Morten Riise Klegseth

En vurdering av vurdering

Emneprøver som måleinstrument for matematisk kompetanse

Masteroppgave i matematikkdidaktikk
Trondheim, mai 2018

Norges teknisk-naturvitenskapelige universitet
Fakultet for samfunns- og utdanningsvitenskap
Institutt for lærerutdanning

NTNU
Kunnskap for en bedre verden

Forord

Å skrive denne masteroppgaven har vært en prosess som best kan beskrives som krevende, morsom, utfordrende, slitsom, lystbetont, utmattende, lærerik, frustrerende og givende. Når den nå foreligger ferdigskrevet og klar for innlevering er det på sin plass å takke de som har hjulpet meg på veien.

Aller først vil jeg rette en takk til alle lærerne og elevene som sa seg villige til å bidra til prosjektet mitt. Uten dere hadde ikke denne oppgaven blitt noe av.

En stor takk til min meget dyktige veileder Trygve Solstad for grundige tilbakemeldinger, gode råd og oppmuntrende ord gjennom hele prosessen. Takk også til Eivind Kaspersen for gode workshops og svar på spørsmål som dukket opp underveis.

En takk også til arbeidsplassen min for god støtte og tilrettelegging de siste tre årene.

Til slutt vil jeg rette en ekstra stor takk til min kjæreste Ida som i lange perioder har måttet ta et kjempeansvar på hjemmebane. Jeg gleder meg veldig til å kunne bidra mer igjen, men mest av alt gleder jeg meg nå til masse tid til latter, lek og moro sammen med både deg og de to små røverne våre!

Morten R. Klegseth

Trondheim, mai 2018

Innhold

Figuroversikt

Tabelloversikt

1 Innledning

I mine snart 14 år som lærer har jeg vært med på å utforme mange prøver, og vurdert et tusentalls flere besvarelser. Selv om en etter hvert har opparbeidet seg en god del erfaring er det fortsatt med en porsjon usikkerhet at en forsøker å svare på spørsmålene: *Hvordan vet vi at dette er en god prøve? Hvordan kan vi vite at denne prøven hjelper oss å sortere elevenes kompetanse på en hensiktsmessig måte? Er prøven lett? Vanskelig? Rettferdig?*

De senere årene har ferdig utformede prøver fra forlagene skolene kjøper lærebøker fra blitt mer og mer tatt i bruk, enten i sin helhet eller delvis, uten at dette eliminerer usikkerheten rundt svarene på spørsmålene over. I en spørreundersøkelse tilknyttet evaluering av matematikkeksamen for 10. trinn 2017 rapporterer 85 % av lærerne at de benytter seg av forlagsgitte prøver (Andersen, Fossum, Rogstad & Smestad, 2017, s. 63). Mens det foreligger grundige evalueringer av sentralgitte prøver, slik som for eksempel av tidligere nasjonale prøver (Hopfenbeck, T.N., Ibsen, E., Turmo, A. & Lie, S. (2003), er det lite å finne om kvaliteten til andre typer prøver som benyttes i skolen. Søk i ulike databaser etter studier spesifikt rettet mot egenskaper til de forlagsgitte prøvene har resultert i null relevante treff, og gir således inntrykk av at slike studier til nå ikke er utført. Dette er i så fall uheldig, ettersom det i andre studier er fremkommet at slike prøver benyttes som grunnlag for fastsetting av karakter i matematikkfaget (Prøitz & Borgen, 2010).

I studier som har undersøkt hva lærere legger til grunn for bestemmelse av standpunktkarakterer i matematikk kommer skriftlige tester/prøver ut som det som vektlegges mest, både her til lands (Prøitz & Borgen, 2010) og i andre land (Brookhart, 1994). I norsk skole er det gjerne terminprøver (heldagsprøver) og andre mindre emneprøver som utgjør hovedgrunnlaget for standpunktkarakteren (Prøitz & Borgen, 2010). Kvaliteten til disse prøvenes måleegenskaper vil dermed være av betydning for vurderingen av elevenes matematikkfaglige kompetanse.

En prøves kvalitet forstås som både dens presisjon og anvendbarhet (Wu & Adams, 2007). En emneprøve benyttes gjerne med både et summativt og formativt vurderingsformål, noe som betinger begge de nevnte egenskapene. Med summativ menes at prøven har til hensikt å måle elevenes grad av oppnådde kompetansemål, og gjøres primært ut ifra et rapporteringskrav,

mens formativ vurdering har som formål å virke læringsfremmende, og gis underveis i undervisningsløpet (Harlen, 2006). De to begrepene summativ og formativ forbindes gjerne synonymt med uttrykkene *vurdering av læring* og *vurdering for læring* (Bueie, 2015).

I forkant av innføringen av ny læreplan (LK06) falt godkjenningsordningen for læreverk bort. Kombinert med læreplanens innføring av kompetansemål fremfor innholdsangivelser gir dette stort tolkningsrom for læremiddelutviklere (Juuhl, Hontvedt & Skjelbred, 2010). Ettersom emneprøvene er knyttet til ulike læreverk er derfor grunnlag for å anta at forlagsgitte emneprøver vil kunne være ganske ulike i både utforming og kvalitet.

I denne studien tar jeg for meg oppgaver hentet fra tre forlagsgitte prøver innenfor emnet funksjoner. Funksjoner er et sentralt emne i læreplanen som er nært knyttet til algebra (Sierpinska, 1992) – et annet sentralt emne hvor norske elever ved internasjonale undersøkelser presterer særlig svakt sammenlignet med andre land (Grønmo & Hole, 2017). Videre er funksjoner vanligvis på skolenes årsplaner i 10.klassetrinn som også er tidspunktet for summativ vurdering av sluttkompetanse i den norske læreplanen. Å studere 10.klassingers besvarelser på oppgaver fra emneprøver i funksjoner gir dermed både A) kunnskap om emneprøver i et stort matematisk emne, B) muligheten til å sikre at studien blir gjort omtrent like lenge og relativt kort tid etter undervisningstidspunktet på ulike skoler og C) et nærmest mulig bilde på sluttkompetansen emneprøvene skal vurdere.

Ved å gjennomføre en kvantitativ undersøkelse av besvarelser fra elever i 10.klasse ved flere ulike skoler vil jeg ved bruk av Rasch-analyser forsøke å si noe om oppgavenes kvalitative egenskaper. Hensikten med studien er å undersøke måleegenskapene til de forlagsgitte emneprøvene spesielt, og derigjennom finne svar på de innledende spørsmålene omkring bruk av emneprøver generelt. Undersøkelsen søker å finne svar på følgende spørsmål:

1. Hvordan operasjonaliseres funksjonskompetanse i de forlagsgitte emneprøvene?

2. Hvor egnet er oppgavene til å måle det de er ment å måle?

3. Hvilke tiltak kan gi bedre emneprøver i funksjoner?

2 Kompetanse for funksjoner

Målet for denne studien kan deles i to hoveddeler. I den ene delen undersøkes det hvordan kompetanse innenfor emnet funksjoner forsøkes målt gjennom oppgaver fra emneprøver som er i bruk i Trondheimsskolen i dag. I den andre delen undersøkes det hvordan måleinstrumentet kan forbedres med hensyn på kvalitet og presisjon. På bakgrunn av dette vil det bli nødvendig å redegjøre både for hva som legges i begrepet matematisk kompetanse generelt, samt innenfor funksjoner spesielt. Videre må det redegjøres for sentrale områder innenfor måleteori. Disse to teoretiske rammeverkene er ikke direkte knyttet til hverandre, og har derfor fått hvert sitt teorikapittel.

2.1 Vurdering av kompetanse i et sosiokulturelt perspektiv

Ettersom studiens primære fokus er å analysere skriftlige oppgaver og tilhørende skriftlige besvarelser, er det først nødvendig å avklare noen rammer for hvordan disse bør sees og tolkes. Et sosiokulturelt læringssyn, som tungt vektlegger betydningen av begrepet *mediering* når kunnskap både skal tilegnes og formidles, tilbyr gode rammer for hvordan vi kan forstå kompetansen som søkes målt.

> *Begrepet mediere – som kommer fra det tyske Vermittlung (formidle) - antyder at mennesker ikke står i direkte, umiddelbar og ufortolket kontakt med omverdenen. Tvert i mot håndterer vi den ved hjelp av ulike fysiske og intellektuelle redskaper som utgjør integrerte deler av våre sosiale praksiser (Säljö, 2001, s. 83).*

I et sosiokulturelt perspektiv er altså sosiale og kulturelle aspekter helt essensielle for læring. Den sosiale interaksjonen med andre er det som skaper det selvstendige, tenkende individet, og av definisjonen over ser vi at dette inkluderer bruk av ulike fysiske og intellektuelle redskaper. Dette innebærer at redskapene, herunder både skriftlig og muntlig språk, blir *bindeleddet mellom individets indre tenking og dets ytre kommunikasjon av dets tenking (Säljö, 2001).* I skolesammenheng, som i samfunnet for øvrig, gjelder dette altså gjensidig for både lærer og den lærende, i møte med hverandre, faglitteratur, illustrasjoner m.m.

Det at vi ikke står i direkte, umiddelbar og ufortolket kontakt med omverden betyr at vi ikke har *direkte* tilgang til en elevs kunnskaper og kompetanse. Denne er tett knyttet til elevens

indre tankevirksomhet og er i et sosiokulturelt perspektiv utilgjengelig for oss som måtte ha ønske om å måle den. Det eneste vi kan måle er den kunnskapen og kompetansen som kommer til syne gjennom de medierende redskapene eleven har tilgang til. I denne masteroppgaven, som tar for seg kompetanse innenfor emnet funksjoner, presiseres det derfor at en elevs kompetanse forstås som den kompetansen som *kan observeres* i elevens besvarelser, og ikke den kompetansen eleven måtte ha eller ikke ha utover dette. Videre må vi også forstå oppgavene som benyttes i lys av det sosiokulturelle perspektivet, hvor oppgavenes medierende egenskaper er viktige for i hvilken grad tegn på den ønskede kompetansen kommer til syne i elevens besvarelse. Dette vil være et sentralt kvalitativt aspekt ved en oppgave som det kan bli naturlig å vurdere for oppgaver som ikke fungerer som tiltenkt eller på andre måter ikke oppfyller ønskede krav.

2.2 Matematisk kompetanse

Kompetanse er et vidt begrep, også innenfor matematikk. Flere har gått grundig til verks for å beskrive hva det vil si å inneha matematisk kompetanse og bidrar hver med sine synspunkter, klassifikasjoner og modeller.

Niss & Jensen (2002) legger åtte delkompetanser, inndelt i to hovedgrupper, til grunn i sin beskrivelse av hva det vil si å være matematisk kompetent. De to hovedgruppene er 1) *å kunne spørre og svare i og med matematikk,* og *2) å kunne håndtere matematikkens språk og redskaper*. Hver av hovedgruppene rommer fire av delkompetansene, og gjengis i korte trekk under.

Å kunne spørre og svare i og med matematikk:

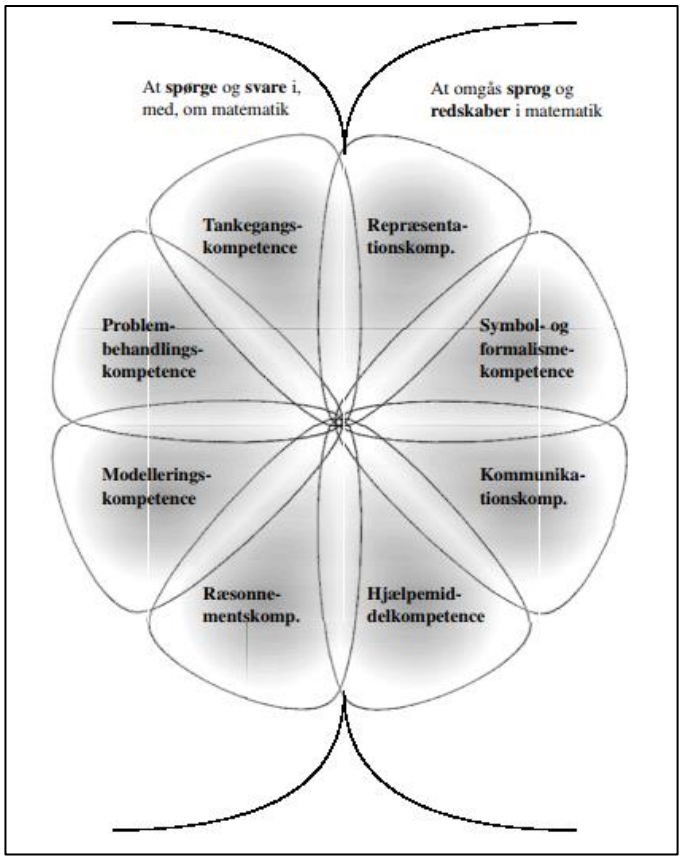
1) *Tankegangskompetanse* innebærer å kjenne til, kunne stille og kunne forutsi svar på spørsmål som er karakteristiske for matematikkfaget. I tillegg innebærer denne kompetansen å kjenne til, å forstå og å kunne håndtere rekkevidden av matematiske begreper, samt ens evne til å håndtere forskjellige typer matematiske utsagn.
2) *Problembehandlingskompetanse* innebærer å kunne oppstille og løse matematiske problemer, det være seg egne formulerte problemer eller de som er formulert av andre.
3) *Modelleringskompetanse* innebærer å kunne analysere eksisterende modeller, samt å kunne bedrive konstruksjon av egne modeller.

4) *Resonnementskompetanse* innebærer på den ene siden å kunne følge og vurdere matematiske resonnementer, herunder forstå hva et bevis er, og på den andre siden å kunne tenke ut og gjennomføre uformelle og formelle resonnementer selv.

Å kunne håndtere matematikkens språk og redskaper:

5) *Representasjonskompetanse* innebærer å kunne forstå og benytte seg av ulike matematiske representasjoner, samt å kunne se forbindelser og å kunne oversette mellom ulike typer representasjoner.
6) *Symbol og formalismekompetanse* innebærer å kunne tolke symbol- og formelspråk, kunne oversette mellom det formelle matematiske språket og naturlig språk, samt å kunne behandle og benytte seg av symbolholdige formler og uttrykk.
7) *Kommunikasjonskompetanse* innebærer å kunne sette seg inn i og tolke andres matematiske kommunikasjon av både skriftlig, muntlig og/eller visuell karakter, samt selv å kunne kommunisere matematikk på forskjellige måter og nivåer overfor et vidt spekter av mottakere.
8) *Hjelpemiddelkompetanse* innebærer både å kjenne til, se muligheter og begrensninger i, samt å kunne benytte seg av hjelpemidler på hensiktsmessige måter ut i fra hva en aktuell situasjon krever.

De forskjellige delkompetansene, så vel som de to ulike hovedgruppene, går naturlig nok inn i hverandre og må således ikke sees på som atskilte og uavhengige av hverandre. For eksempel er representasjonskompetanse og symbol- og formalismekompetanse nært forbundet med hverandre. Likevel vektlegger hver av dem forskjellige sider av matematikkens språk og redskapskasse. I representasjonskompetansen vektlegges selve representasjonen av et matematisk objekt, mens i symbol- og formalismekompetansen er det symbolspråket og de formelle systemers «spilleregler» det legges vekt på. Videre inngår begge disse kompetansene som bestanddeler av en persons kommunikasjonskompetanse, men sistnevnte kompetanse vektlegger bl.a. i større grad sender-mottakeraspekter ved kommunikasjon (Niss & Jensen, 2002). De ulike delkompetansene har altså alle med hverandre å gjøre, og slik utgjør de en helhetskompetanse innenfor matematikk, samtidig som de har klare særtrekk som gjør at de kan betraktes og diskuteres isolert fra de øvrige delkompetansene.



Figur 1: En visuell fremstilling av kompetanseområdene. Hentet fra Niss & Jensen, 2002, s. 45.

2.3 Funksjonskompetanse

En rekke studier er gjort for å kunne si noe om hva funksjonskompetanse er og ulike delkompetanser som begrepet inneholder. En relativt stor andel har sitt fokus på de ulike representasjonsformene funksjoner har, og elevenes utfordringer knyttet til tolkning av og oversettelser mellom forskjellige typer representasjoner (Bossé, Adu-Gyamfi & Cheetham, 2011; Gagatsis & Shiakalli, 2004; Wilmot, Schoenfeld, Wilson, Champney & Zahner, 2011). En av årsakene til dette er at funksjonene kun er tilgjengelige gjennom sine representasjoner og dermed uunngåelig vil være en svært betydningsfull og grunnleggende del av funksjonskompetansen. Et sentralt funn knyttet til oversettelser er at det har mye å si om den aktuelle representasjonen er en kilderepresentasjon eller målrepresentasjon. Med kilderepresentasjon menes den typen representasjon som er utgangspunktet og må tolkes, den som det ved oversettelsesoppgaver skal oversettes *fra*, mens målrepresentasjon er den typen representasjon som det skal oversettes *til*.

Andre, for eksempel Even (1990), Hartter (2009), Zachariades (2002) vektlegger betydningen av en god forståelse for definisjonen av en funksjon, og problematiserer konsekvenser av en mangelfull bevissthet og/eller forståelse av denne. Ytterligere andre, blant annet Sfard (1991) poengterer viktigheten av at funksjonene blir *tingliggjort*, dvs. at de går fra å bli betraktet som en prosess til å bli betraktet som et objekt.

Felles for de nevnte studiene er at de fokuserer på hver sin avgrensede del av det vide begrepet funksjonskompetanse. Som i Niss & Jensens modell av matematisk kompetanse kan disse sees på som ulike delkompetanser innenfor noen større hovedgrupper av funksjonskompetanse. I en studie av Brian R. O'Callaghan (1998) deles kompetanse innenfor emnet funksjoner inn i fire hovedkomponenter: *Modellering, tolkning, oversetting og tingliggjøring.* I tillegg vektlegges det et femte område av betydning; *prosedyrekunnskaper*.

I studien defineres modelleringskompetanse som ens evne til å representere en problemsituasjon ved bruk av funksjoner, og bygger på en bredere definisjon av modelleringskompetanse som evnen til å oversette en situasjonsbeskrivelse til en matematisk representasjon av situasjonen, f.eks. i form av likninger, tabeller og/eller grafer.

Tolkning ansees som den omvendte prosedyren av modellering, altså å tolke matematiske representasjoner opp imot konkrete situasjonsbeskrivelser. Denne inkluderer både lokale tolkninger (individuelle punkter på en graf, konstantleddet i et funksjonsuttrykk, m.m.) og mer globale tolkninger (funksjonsuttrykk/tabell/graf som helhet).

Den tredje komponenten er oversetting, og omhandler ens evne til å gå fra én type representasjon av en funksjon til en annen representasjon. De representasjonsformene det her er snakk om er symboler, tabeller og grafer, de samme tre som er nevnt i avsnittet om modelleringskompetanse, og som Kaput (1989) omtaler som *the three core representational systems.*

Den siste komponenten, tingliggjøring, innebærer at det som en tidligere så på som en prosess eller prosedyre nå betraktes som et mentalt, matematisk objekt. Dette objektet, eller tingen (herav tingliggjøring), betraktes nå som bærer av bestemte egenskaper, og som det videre kan gjøres matematiske operasjoner på, som for eksempel transformasjoner eller sammensetting med andre funksjoner. Komponenten betraktes som et øverste nivå i en hierarkisk modell av konseptuell forståelse (Sfard, 1991).

Forbundet med de fire komponentene som utgjør den konseptuelle funksjonskompetansen finner vi et sett av prosedyrekunnskaper. Dette er kunnskaper som gjør det mulig for elevene å operere innenfor et representasjonssystem. Eksempler kan være algebrakunnskaper eller kunnskaper knyttet til koordinatsystemet, for hhv. algebraiske og grafiske representasjoner.

En slik inndeling av delkompetanser viste seg å være nyttig for denne studien. I emneprøvene var det en overvekt av oppgaver innenfor de tre første hovedkomponentene, i tillegg til noen oppgaver som omhandlet det som omtales som prosedyrekunnskaper. Hva gjelder komponenten *tingliggjøring* var denne vanskelig å inkludere i en modell for kompetanse målt gjennom de skriftlige prøvene som undersøkes her. I emneprøvene var det derimot mange oppgaver som rettet seg mot sentrale begreper innenfor funksjoner (stigningstall, proporsjonalitet, m.m.).

Niss & Jensen (2002) beskriver begrepskompetanse som en viktig del av tankegangskompetansen. Det å kunne redegjøre for betydningen av sentrale fagbegreper er viktig både for å få tilgang til og for å være i stand til å videreutvikle forståelsen av et matematisk konsept. Usiskin (2015) hevder: *Dealing with the written and spoken vocabulary of a concept is an essential part of its understanding that transcends all aspects of that understanding* (Usiskin, 2015, s. 824). Kjennskap til matematikkens vokabular er en forløper til all forståelse hevder han. På bakgrunn av dette utgjør *begrepskompetanse* den fjerde komponenten i kompetansemodellen for funksjoner i denne masteroppgaven.

2.4 Kompetansemodellen og læreplanens kompetansemål

Ettersom emneprøvene jeg undersøker er knyttet til læreverk som alle er skrevet i samsvar med gjeldende læreplan i faget er det naturlig å se på hva læreplanen definerer som funksjonskompetanse. I denne kan vi lese at elever etter 10. trinn skal kunne

- *lage funksjonar som beskriv numeriske samanhengar og praktiske situasjonar, med og utan digitale verktøy, beskrive og tolke dei og omsetje mellom ulike representasjonar av funksjonar, som grafar, tabellar, formlar og tekstar*

- *identifisere og utnytte eigenskapane til proporsjonale, omvendt proporsjonale, lineære og kvadratiske funksjonar og gje døme på praktiske situasjonar som kan beskrivast med desse funksjonane*

  (Utdanningsdirektoratet, 2013).

Sammenligner vi læreplanens kompetansemål med kompetansemodellen som legges til grunn i denne masteroppgaven vil det nok fint være mulig å forsvare at komponenten *modelleringskompetanse* vil kunne være synonym med kompetansen å kunne *lage funksjoner,* og at *tolkningskompetansen* vil kunne være synonym med kompetansen å kunne *beskrive og tolke funksjoner,* samt trolig også å kunne *identifisere egenskaper* til ulike funksjoner. At *oversettingskompetansen* kan sees på som synonym med læreplanens *oversettingskompetanse* vil vel mest sannsynlig være en ukontroversiell påstand. Videre forstår jeg læreplanens mål om å kunne *utnytte egenskapene* til de nevnte typene funksjoner som direkte knyttet til elevenes evne til å modellere og løse ulike problemsituasjoner med funksjoner. På bakgrunn av denne tolkningen vil jeg påstå at også dette kompetansemålet vil kunne relateres til komponenten *modelleringskompetanse,* mens komponenten *begrepskompetanse* kan sees på som et middel, eller verktøy om du vil, for å understøtte utviklingen av de øvrige delkompetansene.

2.5 Representasjonskompetanse

Matematiske objekter er kun tilgjengelige gjennom ulike former for representasjoner, altså gjennom ulike typer tegn, det være seg språklige, ikoniske, symbolske m.m. Dette er særskilt for matematikken sammenlignet med andre fagområder hvor objekter kan observeres direkte (Duval, 2006, s. 107). Innenfor matematikk er det å kunne gjennomføre overganger mellom ulike former for representasjoner en avgjørende terskel for matematisk forståelse for elever, uavhengig av på hvilket trinn i utdanningsløpet de befinner seg (Duval, 2006). En studie utført av De Bock, Van Dooren & Verschaffel (2013) fant at både antall feil og typer feil som elevene gjorde i arbeid med funksjoner var svært avhengig av hvilken representasjonsform av funksjonen elevene måtte forholde seg til. Evnen til å tolke eller konstruere representasjoner, samt å kunne gå fra en type representasjon til en annen, omtales som *«representational fluency»* av Ainsworth, Bibby & Wood (1998), og de antyder at en god konseptuell forståelse

er sterkt knyttet til elevenes evne til å kunne benytte seg av, og forflytte seg mellom, ulike typer representasjoner.

Å inneha god funksjonskompetanse slik den er definert i denne masteroppgaven vil på mange måter være det samme som å inneha god representasjonskompetanse. Å modellere (eller *å lage en funksjon* slik tilsvarende kompetanse er formulert i læreplanen) innebærer å konstruere en eller flere representasjoner av en konkret situasjonsbeskrivelse, å tolke en funksjon vil egentlig være å tolke en eller flere representasjoner av den aktuelle funksjonen og å oversette innebærer å erstatte en representasjon med en annen. Her vil disse representasjonene være avgrenset til tabeller, grafer og algebraiske symboler.

2.6 Grad av funksjonskompetanse

Dersom vi ser funksjonskompetanse og representasjonskompetanse som to sider av samme sak, eller i alle fall forutsetter at grad av utvikling av førstnevnte kompetanse er en forutsetning for grad av utvikling av sistnevnte kompetanse, melder det seg et behov for å klassifisere grader av denne typen kompetanse. I en studie kalt *Students' competencies in working with functions in secondary mathematics education – empirical examination of a competence structure model* (Nitsch et al., 2015) deles de ulike oppgavene elevene i undersøkelsen blir utsatt for inn i ulike kategorier ut ifra hvilke kognitive krav de forskjellige oppgavene betinger. Studien var rettet mot elevers oversetting mellom representasjoner av funksjoner og oppgavene ble klassifisert ut ifra de kognitive handlingene *identifisering, konstruksjon og bekrivelse/forklaring.*

De fant i studien støtte for at en slik tre-delt kategorisering av kognitive handlinger er velegnet til å beskrive elevenes kompetanse i arbeidet med funksjoner. Dette begrunnes blant annet med at identifisering av relevante verdier eller kjennetegn innenfor én eller flere representasjonsformer er grunnleggende for alle andre handlinger, at konstruksjon skiller seg fra identifikasjon fordi en målrepresentasjon ikke er oppgitt, men må bygges, og at det å kunne beskrive/forklare er nødvendig for å kunne oppnå en dypere forståelse av sammenhenger i og mellom ulike typer funksjoner. Fra andre undersøkelser vet vi også at identifiserings- og konstruksjonsoppgaver er vesensforskjellige, hvor den sistnevnte typen stiller større kognitive krav til eleven enn sistnevnte (Bossé, Adu-Gyamfi & Cheetham, 2011; Hattikudur et al., 2012; Leinhardt, Zaslavsky & Stein, 1990)

I studien til Leinhardt et al. (1990) skilles det imidlertid ikke mellom identifisering og konstruksjon, men mellom *tolkning* og konstruksjon. Meningsinnholdet i kategoriene er på mange måter den samme, men ikke spesifikt knyttet til oversettelsesaspektet ved funksjonskompetansen. Tolkning innebærer å skape mening av deler av eller helheten til en representasjon av en funksjon, og konstruksjon kan i tillegg til rene oversettelseshandlinger innebære det å oppgi et eksempel på en funksjon, evt. et eksempel på en bestemt type funksjon (f.eks. lineære, omvendt proporsjonale etc.). Konstruksjon betyr i bunn og grunn at eleven må skape noe utover det som er oppgitt (Leinhardt et al., 1990). Dette meningsinnholdet i begrepene tolkning og konstruksjon legges til grunn i denne studien.

2.6.1 Lokale og globale prosesser

Når det kommer til tolkning av representasjoner skiller litteraturen mellom lokale og globale betraktninger (Duval, 2006; Leinhardt et al., 1990). Med lokale tolkninger menes en punktvis tilnærming til en representasjon, slik at informasjonen en henter ut av representasjonen er i form av enkeltverdier. Den motsatte måten å tilnærme seg en funksjon på kalles for global tolkning. Da er fokuset i stedet rettet mot intervaller eller funksjonen i sin helhet, for eksempel å tolke en grafs generelle form eller intervaller hvor funksjonen øker/minker. Evnen til å gjøre globale tolkninger er viktig for tilgangen til mer avansert matematikk senere i undervisningsløpet (Leinhardt et al., 1990).

Også representasjonsoverganger, eller det som i oppgaven her omtales som oversettelser, har fått betegnelser som lokale og globale (Bossé et al., 2011; Gagatsis & Shiakalli, 2004). Begrepene er sterkt knyttet til det som omtales som lokale og globale tolkninger, der lokale oversettelser er å forstå som oversettelser som kan gjøres punkt for punkt, mens globale krever at en må identifisere samvariasjonen mellom de to variablene. Bossé et al. (2011) har i sin forening av eksisterende litteratur utarbeidet et generelt hierarki over ulike typer oversettelser. Forenklet gjengitt øker oversettelser i vanskegrad i en rekkefølge fra lokale oversettelser (enklest), til oversettelser som i utgangspunktet er globale, men hvor en global oversettelse kan brytes ned til to lokale (middels vanskelig), og til rene globale oversettelser (vanskeligst). Til grunn for denne rekkefølgen ligger det blant annet en erkjennelse av at rekkefølgen har betydning, slik bl.a. De Bock et al. (2013) også påpekte. En oversettelse fra for eksempel et funksjonsuttrykk til en graf er altså vesensforskjellig fra en oversettelse fra

graf til et funksjonsuttrykk, hvor sistnevnte betraktes som generelt vanskeligere enn førstnevnte.

2.6.2 Kvantitative og kvalitative representasjoner

Det er blitt foreslått at elever har en tendens til å innta en punktvis tilnærming til grafiske representasjoner av funksjoner (Hattikudur et al., 2012). Det er også blitt påpekt at hyppigere bruk av kvalitative representasjoner i undervisningen kan være et middel for å trene opp elevenes evne til i større grad å innta en global tilnærming (Goldenberg, 1987, referert i Hattikudur et al., 2012). Kvalitative representasjoner inneholder ingen numeriske verdier, slik kvantitative representasjoner gjør, og dermed vil en punktvis tilnærming gi liten mening når en betrakter grafen opp imot situasjonen den er en representasjon av. Slike representasjoner vil altså naturlig flytte fokus over på variablenes innbyrdes relasjon. Oppgaver som er kvalitative av natur oppleves gjerne forvirrende for elever som ikke er vant med slike, og det har vist seg at langt flere elever unngår å avgi svar på slike oppgaver sammenlignet med kvantitative (Hattikudur et al., 2012).

2.6.3 Antall operasjoner

Det finnes for øvrig flere faktorer som kompliserer eller forenkler en oppgave. Ved studier rettet mot å finne slike faktorer påpekes blant annet antall operasjoner som kreves fram mot en løsning som et generelt kognitivt krav som påvirker vanskegraden til matematikkoppgaver (Hart, 1981 referert i Fisher-Hoch & Hughes, 1996; OECD, 2013). I sammenheng med funksjonsoppgavene som undersøkes i denne studien vil betegnelsen *flere operasjoner* forstås som at eleven må gjøre mer enn én behandling, f.eks. der hvor en lokal tolkningsoppgave krever en transformasjon/omskriving først (evt ikke er direkte observerbar). For konstruksjonsoppgaver forstås betegnelsen som at eleven må gjøre behandlinger utover en direkte oversettelse, eller der hvor en må gjøre samme type operasjon flere ganger.

