultater blir da ikke sett på som en del av validitetsbegrepet, men heller formålet til valideringsprosessen. Arbeidet med testvalidering er hele veien gjennomsyret av det overhengende målet om å sikre at vi kan ta godt begrunnede slutninger som har et solid fundament i testteoretisk grunnlag og konstruktet prøven skal teste. Uønskede konsekvenser av hvordan prøveresultater benyttes er dermed av interesse i en valideringsundersøkelse, og det vil bli relevant for denne oppgaven. Av spesiell interesse er muligheten av at prøven kan ramme forskjellige grupper av elevmassen ulikt.

2.4.4 En avslutning av validitetspresentasjonen

Flere bøker har blitt skrevet om alle de forskjellige sidene og aspektene til validitet som benyttes i forsknings- og målingssammenheng, og en total gjennomgang av all litteratur er utenfor rekkevidde for denne oppgaven. Presentasjonen av teori bak validitet er gjort på bakgrunn av hva som er mest relevant for undersøkelsene som gjøres i denne studien, innhold- og konstruktvaliditet. En gjennomgang av det nyere konsekvensfokuserte synet på validering var også nødvendig, da denne oppgaven ønsker å undersøke om det er bias i eksamensoppgavene.

Årsaken til valget av disse tre aspektene er hovedsakelig at jeg mener de er relevante for å kunne si noe om validiteten til eksamen som helhet, og at det er disse validitetsaspektene det er mulig å si noe om med datamaterialet som er tilgjengelig for denne oppgaven. Som en avslutning, og oppsummerende retningslinje for begrepet validitet vil jeg nevne definisjonen fra *Standards in educational and psychological testing*;

"Validity refers to the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests." (AERA, APA, & NCME, 2014, s. 11).

Denne beskrivelsen tar den opprinnelige tankegangen bak validitet, at validitet handler om i hvor stor grad en test måler det den forsøker å måle, men setter søkelys på problemstillingen rundt konsekvenser. Hvis vi ikke har tilstrekkelig forståelse av hva en test måler bør den ikke benyttes som slutningsgrunnlag for valg som kan ha store konsekvenser for en persons liv og fremtid. I neste seksjon vil dette belyses ytterligere.

2.5 Bias og equity i prøver

En prøve har som formål å si noe om egenskapene hos en person som besvarer prøven. Hvis testtakeren har høy kompetanse i faget vil vedkommende ideelt få en høy score på prøven, mens en person med lavere kompetanse vil få en lavere score. En rettferdig prøve er en prøve som fungerer på denne måten; en kandidat som besvarer en rettferdig prøve vil få en score, eller et resultat, som er avhengig kun av kandidatens kompetanse i det som testes. Prinsippet bak en rettferdig prøve kan videreføres til de individuelle oppgavene prøven består av. En rettferdig oppgave bør fungere slik at en person med høy kompetanse vil ha høyere sannsynlighet for å besvare oppgaven korrekt, mens en med lavere kompetanse bør klare dette med lavere sannsynlighet. Av og til inneholder tester allikevel oppgaver som fungerer annerledes, oppgaver hvor to personer med lik ferdighet har ulik forutsetning for å klare oppgaven. Tester, eller oppgaver, som fungerer på denne måten utviser bias. Begrepene bias og rettferdig prøve defineres nå i henholdsvis punkt 1) og 2).

1) En prøve utviser bias hvis to kandidater med lik ferdighet har ulik forutsetning for å lykkes.
2) En rettferdig prøve er en prøve som ikke utviser bias.

Det er viktig å presisere at bias ikke er et resultat av tilfeldigheter, og det kan ikke konkluderes at en prøve utviser bias kun fordi én gruppe har gjort det bedre på prøven enn en annen (Camilli & Shepard, 1994). Reynolds, Livingston, Willson, and Willson (2010) beskriver bias som et fenomen som oppstår når en test systematisk over- eller undervurderer ferdigheten den forsøker å måle. Et eksempel er en prøve som forsøker å finne ut noe om engelskferdigheter. Hvis en slik prøve gis til to grupper, en gruppe elever med engelsk som morsmål, og en gruppe med engelsk som andrespråk, kan man forvente at den engelskspråklige gruppen systematisk gjør det bedre. Det betyr allikevel ikke at prøven utviser bias, nettopp fordi det er engelskferdigheter man prøver å måle; og man kan forvente at en gruppe elever med et annet morsmål vil ha lavere forutsetninger for å lykkes i en slik prøve. Hvis man derimot forsøker å måle intelligensen til to grupper med forskjellig morsmål, og en gruppe systematisk scorer høyere enn den andre, er det grunn til å mistenke at språket i prøven fører til bias; eksempelet med språklig bias i IQ-tester er velkjent i historisk sammenheng, og en viktig påminnelse om hvorfor bias er viktig å være på utkikk etter i utdanning og testsammenheng (Camilli & Shepard, 1994).

Bias i en prøve kan ha flere årsaker, og kan ofte være vanskelige å forutse. Som nevnt er språk en mulig kilde til bias. Kultur, som ofte er nært knyttet språk, kan også føre til bias (Banks, 2006). En annen mulig kilde kan være prøvens format. Leder and Forgasz (2018) skriver om hvordan det var systematisk kjønnsforskjell i matematikkeksamen for avgangselever i matematikk basert på formatet prøven var laget i. Alle avgangsfagene i matematikk de undersøkte benyttet tre separate

vurderingssituasjoner. En hjemmeeksamen som gikk over flere uker, en tradisjonell prøve med flervalgs- og kortsvaroppgaver, og en tradisjonell prøve som krevde mer utfyllende svar. Gutter gjorde det systematisk bedre i de to sistnevnte vurderingssituasjonene i faget «Mathematical Methods», mens jenter gjorde det bedre enn gutter på den første vurderingsformen. Prøveangst har også blitt nevnt som en kilde til bias i prøver (Meijer & Coping, 2001). De argumenterer for at tradisjonelle skoleeksamener favoriserer en viss type personlighet, og at elever som opplever mye angst vil gjøre det dårligere i denne typen vurderingssammenheng. Med så mange mulige kilder kan det ofte være vanskelig å si noe om årsakene til at en prøve utviser bias, men selv om man ikke kan fastslå årsakene bak bias ligger det verdi i å kunne identifisere at bias oppstår. Hvis man vet at en prøve ikke er rettferdig kan man ta hensyn til dette når resultatene skal behandles eller benyttes for et formål, for eksempel inntak til et lærested eller som grunn for ansettelse til en stilling. Bias er derfor nært knyttet valideringsprosessen diskutert tidligere, og utgjør en direkte trussel til formålet med prøver i skolen.

Her er det verdt å introdusere begrepet equity til diskusjonen. Rousseau and Tate (2003) beskriver to forskjellige syn på rettferdighet, forfatterne bruker ordet «equality», i sin artikkel. Det ene synet setter søkelys på rettferdighet som en prosess, og argumenterer for at rettferdighet i hovedsak handler om å behandle elever likt. Det andre synet ser på rettferdighet i lys av utfall, og mener dermed at lik behandling ikke er rettferdig hvis utfallet av det blir forskjellig for enkelte grupper. Dette andre synet på rettferdighet omtales som equity. Equity handler også om rettferdighet, men fokuserer mer på at alle skal ha lik mulighet til å lære, og det medfører nødvendigheten av å anerkjenne at elever kan ha behov for forskjellig oppfølging og behandling ifølge Nortvedt & Buchholtz (2018). De skriver at equity i vurderingssammenheng betyr at alle elever skal ha mulighet til å demonstrere sin kompetanse i faget som testes, og equity blir derfor truet når bias oppstår i vurderingssituasjoner.

For eksamen i matematikk er det primært to typer bias jeg mener er interessante å identifisere i prøven. Det ene er bias som oppstår på grunnlag av elevens kjønn. Hvis eksamen systematisk favoriserer ett kjønn fører dette til redusert opptaksgrunnlag til videregående og høyere utdanning for halve populasjonen, og påvirker validiteten til prøven direkte. Kjønnsforskjeller i matematikkprestasjoner er et veldokumentert fenomen. Nortvedt (2013) forteller at gutter systematisk gjør det bedre enn jenter på norske nasjonale prøver i matematikk (5., 8. og 9. trinn), mens jenter gjør det bedre på skriftlige eksamener ved 10. trinn og på videregående. I siste PISA-undersøkelse (2018) var det også signifikant kjønnsforskjell mellom gutter og jenter i Norge, i jentenes favør (OECD., 2019). At det finnes prestasjonsforskjeller mellom kjønn er ikke i seg selv bevis for bias, men det bør sees på som en indikator for at bias kan være tilstedeværende.

Den andre kilden til bias som kan ha stor påvirkning, og derfor er av interesse å identifisere, er bias på grunnlag av morsmål. Det er mange elever i Norge som har norsk som andrespråk, og som ikke prater norsk i hjemmet, og dette kan påvirke deres forutsetninger for mestring på eksamen i matematikk. Reynolds et al. (2010), som skriver om utdanningsmiljøet i USA, forteller hvordan tester som har blitt benyttet for å måle egnethet for utdanning ofte bruker vokabular og kulturelle referanser som er tilpasset elever fra hovedsakelig hvite middelklassefamilier. Slike prøver, som en gruppe av elevstanden er dårligere rustet for å forstå, kan dermed indikere forskjeller i ferdigheter som ikke er realistiske, og kan være en betydelig kilde til bias.

3 Forskningsspørsmål

Nå som vi har viktige teoretiske begreper på plass er det på tide å presentere forskningsspørsmålene som vil styre veien videre. Den overhengende motivasjonen bak denne oppgaven er å undersøke validiteten til eksamen i MAT0010 våren 2019. Vi står midt i en fornyelse av læreplanen, og kjerneelementene er på vei inn i matematikkundervisningen, er det både viktig og ønskelig at vurderingsformer endres for å være representative for de nye målene til læreplanen.

Vurderingsformer må tilfredsstille kriteriene for reliabilitet, objektivitet og validitet. Da reliabilitet handler om nøyaktigheten til resultater, og objektivitet tar for seg at resultater skal være uavhengige av personen som gjennomfører målingen (i dette tilfelle sensor som retter en eksamen), er begge disse aspekter av validitet. Hvis objektivitet eller reliabilitet ikke er til stede kan man heller ikke si at resultatene er valide, og de bør ikke brukes for å trekke pålitelige slutninger fra målingen.

Kan det finnes evidens for å si at resultatene fra eksamen i MAT0010 våren 2019 er valide? Denne setningen brukes som overstyrende problemstilling for oppgaven, men det er en for omfattende problemstilling for å kunne besvares direkte. For å undersøke i hvilken grad eksamen gir valide resultater om en elevs matematiske ferdighet ble den overstyrende problemstillingen brutt ned i tre mindre forskningsspørsmål.

1) I hvilken grad representerer oppgavene på eksamen i MAT0010 våren 2019 kompetansekonstruktet som presenteres i kjerneelementene og MEG-skjemaet fra Turner et al. (2015)?
2) Er konstruktet «matematikkferdigheter» som eksamen tester et endimensjonalt konstrukt?
3) Eksisterer det kjønnsbasert bias i oppgavene gitt på eksamen i MAT0010 våren 2019?

Det første forskningsspørsmålet undersøker validitet fra perspektivet til eksamens innhold. Både de grunnleggende ferdighetene og de nye kjerneelementene har mye til felles med funn fra matematikkdidaktisk forskning rundt matematisk kompetanse, og bør derfor være representert på eksamen.

Det andre spørsmålet undersøker dimensjonaliteten til konstruktet «matematikkferdigheter». En eksamen i tråd med de nye kjerneelementene bør representere disse forskjellige dimensjonene, og hvis konstruktet eksamen tester viser seg å være endimensjonalt kan dette være et tegn på at noen av kompetansene dominerer i vurderingsformen.

Relevansen til *forskningsspørsmål tre* burde være åpenbar. Mangel av bias er et direkte krav for at man trekker gyldige slutninger fra eksamensresultater, og derfor for at eksamen skal kunne sies å være valid.

Da forskningsspørsmålene er relativt forskjellige, og dermed vil kreve forskjellige metoder for å finne gode svar, er studien delt opp i tre forskjellige deler som tar for seg hvert sitt forskningsspørsmål. I seksjon 2.4.1 og 0 ble det fortalt om hvordan innholdsvaliditet undersøkes kvalitativt ved å vurdere oppgavene på prøven mens konstruktvaliditet må undersøkes empirisk med kvantitative metoder. Selve metoden bak undersøkelsene vil bli beskrevet i påfølgende seksjon. Datamaterialet som benyttes er oppgaven gitt ved skriftlig eksamen i MAT0010 våren 2019 og et datasett med responsmønstre fra 3035 elever som tok eksamen.

4 Metode

For å besvare forskningsspørsmålene ble det i løpet av denne oppgaven utført tre separate analyser av datamaterialet. Dette fordi forskningsspørsmålene dekker et bredt felt av validitetsundersøkelser, og kan ikke lett besvares med én undersøkelse.

For å kunne vurdere innholdsvaliditeten til eksamen, med andre ord å undersøke om innholdet eksamen tester er representativ for innholdet i læreplanen, må oppgavene undersøkes individuelt. Som nevnt i seksjon 2.4.1 er det generelt ikke mulig å kvantitativt måle innholdsvaliditet, og innholdet i en prøve må avgjøres av eksperter med kjennskap til konstruktet prøven skal teste. Den beste måten å vurdere innholdet til prøven ble vurdert til å være en kvalitativ undersøkelse av oppgavene fra eksamen, og presenteres i mer detalj i seksjon 4.1.

Datagrunnlaget for resten av analysen er et sett med responsmønstre fra 3035 elever som besvarte eksamen våren 2019. Responsmønstrene er i formen poeng per deloppgave. Hver rad representerer en elev som tok eksamen. I cellene for denne elevens rad kan man lese hvor mange poeng denne eleven ble tildelt for sin besvarelse på hver enkelt deloppgave fra eksamen. For de fleste oppgaver vil dette være et tall i intervallet 0-2. Hvis en celle står markert som 9 betyr dette at oppgaven ikke er besvart. I tillegg til responsmønsteret inneholder datasettet en kolonne med verdien G eller J for å indikere om eleven er en gutt eller jente. Som et illustrerende eksempel bes leser se Tabell 1.

	Kjønn	Oppg.1	Oppg.2	Oppg.3a	Oppg.3b	Oppg.4a	Oppg.4b	Oppg.5	Oppg.6
Elev 1	G	1	1	1	1	1	1	1	9
Elev 2	G	1	1	1	0	1	0	1	1
Elev 3	J	0	1	1	0	2	0	1	0

Tabell 1: Eksempel på responsmønster

Med et slikt datagrunnlag er kvantitative metoder best egnet for å si noe om oppgavesettet, og statistiske metoder basert på Item Response Theory blir ofte sett på som det mest hardføre rammeverket for denne typer analyser. For å besvare forskningsspørsmålene ble det utført to analyser med basis i dette paradigmet. Én endimensjonal IRT-analyse for å vurdere dimensjonaliteten til konstruktet i prøven, og *en DIF-analyse* basert på kjønn. Begge analysene benytter samme datasett med responsmønstre, og presenteres i seksjonene 4.2 og 4.3. Bakgrunnen for valget av disse metodene springer ut av forskningsspørsmålene. Ønsket er å undersøke

konstruktvaliditeten til eksamen, og å undersøke om eksamensresultatene kan ha uønskede konsekvenser for elevene. For å undersøke dette må kvantitative metoder benyttes for at resultatene skal ha tilstrekkelig reliabilitet.

Metoden bak disse undersøkelsene er i mine øyne den beste måten å finne svar på forskningsspørsmålene. Relevant teori sier at innhold og konstrukt krever henholdsvis kvalitativ og kvantitativ undersøkelse. For å undersøke mulig bias i oppgavene vil jeg også argumentere for at en kvantitativ tilnærming er best tilpasset. Bias er nært knyttet konsekvensaspektet ved validering, og hvis man vil argumentere for at eksamen har, eller ikke har, negative konsekvenser som følge av bias må dette kunne sies generelt. Validering er et såpass stort felt at flere forskjellige metoder må benyttes for undersøkelsen. Når både oppgaver og responsmønstre skal undersøkes krever dette forskjellige angrepsmetoder.

4.1 Koding av matematisk kompetanse

For å si noe om innholdsvaliditeten til eksamen ble det utført en koding av kompetansekrav for eksamensoppgavene. Rammeverket utviklet at MEG-gruppen, og beskrevet i Turner et al. (2015), ble benyttet som grunnlag for denne analysen. Rammeverket, lagt ved som vedlegg B beskriver de seks kompetansene som ble presentert i seksjon 2.2.2. For hver kompetanse er det beskrevet fire mulige nivåer for kompetanseaktivering. Som illustrasjon kan beskrivelsen av nivåene for kompetansen matematisering sees under i Figur 4.

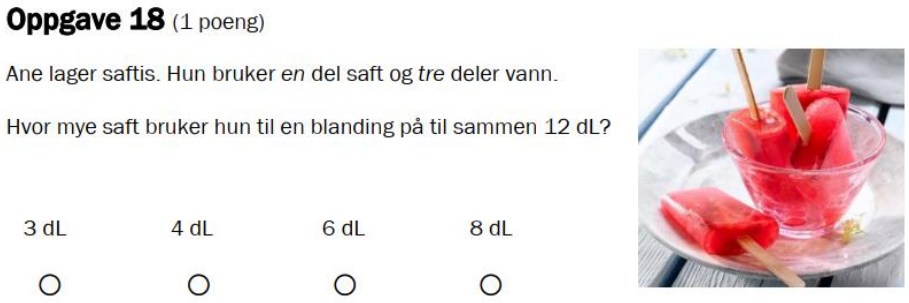
Definition: **Translating** an extra-mathematical situation into a mathematical model, **interpreting** outcomes from using a model in relation to the problem situation, or **validating** the adequacy of the model in relation to the problem situation.
0: Either the situation is purely intra-mathematical, or the relationship between the extra-mathematical situation and the model is not relevant to solving the problem 1: Construct a model where the required assumptions, variables, relationships and constraints are given; or draw conclusions about the situation directly from a given model or from the mathematical results 2: Construct a model where the required assumptions, variables, relationships and constraints can be readily identified; or modify a given model to satisfy changed conditions; or interpret a model or mathematical results where consideration of the problem situation is essential 3: Construct a model in a situation where the assumptions, variables, relationships and constraints need to be defined; or validate or evaluate models in relation to the problem situation; or link or compare different models

Figur 4: Kompetansenivåer for matematisering (Turner et al., 2015, s.112)

En høyere vurdering betyr altså at kompetansen vurderes som nødvendig på et høyere nivå. For kodingsprosessen ble nivå 1, 2, og 3 slått sammen, mens nivå 0 ble stående som beskrevet i Turner et al. (2015). I vurderingen av oppgavene har det altså blitt undersøkt om den relevante kompetansen er nødvendig eller ikke, og for hver oppgave ble alle seks kompetanser i rammeverket vurdert. Valget bak å kombinere de tre høyeste nivåene til én kode ble gjort primært for å gjøre

kodingsprosessen enklere og mindre tidskrevende, da den bare er én av flere undersøkelser. Det var tidvis vanskelig å avgjøre hvilket av rammeverkets nivåer som var mest representativt for en oppgave, og det ble derfor valgt å kun vurdere hver kompetanse som tilstedeværende eller ikke tilstedeværende. I samarbeid med veileder ble det vurdert å kombinere de to laveste nivåene (0 og 1) til en ny kategori 0, og de to høyeste nivåene (2 og 3) til en ny kategori 1. Etter en diskusjon ble det avgjort at det første av alternativene var det meste passende. Den viktigste årsaken bak dette valget var en enighet om at få oppgaver kvalifiserte til kategori 2 og 3 i flertallet av kategorier. Hvis oppgaver med kompetansekrav 1 skulle kodes som 0, at kompetansen ikke er nødvendig, ville de fleste oppgaver blitt vurdert til kompetansekrav 0 i nesten alle kategorier; og den totale vurderingen av eksamen ville blitt at ingen kompetanser er nødvendige.

