UiT Norges arktiske universitet

Fakultet for humaniora, samfunnsvitenskap og lærerutdanning

Validitet og reliabilitet i to åpne oppgaver fra de nasjonale prøver i lesing 2023

Cecilie Hansen og Christina Mlinnikova Kjølås

Mastergradsoppgave i norskdidaktikk, LER-3901, mai 2024.

Innholdsfortegnelse

Tabelliste

Figurliste

Vedlegg

Forord

Denne masteroppgaven markerer slutten på vår fagdidaktiske videreutdanning i norsk. Vi er veldig glade og stolte over at prosjektet kom i mål. Det å ta videreutdanning i kombinasjon med familieliv og full lærerjobb har krevd harde prioriteringer de siste årene. Vi opplever å sitte igjen med mye god kunnskap som vi tar med oss tilbake til elevene våre og profesjonsfelleskapet. Det var helt klart verdt strevet!

Å samarbeide om et masterprosjekt kan være krevende, men for oss har det vært positivt og en styrke. Vi har utfylt hverandre faglig og i kraft av våre personligheter på en god måte. Det har vært godt å være to, fordi vi har utfordret, oppmuntret og motivert hverandre.

Vi vil rette en stor takk til vår veileder Morten Bartnæs. Tusen takk for at du har vært så tilgjengelig, fleksibel og veiledet oss fram til målet. Vi har satt stor pris på dine faglige innspill, gode råd og ditt kritiske blikk.

Takk til Utdanningsdirektoratet ved Ga Young Yoon og Hilde Hultin som har bistått med empiri til denne masteroppgaven. Dere har vært tilgjengelig og hjelpsomme underveis i prosessen, det har vi satt stor pris på.

Vi vil også rette en stor takk til våre arbeidsplasser og kollegaer som alltid har heiet oss framover. Det har hatt mye å si at dere har vært så positive!

Til sist vil vi rette en uendelig takk til våre familier. Det er mange som skal ha takksigelser fra oss, men mest av alt; Jon og Jevgeni – takk for at dere har tatt støyten på hjemmebane det siste året, det hadde ikke gått uten dere! Tusen takk!

Til Edvard og Aksel, Aleksandra og Nikolai – takk for at dere har vært så tålmodige og vist forståelse for at vi har vært så mye borte.

Tromsø, 14. mai, 2024.

Cecilie Hansen og Christina Mlinnikova Kjølås

Sammendrag

Tema for denne masteroppgaven er validitet og reliabilitet i åpne oppgaver på nasjonale prøver i lesing. Nordiske studier har vist at vurdering og skåring av elevsvar på åpne oppgaver kan være utfordrende. Problemstillingen for denne masteroppgaven er: I hvilken grad gir åpne oppgaver på nasjonale prøver i lesing valide målinger og reliable resultater?

For å belyse dette tar studien utgangspunkt i to forskningsspørsmål: «I hvilken grad måler åpne oppgaver elevenes lesekompetanse i å tolke og sammenholde informasjon, og å finne informasjon i tekster?» og «Hvor presise, dekkende og anvendelige er kriteriene for skåring som gis i vurderingsveilederen, og i hvilken grad svarer skårernes vurderingspraksis til disse kriteriene?»

Forskningsspørsmålene undersøkes gjennom en kvalitativ tekstgjennomgang knyttet til prøvetekstene *Stargate* og *Snikfotografen.* Dette kombineres med en kvantitativ innholdsanalyse. Statistiske beregninger foretas for å undersøke graden av samsvar i skåring av åpne oppgaver. Utvalget består av 1200 elevsvar, fra 8. og 9. trinn fra nasjonale prøver høsten 2023.

Kvalitative funn i gjennomgang av de to tekstene viser at elevene må lese og forstå tekst med både implisitt og eksplisitt meningsinnhold. Tolkningsrommet som oppstår mellom teksten, oppgavene, vurderingsveiledere og oss som vurderer, står sentralt i dette prosjektet.

Funn viser at sensorreliabiliteten utfordres i noen grad i vurdering og skåring av tolkningsoppgaven i *Stargate*. I skåring av åpen oppgave *Snikfotografen* er sensorreliabiliteten derimot svært god.

Det finnes et potensial i å prioritere forberedelser i profesjonsfellesskapet før vurdering og skåring av elevsvar på de åpne oppgavene.

1 Innledning[1]

1.1 Bakgrunn for oppgaven

Nasjonale prøver i lesing er en landsomfattende prøve, der det primære formålet er å kartlegge norske elevers lesekompetanse på 5., 8. og 9. trinn. De nasjonale prøvene skal gi skolen kunnskap om elevens ferdigheter i lesing og prøvene gjennomføres årlig. Resultatene publiseres på Utdanningsdirektoratet sine hjemmesider. Resultatene gir informasjon om enkelt elever, grupper, trinn og skoler (Utdanningsdirektoratet, 2022, s. 2). I Norge foregår det for tiden en diskusjon om, og eventuelt på hvilken måte de nasjonale prøvene skal endres. Høsten 2023 kom Prøitz- utvalget med sin anbefaling om endringer i prøver og undersøkelser i skolen. De foreslår at de nasjonale prøvene skal erstattes med nye læringsstøttende prøver, som kan være til bruk i det faglige pedagogiske arbeidet i skolene (NOU 2023:27, s.15). Ifølge Astrid Roe m. fl. (2018) er måling av kunnskap og ferdigheter ikke noe nytt i norsk skole, og elever har blitt testet til alle tider. Nasjonale prøver skal gi mer nyansert informasjon om leseferdighetene til alle elevene på alle nivåer. Prøven har stor spredning når det gjelder tekster og oppgavenes vanskelighetsgrad (Roe, Ryen, & Weyergang, 2018, s. 20).

Den nasjonale prøven i lesing består av to oppgavetyper, flervalgsoppgaver og åpne oppgaver. I 2001 ble åpne oppgaver benyttet som oppgavetype i PIRLS. Åpne oppgaver gir elevene mulighet til å uttrykke seg selvstendig og uttrykke egne tanker (Solheim & Skaftun, 2009, s. 149). En leseprøve uten åpne oppgaver vil kanskje måle en snevrere del av lesekompetansen, noe som igjen kan svekke validiteten til prøven (Roe, Ryen, & Weyergang, 2018, s. 192). Som norsklærere og gjennom denne videreutdanningen har vi utviklet en stor interesse for de åpne oppgavene ved nasjonale prøvene i lesing. De åpne oppgavene, der elevene skal formulere svaret med egne ord, vil kunne føre til uenighet blant lærerne når de skal vurderes og skåres (Roe, Ryen, & Weyergang, 2018, s. 29). I vår jobb har vi lagt til rette for gjennomføring av nasjonale prøver i lesing for 8. og 9. trinn. I tillegg har vi erfaring med vurdering og skåring. I skåring av åpne oppgaver har vi fått innblikk i hvordan elevene tolker, reflekterer og trekker slutninger, i møte med tekst. Ved flere anledninger har vi opplevd å

[1] I dette avsnittet og seinere på s. 1- 4 gjenbruker vi noen passasjer fra et arbeidskrav levert innenfor emnet LER-3501 Metoden, høsten 2023. Vi henviser også til eget FoU-prosjekt, med innlevering 6. desember 2021 fra emnet LER-3680 Prosjektoppgaven.

møte elevsvar som på ulike vis faller utenfor vurderingsveilederens kriterier. Det kan være upresise svar, ubesvarte spørsmål og elevuttrykk som for eksempel *«slutt med alle disse spørsmålene», «idk»* (Engelsk: I don't know) eller «*hdlfsdfkdf»* (tilfeldige tastetrykk).

I mange tilfeller møter vi også elevbesvarelser som er vanskelig å skåre, som for eksempel *«For ay bare Faren seal få jobb»* eller *«bildene har forskjellig kvalitet på bildene».* Elevsvar som er «vanskelig å skåre» omtaler vi videre i oppgaven som *tvilstilfeller*. Det kan være krevende å vurdere om tvilstilfellene skal ha poeng i skåring eller ikke. Vurderingsveilederen til de åpne oppgavene påpeker riktignok at tvilen skal komme eleven til gode, men er det så enkelt? Å være i tvil er en subjektiv opplevelse. Med andre ord kan vi og våre kollegaer oppleve tvilen ulikt. Derav finnes det også en risiko for at vi skårer elevsvar som ikke helt treffer vurderingsveilederens beskrivelser, ulikt. I forbindelse med vår fagdidaktiske videreutdanning i norsk høsten 2021 undersøkte vi sensorreliabiliteten i skåringen av de åpne oppgavene. Sensorreliabilitet i skåring handler om graden av overenstemmelse mellom lærernes vurderinger av de samme elevsvarene. Våre funn fra denne undersøkelsen viste at sensorreliabiliteten varierte. Det vil si at noen elevsvar hadde høy grad av samsvar mellom skårere og andre noe lavere (Kjølås & Hansen, 2021, ss. 22-23).

Et av funnene i undersøkelsen viste at det i noen tilfeller oppstår konflikt mellom vurderingsveilederen og lærerens egen autonomi og integritet i vurderingen. En lærer skrev dette til oss som en kommentar i nettskjemaet:

> Hei. Jeg syns ikke alltid [vurderingsveilederen] gir nok rom for fordelaktig tolkning for elevene. Et eksempel er et elevsvar til oppgave 3 i dette spørreskjemaet. Der svarer eleven godt i forhold til hva man kan lese i teksten (om trusler mot hvalarten) men ikke 100 % korrekt i forhold til forståelsen av piktogram/symbol. Og da krysser jeg av på «galt svar», enda min personlige vurdering er at svaret viser veldig god lesekompetanse (Kjølås & Hansen, 2021, s. 16).