3 Måling

I den norske læreplanen for grunnskolen er kompetanse innenfor fagene inndelt i en rekke ulike kompetansemål. Et kompetansemål er naturligvis ikke slik at det nås eller ikke nås, men vil av hver enkelt elev oppnås i en eller annen *grad*. For å kunne si noe om graden av oppnådd kompetanse betinger dette naturlig nok en eller annen form for måling. I dette kapittelet vil jeg først ta for meg noen aspekter ved måling generelt, før jeg etter hvert går over til å presentere Rasch-modellen spesielt.

3.1 Måling av fysiske størrelser

Måling av mange forskjellige fysiske fenomener er en så integrert del av vårt daglige liv at det er tilnærmet umulig å forestille seg en verden uten. De er overalt rundt oss. På daglig basis nyttiggjør vi oss av en felles oppfatning av mål på tid, strekning, fart, vekt, temperatur, osv. Årsaken til at vi kan nyttiggjøre oss disse målene, både på individ- og samfunnsplan, er at målene er et resultat av veletablerte og standardiserte instrumenter og skalaer. Denne standardiseringen gjør målinger av ulike objekter og fenomener både pålitelige og lett sammenlignbare. Når du for eksempel på en reise rapporter hjem at det er 32°C der du er, får de hjemme en grei anelse av hvor varmt dette er uten at de trenger å komme til deg for å finne det ut. Videre kan de både avgjøre om dette er varmere eller kaldere enn der hvor de selv er, og også *hvor mye* varmere eller kaldere.

3.2 Måling av psyko-sosiale størrelser

Målinger av ulike psyko-sosiale størrelser er kanskje ikke så veletablerte og integrerte i vårt daglige liv. De er likevel tallrike og eksisterer «overalt rundt oss», og det er blant annet i denne kategorien vi finner størrelsen kompetanse. Utfordringen med målinger av psyko-sosiale størrelser er at egenskapene vi er interesserte i ikke er direkte synlige for oss slik som fysiske objekter er. Det er bare gjennom observerbare indikatorer, såkalte *latente trekk*, på egenskapene at målinger kan gjøres. For eksempel kan søvnløshet og spiseforstyrrelser være symptomer på depresjon. Gjennom observasjoner av symptomer på depresjon kan en utvikle

et måleinstrument og en skala for grader av depresjon. På samme måte må en finne ut hva en person vet og kan gjøre innenfor et fagområde for å kunne komme opp med et mål på personens kompetanse på dette området. En kan ikke «se» fagkompetanse på samme måte som en kan se dimensjonene til et hus. En kan bare måle kompetansen gjennom latente trekk, som for eksempel hvilke oppgaver personen kan utføre (Wu & Adams, 2007, s. 4).

En annen utfordring er at mens det for eksempel er temmelig klart for alle hvilken egenskap ved et fysisk objekt som er blitt målt ved formidlingen av dets lengdemål, er det mindre opplagt hva som er blitt målt dersom en oppgir en persons mål på kompetanse. Denne egenskapen må klargjøres og defineres før den kan måles.

3.3 Ulike nivåer av måling

Av S. S. Stevens (1946) fikk vi en teori om måleskalaer som er blitt både anerkjent og etablert. Her deles ulike måleskalaer inn i fire typer målenivå ut ifra hvor mye informasjon de tilbyr og hvilke regneoperasjoner det er mulig å utføre med dem. De fire nivåene er *nominal, ordinal, intervall* og *forhold*.

Stevens idé var at måling involverer en numerisk modellering av aspekter ved en empirisk verden. Aspektene som modelleres vil variere i kompleksitet og gi opphav til ulike typer skalaer (Michell, 2002, s. 99).. Dersom modelleringen er av klassifiseringer, altså av typen *tilhører en bestemt gruppe*, vil det produseres en nominal skala. Modellering av rangeringsnivåer, av typen *mer enn /mindre enn*, produserer en ordinal skala. Modellering av differanser mellom nivåer av en egenskap, av typen *så mye mer enn /så mye mindre enn*, vil produsere en intervallskala. Mens modellering av forhold mellom nivåer av en egenskap, av typen *så mye større/så mye mindre*, gir oss en forholdsskala.

3.4 Krav til måling

3.4.1 Én-dimensjonalitet og konstruksjonen av en variabel.

Et fundamentalt krav til måling er kravet om én-dimensjonalitet. Dette innebærer at variabelen vi ønsker å måle, for eksempel en spesifikk kompetanse, må kunne sees som en

linje (et kontinuum), og at hvert enkelt mål svarer til et punkt på denne linjen. Når vi tester en persons kompetanse søker vi å beregne hvor på denne linjen personen befinner seg (Wright & Stone, 1979). For å kunne få til dette må vi først definere den kompetansen som utgjør linjen gjennom å konstruere et måleinstrument som tillater at størrelser av den definerte kompetansen kan plasseres langs linjen. Instrumentet vil utgjøres av oppgaver med forskjellig vanskegrad (Andrich, 1989), og oppgavene vil altså både utgjøre instrumentet og definere variabelen.

Det er verdt å bemerke at når vi måler latente størrelser, som for eksempel en spesifikk kompetanse, vil denne trolig bestå av flere ulike dimensjoner. For eksempel har vi i denne studien definert funksjonskompetanse som bestående av flere ulike delkompetanser. Hvis vi imidlertid har lyktes med definisjonen vår slik at den kan betraktes som en egen «enhet» kan denne også betraktes som å utgjøre en én-dimensjonal skala (Linacre, 1998; Sjaastad, 2014). Dersom noen av oppgavene som utgjør måleinstrumentet viser seg å måle en annen egenskap istedenfor, eller i tillegg til, den egenskapen vi søker å måle vil ikke måleinstrumentet lenger tilfredsstille kravet om én-dimensjonalitet. Som en følge kan måleresultatene bli vanskelige å tolke, ettersom det vil være umulig å avgjøre om måleresultatet vi har fått er et produkt av den ene eller den andre egenskapen, eller en kombinasjon av begge.

3.4.2 Additivitet

Et annet krav til måling er additivitet. Dette innebærer at avstanden mellom oppgavene som utgjør måleinstrumentet er konsistent (innenfor verdiene til standardfeilen av målingene). Hvis avstanden mellom måleverdien til en oppgave δ_1 og måleverdien til en oppgave δ_2 har størrelsen d_{12}, og avstanden mellom måleverdien til oppgave δ_2 og en oppgave δ_3 har størrelsen d_{23}, så skal avstanden mellom måleverdien til oppgave δ_1 og oppgave δ_3 være lik $d_{12} + d_{23}$. Hvis dette ikke er tilfelle er konsekvensen at «validiteten til det antatte kontinuumet» avvises, ettersom en variabel beviselig ikke er konstruert (Andrich, 1989, s. 9).

3.4.3 Invarians

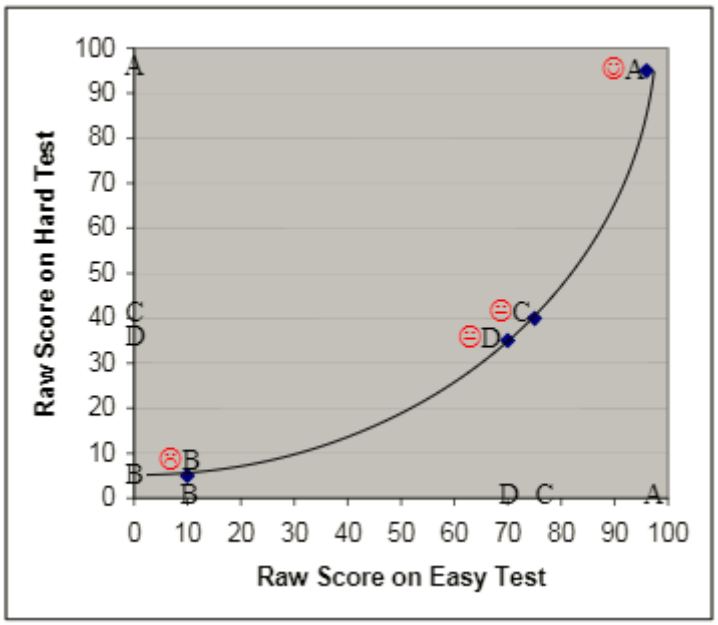
"You can't measure change with a measure that changes" (Bond & Fox, 2001, s. 83).

Et tredje krav til måling er kravet om invarians. Vi må kunne forvente at et måleinstrument produserer forholdsvis nøyaktige målinger, uavhengig av i hvilken kontekst målingen blir gjort (så lenge konteksten samsvarer med intensjonen til målingen). Videre må vi også kunne forvente at ethvert instrument som er ment å måle det samme produserer forholdsvis nøyaktige målinger, uavhengig av hvilket instrument vi velger. For eksempel forventer vi at et termometer vil gi oss rimelig nøyaktige målinger av lufttemperaturen uavhengig av hvilket sted på jorden vi befinner oss, og at vi vil få tilnærmet de samme temperaturmålingene uavhengig av hvilket termometer vi velger å bruke. På samme måten må vi kunne forvente at en test gir rimelig nøyaktige målinger av en persons kompetanse uavhengig av hvem vi tester, og at personens kompetanse er tilnærmet den samme uavhengig av hvilken test vi bruker. Dette betyr igjen at alle personer, uavhengig av kompetansenivå, opplever de lette oppgavene som lette, og de vanskelige oppgavene som vanskelige (Sjaastad, 2014, s.214).

3.5 Klassisk testteori

De fleste prøver som benyttes i skolen i dag er basert på klassisk testteori. Ved bruk av klassisk testteori (CTT) vil oppgavers vanskegrad utelukkende bestemmes ut ifra hvor stor andel av et personutvalg som lykkes med oppgaven. Kompetansen til de som tok testen uttrykkes typisk som hvor stor andel av testens totalscore personen oppnådde. Ulempen med den klassiske testteorien er at elever som besvarer like mange oppgaver med ulik vanskegrad vil få samme resultat. F.eks. vil to elever som på en test besvarer 2 av 5 oppgaver korrekt få samme score selv om de to oppgavene den ene eleven lyktes med var vesentlig enklere enn de to oppgavene den andre eleven lyktes med. Poengsummen vil dermed ikke fullt ut representere elevens egentlige ferdighetsnivå (Utdanningsdirektoratet, 2016). Et annet fundamentalt problem med den klassiske testteorien er at vanskegraden til oppgavene er avhengig av hvilke elever som tok testen, og bestemmelsen av elevenes kompetansenivå er avhengig av hvilke oppgaver testen besto av. Hadde en valgt en annen sammensetning av oppgaver ville muligens en elevs kompetansemål også blitt et annet, og hadde testen blitt gitt til en annen sammensetning av elever ville muligens en oppgaves vanskegrad blitt en annen.

For de lavest presterende og de høyest presterende elevene vil trolig ikke dette være av særlig betydning, men for elevene et eller annet sted i mellom kan det godt tenkes at det vil ha relativt stor betydning. Figur 2 illustrerer dette på en god måte.



Figur 2: Fire elevers score på to forskjellige tester. Hentet fra Wu & Adams, 2007, s. 11.

Av figuren kan vi se fire elevers score på en lett test (langs x-aksen) og en vanskelig test (langs y-aksen). Vi ser at elev A, som er en svært kompetent elev, scorer høyt på begge testene, mens elev B, som har ganske lav kompetanse, scorer lavt på begge testene. Elev C og D har begge en mer gjennomsnittlig grad av kompetanse, og får naturlig nok en noe høyere score på den lette testen enn på den vanskelige. På den lette testen er imidlertid elev C og D mye nærmere elev A enn elev B. Mens på den vanskelige testen er elev C og D nærmere elev B enn elev A. Dersom begge testene måler den samme kompetansen burde en kunne forvente at avstanden hadde vært den samme, uavhengig av hvilken test som ble gitt elevene (Wu & Adams, 2007).

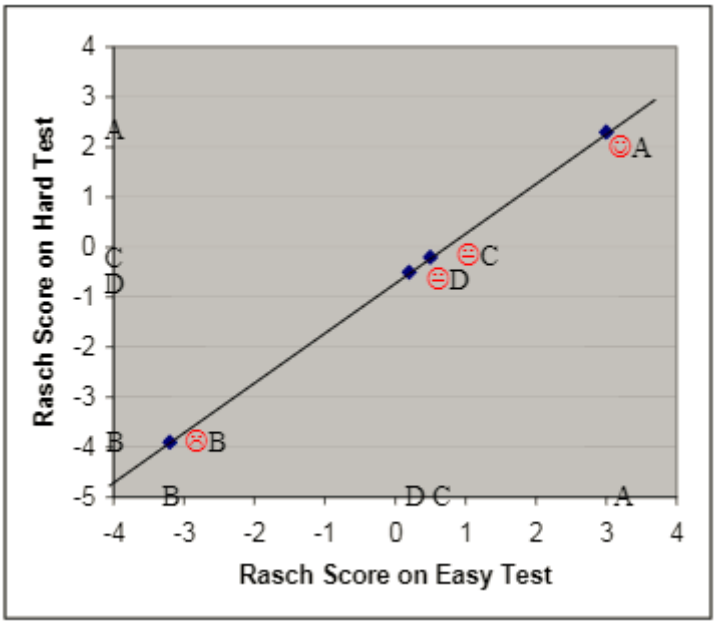
Det samme vil gjelde bestemmelsen av mål på vanskegrad av oppgaver. Dersom A, B, C og D var oppgaver som ble gitt til hhv en elevgruppe med høy kompetanse og en elevgruppe med lav kompetanse ville oppgave C og D kunne tolkes som relativt enkle ut ifra besvarelsene til elevgruppen med lav kompetanse, og som relativt vanskelige ut ifra besvarelsene til elevgruppen med høy kompetanse.

Som vi ser kan mål angitt som totalscore eller andelen riktige svar være vanskelige å tolke utenfor rammen av den aktuelle testen, samt vanskelige å sammenligne med resultater av andre tester som tar sikte på å måle den samme kompetansen. Ettersom de relative avstandene ikke beholdes fra test til test vil målingene i større grad produsere ordinale skalaer enn intervaller. Dersom en ønsker et mest mulig nøyaktig svar på hvor godt en test måler en definert kompetanse er det ønskelig at variabelen kompetanse ikke endrer seg avhengig av hvilke oppgaver som gis, og at variabelen vanskegrad ikke endrer seg avhengig av hvilke elever som responderer på oppgaven.

3.6 Item Response Theory

Item Response Theory (IRT) tilbyr en løsning på denne problematikken. IRT forutsetter at det er mulig å måle spesifikke latente egenskaper som ikke er direkte observerbare. Modellen går ut ifra at det er en sammenheng mellom en persons kompetansenivå og personens respons på en oppgave (Cohen et al., 2011, s. 480). Ved å benytte IRT vil en kunne oppnå målinger som er uavhengige av både oppgaveutvalget og personutvalget, noe som både bidrar til mer presise mål og høyere grad av generaliserbarhet. Der hvor klassisk testteori gir oss ordinale mål på de enkelte testtakernes kompetanse (altså at Elev 1 er mer kompetent enn Elev 2, og at Oppgave 1 er vanskeligere enn Oppgave 2), gir IRT oss intervallmål (altså *hvor mye mer* kompetent Elev 1 er sammenlignet med Elev 2, og *hvor mye vanskeligere* Oppgave 1 er sammenlignet med Oppgave 2). På bakgrunn av blant annet disse egenskapene foretrekkes IRT-modeller i mange storskala-prøver, som for eksempel PISA og TIMMS, og fra 2014 gikk også Utdanningsdirektoratet over til å benytte IRT i utviklingen av de nasjonale prøvene.

Rasch-modellen kan sees på som en én-parameter IRT-modell (Wu & Adams, 2007). Rasch-modellen, lik andre IRT-modeller, transformerer totalscoren til en elev på en slik måte at avstanden mellom de enkelte elevene bevares, uavhengig av de spesifikke oppgavene testen besto av. Resultatet er at kurven i figur 2 blir en rett linje:



Figur 3: De samme fire elevenes score på de samme to prøvene. Hentet fra Wu & Adams, 2007, s. 16.

Vi kan av figuren se at de fire elevene oppnår ulike måltall på de to testene. Det viktige er imidlertid at avstandene mellom de enkelte elevene er uforandret. Konsekvensen av dette er at vi nå i tillegg til å kunne si noe om hvem som er mer kompetent enn andre også kan si noe om *hvor mye mer* kompetent en elev er sammenlignet med en annen. Det samme vil gjelde dersom A, B, C og D var oppgaver. I tillegg til å kunne si at oppgave A er vanskeligere enn oppgave C, kan vi også si noe om hvor mye vanskeligere den er. For begge variablene oppnår vi en måleskala som ivaretar kravet om additivitet (se side 15).

3.7 Rasch-modellen

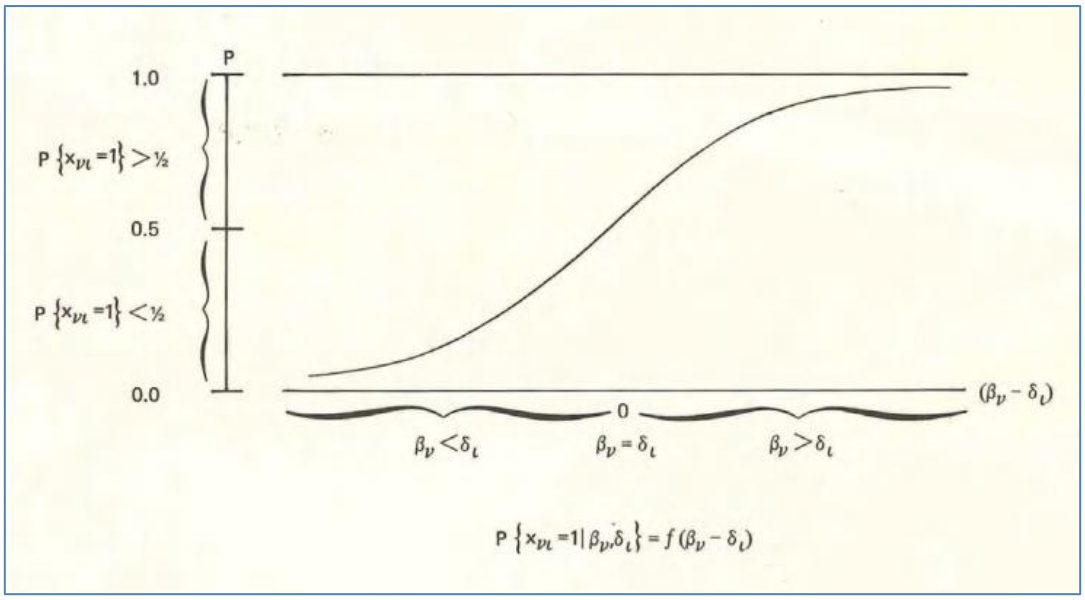
Rasch-modellen legger til grunn at det som avgjør hvordan en person responderer på en oppgave avhenger av to parametere; personens kompetanse og oppgavens vanskegrad. Rasch-modellen estimerer personers kompetanse og oppgavers vanskegrad uavhengig av hverandre og plasserer de på én og samme linje. Resultatet av dette er at det vil bli mulig å si noe om relasjonen mellom en persons kompetanse β_v og en oppgaves vanskegrad δ_i, og at denne kan uttrykkes som differansen mellom de to $(\beta_v - \delta_i)$. Det er imidlertid rimelig å forvente at middels kompetente personer innimellom vil svare feil på en svært enkel oppgave, på samme

måte som det er rimelig å forvente at de innimellom vil svare korrekt på en svært vanskelig oppgave. Denne erkjennelsen har så gitt opphavet til en sannsynlighetsmodell som lar oss beregne sannsynligheten for en gitt respons på enhver oppgave ut ifra størrelsen på differansen til de to parameterne $(\beta_v - \delta\iota)$ (Wright & Stone, 1979, s.12). Konstruksjonen av modellen kan stegvis forklares slik:

1. $\boldsymbol{\beta v - \delta\iota}$ gir oss verdier fra -∞ til +∞.
2. $\boldsymbol{e^{(\beta v - \delta\iota)}}$ gir oss verdier fra 0 til +∞
3. $\dfrac{\boldsymbol{e^{(\beta v - \delta\iota)}}}{\boldsymbol{1 + e^{(\beta v - \delta\iota)}}}$ gir oss verdier fra 0 til 1
4. $\boldsymbol{P\{X_{v\iota} = 1|\beta_v, \delta_\iota\} = \dfrac{e^{(\beta v - \delta\iota)}}{1 + e^{(\beta v - \delta\iota)}}}$ Rasch-modellen

Av Rasch-modellen ser vi at jo større forskjellen mellom en persons kompetanse og en oppgaves vanskegrad blir, jo større eller mindre vil sannsynligheten for at en person klarer oppgaven bli, altså vil sannsynligheten gå mot hhv. 1 og 0. Av dette kommer det også at jo mindre forskjellen er, jo nærmere 0,5 vil sannsynligheten være for riktig eller galt svar. Når en person forsøker å løse en oppgave med akkurat samme mål som seg selv vil vi altså forvente en sannsynlighet på 0,5 for at den blir besvart riktig (Wright & Stone, 1979, s.12).

Figur 4 viser alle mulige utfall når en person med kompetanse β_v besvarer en oppgave med vanskegrad $\delta\iota$. Den viser også alle mulige utfall når en oppgave med vanskegrad $\delta\iota$ besvares av en person med kompetanse β_v. Når β benyttes som variabel kalles utfallskurven for ICC (item characteristic curve), og når δ benyttes som variabel kalles den for PCC (person characteristic curve).



Figur 4: Utfallskurve for både βv og δı. Hentet fra Wright & Stone, 1979, s.14.

3.7.1 Rasch-modellens måleenhet og egenskapen spesifikk objektivitet

Når både kompetanse og vanskegrad er plassert langs samme skala følger det at de også får samme måleenhet. I Rasch-modellen heter måleenheten for logit, som er en sammentrekning av ordene «log odds unit». En persons odds er forholdet mellom personens prosentandel korrekte svar (p) og personens prosentandel feilaktige svar (1-p). Personens kompetansemål i logits er den naturlige logaritmen til denne oddsen. Tilsvarende prosedyre gjennomføres for oppgavene, hvor oddsen er forholdet mellom prosentandelen personer som svarte korrekt på oppgaven og prosentandelen som svarte feil på oppgaven (Bond & Fox, 2001). Deretter gjennomføres en re-skalering slik at oppgavenes gjennomsnittlige vanskegrad blir satt til 0. En persons kompetansemål i logits er den naturlige logaritmen av personens odds for å besvare de(n) oppgaven(e) som definerer nullpunktet på vanskegradsskalaen (Wu & Adams, 2007, s.17).

Ved å omskrive Rasch-modellen kan vi vise hvordan logit-verdien svarer til differansen mellom en persons kompetanse og en oppgaves vanskegrad (Wu & Adams, 2007):

$$\log\left(\frac{p}{1-p}\right) = \beta - \delta$$

Denne omskrivingen lar oss også vise en spesiell egenskap ved modellen kalt *spesifikk objektivitet*. Spesifikk objektivitet innebærer at sammenligningen av to objekter ikke blir påvirket av instrumentet som gjorde sammenligningen mulig. For eksempel skal ikke sammenligningen av to personers kompetanse være påvirket av oppgaven som var grunnlaget for sammenligningen (Wu & Adams, 2007, s.29). Vi kan demonstrere denne egenskapen ved å sammenligne to personer med kompetanse β_1 og β_2 på en oppgave med vanskegrad δ. Vi lar p_1 og p_2 være sannsynligheten for korrekt svar på oppgaven fra hhv. person 1 og person 2.

$$\log\left(\frac{p_1}{1-p_1}\right) = \beta 1 - \delta$$

$$\log\left(\frac{p_2}{1-p_2}\right) = \beta 2 - \delta$$

$$\log\left(\frac{p_1}{1-p_1}\right) - \log\left(\frac{p_2}{1-p_2}\right) = \beta 1 - \delta - (\beta 2 - \delta) = \beta 1 - \beta 2$$

Som vi ser sammenlignes de to elevenes kompetansemål uavhengig av oppgaven.

3.8 Rasch-analyser

I motsetning til andre IRT-modeller, som søker å finne en modell som passer datasettet, forutsetter Rasch-modellen at datasettet skal passe til modellen (Bond & Fox, 2001). Dermed er dette det grunnleggende kriteriet som avgjør om målingen vår kan betraktes som pålitelig eller ikke, jf. de fundamentale kravene til måling. For å finne ut av dette kan vi utføre en rekke analyser.

3.8.1 Variabelkart

I presentasjonen av Rasch-modellen har jeg fremhevet noen sentrale aspekter ved modellen, som at den ved å plassere logit-verdiene til både testdeltakerne og oppgavene på samme linje gjør det enklere for oss å se hvordan disse to størrelsene kan relateres til hverandre. De fleste Rasch-analyseverktøy kan produsere en grafisk fremstilling av denne linjen, og kalles gjerne da et variabelkart («variable map»). Kartet gir oss et bilde av helheten av testen vår hvor vi

kan studere hvordan både personene og oppgavene forholder seg til den underliggende variabelen (Bond & Fox, 2015, s.330).

3.8.2 Fit-verdier og utfallskurver (ICC-er)

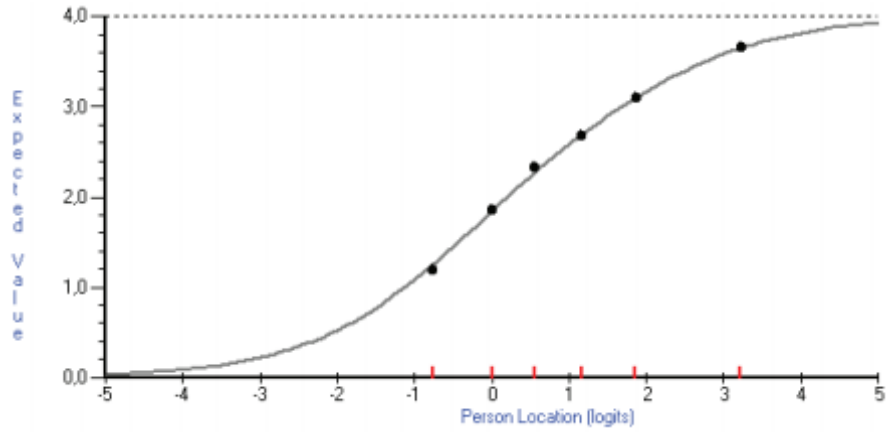
Rasch-modellen forutsetter at alle oppgaver og alle testdeltakere passer til modellen. Sagt på en annen måte forventer den at sannsynligheten for at en oppgave blir besvart korrekt øker når personens kompetansenivå øker, og at sannsynligheten for at en person besvarer en oppgave korrekt øker når vanskegraden synker. Ved bruk av et analyseprogram kan en undersøke i hvor stor grad denne forutsetningen møtes eller ikke av oppgavene i den aktuelle testen, i form av ulike typer fit-verdier og ved å studere forventede ICC-er opp imot reelle ICC-er.

To ulike typer fit-verdier som benyttes er infit og outfit mean squared (infit/outfit MNSQ) som er gjennomsnittlige kvadratavvik mellom forventede responser på en oppgave og de faktiske responser på oppgaven. Forskjellen på de to er at det i estimeringen av infit-verdier legges mer vekt på responser fra personer med mål nær en oppgaves vanskegrad enn de som er langt ifra. Ved estimeringen av outfit-verdier vektlegges alle responser likt. Den forventede mean square-verdien når en oppgave passer perfekt til modellen er 1. Verdier på >1 indikerer såkalt «underfit» og forstås som at det er mer variasjon i de observerte responsene enn det som er forventet av modellen. Verdier mellom 0 og 1 indikerer «underfit», noe som betyr at det er mindre variasjon i de observerte responsene (Bond & Fox, 2015, s.269). Begge typene fit-verdier bør tas hensyn til i vurderingen av hvor godt en oppgave passer, men generelt legges det mer vekt på infit-verdiene enn outfit-verdiene når slike avgjørelser tas (Bond & Fox, 2015).

Litteraturen er ikke entydig på hvilke grenseverdier for hva som kan sees som akseptabelt. Infit - eller outfit MNSQ under 0,7 eller over 1,3 betraktes i følge Bond og Fox (2015) som «misfit» i undersøkelser hvor deltakertallet er under 500, mens verdier mellom 0,5 og 1,5 er akseptable i følge Linacre (2017a). Grenseverdiene på 0,7 og 1,3 er imidlertid anbefalt for rene flervalgstester. En noe større variasjonsbredde vil kunne aksepteres for tester som inneholder konstruksjonsoppgaver i stedet.

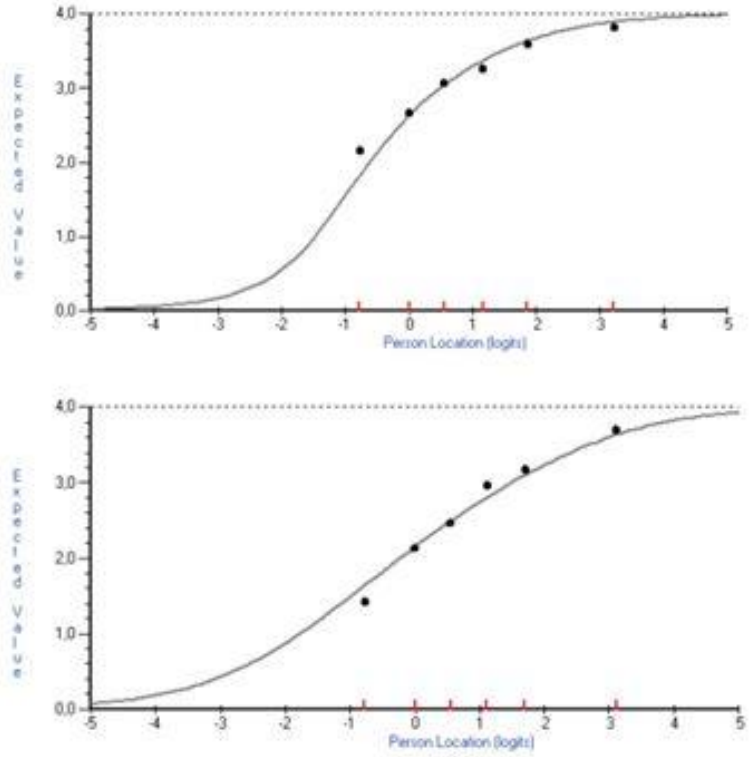
En sammenligning av empiriske og teoretiske ICC-kurver kan gi ytterligere informasjon om hvor godt en oppgave passer modellen. Her deles alle respondenter inn i forskjellige grupper

basert på estimert kompetansenivå. Ved å sammenligne hvordan gjennomsnittsscoren i hver gruppe passer med forventningene til modellen kan en si noe om hvor godt oppgaven passer modellen. Dersom gjennomsnittsscoren til hver gruppe legger seg tilnærmet langs den forventede kurven passer oppgaven modellen godt, noe som betyr at økningen i kompetansenivå gir den forventede økningen i poengscore.



Figur 5: Gjennomsnittscore for hver gruppe (punkter) og ICC samsvarer godt. Her går verdiene langs y-aksen fra 0 til 4, men for en dikotom Rasch-modell som benyttes i denne undersøkelsen vil verdiene gå fra 0 til 1. Hentet fra Sjaastad, 2014, s.220.