Det er viktig å understreke at denne prosessen ble gjennomført for hver av eksamensettets 54 deloppgaver, og ikke for hver «hovedoppgave». Det er to viktige årsaker til dette. Selv om de forskjellige deloppgavene under en eksamensoppgave ofte omhandler samme tema, og derfor ofte er satt i samme kontekst, kan de omhandle forskjellige kompetansemål i læreplanen og også kreve forskjellige kompetanser. For en innholdsvurdering av eksamen er det derfor viktig å se på hver enkelt deloppgave for seg selv. I tillegg til dette er det viktig når resultatene fra analysen skal sees i sammenheng med den kvantitative analysen som beskrives i seksjon 4.2 og 4.3. Kodingsprosessen ble til å begynne med utført både av forfatter og av veileder separat; uten noen form for felles forberedelse, men med samme veiledende skjema for vurderingen. Totalt har altså begge kodere vurdert 54 oppgaver for seks forskjellige kompetanser. Til illustrasjon vil kodingsprosessen vises med en eksempeloppgave fra eksamen. Oppgaven kan sees i Figur 5, og er fra eksamen i MAT0010 våren 2019.



Figur 5: Oppgave 18, del 1 (Utdanningsdirektoratet, 2019d)

Oppgaven inneholder lite tekst. Ingen unødvendig informasjon er inkludert i oppgaveteksten (navnet på personen som lager is er kun med for å skape en kontekst rundt oppgaven). Oppgaven skal kun besvares ved å krysse av i riktig rute. Kommunikasjon vurderes dermed til nivå 0. En løsningsstrategi må utvikles. Den mest naturlige strategien er å se at blandingen består av totalt 1+3 = 4 deler væske,

og saften utgjør én av disse fire delene. Dette er en strategi som ikke er eksplisitt nevnt eller åpenbar, og strategiutvikling vurderes dermed til nivå 1. Den oppgitte situasjonen er en modell for blanding av saft hvor all informasjon er gitt. Å løse oppgaven krever at eleven kan tolke denne modellen, og matematisering ble vurdert til nivå 1. Her kan man argumentere for at koblingen mellom oppgaven og den ekstramatematiske situasjonen ikke er relevant nok til at nivå 1 er fortjent, så oppgaven kan også tolkes til nivå 0. Oppgaven inneholder ingen representasjoner og alle tall man regner med er små og relativt enkle å forholde seg til; representasjons- og SOF-kompetanse blir dermed begge vurdert til nivå 0. Oppgaven krever ingen resonnering som ikke dekkes av matematisering og strategiutvikling. Resonneringskompetanse vurderes dermed også til nivå 0.

For å kvalitetssikre prosessen ble verdien for Cohen's Kappa (heretter kun Kappa) regnet ut. Kappa er et mål som brukes for å si noe om enigheten til to kodere, og tar hensyn til muligheten for at koderne kan være enige eller uenige ved tilfeldighet. Kappa har alltid en verdi i intervallet fra -1 til 1, hvor en verdi på 0 er en indikasjon på at koderne ikke er mer enige enn man kan forvente ved tilfeldigheter. Alle verdier over 0 er indikasjoner på at koderne er i en viss grad av enighet, spørsmålet er hvor enige man bør være for å kunne si at analysen holder god nok kvalitet. Bakeman, McArthur, Quera, and Robinson (1997) konkluderte med at det ikke finnes noen universell verdi Kappa burde overskride som kan anvendes i alle situasjoner, så det kan være vanskelig å si nøyaktig hvor høy den burde være for at kodingen av kompetanser skal være pålitelig. Analysen av kodingen for de seks matematiske kompetansene i eksamenens oppgaver (én Kappa ble regnet ut ved hjelp av SPSS for hver kompetanse) viste at våre rangeringer av kompetanse ga en Kappa som hadde verdier i intervallet mellom 0 og 0.3. Disse verdiene for Kappa indikerer en viss grad av enighet, men er ikke høye nok til å kvalitetssikre kodingsprosessen. Landis and Koch (1977) publiserte retningslinjer for verdier av Kappa, og deres vurdering indikerer at resultatene fra kodingen av kompetanser faller i kategoriene «slight agreement» og «fair agreement». Disse refererer til verdier av Kappa hhv. i intervallene (0, 0.2) og (0.2, 0.4), men inndelingen av kategoriene er kun basert på deres egne vurderinger. Min vurdering er at verdiene fra kodingen av kompetanser ikke er høy nok til å indikere enighet, og at den primære årsaken er at vi har vært uenige i hvordan rammeverket skulle tolkes.

På bakgrunn av denne uenigheten ble det utført en ny koding i fellesskap, hvor alle oppgavene ble gått gjennom og punkter hvor vi var uenige ble diskutert. Mange av årsakene bak den opprinnelige uenigheten ble også tydelig i løpet av denne prosessen. For flere av kompetansene var det ikke lett å se et tydelig skille mellom flere av nivåene, og det var derfor flere oppgaver hvor vi i diskusjonen fant ut at vi var enige i hva oppgaven krevde, men allikevel hadde gitt forskjellige koder. Som eksempel er nivå 0 for strategiutvikling beskrevet som oppgaver hvor fremgangsmåten er eksplisitt skrevet eller åpenbar. Om løsningsstrategien bak en oppgave er åpenbar er i stor grad subjektivt, og

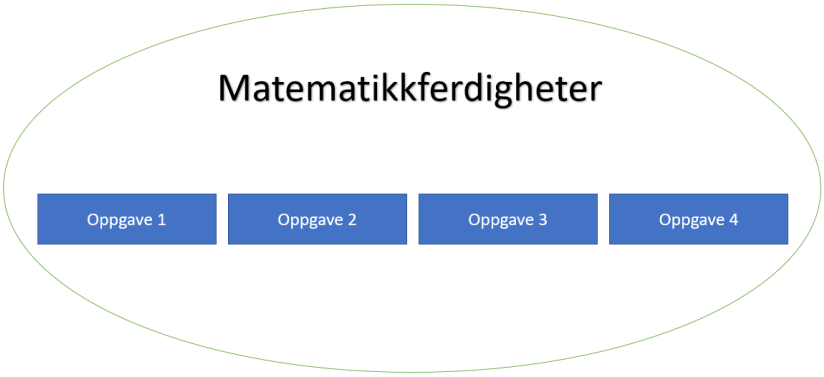
forskjellige tolkninger av dette var en av kildene til uenighet i kodingen. Det var også flere oppgaver hvor vi hadde hatt forskjellig oppfatning om hvilken kompetanse som best beskrev den kognitive delen av løsningsprosessen. Denne oppfatningsforskjellen førte dermed til situasjoner hvor begge kodere var enige i at oppgaven krevde en komponent av matematisk kompetanse, men én koder kunne vurdere oppgaven til å trenge resonneringskompetanse mens den andre koderen vurderte matematiseringskompetanse til å bedre beskrive løsningsprosessen.

Det er resultatene fra denne siste felles kodingsprosessen som danner grunnlag for videre analyser i oppgaven. Resultatene er i seg selv viktige for valideringsprosessen av eksamen. De sier noe om innholdet i eksamensoppgavene, og er dermed med på å validere innholdet i prøven. Dette blir spesielt viktig når eksamener skal lages etter fagfornyelsen, og de nye kjerneelementene blir en eksplisitt del av læreplanen. Resultatene vil også brukes i sammenheng med resultatene fra de kvantitative analysene. Ved å koble disse resultatene vil vi kunne undersøke om det finnes en sammenheng mellom oppgavers vanskegrad og kompetansekrav, for eksempel om noen av kompetansene ikke representeres i vanskelige eller lette oppgaver. En tilsvarende sammenlikning kan gjøres for DIF-analysen ved å se etter koblinger mellom kompetansekrav og forskjellig vanskelighet for gutter og jenter.

4.2 Endimensjonal Item Response-analyse

Analysen i forrige seksjon baserer seg utelukkende på de faktiske eksamensoppgavene fra våren 2019, og brukte en kvalitativ tilnærming for å tolke kompetansekrav til oppgavene. For å kunne se på oppgavene fra et annet empirisk perspektiv ble det også utført en kvantitativ oppgaveanalyse for å undersøke dimensjonaliteten av kompetanseskala og vanskegrad for oppgavene. Analyse ble gjennomført ved hjelp av ConQuest (Wu & Adams, 2007); ConQuest er en programvare som benytter seg av modeller fra *Item Response Theory* (heretter IRT). Datamaterialet som benyttes i denne analysen er tabellen med responsmønstre som ble illustrert i Tabell 1. I en IRT-analyse brukes alle disse verdiene for å kunne si noe om både vanskegraden til oppgavene og ferdigheten til hver enkelt elev.

Analysen baserer seg på et par antagelser som bør nevnes før metoden beskrives ytterligere. I psykometrilitteraturen kalles en egenskap som ikke kan måles direkte for en latent variabel (Baker & Kim, 2017). I *endimensjonale* analyser, derav navnet, antas det at alle oppgavene fra prøven tester den samme latente variabelen, eller samme dimensjon av et konstrukt. Alle oppgavene i eksamen antas derfor å teste den samme egenskapen hos elevene, og vi kan kalle denne egenskapen for matematikkferdigheter, slik det illustreres i Figur 6.

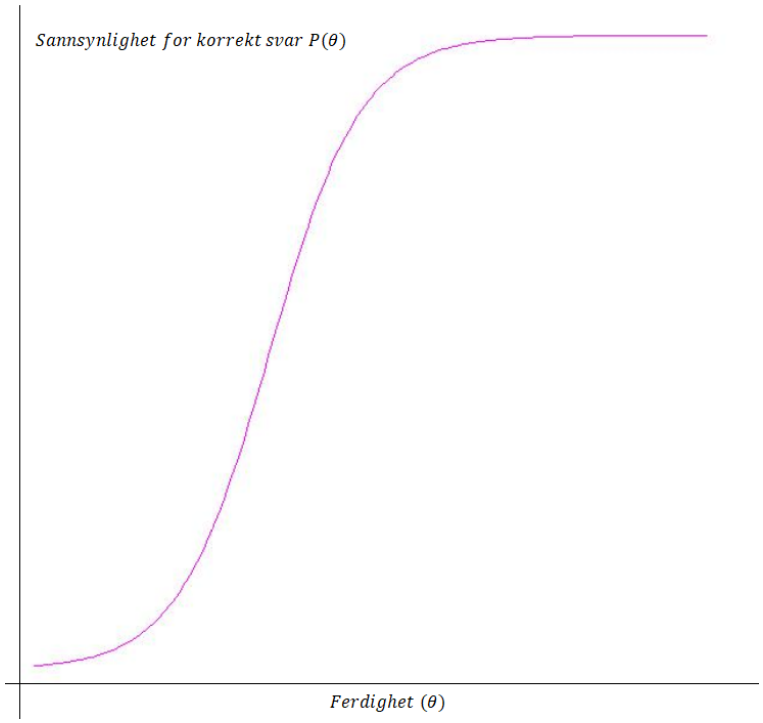


Figur 6: Illustrasjon av testens antatte struktur

Antagelsen om endimensjonalitet krever en forklaring. Hittil i oppgaven har konstruktet matematisk ferdighet blitt beskrevet som et multidimensjonalt konstrukt med flere forskjellige aspekter. Det kan dermed virke rart at det flerdimensjonale bildet av matematikkferdigheter forkastes i denne delen av analysen, og at matematisk ferdighet nå ses på som et endimensjonalt konstrukt. Dette er en rimelig kritikk av metoden; og en multidimensjonal IRT-analyse, hvor man analyserer om oppgavene representerer flere egenskaper, er mulig å gjennomføre. Det ble allikevel valgt å ikke utføre en multidimensjonal analyse pga. restriksjoner i både tid og oppgaveomfang. Det multidimensjonale aspektet ved konstruktet blir derfor undersøkt kvalitativt som beskrevet i seksjon 4.1. Niss and Jensen (2002) og Kilpatrick et al. (2001), som begge har utviklet rammeverk for å beskrive matematisk kompetanse, understreker også at de forskjellige aspektene ved matematisk kompetanse ikke er uavhengige av hverandre, men bør sees på som aspekter av et helhetlig konstrukt. Derfor er det ikke er en urimelig antagelse å benytte en endimensjonal analyse for denne delen av undersøkelsen. Et annet argument for antagelsen er ønsket om å undersøke dimensjonaliteten til eksamensoppgavene. Hvis en endimensjonal analyse gir en modell som er en god tilnærming til dataene er dette en indikasjon på at eksamen faktisk tester ett endimensjonalt konstrukt.

Alle oppgavene antas altså å teste den latente egenskapen matematikkferdigheter, og det er rimelig å forvente at alle testtagerne besitter et nivå av denne egenskapen. Dette nivået kan tilordnes en verdi, og det er vanlig å bruke bokstaven θ for å betegne denne verdien (Baker & Kim, 2017). I IRT-analysen antas det at denne ferdigheten påvirker sannsynligheten, P(θ), for at en testtager besvarer en oppgave korrekt, og at det dermed er mulig å finne en funksjon som beskriver hvordan denne sannsynligheten avhenger av ferdigheten θ. For høyere ferdighet bør sannsynligheten for korrekt

svar øke, og den bør nærme seg 1 når ferdigheten blir veldig høy. For et eller annet ferdighetsnivå vil sannsynligheten for at en oppgave besvares korrekt være lik 50%. I de fleste IRT-modeller kalles denne verdien for δ og betegner oppgavens vanskegrad. På denne måten kan testtagere og oppgaver plasseres på samme skala, og man har dermed et sterkt grunnlag for å sammenlikne oppgavers vanskegrad og ferdigheter hos testtagerne. Når de nødvendige parametere er regnet ut kan sannsynligheten for at en person klarer en oppgave plottes som en funksjon av ferdighet. Denne funksjonen bør være strengt økende og tilnærme seg 100% når ferdighet blir høy. Resultatet blir en logistisk s-kurve, og kalles oppgavens karakteristiske kurve (ICC) (Baker & Kim, 2017). Se *Figur 7* for et eksempel på en karakteristisk kurve.



Figur 7: Eksempel på karakteristisk kurve

Den andre av de to antagelsene som ligger bak analysen er at alle oppgavene er lokalt uavhengige av hverandre. Mestring av en oppgave skal altså ikke påvirke sannsynligheten for å mestre etterfølgende oppgaver. Analysen av oppgavene som beskrives i seksjon 4.1 undersøkte også dette, men da det ikke var hovedmålet bak analysen er det ikke inkludert i beskrivelsen eller i resultatene. Alle deloppgavene på eksamen kan løses individuelt uten å ha klart tidligere oppgaver. Det er allikevel ett eksempel på en deloppgave hvor det virker rimelig å anta at sannsynligheten for å klare oppgaven øker hvis eleven har klart den foregående deloppgaven. Dette er oppgave 11c fra eksamen del 1. Den har blitt inkludert i analysen da den er mulig å løse uten ytterligere informasjon, men den er allikevel verdt å nevne.

Analysen ble gjennomført i to steg. I det første steget bruker ConQuest svarene fra alle elevene for å estimere en verdi for vanskegraden til alle oppgavene. I dette steget ble alle tilfeller hvor elever ikke hadde besvart en oppgave (kodet som verdi 9 i regnearket) satt til å være manglende, og ikke til å telle som feil svar. Wu & Adams (2007) skriver at hvordan manglende svar skal behandles kan variere etter formålet ved testen. Noen ganger er det hensiktsmessig å behandle dem som ikke korrekt, andre ganger som manglende. Dette vil i prinsipp si at analysen tolker det som om noen testtagere ikke fikk utdelt alle oppgavene. I denne oppgavens analyse ble det gjort for å forhindre at oppgaver som mange elever ikke har besvart blir estimert til å være vanskeligere enn de er. Disse resultatene benyttes så videre i steg 2. I dette steget bruker ConQuest de samme dataene som i steg 1 for å estimere ferdigheten til alle elevene i datasettet. I dette steget ble avgjørelsen angående manglende besvarelser behandles annerledes. Tilfeller hvor elever ikke hadde besvart en oppgave ble i dette steget kodet som feil svar. Det er ikke nødvendigvis en sikker antagelse at en elev ikke ville ha klart en oppgave de ikke har besvart, og det er i dette steget også mulig å kode dataene som manglende. Allikevel ble det valgt å kode manglende besvarelser som feil svar på bakgrunn av at elevene hadde tilgang til alle oppgavene. Det er en rimelig antagelse at årsaken til å velge å ikke besvare en oppgave er at eleven vurderte det til en oppgave de sannsynligvis ikke ville mestre, eller i det minste ville ha en lavere sannsynlighet for å mestre enn andre oppgaver.

Analysen viste en EAP/PV (Expected A Posteriori/Plausible Value) skala reliabilitet på 0.915. Denne verdien er en indikator på analysens reliabilitet når feilmarginer tas hensyn til, og ligger godt over grensen for aksepterte verdier. Dette er også en god indikator på at en endimensjonal modell er en god tilnærming, og at det dermed kan forsvares å anta at alle oppgavene på eksamen tester ett og samme konstrukt.

4.3 Analyse av differential item functioning (DIF)

I seksjon 2.5 ble bias i prøver diskutert, og trusselen bias kan utgjøre mot en tests validitet ble tatt opp. Det er derfor av stor interesse å undersøke om bias eksisterer i norske eksamensoppgaver. Det er hva den tredje analysen ønsker å avdekke. I forrige seksjon (4.2) ble parameterne θ og δ beskrevet, som verdier for elevers ferdighetsnivå og oppgavers vanskegrad. Fordi disse parameterne er utregnet empirisk er det ikke nødvendig at en analyse vil returnere samme verdier når forskjellige sett med responsmønstre brukes som datagrunnlag, og en oppgaves vanskegrad trenger derfor ikke å være lik for alle grupper.

Denne muligheten vil nå undersøkes grundigere gjennom en DIF-analyse. DIF er en forkortelse for differential item functioning, og beskriver et fenomen som kan oppstå når to individer med lik

estimert ferdighet ikke har lik sannsynlighet for å mestre en oppgave (Zumbo, 1999). Det viktige å huske fra denne setningen er «lik estimert ferdighet». Det er forventet at to personer har ulike forutsetninger for å mestre en test; hvis vi derimot vet at de innehar samme nivå av den underliggende ferdigheten θ som testes, men fortsatt har forskjellig forutsetning for å mestre en oppgave på prøven, utviser denne oppgaven DIF. Implikasjonene dette har for equity og testens validitet bør være åpenbare, da det indikerer at en oppgave er lettere for noen enn den er for andre med samme nivå av ferdighet, og det er derfor ønskelig at DIF ikke skal oppstå på noen av testens oppgaver.

Datagrunnlaget for analysen inneholder variabler for kjønn for alle testtagerne, så den eneste formen for DIF som lar seg undersøke er om oppgaver oppfører seg annerledes for jenter enn de gjør for gutter. For å undersøke om oppgaver på eksamen viser DIF ble et program skrevet i programmeringsspråket R. Hele programmet kan finnes lagt ved som vedlegg A (seksjon 9.1).