Slik vi tolket læreren, kan det virke som at hun følte seg forpliktet å skåre elevsvaret til 0 poeng ut fra vurderingsveilederens beskrivelse. Dette til tross for at lesekompetansen ble vurdert som god. Poenget som ble aktualisert i denne kommentaren, var noe vi kjente igjen.

Michael Tengberg, er en av forskerne som har skrevet en rekke artikler om validitet og reliabilitet i leseprøver. Flere artikler viser til funn om varierende sensorreliabilitet i skåring av åpne oppgaver, både i Norge og Sverige (Tengberg, Roe & Skar, 2018; Tengberg, 2017). Artiklene drøfter også hvordan reliabilitet i skåring kan påvirke validiteten i leseprøvene, og viser blant annet til at det har vært store diskusjoner om prøvenes validitet og reliabilitet i vurdering av åpne oppgaver i Sverige (Tengberg & Skar, 2017, s. 114).

Solheim og Skaftun (2009) har utforsket oppgavekonstruksjonen i leseprøver til PISA. Forfatterne diskuterer betydningen av skåringsguidene [vurderingsveilederen]. Det ses på hvilken sammenheng det er mellom prøveteksten og oppgavene på den ene siden og skåringsguidene på den andre siden. Videre gjennomføres det en diskusjon rundt problematiske aspekt i operasjonaliseringen mellom intensjonene som er uttrykt i det testteoretiske rammeverket. Artikkelen viser til tilfeller hvor formuleringene i skåringsguiden begrenser den semantiske åpenheten. Da vil det kunne resultere i en konflikt med intensjon for leseaspektet oppgaven skal måle (Solheim & Skaftun, 2009). Dette var spesielt interessant å se til i vårt prosjekt, som tar utgangspunkt i to åpne oppgaver fra nasjonale prøver i lesing 2023. Basert på våre erfaringer med skåring, ny innsikt i faglitteraturen og egne funn i FoU-oppgaven, ble validitet og reliabilitet i de åpne oppgavene på nasjonale prøver i lesing tema for vår masteroppgave.

1.2 Forskningsspørsmål

I vår masteroppgave har vi som mål å få større innsikt i de åpne oppgavenes validitet og reliabilitet. Samtidig søker vi en dypere forståelse av hva prøvetekstene og de åpne oppgavene krever av elevene sett opp mot prøvens rammeverk, og hvordan dette blir vurdert. I tillegg ønsker vi å bidra med empiri som kan være et innspill i diskusjonen omkring vurdering og skåring av de åpne oppgavene fra et lærerperspektiv. Problemstillingen for denne masteroppgaven er: I hvilken grad gir åpne oppgaver på nasjonale prøver i lesing valide målinger og reliable resultater?

For å belyse denne problemstillingen har vi utviklet to forskningsspørsmål:

1. I hvilken grad måler åpne oppgaver elevenes lesekompetanse i å tolke og sammenholde informasjon, og å finne informasjon i tekster?
2. Hvor presise, dekkende og anvendelige er kriteriene for skåring som gis i vurderingsveilederen, og i hvilken grad svarer skårernes vurderingspraksis til disse kriteriene?

1.3 Forskningsdesign

Et godt forskningsdesign kjennetegnes av at de ulike delene i masteroppgaven henger godt sammen (Gleiss & Sæther, 2022, s. 26). I vårt masterprosjekt har vi valgt å kombinere en kvalitativ og en kvantitativ tilnærming for å besvare våre forskningsspørsmål. Først utfører vi en kvalitativ tekstgjennomgang av prøvetekstene *Stargate* og *Snikfotografen*, oppgaveteksten og vurderingsveilederne, fra nasjonale prøver i lesing høsten 2023.

Deretter foretar vi en kvantitativ innholdsanalyse ved selv å skåre og kategorisere elevsvar fra et utvalg på totalt 1200 elevbesvarelser fra de to åpne oppgavene. Videre sammenligner vi resultatene fra skåring mellom skårer 1 Cecilie og skårer 2 Christina, og rapporterer fra vurderingsarbeidet fram til en omforent vurdering. Deretter blir omforent skår sammenlignet med de autentiske resultatene som foreligger i datautvalget. Vi omtaler lærere som har gitt sine vurderinger som *Lærer x*. Denne betegnelsen representerer alle lærerne som har utført vurdering og skåring i vårt datasett.

Avslutningsvis utfører vi statistiske beregninger av reliabilitetsmål i skåring av de åpne oppgavene ved bruk av Cohens kappa og Fleiss kappa.

1.4 Oppgavens struktur

Oppgaven er delt inn i seks hoveddeler. Etter innledningen presenterer vi et teorikapittel som tar for seg lesing og nasjonale prøver. Teoridelen er delt inn i fire underkapitler. Vi starter med å redegjøre for hva lesing er, og de tre leseprosessene.

Deretter presenterer vi et historisk og politisk perspektiv på prøvene i Norge. Videre beskrives det teoretiske rammeverket som prøvene er tuftet på, de ulike elementene nasjonale prøver består av og hvilke kvalitetskrav som settes til prøvene. Til sist viser vi til hvordan åpne oppgaver er omtalt i forskningslitteraturen og hvordan validitet og reliabilitet er definert i vitenskapsteorien.

I metodekapittelet blir vitenskapsteoretisk forankring, begrunnelse for våre metodevalg, datainnsamling og utvalg gjort rede for. Studiens kvalitet blir deretter drøftet gjennom en validitet- og reliabilitetsdiskusjon, der vi gjennomgår våre roller som skårere i masterprosjektet. Forskningsetiske refleksjoner spesielt knyttet til ivaretakelse av elevenes personvern avslutter kapittelet.

Det fjerde kapitelet er delt i to analyser: *Stargate* og *Snikfotografen*. De to analysene følger samme struktur. Underkapitlene behandler presentasjon av tekstgjennomgang, data og funn. Vi starter med en tekstgjennomgang av prøvetekstene, og veien fram til det åpne spørsmålet. Deretter følger resultater og funn fra kategorisering, før en analyse av kvalitative funn i diskusjon fram til omforent skår blir presentert. Videre fremstiller vi kvantitative resultater fra skåring av åpne oppgaver for 8. og 9. trinn. Graden av samsvar i skåring presenteres gjennom statistiske beregninger av Cohens kappa og Fleiss kappa.

I kapittel fem diskuterer vi tolkningsrommet og lærerens bruk av skjønn i vurderingsarbeid, og fører en validitet- og reliabilitetsdiskusjon. Kapittel seks avslutter denne masteroppgaven og vi svarer på våre forskningsspørsmål. Til sist retter vi blikket mot nye muligheter for de åpne oppgavene i nasjonale prøver.

2 Lesing og nasjonale prøver

I dette kapittelet vil vi redegjøre kort for ulike retninger innenfor leseteori og beskriver lesing som grunnleggende ferdighet. Vi vil forklare hvordan de tre aspektene for lesing er beskrevet i rammeverket, før vi beskriver hvordan nasjonale prøver i lesing har vært utfordret, endret og debattert. Deretter presenterer vi åpne oppgaver som forskningsobjekt og viser til studier som har satt søkelys på de åpne oppgavenes reliabilitet i vurdering og skåring. Videre gjennomgår vi validitet og reliabilitet i vitenskapsteori, samt hvordan dette kan være en utfordring i prøver og tester avslutter teorikapittelet.

2.1 Å kunne lese

Forståelse knyttet til hva det vil si å kunne lese har endret seg over tid og ulike forskningsteorier vektlegger forskjellige aspekt. Forskning har gått fra å være tekstfokusert til å i større grad se hvilket arbeid den kognitive prosessen med å skape mening og forståelse i tekst krever. I den sosiokulturelle tilnærmingen er oppmerksomheten rettet mot konteksten lesingen skjer i (Weyergang, Siljan, & Frønes, 2023, s. 118). Kulbrandstad beskriver lesing som både en individuell ferdighet og en sosial og kulturell praksis. Både kognitive og sosiokulturelle teorier er viktig for å forstå leseutvikling til elever (Kulbrandstad, 2022, s. 69).

Det er likheter mellom måten lesing er definert i PISA og literacy-tenkningen i Kunnskapsløfte fra 2006. Likheten er at lesing regnes som en grunnleggende ferdighet på tvers av fag, og lesing blir sett på som en avgjørende kompetanse for enkeltindividet og samfunnet. I Fagfornyelsen (LK20) ble det foretatt noen endringer i hvordan de grunnleggende ferdighetene er beskrevet i fagplanen. Alle de grunnleggende ferdighetene for lesing, regning, skriving, muntlige og digitale ferdigheter beskrives opp mot fagplanen med fagspesifikke ferdigheter (Jensen, Frønes, Kjærnsli, & Roe, 2020, s. 22). De nasjonale prøvene måler de grunnleggende ferdighetene i lesing, og læreplanen i norsk beskriver følgende:

> Norsk har et særlig ansvar for opplæringen i å kunne lese. Utviklingen av å kunne lese i norsk går fra den grunnleggende avkodingen til å lese, tolke og reflektere over tekster i ulike sjangre, for ulike formål og av ulik lengde og kompleksitet (Utdanningsdirektoratet , 2020, s. 5)

Å kunne lese som grunnleggende ferdighet er avgjørende for læring i alle fag. Rammeverket omtaler lesing som: «(…) å kunne forstå, bruke, reflektere over, vurdere og engasjere seg i

tekster» (Utdanningsdirektoratet, 2022, s. 7). Videre fremheves det at lesing i dagens informasjonssamfunn krever lesing for ulike formål og av mange typer tekster. Rammeverket uttrykker også at leseren må ha kunnskap om lesestrategier i møte med ulike tekster. Med bakgrunn i denne forståelsen, er leseforståelse delt inn i tre leseprosesser i de nasjonale prøvene i lesing. Å finne informasjon, å tolke og sammenholde informasjon og, å reflektere over og vurdere teksters form og innhold (Utdanningsdirektoratet, 2022, s. 7). Disse tre leseprosessene vil vi redegjøre nærmere for under.