Om gjennomsnittsverdiene utgjør en «slakere» kurve eller en «brattere» kurve enn forventet av modellen sier vi at oppgavene hhv. underdiskriminerer eller overdiskriminerer. Underdiskriminering innebærer at økningen i poengscore er mindre enn forventet sammenlignet med økningen i kompetansenivå, og overdiskriminering betyr at økningen i poengscore er større enn forventet sammenlignet med økningen i kompetansenivå (Sjaastad, 2014). Figur 6 viser et eksempel på en oppgave som er noe underdiskriminerende og en oppgave som er noe overdiskriminerende.



Figur 6: Øverst: Oppgave som underdiskriminerer noe. Nederst: Oppgave som overdiskriminerer noe. Hentet fra Sjaastad, 2014, s. 221.

3.8.3 DIF og PCA

Som tidligere nevnt er både én-dimensjonalitet og invarians to fundamentale krav til ethvert måleinstrument. Disse to kravene er nært forbundet med hverandre (Bond & Fox, 2015). Dersom en test inneholder oppgaver som varierer i vanskegrad av årsaker andre enn forskjeller i personenes kompetansenivå brytes kravet om invariante målinger. Dersom kravet om invariante målinger brytes er det sannsynlig at vi måler mer enn kun den ene variabelen vi mener å måle, og kravet til én-dimensjonalitet er brutt.

Når oppgaver oppleves ulikt av ulike personer kalles det «differential item functioning» (DIF), og årsaker til DIF kan knyttes til nasjonalitet, kjønn, yrke, fritidsinteresser, m.m. (Sjaastad, 2014). Ved å sammenligne responsene til ulike definerte undergrupper kan vi identifisere oppgaver som oppfører seg ulikt på tvers av gruppene. Deretter må vi forsøke å finne mulige årsaker til at oppgaven oppfører seg som den gjør.

En «principal component analysis» (PCA) er en type faktoranalyse som benyttes for å avdekke mulige underdimensjoner i variabelen vi ønsker å måle. Enkelt sagt leter den etter systematiske korrelasjoner mellom oppgaver som kan indikere at enkelte oppgaver utgjør undergrupper som har noe til felles utover den latente variabelen vi ønsker å måle. Deretter kan vi gjennomføre en uavhengig t-test for å finne ut av om denne undergruppen av oppgaver gir signifikante forskjeller i testdeltakernes poengscore sammenlignet med de andre oppgavene i testen (Sjaastad, 2014, s.223).

3.9 Validitet og reliabilitet hos et måleinstrument

Dersom vår intensjon gjennom en test er å måle en bestemt egenskap/evne, for eksempel funksjonskompetanse, så vil vi ønske to kvaliteter i denne testen, at målingen er så presis som mulig og at målingen er så anvendbar som mulig i forhold til vår intensjon. I begrepet «presis» ligger det at vi kan stole på resultatene av målingen; den er reliabel. I begrepet «anvendbar» ligger det at vi kan bruke resultatene til å si noe meningsfullt om den egenskapen/evnen vi mente å måle; målingen er valid (Wu & Adams, 2007).

Det er sagt at reliabilitet er en nødvendig, men utilstrekkelig betingelse for validitet (Cohen et al., 2011, s.179). Wolfe & Smith (2007) identifiserer i alt åtte forskjellige områder hvor en kan innhente dokumentasjon for validiteten til et måleinstrument, og hvor reliabilitetsaspektet inngår som en del av disse områdene. De åtte områdene er *innhold* (content), f.eks. grad av relevans og representativitet i forhold til intensjon, oppgavenes tekniske kvalitet; *substansiell* (substantive), f.eks. grad av teoretisk forankring; *strukturell* (structural), f.eks. grad av endimensjonalitet; *generaliserbarhet* (generalizability), f.eks. grad av generaliserbarhet utover utvalg og kontekst, reliabilitet; *ekstern (*external), f.eks. grad av relasjon til andre målinger av samme/lignende begrep/variabel; *konsekvensiell* (consequential), f.eks. grad av reelle og/eller potensielle konsekvenser som kan komme av å bruke instrumentet; *responsivitet (*responsiveness), f.eks. grad av evne til å avdekke endring før og etter en form for intervensjon; *tolkbarhet* (interpretability), f.eks. grad av relasjon mellom kvantitativ måling og kvalitativ mening.

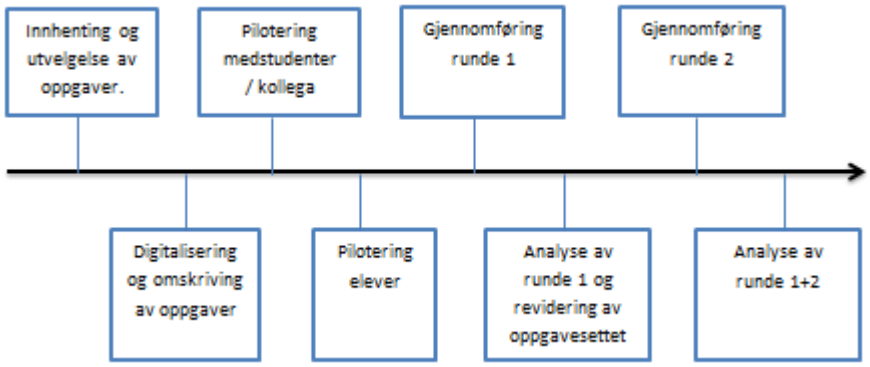
Det er umulig for forskning å være 100% valid og reliabel, og i følge Wolfe & Smith (2007) er det ikke et spørsmål om hvorvidt forskningen er valid eller ikke, men snarere *hvor* valid

forskningen kan sies å være. Å etterstrebe validitet og reliabilitet innebærer altså å forsøke å oppnå så høy grad av validitet og reliabilitet som en kan.

4 Metode

Forskningsmetode for denne undersøkelsen var delvis gitt ut ifra karakteren til forskningsspørsmålene. For å kunne svare på hvordan oppgavene målte funksjonskompetanse var det nødvendig å forsøke å definere hva funksjonskompetanse er. Denne definisjonen vil det imidlertid være ønskelig å verifisere på en eller annen måte utover ens egen tolkning av andres definisjoner. Rasch-modellen med sitt krav til endimensjonalitet lar oss undersøke egnetheten til både definisjon og oppgaver.

Valget av Rasch-modellen, med sine tilhørende statistikker, lar oss analysere de kvantitative sidene ved enkeltoppgaver og måleinstrumentet som helhet. Og dokumentasjon omkring de fleste validitetsaspektene presentert av Wolfe & Smith (2007) vil kunne innhentes gjennom disse statistikkene. Det er imidlertid også kvalitative sider ved selve designet og gjennomføringen av undersøkelsen som må adresseres for å underbygge validiteten og reliabiliteten til undersøkelsen som helhet. Eksempler på slike krav kan være kontrollerbarhet, replikerbarhet, forutsigbarhet og objektivitet (Cohen et al., 2011, s. 180). Jeg vil i dette kapittelet redegjøre for metodiske valg som ble gjort i forkant av, og ved gjennomføringen av undersøkelsens to datainnsamlingsrunder.



Figur 7: Illustrasjon av datainnsamlingsprosessen.

4.1 Forarbeid til undersøkelsen

For i det hele tatt å finne ut om det var grunnlag for å gå videre med den planlagte undersøkelsen ble det tidlig tatt kontakt med matematikklærere på 10. trinn ved ulike

ungdomsskoler i Trondheim. Der hvor det var interesse ble også en eventuell gjennomføring av undersøkelsen avklart med skolens ledelse. Samtidig ble lærerne forespurt om å sende meg emneprøven som var tilknyttet læreverket de benyttet. Prøver som av lærere ble bekreftet benyttet, enten delvis eller i sin helhet, utgjorde databasen for mulige oppgaver i undersøkelsen. Dette var emneprøver tilknyttet læreverkene Faktor 10, Grunntall 10 og Sirkel 10.

På bakgrunn av teori presentert i kapittel 2.3 ble oppgaver i alle tre emneprøvene kategorisert i gruppene *modellering*, *tolkning*, *oversettelse* og *begrepskompetanse*. Oppgavekategoriene ble opprettet med hensyn på de typer oppgaver som utgjorde majoriteten av oppgavene i emneprøvene. Det innebar at øvrige oppgavetyper ble utelatt. Hovedsakelig gjaldt dette oppgaver som omhandlet kvadratiske funksjoner, samt grafiske og algebraiske løsninger av likninger (to av emneprøvene omhandlet både funksjoner og likninger). Der hvor to oppgaver ble betraktet som svært like ble en av disse valgt. Der hvor aspekter ved den definerte funksjonskompetansen ikke ble fanget opp i oppgaveutvalget, ble oppgaver lagt til. Oppgave T2b og T4 ble lagt til oppgavesettet for å kunne si noe om elevenes evne til å gjøre globale tolkninger av grafer.

Tabell 1: Oversikt over hovedkategorier av oppgaver. Oppgaver merket med «0» var kun med i oppgavesettet i runde 1. Oppgave G1 skiller seg ut ved at den tester forkunnskaper om koordinater/koordinatsystemer og faller utenfor alle de fire kategoriene.

Modellere	Tolke	Oversette	Begreper
M1	T1a	A1	F1
M2	T1b	A2	F2
M5a	T1c	A3	F4a
M5b	T1d	A4	F4b
0M3	T2a	A5	F6
0M4	T2b	A6	0F3
E1	T3	A7	0F5
	T4	A8	
G1*	T5	A9	

Videre ble oppgavene inndelt i kategoriene *lav*, *middels* og *høy* ut ifra forventet vanskegrad. Denne inndelingen tilsvarte en inndeling etter oppgavetypene *tolkning (identifisering)*, *konstruksjon* og *forklaring/begrunnelse*, jf. rammeverket fra Nitsch et al. (2015) og Leinhardt et al. (1990). Deretter ble oppgavene justert opp eller ned i vanskegrad ut ifra om oppgaven krevde en lokal eller global tolkning, og ut ifra om oppgaven krevde én eller flere operasjoner.

Denne grovinndelingen av forventet vanskegrad var ment å gi oss en viss idé om den relative spredningen av oppgavene i settet. Alle oppgavene er å finne i appendiks C.

Tabell 2: En grovinndeling av vanskegrad. For oppgaver som ble justert indikerer lys, grå skrift plassering før justering, og sort skrift viser endelig forventet plassering.

← Lokal	Global →
← En operasjon	Flere operasjoner →

Enkel	Middels	Vanskelig
		F1
	A1 →	A1
T3		
T1a		
T1b		
T1c		
T1d		
T2a →	T2a	
T2b →	T2b →	T2b
	A2	
T5*		
	M1 →	M1
	A3	
F4a ←	F4a ←	F4a
F4b →	F4b	
F6		
	A4	
	A5	
	A6	
	A7	
	M2	
G1		
	E1 →	E1
	A8	
	T4 ←	T4
	M5a	
	M5b	
	A9	
F2 →	F2	
		0M3
	0M4	

*I kategorien «forklaring/begrunnelse», men med forenklende elementer.

Til gjennomføringen av undersøkelsen ble prøveverktøyet Matistikk benyttet. Dette er en digital webplattform for innsamling av matematiske tekster som er under utvikling av Institutt for lærerutdanning ved NTNU. Ved å bruke et digitalt verktøy til datainnsamlingen ble det lettere å administrere testen og besvarelsene enn om dataene ble samlet inn på papirform. I Matistikk kunne elevene gjennomføre testen anonymt uten at vi behøvde å samle inn

personidentifiserende informasjon. Bruken av Matistikk innebar at oppgavene måtte digitaliseres og tilpasses denne plattformen.

Ettersom emneprøvene var tilknyttet læreverk som flere av elevene i undersøkelsen benyttet ble oppgavene også omskrevet noe for å unngå at elevene skulle ha gjort samme oppgave før. Denne omskrivingen begrenset seg hovedsakelig til enten å endre tallverdier på oppgaver uten kontekst, eller å endre situasjonsbeskrivelsen til oppgaver med kontekst. Endringene ble forsøkt gjort slik at oppgaveinnholdet i så stor grad som mulig forble det samme.

Figur 8: Eksempel på omskriving av oppgave A4. Opprinnelig oppgave til venstre.

Figur 9: Eksempel på omskriving av oppgave M5a og M5b. Opprinnelig oppgave til venstre.

Oppgave E1 gjennomgikk imidlertid en større transformasjon, da denne ble omskrevet fra å være en kvantitativ oppgave til å bli en kvalitativ oppgave. Med det menes at situasjonen som var beskrevet i oppgaven gikk fra å inneholde kvantitativ informasjon til å være fri for kvantitativ informasjon (se s. 12). Bakgrunnen for omskrivingen var både fordi kvalitative oppgaver var fraværende i alle de tre emneprøvene, og fordi jeg ønsket at elevenes digitale kompetanse i minst mulig grad skulle påvirke resultatene. Situasjonen i oppgave E1 beskrev en omvendt proporsjonal sammenheng, og ville krevd et svar på brøkform ($y = \frac{2000}{x}$). Dersom elevene ikke er vant med å skrive brøker i arbeid med digitale verktøy kunne det ha medvirket til at elever hadde hoppet over oppgaven eller brukt uønsket ekstra tid på finne ut av hvordan de skulle få til å kommunisere løsningen. Ved å omskrive oppgaven ble dette problemet omgått og oppgavesettet fikk en kvalitativ oppgave.

4.1.1 Elevutvalg

Deltakerne i undersøkelsen var elever på 10. klassetrinn fra totalt fem forskjellige ungdomsskoler i Trondheim kommune. Totalt 250 elevbesvarelser fordelt på 11 ulike klasser utgjør datagrunnlaget for analysene.

4.2 Erfaringer fra piloteringer

For å se om oppgavene fungerte som tiltenkt på den digitale plattformen fikk jeg to medstudenter og en kollega til å besvare oppgavesettet. Formålet med denne piloten var å få tilbakemeldinger på hvordan de opplevde at oppgavene ble presentert i sin digitale form, hvordan de opplevde brukervennligheten til prøveverktøyet og for å få en liten pekepinn på tidsbruken. Tilbakemeldingene var stort sett positive, det tekniske så ut til å fungere, og kun små justeringer i utformingen av noen oppgaver ble gjort i etterkant. Deretter ble det forsøkt gjennomført en pilot i en 10. klasse ved en skole i Trondheim. I Trondheimsskolen har en de siste årene gått over fra at elevene bruker PC til at de bruker Google Chromebooks til digitalt arbeid på skolen. Hovedformålet med piloten var å se hvordan det tekniske fungerte på disse enhetene, samt en ytterligere tilbakemelding på om oppgavesettet virket å være tilpasset en tidsramme på i underkant av 60 min. Her ble det fullstendig stopp med en gang elevene skulle laste opp prøven, og en feilmelding «connection timed out» var det eneste de fikk opp på skjermen. Etter omfattende feilsøking fra flere hold, og nok et mislykket forsøk på gjennomføring av pilot, ble det avdekket at oppgavesettet var for stort til at Chromebookene kunne laste inn hele oppgavesettet på én gang. Løsningen ble derfor å dele oppgavesettet i tre deler, slik at elevene kunne laste opp en del av gangen. Som en konsekvens av dette måtte det opprettes et kodeord som kunne identifisere elevene på tvers av de tre delene, og på den måten få på plass fullstendige besvarelser. En oppgave hvor elevene ble bedt om å skrive inn sitt kodeord ble derfor lagt til i hver av de tre delene (se side 90). På grunn av tiden det tok å løse de uforutsette tekniske problemene ble det ikke gjennomført noen ny pilot blant elevene før første runde av hovedundersøkelsen.

4.3 Endringer fra første til andre runde

Første runde av undersøkelsen ble gjennomført i fem forskjellige elevgrupper på 10. trinn fordelt på tre ungdomsskoler i Trondheim. Analyser av datasettet fra denne runden førte til endringer av noen av oppgavene. Årsaker til, og konsekvenser av endringene vil jeg komme nærmere inn på i analysekapittelet, men nevnes kort her ettersom de er vesentlige for datainnsamlingen i runde 2. Oppgavene som omhandlet fagbegreper, altså i kategorien begrepskompetanse, var opprinnelig flervalgsoppgaver alle sammen, og dermed også rene identifikasjonsoppgaver. På bakgrunn av analysestatistikker, som høye outfit-verdier og noe utilfredsstillende diskrimineringsegenskaper, ble disse omskrevet til å være oppgaver som ba elevene forklare/begrunne påstander i stedet. Ettersom denne typen oppgaver gjerne tar mer tid å besvare enn identifikasjonsoppgaver ble to av oppgavene, F3 og F5, tatt ut av oppgavesettet. Oppgavene F3 og F5 er derfor også utelatt fra tabell 2 da denne inkluderer de oppgavene som inngår i den endelige analysen. Oppgave T4 var også i utgangspunktet en flervalgsoppgave som ble omskrevet til å be elevene begrunne en påstand, og oppgave T5 ble omskrevet etter indikasjoner på at oppgaven ga litt for stor variasjon i elevenes svarmønster i forhold til Rasch-modellens forventninger og ikke diskriminerte godt nok mellom elevene. Som en konsekvens inneholdt oppgavesettet i runde 2 ingen flervalgsoppgaver.

To oppgaver ble lagt til oppgavesettet, og en oppgave fikk endret tallverdiene som var oppgitt. Oppgave F4b ble lagt til for å innhente informasjon som kunne falle bort da vi endret flervalgsoppgaven F4. Oppgave F4b spurte etter stigningstallet til et linjestykke. Oppgave A9 ble lagt til for å se om dette kunne være en enklere oversettelsesoppgave enn de som allerede var i oppgavesettet. Bakgrunnen for dette var at en dimensjonsanalyse (PCA) indikerte at oversettelsesoppgavene utgjorde en egen dimensjon, men viste samtidig at alle oversettelsesoppgavene hadde en vanskegrad fra gjennomsnittet til oppgavesettet og oppover. Ved å legge til oppgave A9 kunne vi få tydeligere indikasjon på om dette var gjeldende for oversettelsesoppgaver generelt. Oppgave A6 fikk endret på tallverdiene i en tabell, slik at stigningstallet skulle kunne leses direkte ut fra tabellen. Årsaken til endringen var den samme som for inkluderingen av oppgave A9.

4.4 Gjennomføring av undersøkelsen

Den første runden ble som nevnt gjennomført i fem elevgrupper på 10. trinn, fordelt på tre skoler. Den andre runden ble gjennomført to måneder senere i ni andre elevgrupper på 10. trinn, fordelt på fem skoler. Totalt 250 elevbesvarelser danner grunnlaget for den endelige analysen.

For å sikre at rammene rundt gjennomføringen ble holdt så like som mulig i hver elevgruppe administrerte jeg undersøkelsen selv. Alle gruppene fikk først en fem minutter introduksjon som inkluderte informasjon om formålet med studien, hvordan undersøkelsen ivaretok anonymiteten deres, at det ikke var en plikt å bidra til undersøkelsen, men en forespørsel, samt praktisk informasjon om hvordan undersøkelsen ble gjennomført. Det siste inkluderte en fremvisning av en eksempeloppgave og en kort demonstrasjon av bruk av verktøyene som var tilgjengelige. Formidlingen av oppgavesettets tre deler ble gjort i form av tre ulike lenker som faglærer hadde gjort tilgjengelig for elevene på matematikkfagets nettside. For å minimere muligheten for samarbeid eller kopiering av andres besvarelser ble alle oppgavene gitt i tilfeldig rekkefølge, og for å sikre en tilnærmet jevn fordeling av besvarelser utover alle oppgavene i testen begynte en tredel av elevene med del 1, en tredel med del 2 og en tredel med del 3. Med en fem minutter introduksjon ble den resterende tiden på 55 minutter satt av til at elevene besvarte oppgavene. De elevene som ikke ble ferdige fikk beskjed om å sikre at de hadde skrevet inn kodeordet på den delen de ikke var ferdige med, og deretter om å levere.

4.5 Scoring av oppgaver og valg av analysemodell

Rasch-modellen som er beskrevet tidligere i denne oppgaven er en såkalt dikotom modell. Det vil si at besvarelser på oppgaver scores enten riktig (1) eller galt (0). Det finnes imidlertid andre Rasch-modeller og én av disse, partial credit model (PCM), ble benyttet i analysen av datasettet fra runde 1, men ikke runde 2. PCM benyttes der hvor oppgaver også kan være delvis riktige, og disse kan altså scores fra 0-2 eller 0-3 osv., alt etter hva som er mest hensiktsmessig for den aktuelle oppgaven (Bond & Fox, 2015). Årsaken til at PCM ble benyttet i første runde var at en av flervalgsoppgavene, oppgave F1, hadde flere enn ett riktig alternativ. Da denne ble omskrevet i forkant av runde 2, falt også behovet for bruk av PCM bort.

Dersom delvis riktige svar skal belønnes krever dette en systematisk scoring slik at hver enkelt økning i poeng representerer en definert økning i den aktuelle ferdigheten som testes. Dette er et prinsipp ikke bare ved bruk av Rasch-modellen, men et prinsipp ved måling generelt (Bond & Fox, 2015, s.141). De øvrige oppgavene som var å finne i emneprøvene var hovedsakelig utformet slik at de ba om ett enkelt svar. For vår del ble gradering av besvarelser vurdert som lite hensiktsmessig sett opp imot dette prinsippet, og også sett opp imot formålet med bl.a. å bestemme oppgavenes vanskegrad.

I analysearbeidet ble dataprogrammet WINSTEPS (Linacre, 2017b) benyttet. Ettersom alle oppgaver ble scoret som enten 1 eller 0, kunne andre siffer benyttes som kode for «ikke rukket» eller «hoppet over». Der hvor det var oppgaver som eleven ikke hadde rukket å besvare, eller av andre årsaker hadde brukt mindre enn to sekunder på oppgaven (en annen fordel ved bruk av Matistikk var at vi får oppgitt tidsbruk på hver enkelt oppgave), ble besvarelsene scoret «9». WINSTEPS tolker dette som «missing data» og Rasch-modellen blir ikke særlig berørt av besvarelser som ikke er komplette. WINSTEPS klarer derfor å estimere både kompetansemål og vanskegradsmål til tross for at elever ikke svarer på alle oppgavene. Dersom besvarelsene skulle bli veldig langt unna komplette vil naturlig nok presisjonen på estimeringene bli mer unøyaktige (Bond & Fox, 2015). For denne undersøkelsen ble elever som manglet å levere/besvare mer enn en tredel av oppgavene utelatt fra datasettet.

4.6 Metodekritikk

Selv om gjennomføringene ble forsøkt holdt så like som mulig var det for eksempel ikke alltid mulig å påvirke når på dagen, og hvor langt i etterkant av endt undervisning i emnet undersøkelsen kunne tas av de enkelte elevgruppene. For runde 1 sprikte tidspunkt på dagen fra andre til femte skoletime, men ved runde 2 gjennomførte alle elevene oppgavesettet mellom klokken 0815 og 1130.

Ingen kunne delta før etter å ha deltatt i undervisningen om funksjoner. Det innebar at to skoler ikke kunne være en del av runde 1, men måtte vente til runde 2. Som en konsekvens var det større forskjell i hvor lenge siden elevene ved de forskjellige skolene hadde hatt undervisning om emnet. Selv om dette trolig ikke har all verden å si kan vi ikke se bort ifra at det kan ha hatt en påvirkning på prestasjonene, hovedsakelig hos elever som mangler konseptuell forståelse, og dermed er mer avhengig av å huske innøvde prosedyrer. Det må

likevel poengteres at samtlige elever hadde arbeidet med dette emnet relativt nylig, så tidspunktet for undersøkelsen må kunne sies å være gunstig for å måle elevenes kompetanse for funksjoner.

Omskrivingen av oppgaver ble som nevnt forsøkt holdt til et minimum. Likevel kan dette ha endret noen av oppgavene, og i så fall påvirket resultatene noe. Det samme gjelder overgangen fra det opprinnelige papirformatet til digitalt format på testen. Det var imidlertid presisert på den ene av emneprøvene at flere av oppgavene kunne løses digitalt, så for disse oppgavene vil det ikke være noen særlig endring på dette området.

5 Resultater

Resultatkapittelet er inndelt i tre deler. Innledningsvis presenteres en analyse av hvordan oppgavene i hver av de tre emneprøvene fordeler seg på de definerte delkompetanse-kategoriene. Gjennom denne ser jeg på hva slags kompetanse emneprøvene vektlegger. Deretter undersøkes emneprøvenes oppgaver med hensyn på deres måleegenskaper. I denne delen av resultatkapittelet vil ulike analyseresultater som adresserer reliabilitet og validitetsaspekter fremsatt av Wolfe og Smith (2007) presenteres. Gjennom disse søker jeg å finne svar på hvordan, og med hvilken presisjon, oppgavene fra de utvalgte emneprøvene måler funksjonskompetanse. Avslutningsvis presenteres resultater av kvalitative analyser som belyser sentrale bakenforliggende årsaker til noen av de kvantitative resultatene.

5.1 Funksjonskompetanse i emneprøvene

I hvilken grad kan emneprøvene som helhet sies å dekke funksjonskompetanse som definert i teori og læreplan? Oppgavene i emneprøvene ble kategorisert i de fire delkompetansene modellering, tolkning, oversettelse og begrepskompetanse. Oppgaver innenfor de tre første delkompetansene-kategoriene ble ansett som et minimum for å dekke kompetansemålene i læreplanen (se s. 9). Begrepskompetanse er sett på som sentralt for tilegnelsen av de tre andre delkompetansene (Usiskin, 2015).

Ved sammenligning av de tre ulike forlagsgitte prøvene kom flere forskjeller til syne, både i prøvenes totale antall oppgaver og i hvordan oppgavene fordeler seg på de definerte kategoriene (se tabell 3). Emneprøve A har en relativt jevn fordeling på tvers av kompetanse-kategoriene. I emneprøve B er det en hovedvekt på tolkningsoppgaver. Kun én av oppgavene er av typen oversettelse, og ingen oppgaver er spesifikt rettet mot å måle begrepskompetansen til elevene. For emneprøve C er vektingen av kategoriene tolkning og oversettelse omtrent motsatt av hva den er i emneprøve B, med kun én tolkningsoppgave og seks oversettelsesoppgaver. Emneprøve C inneholder som emneprøve A fire oppgaver av typen begrepskompetanse. Tabellen viser at modelleringsoppgaver er representert ved hhv. to, tre og fire oppgaver i hver av de tre emneprøvene.

I kategorien *forkunnskap* finner vi oppgaver som omhandler koordinatsystemet, mens kategorien *andre* inneholder oppgaver som omhandler kvadratiske funksjoner, samt oppgaver som ble vurdert til å i større grad adressere digital kompetanse enn funksjonskompetanse. Sistnevnte var oversettelsesoppgaver fra funksjonsuttrykk til graf, hvor elevene kunne skrive funksjonsuttrykket i et inntastingsfelt og en digital graftegner gjorde oversettelsen.

Tabell 3: Fordeling av oppgaver på hver av kompetanse-kategoriene for de tre emneprøvene.

Emneprøve	Modellering	Tolkning	Oversettelse	Begreper	Forkunnskap	Andre
A	2	4	5	4	1	5
B	3	7	1	0	1	3
C	4	1	6	4	4	5

5.2 Instrumentets måleegenskaper

I analysene av instrumentets måleegenskaper innhentes det dokumentasjon som gjør oss i stand til å svare på i hvilken grad oppgavene gir et presist og rettferdig bilde av elevenes kompetanse innenfor emnet funksjoner. Her analyseres måleinstrumentet med hensyn på validitetsaspekter som ble presentert i måleteorikapittelet (s. 26).

5.2.1 Én-dimensjonalitet

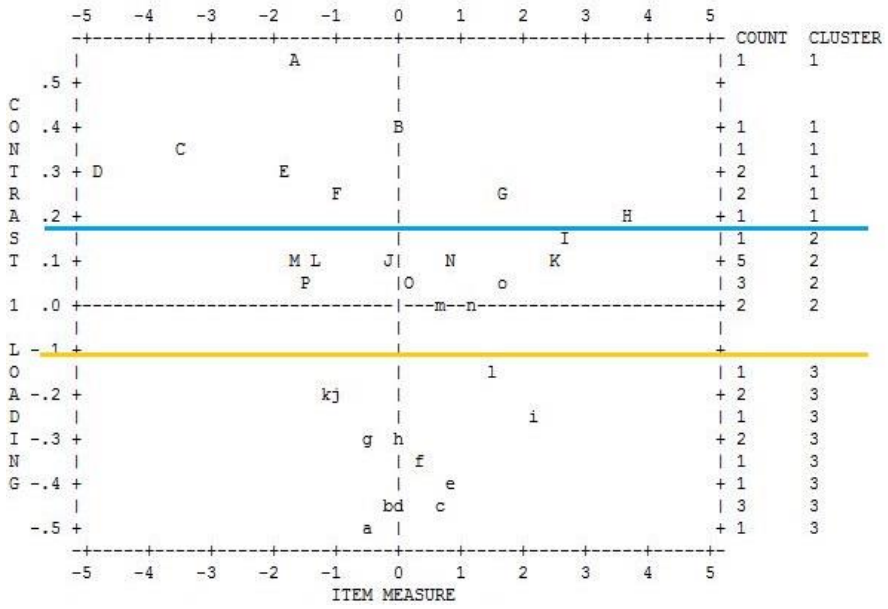
Et fundamentalt krav til ethvert måleinstrument er kravet om én-dimensjonalitet (Wright & Stone, 1979). Det er derfor viktig at oppgavesettet ikke måler helt urelaterte aspekter som for eksempel evnen til å tolke tekst. Funksjonskompetanse beskrives i teorien som sammensatt av flere delkompetanser. Det er imidlertid uklart hvor uavhengige de er, og om dette har noen betydning for vurdering av funksjonskompetanse. Ved å gjennomføre dimensjonsanalyser undersøkes instrumentets strukturelle validitet (Wolfe & Smith, 2007).

En PCA (s. 26) av datasettet indikerte at to undergrupper av den latente variabelen skilte seg ut som mulige dimensjoner.

Tabell 4: Resultat av PCA.

Table of STANDARDIZED RESIDUAL variance in Eigenvalue units = ITEM information units			
	Eigenvalue	Observed	Expected
Total raw variance in observations =	60.5128	100.0%	100.0%
Raw variance explained by measures =	29.5128	48.8%	49.0%
Raw variance explained by persons =	16.2621	26.9%	27.0%
Raw Variance explained by items =	13.2507	21.9%	22.0%
Raw unexplained variance (total) =	31.0000	51.2% 100.0%	51.0%
Unexplned variance in 1st contrast =	2.4508	4.1% 7.9%	
Unexplned variance in 2nd contrast =	1.9434	3.2% 6.3%	
Unexplned variance in 3rd contrast =	1.6385	2.7% 5.3%	
Unexplned variance in 4th contrast =	1.5855	2.6% 5.1%	
Unexplned variance in 5th contrast =	1.4980	2.5% 4.8%	

Tabell 4 viser resultatet av variansanalysen. Vi ser at variansen i datasettet som kan forklares av målingene (raw variance explained by measures) er noe mindre enn variansen som ikke kan forklares (49% mot 51%). Det indikerer at det finnes uønskede faktorer som påvirker målingene. Videre viser analysen i form av en egenverdi (eigenvalue) på over 2,0 i første kontrast at det kan finnes oppgaver i oppgavesettet som trolig har noe annet til felles utover å være en del av funksjonskompetansen (Linacre, 2017a). Figur 10 viser hvordan oppgavene i oppgavesettet inndeles i en hovedgruppe og to undergrupper. Den blå linjen markerer skillet mellom undergruppe 1 og hovedgruppen (2), og den oransje markerer skillet mellom undergruppe 3 og hovedgruppen (gruppe = cluster).