Programmet bruker flere steg for å oppnå dette. Gutter og jenter separeres i to grupper, jenter defineres som referansegruppe, gutter defineres som fokusgruppe. Deretter analyseres elevsvarene med modeller fra IRT. Dette er i prinsipp det samme som har blitt utført i den forrige seksjonen, men utført i et annet program. Undersøkelsen som forsøker å lokalisere DIF-oppgaver begynner i neste steg. Dette steget går gjennom samme utregninger som første steg, men i flere runder. For hver gjennomgang utregner programmet, basert på alle elevsvarene, parametere for vanskegrad på oppgaver og ferdigheter til testtagerne. Forskjellen fra forrige steg er at programmet utregner en separat vanskegradsparameter, δ, for gutter og jenter. Utregningen gjøres en gang for hver oppgave på eksamen, og i hver gjennomgang antas det at i alle de andre oppgavene er vanskegraden lik for både gutter og jenter. Målet med steget er å identifisere oppgaver som ikke utviser DIF. Utregningen finner ofte falske positiver, så vi prøver derfor å finne oppgavene med lavest DIF slik at disse oppgavene kan benyttes som grunnlag for resten av analysen. De fem oppgavene som viser minst tegn til DIF noteres, og benyttes som anker videre i analysen. At de brukes som anker betyr i denne sammenhengen at de tvinges til å ikke ha forskjellig vanskegrad for fokus og referansegruppen, og kan dermed brukes som grunnlag for å identifisere oppgaver som faktisk har forskjellig vanskegrad.

I neste steg (steg 4 i programkoden) utregnes en ny modell hvor disse fem oppgavene settes til å være ankere og et nytt sett med parametere for oppgavene estimeres. Parameterne av interesse som regnes ut er vanskegrad for begge grupper på alle oppgavene. Programmet starter nå en utregningsløkke (steg 5 i koden). For hver gjennomgang av løkken tvinges ankeroppgavene og ytterligere én oppgave til å være like vanskelig for begge gruppene. Parametere for vanskegrad utregnes så for de resterende oppgavene, og programmet sammenlikner nå hvor godt den

utregnede modellen stemmer med responsmønsteret fra eksamen. For gjennomgangen av løkken som forsøker å avgjøre om en gitt oppgave utviser DIF ser programmet altså på sannsynligheten for at de observerte svarene ville oppstått dersom oppgaven faktisk ikke utviser DIF. Hvis sannsynligheten for de observerte resultatene er høy er det en rimelig antagelse at den valgte oppgaven ikke utviser DIF, sannsynligheten er lav blir oppgaven flagget som en oppgave hvor DIF eksisterer. Grensen for om en oppgave oppfører seg forskjellig for de to gruppene er en p-verdi på 0.05 eller høyere. En oppgave flagges altså som en DIF-oppgave hvis sannsynligheten for det observerte responsmønsteret er lavere enn 5%. Siste steg (steg 6 i programmet) utregner en endelig modell med vanskegrad for de to gruppene etter at alle oppgaver med DIF er identifisert og tvunget til å være like vanskelige for gutter og jenter. Programmet plotter også de karakteristiske kurvene for alle oppgavene og skriver disse ut i egne filer.

5 Resultater og analyse

I dette kapittelet vil resultatene fra de tre analysene bli presentert. En full presentasjon av alle resultatene fra de kvantitative analysene er ikke inkludert, da det er for mye data å legge ved i oppgaven, og fokus vil være på globale trender med illustrerende eksempler. En mer fullstendig oversikt over resultatene kan finnes i vedleggene.

5.1 Analyse av oppgavers kompetansekrav

Funnene fra analysen av kompetansekrav er de første resultatene jeg vil presentere, da disse er interessante og leder til flere nye spørsmål. Selv om alle kompetansene i MEG-rammeverket var representert i eksamensoppgavene var det helt klart ikke en jevn fordeling, og noen kompetanser virker derfor viktigere enn andre. Den komplette listen over kompetansekrav for hver enkelt deloppgave i eksamen er lagt ved denne oppgaven i vedlegg C. Det var noen resultater som imidlertid gjorde seg merkbare, og jeg vil nevne disse i denne delen.

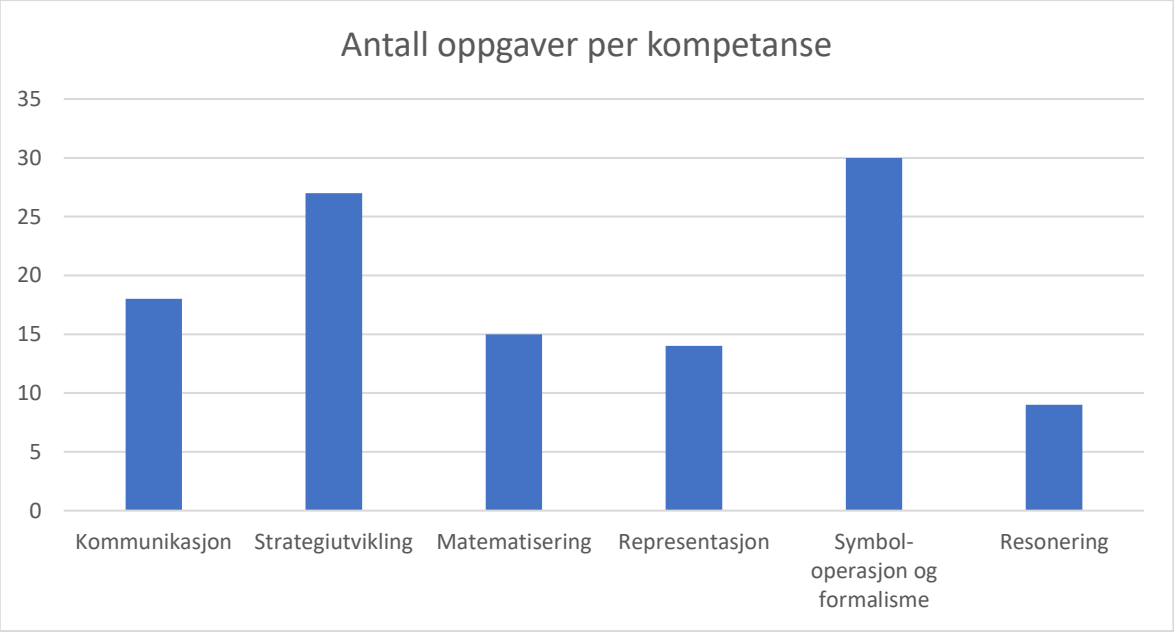
Først er det viktig å understreke at arbeidet med å klassifisere kompetansekrav for eksamensoppgavene var en stor utfordring. Både undertegnede og veileder syntes at oppgavene var veldig like hverandre, og at det tidvis var vanskelig å vurdere hvilke kompetanser fra MEG-skjemaet som var representert i hver enkelt oppgave. De seks kompetansene i MEG-rammeverket er, som nevnt i teoridelen*, kommunikasjon, strategiutvikling, matematisering, representasjon, symbol-operasjon og formalisme (SOF), og resonnering*. Min analyse av eksamensoppgavene viser at kompetansekravene for oppgavesettet presenteres under, og kan også sees representert i Figur 8.

- Kommunikasjon er representert i 18 av 54 oppgaver.
- Strategiutvikling er representert i 27 av 54 oppgaver.
- Matematisering er representert i 15 av 54 oppgaver.
- Representasjon er representert i 14 av 54 oppgaver.
- Symbol- operasjon og formalisme er representert i 30 av 54 oppgaver.
- Resonnering er representert i 9 av 54 oppgaver.

Viktige punkter å legge merke til foreløpig er at alle de seks kompetansene er representert i eksamensoppgavene, så ingen aspekter ved matematisk kompetanse er totalt fraværende. Dette er et godt funn, da det tyder på at eksamen faktisk tester konstruktet matematisk kompetanse bredt, og ikke fokuserer kun på enkelte aspekter av det. Man kan allikevel se at noen kompetanser er hyppigere representert enn andre. Spesielt dominerende er kompetansene strategiutvikling og SOF,

som begge er representert i minst halvparten av eksamenenes oppgaver. Da SOF-kompetanse er den kompetansen som ligger nærmest de fleste matematikkoppgaver i skolen burde ikke dette være overraskende. For strategiutvikling er det også verdt å minne om kriteriet for at kompetansen skulle vurderes til nivå 1 eller høyere. Nivå 0 er beskrevet som oppgaver hvor løsningsstrategien er åpenbar eller eksplisitt nevnt. De fleste oppgaver som ikke er ferdig oppstilte regnestykker blir altså vurdert til å kreve kompetanse i strategiutvikling.

Et annet viktig funn å legge merke til er at resonneringskompetanse er det aspektet ved matematisk kompetanse som viser seg å være lavest representert. Denne kompetansen er primært gjeldende i oppgaver som krever argumentasjon i løpet av løsningsprosessen, som f.eks. oppgave 8c del 2, og er den kompetansen som er nærmest knyttet den logiske delen av matematikkfaget.



Figur 8: Antall oppgaver som aktiverer hver kompetanse

Resultatene viser også at matematisering, representasjon og kommunikasjon er betraktelig mindre tilstedeværende enn strategiutvikling og SOF. Det er flere eksamensoppgaver som ikke bruker representasjoner (de er nesten ikke tilstedeværende i del 1), og der de benyttes er det ofte kun nødvendig å lese av ett tall for å kunne løse oppgaven. Matematisering, som omhandler å kunne oversette en reell problemstilling til en matematisk oppgave er også noe vi ikke finner i det fleste oppgavene. Kommunikasjon er noe representert, dette gjelder som regel oppgaver der det er nødvendig å lese oppgaveteksten grundig, eller flere ganger, i løpet av løsningsprosessen.

Disse resultatene leder til hypotesen at oppgavene på eksamen er like nok til at de tester én dimensjon. Sagt med andre ord vil en person som har høy sannsynlighet for å klare én av oppgavene også kunne klare de resterende med høy sannsynlighet. Dette er ikke urimelig for en eksamen som

tross alt skal gi karakter i ett fag, og at det er korrelasjon mellom oppgaver er forventet. Hvis kompetansespekteret ved eksamen blir for snevert kan det allikevel bety at matematikk er et fag man enten mestrer eller ikke. Læreplanen sier at matematikk er et fag sammensatt av flere kompetanser, noe som er enda tydeligere nevnt i de nye læreplanene, og en analyse av oppgavene viser at alle kompetansene i MEG-rammeverket er representert. De virker allikevel veldig like hverandre, som støtter hypotesen at eksamen tester et endimensjonalt konstrukt. For å teste denne hypotesen ble det utført en endimensjonal IRT-analyse, og resultatene presenteres i neste seksjon.

5.2 Endimensjonal IRT-analyse av eksamensresultatene.

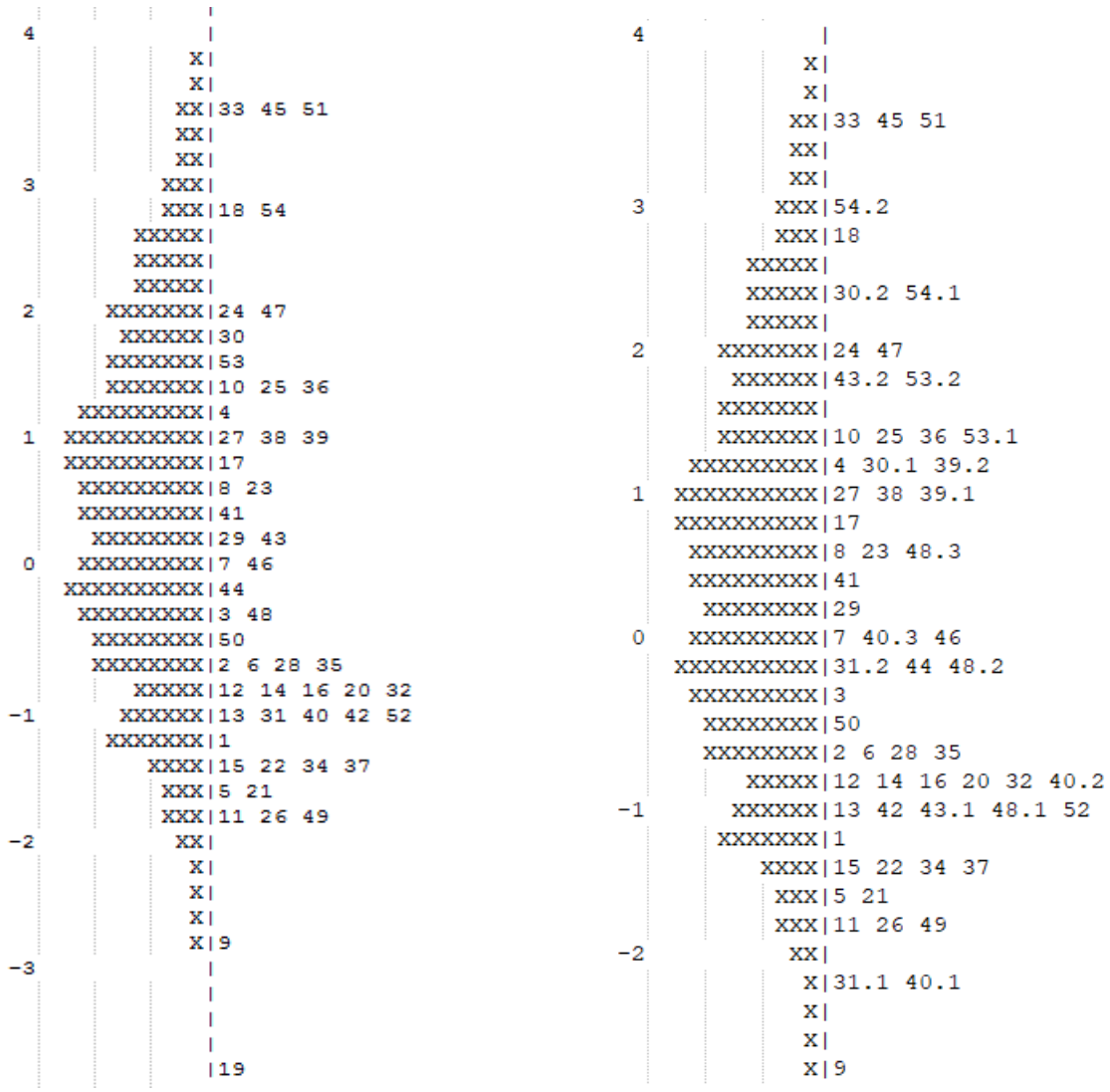
Den endimensjonale IRT-analysen ble utført ved hjelp av programvaren Conquest. I utgangspunktet var hensikten med denne analysen å kunne si noe om kompetansekrav til oppgaver av forskjellig vanskegrad. Spørsmål av interesse var

- Testes alle kompetanser i MEG-skjemaet på eksamen?
- Testes alle kompetanser for alle vanskegrader? Hvis ikke, hvilke kompetanser testes ikke?
- Hvilke aspekter av matematisk kompetanse kjennetegner lette/vanskelige oppgaver?
- Er oppgavene til eksamen fordelt langs hele ferdighetsspekteret, eller er de i større grad lette/vanskelige?

Den utregnede modellen stemte veldig godt med resultatene, og hadde en utregnet EAP/PV reliabilitet på 0.941. Dette er en høy reliabilitet, og indikerer at en endimensjonal modell gir en veldig god tilpassing til dataene, og dermed at oppgavene tester samme egenskap. Resultatene fra analysen presenteres her som et Wright-map i *Figur 9*. Et Wright-map er en grafisk representasjon av oppgaver og elever, og plasserer dem langs samme skala. Hvert kryss på venstre side av figuren representerer tilnærmet 17 elever, mens tallene på høyre side representerer en deloppgave fra eksamen. Den vertikale aksen representerer vanskegrad på oppgaver og ferdighet til elevene. *Figur 9* inneholder resultatene fra to forskjellige modeller. Den venstre halvdelen av figuren viser resultatene for modellen hvor alle tilfeller av delvis uttelling har blitt behandlet som om eleven fikk null poeng. Høyre halvdel av figuren viser en modell hvor delvis uttelling (f. eks. hvor en elev har fått ett av to mulige poeng) har blitt tatt hensyn til.

Som man kan se i Figur 9 er oppgavene godt fordelt langs hele ferdighetsspekteret. Den letteste oppgaven (Item 19, Oppgave 12a Del 1), er en oppgave man kan forvente at så godt som alle elever vil klare, mens de vanskeligste oppgavene (Items 33, 45 og 51; Oppgaver 1c, 5d og 8c Del 2) er

oppgaver et fåtall av elevene vil mestre. Det er en liten overvekt av lettere oppgaver, men over prøven sett som helhet er det en rimelig fordeling av vanskegrad.



Figur 9 – Endimensjonal IRT-analyse. Hver X representerer 16.67 elever

Modellen som tar hensyn til delvis uttelling gir resultater som er relativt like den første modellen, men det er enkelte forskjeller som er interessante. Jeg vil påpeke noen av dem her.

Item 39 (oppgave 3c, Del 2) kunne gi to poeng til elever som løste den tilfredsstillende, med mulighet for ett poeng for de elevene som kun klarte deler av fremgangsmåten. Hvis man ser på resultatene fra modellen for delvis uttelling i Figur 9, hvor dette modelleres som to forskjellige oppgaver, ser man at de har nesten helt like vanskegrad. De eksakte tallene for denne oppgaven var fordelt slik; 1049 elever fikk ingen uttelling, 123 elever fikk ett poeng, 1123 elever fikk full uttelling (to poeng).

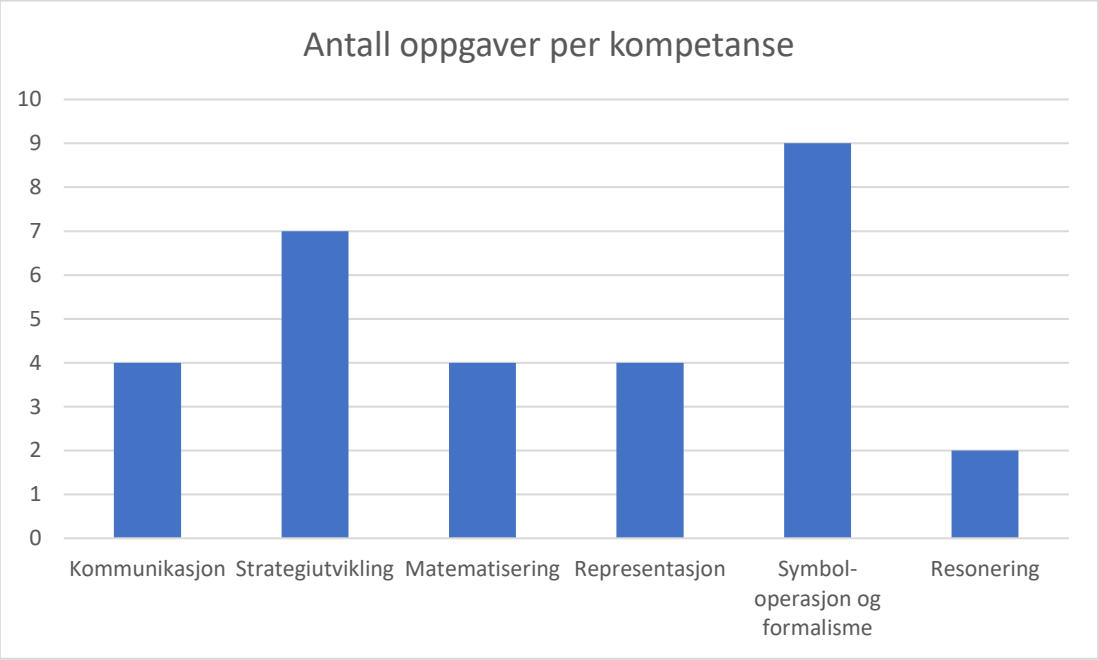
Som kontrast kan man se at item 40 (oppgave 4a, Del 2) har sine grenser for delvis uttelling mye skarpere adskilt. Ett poeng (item 40.1) har en vanskegrad på -2.23, to poeng (item 40.2) har en vanskegrad på -0.88, og full uttelling (item 40.3) har en vanskegrad på 0.13. Fordelingen av elever per nivå er i denne oppgaven strengt stigende; 81 elever fikk ingen uttelling, 292 elever fikk ett poeng, 600 elever fikk to poeng og 1885 elever fikk tre poeng. Disse resultatene er interessante, selv om de ikke er strengt relevante for forskningsspørsmålene, da det viser hvordan delvis uttelling kan oppføre seg forskjellig. Hvis de to poengfordelingene har tilnærmet lik vanskegrad, bør det stilles spørsmål til om det er nødvendig å adskille dem i det hele tatt. Hvis en elev har delvis uttelling på denne oppgaven, mens en annen elev har full uttelling, har de fått utdelt forskjellig antall poeng for å ha klart å løse oppgaver som er estimert til å være omtrent like vanskelige.