2.1.1 Å finne informasjon i teksten

Å finne informasjon i tekster krever at elevene skal kunne finne informasjon som er både eksplisitt og implisitt uttrykt. Eleven må forstå hva som er relevante innholdselementer og forstå tekstens organisering (Utdanningsdirektoratet, 2022, s. 7). Oppgaver designet for å teste denne delferdigheten på nasjonale prøver i lesing, er gradert i vanskelighetsgrad og plassert inn på ulike mestringsnivå. På 8. og 9. trinn er det fem mestringsnivåer som spenner fra 1 til 5. Denne progresjon kommer frem av beskrivelsene i rammeverket. Vanskegrad øker i forhold til konkurrerende informasjon og hvor mye informasjon som må lokaliseres eller gjenkjennes. Det vil si at det er flere faktorer som avgjør vanskegraden på en gitt oppgave (Roe, Ryen, & Weyergang, 2018, ss. 30-32). I vårt prosjekt representerer teksten *Snikfotografen* oppgave 3 leseformålet «å finne informasjon», mestringsnivå 4.

2.1.2 Å tolke og sammenholde informasjon

Vanskegraden til dette leseformålet avhenger av hvilken tolkning som kreves. Fra informasjon som er eksplisitt uttrykt til å i større grad bruke forkunnskap for å skape mening og sammenheng (Utdanningsdirektoratet, 2022, s. 8). Tidligere ble dette leseformålet omtalt som «å tolke og forstå». I nåværende rammeverket er «forstå» erstattet med «å sammenholde informasjon».

I vårt masterprosjekt er leseformålet «å tolke og sammenholde informasjon» er knyttet opp til teksten *Stargate*, oppgave 5. I rammeverket beskrives dette for 8. trinn, mestringsnivå 4: «forstå hvordan ikke tydelige informasjonselementer i en eller flere tekster henge sammen, og/eller hvordan disse henger sammen med teksten som helhet. Å forstå meningsinnhold som står i motsetning til det forventede» (Utdanningsdirektoratet, 2022, ss. 21-22). I arbeid med prøvetekstene oppdaget vi at oppgavens formål i *Stargate* er beskrevet som «å kunne tolke og trekke slutninger på bakgrunn av informasjon i tekst». I litteraturen om de nasjonale prøvene

har vi ikke funnet noen forklaring på denne endringen i ordlyd, men velger å trekke det frem for å belyse at vi har observert dette.

2.1.3 Å reflektere og vurdere teksters form og innhold

Elevene må forholde seg selvstendig og kritisk til teksten de leser. De må kunne kommentere innholdet i teksten, være kritisk, begrunne egne synspunkter, analysere og kunne vurdere tekstene (Utdanningsdirektoratet, 2022, s. 8). Hvor vanskelig oppgavene er, vil avhenge av hvilken refleksjon som kreves. Å begrunne synspunkter om tekstens form er ofte vanskeligere, enn gjøre sammenligninger og forklaringer som skal relateres til elevens erfaringer (Roe, Ryen, & Weyergang, 2018, s. 30). I vårt prosjekt er dette leseformålet knyttet til den åpne oppgaven nr. 5 til teksten *Skulearbeid*. Den åpne oppgaven til denne teksten ble skåret av oss, men ble tatt ut av prosjektet grunnet plasshensyn.

2.1.4 Leseforståelse

Vi forstår det slik at disse hovedkategoriene som leseprosessene deles inn i, er operasjonaliseringen som ligger til grunn for måling av lesing i nasjonale prøver. Sammenhengene i tekst er ikke alltid tydelig, og leserne må selv trekke slutninger. Hva skal til for at elevene skal forstå tekstene som benyttes i de nasjonale prøvene? Denne prosessen kalles ofte inferens. Det kan være fra det enkle å se sammenhenger mellom ulike deler av teksten, til å bruke bakgrunnskunnskap for å gi mening til teksten (Frønes & Ryen, 2020, s. 138). I vårt prosjekt er leseforståelsen sentral i forhold til hva som kreves av elevene for å lese *Stargate* og *Snikfotografen* og til å besvare de åpne oppgavene. I våre tekstgjennomganger, kapittel 4, vil vi komme tilbake til dette.

Tidligere kunnskap og erfaringer har betydning for hvordan vi oppfatter en tekst. Det er bred enighet om at forkunnskap påvirker tolkning av tekst. Innenfor ulike vitenskapelige retninger finnes det ulike måter å beskrive disse fenomenene på. Wolfgang Iser representerte de tyske resepsjonsestetikerne. Han er kjent for sin teori om tekstenes tomme plasser (Roe & Blikstad-Balas, 2022, s. 33). Ifølge Frønes og Ryen (2020) har Iser en fenomenologisk tilnærming til lesing. Han er opptatt av leseprosessen og hvordan leseren skaper menig i møte med teksten. Ifølge Iser har alle tekster et tomrom «gaps» leseren må fylle. Han beskriver lesing som en dynamisk prosess mellom tekst og leser. Det vil kunne være ulikt hva de forskjellelige leserne fyller inn i tomrommet. Hva leseren forstår påvirkes av lese- og livserfaring. Forståelse vil være ulik, men innenfor noen rammer og grenser som forfatteren har skapt. Disse grensene for tolkning inviterer leseren til å lese teksten på en bestemt måte og med en leseposisjon.

Denne posisjonen kaller Iser for den implisitte leser, dette er ikke en reel leser, men en tenkt struktur leseren kan bruke i møte med tekst for å skape mening (Frønes & Ryen, 2020, s. 139). Begrepet modelleser ligner på Isers implisitte leser. Umberto Eco tar også utgangspunkt i at alle tekster har små hull som må fylles med mening av leseren. Disse tomrommene vil fylles på ulike måter avhengig av blant annet leserens livs- og teksterfaring i tillegg til forhold i lesesituasjonen. En modelleser av en tekst vil lese teksten velvillig og etter konvensjonen til sjangeren. Ulike spor i teksten vil bli oppfattet og forstått på en bestemt måte (Frønes & Ryen, 2020, s. 139).

Modelleseren er med andre ord en teoretisk modell som kan brukes for å begrunne hvordan måling av elevers lesekompetanse kan gjøres. Nasjonale prøver i lesing bygger på prinsippet om at det finnes noen måter å lese på som er bedre og mer korrekt enn andre. Modelleseren representere et tenkt ideal, som fanger opp alle intensjoner i teksten og realiseres den fulle potensial og meningsinnhold. Ingen lesere er en perfekt leser, vi har en ufullkommenhet som kan skyldes dårlige leseferdigheter, lite forkunnskaper, lav selvtillit og manglende motivasjon (Roe, Ryen, & Weyergang, 2018, s. 64). I flervalgsoppgavene vil det være et riktig svar, i henhold til en tenkt modelleserstrategi. I de åpne oppgavene hvor elevene må formulere svarene sine skriftlig, er det flere svar som kan være riktig. Vurderingsveiledningen inneholder flere svar som representere ulike modelleserposisjoner som kan aksepteres (Frønes & Ryen, 2020, s. 139).

2.2 Nasjonale prøver i lesing

2.2.1 Nasjonale prøver i Norge

I den norske skolen har vi et vurderingssystem som består av lærerbaserte standpunktkarakterer og eksamener. Disse har lange tradisjoner. På begynnelsen av 2000-tallet ble også kartleggingsprøver og nasjonale prøver innført (Blömeke & Rolf, 2018). Norge deltar også i Pisa-undersøkelsen. PISA gjennomføres hvert tredje år, og ble gjennomført for første gang i 2000. Denne undersøkelsen måler elevenes kompetanse i lesing, matematikk og naturfag, men de tre fagene veksler om å være hovedfagområde (Frønes & Ryen, 2020, s. 14).

PISA-resultatene i 2003 tydet på at norske elevers kompetansenivå lå litt over gjennomsnittet i OECD, men under Finland, Sverige og New Zealand. Dette bidro til endring i norsk kunnskaps- og skolepolitikk. Slik bidro PISA til at de grunnleggende ferdighetene ble grunnlag for læring i alle fag (Matre, et al., 2021, s. 24). Videre ble det innført et nasjonalt kvalitetsvurderingssystem for grunnopplæringen (NKVS) for grunnskolen, vedtatt av Stortinget (St.meld. nr. 30, 2003-2004). Systemet ble utviklet for å overvåke og forbedre kvaliteten på utdanningen gjennom regelmessig evaluering av skoleprestasjoner. Dette kan ses på som et skifte i skolepolitikken i Norge. Kvaliteten i skolene skulle nå i større grad kontrolleres eksternt (Vestheim & Sem, 2019, s. 28). De sentrale målingsverktøyene for det nasjonale kvalitetsvurderingssystemet er kartleggingsprøver, nasjonale prøver og elevundersøkelsen. Slik er det fortsatt.