Figur 10: Analyse av dimensjoner. Analysen indikerer at funksjonskompetanse kan bestå av underdimensjoner.

De oppgavene som er av størst interesse for oss er de som ligger lengst fra hverandre, altså de oppgavene med størst avstand fra «midtlinjen» vår i hhv. undergruppe 1 og 3 (Bond & Fox, 2015). Først ble det undersøkt hva som kunne forklare dannelsen av undergruppe 1, deretter hva som kunne beskrive undergruppe 3, før vi sammenlignet oppgavene lengst fra hverandre for å se om de kunne sies å kreve vesentlig forskjellige typer kompetanse. Ettersom oppgavene var navngitt etter hvilken av de fire kategoriene modellering, tolkning, oversettelse og begrepskompetanse de var en del av, gir tabell 5 en tydelig indikasjon på hva som kjennetegner hver av de to hovedgruppene.

Tabell 5: Oversikt over hvilke oppgaver som utgjør mulige underdimensjoner.

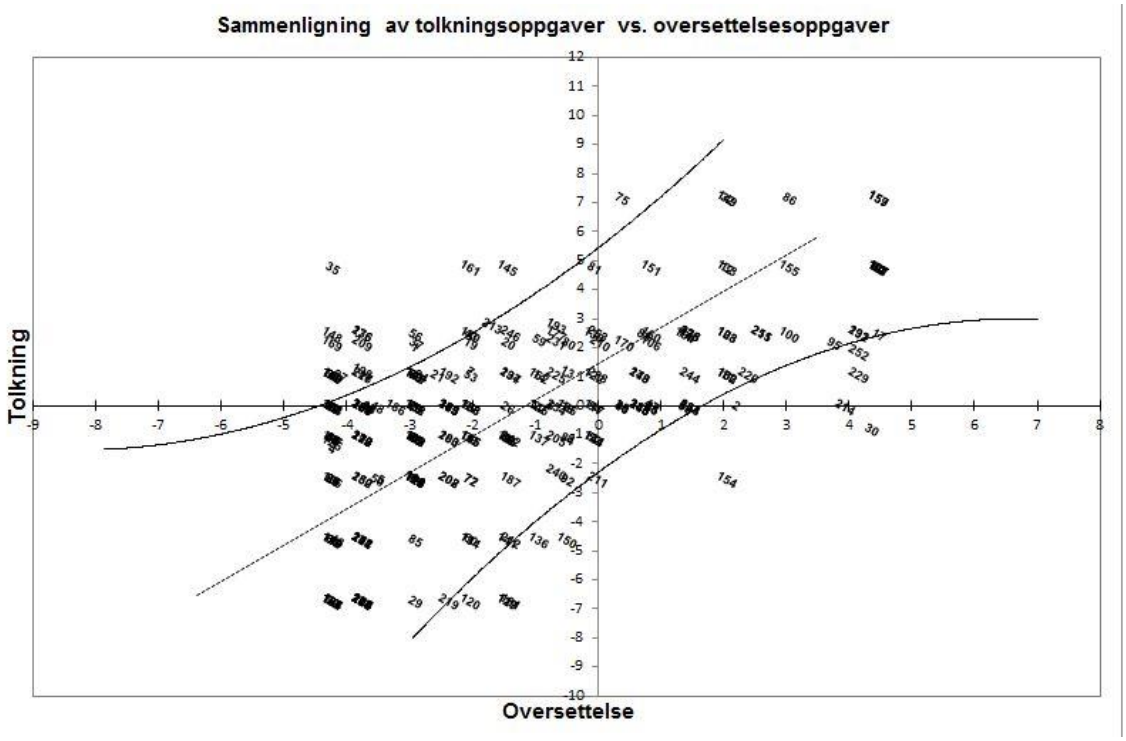
CON-TRAST	LOADING	INFIT MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITE	LOADING	INFIT MEASURE	INFIT MNSQ	OUTFIT MNSQ	ENTRY NUMBER	ITE
1	.53	-1.62	1.18	1.39	A 3	T3	-.49	-.43	.75	.71	a 17	A4
1	.39	-.05	1.24	1.15	B 8	T2a	-.47	-.09	.70	.55	b 28	A9
1	.35	-3.55	.93	.70	C 6	T1c	-.46	.61	.70	.49	c 13	A3
1	.30	-4.87	1.01	.71	D 4	T1a	-.43	.05	.74	.58	d 19	A6
1	.29	-1.76	1.11	1.13	E 7	T1d	-.38	.87	.75	.41	e 20	A7
1	.25	-.93	1.24	1.39	F 5	T1b	-.36	.29	.74	.59	f 24	A8
1	.24	1.69	1.19	1.36	G 12	M1	-.32	-.57	.81	.74	g 26	M5a
1	.20	3.67	1.07	.30	H 1	F1	-.28	.03	1.21	1.37	h 16	F6
1	.14	2.59	.96	.42	I 29	F2	-.25	2.15	.84	.40	i 2	A1
1	.12	-.13	1.01	1.00	J 22	G1	-.22	-1.03	.99	.94	j 11	T5
1	.11	2.43	1.06	.73	K 9	T2b	-.20	-1.10	1.26	1.41	k 21	M2
1	.11	-1.37	1.21	1.50	L 15	F4b	-.16	1.44	1.03	.90	l 18	A5
1	.09	-1.67	.99	1.03	M 27	M5b	-.01	.73	.91	.84	m 10	A2
1	.08	.85	1.07	.85	N 25	T4						
1	.06	.21	1.17	1.16	O 38	0M4						
1	.04	-1.43	1.03	1.04	P 14	F4a						
1	.04	1.75	.98	.68	o 23	E1						
1	.01	1.23	1.27	1.29	n 37	0M3						

Oppgavene i de blå rammene er de som ligger lengst fra hverandre i hhv. undergruppe 1 (venstre ramme) og undergruppe 3 (høyre ramme). Av første bokstav i oppgavenavnet kan en se at alle oppgavene i undergruppe 1 er i kategorien tolkning. En nærmere undersøkelse viste at oppgavene i tillegg primært adresserer lokale tolkninger av grafer. De to oppgavene T2b og T4 som ikke inngår i denne undergruppen skiller seg fra de andre ved at tolkningskravet er mer globalt. Den siste, oppgave T5, er tolkning av et funksjonsuttrykk. Felles for oppgavene i den blå rammen til høyre er at de alle er oversettelsesoppgaver. Et annet funn var at samtlige

oppgaver på høyre side (gul ramme) involverer et funksjonsuttrykk, enten som kilderepresentasjon eller målrepresentasjon, og at alle oversettelsesoppgaver er å finne innenfor denne rammen.

5.2.2 Underdimensjoner av funksjonskompetanse

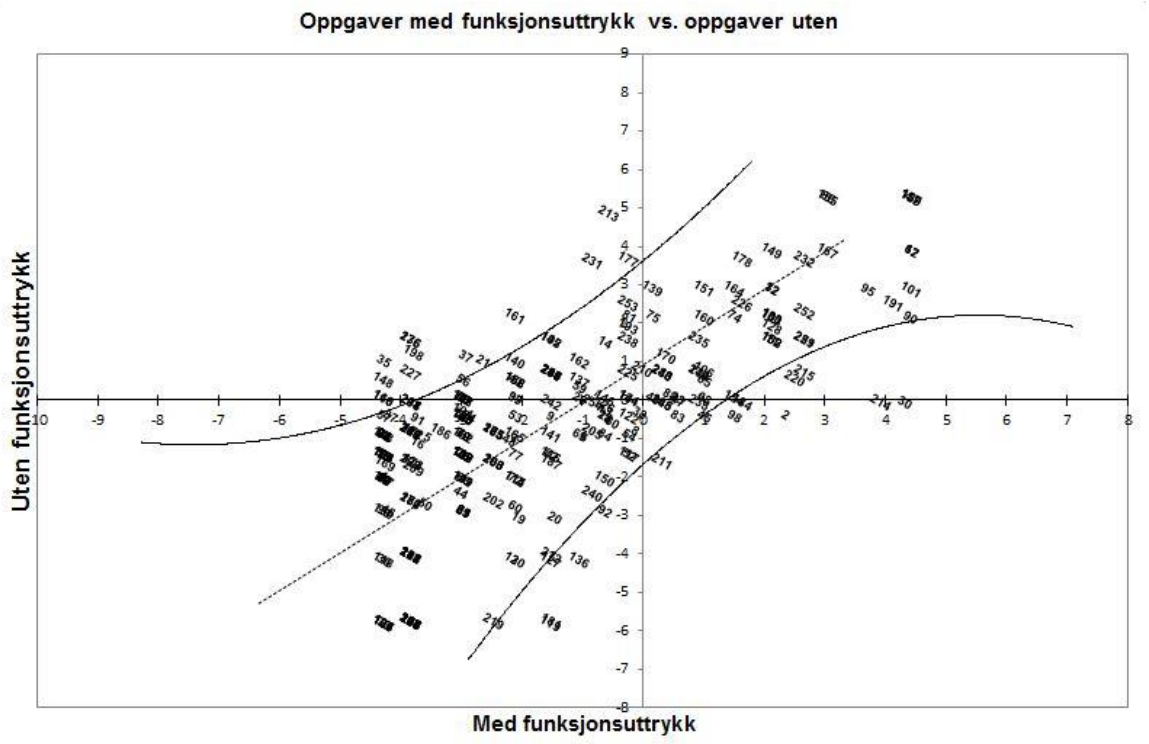
Hvilken betydning har så de identifiserte underdimensjonene? Årsaken til at det danner seg undergrupper er at noen elever presterer ulikt på oppgaver i hver av undergruppene, altså at noen elever gjør det merkbart bedre på oppgaver i for eksempel undergruppe 1 sammenlignet med oppgavene i undergruppe 3. For å finne ut av hvilken betydning dette har for målingen ble elevmålene ut ifra oppgavene i hver av de to undergruppene sammenlignet med hverandre. Figur 11 viser en sammenligning mellom mål elevene oppnår dersom vi kun gir de tolkningsoppgaver og mål de oppnår hvis de kun får oversettelsesoppgaver. De elevene som faller utenfor 95 % konfidensintervallet (buede linjer) får statistisk signifikante endrede mål fra en undergruppe til en annen.



Figur 11: Mål på elever, oversettelsesoppgaver (x-akse) mot tolkningsoppgaver (y-akse).

Vi ser av analysene at flere elever oppnår ulike mål avhengig av hvilken undergruppe oppgavene tilhører. Ved sammenligningen av tolkningsoppgaver og oversettelsesoppgaver får 12% (30 av 250) av elevene signifikant endrede mål. Et ekstremt tilfelle finner vi i elev nr. 35, som får et kompetansemål som varierer mellom 5 og -4 logits avhengig av om eleven får tolkningsoppgaver eller oversettelsesoppgaver. Vi ser at det både finnes elever som oppnår bedre resultater på tolkningsoppgavene, og at det finnes elever som oppnår bedre resultater på oversettelsesoppgavene.

For å finne ut om funksjonsuttrykk er en mulig kompetanse-kategori i seg selv ble en tilsvarende sammenligning som den over gjort mellom oppgaver som omhandler representasjonsformen funksjonsuttrykk på den ene siden, og oppgaver fri for denne representasjonsformen på den andre siden. Figur 12 viser forskjeller i elevmål også her, og at noen elever viser signifikant høyere kompetanse i den ene kategorien, mens andre viser signifikant høyere kompetanse i den andre.

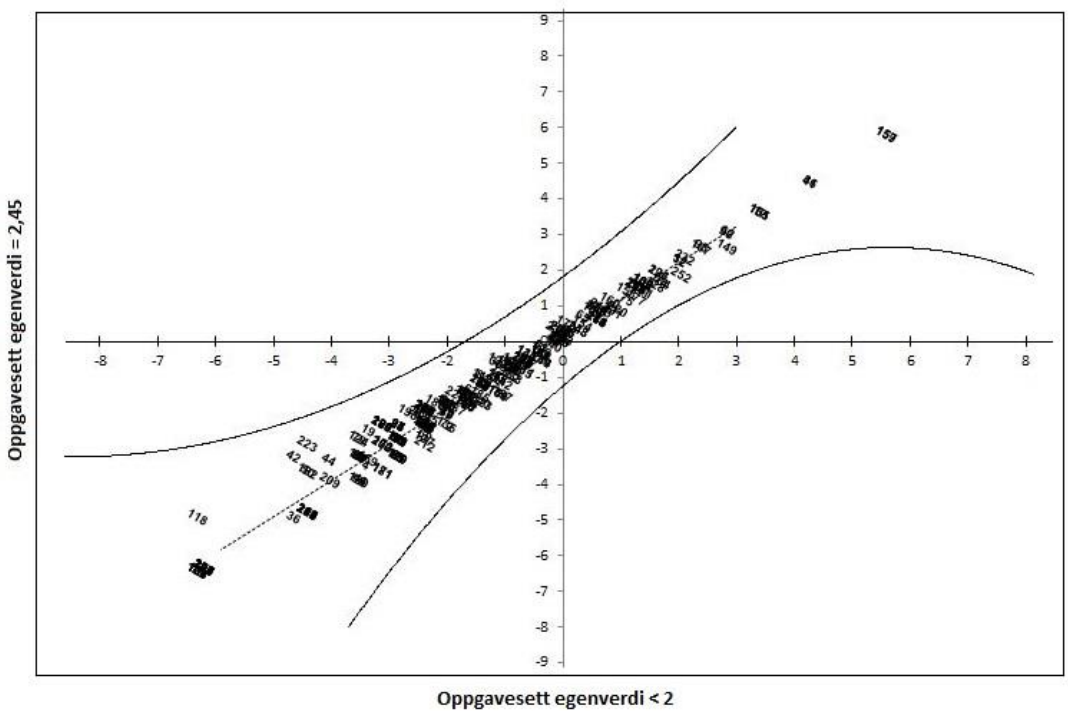


Figur 12: Mål på elever, oppgaver med funksjonsuttrykk (x-akse) mot oppgaver uten funksjonsuttrykk (y-akse).

Av analysene ser vi altså at elevmålene varierer fra en oppgavekategori til en annen, selv om det må bemerkes at endringene for de aller fleste elevene ikke er statistisk signifikante. Analysene indikerer at tolkingsoppgaver og oversettelsesoppgaver er vesensforskjellige, og at oppgaver som stiller krav til algebraisk symbolkompetanse også skiller seg fra de øvrige oppgavene. Dette er av teoretisk interesse ettersom det underbygger at funksjonskompetanse består av flere ulike delkompetanser. Hvor stor praktisk betydning dette har for målingen er imidlertid en annen sak, og må vurderes ut ifra hva som er formålstjenlig i forhold til hensikten med måleinstrumentet.

5.2.3 Underdimensjonenes praktiske betydning for instrumentet

Et instrument er sannsynligvis å betrakte som én-dimensjonalt dersom analysen viser at en mulig andre dimensjon har en egenverdi under 3 (Linacre, 2017a, s. 555). På bakgrunn av det kan en konkludere med at oppgavesettet som undersøkes her kan betraktes som én-dimensjonalt for måling av funksjonskompetanse. Mange benytter imidlertid ofte en mer rigid grense på egenverdi over 2 for om en bør vurdere om måleinstrumentet kanskje inneholder underdimensjoner (Linacre, 2017a). For sikkerhets skyld ble oppgaver lengst fra hverandre i hhv. undergruppe 1 og 3 tatt ut av datasettet helt til egenverdien kom under 2, for så å sammenlignes med det opprinnelige datasettet.



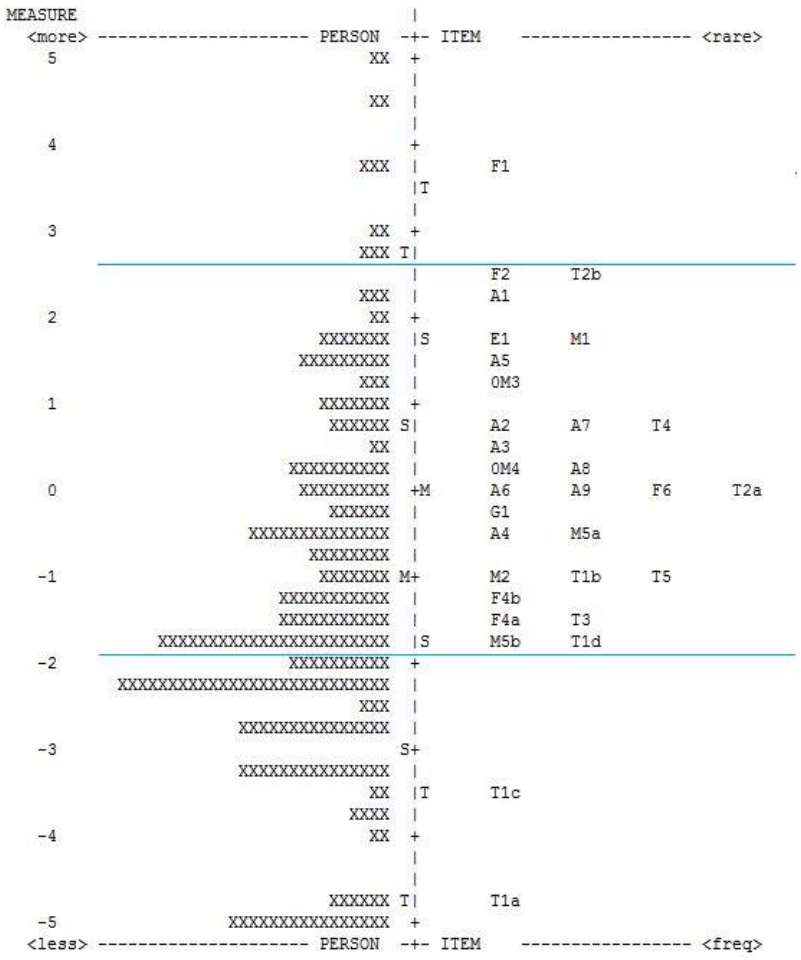
Figur 13: Mål på elever, redusert oppgavesett (x-akse) mot komplett oppgavesett (y-akse).

Figur XX viser at ingen av elevene får endrede mål som følge av fjerning av oppgaver fra de indikerte underdimensjonene. Analysen viser at selv om det har betydning for elevresultatene dersom elevene kun får oppgaver fra én dimensjon sammenlignet med om de kun får oppgaver fra en annen dimensjon, er den praktiske betydningen for målingene neglisjerbar så lenge det i måleinstrumentet er balanse mellom antall oppgaver fra hver dimensjon. At underdimensjonene kan betraktes som «strenger» innenfor variabelen funksjonskompetanse, og ikke som egne dimensjoner, ser ut til å være en rimelig konklusjon for dette instrumentet.

5.2.4 Instrumentets evne til å skille elevenes kompetansenivå

En reliabilitetskoeffisient på 0,98 og 0,86 for hhv. oppgaver og elever antyder at oppgavene er pålitelig estimert og at oppgavesettet som helhet gjør en god jobb med å skille elevene (Bond & Fox, 2015, s. 70-73). Reliabilitetskoeffisienten forteller oss imidlertid ikke noe om instrumentet gjør en like god jobb langs hele variabelen, noe vi ut ifra et vurderingsperspektiv i så stor grad som mulig ønsker at det skal.

Variabelkartet (figur 14) gir oss et visuelt bilde på hvor godt instrumentet vårt definerer variabelen og dermed også måler kompetansen til elevene. Elevene er angitt som X på venstre side av den vertikale, stiplete linjen i «midten», og oppgavene er plassert på høyre side av linjen. Helt til venstre ser vi måleskalaen i logits. Denne angir elevenes og oppgavenes plassering langs variabelen, og går i dette tilfellet fra -5 til +5. Jo lengre opp på skalaen, jo mer kompetent er eleven/ vanskeligere er oppgaven, og motsatt. Den nederste raden av elever (X-er) er de som ikke fikk til en eneste av testens oppgaver, mens de to øverste besvarte alle oppgavene korrekt. Disse elevenes kompetanse vil vi ikke kunne få et nøyaktig estimat på.



Figur 14: Variabelkart som viser elevenes kompetansemål og oppgavenes vanskegradsmål langs samme variabel.

Av variabelkartet kan vi se at elevene er forholdsvis godt fordelt langs variabelen. Det indikerer at vi har en god spredning av oppgaver med ulik vanskegrad, og at de gjør en brukbar jobb med å skille elevene (Wright & Stone, 1979, s. 9). Videre viser kartet at 28 av

31 oppgaver ligger i området -2 til +2,5 logits (avgrenset med blå linjer), noe som vil si at oppgavene skiller svært godt mellom elever som har kompetansemål i dette området.

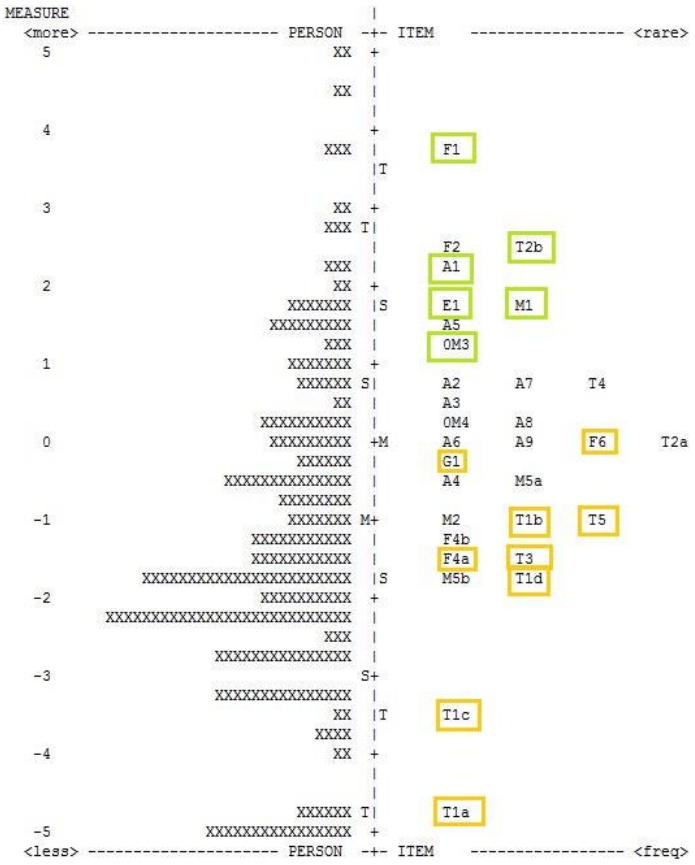
Vi registrerer imidlertid at 102 elever ligger utenfor disse grenseområdene, og at 60 av disse elevene plasserer seg mellom målene til oppgave M5b /T1d og T1c. Konsekvensen av dette er at det for disse elevene vil være vanskelig å angi en nøyaktig posisjon langs variabelen, og dermed vanskelig å si noe om hva som skiller disse elevene i form av kompetanse.

Vi observerer også at gjennomsnittsmålet på elevene og gjennomsnittsmålet på oppgavene (angitt som en M langs variabelen) spriker med over én logit. Dette forteller oss at testen som helhet er for krevende sammenlignet med elevenes kompetansenivå. Dersom det er vårt ønske både å kunne skille elevene og å kunne si mest mulig om den enkeltes kompetanse, kan vi tyde variabelkartet som at instrumentet har behov for flere enkle oppgaver som plasserer seg mellom den tredje letteste oppgaven og den nest letteste. Dernest kan det kanskje prioriteres å forsøke å finne oppgaver mellom de to letteste oppgavene.

5.2.5 Oppgavetype gir kvalitativ mening til ulike kompetansenivå

Ved kategorisering og utvelgelse av oppgaver ble det ut i fra teori anslått i hvilken av kategoriene lett, middels og vanskelig hver enkelt oppgave ville plassere seg. Her ble det i samsvar med Nitch et al. (2015) og Leinhardt et al. (1990) vektlagt om oppgavene var av typen tolkning/identifisering, konstruksjon eller forklaring/begrunnelse. I tillegg ble det hensynstatt om oppgavene kunne betegnes som lokale eller globale (Leinhardt et al., 1990; Bossé et al., 2011), samt mer generelle kognitive krav som hvor mange operasjoner oppgaven krevde fram mot en løsning (OECD, 2013). Av variabelkartet kan en nå sammenligne oppgavenes relative plassering i forhold til forventingene ut ifra teori. Gjennom dette undersøkes instrumentets substansielle validitet (Wolfe & Smith, 2007).

I figur 15 er oppgavene vi forventet skulle være de vanskeligste markert med grønn ramme, og de oppgavene vi forventet skulle være lettest er markert med oransje ramme.



Figur 15: Oppgavenes empiriske plassering langs variabelen. Grønne var forventet vanskelige, oransje var forventet enklest.

Variabelkartet viser at oppgavenes plassering i stor grad kan forklares ut ifra teorien som lå til grunn for forventningene, og således kan teorien være med på å gi elevmålene kvalitativ mening. De største avvikene er det oppgave G1, hvis intensjon var å måle forkunnskaper knyttet til koordinatsystemet, og begrepsoppgavene F2 og F6 som står for. Oppgave F2 og F6 er nesten identiske oppgaver der elevene skal finne stigningstallet til et funksjonsuttrykk. Det som skiller dem er at stigningstallet i F2 er på brøkform, som gjør denne oppgaven vanskeligere enn F6. En utbredt misoppfatning ser ut til å ha gjort disse oppgavene vanskeligere enn forventet ut ifra teori. Det samme gjelder oppgave G1. Disse vil bli kommentert i siste del av resultatkapittelet.

En interessant observasjon var at gjennomsnittsmålet for elevene markerer et skille der hvor oppgavene går over til å kreve kunnskap om det algebraiske symbolspråket. Foruten oppgave M2 krever ingen av oppgavene med mål fra -1 logit og nedover verken tolking av, eller

konstruksjon av et funksjonsuttrykk. Dette kan tyde på at algebraisk symbolkompetanse kan utgjøre en terskel for videre avansement langs variabelen.

Oppgave M2 skiller seg imidlertid fra de vanskeligere oppgavene ved at det er den eneste oppgaven i testen som ber om en «likning» istedenfor et «funksjonsuttrykk» i oppgavebestillingen. Ettersom oppgaven ber elevene uttrykke en likning som beskriver sammenhengen mellom to personers alder, hvor den ene er syv år yngre enn den andre, skiller oppgave M2 seg fra de andre modelleringsoppgavene ved at en i større grad kan forholde seg til variablene i oppgaven som statiske, konstante verdier. En feilaktig tolking av variablene som at de for eksempel er representasjoner for *personen* x og *personen* y, istedenfor å representere person x og y *sin alder*, vil for eksempel føre til en korrekt løsning på oppgaven. Av dette bemerkes det at elevene ved denne oppgaven ikke nødvendigvis behøver å ta hensyn til et forholdstall som beskriver sammenhengen mellom x og y. En additiv tankegang (person x + 7= person y) vil føre til riktig løsning.

5.2.6 Fordeling av vanskegrad på de fire delkompetansekategoriene

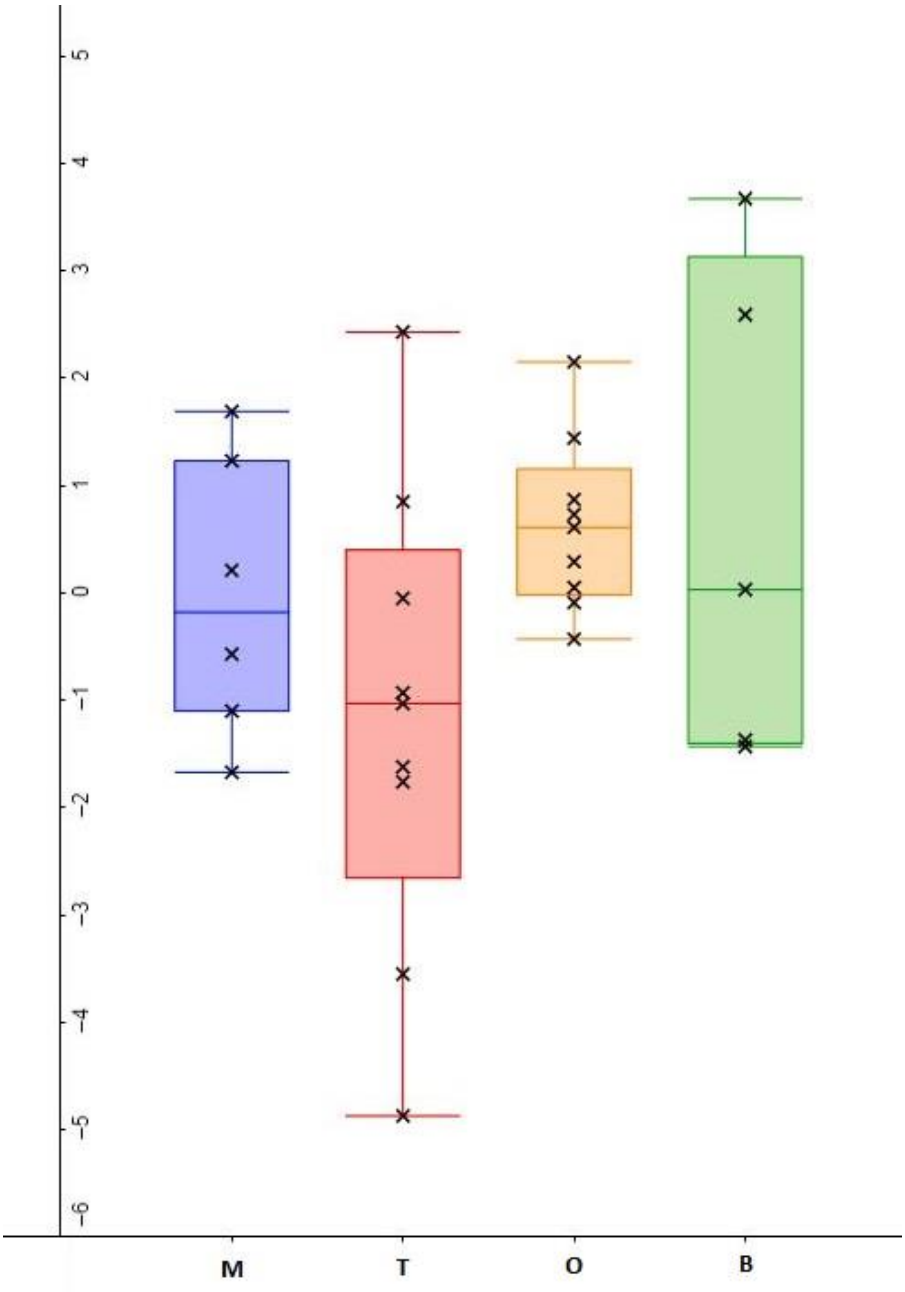
Funksjonskompetanse er i denne studien inndelt i de fire hovedkategoriene modellerings-, tolknings-, oversettelses- og begrepskompetanse. Da er det interessant å se hvorvidt oppgavenes vanskegrad har lik fordeling i alle kategorier.