5.3 Kompetansekrav og vanskegrad

Det er nå naturlig å se på hvordan resultatene fra seksjon 5.1 og 5.2 sammen lager et bilde av innholdet i oppgaver fra forskjellige vanskegrader. Seksjon 5.1 viser resultatene av kompetanseanalysen for eksamensoppgavene, mens seksjon 5.2 gir en veldig reliabel inndeling av oppgavers vanskegrad, og disse resultatene kan nå sees i sammenheng. Som man kan se i Figur 9 faller nesten hele elevmassen i ferdighetsintervallet mellom nivå -3 og 3. Dette intervallet blir nå delt opp i tre separate deler som kalles lette oppgaver, middels vanskelige oppgaver og vanskelige oppgaver for å se hvordan oppgaver av forskjellige vanskegrader passer inn i kompetanseskjemaet fra analysen. Det ønskede resultatet er at alle kompetanser finnes i alle tre intervaller, fordi det vil indikere at det finnes muligheter for å demonstrere kompetanser for flertallet av elevene.

5.3.1 Lette oppgaver. Vanskegrad lavere enn θ = -1.

I dette intervallet finner vi 16 deloppgaver. Jeg har ikke inkludert delvis uttelling på oppgaver i denne seksjonen, så de 16 oppgavene er oppgavene fra modellen som kun inkluderer full uttelling. Oppgavene i dette intervallet er oppgave 1a, 3, 6a, 7, 8b, 10, 12a, 12c, 13 og 17a i Del 1; og 1a, 2a, 3a, 5a, 8a og 9a i Del 2. Fordelingen av kompetanser for disse oppgavene kan sees under i Figur 10.

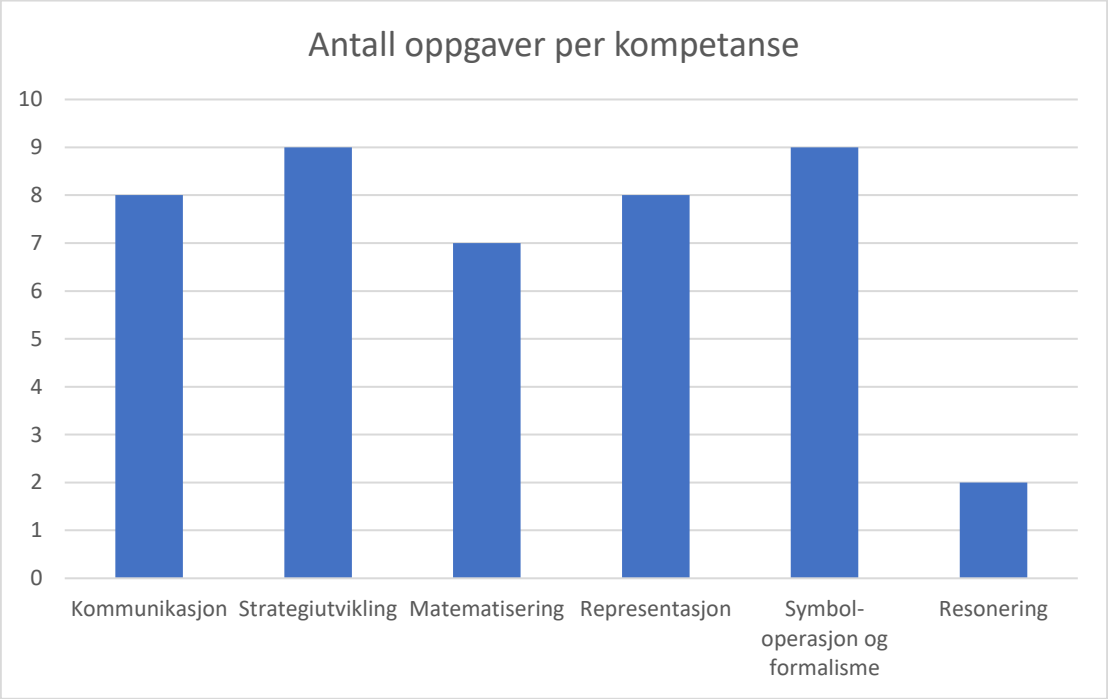


Figur 10 Kompetansefordeling i lettere oppgaver

Figuren viser at vi i dette intervallet finner alle aspektene ved kompetanser representert. Det ser altså ut som om at selv for svakere elever er det mulig å demonstrere alle aspektene ved sin matematiske kompetanse. Fordelingen er også relativt lik oversikten for eksamenssettet som helhet; SOF-kompetanse og strategiutvikling er forholdsvis kraftigere representert enn de fire øvrige kompetansene, og resonneringskompetanse er relativt sjelden.

5.3.2 Middels vanskelige oppgaver. Vanskegrad mellom θ = -1 og θ = 1

I det middels vanskelige intervallet finner vi 23 deloppgaver. Som tidligere har jeg kun inkludert varianten av oppgaven fra modellen som kun teller full uttelling. Oppgavene i dette intervallet er oppgave 1b, 2a, 4a, 4b, 5, 8a, 9, 11a, 11b, 12b, 14, 17b, 18 og 19 i Del 1; og 1b, 2b, 4a, 4b, 5b, 5c, 6a, 7 og 8b i Del 2. Fordelingen av kompetanser i disse oppgavene kan sees i Figur 11.

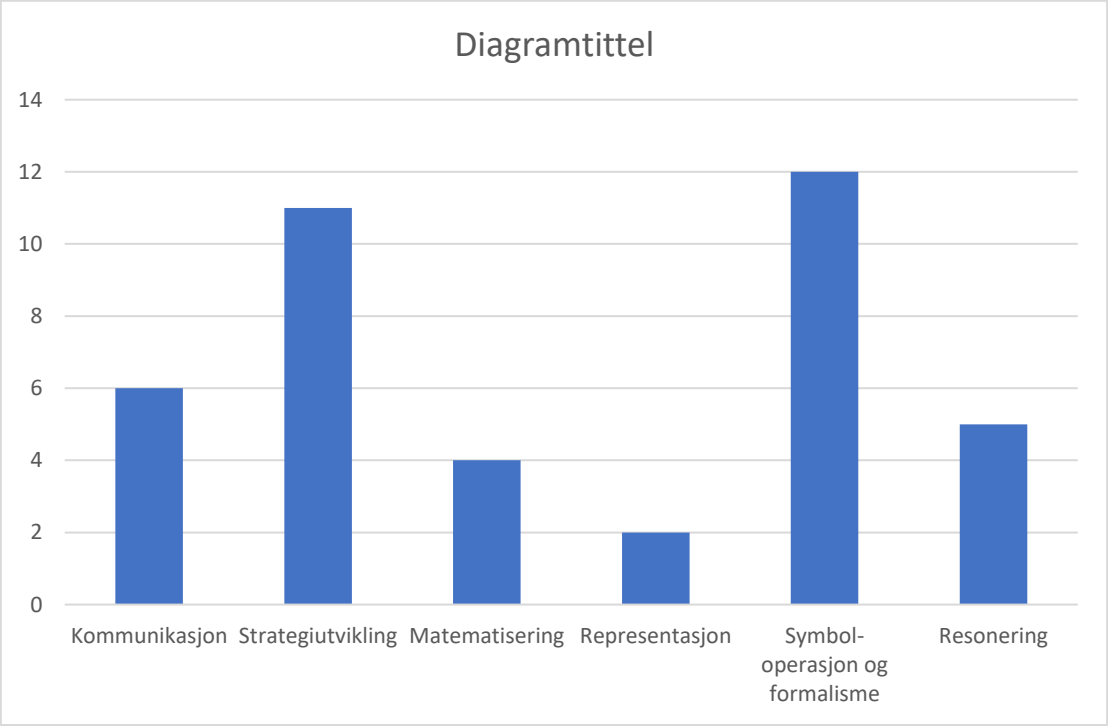


Figur 11 Kompetansefordeling i middels vanskelige oppgaver

I det middels vanskelige intervallet (som er intervallet med flest oppgaver), finner vi også oppgaver som representerer alle kompetanser. Funn det er verdt å legge merke til er at resonneringskompetanse fortsatt er lavt representert (2 av 23 oppgaver), som forholdsmessig er en lavere representasjon enn i intervallet med lavest vanskelighet. I tillegg ser vi at selv om SOF og strategiutvikling fortsatt er kompetansene som står sterkest er de ikke like dominerende i forhold til kommunikasjon, matematisering og representasjonskompetanse. Med unntak av resonneringskompetansen virker de middels vanskelige oppgaven som om de utgjør en veldig jevn fordeling av forskjellige aspekter av kompetanse.

5.3.3 Vanskelige oppgaver. Vanskegrad større enn θ = 1

I det vanskeligste intervallet finner vi 15 deloppgaver. Disse er oppgave 2b, 6b, 11c, 15, 16 og 20 fra Del 1; og 1c, 2c, 3b, 3c, 5d, 6b, 8c, 9b, 9c fra Del 2. Antallet oppgaver som krever de ulike kompetansene kan sees under i Figur 12.

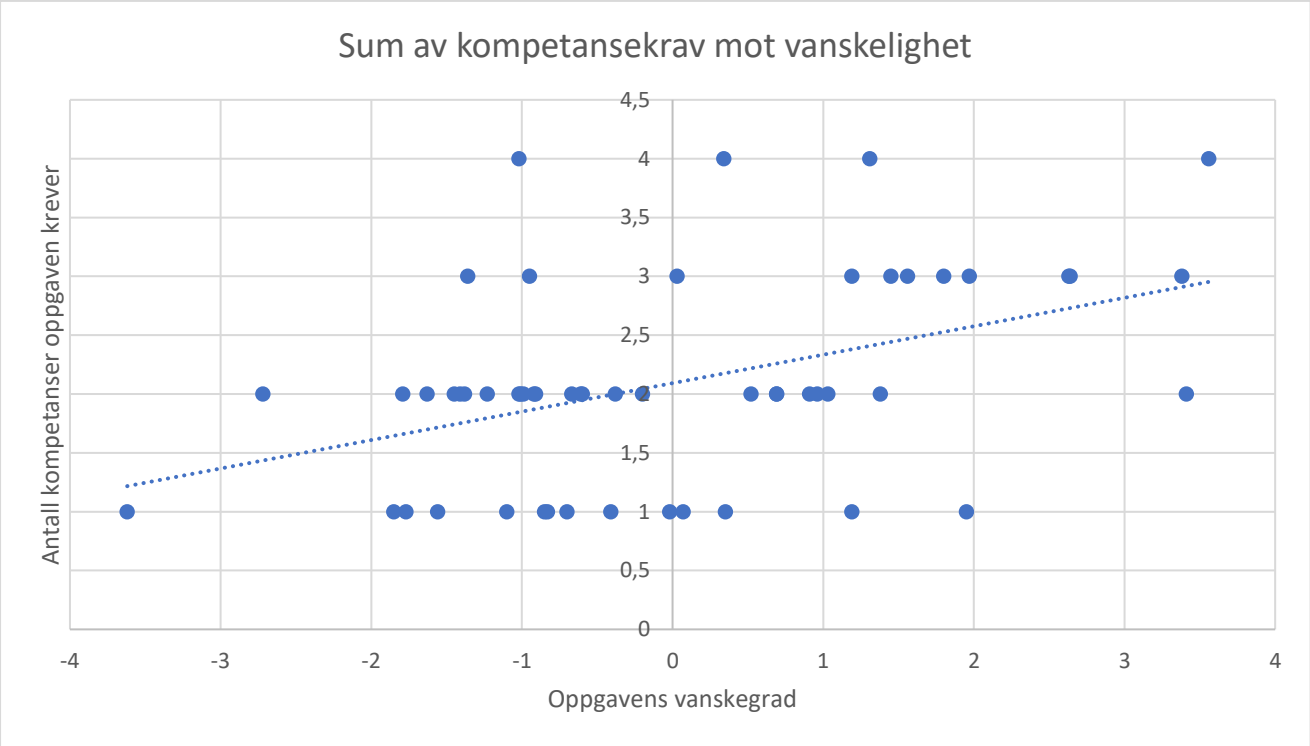


Figur 12 Kompetansefordeling i vanskelige oppgaver

Nok en gang kan vi se at alle kompetanser fra rammeverket er representert, men distribusjonen er her på mange måter annerledes. SOF og strategiutvikling dominerer igjen, men enda kraftigere enn for eksamenssettet som helhet. Hele 12 av 15 oppgaver krever SOF-kompetanse ifølge min analyse. Representasjonskompetanse er nesten fraværende, og matematisering er også lavt representert. Resonneringskompetanse er derimot sterkere representert for vanskelige oppgaver. 5 av 15, en tredjedel, av oppgavene i dette intervallet krever denne kompetansen. Sammenlignet med 9 av 54, en sjettedel, for eksamenssettet som helhet er dette en betraktelig økning.

5.3.4 Totalt kompetansekrav per oppgave

Figur 13 viser et plott av alle deloppgaver i eksamenssettet. Hvert punkt i figuren representerer en deloppgave, hovedaksen viser deloppgavens utregnede vanskegrad og andreaksen viser hvor mange aspekter av matematisk kompetanse oppgaven krever (0-6). Figuren viser også en utregnet trendlinje for datapunktene. Det vises til en viss grad at vanskeligere oppgaver generelt aktiverer flere kompetanser enn lette oppgaver. I det vanskeligste intervallet er det kun to deloppgaver som er vurdert til å kreve kun én kompetanse, mens det letteste intervallet har fem slike deloppgaver. Av de 14 oppgavene som er vurdert til å kreve tre eller fire kompetanser finner vi ni av dem i det vanskelige intervallet, mens kun to ligger i intervallet for lette oppgaver.



Figur 13

Hva sier disse resultatene? IRT-analysen viser ganske tydelige tegn på et endimensjonalt konstrukt, mens den kvalitative analysen tyder på at flere kompetanser er representert. En mulig måte å koble disse to delvis motstridende resultatene kan være representert i figuren over. Flere kompetanser representert i samme oppgave kan manifester seg som en oppgave med høyere vanskegrad i konstruktet. Pettersen & Braeken (2019) fant at en vektet IRT-modell med kompetanserammeverket til PISA som grunnlag ga god tilpassing til datasettet fra eksamen da de brukte læreres kompetansevurderinger av oppgaver som basis for sin modell. Turner et al. (2013) fant også at

kompetansevurderinger kunne brukes som prediksjon av oppgavers vanskegrad. Det er derfor ikke overraskende å se at hvilke kompetanser som kreves samsvarer med hvor vanskelig en oppgave er.

5.4 Resultater fra analyse av Differential Item Functioning

Tidligere nevnte vi flere forskjellige aspekter ved validitet. For eksamen, som brukes som del av opptaksgrunnlag for videre utdanning er konsekvensaspektet kanskje noe av det viktigste å vurdere. For å kunne si noe mer detaljert om dette valgte jeg å gjennomføre en DIF-analyse. Ved siden av poengfordeling inneholder datasettet kun parametere for kjønn, så dette er den eneste mulige formen for DIF som lot seg undersøke. Resultat av analysen som helhet kan finnes i vedlegg D, her presenteres bare et helhetlig bilde og resultatet fra enkelte deskriptive oppgaver. Før resultatene presenteres må jeg minne om at DIF-analysen og den opprinnelige endimensjonale IRT-analysen ble utført med to forskjellige programvarer. Tallene i denne seksjonen kan derfor ikke sammenliknes direkte med resultatene i seksjon 5.2.

Oppgave	Vanskegrad Jenter	Vanskegrad Gutter	Forskjell Vanskegrad
1a, Del 1	-2,456	-1,445	-1,011
4a, Del 1	-0,960	-0,651	-0,308
5, Del 1	-0,263	0,143	-0,406
6a, Del 1	-2,607	-1,831	-0,776
8a, Del 1	-1,537	-0,969	-0,567
9, Del 1	-1,553	-1,095	-0,459
16, Del 1	0,401	0,766	-0,364
17b, Del 1	0,348	0,187	0,161
20, Del 1	0,748	0,547	0,201
1b, Del 2	-0,594	-0,902	0,307
1c, Del 2	1,808	2,565	-0,758
2c, Del 2	0,551	0,630	-0,079
3c, Del 2	0,236	0,445	-0,209
5b, Del 2	-0,118	-0,325	0,206
5c, Del 2	0,312	0,110	0,201
7, Del 2	-0,848	-1,099	0,250

Figur 14: Oppgaver som viser tegn til DIF

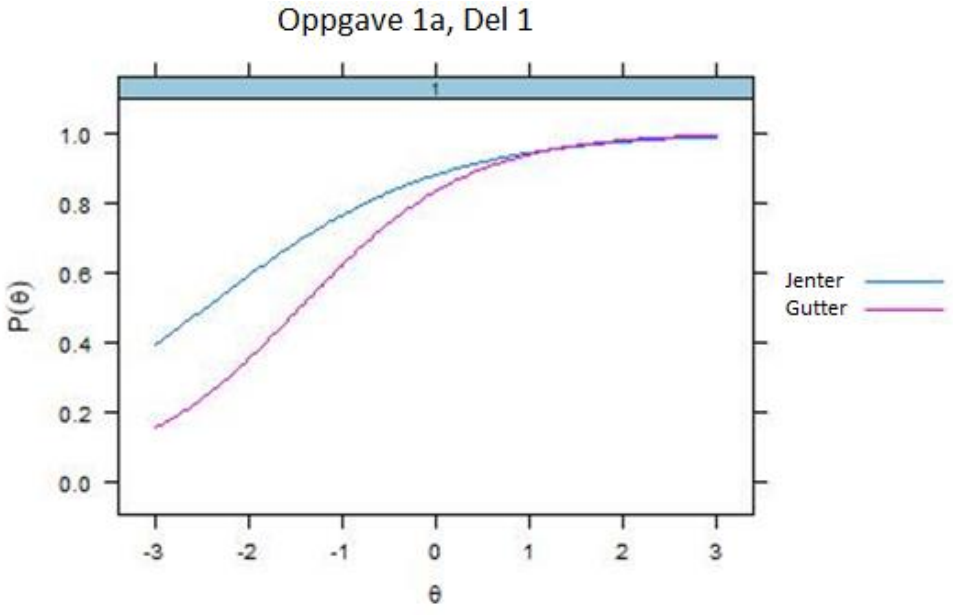
Flere av oppgavene i eksamen viser tegn til DIF, og for enkelte av dem er effekten meget kraftig. Eksamen består av 54 deloppgaver, og av disse kunne man i 16, nesten en tredjedel av eksamen, se kraftig nok forskjell i oppførsel mellom gutter og jenter til å si at oppgaven har forskjellig vanskegrad for kjønnene under forutsetningen om at eksamen tester ett konstrukt. Det er spesielt kolonnen lengst til høyre i Figur 14 som er interessant. Denne kolonnen viser forskjellen i utregnet vanskegrad for kjønnene for den aktuelle oppgaven. Et negativt tall svarer til at oppgaven var lettere for jenter enn for gutter, og markeres med fargen rød. Et positivt tall indikerer at oppgaven var lettest for

guttene, og markeres med fargen blå. Sterkere farge betyr større forskjell i vanskegrad enn svakere fargelegging.