Den første nasjonale prøven i lesing ble gjennomført i våren 2004 for alle elever på fjerde og tiende trinn. I 2005 ble prøven utvidet til å gjelde sjuende trinn og VG1. I tillegg ble skriving innført som et nytt fagområde. Prøvene ble lagt til slutten av barne- og ungdomstrinnet, men ble kritisert. Argumentasjonene dreide seg om at kartleggingen i hovedsak hadde en summativ funksjon, og i liten grad ble et verktøy i utviklingsøyemed for elevene og lærerne. Prøvene ble også gjenstand for faglig kritikk i evalueringsrapporter. Kritikken var rettet mot måten vanskegraden og poengene ble beregnet på, og at prøvene i liten grad la til rette for formativ vurdering. De nasjonale prøvene ble satt på pause i 2006, men høsten 2007 var prøvene tilbake. De ble da lagt til høstsemestret for 5. og 8. trinn. På dette tidspunktet var prøven utformet med utgangspunkt i et rammeverk som korresponderte med Kunnskapsløftet 2006 (Roe, Ryen, & Weyergang, 2018, ss. 22-23). Det var kritikk og en del motstand mot nasjonale prøver da de ble innført. Faglig kom frem gjennom to evalueringsrapporter. Kritikken var rettet mot prøvenes måte å beregne vanskegrad på, i tillegg til at de i for liten grad la til rette for formativ vurdering (Lie, Hopfebeck, Ibsen, & Turmo, 2005).

Debatten rundt prøvene fikk oppmerksomhet i det offentlige rom. Norske riks- og regionaviser satt søkelys på blant annet resultatene. Skoler, kommuner og fylker ble satt opp mot hverandre. Noe av kritikken var at det skapte et ensidig søkelys på testingen (Vestheim & Sem, 2019). Den politiske debatten og retorikken mellom høyre- og venstresiden har og vil nok fortsette å prege samfunns- og skoledebatten i fremtiden. I 2005 var Øystein Djupedal (AP) kunnskapsminister, han ville gjøre endringer for å legge mer vekt på den formative vurderingen. Djupedal som representerte Sosialistisk Venstreparti, kritiserte Kristin Clemets «prestisjeprosjekt» Skoleporten, hvor resultatene skulle publiseres (Vestheim & Sem, 2019).

Skaftun (2006) skriver at de nasjonale prøvene har oppstått i et spenningsfelt mellom sprikende interesser mellom fag og politikk. Politiske partier og interesseorganisasjoner har meninger om hvordan skolen bør være. (Skaftun, 2006, s. 123). Motsetningene oppstår mellom oppfatningen av nasjonale prøver som verktøy for ekstern evaluering, og prøver som kan brukes formativt og læringsstøttende. Dette er viktig i et overordnet perspektiv, samtidig som vi velger å forholde oss nøytrale til det politiske ordskifte i denne masteroppgaven.

De nasjonale prøvene har gjennomgått flere endringer fra oppstarten til nå i dag. Nå gjennomføres de nasjonale prøvene på 5. og 8. trinn i regning, engelsk og lesing hvert år. Prøvene i regning og lesing gjennomføres også årlig på 9. trinn. Det er Utdanningsdirektoratet som har ansvar for prøvene. De nasjonale prøvene gir informasjon på flere nivåer. Det første formålet er som styrings- og kontrolldokument for skole, kommune, fylke og nasjonalt. Det andre formålet er formativt vurdering for videre utvikling og planlegging av undervisning og læring (Blömeke & Rolf, 2018).

Prøvene ble gjennomført på papir frem til 2016, i dag er prøvene digitale. Fra perioden 2007 til 2013 ble det laget nye prøver for hvert år. Dermed var det ikke grunnlag for å sammenligne utviklingen over tid. Fra 2014 til 2021 ble det benyttet et nytt analysegrunnlag, IRT, som grunnmetode. Ankeroppgaver med samme utforming ble lagt inn for å kunne følge utviklingen over tid. Den siste endringen i forhold til ankeroppgaver kom i 2022 (Bjørnsson, 2022, s. 11).

Høsten 2023 kom det et nytt PISA-sjokk, da resultatene for 2022 viste betydelig nedgang i både lesing, regning og naturfag sammenlignet med resultatene i 2018. I tillegg økte andelen elever som prestere på det laveste mestringsnivået (Jensen, et al., 2023, s. 3).

Utdanningsforbundets landsmøte vedtok høsten 2023 at de ønsker å gå inn for å avvikle de nasjonale prøvene og erstatte de med et faglig- pedagogisk verktøy (Helland & Tresse, 2023).

I tillegg fulgte en rapport fra utvalget for kvalitetsutvikling, ledet av Tine Prøitz. Rapporten konkluderer med at dagens ordning er ressurskrevende og bidrar til målforskyving i opplæringen (Ruud, 2023).

> Utvalget foreslår å avvikle dagens nasjonale prøver og utvikle nye læringsstøttende prøver med formål om å gi informasjon om elevenes grunnleggende ferdigheter, for å bruke det i faglig-pedagogiske kvalitetsutviklingsarbeidet (NOU 2023: 27, s. 15).

2.2.2 Rammeverket til nasjonale prøver i lesing

Nasjonale prøver utvikles og gjennomføres av Utdanningsdirektoratet på oppdrag fra Kunnskapsdepartementet. Prøveutviklerne lager også tekniske rapporter og innhold til støtteressurser (Utdanningsdirektoratet, 2022). De nasjonale prøvene i lesing er utformet etter et rammeverk som inneholder en detaljert beskrivelse av det aktuelle fagområdet. Prøvene utvikles av en ekspertgruppe som er sammensatt av forskere og fagpersoner innenfor de ulike fagområdene (Frønes & Ryen, 2020, s. 14). De nasjonale prøvene i lesing «skal måle i hvilken grad elevenes leseferdigheter er i samsvar med bekrivelse av lesing som grunnleggende ferdighet» (Utdanningsdirektoratet, 2022, s. 7). Prøven måler ikke bare lesing i norskfaget, men lesing som grunnleggende ferdighet i alle fag.

Tekstene som benyttes i prøvene, skal representere et mangfold av tekster elevene møter i fagene og i samfunnet ellers. Det er både skjønnlitterære- og sakprosatekster, og begge målformene er representert. Målene i læreplanverket gjenspeiles i prøvene. Tekstene for nasjonale prøver i lesing 2023 er prøvetekster som i utgangspunktet er autentiske tekster. Det vil si at de ikke er spesialskrevet til prøvene. I tekstutdragene vi studerte, fant vi at det var gjort enkelte tilpasninger og strukturelle endringer. Disse vil omtales i kapittel 4., presentasjon og analyse av data.

Den nasjonale prøven i lesing for høsten 2023 bestod av syv tekster på 9. trinn og seks tekster i tillegg en ankertekst på 8.trinn. Elevene må besvare både flervalgsoppgaver og åpne oppgaver der elevene måtte formulere skriftlige svar. På flervalgsoppgavene skal elevene velge mellom fire svaralternativ, hvor ett er riktig og tre er feil. De åpne oppgavene benyttes ofte på tolknings- og refleksjonsoppgaver. Høsten 2023 var det syv åpne oppgaver på 9.trinn. I rammeverket står det at de åpne oppgavene ikke skal utgjøre mer enn 25 % av oppgavene totalt i prøven (Utdanningsdirektoratet, 2022).

Under «Lesing som grunnleggende ferdighet» omtaler rammeverket lesing som en komplisert aktivitet: «Lesing er en sammensatt og komplisert aktivitet, som blant annet påvirkes av avkodingsferdigheter, lesehastighet, flyt, vokabular- og begrepsforståelse, samt kunnskap om tekststruktur og tekstens tema» (Utdanningsdirektoratet, 2022, s. 7). Leseren må også kunne identifisere formålet med teksten og velge en strategi for å forstå ulike tekster. Leseforståelse måles i de nasjonale prøvene i lesing som følgende tre leseprosesser: 1. Å finne informasjon i tekst, 2. Å tolke og sammenholde informasjon og 3. Å reflektere over og vurdere teksters form og innhold. Det er viktig for oss å prøve å forstå de grunnleggende aspektene ved disse tre leseprosessene som rammeverket deler leseforståelse inn i.

I vår gjennomgang av tekster, vurderingsveilederen og skåringen av elevsvarene, vil vi legge vekt på å identifisere forholdet mellom validitet og reliabilitet. Teksten og konstruksjon av det åpne spørsmålet påvirker elevsvarene. Elevsvaret, sett opp mot vurderingsveilederen påvirker vurdereren sitt arbeid i skåringen.