Figur 16 viser at fordelingen er ganske lik på tvers av de fire kategoriene. Unntakene er de to letteste oppgavene i settet og den vanskeligste. Det er også tydelig at den ene kategorien, oversettelse, har en litt mindre variasjonsbredde enn de andre.

En Levene's test viste at variansen i gruppene ikke er signifikant heterogen ($W_{2,99} = 1{,}94$; $p = 0{,}148$), og en enveis ANOVA viste at gruppenes gjennomsnittlige vanskegrad ikke er signifikant forskjellig ($F_{2,99} = 2{,}05$; $p = 0{,}132$). Oppgavene i hver av kategoriene kan således betraktes som å utgjøre komplementære deler av funksjonskompetansen.

I planleggingen av undersøkelsen ble det imidlertid lagt til to oppgaver (T2b og T4) innenfor kategorien tolkning. Årsaken var at det manglet oppgaver som krevde mer globale tolkninger av grafene. I endringen av oppgavesettet fra den første til den andre runden ble det også lagt til en oppgave i kategorien oversettelse (A9). Uten disse tre oppgavene er det signifikant

forskjell mellom kategorienes gjennomsnittlige vanskegrad (Enveis ANOVA; $F_{3,05} = 4,67$; $p = 0,011$). Gruppenes varians var ikke signifikant heterogen (Levenes test; ($W_{3,05} = 1,79$; $p = 0,179$). Tolkningsoppgavene i de opprinnelige emneprøvene var med andre ord enklere enn de andre oppgavene.



Figur 16: Boksplott som viser oppgavefordelingen ut ifra vanskegrad innad i hver delkompetanse-kategori. Kryss er oppgaver. Farget område viser interkvartil bredde. Streken i hver boks viser medianen.

5.2.7 Enkeltoppgavenes måleegenskaper

Analyser av infit og outfit MNSQ forteller oss hvor godt oppgavene passer Rasch-modellen. At oppgavene produserer data som i tilstrekkelig grad passer modellen er nødvendig dersom en skal kunne oppnå invariante mål på intervall-nivå (Bond & Fox, 2015, s. 265). Å undersøke hvor godt oppgavene passer Rasch-modellen vil være en undersøkelse av oppgavenes tekniske kvalitet, og knyttes til validitetsaspektet *innhold* (Wolfe & Smith, 2007).

Analysene viste at oppgavenes infit MNSQ-verdier lå mellom 0,71 og 1,27, noe som betraktes som gode. Verdiene for outfit MNSQ hadde litt større variasjon, med verdier mellom 0,3 og 1,5. I motsetning til ved estimeringen av infit MNSQ vektlegges alle responser likt ved estimeringen av outfit MNSQ (Bond & Fox, 2015). Det betyr at outfit MNSQ-verdiene i større grad påvirkes av uventede responser fra personer med kompetansemål langt over eller langt under oppgavens vanskegrad.

Av tabell 6 ser vi at oppgave F4b har en outfit-verdi som ligger i grenseland for hva som kan ansees som produktivt for måling, Med en outfit MNSQ på 1,5 har oppgaven 50% mer variasjon i svarmønsteret enn hva Rasch-modellen forventet. Høye fit-verdier kan være av større trussel for måleinstrumentet enn lave verdier (Linacre, 2017a). En nærmere undersøkelse identifiserte to rimelig kompetente elever som på mange måter hadde tenkt riktig, men likevel formulert et svar på oppgaven som ikke kunne godtas. Ved midlertidig å omkode disse responsene til 9 (missing data) fikk oppgaven en infit/outfit på 1,14/1,10. Dermed ser det ut til at oppgaven fører med seg vesentlig mindre støy enn hva fit-analysen først indikerte.

Fire oppgaver har meget lave outfit-verdier, mellom 0,3 og 0,42. Fit-verdier i dette området antyder at oppgavene frembringer et svarmønster som er for godt i forhold til hva Rasch-modellen forventet. Et for godt svarmønster vil innebære at elevene som har mål under denne oppgaven i stor grad feiler på oppgaven, mens elever som har mål over denne oppgaven i stor grad avgir korrekte svar. Dette kan i utgangspunktet betraktes som en ønskelig egenskap hos en oppgave, men Rasch-modellen forventer at det er mer realistisk at det «svinger» litt mer der oppgavens vanskegrad er nær en elevs kompetansemål (Bond & Fox, 2015, s. 272). Oppgaver med lave fit-verdier har imidlertid sjelden noe praktisk betydning for måleinstrumentet i negativ forstand, og er å foretrekke fremfor oppgaver med høye fit-verdier (Bond & Fox, 2015, s.271). I dette tilfellet kan det se ut som at årsaken til verdiene skyldes at

oppgavene er blant de vanskeligste i testen, og en undersøkelse av oppgavenes ICC-er viste at nesten alle elevene mislykkes med oppgaven, inntil vi når et vendepunkt hvor et resterende mindretall lykkes med oppgaven. På bakgrunn av det kan vi si at oppgavene skaper et ganske skarpt skille mellom elever med høy kompetanse og resten, men at dette ikke forringer måleinstrumentet nevneverdig.

Med infit MNSQ i området 0,7-1,3 og outfit MNSQ < 1,5 antydes det at oppgavene kan utgjøre et brukbart instrument for måling av funksjonskompetanse.

Tabell 6: Oppgaver sortert i synkende rekkefølge etter grad av "misfit". Infit/outfit MNSQ markert med grønn ramme. Rød og gul ramme markerer oppgavene med hhv. høyest og lavest outfit MNSQ.

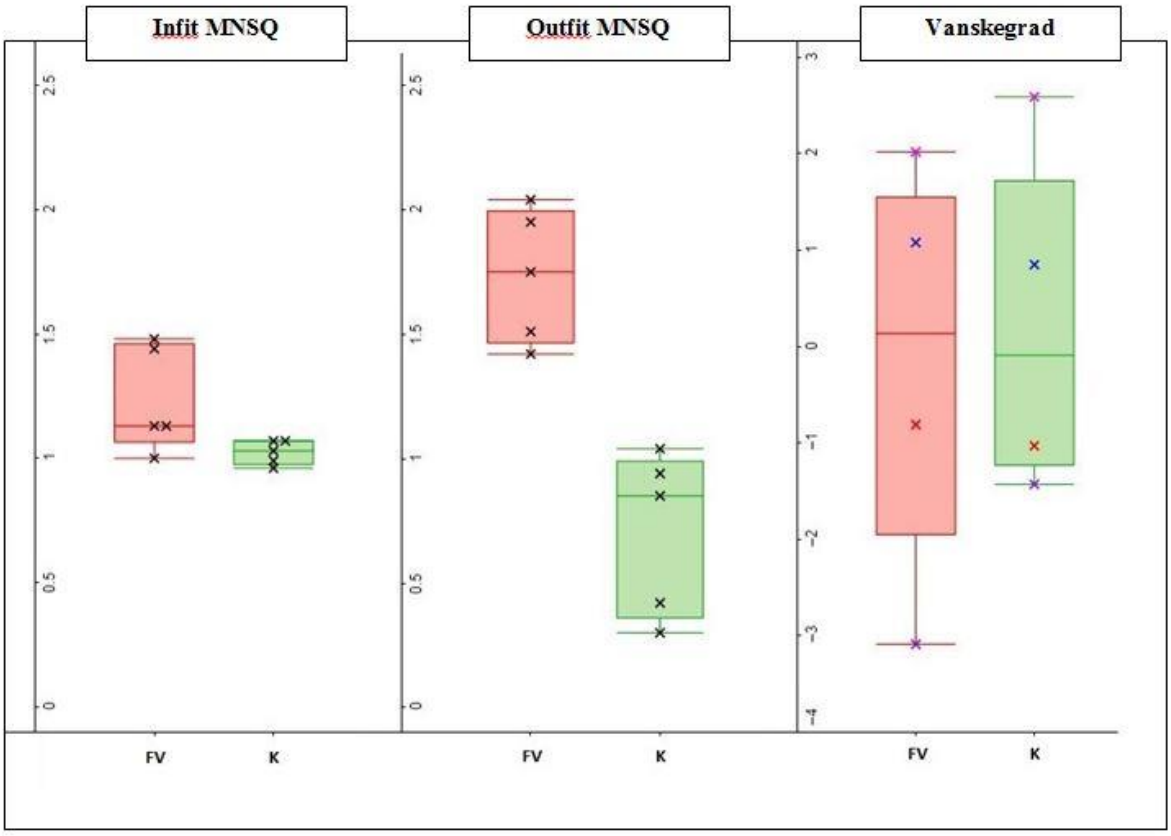
ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ITEM
15	79	158	-1.37	.20	1.21	2.3	1.50	1.7	A .48	.57	71.5	75.2	F4b
21	111	240	-1.10	.17	1.26	3.0	1.41	1.8	B .52	.60	71.4	77.0	M2
3	127	236	-1.62	.17	1.18	2.3	1.39	1.7	C .53	.60	70.6	76.1	T3
5	99	230	-.93	.17	1.24	2.6	1.39	1.8	D .53	.60	72.1	77.6	T1b
16	75	245	.03	.18	1.21	2.0	1.37	1.5	E .53	.60	76.9	81.7	F6
12	35	242	1.69	.24	1.19	1.4	1.36	.9	F .47	.53	88.9	88.8	M1
37	14	82	1.23	.35	1.27	1.5	1.29	.6	G .37	.45	76.7	83.7	0M3
8	75	235	-.05	.18	1.24	2.3	1.15	.7	H .53	.60	74.1	81.3	T2a
38	23	82	.21	.31	1.17	1.1	1.16	.5	I .51	.56	75.3	79.8	0M4
7	129	232	-1.76	.17	1.11	1.5	1.13	.6	J .56	.59	72.4	75.9	T1d
1	9	160	3.67	.46	1.07	.3	.30	-.8	K .50	.49	95.4	95.9	F1
25	34	158	.85	.26	1.07	.5	.85	-.3	L .60	.61	86.1	87.0	T4
9	23	235	2.43	.28	1.06	.4	.73	-.3	M .47	.48	90.9	92.2	T2b
14	83	161	-1.43	.20	1.03	.4	1.04	.3	N .55	.57	74.0	75.1	F4a
18	28	163	1.44	.28	1.03	.2	.90	-.1	O .60	.61	89.7	89.0	A5
27	132	241	-1.67	.17	.99	-.1	1.03	.2	P .59	.59	76.4	75.9	M5b
4	201	228	-4.87	.30	1.01	.1	.71	-.4	o .54	.54	93.4	93.4	T1a
22	52	158	-.13	.22	1.01	.1	1.00	.1	n .61	.61	79.6	81.4	G1
11	74	163	-1.03	.20	.99	.0	.94	-.1	m .59	.59	77.6	76.9	T5
23	34	241	1.75	.24	.98	-.1	.68	-.7	l .54	.52	89.4	89.1	E1
29	16	161	2.59	.36	.96	-.1	.42	-.8	k .58	.55	92.2	93.1	F2
6	178	227	-3.55	.21	.93	-.5	.70	-.7	j .58	.56	87.3	85.4	T1c
10	55	238	.73	.20	.91	-.8	.84	-.4	i .60	.57	86.5	84.7	A2
2	27	240	2.15	.26	.84	-1.0	.40	-1.4	h .55	.50	92.4	91.1	A1
26	93	241	-.57	.17	.81	-2.2	.74	-1.4	g .66	.60	81.8	78.6	M5a
17	90	244	-.43	.17	.75	-2.9	.71	-1.5	f .68	.61	86.4	79.6	A4
20	52	243	.87	.21	.75	-2.3	.41	-2.1	e .66	.57	88.1	85.6	A7
19	50	162	.05	.23	.74	-2.2	.58	-1.8	d .70	.62	87.7	82.8	A6
24	67	241	.29	.19	.74	-2.7	.59	-1.8	c .67	.59	88.4	82.6	A8
13	59	244	.61	.20	.70	-2.9	.49	-2.0	b .67	.58	88.2	84.3	A3
28	52	159	-.09	.22	.70	-2.7	.55	-2.0	a .71	.61	87.5	81.7	A9

5.2.8 Konstruksjonsoppgaver bedre enn flervalgsoppgaver

Som beskrevet i metodekapittelet ble det på bakgrunn av analyser av datasettet etter runde 1 gjort noen endringer på enkelte oppgaver. Oppgavene som ble endret var primært de som var

flervalgsoppgaver, og årsaken var at alle disse toppet listen over «most misfitted items», med infit MNSQ-verdier mellom 1,00 og 1,47, og outfit MNSQ-verdier mellom 1,47 og 2,86. Ved å endre disse til konstruksjonsoppgaver ble det antatt at oppgavene ville få bedre diskrimineringsegenskaper. Variabelkartet viste imidlertid at oppgavesettet ikke traff elevene så presist som en skulle ønske. For å få mer informasjon om kompetansen til elevene med mål under oppgave M5b og T1d hadde testen behov for flere enkle oppgaver. Fra presentert teori av Nitsch et al (2015) var det grunn til å forvente at en endring fra oppgavekategorien identifikasjon til konstruksjon derimot ville gjøre oppgavene vanskeligere. Samme konsekvens kunne forventes av en endring fra kategorien tolkning/identifikasjon til kategorien forklare/begrunne. En sammenligning av oppgavenes vanskegradsmål før og etter endring er derfor også relevant for å se hvor stor betydningen av endringene totalt sett var for måleinstrumentet.

Figur 17 viser infit MNSQ-verdier, outfit MNSQ-verdier og vanskegradsmål til oppgavene før og etter endring. Fem flervalgsoppgaver ble omskrevet og videreført fra det første oppgavesett til det andre. Sammenligningen av fit-verdier er for disse fem oppgavene. Ved sammenligningen av vanskegrad er oppgave F1 utelatt ettersom omgjøringen medførte endring av scoringen av oppgaven. Dermed er det de fire oppgavene F2, T4, F4a og T5 som undersøkes her.



Figur 17: Før og etter endring av oppgavene. Flervalgsoppgaver = FV, Konstruksjonsoppgaver = K. Kryss markerer oppgavenes plassering før og etter endring. De fargelagte kryssene der hvor endring i vanskegrad illustreres viser hver enkelt oppgaves plassering før og etter den ble endret.

Omgjøringen fra flervalgsoppgave til konstruksjonsoppgave medførte merkbare endringer av både infit- og outfit-verdier. Samtlige fem oppgaver la seg tett omkring infit MNSQ = 1, med verdier mellom 0,96 og1,07. En parvis t-test viste også at forskjellene var signifikante, p = 0,047.

For outfit MNSQ var effekten av oppgaveomgjøringene enda større. En parvis t-test viser at også disse forskjellene er signifikante, p = 0,013. Det innebærer at oppgavene frembringer langt færre uventede responser nå enn tidligere. To av oppgavene fikk imidlertid i overkant lave outfit-verdier, men noen lave er som tidligere nevnt å foretrekke fremfor høye. Oppgavenes ICC-er gir ytterligere informasjon om hvordan diskrimineringsegenskapene for oppgavene bedret seg, og en sammenligning av disse for oppgave F4 og T4 kan sees i appendiks A.

Hva så med vanskegraden? To av oppgavene ble betydelig vanskeligere, mens de to andre faktisk fikk noe lavere verdier. De to sistnevnte oppgavenes nye verdier er likevel innenfor standardfeilen til estimeringen av de opprinnelige, så en kan si at omgjøringen trolig ikke endret vanskegraden til disse oppgavene. Samtidig bemerkes det at standardfeilen ikke tas hensyn til ved sammenligningen her; middelverdien av denne er lagt til grunn. Sammenligner vi oppgavene som gruppe kan vi si at testen som helhet ikke ser ut til å ha blitt vesentlig vanskeligere av å endre disse oppgavene, og en parvis t-test viser også at endringene ikke er signifikante, p = 0,39.

På bakgrunn av at infit- og outfit-verdiene til oppgavene ble signifikant forbedret, samtidig som vanskegraden ikke ble signifikant endret, må vi kunne si at endringene totalt sett ser ut til å være til det bedre for måleinstrumentet.

5.2.9 Oppgavesettet er invariant ift kompetansenivå

Lik som på instrumentnivå ønsker vi også på enkeltoppgavenivå å oppnå invariante mål, og som tidligere nevnt på side 16 innebærer bl.a. dette at en enkel oppgave bør oppleves enklere for alle, mens en vanskelig oppgave bør oppleves vanskeligere for alle, uavhengig av kompetansenivå (Sjaastad, 2014). Dersom dette ikke er tilfelle vil det kunne ha en negativ konsekvens for generaliserbarhetsaspektet ved instrumentets validitet (Wolfe & Smith, 2007). En parvis DIF-analyse (differential item functioning) for alle oppgavene, hvor vi delte respondentene inn i en øvre og en nedre halvdel basert på kompetansemål, ble derfor gjennomført. Her ble Rasch-Welch-metoden, som er en logistisk regresjonsmodell, benyttet. I henhold til Linacre (2017a, s. 538) ble en grense på 0,64 logits i «DIF-contrast» (forskjellen i oppgavens vanskegrad mellom de to elevgruppene) sammen med en Bonferroni-korrigert p-verdi <0,0016 (ettersom det gjennomføres en hypotese-test for hver enkelt oppgave) lagt til grunn for nærmere undersøkelse av mulige problematiske oppgaver.

DIF-analysen viste at instrumentet som helhet består av invariante oppgaver. To oppgaver ble imidlertid identifisert med verdier utenfor grensene som ble satt. En kvalitativ undersøkelse av de to oppgavene ble derfor utført.

Oppgave A3 favoriserte gruppen med over middels kompetanse. Dette var en oversettelsesoppgave hvor den grafiske representasjonen av funksjonen $f(x) = -x + 3$

skulle oversettes til dens funksjonsuttrykk. Oppgavens ICC viste også tydelig hvordan oppgaven underdiskriminerer for elever fra gjennomsnittlig kompetansemål og nedover, for så å overdiskriminere for elever med kompetansemål fra rundt gjennomsnittet og opp (Appendiks B). Kvalitativt skiller oppgaven seg ut ved at funksjonen har et negativt stigningstall, og at det er dette som gjør oppgaven tilnærmet uoppnåelig for elevene i nedre halvdel av kompetanseskalaen.

Oppgave T3, som ut ifra en oppgitt x-verdi ba elevene finne tilhørende y-verdi, favoriserte derimot gruppen med under middels kompetanse. Inspeksjon av oppgavens ICC viste også hvordan oppgaven tydelig overdiskriminerer for elever med kompetansemål -3 til 0 logits. Den mest åpenbare årsaken synes å være en teknisk mangel ved administrasjonen av oppgaven, og ikke oppgaven i seg selv. Y-verdien skulle ikke være mulig å lese av direkte, men var ment å måtte identifiseres ut ifra skala-inndelingen og sees relativt til andre oppgitte verdier langs aksen. Med touch-screen på elevenes Chromebook viste det seg imidlertid mulig å zoome inn helt til den eksakte verdien åpenbarte seg, noe som endret rammene fullstendig. Det synes rimelig å forvente at elever som mangler andre strategier for å finne verdien vil være de som er mest tilbøyelige til å undersøke muligheten for å zoome, og at dette kan være en plausibel forklaring på at oppgaven viste statistisk signifikant DIF. Dette skaper før øvrig i tillegg usikkerhet om hvorvidt vanskegraden til denne oppgaven er pålitelig estimert.

5.3 Uventet vanskelige oppgaver og utbredte misoppfatninger

Ved sammenligningen av oppgavenes empiriske vanskegrad opp imot forventet vanskegrad ble det påpekt at oppgave F2, F6 og G1 var vesentlig vanskeligere enn på forhånd antatt. En kvalitativ undersøkelse av disse oppgavenes besvarelser ble derfor gjort.

Oppgave F2 og F6 ba elevene identifisere signingstallet i et funksjonsuttrykk. Der hvor elevene har avgitt svar på disse oppgavene kan majoriteten av de feilaktige svarene deles inn i to grupper. Den ene gruppen består av besvarelser hvor elevene ikke skiller mellom stigningstallet og den uavhengige variabelen x, slik at stigningstallet til hhv. oppgave F2 og F6 oppgis som $\frac{x}{4}$ og $3x$ istedenfor $\frac{1}{4}$ og 3. Den andre gruppen består av besvarelser hvor funksjonens konstantledd oppgis som funksjonens stigningstall. Denne siste gruppen besvarelser ble sett i sammenheng med øvrige oppgaver hvor elevene må foreta en overgang

til eller fra et funksjonsuttrykk. Denne undersøkelsen avdekket at omkring hver femte elev viste tegn til å forveksle funksjonens endringsrate med funksjonens konstantledd.

Ved oppgave G1, hvor elevene skulle tegne en rett linje gjennom to oppgitte punktkoordinater, viste igjen omkring en av fem elever tegn på samme type misoppfatning. De to koordinatene (-1,3) og (1,0) tolkes av disse elevene som fire punkter i koordinatsystemet istedenfor to. Resultatet blir at det tegnes to rette linjer gjennom fire punkter.

Disse utbredte misoppfatningene ser ut til å forklare hovedårsaken til at de tre oppgavene avvek så mye fra det som på forhånd var forventet. De tyder også på at mange elever mangler nødvendige forkunnskaper, og at manglene er store innenfor de to kompetanseområdene symbolkompetanse og representasjonskompetanse i modellen til Niss og Jensen (2002).

6 Drøfting

Hensikten med denne studien var å forsøke å besvare 1) hvordan funksjonskompetanse operasjonaliseres i de tre forlagsgitte emneprøvene, 2) hvor egnet emneprøvenes oppgaver er til å måle det de søker å måle og 3) hvilke tiltak som eventuelt kan gi bedre emneprøver.

Jeg fant at de forlagsgitte emneprøvene hadde mange like oppgavetyper som lot seg inndele i fire delkompetanse-kategorier, samt en forkunnskap-kategori, og at disse oppgavene sammen utgjorde et brukbart måleinstrument for funksjoner. Samtidig fant jeg at det var ulikt hvordan delkompetansene var vektlagt i hver av de tre prøvene, og dermed også at fordelingen av enkle og vanskeligere oppgaver varierte mellom prøvene. Dette peker mot at prøvene kan ha ganske forskjellige måleegenskaper, og noen tiltak for forbedringer vil på bakgrunn av analysene bli foreslått. På enkeltoppgavenivå fant jeg at oppgavene, med unntak av noen flervalgsoppgaver, gjorde en tilfredsstillende jobb med å måle kompetanse innenfor funksjoner. I dette kapittelet vil jeg utdype disse funnene nærmere og diskutere hva de impliserer med tanke på bruk av emneprøver som måleinstrument for matematikkompetanse. Det vil også bli viet et delkapittel til noen utbredte misoppfatninger som ble avdekket blant elevene.

6.1 Funksjonskompetanse – måleinstrumentet og teorien

Ut fra det teoretiske rammeverket som ble lagt til grunn for studien hadde vi en forventning om hvordan de ulike oppgavene ville plassere seg langs kompetansevariabelen. Det var blant annet forventet at lokale tolkninger ville være lettere enn mer globale tolkninger (Duval, 2006; Leinhardt et al. 1990), og at tolkningsoppgaver generelt ville være lettere enn konstruksjonsoppgaver (Bossé et al., 2011; Hattikudur et al., 2012; Leinhardt et al., 1990). Både innad i kategorien tolkning av representasjoner og innad i kategorien oversettelser mellom representasjoner viser tidligere studier at vanskegraden til oppgaver øker etter hvert som oppgavene går fra å betegnes som lokale til å betegnes som globale aktiviteter (Bossé et al., 2011; Leinhardt et al., 1990). Oppgaver som krevde at elevene måtte forklare eller begrunne var forventet å være de vanskeligste (Nitsch et al., 2015). Disse forventningene ble i stor grad møtt da vanskegraden for oppgavene i emneprøvene ble beregnet med Rasch-

analyse. I de tre tilfellene det var avvik ble det gjennom kvalitative undersøkelser identifisert sannsynlige årsaker til avvikene.

Den nedre delen av variabelkartet domineres av oppgaver som stiller krav til lokale tolkninger, den midterste domineres av oversettelsesoppgaver som kan karakteriseres som globale oversettelsesaktiviteter, men hvor flere kan brytes ned til to lokale oversettelsesaktiviteter (Bossé et al., 2011), og den øverste delen domineres av oppgaver som stiller krav til globale betraktninger av funksjonene. På bakgrunn av dette kan oppgavenes fordeling langs variabelen sies å samsvare godt med nevnte studier, selv om disse studienes resultater har sitt utspring fra andre metodegrunnlag enn det denne har. Ut fra resultatene i denne studien kan vi beskrive en kompetansemodell hvor økning i grad av kompetanse beskrives som en forflytning fra å kunne identifisere/tolke representasjoner, til å kunne konstruere representasjonsoverganger, videre til også å kunne forklare og begrunne sentrale aspekter ved representasjonene (parallelt med en gradvis forflytning fra lokale til globale betraktninger og aktiviteter). Det påpekes at dette er en generell betraktning hovedsakelig på bakgrunn av kvantitative data, og at det vil være flere faktorer som spiller inn på en oppgaves vanskegrad. Økning i antall operasjoner som kreves frem mot en korrekt løsning er ett eksempel på en slik faktor, og denne har også sin plass i det totale bildet her.

6.1.1 Emneprøvenes dekning av funksjonskompetanse

Fordelingen av oppgaver i kompetanse-kategorier for hver av de tre emneprøvene var ulik for noen av kategoriene (se tabell 3). Størst forskjell var det mellom emneprøve B og C, hvor førstnevnte hadde langt flere oppgaver innenfor kategorien tolkning enn innenfor oversettelse. For emneprøve C var det motsatt. Kategorien begrepskompetanse var ikke representert i emneprøve B.

Variansanalysen viste at det ikke var signifikant forskjell i gjennomsnittlig vanskegrad på tvers av kompetanse-kategoriene for måleinstrumentet som ble benyttet i studien. Dette tyder på at hver av kategoriene utgjorde komplementære komponenter av funksjonskompetanse, snarere enn stegvise kompetanser som bygger på hverandre (O'Callaghan, 1998). Da tillagte oppgaver ble tatt ut av datasettet, slik at instrumentet kun besto av oppgaver fra emneprøvene, viste variansanalysen at forskjellen var signifikant. Ettersom det var de to vanskeligste

tolkningsoppgavene som ble tatt ut betyr det at det var i denne kategorien den gjennomsnittlige vanskegraden ble signifikant lavere. Selv om vi i denne studien ikke har mulighet til å sammenligne hver av emneprøvene mot hverandre, kan dette indikere at emneprøvene spriker både i dekningsgrad av delkompetansene og vanskegrad på prøvene. I en vurderingssammenheng i skolen kan dette i så fall bety at elevenes kompetansemål på tvers av skoler vil kunne vurderes ulikt avhengig av hvilken emneprøve som benyttes. Det kan også bety at elevenes kompetansemål relativt til hverandre innad i en elevgruppe vil kunne bli ulik avhengig av hvilken emneprøve som benyttes. Som en konsekvens (dersom vi forutsetter ukritisk bruk av emneprøver) kan dette i så fall medføre utslag i form av ulike karakterer på en og samme kompetanse.

6.2 Instrumentets måleegenskaper

Hvor gode måleegenskapene til instrumentet kan sies å være vurderes ut ifra dokumentasjonen som er innhentet med hensyn på instrumentets reliabilitet og validitet (Wu & Adams, 2007). Over er instrumentets substansielle validitet adressert **(6.1).** I dette delkapittelet vil måleinstrumentets egenskaper drøftes med hensyn på øvrige validitetsaspekter som *innhold, strukturell validitet, generaliserbarhet, konsekvensiell validitet* og *tolkbarhet.*

6.2.1 Måleinstrumentet og kravene til måling

En kort oppsummering av analyseresultater underbygger et overordnet bilde av at emneprøvenes oppgaver gjorde en god jobb med å måle funksjonskompetanse. Infit MNSQ fra 0,7 til 1,27 og outfit MNSQ < 1,5 indikerte at oppgavene ga forholdsvis invariante mål. Tilfredsstillende infit-verdier tyder også på at oppgavene definerer sentrale deler av hva som kan betraktes som funksjonskompetanse. I tillegg kunne instrumentet sies å være tilstrekkelig én-dimensjonalt tatt i betraktning dets hensikt. Spredningen av elevene var også stor langs variabelen, med elever som både får mål over den vanskeligste oppgaven og elever som får mål under den letteste. En reliabilitetskoeffisient for elevene på 0,86 underbygger at instrumentets evne til å skille elevene må betraktes som akseptabel.

Fordelingen av oppgaver i form av vanskegrad var god i området -2 til +2,5 logits. Her gjør altså oppgavene en god jobb med å skille elevene. Når den innbyrdes avstanden mellom oppgavene er tilnærmet lik vil måleinstrumentet ha den fordelen at hver korrekte løsning på en oppgave medfører den samme intervalløkingen langs variabelen. Hvert ekstra poeng vil dermed tilsvare en tilnærmet konstant økning i kompetanse (Wright & Stone, 1979). Dersom dette gjelder for alle oppgavene som utgjør instrumentet vil vi på mange måter ha et tilnærmet ideelt måleinstrument.

En vanlig måte å vurdere matematikkprøver på er å telle opp antall oppnådde poeng (se side 16). Det er greit der hvor det er lik avstand mellom oppgavenes vanskegrad. For instrumentet som undersøkes her er dette imidlertid ikke tilfelle, ettersom det inneholder to oppgaver som er vesentlig lettere, og en oppgave som er vesentlig vanskeligere, enn instrumentets øvrige oppgaver. Som en konsekvens vil et hopp fra å få til de to første oppgavene til å få til de tre første innebære en langt større økning i kompetanse enn et hopp fra for eksempel de 13 første til de 17 første oppgavene (Wu & Adams, 2007; Bond & Fox, 2015). Dersom en benytter råscore (antall poeng) som mål på kompetanse vil det ofte tolkes motsatt, ettersom det første hoppet tilsvarer 3 % økning i poeng, mens det andre tilsvarer 13 % økning i poeng. I ytterste konsekvens kan dette medføre ulike karakterer på elever som egentlig kan sies å vise tilnærmet lik kompetanse. Dersom slike uregelmessige avstander befinner seg flere/andre steder langs variabelen kan dette også medføre at elever som burde hatt ulike karakterer får like, simpelthen fordi avstanden opp til neste mulige poeng er uforholdsmessig stor. For at vurdering gjennom bruk av råscore skal kunne betraktes som en mest mulig rettferdig vurderingsform for dette måleinstrumentet har vi, som tidligere nevnt, behov for flere oppgaver i den nedre delen. I tillegg bør oppgaver med omtrent lik vanskegrad tas bort, revideres eller erstattes av andre oppgaver.