Totalt 16 av de 54 deloppgavene viser altså tegn til DIF, og er derfor urettferdige under forutsetningen at eksamen tester ett konstrukt. Av disse 16 oppgavene finnes det oppgaver som favoriserer begge kjønn. Totalt 9 av 16 deloppgaver favoriserer jenter, mens 6 av 16 deloppgaver favoriserer gutter. Én oppgave favoriserer begge kjønn, men i forskjellig ende av ferdighetsspekteret. Fordelingen av oppgaver kan sees i tabellen under.

DIF i favør jenter	DIF i favør gutter	DIF i favør begge
Del 1: 1a, 4a, 5, 6a, 8a, 9, 16 Del 2: 1c, 3c,	Del 1: 17b, 20 Del 2: 1b, 5b, 5c, 7	 Del 2: 2c

Hvis man tar hensyn til utslagene for alle oppgavene som viser tegn til DIF kan man regne ut oppførselen til prøven som helhet. Antallet oppgaver som favoriserer jenter er større enn antallet oppgaver som favoriserer gutter. I tillegg til dette er effekten av DIF ofte større i oppgavene som favoriserer jenter, så disse oppgavene påvirker den totale oppførselen i større grad enn oppgavene som favoriserer gutter.



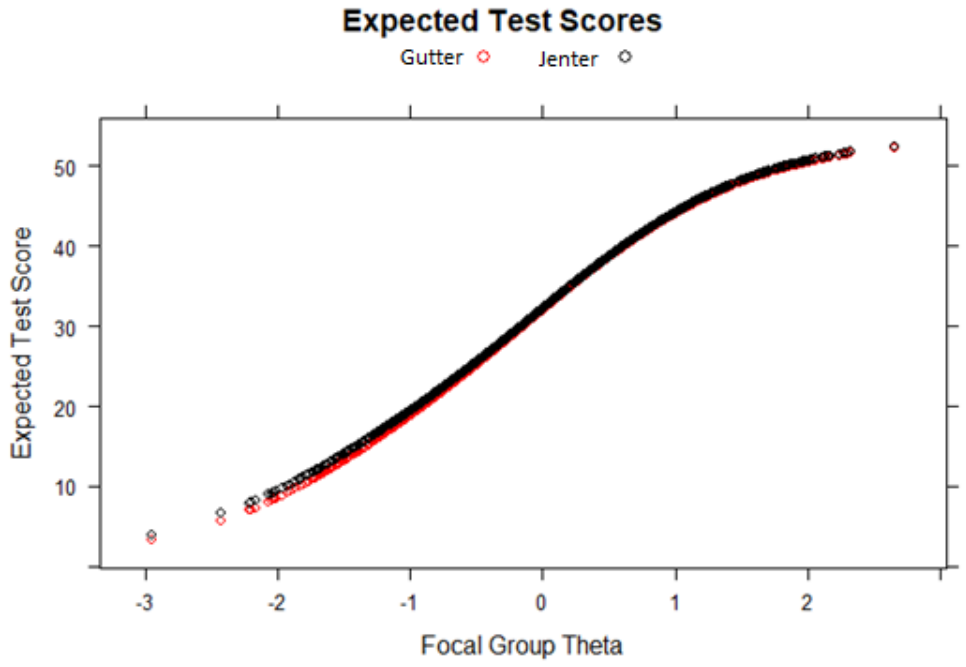
Figur 15

Som et eksempel inkluderes her utregnet ICC fra én av eksamens oppgaver. Figur 15 viser resultatet av analysen for oppgave 1a fra eksamen Del 1. Fra grafen kan man lese at jenter på ferdighetsnivå -2

klarer oppgaven omtrent med omtrent 60% sannsynlighet, mens gutter på tilsvarende ferdighetsnivå vil klare den i omtrent 30% av tilfellene. For lavere ferdighetsnivå viser denne oppgaven altså veldig sterk DIF i favør jenter, og man vil forvente at jenter klarer oppgaven oftere enn gutter gjør. For høye ferdighetsnivåer sammenfaller grafene, sterkere elever får til denne oppgaven uavhengig av kjønn. Implikasjonen av dette er at man kan forvente at jenter totalt sett får flere poeng på denne oppgaven enn gutter gjør, og at den dermed favoriserer dem. Denne oppgaven vil vi komme tilbake til i avsnitt 0.

5.4.1 Forventet prøveresultat

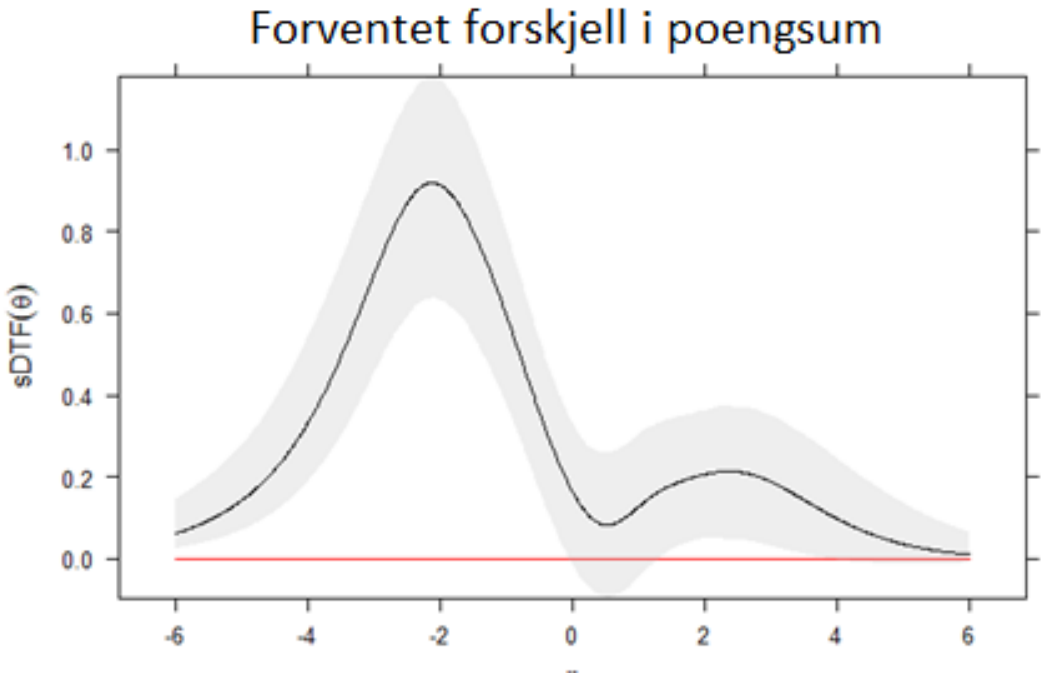
Fordi enkelte av prøvens oppgaver har ulik vanskegrad for kjønnene er det derfor mulig at gutter og jenter ikke forventes å oppnå samme poengsum på eksamen. Oppgaver som er vanskeligere for ett av kjønnene vil totalt bidra til færre poeng for dem enn det for det andre. Med de utregnede vanskegradene for gutter og jenter kan vi se hva forventet poengsum vil være for de to gruppene. Det er denne forventede poengsummen man kan se i *Figur* 16. Den svarte grafen viser forventet poengsum for jenter, og for lavere ferdighetsnivåer kan det sees at jenter forventes å få høyere resultat.



Figur 16: Forventet score

Så hvor stor er forskjellen i forventet poengsum? Rundt ferdighetsnivå -2, som er i lavere ende av sjiktet og svarer til karakterskillet 1/2 er differansen omtrent ett poeng. For høyere ferdighetsnivåer er det også en svak forventet forskjell, men den utgjør ikke nok til å anta at det vil føre til en forskjell

i poeng tildelt av sensor. Figur *17* viser forskjellen i forventet poengsum basert på grafen i Figur 16 med tilhørende usikkerhetsmarginer markert i grått.



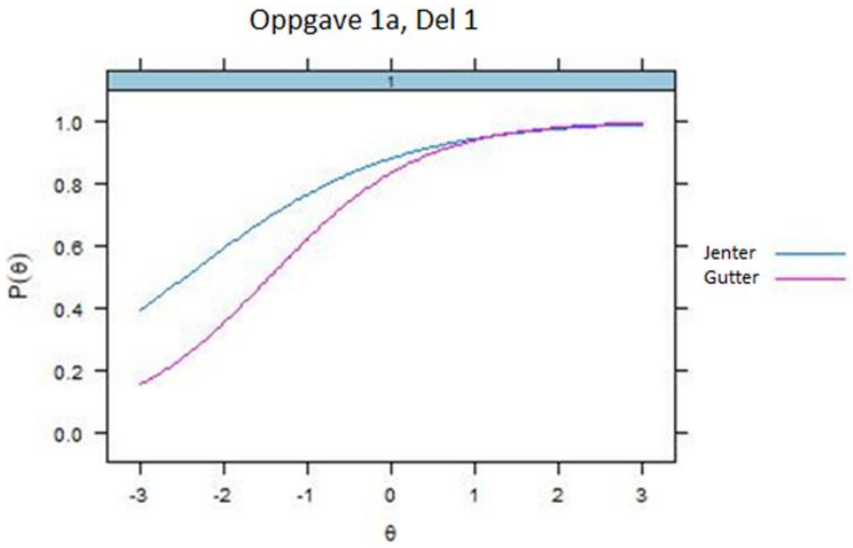
Figur 17: Forskjell i forventet poengsum

Selv om effekten ikke er større enn ett poeng for lavere ferdighetsnivå er det verdt å nevne. Dette er elever som ikke mestrer matematikk, og det ene poenget kan utgjøre forskjellen mellom to karakterer. Dette har implikasjoner for både deres egen mestringsfølelse i faget, og for videre skolegang. Etter denne oppsummeringen av den totale effekten av DIF i oppgavesettet skal vi nå se på noen enkeltoppgaver som gjorde seg bemerket i analysen.

5.5 Resultater av DIF fra et utvalg oppgaver

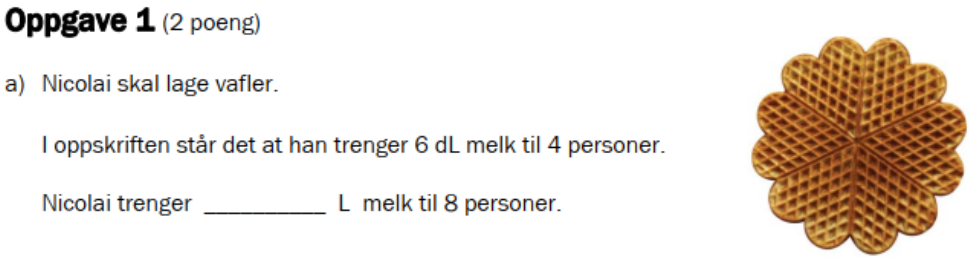
I denne seksjonen vil enkelte av oppgavene fra eksamen presenteres. Oppgavene er blant utvalget som viser sterkest tegn til DIF, og er derfor de mest interessante å se på. Alle oppgavene vil presenteres i samme format. Først vises en graf som presenterer ICC for både gutter og jenter. Deretter presenteres oppgaveteksten, relevante tall fra DIF-analysen og hvilke kompetanser oppgaven krever. Som avsluttende kommentar vil mulige årsaker til DIF diskuteres.

5.5.1 Oppgave 1a, Del 1



Figur 18 Oppgave 1a, Del 1

Figur 18 ble presentert tidligere i resultatseksjonen, og viser ICC for jenter og gutter på for den aller første oppgaven i eksamenssettet. Denne har en av de sterkeste utslagene av DIF blant hele oppgavesettet. Selve oppgaven fra eksamen kan sees under i Figur 19.



Oppgave 1 (2 poeng)

a) Nicolai skal lage vafler.

I oppskriften står det at han trenger 6 dL melk til 4 personer.

Nicolai trenger _____ L melk til 8 personer.

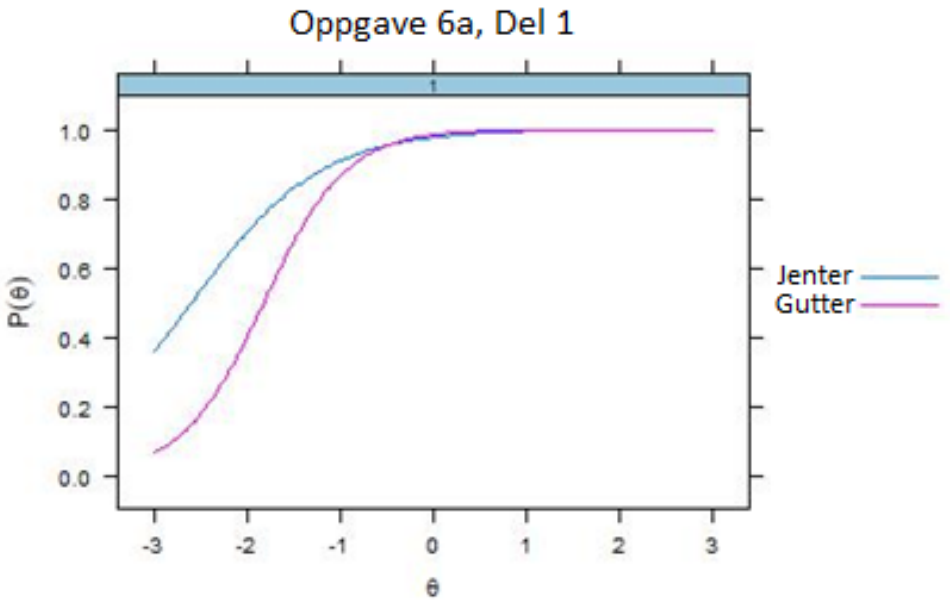
Figur 19: 1a, Del 1 (Utdanningsdirektoratet, 2019d)

Oppgaven var blant de lettere oppgavene på eksamen. Første steg av DIF-analysen viser at den har en vanskegrad på -1.95, som plasserer den godt under gjennomsnittlig vanskegrad. Etter siste steg ble det derimot utregnet en vanskegrad på -1.44 for gutter, og -2.46 for jenter. Dette betyr at for elever på middels og høyt ferdighetsnivå er det minimale forskjeller, og begge kjønn klarer oppgaven med tilnærmet lik sannsynlighet. For den lavere halvdel av skalaen viser grafen derimot tydelige forskjeller. I karakterområdet 1-2, som svarer omtrent til ferdighetsnivå -2, er sannsynligheten for at en jente klarer oppgaven dobbelt så høy som for gutter.

Min analyse av kompetansekrav for denne oppgaven er følgende. Oppgaven krever strategiutvikling. Selv om oppgaven er relativt enkel må elevene bruke den oppgitte informasjonen for å finne ut hvor mye røre som kreves for åtte personer. De to mest sannsynlige strategiene er å bruke den oppgitte mengden som kreves for fire personer for å regne ut hvor mye røre hver person trenger, for så å multiplisere med åtte, eller ved å se at en dobling av antall personer vil føre til et behov for dobbelt så mye røre. Oppgaven krever også matematisering da det innebærer å tolke en modell av en reell situasjon.

Tallene man regner med er oversiktlige, og språket er lite komplisert, så analysen konkluderte med at andre kompetanser ikke er nødvendige. At få kompetanser er nødvendig er ikke overraskende med tanke på oppgavens vanskegrad. Hvorfor er da DIF så kraftig for denne oppgaven? Ett element ved oppgaven som ikke har blitt nevnt foreløpig er enhetsomregningen i svaret. En dobling av 6 dL gir 12 dL, men oppgaven ber eleven oppgi svaret i liter. Sensorveiledningen for oppgaven sier at svaret 12 ikke gir uttelling, men tas med i helhetsvurdering av besvarelsen (Utdanningsdirektoratet, 2019c). Dette betyr at elever som ikke leser oppgaven grundig nok til å se at oppgitt informasjon og svaret har forskjellige enheter ikke vil få poeng på denne oppgaven. Resultatene fra PISA-undersøkelsen 2018 viser at jenter gjør det statistisk signifikant bedre enn gutter i leseforståelse (OECD, 2019). Elevene som tok eksamen i MAT0010 våren 2019 er ikke av samme årskull som elevene som gjennomførte PISA-undersøkelsen, men jeg mener det er forsvarlig å anta at resultatene er tilsvarende for disse elevene som er ett år yngre. En mulig forklaring av den store forskjellen mellom kjønnene er altså at gutter ikke vil få med seg at svaret skal ha en annen enhet enn den oppgitte informasjonen, og derfor ikke vil få uttelling for besvarelsen sin. Hvis denne hypotesen stemmer er det en indikasjon på at oppgaven krever kommunikasjonskompetanse selv om oppgaveanalysen vurderte den til å ikke gjøre det, og at andre oppgaver med informasjon som er vanskelig å tolke kan vise liknende tegn til DIF.

5.5.2 Oppgave 6a, Del 1



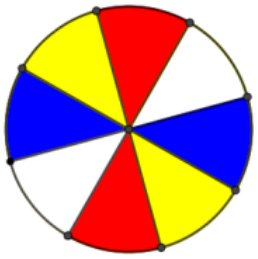
Figur 20: Oppgave 6a, Del 1

Figur 20 viser resultatet av DIF-analysen for oppgave 6a, Del 1. Resultatene viser at den har tydelige tegn til DIF for lavere ferdighetsnivå på samme vis som oppgave 1a, Jenter har altså høyere sannsynlighet enn gutter for å løse oppgaven. Oppgaveteksten kan sees i Figur 21.

Oppgave 6 (2 poeng)

Et lykkehjul har 8 like store felt

- 2 røde
- 2 gule
- 2 blå
- 2 hvite



a) Bestem sannsynligheten for at lykkehjulet stopper på et rødt felt.

Svar: _____

Figur 21: 6a, Del 1 (Utdanningsdirektoratet, 2019d)

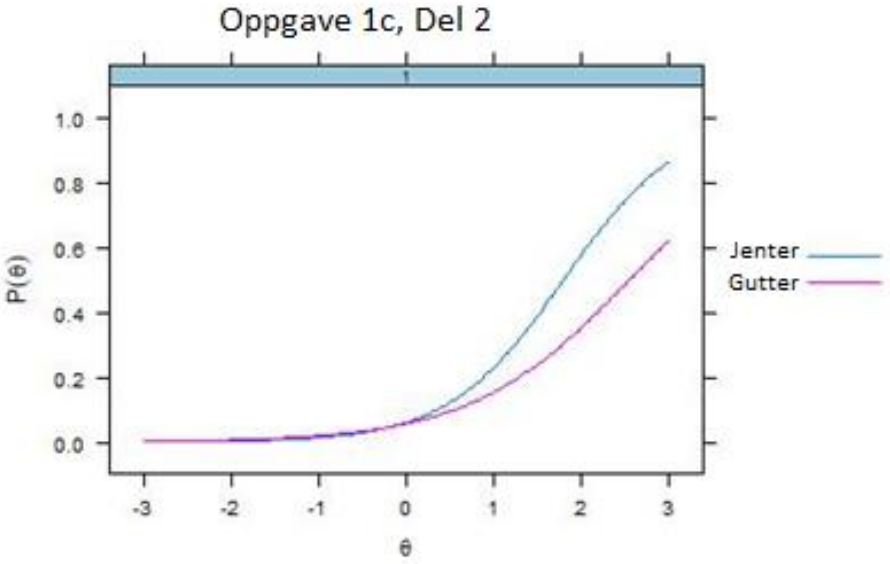
Denne oppgaven var blant de letteste på hele eksamenssettet med en vanskegrad på -2.2. Det er derfor ikke overraskende at forskjellen mellom ICC for gutter og jenter er størst på lave

ferdighetsnivå, da så godt som alle elever på høyere nivåer vil ha klart oppgaven. For gutter er utregnet vanskegrad -1.83, og for jenter er den -2.61.