2.2.3 Skriveprøver i Norge

Vi synes det er viktig å ta med historikken til skriveprøvene, da utfordringen i måling av komplekse kognitive ferdigheter som skriving og lesing har noen fellestrekk. Elevene må besvare de åpne oppgavene skriftlig. For vårt prosjekt er det derfor nyttig å legge merke til hvilke utfordringer som oppstod med validitet og reliabilitet i vurderingsarbeidet for skriveprøvene. Med innføringen av de grunnleggende ferdighetene vokste det også frem et behov for å kunne kartlegge skriveferdigheter. På begynnelsen av 2000-tallet var det økt fokus på skriving som kompetanse. I Kunnskapsløftet 2006 ble skriving definert som en av fem grunnleggende ferdigheter. For å måle denne ferdigheten i skolen ble det i 2005 utformet nasjonale prøver i skriving. Allerede i 2006 ble skriveprøvene avviklet på grunn av manglende normer og tolkningsfellesskap mellom lærerne som skulle vurdere prøvene. (Skar & Aasen, 2018). I 2009, begynte arbeidet med å utvikle utvalgsprøver og læringsstøttende prøver i skriving for 4. og 7.trinn. Det ble også arbeidet med skolering av et vurderingspanel. Den første utvalgsprøven ble gjennomført i 2012 og ble grunnlag for utviklingen av læringsstøttende prøver i skriving. I 2016 ble læringsstøttende prøver i skriving offentliggjort og var tilgjengelig for alle skoler gjennom Skrivesenteret sin nettside. Prøven var en ressurs som det var frivillig for lærerne å benytte som støtte i skriveopplæringen (Skrivesenteret, 2016). Ordningen med utvalgsprøver i skriving ble avviklet i 2016 på grunn av økonomiske hensyn og manglende budsjettering. Prøvene var ikke en del av nasjonalt kvalitetsvurderingssystem for skolen.

Det har vist seg vanskelig å teste skriving med god validitet og reliabilitet (Skar & Aasen, 2018). Det resulterte i at de nasjonale skriveprøvene ble avviklet av Utdanningsdirektoratet uten motstand fra Kunnskapsdepartementet, etter at Stortinget skar ned på bevilgingen. 40 millioner var brukt på utviklingsarbeidet av skriveprøvene, men disse ble ikke videreført. Det mangler i dag en nasjonal strategi for å styrke elevers skriveferdigheter (Berge, 2019, s. 176). I dette masterprosjektet er måling og vurdering i åpne oppgaver av lesekompetanse hovedfokus.

Utfordringer med validitet og reliabilitet i testing av skriving er beskrevet av Skar og Aasen (2018). Forskerne viser til problematikken rundt vektlegging av validitet og reliabilitet i vurdering av skriving. Reliabiliteten omtales som konsistensen i vurderingen. Vurderingen skal være mest mulig lik uavhengig av hvem som vurderer den og når prøven blir gjennomført. Validitet sett i forhold til at det er mulig å tolke testresultatet meningsfullt i henhold til konstruktdefinisjonen. Videre omtaler de at prioriteringen mellom validitet og reliabilitet, får konsekvenser for hvordan skrivekompetansen måles. Skrivekompetanse kan måles indirekte ved at eleven for eksempel skal finne feil (Breland, Bridgman & Fowles- referert av Skar og Aasen), eller dirkete gjennom oppgaver hvor eleven produserer egne tekster. Valget mellom disse to refereres til som et valg mellom reliabilitet og begrepsvaliditet (Skar & Aasen, 2018, ss. 8-9).

Forfatterne løfter også frem kompleksiteten i å vurdere skriving i forhold til de ulike komponentene som inngår i en slik vurderingssituasjon. Dette illustrerer de gjennom McNamaras modell for språkprøver. Vi finner likheter til denne og komponentene som inngår i måling av leseferdigheter gjennom de åpne oppgavene i nasjonale prøver i lesing (Skar & Aasen, 2018). I vårt prosjekt bruker vi begrepet «delene» for å beskrive de ulike komponentene; teksten, oppgaveteksten og vurderingsveilederen.

2.3 Nasjonale prøver i lesing oppbygging

2.3.1 Åpne oppgaver

I de åpne oppgavene skal elevene formulere skriftlige svar med egne ord. Noen oppgaver krever korte svar som tall eller navn, mens andre krever kortere begrunnelser og forklaringer. Svarene på de åpne oppgavene vurderes og skåres av ulike lærere på skolen. De fleste lærere skårer elevbesvarelsene til sine egne elever. Dette skjer anonymt i PAS (Roe, Ryen, & Weyergang, 2018, s. 185). Det vil si at lærerne får elevsvarene til vurdering på nettsiden, uten at de kan knytte elevsvaret opp mot elevens identitet.

Skaftun et al (2006) tematiserer det teoretiske grunnlaget de nasjonale prøvene er basert på i sin artikkel «Tilnærminger til et teoretisk rammeverk for de nasjonale prøvene i lesing». Her legges det vekt på at krav om høy sensorreliabilitet i skåring av åpne oppgaver, innebærer vurderingsveiledere som er mest mulig entydige. Dersom vurderingsveilederne til skåring ikke skal bygge på tolking og skjønn hos vurdereren, må kriteriene for poenggiving være svært nøyaktig i sin formulering, slik som i PIRLS og PISA (Skaftun et al. 2006 s. 361). Samtidig setter de søkelys på hvordan tolknings- og refleksjonsoppgaver, som gjør det mulig for eleven å reflektere selvstendig, ofte vil falle ut, hvis kriteriene for skåring blir for rigide. Dette begrunnes med at kriteriene for skåring forutsetter kvalifisert skjønn av vurdereren, noe som kan utfordre sensorreliabiliteten. Et viktig poeng Skaftun et al legger vekt på, er at læreren er personer som antas å ha dette kvalifiserte skjønnet (Skaftun, Roe, Narvhus, & Solheim, 2006, s. 361).

I vårt arbeid med masterprosjektet forsøkte vi å finne ut hvilke kvalifikasjoner lærere som skårer de åpne oppgavene, skal inneha. I rapporten utarbeidet av Nordisk institutt for studier av innovasjon, forskning og utdanning (2013) analyserte en arbeidsgruppe arbeid og erfaringer med nasjonale prøver som system. I utarbeiding av en lærersurvey fant arbeidsgruppen at det ikke finnes noen «(…) tilgjengelige opplysninger om hvordan populasjonen av norske lærere, som er involvert i gjennomføringen av nasjonale prøver, er sammensatt (Selan, Vibe, & Hovdhaugen, 2013, s. 33)». Vårt inntrykk er at skolens ledelse sørger for denne kvalitetssikringen, gjennom å tildele skåringsansvaret til utvalgte lærere i PAS[2]. Vi synes dette er interessant, og vi undrer oss over at det tilsynelatende mangler et overordnet system som ivaretar skåringsprosessen og sikrer objektiviteten i vurdering av de åpne oppgavene. Denne problemstillingen faller noe utenfor vår oppgave, men vi velger å belyse at det kan utvikle seg ulike praksiser på ulike skoler.

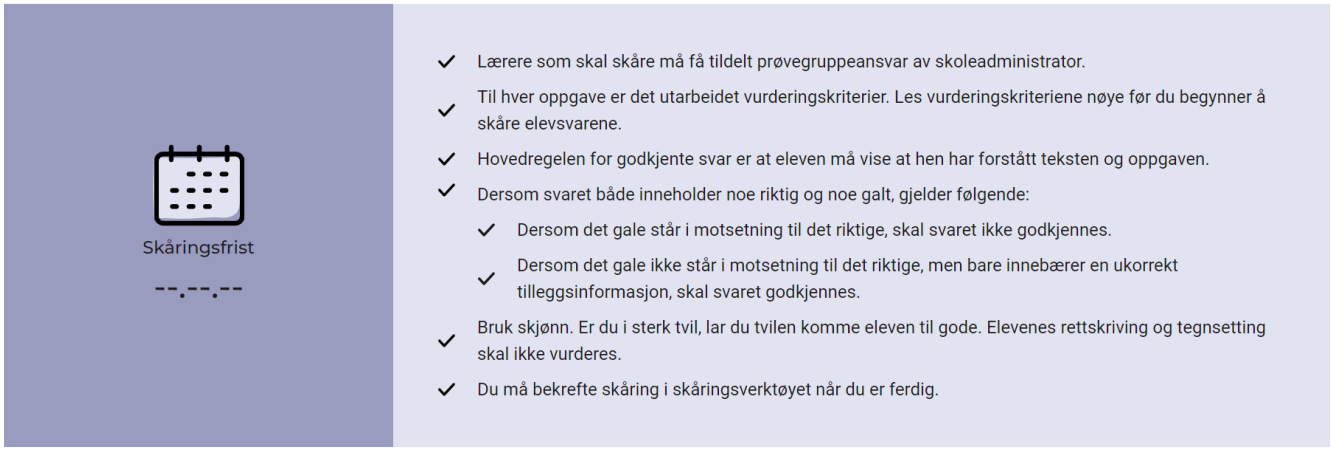
Skåringen av de åpne oppgavene registreres inn i Utdanningsdirektoratet sitt prøveadministrative system PAS. Utdanningsdirektoratet har også laget en eksempelvideo som forklarer gangen i skåringen, slik at lærere vet hvordan skåringen foregår og hva som forventes av dem i arbeidet med skåring[3]. Lærere som skal skåre åpne oppgaver, møter et skjermbilde med informasjon om skåring, samt oversikt over elevgruppene en har ansvar for.

---

[2] PAS er Utdanningsdirektoratets elektroniske prøveadministrative system.

[3] https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/administrere-nasjonale-prover2/#a124951 Sist sett 6. mars 2024.

Her presiserer Utdanningsdirektoratet hvilke retningslinjer skårere skal følge. Se figur 1 under.