6.2.2 Måleinstrumentet og elevenes funksjonskompetanse

Det er ikke til å komme fra at testen som helhet var for krevende for en svært stor andel elever. Konsekvensene av dette er flere. For det første får vi i liten grad målt hva disse elevene får til, men nesten utelukkende hva de ikke får til. Dette er uheldig både for lærer og elev, ettersom det vil bli vanskelig beskrive elevens kompetanse. Da vil det også være vanskelig å kunne bruke resultatene til noe produktivt for videre læring. For det andre vil

mange elever oppnå like kompetansemål, selv om det er rimelig å anta at det finnes betydelige nyanser av funksjonskompetanse også for disse elevene. Spesielt ugunstig kan en si at avstanden på et helt standardavvik mellom den tredje letteste oppgaven og den nest letteste oppgaven viste seg å være. Før en diskuterer hva slags type oppgaver som kanskje kan inkluderes for å utfylle dette gapet synes det rimelig å diskutere elevresultatene opp imot den norske læreplanens kompetansemål for funksjoner. Disse kompetansemålene er tross alt styrende for utviklingen av læreverkene og dermed også de tilhørende emneprøvene, og det har tidligere blitt kommentert hvordan oppgavene kan sies å måle flere av læreplanens kompetansemål. På bakgrunn av det er det relevant å spørre om det heller var slik at elevene ikke traff testen, fremfor å ta utgangspunkt i at det var testen som ikke traff elevene. Det kan for eksempel argumenteres for at de to letteste oppgavene ikke betinger at elevene må forholde seg til en relasjon mellom de to variablene x og y. Kanskje er kompetansen som utvises gjennom disse to oppgavene da helt i grenseland mellom hva vi kan betegne som funksjonskompetanse og hva vi kan betegne som forkunnskaper? I så fall vil vi kunne si at oppgavesettet begynner målingen av funksjonskompetanse først ved oppgave M5a og T1d, og at det derifra må sies å gjøre en god jobb med både å definere variabelen og med å skille elevene. Det innebærer i så fall at et betydelig antall elever ikke viser tegn til å besitte det vi her har definert som funksjonskompetanse.

På en annen side så er oppgavesettet i denne studien kun en slags modell av de tre emneprøvene. Ved å velge ut oppgaver fra fellestrekk på tvers av emneprøvene ble flere oppgaver utelatt. Selv om det kan argumenteres for at oppgavesettet favner en betydelig andel av emneprøvenes totale antall oppgaver, kan en ikke se bort ifra at utelatte oppgaver kunne hatt påvirkning på resultatet hadde disse blitt inkludert. Eksempelvis er det blant annet blitt påpekt at alle oversettelsesoppgavene i settet var av typen globale oversettelsesaktiviteter. Av litteraturen fremkommer det at lokale oversettelsesaktiviteter generelt oppleves som enklere for elever enn de globale (Bossé et al, 2011). Et innslag av denne typen oppgaver hadde derfor trolig plassert seg lavere ned på måleskalaen (selv om det er vanskelig å anslå hvor). Det var imidlertid kun én slik oppgave i de tre emneprøvene til sammen, og den ble i denne omgang utelatt på grunn av utfordringer med å administrere oppgaven gjennom prøveverktøyet som ble benyttet.

6.2.3 Elevenes funksjonskompetanse

Jeg vil her utdype to misoppfatninger som viste seg å være utbredt blant elevenes besvarelser i denne undersøkelsen; en feilaktig tolkning og/eller kommunikasjon av en funksjons stigningstall og en feilaktig tolkning av koordinatangivelser. Den første misoppfatningen gikk ut på at elevene ved oppgaver som omhandlet lineære funksjoner identifiserte og/eller oppga konstantleddet som funksjonens stigningstall. Besvarelsene tyder på at elevene identifiserer variablenes samvariasjon, men leser og/eller uttrykker denne som en addisjon. To eksempler er en funksjon $y = 4x$ som angis som $y = x + 4$ og en funksjon $y = 30x + 20$ som tegnes som om funksjonen var $y = 20x + 30$. Det kan være ulike årsaker til dette fra elev til elev, og det er ikke grunnlag i denne studien for å spekulere videre omkring disse, men at en slik misoppfatning indikerer en mangelfull forståelse for det algebraiske symbolspråket synes rimelig å påstå. For mange av elevene med denne misoppfatningen førte den til feilaktige svar på mange av settets oppgaver. Et av de mer ekstreme eksemplene finner vi i elev nr. 158 som viser tydelige tegn på denne misoppfatningen ved hele 9 av instrumentets 31 oppgaver.

Oppgave G1 ba elevene tegne en rett linje gjennom punktene (-1,3) og (1,0), og ble av omtrent 1 av 4 elever besvart i form av to linjer, hvor den ene stort sett gikk gjennom punktet (-1,0) og (0,3) og den andre stort sett gikk gjennom punktet (1,0) og (0,0). Omfanget av denne misoppfatningen var overraskende, men belyser kanskje en utfordring ved at visse forkunnskaper som det kanskje forventes at er til stede hos elevene ikke er det likevel.

Utbredte misoppfatninger som de nevnte vil naturlig nok både påvirke estimeringen av oppgavenes vanskegrad og de aktuelle elevenes kompetansemål Mens den siste misoppfatningen tilbyr en forklaring på hvorfor oppgave G1 var så mye vanskeligere enn forventet, er begge typene misoppfatninger med på å underbygge et inntrykk av at elevenes lave kompetansemål er knyttet til at nødvendige forkunnskaper synes å ikke være på plass. De forkunnskapene som fremheves knyttes her i første rekke til aspekter ved en symbol- og formalismekompetanse som ansees som nødvendig for tolkning og anvendelse av sentrale representasjoner av funksjoner (Niss & Jensen, 2002; O'Callaghan, 1998). Svært mange elever utviser lave algebrakunnskaper i besvarelsene sine, og er således i tråd med resultatene fra internasjonale prøver (Grønmo & Hole, 2017).

6.3 Nye emneprøver med gode måleegenskaper

Vi har sett at oppgavesamlingen som helhet har gode måleegenskaper samtidig som emneprøvene hver for seg viste noe ubalanse i dekningen av kompetansekategorier og sannsynligvis varierte noe i vanskegrad. Hvordan kan vi lage prøver som gir en best mulig summativ vurdering av elevens kompetansenivå innen funksjoner og samtidig gir både elev og lærer mest mulig informasjon om hva som kan hjelpe eleven videre?

Dimensjonsanalysene indikerte at vi kunne betrakte måleinstrumentet som tilstrekkelig én-dimensjonalt for måling av funksjonskompetanse. Analysene viste imidlertid at det var signifikante forskjeller i mål flere av elevene fikk avhengig av hvilken undergruppe av oppgaver de responderte på. Forskjellene var for enkelte elever store mellom lokale tolkningsoppgaver og oversettelsesoppgaver, og analysene indikerte også at forskjellene var tilsvarende store mellom oppgaver med algebraisk symbolspråk sammenlignet med oppgaver uten. Fordi disse kategoriene er delvis uavhengige er det tilrådelig å balansere oppgavene fra disse kategoriene både i antall og i vanskegrad.

To av de tre emneprøvene som utgjorde oppgavedatabasen for undersøkelsen hadde innslag av flervalgsoppgaver. Disse oppgavene var i sin opprinnelige form en del av instrumentet i den første av de to undersøkelsene. Analysene viste at flervalgsoppgavene var de oppgavene som passet Rasch-modellen minst, hovedsakelig på grunn av høye outfit MNSQ verdier. Høye verdier indikerer uforutsigbare svarmønstre og er ofte et resultat av gjetting (Linacre, 2017a). Det var hovedsakelig oppgaver innenfor kategorien begrepskompetanse som var flervalgsoppgaver, og kategorien begrepskompetanse var faktisk ikke representert med andre enn denne oppgavetypen. På bakgrunn av disse observasjonene kan en diskutere to problematiske sider ved å inkludere disse oppgavene i en test. Den ene er hvorvidt vi ønsker å legge til rette for at gjetting kan gi uttelling, og den andre er hvorvidt det er formålstjenlig å la denne typen oppgaver være hovedkilden til informasjon om et av de sentrale aspektene ved kompetansen som måles. Det er både tidkrevende og kostbart å utvikle flervalgsoppgaver med gode måleegenskaper som i nasjonale og internasjonale prøver. Det er derfor ikke urimelig å forvente at flervalgsoppgaver i forlagsgitte emneprøver, så vel som i lokale lærerutviklete emneprøver, vil være av varierende kvalitet. I oppgavesettet som ble undersøkt her var noen gode, mens andre var mindre gode. Flervalgsoppgaver er tidsbesparende å rette, men tilbyr gjerne også en mer begrenset informasjon om en elevs kompetanse. Gjennom konstruksjonsoppgaver vil en potensielt kunne måle et bredere spekter av elevens

matematiske kompetanse, som for eksempel tankegangs- og resonnementskompetanse (Niss & Jensen, 2002).

Analysene av det endelige datasettet fra denne undersøkelsen tilsier at det er en fordel for instrumentet at flervalgsoppgaver blir endret til å være konstruksjonsoppgaver i stedet. Ved å velge konstruksjonsoppgaver over flervalgsoppgaver endres ikke vanskegraden, problemet med gjetting blir unngått og en får oppgaver med bedre måleegenskaper. I tillegg kan en få ekstra informasjon om elevenes kompetanse til formativ vurdering. Disse fordelene bør veie opp for den eventuelle besparelsen i tid med å rette flervalgsoppgaver.

Til sist kan det nevnes at instrumentet gir liten informasjon om elever som oppnår mål under den tredje letteste oppgaven i oppgavesettet. I denne studien gjaldt dette 40 % av elevene (100 av 250), og spesielt med tanke på at emneprøver gjerne har både en summativ og formativ funksjon i et vurderingsperspektiv kunne en med fordel vurdere å inkludere enklere oppgaver. Ut ifra både det teoretiske rammeverket og empiri fra denne studien ser vi at det vil være mest hensiktsmessig å finne oppgaver innenfor kategorien lokale tolkninger. For formativ vurdering kan også oppgaver som adresserer sentrale forkunnskaper vurderes inkludert.

6.4 Implikasjoner for bruk av emneprøver

Utgangspunktet for denne studien var en mangel på dokumentasjon av emneprøvers generelle kvalitet med hensyn på presisjon og rettferdig vurdering. Gjennom bruk av Rasch-analyser har sentrale måleegenskaper både ved det aktuelle instrumentet spesielt, men derigjennom også ved prøver generelt, blitt belyst. Selv om flere av måleegenskapene er vanskelige å identifisere uten å gå så grundig til verks som det har blitt gjort her mener jeg at noen slutninger kan trekkes ut fra resultatene som generelle råd for hva det er viktig å være bevisst ved tillaging og bruk av emneprøver.

Dimensjonsanalysene av instrumentet i denne studien belyser en helt fundamental forutsetning for et rettferdig og vurderingsmessig meningsfullt måleinstrument. En matematikkprøve vil stort sett være flerdimensjonal av natur (Sjaastad, 2014, s. 214). For studiens instrument ble det påvist at tolkningsoppgaver, oversettelsesoppgaver og oppgaver som involverte funksjonsuttrykk potensielt kunne utgjøre egne dimensjoner. Sett i lys av en

vurderingspraksis hvor elevers prestasjoner i stor grad vurderes ut ifra antall poeng oppnådd på testen vil en overvekt av en av disse oppgavetypene kunne være i betydelig disfavør for elever som har sine styrker innenfor andre deler av kompetansen som måles. Å forsøke å sikre en riktig balanse av oppgavetyper vil derfor være av stor betydning for å skape et rettferdig måleinstrument (Linacre, 2017a). Med balansering av oppgavetyper menes at en bør sikre at testen ikke inneholder en overvekt av oppgaver fra den ene eller den andre oppgavekategorien, men en jevn fordeling.

Dersom råscore skal benyttes som mål på elevenes kompetanse bør det også så langt som mulig sikres en mest mulig lineær relasjon mellom oppgavenes vanskegrad og grad av kompetanse. I dette ligger det at avstanden mellom oppgavene som definerer variabelen som søkes målt er forholdsvis lik langs hele målestaven (Wright & Stone, 1979). I en lærers daglige virke er det selvsagt urealistisk at emneprøver kan analyseres slik de er gjort i denne studien. Imidlertid belyser studien viktigheten av at tid til kvalitative analyser av oppgaver ut ifra egen kompetanse og teori bør prioriteres for å tilstrebe en målestav med forholdsvis like avstander mellom oppgavene. Oppgaver som kan begrunnes å være tilnærmet like krevende bør det således ikke inkluderes så mange av, da en marginal økning i kompetanse vil kunne belønnes med uforholdsmessig mange poeng. Av analysene av undersøkelsens måleinstrument kan det for eksempel argumenteres for at enkelte av oversettelsesoppgavene i så måte bør tas ut eller helst erstattes av andre mindre krevende oppgaver. Ved å gjøre disse vurderingene etter å ha definert og kategorisert sentrale delkompetanser vil en kunne øke muligheten for at vanskegraden er godt fordelt innad i de ulike delkompetanse-kategoriene, noe som vil kunne utgjøre mye for rettferdighetsaspektet ved prøven og ikke minst for prøvens evne til å gi kvalitativt meningsinnhold til elevenes kvantitative mål.

I etterkant av studien foreligger det nå et oppgavesett som har blitt grundig analysert og evaluert. Dette oppgavesettet kan dermed være et utgangspunkt for prøver der en kan plukke oppgaver slik at de blir så jevnt fordelt på vanskegrad og kompetansekategorier som mulig.

Til sist bør det nevnes en økt bevissthet omkring innslag av flervalgsoppgaver på prøven. Å lage gode flervalgsoppgaver er ingen enkel sak. Det gjør at denne typen oppgaver er sårbare for gjetting, og dermed bidrar med mindre (og i verste fall misvisende) informasjon om en elevs kompetanse. Ved å benytte konstruksjonsoppgaver i stedet vil en også kunne innhente dokumentasjon på flere områder av en elevs samlede matematiske kompetanse, og det uten at oppgavene nødvendigvis blir vanskeligere, noe resultatene i denne studien er et eksempel på.

6.5 Avslutning

Ved å gjennomføre denne studien har jeg, så vidt jeg kan se, bidratt med forskning som til nå ikke er blitt gjort. Ettersom læremiddelutviklere potensielt kan være betydelige premissleverandører for hvordan måling av elevkompetanse utføres i skolene virker det nødvendig å belyse ulike sider av hva dette kan innebære. Denne studien er et bidrag i så måte.

Studien antyder at det kan være forskjeller mellom emneprøvene som har vært gjenstand for analyser her, men samtidig at de ikke er veldig store. Studien har imidlertid avgrenset seg til de mest sentrale delene i et utvalg på tre emneprøver om funksjoner. Følgelig hadde det vært fint om flere prøver kunne blitt undersøkt på tilsvarende måte (evt. på andre måter), gjerne også innenfor andre emner enn funksjoner.

Noen valg som begrenser utforskningen av hva funksjonskompetanse kan inneholde er også gjort i denne studien. Kategoriseringen og utvelgelsen av oppgaver ble gjort med hensyn på alle de tre emneprøvene. Én oppgave som ble utelatt fordi den ikke passet inn i kompetanse-kategoriene som utgjorde rammeverket kan i denne sammenhengen nevnes. Dette var en oppgave hvor elevene skulle skille mellom et uttrykk, en likning og en funksjon, og kan sies å være en type kategoriseringsoppgave. I tillegg ble det på grunn av hensyn til mengde og prøvevarighet gjort en avgrensning hvor oppgaver om kvadratiske funksjoner ble besluttet holdt utenfor oppgavesettet. Oppgaver som ble utelatt kan også passe inn og på den måten utvide kompetansebegrepet knyttet til funksjoner. Det kan gjerne gjennomføres studier som undersøker hvordan andre deler av funksjonskompetanse, samt andre oppgavetyper, passer sammen med de som er undersøkt her.

Resultatene av studien gir et bilde av at svært mange elever mangler helt grunnleggende kompetanse for funksjoner. Kvalitative analyser av elevenes besvarelser peker mot at nødvendige forkunnskaper ikke er på plass. Mangelfull algebraisk representasjonskompetanse er blitt fremhevet som særlig fremtredende. Dette er også i tråd med hva større internasjonale undersøkelser gang på gang har avdekket som den største «mangelen» ved norske elevers matematiske kompetanse. Identifisering av sentrale årsaker til hvorfor denne kompetansen er så lav virker å være avgjørende for å kunne øke elevenes kompetanse for funksjoner.

Avslutningsvis bør det nevnes at resultatene i sin helhet tyder på at prøvene er tilfredsstillende instrumenter for måling av funksjonskompetanse. Ved å bruke et teoretisk rammeverk for kompetansekategorier og vanskegrader ut ifra oppgavetyper for å velge ut et mest mulig «balansert» oppgavesett slik det er brukt i denne studien vil en imidlertid på en lite ressurskrevende måte kunne lage enda bedre prøver i fremtiden.

7 Litteraturliste

Ainsworth, S., Bibby, P. & Wood, D. (1998). Analyzing the costs and benefits of multiple-representational learning environments. I M. W. Someren, P. Reimann, H.P.A. Boshuizen & T. Jong (Red.), *Learning with Multiple Representations*, s. 120-133. New York: Pergamon.

Andersen, S., Fossum, A., Rogstad, J. & Smestad, B. (2017). *På prøve: Evaluering av matematikkeksamen på 10. trinn våren 2017.* (Fafo rapport 36/2017). Hentet fra http://www.fafo.no/images/pub/2017/20644.pdf

Andrich, D. (1989). Distinctions Between Assumptions and Requirements in Measurement in the Social Sciences. I J.A. Keats, R. Taft, R.A. Heath & S.H. Lovibond (Red.), *Proceedings of the XXIVth International Congress of Psychology 4. Mathematical and Theoretical Systems* (s. 7-16). North-Holland: Elsevier Science Publisher BV.

Bond, T. & Fox, C. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, N.J: L. Erlbaum.

Bond, T. G. & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3. utg.). New York: Routledge.

Bossé, M. J., Adu-Gyamfi, K. & Cheetham, M. R. (2011). Assessing the difficulty of mathematical translations: Synthesizing the literature and novel findings. *International Electronic Journal of Mathematics Education, 6*(3), 113-133.

Brookhart, S. M. (1994). Teachers' grading: Practice and theory. *Applied Measurement in Education, 7*, 279–301. doi:10.1207/s15324818ame0704_2

Bueie, A. (2015). Summativ vurdering i formativ drakt; elevperspektiv på tilbakemelding fra heldagsprøver i norsk. *Acta Didactica Norge [elektronisk Ressurs], 9*(1), 1-21.

Cohen, L., Manion, L. & Morrison, K. (2011). *Research methods in education* (7.utg.). London: Routledge.

De Bock, D., Van Dooren, W. & Verschaffel, L. (2013). Students' understanding of proportional, inverse proportional and affine functions: two studies on the role of external

representations. *International Journal of Science and Mathematics Education, 13*(1), 47-69. doi: 10.1007/s10763-013-9475-z.

Duval, R. (2006). A Cognitive Analysis of Problems of Comprehension in a Learning of Mathematics. *An International Journal, 61*(1), 103-131. doi: 10.1007/s10649-006-0400-z

Even, R. (1990). Subject matter knowledge for teaching and the case of functions. *Educational Studies in Mathematics, 21*(6), 521-544.

Fisher-Hoch, H. & Hughes, S. (1996). *What makes mathematics exam questions difficult?* Paper presentert på British Educational Research Association, University of Lancaster, England.

Gagatsis, A., & Shiakalli, M. (2004). Ability to Translate from One Representation of the Concept of Function to Another and Mathematical Problem Solving. *Educational Psychology, 24*(5), 645-657. doi: 10.1080/0144341042000262953

Goldenberg, E. P. (1987). Believing is seeing: How preconceptions influence the perception of graphs. I J. C. Bergeron, N. Herscovics & C. Kieran (Red.), *Proceedings of the 11th International Conference for the Psychology of Mathematics Education* (s. 197–203). Montreal, Canada

Grønmo, L.S. & Hole, A. (2017). *Prioritering og progresjon i skolematematikken : En nøkkel til å lykkes i realfag. Analyser av TIMMS Advanced og andre internasjonale studier*. Oslo: Cappelen Damm Akademisk/NOASP (Nordic Open Access Scholarly Publishing).

Harlen, W. (2006). On the relationship between assessment for formative and summative purposes. I J. Gardner (Red.), *Assessment and learning* (95-110). London: Sage Publications Ltd.

Hartter, B. (2009). A Function or Not a Function? That Is the Question. *The Mathematics Teacher, 103*(3), 200-205.

Hattikudur, Shanta, Prather, Richard W., Asquith, Pamela, Alibali, Martha W., Knuth, Eric J., & Nathan, Mitchell. (2012). Constructing Graphical Representations: Middle Schoolers' Intuitions and Developing Knowledge about Slope and Y-Intercept. *School Science and Mathematics, 112*(4), 230-240.

Hopfenbeck, T.N., Ibsen, E., Turmo, A. & Lie, S. (2003). *Nasjonale prøver på ny prøve: Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005* (ILS rapport 1/2005). Hentet fra https://www.duo.uio.no/bitstream/handle/10852/32300/AD0501.pdf?sequence=1&isAllowed=y

Kaput, J. J. (1989). Linking representations in the symbol systems of algebra. I S. Wagner & C. Kieran (Red.), *Research issues in the learning and teaching of algebra* (s. 167-194). Hillsdale, NJ: Erlbaum

Leinhardt, G., Zaslavsky, O. & Stein, M. (1990). Functions, Graphs, and Graphing: Tasks, Learning, and Teaching. *Review of Educational Research, 60*(1), 1-64.

Linacre, J. M. (1998). Detecting multidimensionality: which residual data-type works best? *Journal of outcome measurement, 2*(3), 266-283.

Linacre, J. M. (2017a). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com

Linacre, J. M. (2017b). Winsteps® (Versjon 4.0.1) [Computer Software]. Beaverton, Oregon: Winsteps.com. Hentet fra http://www.winsteps.com/

Michell, J. (2002). Stevens's theory of scales of measurement and its place in modern psychology. *Australian Journal of Psychology, 54*(2), 99-104.

Niss, M., & Jensen, T. H. (2002). *Kompetencer og matematiklæring : Ideer og inspiration til udvikling af matematikundervisning i Danmark* (Vol. Nr 18/2002, Uddannelsesstyrelsens temahæfteserie). København: Undervisningsministeriet.

Nitsch, R., Fredebohm, A., Bruder, R., Kelava, A., Naccarella, D., Leuders, T. & Wirtz, M. (2015). Students' Competencies in Working with Functions in Secondary Mathematics Education - Empirical Examination of a Competence Structure Model. *International Journal of Science and Mathematics Education, 13*(3), 657-682.

O'Callaghan, B. R. (1998). Computer-Intensive Algebra and Students' Conceptual Knowledge of Functions. *Journal for Research in Mathematics Education, 29*(1), 21-40. doi:10.2307/749716

OECD (2013). *PISA 2012 Assessment and Analytical Framework : Mathematics, Reading, Science, Problem Solving and Financial Literacy*. Paris: OECD Publishing

Prøitz, T.S. & Borgen, J.S. (2010). *Rettferdig standpunktvurdering - det (u)muliges kunst? : Læreres setting av standpunktkarakter i fem fag i grunnopplæringen* (NIFU rapport 16/2010). Hentet fra https://brage.bibsys.no/xmlui/bitstream/handle/11250/279131/NIFUrapport2010-16.pdf?sequence=1&isAllowed=y

Sfard, A. (1991). On the Dual Nature of Mathematical Conceptions: Reflections on Processes and Objects as Different Sides of the Same Coin. *Educational Studies in Mathematics, 22*(1), 1-36.

Sierpinska, A. (1992). On understanding the notion of funtion. . I E. Dubinsky & Harel, G. (Red), *The concept of function: Aspects of epistemology and pedagogy* (s. 25-58). Washington D.C.: Mathematical Association Of America

Sjaastad, J. (2014). Enhancing measurement in science education research through Rasch analysis: Rationale and properties. *Nordina: Nordic Studies in Science Education, 10*(2), 212-230.

Stevens, S.S. (1946). On the theory of scales of measurement. *Science, 103*(2684), 677-680. Hentet fra http://www.jstor.org/stable/1671815

Säljö, R. (2010). Læring i praksis. Et sosiokulturelt perspektiv. Oslo: Cappelen akademisk forlag.

Utdanningsdirektoratet. (2013). *Læreplan i matematikk fellesfag. (MAT1-04).* Hentet fra http://www.udir.no/kl06/MAT1-04

Utdanningsdirektoratet. (2016). *Metodegrunnlag for nasjonale prøver*. Hentet fra https://www.udir.no/globalassets/filer/vurdering/nasjonaleprover/metodegrunnlag-for-nasjonale-prover.pdf

Usiskin, Z. (2015). What Does It Mean to Understand Some Mathematics? I S. J. Cho (Red.), *Selected Regular Lectures from the 12th International Congress on Mathematical Education* (s. 821-841): Springer International. doi: 10.1007/978-3-319-17187-6

Wilmot, D. B., Schoenfeld, A., Wilson, M., Champney, D. & Zahner, W. (2011). Validating a Learning Progression in Mathematical Functions for College Readiness. *Mathematical Thinking and Learning: An International Journal, 13*(4), 259-291.

Wolfe, E. W. & Smith, J. E. (2007a). Instrument development tools and activities for measure validation using Rasch models: part I-instrument development tools. *Journal of applied measurement, 8*(1), 97-123.

Wolfe, E. W. & Smith, J. E. (2007b). Instrument development tools and activities for measure validation using Rasch models: part II-validation activities. *Journal of applied measurement, 8*(2), 204-234.

Wright, B. D. & Stone, M. H. (1979). The measurement model. Best Test Design. Rasch Measurement. Chicago, IL: Mesa Press.

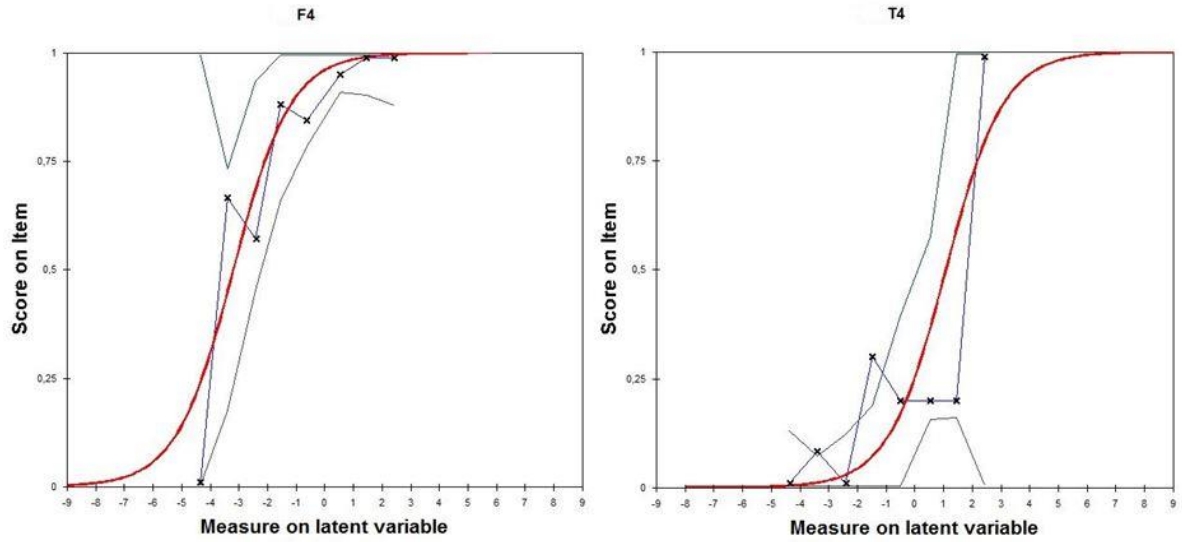
Wu, M., & Adams, R. (2007). Applying the Rasch model to psycho-social measurement: A practical approach. Melbourne: Educational Measurement Solutions.

Zachariades, T., Christou, C. & Papageorgiou, E. (2002). *The difficulties and reasoning of undergraduate mathematics students in the identification of functions.* Paper presentert på *Proceedings in the 10th ICME Conference, Crete, Greece*.

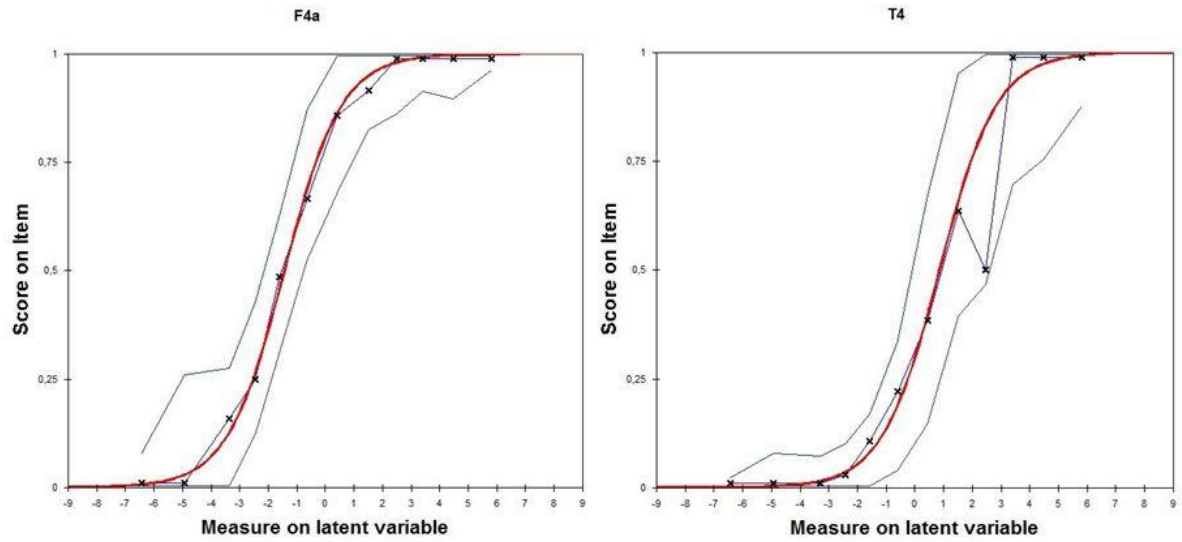
8 Appendiks

8.1 Appendiks A: ICC-er før og etter omgjøring fra FV til K

Sammenligning av ICC for oppgave F4 og T4 før og etter omgjøring fra flervalgsoppgave til konstruksjonsoppgave. Oppgave F4 fikk navnet F4a etter omgjøringen.