Min analyse av oppgaven tilsier at den krever kommunikasjonskompetanse. Teksten er ikke vanskelig, og det er ikke informasjon som er direkte irrelevant, så man kan argumentere for at kommunikasjonskompetanse ikke er nødvendig. Det er allikevel mer informasjon enn i forrige eksempeloppgave, og tilstedeværelsen av bildet gjorde at den ble vurdert til å trenge kommunikasjonskompetanse. Oppgaven ble også vurdert til å trenge matematisering, da en enkel modell for sannsynlighet må konstrueres for å kunne besvare oppgaven. Ingen andre kompetanser ble vurdert til å være nødvendige.

I motsetning til oppgaven i Figur 19 gis det her full uttelling for flere måter å oppgi svaret. Hvis oppgaven er besvart med $\frac{1}{4}, \frac{2}{8}, 25\% \text{ eller } 0{,}25$ vil eleven få full score. Dette gjør det vanskeligere å komme med en hypotese til hvorfor oppgaven viser sterkere tegn til DIF. En mulighet er at oppgaver som omhandler sannsynlighet er vanskeligere for gutter, men, hvis dette var tilfelle er det merkelig at vi ikke ser en tilsvarende DIF i neste deloppgave. Denne oppgaven omhandler også sannsynlighetsregning, og er basert på samme lykkehjulet som oppgave 6a. De to oppgavene har forskjellig svarformat, 6a besvares med fritekst mens 6b er flervalg, som muligens kan utgjøre en forskjell.

5.5.3 Oppgave 1c, Del 2



Figur 22

Figur 22 er et eksempel på en oppgave som viser tegn til DIF i øvre ende av skalaen. Oppgaveteksten er vist under. Inkludert i oppgaven er en tabell som viser arealet av de fem største landene i verden, men for oppgaven er kun arealet til Russland og Brasil relevant.

c) Hvor mange prosent større er arealet av Russland enn arealet av Brasil?

Figur 23: 1c, Del 2 (Utdanningsdirektoratet, 2019d)

Oppgaven er en av de vanskeligste oppgavene i hele eksamenssettet, med en vanskegrad på 2.11, og det er derfor vi kun ser forskjeller for høyere ferdighetsnivå. Utregnet vanskegrad for jenter er 1.81, for gutter er den 2.57. Min analyse av kompetansekrav for oppgaven er følgende:

Strategiutvikling er nødvendig. Oppgaven krever at eleven utvikler en strategi for å bruke de oppgitte tallene for å regne ut svaret, selv om utregningsmetoden er forholdsvis enkel. Representasjonskompetanse er også krevd for å klare oppgaven, hovedsakelig fordi relevant informasjon for å løse oppgaven må finnes i tilhørende tabell. Det er også nødvendig med SOF-kompetanse i løsningsarbeidet, da oppgaven inkluderer arbeid med brøker og omgjøring av desimaltall til prosent. Det bør merkes at oppgaven finnes i del 2 av eksamen, og det kan derfor argumenteres for at tilgangen til hjelpemidler ikke gjør denne kompetansen nødvendig. Tilstedeværelsen av irrelevant informasjon (arealet til land som ikke er Brasil eller Russland) gjør at det også kan argumenteres for at kommunikasjonskompetanse er nødvendig, men da all informasjon finnes i samme tabell ble representasjonskompetanse vurdert som mer passende.

I motsetning til tidligere oppgaver har ikke denne oppgaven blitt vurdert til å kreve kommunikasjonskompetanse, men måten oppgaveteksten er formulert kan være en mulig forklaring til at denne oppgaven viser tegn til DIF i favør jenter. Arealene til henholdsvis Russland og Brasil er oppgitt som 17 098 240 og 8 515 770 kvadratkilometer. Korrekt svar på denne oppgaven er 101%, Russland er omtrent dobbelt så stort som Brasil. Jeg mener at denne oppgaven kan mistolkes på to forskjellige måter. Hvis man kun deler Russlands areal på Brasils areal, og dermed regner ut hvor stort Russland er i forhold til Brasil får man følgende utregning.

$$\frac{17\ 098\ 240}{8\ 515\ 770} \approx 2.01 = 201\%$$

Omvendt kan man regne ut hvor mange prosent av Russlands areal Brasil utgjør ved å regne som følger.

$$\frac{8\ 515\ 770}{17\ 098\ 240} \approx 0.50 = 50\%$$

Begge disse, men kanskje spesielt den førstnevnte, er mulige mistolkninger av oppgaveteksten. En elev som tolker oppgaven på denne måten vil regne ut hvor mange prosent av Brasils areal Russland utgjør, som er 201%. Kommunikasjonsaspektet ved oppgaven kan derfor være en mulig forklaring til DIF i favør jenter. Sensorveiledningen spesifiserer at et svar på 201% ikke gir uttelling, men tas med i helhetsvurderingen av besvarelsen, som tyder på at denne måten å lese oppgaven på er en vanlig feil.

6 Diskusjon

Det er nå på tide å ta opp trådene fra innledningen av oppgaven, og å knytte disse trådene opp mot både teorien som ligger i grunn for oppgaven, forskningsspørsmålene som prøver å belyse problemene vi forsøker å undersøke, og hvordan resultatene kan brukes for å besvare disse. Oppgavens overordnede mål har vært å undersøke tre aspekter ved validiteten til resultatene fra eksamen i MAT0010 våren 2019. Vi vil undersøke om slutningene som trekkes fra eksamensresultatene trygt kan benyttes for sitt formål, å si noe om eksamenstagernes ferdigheter i matematikk. I denne seksjonen vil vi ta opp de forskjellige trådene til validitetsbegrepet, og si noe om hvilke implikasjoner resultatene har for disse aspektene ved validitet.

6.1 Innholdsvalidering

Innholdsvaliditet, som er beskrevet i mer detalj i seksjon 2.4.1, var et mål på om innholdet eksamen tester samsvarer med innholdet i læreplanen, og dette er noe som generelt undersøkes kvalitativt av eksperter (Newton & Shaw, 2014). Analysen i denne oppgaven forsøkte å undersøke dette ved å se på alle deloppgaver fra eksamen ved hjelp av en forenklet utgave av skjemaet for matematisk kompetanse utviklet for PISA-undersøkelsen, slik det er beskrevet i Turner et al. (2015). Resultatene fra denne undersøkelsen, se seksjon 5.1, er det primære grunnlaget for å kunne si noe om innholdsvaliditeten til eksamen i MAT0010 våren 2019. Kompetanseskjemaet fra PISA ble valgt som verktøy for undersøkelsen fordi det er en operasjonalisering av konstruktet matematisk kompetanse som er representativt for de grunnleggende ferdighetene i læreplanen, og spesielt med de nye kjerneelementene (se seksjon 2.2.3).

Før vi diskuterer resultatene bør det minnes om at det ikke har blitt undersøkt om kompetansemålene i læreplanen er representert. Det antas allerede at det er godt samsvar mellom kompetansemålene og eksamensoppgavene, da de utvikles av en eksamensnemd som er eksperter i læreplanen. Resultatene fra undersøkelsen er lovende. Av de seks kompetansene som ble analysert var alle representert i eksamensoppgavene. Dette tyder på at eksamen tester matematisk kompetanse i sin helhet som konstrukt, og bidrar dermed til at innholdsvaliditeten styrkes. I det pågående arbeidet med å utvikle nye eksamensoppgaver og eksamensformat beskrives valideringsprosessen som skal ligge bak de nye eksamenene (Utdanningsdirektoratet, 2019b), men at det er manglende forskning rundt teamet. Resultatene fra denne analysen vil dermed bidra til påstanden om at eksamensoppgavene fra 2019 representerer de nye kjerneelementene i fagfornyelsen. Det er allikevel verdt å nevne at de forskjellige kompetansene ikke var like godt representert. Symbol- operasjon og formalismekompetansen, som beskriver det mer mekaniske

arbeidet med matematikkoppgaver, er sterkere representert enn de andre, og bidrar til inntrykket av at oppgavene i stor grad virker veldig like hverandre. Resonneringskompetanse, spesielt, var lavere representert enn de andre kompetanseaspektene, og særlig i oppgaver av lav og middels vanskelighet. Av de ni oppgavene hvor resonneringskompetanse ble testet, var fem av dem blant de vanskeligste i eksamenssettet, og dette illustrerer en annen problemstilling ved innholdsvaliditeten til eksamen. Svakere elever har ikke samme mulighet til å demonstrere denne kompetansen hvis det er få oppgaver som tester dette på lavt nivå.

Det er mulig at resonneringskompetanse generelt krever høyere grad av tankeprosesser, og at oppgaver som tester denne kompetansen generelt vil være vanskeligere. Dette samsvarer med funnene til Turner et al. (2013), som fant at resonneringskompetanse var den kompetansen som ga de beste prediksjonene for ferdighetsnivå i sin undersøkelse av PISA-oppgaver fra undersøkelsen i 2012. Det skal likevel være mulig å lage enklere oppgaver som aktiverer denne kompetansen. At svakere elever har færre muligheter til å demonstrere sin resonneringskompetanse er altså uheldig fra et equitysynspunkt.

6.2 Konstruktvalidering

All validering er konstruktvalidering ifølge Messick (1989), så denne seksjonen burde muligens hatt et annet navn. Det er likevel passende for denne oppgaven, for i denne seksjonen vil diskusjonen rundt resultatene fra den endimensjonale IRT-analysen finnes; det er denne analysen som i hovedsak benyttes som grunnlag for undersøkelsen av dimensjonaliteten til konstruktet eksamenssettet tester. Analysen var gjort under antagelsen av at alle oppgavene testet samme konstrukt, eller egenskap, hos elevene, og tilpasset en modell for oppgavenes vanskegrad under denne antagelsen. Disse resultatene passet veldig godt med de observerte resultatene fra prøven, med en EAP/PV reliabilitet på 0.941. Hva betyr egentlig dette? Det viktigste for leser å ta med seg videre er at dette indikerer at eksamen tester et endimensjonalt konstrukt. Sagt med andre ord tester én eksamensoppgave i hovedsak den samme egenskapen som alle de andre oppgavene tester. Hva enn den egenskapen er betyr det jo at den testes grundig, men da må vi stille spørsmål ved hva denne egenskapen som testes faktisk er.

Den kvalitative analysen indikerte at alle kompetanser var representert i oppgavesettet, men at det var en generell overvekt av kompetansene knyttet til strategiutvikling og SOF. Til tross for dette viste analysen at de forskjellige dimensjonene i det overordnede konstruktet matematisk kompetanse var tilstedeværende som helhet, og at det var flere oppgaver som testet hver kompetanse. Den kvantitative analysen ga derimot resultater som pekte i motsatt retning. Analysen i ConQuest tyder

på at konstruktet som testes er sterkt endimensjonalt, og at oppgavene tester samme kompetanse. I beskrivelsen av kodingsprosessen ble det fortalt hvordan koderne syntes mange av oppgavene virket like hverandre, selv om kompetanseskjemaet fra Turner et al. (2015) kunne tolkes som om at de testet forskjellige kompetanser, og den endimensjonale analysen støtter dette inntrykket vi fikk under kodingen. En flerdimensjonal analyse kunne også ha vist en god tilpasning til de observerte resultatene, så det kan ikke bestemmes at eksamen er like endimensjonal som det her argumenteres for, men med tanke på hvor godt modellen passer virker det lite sannsynlig. Jeg vil også nevne igjen at det ikke forventes at høyt nivå i én dimensjon av matematisk kompetanse (f. eks. strategiutvikling) er uavhengig av nivåene hos andre kompetanser. Niss & Jensen (2002) og Kilpatrick et al. (2001) presiserer begge at de forskjellige aspektene ved matematisk kompetanse ikke bør sees på som isolerte fra de andre, men at de alle er med på å bygge opp en mer sammensatt ferdighet. Jeg vil derfor prøve å være forsiktig med uttalelsene mine om hva resultatene sier om validiteten ved prøven, men derimot poengtere at det meritterer ytterligere undersøkelser. En mulig forklaring er at kompetanseskjemaet brukt i undersøkelsen er for snevert. De forskjellige nivåene av kompetanseaktivering ble slått sammen under analysen, og det er derfor ikke mulig å skille mellom oppgaver som aktiverer kompetanser på høyt nivå og lavt nivå. Hvis rammeverket fra Turner et al. (2015) hadde blitt brukt i sin helhet mistenkes det at symbol, operasjon og formalismekompetanse hadde dominert i større grad, og at dette hadde gitt et sterkere inntrykk av et endimensjonalt konstrukt.

Med forrige paragraf i bakhodet vil vi nå diskutere hva disse tilsynelatende motstridige resultatene kan indikere når vi ser på prøvens validitet. Matematikk skal som alle andre fag gi én karakter, og det skal være et mål på din overordnede kompetanse i faget. Vi har tidligere diskutert (se seksjon 2.2) hvordan det er en overordnet enighet om at matematikkompetanse er et flerdimensjonalt konstrukt, med flere aspekter, og dette må representeres i fagets vurderingsformer. At resultatene tyder på en endimensjonal test må derfor tolkes som en mulig trussel mot prøvens validitet, og dette er noe som bør undersøkes grundigere i arbeidet med utvikling av nye eksamensoppgaver. Spesielt med tanke på fagfornyelsen er dette viktig å ta med seg videre, da den nye læreplanen vektlegger det flerdimensjonale aspektet ved matematisk kompetanse enda sterkere enn forrige læreplan. Da det har blitt avgjort at eksamen i matematikk nå skal bli heldigital (Utdanningsdirektoratet, 2020), og dette kan medføre store endringer i hvordan oppgavene ved eksamen vil ende opp med å se ut, er det enda viktigere at dette utforskes ytterligere. De nye oppgavene bør være en god representasjon av den nye læreplanen, både av kompetansemålene og kjerneelementene.

6.3 Validering fra et konsekvensperspektiv

Det tredje av forskningsspørsmålene fra seksjon 3 var «Eksisterer det kjønnsbasert bias i oppgavene gitt på eksamen i MAT0010 våren 2019?», og er inkludert i denne oppgaven for å teste denne dimensjonen av valideringsprosessen. Validitet er et mål på om *slutningene* som trekkes på grunnlag av eksamen har belegg i prøvens resultater, og det er kanskje funnene tilknyttet dette spørsmålet som har størst implikasjoner for arbeidet med fremtidig utvikling av matematikkeksamen.

Når vi snakker om konsekvenser for elevene som tar eksamen snakker vi egentlig om equity. Nortvedt and Buchholtz (2018) diskuterer equity fra et vurderingsperspektiv. Det er sentralt for equity at alle elever får mulighet til å demonstrere sin matematiske kompetanse; eller sagt med andre ord, at de får mulighet til å vise hva de kan. Fra dette synspunktet vil jeg argumentere for at det er to mulige problemstillinger som må undersøkes, og begge disse kan besvares med resultater fra analysene i denne oppgaven. *Den første* er nært knyttet innholdsvaliditet som begrep. Det ble nevnt i seksjon 2.4.1 at de to primære truslene mot innholdsvaliditet er konstruktirrelevans og konstruktunderrepresentasjon. Av disse mener jeg underrepresentasjon av konstruktet har størst mulig innflytelse på manglende muligheter for å vise sin kompetanse i faget. Hvis deler av læreplanen har en forholdsmessig lav tilstedeværelse i eksamensoppgavene vil dette påvirke enkelte elever mer enn andre, og de får ikke samme mulighet til å demonstrere sin kompetanse. *Den andre* av disse er ønsket om at det også, av samme årsak, skal finnes tilstrekkelig med oppgaver av alle vanskegrader. En svak elev vil sannsynligvis ikke få til vanskelige oppgaver, og prøven bør ha nok lette og middels vanskelige oppgaver slik at man har nok informasjon til å kunne gi dem en rettferdig vurdering. Den kvalitative kompetanseanalysen indikerer at det er oppgaver som tester alle kompetanseaspektene. Selv om de ikke er like godt representert er de alle tilstedeværende. Det er også tilstrekkelig med oppgaver for alle vanskegrader. Figur 9 viser hvordan oppgavene fra eksamen fordeler seg langs de ulike estimerte vanskegradene. Man kan se i figuren at det finnes oppgaver langs hele spekteret, og at spredningen faktisk er svakt forskjøvet mot den lettere siden. Hvis noe er det litt få vanskelige oppgaver, men i det store og hele er fordelingen av oppgaver god. Det eneste å kommentere her er det som har blitt nevnt tidligere om hvordan oppgavenes kompetansekrav fordeler seg på ulike vanskeligheter. Blant lettere og middels vanskelige oppgaver er det relativt få oppgaver som krever resonneringskompetanse, så svakere elever har færre muligheter for å vise denne kompetansen.

Det er et større problem at det er så tydelig tilstedeværelse av DIF i eksamensoppgavene. Dette kobler tilbake til teorikapittelet om bias i prøver (seksjon 2.5), og kan være et tegn på at prøven er konstruert på en slik måte at enkelte grupper vil ha en høyere sannsynlighet for å lykkes enn andre. Resultatene fra DIF-analysen indikerer at 16 av prøvens 54 deloppgaver viser tegn til slik oppførsel,

og at det totalt sett ser ut til at prøven favoriserer jenter på lavere kompetansenivåer. Disse resultatene i seg selv betyr ikke at prøven er urettferdig, men de burde være nok til å anerkjenne en mulig trussel mot prøvens validitet. Antagelsen som ligger bak DIF-analysen er at prøvens konstrukt er endimensjonalt (noe den første IRT-analysen støtter), og at det under denne antagelsen er høy sannsynlighet for at gutter med lavt ferdighetsnivå vil gjøre det dårligere enn jenter med tilsvarende ferdighetsnivå på enkelte oppgaver. Oppgavene som viser størst tendens til dette er allerede diskutert i detalj i seksjon 0, og det er disse oppgavene som bidrar i størst grad til den totale effekten som vist i Figur *17*. Forskjellen i forventet poengsum for de to kjønnene varierer langs ferdighetsspekteret, og vi kan se at for svakere elever er summen av effekten fra alle oppgavene så høyt som et helt poeng. Dette poenget kan utgjøre forskjellen mellom to karakterer, og det er en validitetstrussel som bør tas alvorlig. Slik eksamen utvikles i Norge i dag er dette ikke reelt mulig å unngå. Måten DIF oppdages er ved å teste oppgaver hos tilstrekkelig antall elever, og eksamensoppgaver testes ikke ut i stor skala før de utleveres på den faktiske eksamensdatoen. Siden det utvikles nye eksamensoppgaver som skal passe med de nye heldigitale formatene for eksamen i disse dager er dette muligens det mest relevante resultatet å nevne fra denne oppgavens undersøkelser.

At oppgaver viser DIF er ikke alene nok til å påstå at prøven er urettferdig. Martinková et al. (2017) poengterer at tilstedeværelse av DIF bør tolkes som at oppgaver tester en annen egenskap enn hva som primært testes av prøven som helhet, og at denne egenskapen kan være relatert til prøvens konstrukt. Resultatene fra analysen er ikke omfattende nok til å kunne si hva årsaken til DIF i oppgavene er, men det er noen fellestrekk som kan være relevante. To av oppgavene er formulert på en slik måte at de kan mistolkes hvis man ikke leser oppgaveteksten grundig nok, eller misforstår hva den spør om. Dette faller inn under kommunikasjonsaspektet av matematisk kompetanse. Som nevnt tidligere viser de siste resultatene fra PISA at jenter scorer høyere enn gutter i leseforståelse. Kan manglende kompetanse i lesing være grunn til at gutter har lavere forutsetning enn jenter for å klare disse oppgavene? De var i utgangspunktet ikke vurdert til å kreve kommunikasjonskompetanse, da de inneholder lite tekst, men de krever allikevel at du tolker oppgavens tekst på én spesifikk måte for å få uttelling for svaret. Spesielt oppgave 1a fra del 1, som ikke gir uttelling med mindre man får med seg at oppgaven krever at svaret oppgis i enheten liter, straffer de som kanskje forstår problemet og utregningen, men ikke har finlest teksten. Er dette dermed en urettferdig oppgave? Kommunikasjonskompetanse er en del av konstruktet i vurderingsskjemaet, og det står også stadfestet i de nye kjerneelementene, så det er ikke urimelig at oppgaver kan inkludere vanskelig tekst. Fra et equityperspektiv er dette en vanskelig problemstilling. Hvis språk er en kilde til DIF, og dette fører til at noen gjør det dårligere på eksamen i matematikk,

betyr det at de straffes dobbelt. Både i språkfag og i matematikkfag. Det er også mulig at denne DIF-tendensen hadde vært enda sterkere hvis en analyse hadde blitt gjort på basis av morsmål. Her er problemstillingen bak equity muligens enda større, da det kan være en kilde til systematisk forskjellsbehandling mellom minoritetselever og andre.