Figur 1: Retningslinjer for skåring til lærer i PAS

Rektor har ansvar for at skåringsarbeidet blir gjennomført innen fastsatt frist. Oppgaver som ikke blir skåret inne fristen, blir gitt 0 poeng (Utdanningsdirektoratet, 2023).

Å formulere et skriftlig svar på de åpne oppgavene betyr at det kan være en hindring for svake skrivere å uttrykke seg godt nok for å få et godkjent svar. De kan ha forstått teksten og oppgaven, men ikke klare å få det frem i skriftlig uttrykksform. Dette kan da problematiseres som at oppgaven i større grad måler skriveferdigheter og kan argumenteres for å være en trussel mot validiteten (Roe & Blikstad-Balas, 2022, ss. 221-222). Nettopp dette var årsaken til at åpne oppgaver ikke ble videreført for prøvene på femte trinn (Roe, Ryen, & Weyergang, 2018, s. 177).

Ifølge Solheim og Skaftun sin artikkel har det siden 80- tallet vært diskusjon rundt konstruksjon av oppgaveformat i leseprøver. I 2001 ble constructed response -items (CR) benyttet som oppgavetype i PIRLS. Oppgavetypen er viktig for å kunne måle lesing som en generativ prosess. Noe som vi forstår som en aktiv prosess der leseren skaper mening og ikke bare passivt motta gjennom multiple-choice (MC). Åpne oppgaver gir elevene mulighet til å uttrykke seg selvstendig å uttrykke egne tanker (Solheim & Skaftun, 2009). En leseprøve uten åpne oppgaver vil kanskje måle en snevrere del av lesekompetansen og noe som igjen kan sies å svekke validiteten til prøven (Roe, Ryen, & Weyergang, 2018, s. 192). I vårt prosjekt omtaler vi oppgavetypene som, flervalgsoppgaver og åpne oppgaver. Tabellen under viser hvilke åpne oppgaver som var knyttet til de ulike leseformålene for nasjonale prøver i lesing, høsten 2023.

Tabell 1: Oversikt over åpne oppgaver – Nasjonal prøve i lesing, 2023

Tittel på tekst	Faglig plassering	Teksttype	Leseformål	Mestrings-nivå
Stargate oppg. 5	Norsk	Skjønnlitterært utdrag	Å tolke og trekke slutninger på bakgrunn av informasjon i teksten	Nivå 4
Skulearbeid oppg. 5	Samfunnsfag Matematikk	Rapport og diagram	Å reflektere over tekstens form og innhold	Nivå 4
Snikfotografen oppg. 3	Samfunnsfag	Nettartikkel	Å finne informasjon i teksten	Nivå 4
Den korte historien om egget- oppg. 2	Naturfag Matematikk	Fagtekst	Å finne informasjon i teksten	Nivå 2
Et skår i gleden Oppg. 4	Norsk Samfunnsfag	Annonse	Å reflektere over tekstens form og innhold	Nivå 3
Et skår i gleden Oppg. 6	Norsk Samfunnsfag	Annonse	Å finne informasjon i teksten	Nivå 2
Tasmansk pungulv Oppg. 4	Naturfag Samfunnsfag	Nyhetsartikkel Fagtekst	Å tolke og trekke slutninger på bakgrunn av informasjon i teksten	Nivå 3

2.3.2 Vurderingsveilederne

Til nasjonale prøver i lesing utformes det en vurderingsveileder. Denne er knyttet opp til hver av de åpne oppgavene, og lærerne oppfordres til å lese denne nøye før skåringsarbeidet. I denne finnes eksempler på svar som skal vurderes som rett eller galt, og for noen oppgaver er det beskrevet minimumssvar. Den inneholder eksempler på autentiske elevsvar fra pilotering av oppgavene. Målet med veilederen er å øke reliabiliteten i skåringen, det vil si at flere lærere skal kunne vurdere samme svar og være enig om hva som er rett eller ikke godkjent svar (Roe, Ryen, & Weyergang, 2018, s. 185).

I følge Skaftun (2006) vil vurderingsveilederen aldri være entydig, og det vil alltid vil være noe rom for tolkning. Samtidig vil sensorreliabiliteten for de åpne oppgavene aldri bli 100%. Om vi begynner å stole på vurderingsguiden som en autoritativ sannhet har vi, ifølge Skaftun, gått i den «testteoretiske rottefella». Vurderingsguiden har som formål å legge grunnlag for høyere presisjon i tolkings- og vurderingsarbeidet (Skaftun, 2006, s. 40). I de nasjonale prøvene i lesing høsten 2023 ligger vurderingsveiledningen som en egen PDF. Vurderingsveiledningen ligger også som en digital fil man kan åpne underveis i skåringsprosessen hvis den ikke er skrevet ut på forhånd. I forbindelse med skåring av åpne oppgaver for egen klasse høsten 2023 erfarte vi at elevsvarene som skulle skåres ble presentert i tilfeldig rekkefølge. Det vil si at vi fikk syv besvarelser fra *Tasmansk pungulv* til skåring først, for deretter å måtte ta stilling til fire elevbesvarelser fra *Skulearbeid*, deretter fikk vi åtte elevsvar til teksten *Stargate.* Antallet elevsvar tilknyttet de ulike tekstene varierte i oppstarten av skåringsprosessen, og etter hvert økte antallet elevsvar fra samme prøvetekst.

Vår personlige opplevelse var at den tilfeldige rekkefølgen i skåring ble et irritasjonsmoment når vi ikke fikk gjort oss ferdig med elevsvarene tilknyttet en prøvetekst om gangen. Dette resulterte i at vi måtte bla opp i ulike veiledere underveis i skåringen. I skåringsprosessen vi utførte som ansatte i skolen, hadde vi mulighet til å navigere oss fram og tilbake i elevsvarene. Det var en fin mulighet dersom vi måtte korrigere noen skåringer etter hvert som vi tok stilling til flere elevsvar. Ettersom elevsvarene ikke var systematisk sortert etter prøvetekstene var det krevende å skulle gå tilbake for å korrigere en skåring der det var nødvendig. Vårt mål som lærere er alltid å gi elevene våre rett vurdering i skåring, og den tilfeldige rekkefølgen ble en tidkrevende tilleggsbelastning i skåringen.

2.3.3 Resultatene

Roe og Blikstad-Balas (2022) skriver at resultatene på de nasjonale prøvene i lesing presenteres i skalapoeng som er en omregning basert på antall rette svar. Vanskelige oppgaver blir tillagt større vekt, enn lette oppgaver som mange har løst. Dette innebærer at elever med samme antall riktige svar kan få ulik skalapoengsum. I tillegg blir resultatene presentert i form av prosentandel elever som fordeler seg på de fem ulike mestringsnivåene. Det er utarbeidet en beskrivelse av hva som kjennetegner leseferdighetene til gjennomsnittseleven på de ulike mestringsnivåene (Roe & Blikstad-Balas, 2022, s. 211). Ifølge en rapport fra Stiftelsen Frischsenteret for samfunnsøkonomisk forskning har norske elever har hatt et snitt på 50 skalapoeng fra 2014 frem til 2021 (Markussen, Ræder, Røgeberg, & Raaum, 2024).

Etter gjennomføring av de nasjonale prøvene presenterer Utdanningsdirektoratet en rapport med hovedfunnene fra prøven. I presentasjonen av resultatene fra høsten 2023 presenterer Utdanningsdirektoratet hovedtrekkene som blant annet går på hvordan 8. og 9. trinn sammenlignes med tidligere år, prestasjoner knyttet til kjønn og deltakelsesprosenter. Her nevnes ikke resultater knyttet til de ulike oppgavetypene, flervalgsoppgaver og åpne oppgaver, spesielt (Utdanningsdirektoratet, 2023). Lærere har tilgang på detaljerte resultatoversikter for sine elever gjennom utdanningsdirektoratets prøveadministrative system PAS. Her kan læreren se klassens resultater av de ulike oppgavetypene, både flervalgsoppgaver og åpne oppgaver, sammenlignet med landsgjennomsnittet.

I vårt prosjekt vil ikke omtale beregningen av poeng som gjøres i etterkant av prøvene, dette faller utenfor vårt prosjekt. Vi er klar over kritikken som har vært rundt regnefeil og hvilken programvare som er benyttet i dette arbeidet.

2.3.4 Kvalitetskrav til de nasjonale prøvene i lesing

Lesing er en sammensatt og kompleks kognitiv ferdighet som i tillegg til de tekniske leseferdighetene også krever motivasjon. Konsentrasjon, intelligens, erfaringer og språklig bakgrunn. Med andre ord må prøver som skal måle dette ha godt teoretisk rammeverk og tydelige kriterier for hva prøven måler (Roe & Blikstad-Balas, 2022, s. 207).