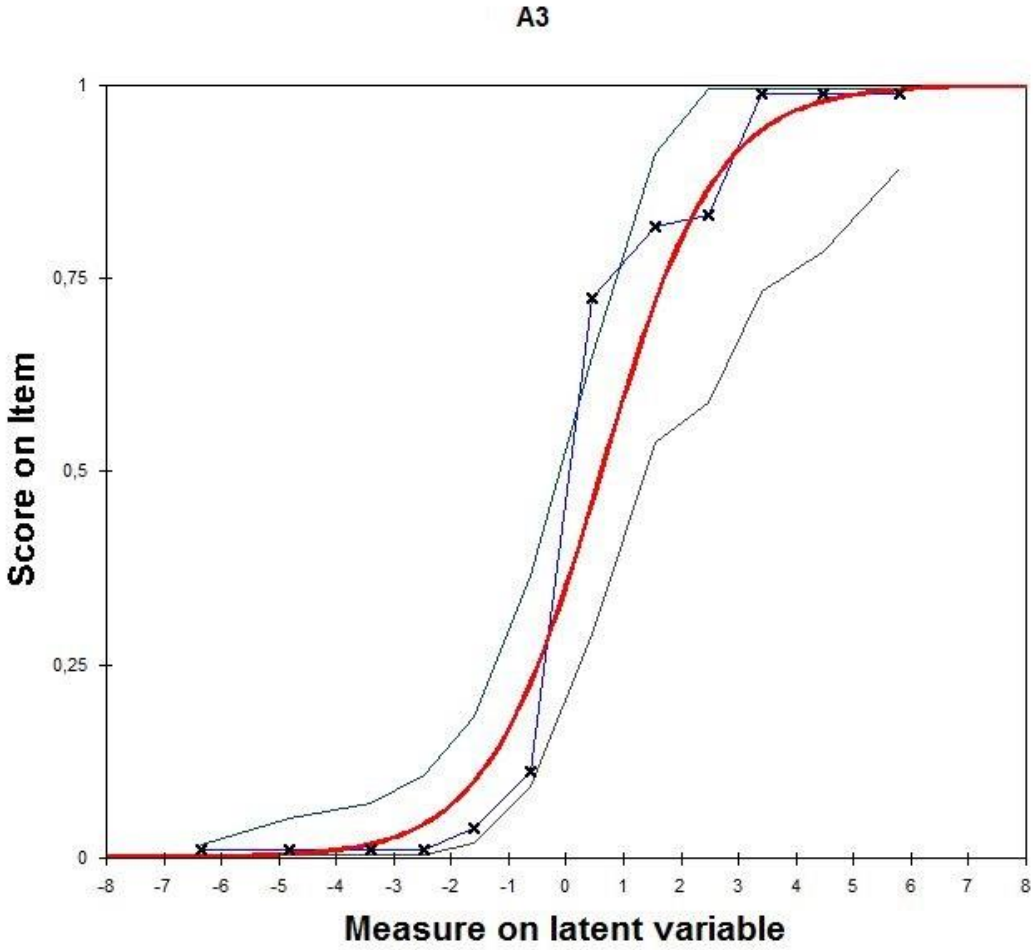


ICC til oppgave F4 og T4 når oppgavene var flervalgsoppgaver.



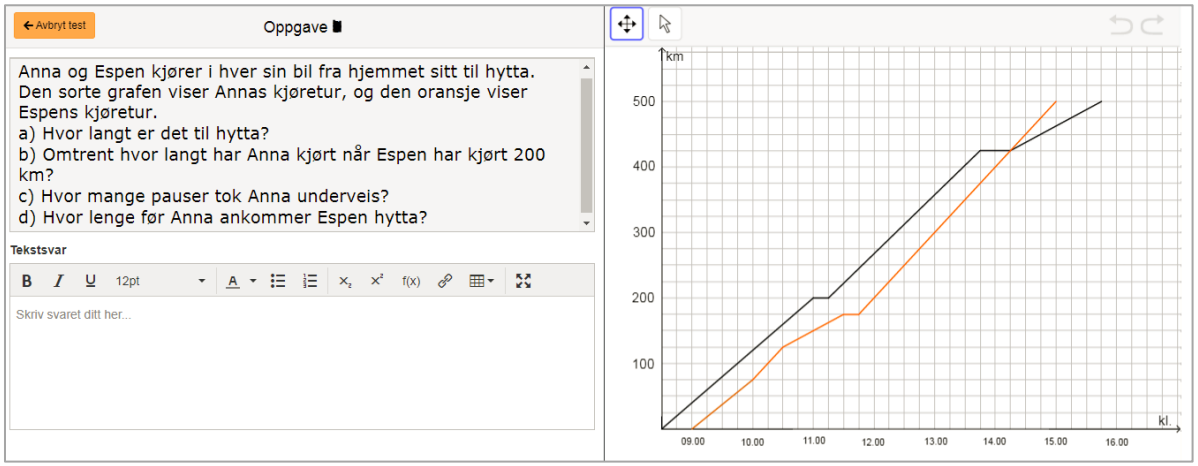
ICC til oppgave F4 og T4 når oppgavene var konstruksjonsoppgaver.

8.2 Appendiks B: ICC til oppgave A3

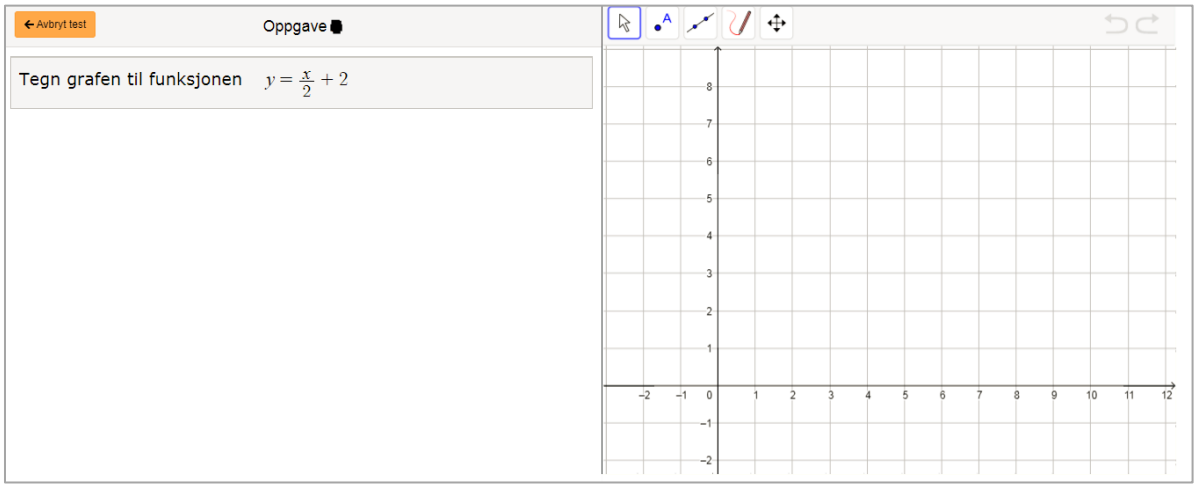


8.3 Appendiks C: Oppgavene

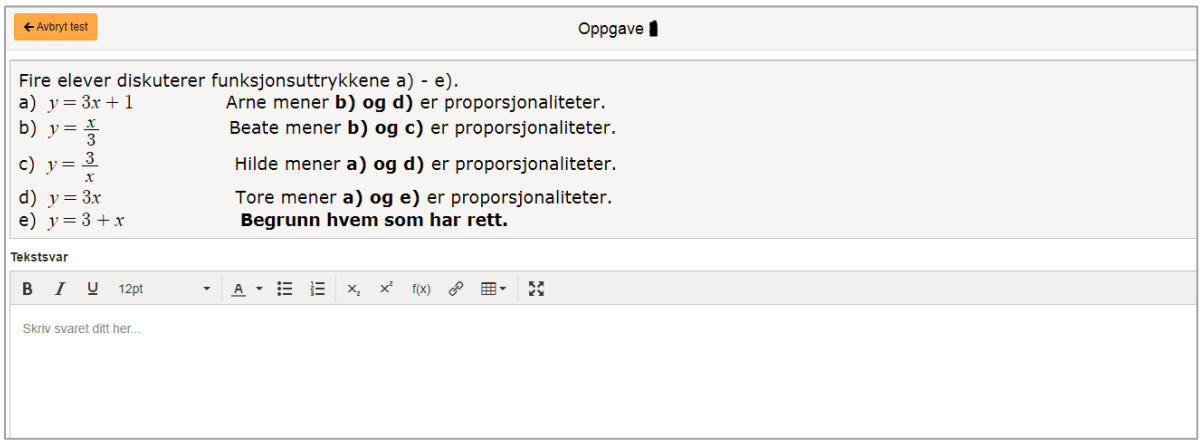
T1a – T1d



A2



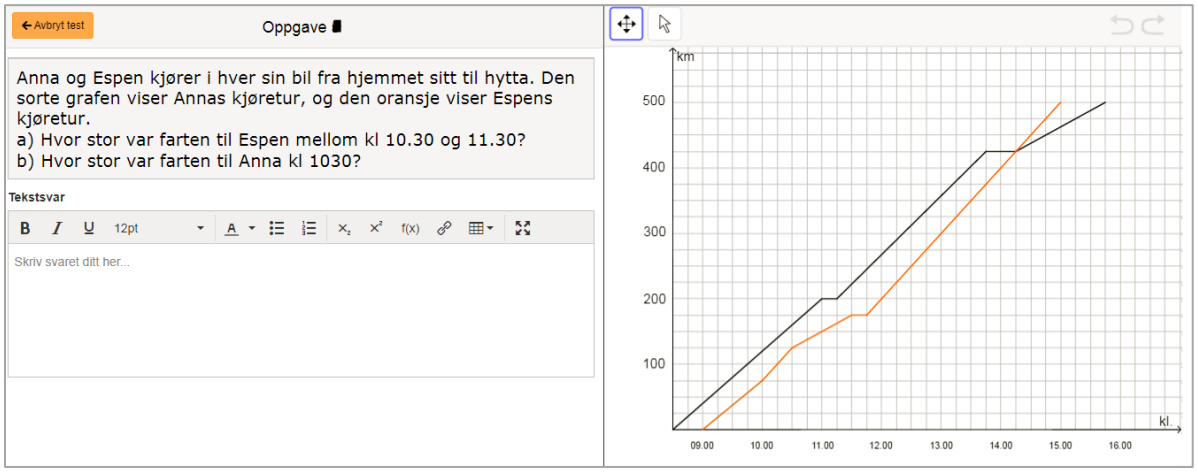
F1


Avbryt test
Oppgave
Fire elever diskuterer funksjonsuttrykkene a) - e).
a) $y = 3x + 1$
b) $y = \frac{x}{3}$
c) $y = \frac{3}{x}$
d) $y = 3x$
e) $y = 3 + x$
Arne mener **b) og d)** er proporsjonaliteter.
Beate mener **b) og c)** er proporsjonaliteter.
Hilde mener **a) og d)** er proporsjonaliteter.
Tore mener **a) og e)** er proporsjonaliteter.
Begrunn hvem som har rett.
Tekstsvar
12pt
Skriv svaret ditt her...

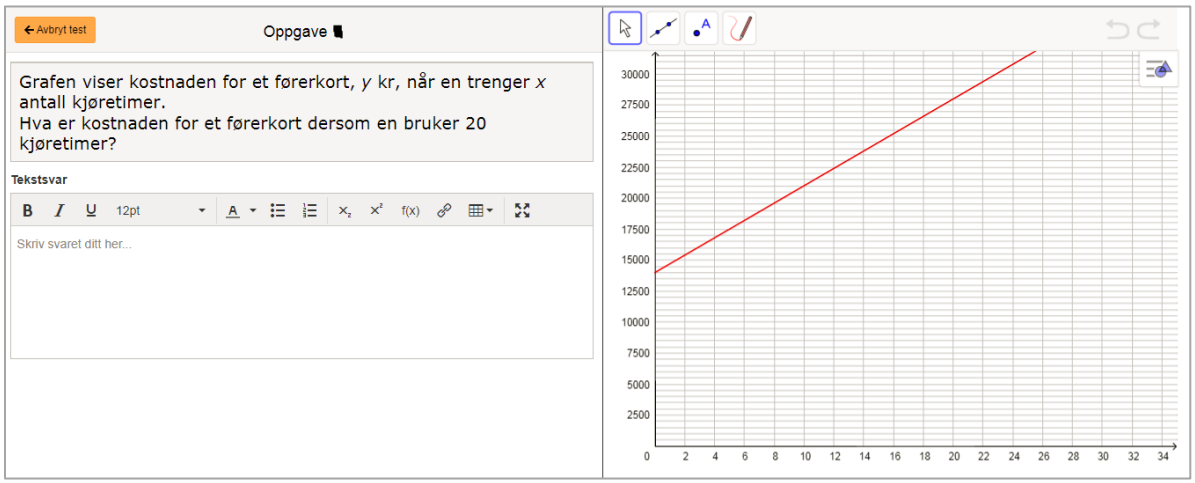
A1


Avbryt test
Oppgave
Ei rett linje går gjennom punktene (2,1) og (0,3).
Lag funksjonsuttrykket for linja.
Tekstsvar
12pt
Skriv svaret ditt her...

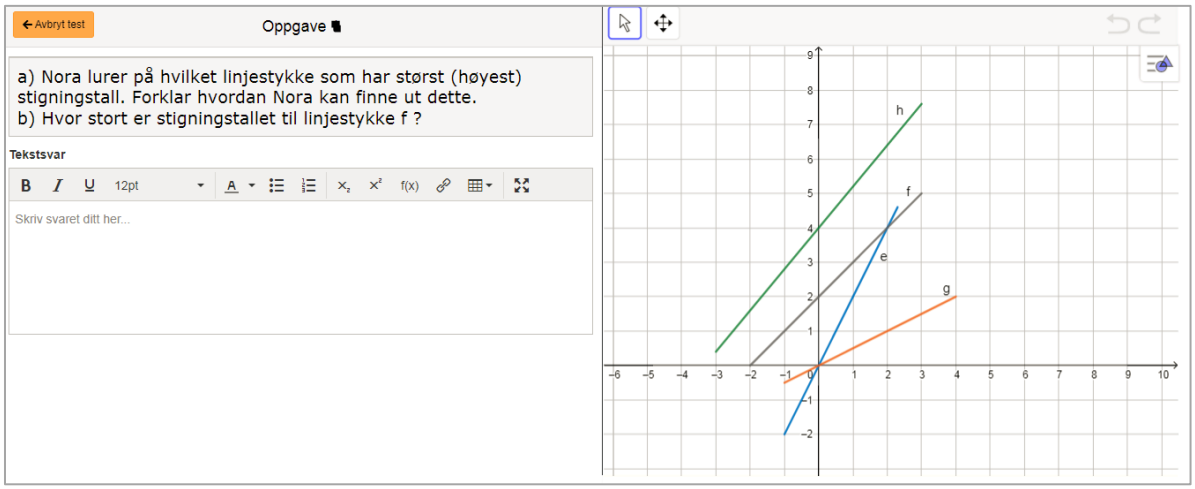
T2a + T2b


Avbryt test
Oppgave
Anna og Espen kjører i hver sin bil fra hjemmet sitt til hytta. Den sorte grafen viser Annas kjøretur, og den oransje viser Espens kjøretur.
a) Hvor stor var farten til Espen mellom kl 10.30 og 11.30?
b) Hvor stor var farten til Anna kl 1030?
Tekstsvar
12pt
Skriv svaret ditt her...
km
500
400
300
200
100
kl
09.00
10.00
11.00
12.00
13.00
14.00
15.00
16.00

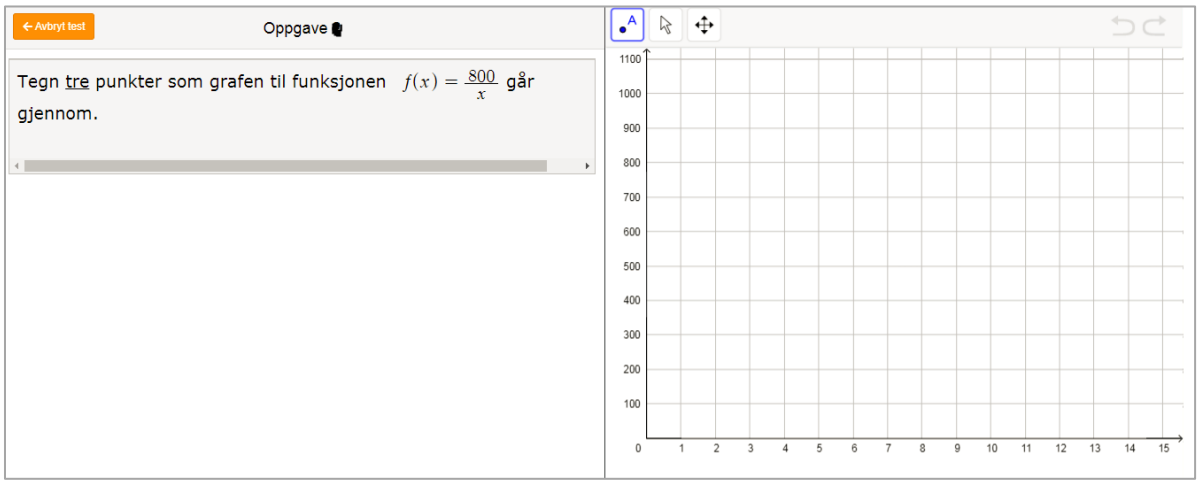
T3


Avbryt test
Oppgave
Grafen viser kostnaden for et førerkort, y kr, når en trenger x antall kjøretimer.
Hva er kostnaden for et førerkort dersom en bruker 20 kjøretimer?
Tekstsvar
12pt
Skriv svaret ditt her...
30000
27500
25000
22500
20000
17500
15000
12500
10000
7500
5000
2500
0
2
4
6
8
10
12
14
16
18
20
22
24
26
28
30
32
34

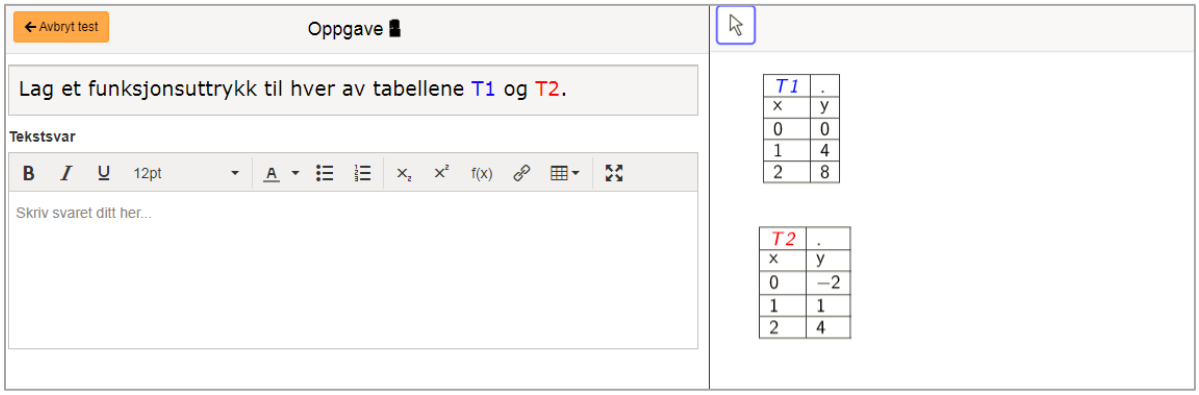
F4a + F4b


Avbryt test
Oppgave
a) Nora lurer på hvilket linjestykke som har størst (høyest) stigningstall. Forklar hvordan Nora kan finne ut dette.
b) Hvor stort er stigningstallet til linjestykke f ?
Tekstsvar
12pt
Skriv svaret ditt her...
h
f
e
g

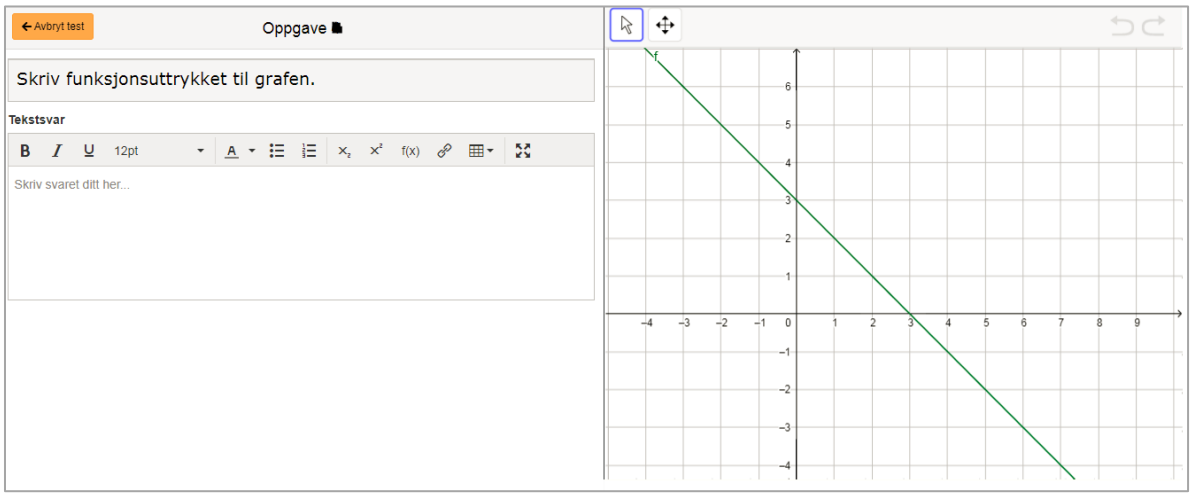
A5


Avbryt test
Oppgave
Tegn tre punkter som grafen til funksjonen $f(x) = \frac{800}{x}$ går gjennom.
1100
1000
900
800
700
600
500
400
300
200
100
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15

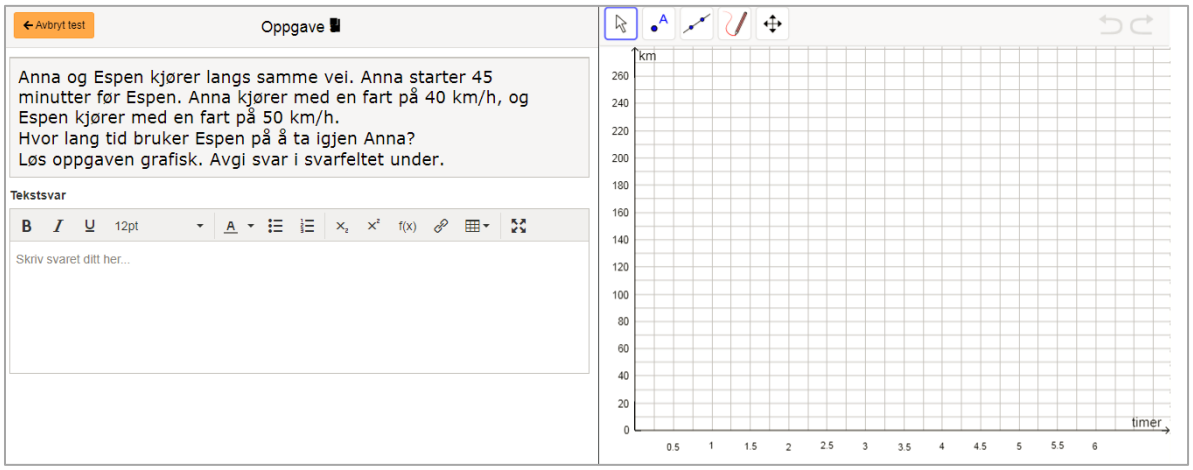
A6 + A7


Avbryt test
Oppgave
Lag et funksjonsuttrykk til hver av tabellene T1 og T2.
Tekstsvar
12pt
f(x)
Skriv svaret ditt her...
T1
x
y
0
0
1
4
2
8
T2
x
y
0
−2
1
1
2
4

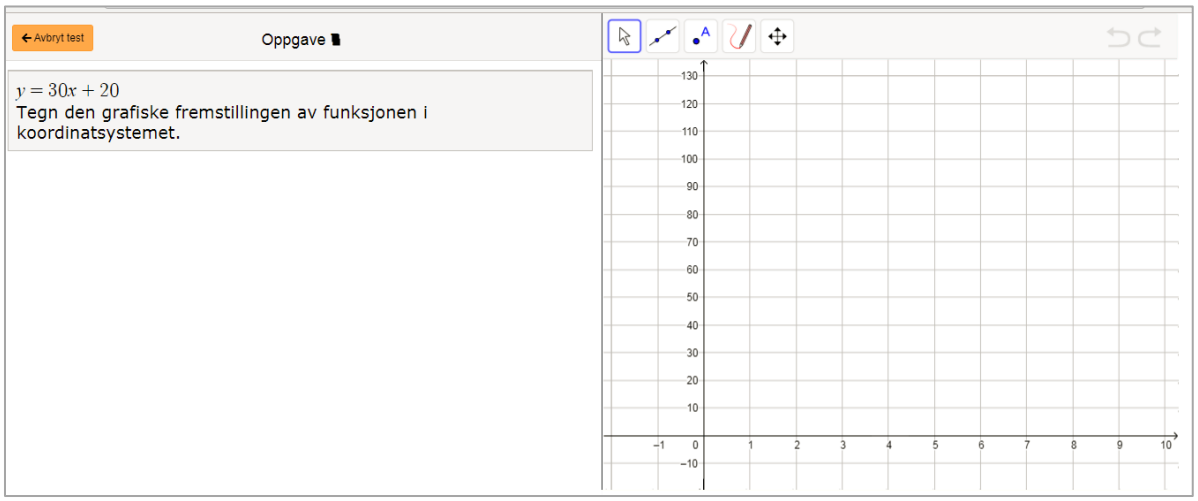
A3


Avbryt test
Oppgave
Skriv funksjonsuttrykket til grafen.
Tekstsvar
12pt
Skriv svaret ditt her...
f

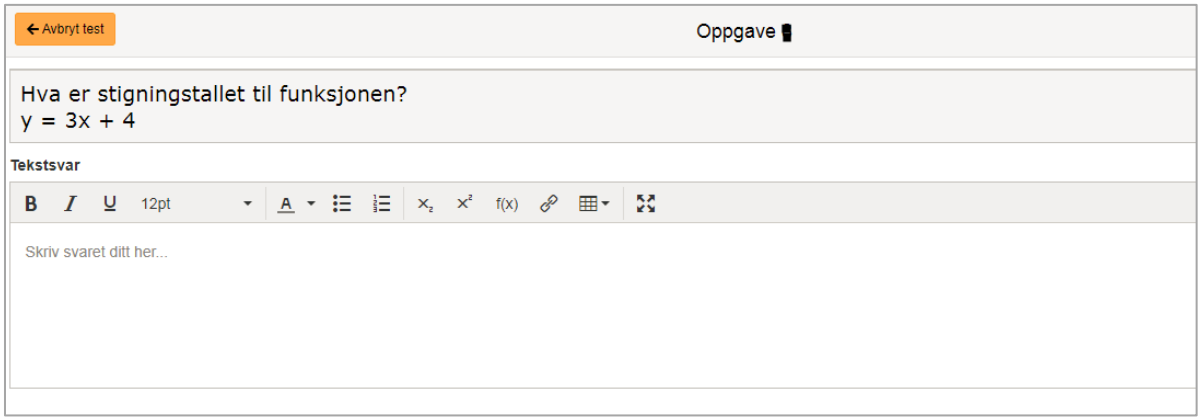
M1


Avbryt test
Oppgave
Anna og Espen kjører langs samme vei. Anna starter 45 minutter før Espen. Anna kjører med en fart på 40 km/h, og Espen kjører med en fart på 50 km/h.
Hvor lang tid bruker Espen på å ta igjen Anna?
Løs oppgaven grafisk. Avgi svar i svarfeltet under.
Tekstsvar
12pt
Skriv svaret ditt her...
km
timer

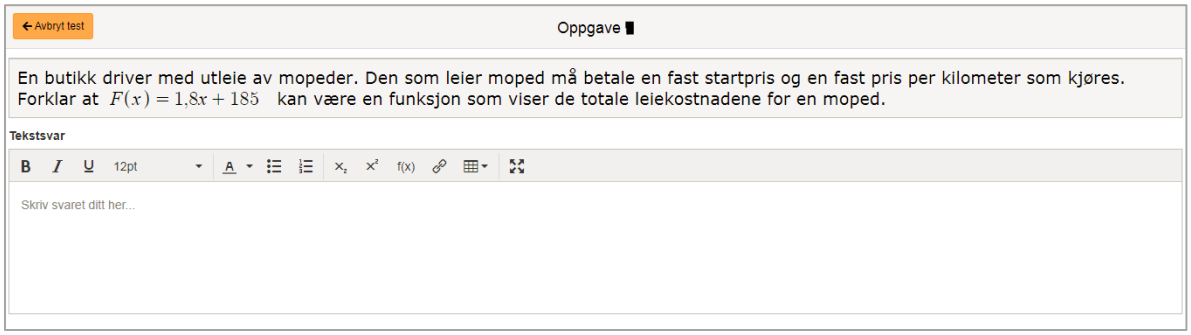
A4


Avbryt test
Oppgave
$y = 30x + 20$
Tegn den grafiske fremstillingen av funksjonen i koordinatsystemet.

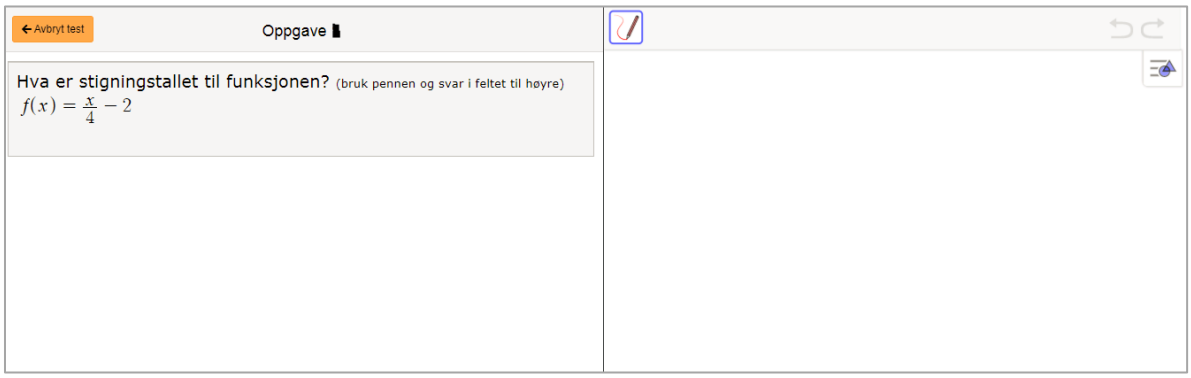
F6


Avbryt test
Oppgave
Hva er stigningstallet til funksjonen?
y = 3x + 4
Tekstsvar
12pt
f(x)
Skriv svaret ditt her...