6.4 Implikasjoner for praksis og videre forskning

Denne oppgaven er lite et bidrag til forskningen rundt eksamen i matematikk, men som i de fleste undersøkelser har det blitt funnet flere nye spørsmål enn svar på de vi allerede hadde. Oppgaven har sett på flere forskjellige aspekter av validitet, og det som mangler er grundigere undersøkelser av de forskjellige aspektene.

Vi har oppdaget DIF i flere av eksamenens oppgaver, og påpekt problemene dette medfører. Det vi ikke kan si noe om er årsakene bak tilstedeværelsen av DIF. En mulig hypotese var språket i oppgaven. Dette er bare én mulighet, og er også en gjetning. Grundigere undersøkelser av senere eksamener burde gjennomføres for å finne ut om dette er et generelt problem, og for å si noe mer om hvorfor det skjer. Det skal ikke være slik at noen grupper har større forutsetning for å lykkes på eksamen enn andre, og dette funnet bør tas alvorlig.

Kompetansekrav i oppgaver er også noe som kan undersøkes nærmere. Analysen i denne oppgaven tyder på at matematisk kompetanse er representert i alle aspekter, men det er flere svakheter ved metoden. Vi benyttet et forenklet skjema som grunnlag for analysen, og kun to kodere deltok. Dette gir resultater av begrenset reliabilitet, og det er ikke sikkert en annen gruppe hadde funnet samme resultater. Vi har heller ikke vurdert hvordan tilgangen til hjelpemidler i oppgaver på eksamens Del 2 har påvirket nødvendige kompetanser. Denne mangelen blir spesielt relevant når eksamen går over til et heldigitalt format, og elever som ikke har samme digitale grunnlag som andre kan få større problemer med eksamensstrukturen.

Funnene av oppgaver som utviser DIF har også implikasjoner for vurderingspraksis i Norge, og hvis dette er et problem med større omfang enn denne oppgaven har undersøkt må det diskuteres om måten eksamen utvikles må endres. Eksamen lages av profesjonelle matematikere og matematikklærere med god kunnskap om læreplan og fagfelt, og dette er ikke ment som en kritikk mot dem, men oppgaver må testes for å undersøke om de fungerer forskjellig for ulike grupper. Dette gjøres ikke i dag, men kan være nødvendig for å sikre likeverdig behandling av elevene i skolen.

7 Avslutning

Som avsluttende kommentar følger her en oppsummerende besvarelse av forskningsspørsmålene med et par ekstra tanker.

Første forskningsspørsmål spør om oppgavene gitt ved eksamen i MAT0010 våren 2019 er en god representasjon av kompetanseskjemaet fra Turner et al. (2015). Analysen viser at alle kompetanseaspekter er representert i eksamensoppgavene, og vi kan dermed konkludere med at eksamen er en god representasjon av rammeverket. Denne konklusjonen må tas med forbehold om at en forenklet versjon av rammeverket ble benyttet i analysen, og at en mer omfattende undersøkelse kan avdekke mer informasjon om graden av representasjon for hver enkelt kompetanse.

Den kvantitative analysen med ConQuest viser at en endimensjonal analyse av besvarelsesmønstrene gir en veldig god tilnærming. Dette indikerer at eksamen tester et endimensjonalt konstrukt, og besvarer det andre forskningsspørsmålet. Dette samsvarer med inntrykket om at oppgavene var veldig like hverandre, og danner dermed et bilde som stiller seg i noe konflikt med resultatene fra den kvalitative kodingen av oppgavenes kompetanser.

Til slutt stilte vi spørsmålet om det eksisterer kjønnsbasert bias ved oppgavene gitt ved eksamen. Resultatene fra DIF-analysen tyder på at det gjør det, og at jenter vil ha noe større forutsetning for å lykkes på eksamen enn gutter. Vi kan ikke direkte konkludere med at prøven derfor er urettferdig. Tilstedeværelsen av DIF tyder på at disse oppgavene tester en annen egenskap enn eksamen som helhet, men det er et bekymrende funn som burde utforskes ytterligere.

Er eksamensresultatene valide? Sannheten er at jeg ikke kan svare på det. Validitet er et komplekst nok begrep til at denne oppgaven ikke er tilstrekkelig til å komme med en konklusjon. Definisjonen fra AERA, APA, &, NCME (2014) poengterer at resultatene må være gyldige for prøvens formål. Jeg vil påstå at prøven tester matematisk kunnskap bredt nok til at resultatene kan forsvares som grunnlag for karakterer, men at forskjellen mellom kjønnene bør tas alvorlig og forskes videre på. Jeg håper at funnene fra analysene er med på å bidra til den faglige diskusjonen rundt vurderingsformene i skolen, og kan inspirere til forskning for å øke troverdigheten til disse i fremtiden.

8 Litteraturliste

American Educational Research Association., American Psychological Association., National Council on Measurement in Education., & Joint Committee on Standards for Educational and Psychological Testing (U.S.). (2014). *Standards for educational and psychological testing*.

Bakeman, R., McArthur, D., Quera, V., & Robinson, B. F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods, 2*(4), 357.

Baker, F. B., & Kim, S.-H. (2017). *The basics of item response theory using R*: Springer.

Banks, K. J. A. M. i. E. (2006). A comprehensive framework for evaluating hypotheses about cultural bias in educational testing. *19*(2), 115-132.

Blum, W. (2006). *Bildungsstandards Mathematik: konkret: Sekundarstufe I: Aufgabenbeispiele, Unterrichtsanregungen, Fortbildungsideen*: Cornelsen Verlag Scriptor.

Blömeke, S. (2013). Validierung als Aufgabe im Forschungsprogramm „Kompetenzmodellierung und Kompetenzerfassung im Hochschulsektor “. *Validation as a Task in the Funding Initiative 'Modeling and Measuring Competencies in Higher Education'](KoKoHs Working Papers, 2). Berlin and Mainz: Humboldt-University and Johannes Gutenberg-University*.

Blömeke, S., Gustafsson, J.-E., & Shavelson, R. J. (2015). Beyond dichotomies. *Zeitschrift für Psychologie*.

Burkhardt, H. (2014). Curriculum Design and Systemic Change. In Y. Li & G. Lappan (Eds.), *Mathematics Curriculum in School Education* (pp. 13-34). Dordrecht: Springer Netherlands.

Burkhardt, H., & Schoenfeld, A. (2018). Assessment in the service of learning: challenges and opportunities or Plus ça Change, Plus c'est la même Chose. *ZDM, 50*(4), 571-585.

Camilli, G., & Shepard, L. A. (1994). *Methods for identifying biased test items* (Vol. vol.4). Thousand Oaks, Calif: Sage.

Embretson, S. (2019). *Explanatory Item Response Theory Models: Impact on Validity and Test Development?*, Cham.

Hamilton, L. S., Stecher, B. M., Marsh, J. A., McCombs, J. S., & Robyn, A. (2007). *Standards-based accountability under No Child Left Behind: Experiences of teachers and administrators in three states*: Rand Corporation.

Kane, M. T. (2006). Validation. I R. L. Brennan (Ed.), *Educational measurement* (4th ed., s. 17-64). Washington, DC: The National Council on Measurement in Education & the American Council on Education.

Kilpatrick, J. (2014). Competency Frameworks in Mathematics Education. In S. Lerman (Ed.), *Encyclopedia of Mathematics Education* (pp. 85-87). Dordrecht: Springer Netherlands.

Kilpatrick, J., Swafford, J., & Findell, B. (2001). The strands of mathematical proficiency. Adding it up: Helping children learn mathematics (s. 115–155). In: Washington, DC: National Academy Press.

Kunnskapsdepartementet. (2013). *Læreplan i matematikk fellesfag (MAT1-04).* Hentet fra https://www.udir.no/kl06/MAT1-04#. Lastet ned 24.10.202.

Kunnskapsdepartementet. (2019). *Læreplan i matematikk 1.–10. trinn (MAT01-05).* Hentet fra https://www.udir.no/lk20/mat01-05?lang=nno. Lastet ned 30.09.2020.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159-174.

Leder, G. C., & Forgasz, H. J. (2018). Measuring who counts: gender and mathematics assessment. *ZDM, 50*(4), 687-697. doi:10.1007/s11858-018-0939-z

Lind Pantzare, A. (2018). *Dimensions of validity: studies of the Swedish national tests in mathematics.* Umeå universitet,

Martinková, P., Drabinová, A., Liaw, Y.-L., Sanders, E. A., McFarland, J. L., & Price, R. M. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education, 16*(2), rm2.

Meijer, J. J. A., stress, & Coping. (2001). Learning potential and anxious tendency: Test anxiety as a bias factor in educational testing. *14*(3), 337-362.

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational researcher, 18*(2), 5-11.

Newton, P. E., & Shaw, S. D. (2014). *Validity in educational & psychological assessment*. Thousand Oaks, Calif: Sage.

Niss, M. (2015). Mathematical competencies and PISA. In *Assessing mathematical literacy* (pp. 35-55): Springer.

Niss, M. A., & Jensen, T. H. (2002). *Kompetencer og matematiklæring: ideer og inspiration til udvikling af matematikundervisning i Danmark*: Undervisningsministeriets forlag.

Nortvedt, G. (2013). Are girls or boys better at mathematics? A commentary on the game of reporting gender differnces. *Proceedings of the International Groups for the Psychology of Mathematics Education*, 385-392.

Nortvedt, G. A., & Buchholtz, N. (2018). Assessment in mathematics education: responding to issues regarding methodology, policy, and equity. *ZDM, 50*(4), 555-570. doi:10.1007/s11858-018-0963-z

OECD (2019), PISA 2018 Results (Volume II): *Where All Students Can Succeed*, PISA, OECD Publishing, Paris, https://doi.org/10.1787/b5fd1b8f-en. Lastet ned 20.09.2020

Pettersen, A. (2019). Towards competency-oriented mathematics education: An investigation of task demands and teachers' knowledge of task demands from a competency perspective (Doktoravhandling). Universitetet i Oslo, Oslo.

Pettersen, A., & Braeken, J. (2019). Mathematical competency demands of assessment items: A search for empirical evidence. *International Journal of Science and Mathematics Education*. 17(2), s. 405-425. doi: 10.1007/s10763-017-9870-y

Pettersen, A., & Nortvedt, G. A. (2018). Identifying competency demands in mathematical tasks: recognising what matters. *International Journal of Science and Mathematics Education, 16*(5), 949-965.

Popham, W. J. (1997). Consequential validity: Right concern-wrong concept. *Educational measurement: Issues and practice, 16*(2), 9-13.

Reynolds, C. R., Livingston, R. B., Willson, V. L., & Willson, V. (2010). *Measurement and assessment in education*: Pearson Education International Upper Saddle River.

Rousseau, C., & Tate, W. F. (2003). No time like the present: Reflecting on equity in school mathematics. *Theory into Practice, 42*(3), 210-216.

Shimizu, Y., Kaur, B., Huang, R., & Clarke, D. (2010). " The Role of Mathematical Tasks in Different Cultures". In *Mathematical Tasks in Classrooms around the World*. Leiden, The Netherlands: Brill | Sense. doi: https://doi.org/10.1163/9789460911507_002

Suurtamm C. et al. (2016) Assessment in Mathematics Education. In: Assessment in Mathematics Education. ICME-13 Topical Surveys. Springer, Cham. https://doi.org/10.1007/978-3-319-32394-7_1

Turner, R., Blum, W., & Niss, M. (2015). *Using competencies to explain mathematical item demand: A work in progress*.

Turner, R., Dossey, J., Blum, W., & Niss, M. (2013). Using mathematical competencies to predict item difficulty in PISA: A MEG study. In *Research on PISA* (pp. 23-37): Springer.

Utdanningsdirektoratet. (2019a). Analyse - grunnskolepoeng og karakterer i grunnskolen 2018-19. Hentet fra https://www.udir.no/tall-og-forskning/finn-forskning/tema/analyse-av-grunnskolepoeng-og-karakterer-for-grunnskolen-skolearet-2018-19/. Lastet ned 20.06.2020

Utdanningsdirektoratet. (2019b). Kunnskapsgrunnlag for evaluering av eksamensordningen. Hentet fra https://www.udir.no/tall-og-forskning/finn-forskning/rapporter/Kunnskapsgrunnlag-for-evaluering-av-eksamensordningen/. Lastet ned 23.09.2020

Utdanningsdirektoratet. (2019c). Sensorveiledning MAT0010 Matematikk. Hentet fra

https://sokeresultat.udir.no/eksamensoppgaver.html#?k=MAT0010&query=mat0010%20sensorveiledning. Lastet ned 10.02.2020

Utdanningsdirektoratet. (2019d). Eksamen MAT0010 Matematikk. Hentet fra https://matematikk.net/side/Eksamensoppgaver. Lastet ned 15.01.2020

Utdanningsdirektoratet. (2020). Eksempeloppgaver i matematikk T. Hentet fra https://www.udir.no/eksamen-og-prover/eksamen/eksempeloppgaver/matematikk-eksempeloppgaver/. Lastet ned 15.11.2020

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach*. Melbourne: Educational Measurement Solutions.

Zumbo, B. D. (1999). A handbook on the theory and methods of differential item functioning (DIF). *Ottawa: National Defense Headquarters*, 1-57.

9 Vedlegg

9.1 Vedlegg A – Program for DIF-analyse

```
library(mirt)

mydata <- read.csv2('MAT0010 V19 anonymisert (1).csv')
mydata[3:56][mydata[3:56] == 9]<- 0
mydata[3:56][mydata[3:56] == 2]<- 1
mydata[3:56][mydata[3:56] == 3]<- 1
apply(mydata[3:56], 2 , table, exclude = NULL)


data1 <- subset(mydata, KjÃ¸nn == "G")
data2 <- subset(mydata, KjÃ¸nn == "J")
ref <- data2
foc <- data1

# Input
reference_group <- ref[3:length(ref)]
focal_group <- foc[3:length(foc)]
technical = list(NCYCLES = 5000)
number_initial_anchor_items <- 5
#####
# Sorting input
# Constructing data
mydata <- rbind(focal_group, reference_group)
# Defining groups
group <- c(rep('1', nrow(focal_group)), rep('2', nrow(reference_group)))
item_names <- colnames(mydata)
shared_items <- intersect(colnames(reference_group), colnames(focal_group))
shared_items_position <- match(shared_items,item_names)
non_shared_items <- item_names[!item_names %in% shared_items]

#####
# Step 1
### Estimating a fully constrained model
model_full_constrain <- multipleGroup(mydata,
                                      1,
                                      group = group,
                                      technical = technical,
                                      #SE = TRUE,
                                      itemtype = '2PL',
                                      invariance = c(item_names, 'free_means', 'free_var'))
jpeg("model_initial.jpeg", width = 6000, height = 4000)
model_initial <- plot(model_full_constrain, type="trace")
print(model_initial)
dev.off()
coef(model_full_constrain, IRTpars = T)
write.csv2(coef(model_full_constrain, IRTpars=T, simplify = T), "model_full_constrain.csv")
### Next two lines are diagnostics - comment out if unvwanted
#fs <- fscores(model_full_constrain)
#write.csv2(fs, "fs_test2.csv")

#####
# Step 2
# Estimating models with one item unconstrained at the time

j = c(1:length(shared_items))
anova_DF_step2 <- data.frame()
for (i in j) {
  # Designates item to unconstrain.
  item_to_unconstrain <- shared_items_position[i]
  print(paste0("Step 2, moodel ", i, " of ", length(j), ". Item ",
               item_names[item_to_unconstrain], " unconstrained"))
  itemS_to_constrain <- item_names[!item_names %in% item_names[item_to_unconstrain]]
  # Runs model with 1 unconstrained item.
  mymodel <- multipleGroup(mydata,
                           1,
                           group = group,
                           technical = technical,
```

```
                    #SE = TRUE,
                    itemtype = '2PL',
                    invariance = c(itemS_to_constrain,
                            'free_means', 'free_var'))
  anova_comp <- anova(model_full_constrain,mymodel, verbose = FALSE)
  anova_df <- anova_comp[2,]
  anova_DF_step2 <- rbind(anova_DF_step2, anova_df)
  rownames(anova_DF_step2)[i] <- shared_items[i]
  # five next lines creates item trace line for unconstrained item. Comment out if unwanted.
  plot_name <- paste0(item_names[item_to_unconstrain], "_step2.jpeg")
  jpeg(plot_name, width = 450, height = 300)
  item_plot <- itemplot(mymodel, item_to_unconstrain, theta_lim = c(-3, 3))
  print(item_plot)
  dev.off()
  # Two next lines gives diagnostic info. Comment out if unwanted
  #print(anova_DF_step2[i,])
  #print(anova(model_full_constrain,mymodel))
  # Two next lines are progress update.
  print(paste(round(100*i/length(j),1), "% of step 2 completed"))
  print("----------------------------")
}
# Item parameters of fully constrained model
items_coef <- coef(model_full_constrain, simplify = TRUE)$'1'$items
# Combining item paramters for shared items with liklihood/ratio test of items
shared_items_stats <- cbind(
  items_coef[rownames(items_coef) %in% shared_items,],
  anova_DF_step2[rownames(anova_DF_step2) %in% shared_items,])

#####
# Step 3
# Select the number chosen initially by "number_initial_anchor_items" items with
# the highest discrimination parameter that does not excibit DIF (maxA items)
anchors <- rownames(subset(shared_items_stats,p >=0.05)
              [order(subset(shared_items_stats, p >=0.05)$a,
                   decreasing = TRUE), ])[1:number_initial_anchor_items]
anchor_items_position <- match(anchors, item_names)

#####
# Step 4
# Estimating model with the anchor items identified

# Combines anchor and non-shared items to have correct number of free paramters.
constrained_items <- c(anchors,non_shared_items)
# Estimates a base-line model with the five anchor items constrained to be equal
model_anchored <- multipleGroup(mydata,
                        1,
                        group = group,
                        technical = technical,
                        SE = TRUE,
                        itemtype = '2PL',
                        invariance = c(constrained_items,
                                'free_means', 'free_var'))

# Item trace overview
jpeg("model_anchored.jpeg", width = 3000, height = 2000)
model_anchor_plot <- plot(model_anchored, type="trace")
print(model_anchor_plot)
dev.off()
# Gives an excel file with model coeficiencies for diagnostic purpose. Comment out if unwanted
coef(model_anchored, IRTpars = T)
# Next line is for diagnostics - comment out if unwanted
write.csv2(coef(model_anchored, IRTpars = T, simplify = T), "model_anchored.csv")
# Plotting item trace of shared non-anchor items. Comment out if unwanted.
unconstrained_items <- item_names[!item_names %in% constrained_items]
j <- c(1:length(unconstrained_items))
for (i in j) {
  plot_name <- paste0(unconstrained_items[i], "_step3.jpeg")
  jpeg(plot_name, width = 450, height = 300)
  item_plot <- itemplot(model_anchored, unconstrained_items[i],theta_lim = c(-3, 3))
  print(item_plot)
  dev.off()
```

```
}

#####
# Step 5
# Estimating models with anchor + 1 item constrained at the time
j = c(1:length(unconstrained_items))
anova_DF_step5 <- data.frame()
for (i in j) {
  item_to_constrain <- unconstrained_items[i]
  # Following three lines are for updating purpose
  print(item_to_constrain)
  print(paste("Step 5, moodel", i, "of", length(j), ". Item",
          item_to_constrain, "constrained"))
  # Defines the items that are to be constrained for a given loop iteration
  itemS_to_constrain <- c(constrained_items, item_to_constrain)
  # Estimates model with anchor + 1 item constrained
  mymodel <- multipleGroup(mydata,
                     1,
                     group = group,
                     technical = technical,
                     #SE = TRUE,
                     itemtype = '2PL',
                     invariance = c(itemS_to_constrain,
                               'free_means', 'free_var'))
  # mirt-anova function does a liklihood ratio test.
  anova_comp <- anova(model_anchored,mymodel, verbose = FALSE)
  # "slice" is a tidyverse function that selects a specific row in a data.frame
  anova_df <- anova_comp[2,]
  # Builds an overview data.frame of the liklihood ratio test.
  anova_DF_step5 <- rbind(anova_DF_step5, anova_df)
  rownames(anova_DF_step5)[i] <- item_to_constrain
  # Two next lines gives diagnostic info. Comment out if unwanted
  #print(anova_DF_step5)
  #print(anova(model_anchored,mymodel))
  # Two next lines are progress update.
  print(paste(round(100*i/length(j),1), "% of step 5 completed"))
  print("----------------------------")
}

#####
# Step 6
# Identifying invariant and non-invariant items

# Selects items that are invariant with a p-value >= 0.05
invariant_items <- c(anchors, rownames(anova_DF_step5[anova_DF_step5$p >=0.05/length(unconstrained_items),]))
invariant_items
length(invariant_items)
# Identifies the non-invariant anchors
non_invariant <- shared_items[! shared_items %in% invariant_items]
non_invariant

#####
# Final model
# Calculates a model with the items identified to be invariant as constrained between groups
model_final <- multipleGroup(mydata,
                      1,
                      group = group,
                      technical = technical,
                      SE = TRUE,
                      itemtype = '2PL',
                      invariance = c(c(invariant_items, non_shared_items),
                                'free_means', 'free_var'))
# Writing .csv with model coefs.
#write.csv2(coef(model_final, simplify = TRUE, IRTpars = T), "model_final.csv")
coef(model_final, IRTpars = T)

# Item trace overview
jpeg("model_final.jpeg", width = 6000, height = 4000)
model_final_plot <- plot(model_final, type="trace")
print(model_final_plot)
dev.off()
```

```
write.csv2(coef(model_final, IRTpars = T, simplify = T), "model_final.csv")
write.csv2(anova_DF_step5, "anova_DF_step5.csv")

# Plotting item trace of shared non-anchor items. Comment out if unwanted.
j <- c(1:length(non_invariant))
for (i in j) {
  plot_name <- paste0(non_invariant[i], "_Final.jpeg")
  jpeg(plot_name, width = 450, height = 300)
  item_plot <- itemplot(model_final, non_invariant[i],theta_lim = c(-3, 3))
  print(item_plot)
  dev.off()
}
```