De nasjonale prøvenes kvalitet omtales av flere. Roe et. al (2018) skriver at dersom de nasjonale leseprøvene skal ha verdi, må prøvene være av høy kvalitet og det må fremgå hva prøvene skal måle (Roe, Ryen, & Weyergang, 2018, s. 175). I rammeverket til prøvene slåes det fast at den skal være valid. Det utdypes med at leseprøven skal måle de grunnleggende ferdighetene i lesing (…) med bakgrunn i læreplanverket og kompetansemål i fag. Videre

utdypes det at oppgavene skal favne ulike ferdighetsnivå, fra det enkle til det komplekse, med høy presisjon. Prøvene skal måle ferdigheter i tråd med prøvens konstrukt opp mot rammeverket. Prøveutviklerne skal rapportere om dette, og de må forholde seg til en rekke strenge krav. De skal motivere likt. (Utdanningsdirektoratet, 2022, s. 15)

Det er viktig at leseprøvene er av høy kvalitet for at de skal ha en verdi, derfor er det svært viktig at kravene til validitet og reliabilitet er oppfylt. I rammeverket er reliabiliteten omtalt for oppgavenivå:

> Oppgaveutviklingen skal skje med klassiske mål for reliabilitet og oppgaveegenskaper, i tillegg til en item respons theory (IRT)-basert analyse av enkeltoppgaver og prøvene i sin helhet. Lenking av prøver mellom år og rapportering av resultater skal skje gjennom bruk av IRT- kalibrerings- og skaleringsmetodologi, med ankeroppgaver, skalerte skårer (skalapoeng), og faste mestringsnivågrenser fra år til år. IRT-kalibreringen skal gjennomføres med 2 PL IRT-modeller for dikotome oppgaver og med en GPCM-modell for polytome oppgaver (…) (Utdanningsdirektoratet, 2022, s. 10).

Det gjennomføres pilotering i flere faser for å sikre kvalitetene til oppgavene i de nasjonale prøvene i lesing. Da utarbeides det rapporter som beskriver å dokumentere både de lesefaglige og psykometriske vurderingene som blir gjort. Vi legger spesielt merke til at rapporten fra 2018 beskriver at «for framtidige prøver tar vi sikte på å pilotere vurderingsveiledningen blant at utvalg lærere og undersøke samsvaret mellom ulike vurdere etter gjennomføring av prøvene» (Fevolden, Thorsen, & Eriksen, 2019, s. 5)

2.4 Forskning av validitet og reliabilitet i åpne oppgaver

I arbeidet med masterprosjektet har vi undersøkt tidligere forskning av validitet og reliabilitet knyttet til åpne oppgaver. Vårt inntrykk er at de åpne oppgavene på nasjonale prøver i lesing er et nokså smalt forskningsfelt i Norge.

Solheim og Skaftun (2009) tar utgangspunkt i åpne oppgaver som oppgaveformat i storskala leseprøver med utgangspunkt i PIRLS 2001. Forfatterne ser på hvordan åpne oppgaver har blitt ansett som en viktig oppgavetype for å fange opp elevenes dypere forståelse av tekster, men påpeker at det er et potensiale for videre utvikling av de åpne oppgavene. For å frigjøre dette potensialet belyser de noen problematiske sider ved bruken av åpne oppgaver.
Det er en fare for at åpne oppgaver ikke måler elevenes dybdeforståelse tilstrekkelig, fordi det begrenses gjennom operasjonaliseringen og vurderingsveilederen (Solheim & Skaftun, 2009).

Michael Tengberg er en av forskerne som har forfattet flere artikler med validitet og reliabilitet knyttet til constructed respons items (åpne oppgaver). Tengberg (2017) redegjør for hovedskillene mellom de nasjonale leseprøvene i Skandinavia. I studien sammenlignes de nasjonale leseprøvene for ungdomsskolen i Norge, Sverige og Danmark for å finne likheter, men også hva som skiller de tre testdomenene fra hverandre. Funnene indikerer at de tre leseprøvene viser ulikheter i antall åpne oppgaver på nasjonale leseprøver. De norske nasjonale prøvene i lesing har en relativt liten andel åpne oppgaver med maksimalt 25 %, sammenlignet med vårt naboland Sverige med om lag 75 % åpne oppgaver (Tengberg, 2017, s. 94). Tengberg indikerer også bekymring knyttet til den lave sensorreliabiliteten i vurdering av åpne oppgaver i de svenske nasjonale prøvene. Resultatene Tengberg (2017) refererer til, er fra egen studie, og viser moderate nivåer av målt interrater reliabilitet med kappaverdi på ,73 (Tengberg & Skar, 2016, s. 8).

I Sverige har den store andelen åpne oppgaver blitt kritisert på grunn av utilfredsstillende interrater-reliabilitet. Men, fordi den svenske nasjonale prøven også inkluderer en større andel flervalgsoppgaver anses dette å øke objektiviteten i prøven (Tengberg, 2017, s. 94). Roe, Ryen og Weyergang beskriver de strenge kravene til den nasjonale leseprøvens validitet og reliabilitet ved å beskrive prøvekonstruksjonen. Forfatterne påpeker at det er lærerne som står for vurderingen av de åpne oppgavene, noe som kan lede til vurderingsskjevheter «(…) i form av at noen er «snille», mens andre er «rigide», selv om det finnes vurderingsveiledere for skåring (Roe, Ryen, & Weyergang, 2018, s. 185). Her peker de også på at andelen åpne oppgaver er relativt lav. Derfor forsvares andelen åpne oppgaver med liknende argumentasjon som i Sverige, ettersom den største andelen av oppgavene i de nasjonale prøvene i lesing er

flervalgsoppgaver og ikke (Tengberg, 2017, s. 97). Tengberg poengterer også at den svenske nasjonale prøven skåres av elevenes lærere og ikke eksterne sensorer, som i Norge.

Tengberg, Roe og Skar (2018) undersøkte sensorreliabiliteten i åpne oppgaver i nasjonale prøver i lesing i Norge. I artikkelen «Interrater reliability of constructed response items in standardized tests of reading» påpeker forfatterne at reliabiliteten i skåring av de åpne oppgavene representerer et område som har begrenset forskningsoppmerksomhet. Studien undersøker sensorreliabiliteten i vurdering av elevsvar på åpne oppgaver fra de nasjonale leseprøvene 2015. Forskerne fokuserer spesifikt på graden av sensorreliabilitet i poenggivingen mellom to grupper sensorer: ungdomsskolelærere og testutviklere (Tengberg, Roe, & Skar, 2018, s. 122). Funn fra studien viste at det var større diskrepans blant lærerparene sammenlignet med testutviklerparene. Testutviklerne hadde større andel av samsvar i skåring og færre tilfeller av diskrepans i elevsvar som var «vanskelig å skåre». For elevbesvarelsene der det oppsto diskrepans, ble det utført en kvalitativ tekstgjennomgang for å identifisere felles trekk ved disse tilfellene. Gjennomgangen viste at sensorer hyppigst opplevde uenighet av elevsvar i kategorien refleksjons- og tolkningsoppgaver (Tengberg, Roe, & Skar, 2018, s. 124). Det finnes fordeler og ulemper med lærere som skårer sine egne elever. Kostnadsulempene kursing og samskåring (co-rating) kan medføre skolene og samfunnet, er en utfordring, samtidig som samskåring gjennom digitale programmer kan være praktisk mulig i framtiden (Tengberg, 2017, s. 133). For å få til dette, må gapet mellom ulike sensorvurderinger være redusert.

Et tilleggsaspekt her er hvorvidt en da skal vurdere elevbesvarelsene uten å vurdere elevens skriftkyndighet (Tengberg, Roe, & Skar, 2018, s. 133). Som et forbedringsforslag legger de fram at vurderingsveilereden bør inneholde flere eksempler på elevbesvarelser som kan godkjennes og ikke godkjennes, slik at rommet for tvil hos sensor blir mest mulig begrenset. Diskusjoner og samarbeid mellom lærere vil være med på å utvikle lærernes faglige utvikling i vurdering, samt bidra til å styrke reliabilitet i skåring av åpne oppgaver (Tengberg, Roe, & Skar, 2018, s. 134).

2.5 Validitet og reliabilitet i vitenskapsteori

I dette underkapittelet vil vi redegjøre for validitet og reliabilitet som vitenskapsteoretiske begreper. Disse begrepene er sentrale i vårt masterprosjekt for å besvare våre forskningsspørsmål og vurdere kvaliteten i eget masterprosjekt. Dette redegjøres for i metodedelen kapittel 3.4 «Studiens kvalitet». Vi har valgt å kombinere en kvalitativ tekstgjennomgang med en kvantitativ innholdsanalyse i vårt prosjekt. I tillegg undersøker vi reliabilitet i skåringen av de åpne oppgavene. Det finnes ulike måter å forstå begrepene på innen de ulike vitenskapsteoretiske retningene (fagdisiplinene). I det følgende vil vi redegjøre for hovedlinjene.

2.5.1 Validitet og reliabilitet i kvantitativ forskning

Validitet og reliabilitet handler om gyldighet og pålitelighet i forskningen (Gleiss & Sæther, 2022, s. 201). Trusler mot validitet og reliabilitet i et prosjekt kan aldri bli helt borte, men dempes ved å være oppmerksom på validitet og reliabilitet gjennom hele forskningsprosjektet (Cohen, Manion, & Morrison, 2007, s. 133).

Validitet kan defineres som forskningens gyldighet. Det handler om hvor godt de ulike delene i forskningsdesignet henger sammen, fra datamaterialet til forskernes fortolkninger og konklusjoner. Likeså dreier det seg om metoden og utvalget egner seg til å besvare forskningsspørsmålene, og svarer man på problemstillingen? (Gleiss & Sæther, 2022, ss. 204-205). Forfatterne påpeker også at fenomener innenfor utdanningsforskning som for eksempel kompetanse er sammensatt og vanskelig å måle. I kvantitative forskningsprosjekter vil en fokusere på om man måler det man ønsker å måle, dette handler om begrepsvaliditet. Begrepsvaliditet handler om i hvilken grad forskeren har klart å operasjonalisere et teoretisk begrep slik at det blir målbart (Gleiss & Sæther, 2022, s. 205). Videre skriver Gleiss og Sæther at validiteten til et forskningsprosjekt også kan styrkes ved at man sammenligner egne funn med tidligere forskningsresultater på feltet. Her er det vanlig å se etter samsvar mellom ulike funn, og dersom man finner samsvar mellom egne funn og tidligere forskningsresultater, er det en styrke for validiteten. Refleksjon over styrker og svakheter ved eget prosjekt, samt vise refleksjoner over hvilken kunnskap bruk av andre metoder ville gitt, er også et grep for å styrke validiteten (Gleiss & Sæther, 2022, s. 206).