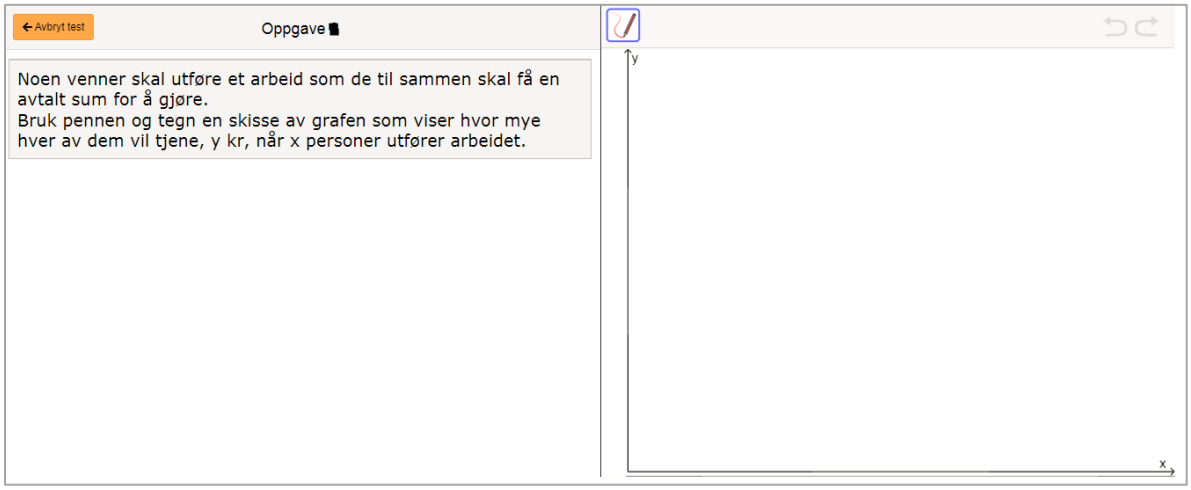
T5


Avbryt test
Oppgave
En butikk driver med utleie av mopeder. Den som leier moped må betale en fast startpris og en fast pris per kilometer som kjøres.
Forklar at $F(x) = 1{,}8x + 185$ kan være en funksjon som viser de totale leiekostnadene for en moped.
Tekstsvar
12pt
f(x)
Skriv svaret ditt her...

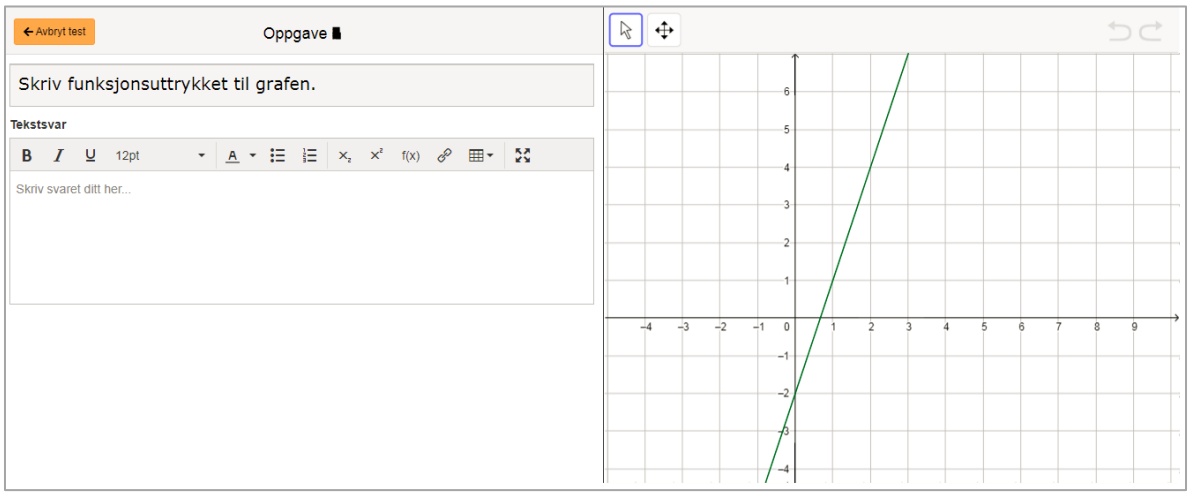
F2


Avbryt test
Oppgave
Hva er stigningstallet til funksjonen? (bruk pennen og svar i feltet til høyre)
$f(x) = \frac{x}{4} - 2$

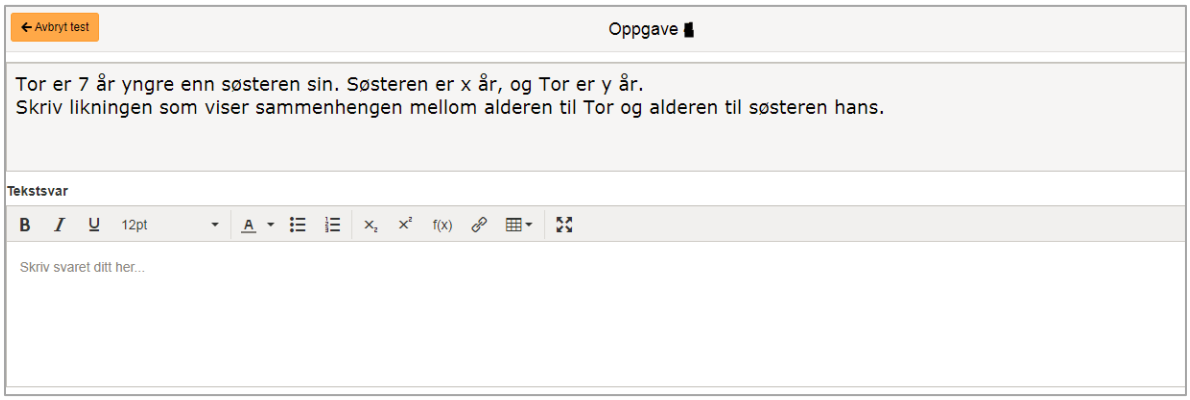
E1


Avbryt test
Oppgave
Noen venner skal utføre et arbeid som de til sammen skal få en avtalt sum for å gjøre.
Bruk pennen og tegn en skisse av grafen som viser hvor mye hver av dem vil tjene, y kr, når x personer utfører arbeidet.
y
x

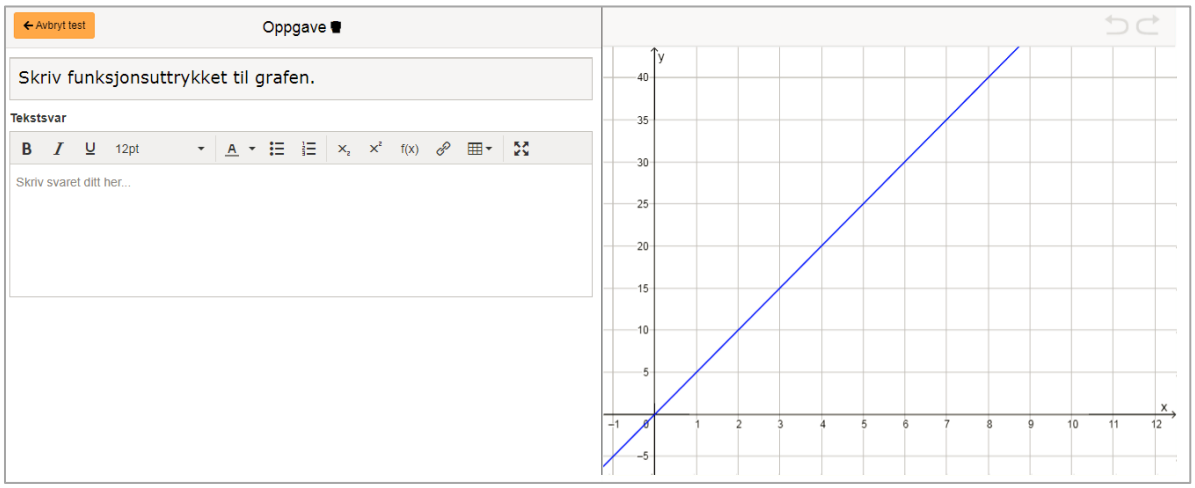
A8


Avbryt test
Oppgave
Skriv funksjonsuttrykket til grafen.
Tekstsvar
12pt
Skriv svaret ditt her...

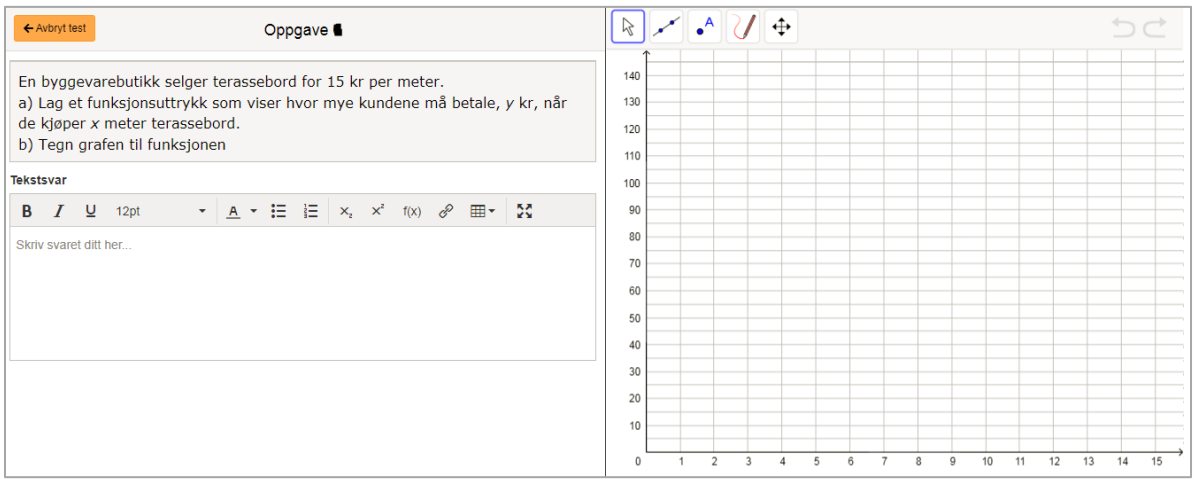
M2


Avbryt test
Oppgave
Tor er 7 år yngre enn søsteren sin. Søsteren er x år, og Tor er y år.
Skriv likningen som viser sammenhengen mellom alderen til Tor og alderen til søsteren hans.
Tekstsvar
12pt
Skriv svaret ditt her...

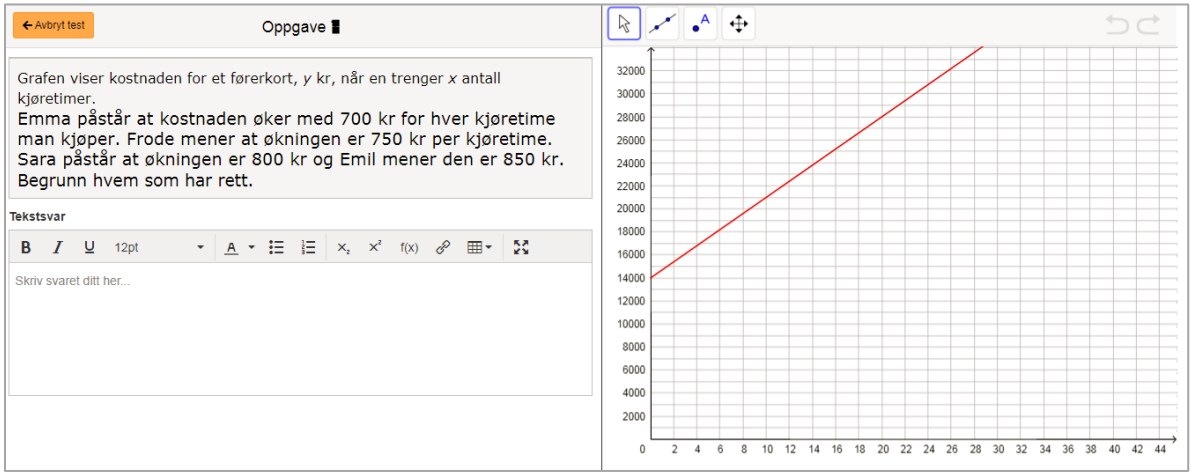
A9


Avbryt test
Oppgave
Skriv funksjonsuttrykket til grafen.
Tekstsvar
12pt
Skriv svaret ditt her...
y
x

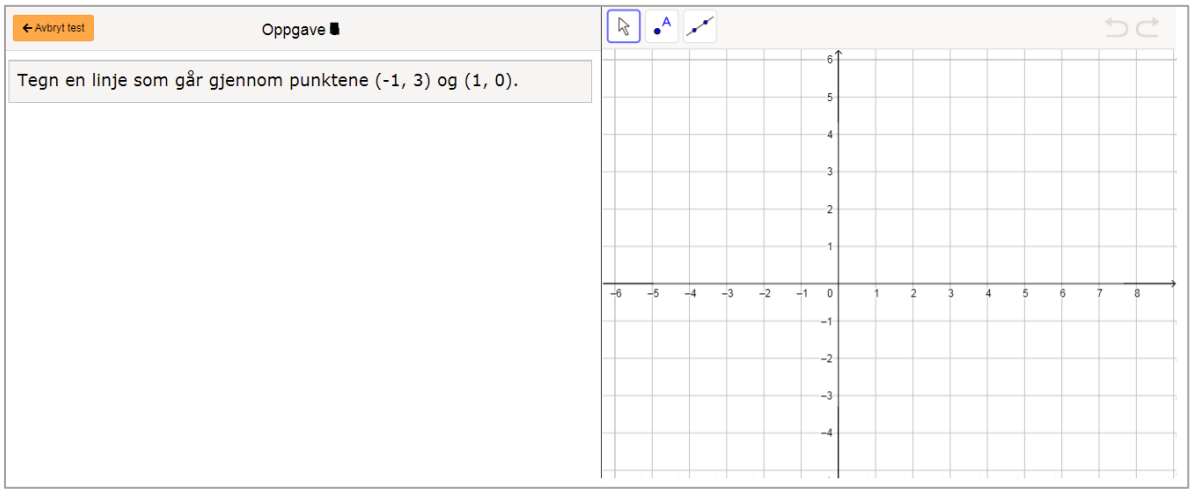
M5a – M5b


Avbryt test
Oppgave
En byggevarebutikk selger terassebord for 15 kr per meter.
a) Lag et funksjonsuttrykk som viser hvor mye kundene må betale, y kr, når de kjøper x meter terassebord.
b) Tegn grafen til funksjonen
Tekstsvar
12pt
Skriv svaret ditt her...

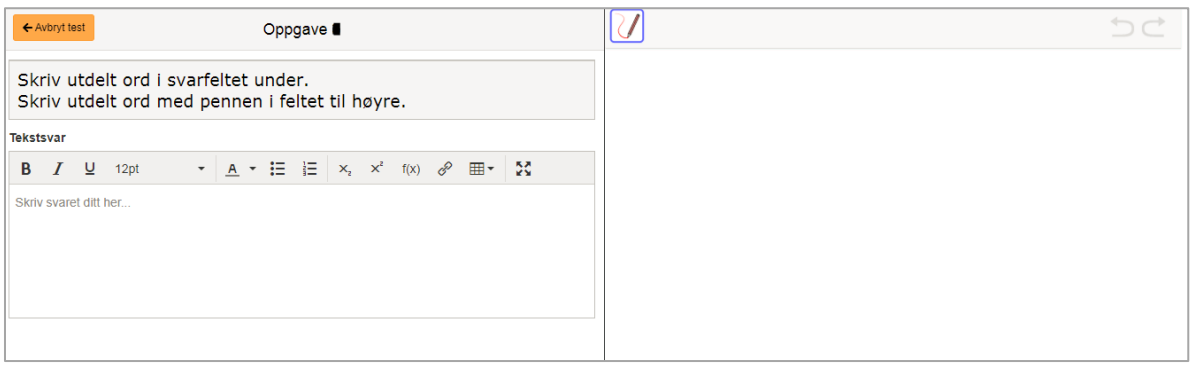
T4



G1



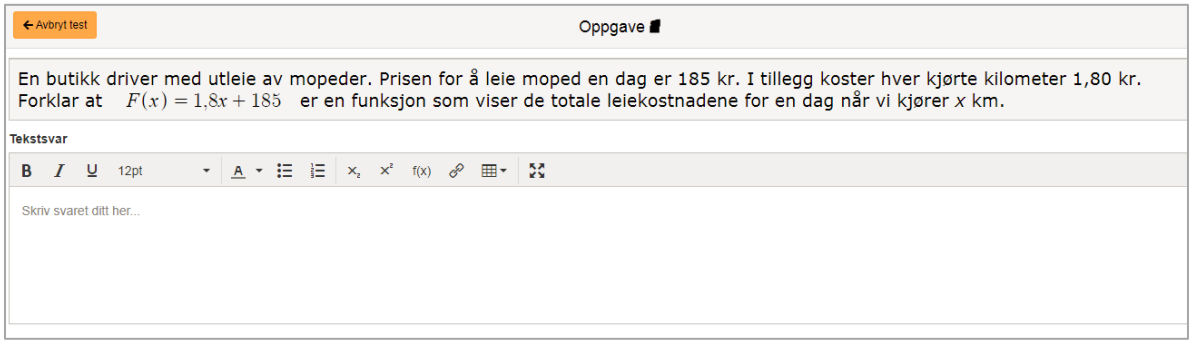
Kodeord



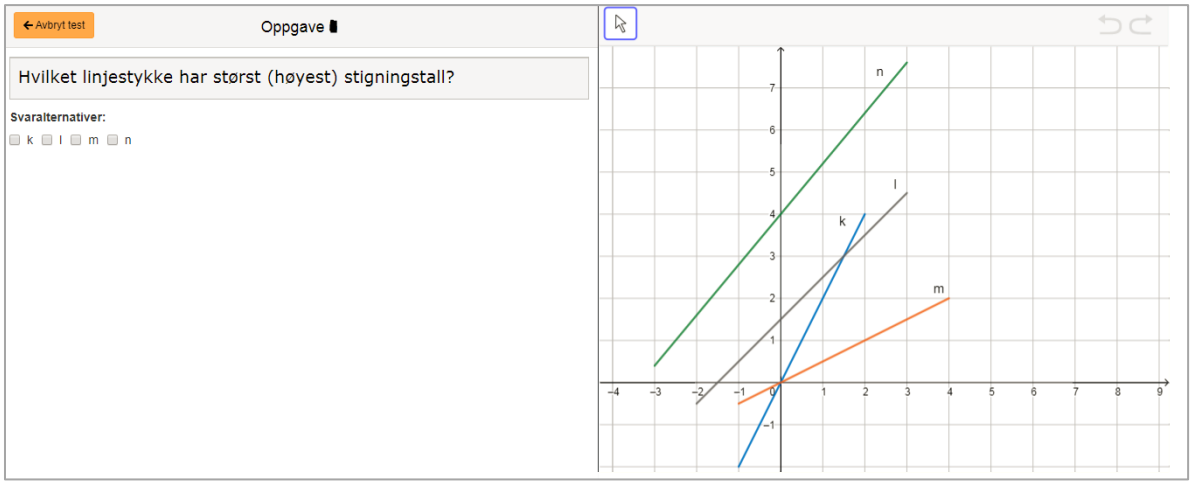
0F1


Avbryt test
Oppgave
Hvilke av disse er proporsjonaliteter?
a) $y = 0{,}3x$
b) $y = \frac{x}{3}$
c) $y = \frac{3}{x}$
d) $y = 3x$
e) $y = 3 - x$
Svaralternativer:
a b c d e Vet ikke

0T5


Avbryt test
Oppgave
En butikk driver med utleie av mopeder. Prisen for å leie moped en dag er 185 kr. I tillegg koster hver kjørte kilometer 1,80 kr.
Forklar at $F(x) = 1{,}8x + 185$ er en funksjon som viser de totale leiekostnadene for en dag når vi kjører x km.
Tekstsvar
12pt
f(x)
Skriv svaret ditt her...

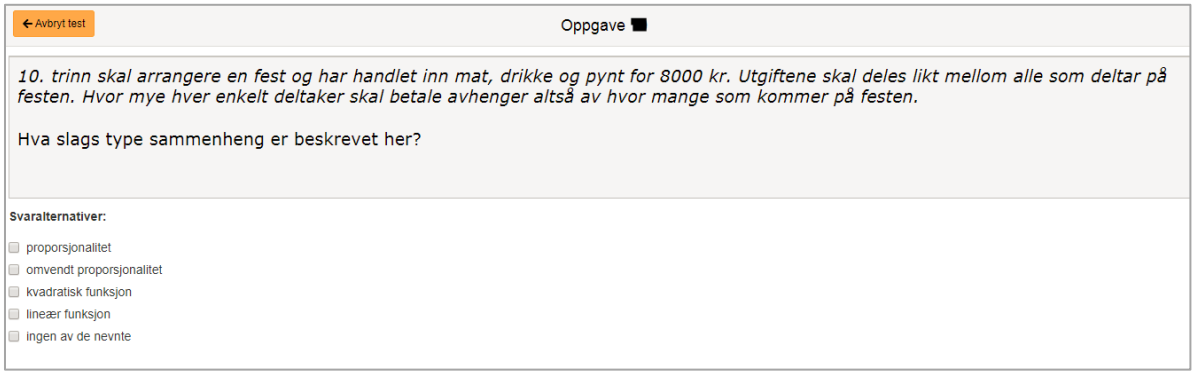
0F4a


Avbryt test
Oppgave
Hvilket linjestykke har størst (høyest) stigningstall?
Svaralternativer:
k l m n
n
l
k
m

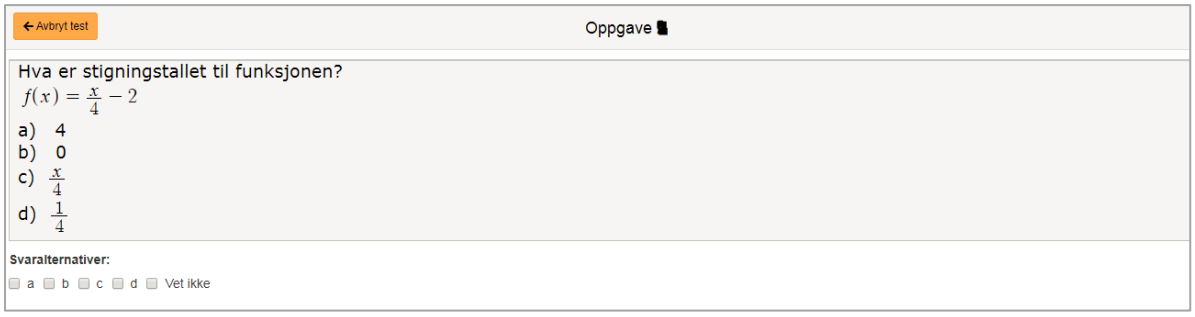
0F5


Avbryt test
Oppgave
I hvilken av funksjonene er x og y omvendt proporsjonale størrelser?
A) $y = 5x + 10$
B) $y = \frac{x}{5} + 10$
C) $y = -5 - 10$
D) $y = \frac{5}{x}$
Svaralternativer:
A B C D ingen

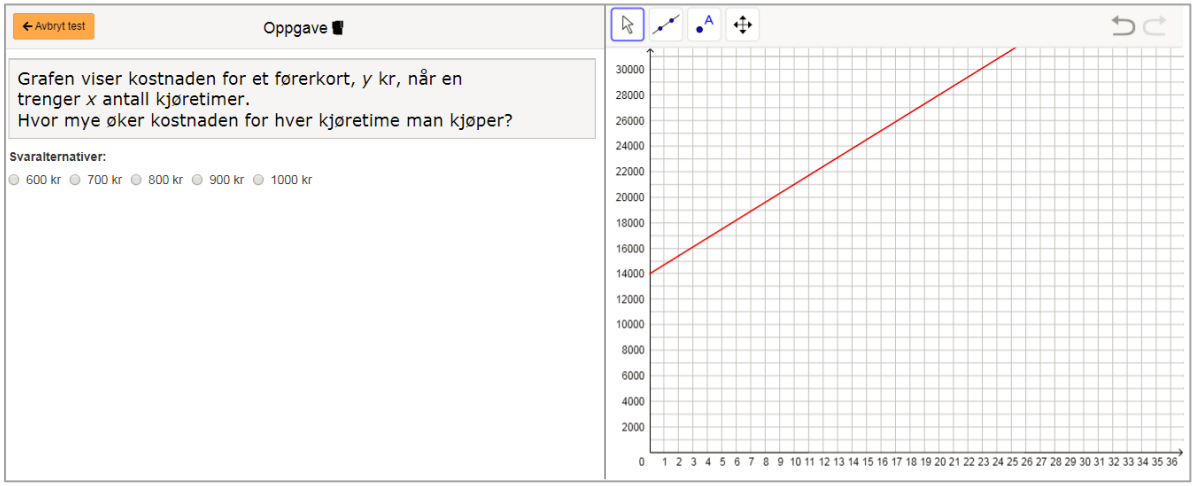
0F3


Avbryt test
Oppgave
10. trinn skal arrangere en fest og har handlet inn mat, drikke og pynt for 8000 kr. Utgiftene skal deles likt mellom alle som deltar på festen. Hvor mye hver enkelt deltaker skal betale avhenger altså av hvor mange som kommer på festen.
Hva slags type sammenheng er beskrevet her?
Svaralternativer:
proporsjonalitet
omvendt proporsjonalitet
kvadratisk funksjon
lineær funksjon
ingen av de nevnte

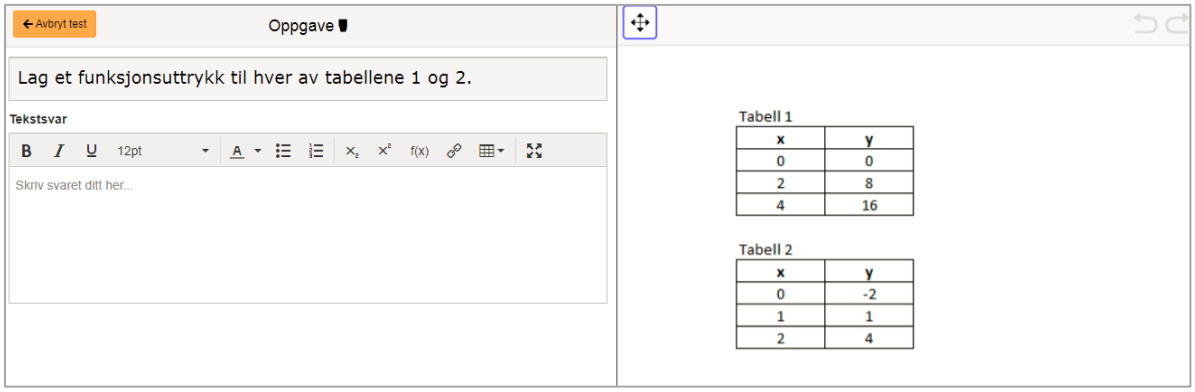
0F2


Avbryt test
Oppgave
Hva er stigningstallet til funksjonen?
$f(x) = \frac{x}{4} - 2$
a) 4
b) 0
c) $\frac{x}{4}$
d) $\frac{1}{4}$
Svaralternativer:
a b c d Vet ikke

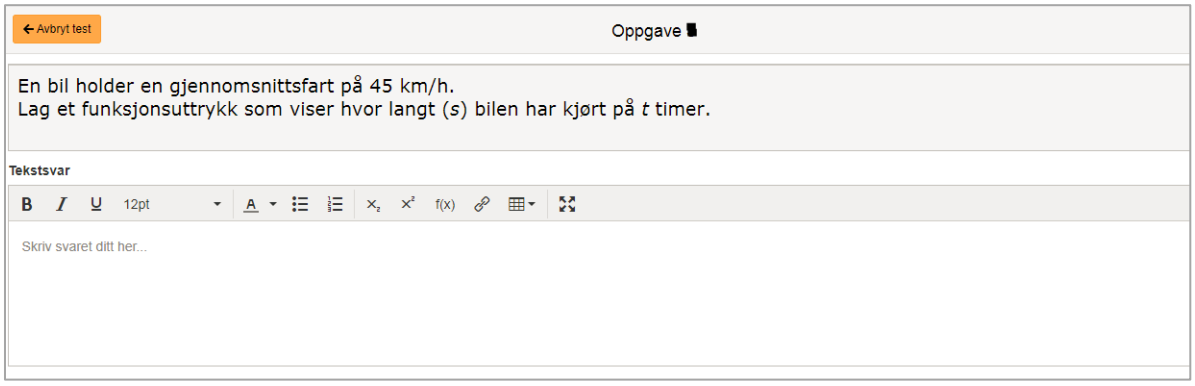
0T4



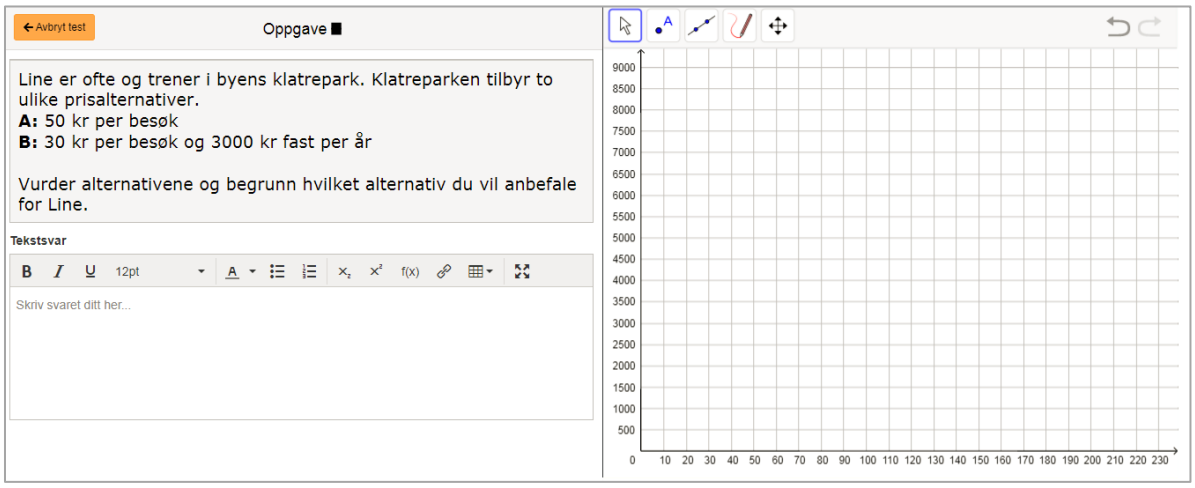
0A6 (tabell 1)



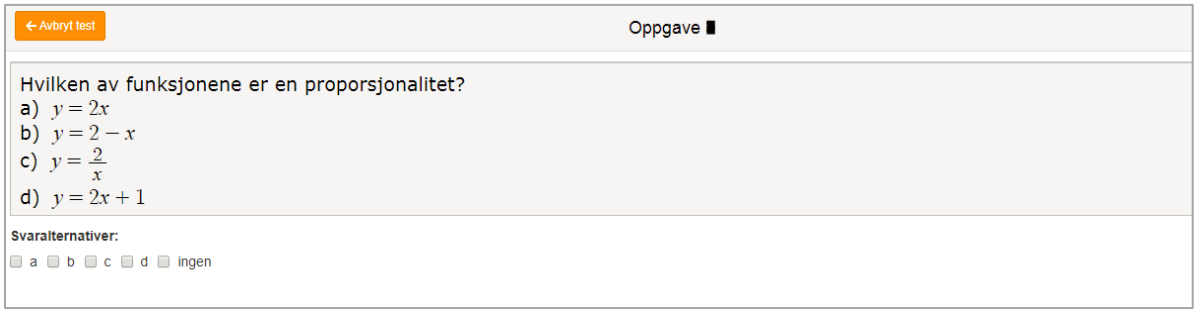
0M4



0M3



0F1m



0F1n