9.2 Vedlegg B – MEG-skjema for innholdsanalyse

Beskrivelser av kompetanser oversatt fra beskrivelsen i Turner et al. (2015, s.110-114), nivåinndelingen beskriver den modifiserte utgaven brukt i denne oppgavens analyse.

Kommunikasjonskompetanse:
Denne kompetansen har både en mottagende og konstruerende komponent. Den mottakende komponenten inkluderer å forstå hva som blir sagt og vist relatert til oppgavens matematiske mål, inkludert matematisk språk som brukes, hvilken informasjon som er relevant, og hva slags type besvarelse som er forventet. Den konstruktive delen består av å presentere et svar som kan inkludere løsningssteg, beskrivelse av resonnering og rettferdiggjøring av svaret som blir gitt. I skrevne og digitalt gitte oppgaver relaterer mottakende kommunikasjon seg til forståelse av tekst og bilder, både stille og bevegende. Tekst inkluderer verbalt presenterte matematiske uttrykk, og kan også finnes i matematiske representasjoner (slik som aksetitler i diagrammer). Kompetansen dekker ikke kunnskapen som kreves for å løse problemet, eller kunnskap om hvordan informasjon i oppgaven skal brukes.
Definisjon: Lesing og tolkning av utsagn, spørsmål, instruksjoner, oppgaver, bilder og objekter; se for seg og forstå situasjonen som presenteres og ta til seg informasjonen som gis, inkludert matematiske uttrykk som brukes; presentere og forklare sitt matematiske arbeid og resonnement.

- Nivå 0: Forstå korte setninger og fraser som gir umiddelbar tilgang til kontekst. All informasjon er direkte relevant for oppgaven, og informasjon oppgis i en rekkefølge som samsvarer med rekkefølgen det benyttes i løsningen. Konstruktiv kommunikasjon involverer presentasjon av et enkelt ord eller tallsvar.
- Nivå 1: Alt som er mer komplisert enn beskrivelsen i nivå 0 faller inn under nivå 1. Dette nivået beskriver oppgaver hvor relevant informasjon for oppgaven må identifiseres for stegene i løsningsprosessen, og irrelevant informasjon kan være inkludert i oppgaven. Mengden informasjon kan være mer kompleks enn korte setninger og uttrykk. Konstrukt del inkluderer å oppgi sitt svar som en setning eller et sett med utregningssteg.

Strategiutviklingskompetanse:
Denne kompetansen omhandler de strategiske delene av problemløsing i matematikk. Dette inkluderer å velge, eller konstruere, en løsningsstrategi og kontroll over implementeringen av denne. En løsningsstrategi betyr her et sett med steg som til sammen former en måte å løse problemet. Hvert steg har egne delmål på vei mot en helhetlig løsning. Kunnskapen og prosessene som trengs for å faktisk utføre stegene er ikke en del av denne kompetansen. Krav for denne kompetanse øker når omfanget av oppgaven, og dermed antall steg som kreves i strategien, blir større.
Definisjon: Velge eller utvikle en strategi for å løse et problem i tillegg til å overvåke eller kontrollere implementasjon av strategien.

- Nivå 0: Ta direkte handling. Løsningsstrategien er direkte gitt eller åpenbar.
- Nivå 1: Hvis oppgaven krever utvikling av en løsningsstrategi, uansett kompleksitet, ble oppgaven kodet til dette nivået.

Matematisering:
Matematisering omhandler den delen av modelleringssyklusen som kobler ekstra-matematisk kontekst med den matematiske representasjonen. Denne kompetansen har dermed to komponenter. En situasjon utenfor matematikken kan måtte oversettes til et matematisk problem. Dette kan inkludere å ta forenklende antagelser, identifisere størrelser som må representeres i modellen og hvordan de henger sammen. Denne prosessen er det som kalles matematisering, men kompetansen inkludere også tolking av matematiske resultater i kontekst av problemet som løses, og validering av løsningen innenfor rammene til den ekstramatematiske situasjonen.
Definisjon: Oversette en ekstra-matematisk situasjon til en matematisk modell, tolke resultater fra modellen i sammenheng med situasjonen, eller validering av modellens gyldighet.

- Nivå 0: Enten er oppgaven rent matematisk, eller så er sammenhengen mellom den ekstra-matematiske situasjonen og modellen ikke relevant for å løse problemet.
- Nivå 1: Dette nivået representerer alle oppgaver hvor en modell må lages for å løse problemet. Nivået representerer alle former for modeller som må dannes, uavhengig av kompleksitet og om variabler og størrelser er oppgitt eller må identifiseres selv.

Representasjonskompetanse:
Representasjonskompetanse ser på egenskapen til å tolke, hente informasjon fra og danne representasjoner av matematiske sammenhenger og størrelser, eller å koble sammen forskjellige

representasjoner for å finne en løsning på problemet. Representasjoner kan være verbale, fysiske, diagrammer, grafer, tabeller eller figurer. Tolkning av rene tekstoppgaver faller ikke inn under denne kompetansen, men oversettelse fra tekst til andre representasjoner er alltid en del av representasjonskompetanse.
Definisjon: Forstå, oversette mellom og gjøre nytte av matematiske representasjoner; velge eller lage representasjoner som beskriver situasjonen eller for å presentere sine resultater.

- Nivå 0: Enten er ingen representasjoner involvert, eller det er begrenset til å lese enkle verdier fra simple representasjoner. For eksempel å lese en enkel verdi fra en tabell eller graf.
- Nivå 1: Mer kompliserte representasjoner eller mer avansert bruk av enkle representasjoner er nødvendig. For eksempel å sammenligne verdier fra forskjellige representasjoner eller se på hvordan verdier endrer seg over tid i en graf. Oppgaver som krever konstruksjon av representasjoner er også inkludert i dette nivået.

Symbol, operasjon og formalismekompetanse:
Denne kompetansen reflekterer evne til å bruke kunnskap om matematiske områder; slik som definisjoner, regler, algoritmer og prosedyrer, manipulasjon av likninger og formler, og formelle regler og operasjoner. Å sette opp likninger som beskriver en situasjon vil være en del av matematiseringskompetansen. Å løse likningen er derimot en del av symbol, operasjon og formalismekompetansen. Matematiske kunnskapsområder, som å vite formler for areal og omkrets av sirkler, Pythagoras' setning for rettvinklede trekanter, eller hvordan man utfører polynomdivisjon vil også være en del av denne kompetansen.
Definisjon: Forstå og implementere matematiske prosedyrer og språk (inkludert symbolske uttrykk, aritmetiske og algebraiske operasjoner), bruke matematiske konvensjoner og reglene som styrer dem; bruke kunnskap om regler, definisjoner og resultater.

- Nivå 0: Bruk enkle matematiske fakta eller definisjoner; utføre korte aritmetiske utregninger med enkle tall. F. eks. regn ut arealet av et rektangel når sidelengdene er gitt, eller skriv ned formelen for arealet av en trekant.
- Nivå 1: Bruk/manipuler matematiske uttrykk med variabler, eller erstatt variabler med kjente verdier. Utfør utregninger med desimaltall eller brøker. Bruk kunnskap som formelen for omkretsen av en sirkel for å regne ut radius til en sirkel med kjent omkrets. Mer komplisert arbeid med symboler og algebraiske uttrykk faller også inn under dette nivået.

Resonneringskompetanse:
Denne kompetansen omhandler å trekke gyldige slutninger basert på intern prosessering av matematisk informasjon, og til å sette sammen slutningene for å begrunne eller bevise et resultat. Mentale prosesser som er involvert i de andre kompetansene faller inn under disse. Resonneringskompetansen setter mer søkelys på å begrunne hvorfor slutninger som trekkes er gyldige.
Definisjon: Trekke slutninger ved å bruke logiske begrunnede tankeprosesser som utforsker og kobler elementer av problemer for å forme, undersøke eller begrunne argumenter og konklusjoner.

- Nivå 0: Trekk slutninger direkte fra informasjon eller instruksjoner som gis.
- Nivå 1: Trekk slutninger fra resonneringer som krever steg med argumentasjon fra den oppgitte informasjonen, eller ved å sette sammen flere biter informasjonen fra forskjellige sider av problemet.

Ny runde vurderinger gjort I fellesskap	Item	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Kompetanse		1a	1b	2a	2b	3	4a	4b	5	6a	6b	7	8a	8b	9	10	11a	11b	11c	12a	12b	12c	13	14	15	16	17a	17b	18	19	20
Kommunikasjon		0	0	0	0	0	1	1	0	1	1	0	1	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	0	1	1
Strategiutvikling		1	1	1	0	0	0	1	0	0	1	0	0	1	1	1	0	0	1	1	0	0	1	0	1	1	0	0	1	1	0
Matematisering		1	1	0	0	0	0	0	1	1	1	0	0	0	1	0	0	0	0	0	1	0	0	1	0	0	0	0	1	0	0
Representasjon		0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0
Symbol- operasjon og formalisme		0	0	1	1	1	0	1	0	0	1	1	0	1	1	1	1	1	1	0	0	0	1	0	1	1	1	1	0	1	1
Resonering		0	0	0	0	1	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	1	1	0	0	0	0	1
Vanskegrad		-1,23	-0,67	-0,4	1,2	-1,6	-0,7	0,03	0,7	-2,7	1,3	-2	-0,9	-1	-1	-1,38	-0,92	0,91	2,63	-3,62	-0,8	-1,6	-1,4	0,7	2	1,45	-1,85	0,96	-0,6	0,3	1,8
Sum kompetanser		2	2	2	1	2	1	3	2	2	4	1	1	4	3	2	2	2	3	1	1	1	2	2	3	3	1	2	2	4	3

30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54
20	1a	1b	1c	2a	2b	2c	3a	3b	3c	4a	4b	5a	5b	5c	5d	6a	6b	7	8a	8b	8c	9a	9b	9c
1	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	1	1	0
0	0	0	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	0	1	0	1	0	1	1
0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1
0	1	1	1	1	1	0	1	0	0	1	1	0	1	1	1	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	1	1	1	1	1	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
1,8	-1,1	-0,9	3,4	-1	-0,6	1,38	-1,36	1,03	1,2	-1	0,5	-1	0,4	-0,02	3,41	0,07	1,95	-0,2	-1,79	-0,41	3,56	-1	1,56	2,6
3	1	2	3	2	2	2	3	2	3	2	2	2	1	1	2	1	1	2	2	1	4	2	3	3

9.4 Vedlegg D – Resultater fra DIF-analyse

	A	B	C	D	E	F	G
1		Diskriminering Jenter	Vanskegrad Jenter	Diskriminering Gutter	Vanskegrad Gutter	Forskjell Diskriminering	Forskjell Vanskegrad
2	X1a_Eks_Del:	0,80620458	-2,45583182	1,103617619	-1,44484127	-0,297413039	-1,01099055
3	X1b_Eks_Del	1,108933771	-1,18174295	1,108933771	-1,18174295	0	0
4	X2a_Eks_Del:	2,334226733	-0,50579723	2,334226733	-0,50579723	0	0
5	X2b_Eks_Del	1,63249993	0,370582434	1,63249993	0,370582434	0	0
6	X3_Eks_Del1	1,62226527	-1,59768917	1,62226527	-1,59768917	0	0
7	X4a_Eks_Del:	1,896268541	-0,95961958	2,106992513	-0,65137741	-0,210723972	-0,30824217
8	X4b_Eks_Del	1,488388225	-0,50531693	1,488388225	-0,50531693	0	0
9	X5_Eks_Del1	1,000169838	-0,26252461	1,1170007	0,14304658	-0,116830862	-0,40557119
10	X6a_Eks_Del:	1,436872504	-2,60690389	2,257951118	-1,83065826	-0,821078614	-0,77624564
11	X6b_Eks_Del	1,539837649	0,391154184	1,539837649	0,391154184	0	0
12	X7_Eks_Del1	1,892629731	-1,57062133	1,892629731	-1,57062133	0	0
13	X8a_Eks_Del:	1,280680609	-1,53661841	1,27067903	-0,96927952	0,010001579	-0,56733889
14	X8b_Eks_Del	1,317951636	-1,33953658	1,317951636	-1,33953658	0	0
15	X9_Eks_Del1	1,402819976	-1,55326994	1,096884719	-1,09469625	0,305935257	-0,45857368
16	X10_Eks_Del:	1,198510885	-1,71848872	1,198510885	-1,71848872	0	0
17	X11a_Eks_De	1,334655797	-1,26232156	1,334655797	-1,26232156	0	0
18	X11b_Eks_De	1,09332421	0,133246902	1,09332421	0,133246902	0	0
19	X11c_Eks_De	2,2679041	1,191499472	2,2679041	1,191499472	0	0
20	X12a_Eks_De	1,673000922	-2,7925505	1,673000922	-2,7925505	0	0
21	X12b_Eks_De	1,269961218	-1,22414762	1,269961218	-1,22414762	0	0
22	X12c_Eks_De	1,225664177	-1,80428244	1,225664177	-1,80428244	0	0
23	X13_Eks_Del:	1,127105775	-1,74618728	1,127105775	-1,74618728	0	0
24	X14_Eks_Del:	0,978217526	-0,07337245	0,978217526	-0,07337245	0	0
25	X15_Eks_Del:	2,502257726	0,907383899	2,502257726	0,907383899	0	0
26	X16_Eks_Del:	1,037365394	0,4011763	1,159228878	0,765607819	-0,121863484	-0,36443152
27	X17a_Eks_De	2,45274179	-1,42718205	2,45274179	-1,42718205	0	0
28	X17b_Eks_De	2,390653477	0,348329735	2,48269638	0,18708684	-0,092042903	0,161242895
29	X18_Eks_Del:	1,579836646	-0,9287171	1,579836646	-0,9287171	0	0
30	X19_Eks_Del:	2,869747174	-0,06361876	2,869747174	-0,06361876	0	0
31	X20_Eks_Del:	2,856536841	0,748194226	2,48994333	0,547182557	0,366593511	0,201011669
32	X1a_Eks_Del:	1,390298838	-2,29837621	1,390298838	-2,29837621	0	0
33	X1b_Eks_Del	1,491046349	-0,5943212	1,598366241	-0,90176129	-0,107319892	0,307440082
34	X1c_Eks_Del2	1,51337291	1,807550735	1,094938961	2,565456307	0,418433949	-0,75790557
35	X2a_Eks_Del:	0,867307308	-2,10891968	0,867307308	-2,10891968	0	0
36	X2b_Eks_Del	1,431545949	-0,9203797	1,431545949	-0,9203797	0	0
37	X2c_Eks_Del2	0,841599181	0,5506335	1,310433102	0,629573805	-0,468833921	-0,07894031
38	X3a_Eks_Del:	1,35055	-1,52890406	1,35055	-1,52890406	0	0
39	X3b_Eks_Del	2,188118653	0,251166428	2,188118653	0,251166428	0	0
40	X3c_Eks_Del2	2,322288104	0,236091191	2,103562461	0,444597363	0,218725643	-0,20850617
41	X4a_Eks_Del:	2,329289052	-1,65085791	2,329289052	-1,65085791	0	0
42	X4b_Eks_Del	1,32968807	-0,01973343	1,32968807	-0,01973343	0	0
43	X5a_Eks_Del:	2,18626518	-0,14300509	2,18626518	-0,14300509	0	0
44	X5b_Eks_Del	2,120463279	-0,11804241	1,96554854	-0,32451899	0,154914739	0,206476574
45	X5c_Eks_Del2	1,896440134	0,311562549	1,811702129	0,110422083	0,084738005	0,201140466
46	X5d_Eks_Del	1,913103016	1,82997818	1,913103016	1,82997818	0	0
47	X6a_Eks_Del:	1,27571003	-0,40548115	1,27571003	-0,40548115	0	0
48	X6b_Eks_Del	1,6121648	1,009398173	1,6121648	1,009398173	0	0
49	X7_Eks_Del2	1,832411127	-0,84844641	2,127930976	-1,09858462	-0,295519849	0,25013821
50	X8a_Eks_Del:	2,340802531	-0,55478681	2,340802531	-0,55478681	0	0
51	X8b_Eks_Del	3,25761953	0,098461632	3,25761953	0,098461632	0	0
52	X8c_Eks_Del2	1,799396529	2,191279236	1,799396529	2,191279236	0	0
53	X9a_Eks_Del:	2,97296711	-0,09412227	2,97296711	-0,09412227	0	0
54	X9b_Eks_Del	3,703718864	0,760963844	3,703718864	0,760963844	0	0
55	X9c_Eks_Del2	3,283313867	1,316176995	3,283313867	1,316176995	0	0