Reliabilitet i kvantitativ forskning handler om hvor *pålitelige* målingene er. Høy reliabilitet betyr at målingene inneholder små målefeil, og uavhengige målinger skal gi tilnærmet

identiske resultater. De ulike leddene i måleprosessen må være fri for unøyaktigheter (Halvorsen, 2008, s. 68). Innenfor en positivistisk tradisjon er det vanlig å vurdere reliabiliteten opp mot kvaliteten på innsamlingen av data, og om forskningsresultatene kan reproduseres av andre forskere. Det er et mål å etterstrebe å være så nøyaktig som mulig, samtidig vil det ikke være mulig å fjerne bias helt. Bias i forskning handler om at resultater og slutninger i en studie er skjeve eller feilaktig. En kan bruke flere metoder for å redusere muligheten for slike skjevheter ved å bruke kombinerte metoder, datakilder eller la flere forskere kode datamaterialet sammen, det siste kalles intercoder-reliabilitet (Gleiss & Sæther, 2022, s. 203).

2.5.2 Interrater reliabilitet

En annen måte å måle reliabiliteten på er interrater reliabilitet. Høy reliabilitet er en forutsetning for høy validitet. For å teste reliabiliteten i observasjonssituasjoner kan to forskere foreta de samme målingene uavhengig av hverandre, dette kalles interrater reliabilitet. Vi kan også forstå dette som sensorreliabilitet. Begge begrepene refererer til graden av samsvar mellom ulike vurderere (sensorer) når de vurderer de samme oppgavene eller elevsvarene (Halvorsen, 2008, s. 68).

Cohen et. al. (2007) beskriver også interrater reliabilitet som en metode for å måle reliabilitet i måling av samme objekt. Det vises til flere faktorer som kan være en trussel for reliabiliteten i tester og eksamener. Det kan for eksempel gi seg utslag som feilføring i registrering av vurderingsresultater, eller at ulike sensorer vurderer den samme elevbesvarelsen ulikt. Det kan resultere i at vurderingen eleven får er avhengig av sensoren. Sensorer kan også befinne seg på ulike stadier i arbeid med et større vurderingsarbeid, særlig hvis sensoren har et høyt antall elevbesvarelser å sensurere. Forfatterne skriver at sensorer kan være strenge i starten av vurderingsarbeidet, og mildere i senere stadier av vurderingsarbeidet. Da kan det variere om sensoren lar elevsvaret når opp til den beste vurderingen, eller velger å sette elevsvaret ned til kategorien under. Videre beskriver Cohen et. al. *The halo effect*. Dette beskrives som en kognitiv skjevhet, der elever i kraft av sine personligheter påvirker vurderingen av andre egenskaper og kunnskaper, for eksempel på prøver (Cohen, Manion, & Morrison, 2007, s. 159).

2.5.3 Validitet og reliabilitet i kvalitativ forskning

Innenfor en sosialkonstruktivistisk tradisjon er begrepene validitet og reliabilitet tilpasset andre måter å tenke om forskningskvaliteten på. Noen forskere innenfor den

sosialkonstruktivistiske tradisjonen velger å bruke begreper som *refleksivitet* for å beskrive reliabilitet, og *troverdighet* for validitet. Dette for å synliggjøre at de har andre kriterier for å beskrive forskningskvalitet (Gleiss & Sæther, 2022, s. 202). Gleiss og Sæther skriver at refleksjon er et grep for å styrke validiteten i kvalitative forskningsprosjekter. Åpenhet rundt studiens styrker og begrensninger, refleksjon over egne fortolkninger og posisjonalitet, vil styrke validiteten (Gleiss & Sæther, 2022, s. 206).

Gleiss og Sæther (2022) løfter fram refleksivitet som et viktig kriterium for god forskning innen en sosialkonstruktivistisk tradisjon. Det innebærer å beskrive, begrunne og reflektere over forskningsprosessen. Videre vil det være viktig å være åpen om retningsskifter, og også fortelle om utfordringer som oppsto underveis i arbeidet. Slik kan en bidra til å styrke reliabiliteten i et sosialkonstruktivistisk perspektiv. Forfatterne påpeker også at repliserbarhet er noe mer utfordrende i kvalitative forskningsprosjekter, ettersom forskerens egen posisjonalitet spiller en sentral del av datainnsamling og analyse. Isteden ligger fokuset på at forskningsprosjektet skal være så gjennomsiktig som mulig, slik at andre kan vurdere forskningsprosessen. Dette omtales også som transparens (Gleiss & Sæther, 2022, s. 204).

2.5.4 Validitet og reliabilitet i prøver og tester

Cohen, Manion og Morrison (2007) redegjør for validitet og reliabilitet i prøver og tester i fagboken «Reasearch methods in Education». De legger vekt på at det er mange faktorer som kan påvirke reliabiliteten i elevenes prøvesituasjon. Elevenes motivasjon for å gjennomføre prøven og i hvilken grad eleven ønsker å prestere er en vesentlig faktor. De problematiserer at motivasjon ofte avhenger av om resultatet er av betydning for eleven. Elevenes motivasjon for å delta i en prøve er ofte sterkest dersom elevene forstår formålet med prøven, og at lærerne er støttende i prøvesituasjon (Cohen, Manion, & Morrison, 2007, s. 160). Dersom elevene ikke er motiverte eller mestrer å svare på oppgavene er det en sjanse for at de gjetter på svaralternativer eller avstår fra å svare.

Cohen et. al påpeker også både fysiske og psykiske rammefaktorene rundt elevene i prøvesituasjonen er viktig. Prøvesituasjonen bør fortrinnsvis skje i klasserommet de vanligvis bruker med kjente lærere rundt seg. Videre kan andre reliabilitetsutfordringer være det datatekniske i gjennomføring av prøvene, kjente eller ukjente oppgavetyper, prøvens lengde og omfang (Cohen, Manion, & Morrison, 2007, s. 159).

3 Metode

Ifølge Knut Halvorsen (2008) er metode en systematisk måte å undersøke virkeligheten på, og er snevert definert som den håndverksmessige siden av vitenskapelig virksomhet (Halvorsen, 2008, s. 20). I det følgende vil vi redegjøre for vår vitenskapsteoretiske forankring som begrunner hvordan dette masterprosjektet plasserer seg innenfor vitenskapene. Deretter redegjør vi for metodevalgene, utvalg og datainnsamling vi bruker for å besvare våre forskningsspørsmål. Videre følger en drøfting av masterprosjektets validitet og reliabilitet. Avslutningsvis vil vi redegjøre for hvordan vi håndterte etiske problemstillinger i prosjektet.

3.1 Vitenskapsteoretisk forankring

Vår masteroppgave har et samfunnsvitenskapelig perspektiv i vår vitenskapsteoretiske tilnærming. Gleiss og Sæther beskriver at forskere har «(…) ulike mål med forskningen og tar utgangspunkt i forskjellige vitenskapsteoretiske antakelser om hvordan verden henger sammen (…) (Gleiss & Sæther, 2022, s. 201)». Videre er det vanlig å skille mellom en positivistisk vitenskapstradisjon og en sosialkonstruktivistisk tradisjon. Gleiss og Sæther viser til at utdanningsforskning ofte posisjonerer seg innenfor et vitenskapsteoretisk spekter, som strekker seg mellom en positivistisk og en sosialkonstruktivistisk forskningstradisjon (Gleiss & Sæther, 2022, s. 202).

Ifølge Halvorsen er det i den positivistiske tradisjonen vanlig å vektlegge objektivitet og kvantitative metoder. De som identifiserer seg med et naturvitenskapelig syn, blir ofte kalt positivister. Idealet innenfor denne tradisjonen er nøytral og verdifri forskning (Halvorsen, 2008, s. 23). Gleiss og Sæther legger vekt på at muligheten for reproduksjon av forskningsresultater er et kjennetegn på god forskning innenfor en positivistisk tradisjon (Gleiss & Sæther, 2022, s. 204). Videre peker Halvorsen på at kritikere av den positivistiske vitenskapstradisjonen vil hevde at i motsetning til naturlovene, finnes det ingen uforanderlige lover om samfunnet og menneskelig atferd. Samfunnsforhold kan ikke studeres på samme måte som naturvitenskapene, og sosiale fenomener kan ikke «tingliggjøres» (Halvorsen, 2008, s. 23). Den sosialkonstruktivistiske tradisjonen, legger vekt på subjektivitet og konstruksjon av kunnskap gjennom sosiale prosesser (Gleiss & Sæther, 2022, ss. 201-207).

Vår masteroppgave vil bevege seg innenfor dette kontinuumet mellom en konstruktivistisk og positivistisk tradisjon. Dette forklares med at arbeid med nasjonale prøver i lesing og de åpne oppgavene spesielt, både har kvalitative og kvantitative aspekt ved seg. Det handler om vurdering av åpne oppgaver og sammenligning av resultater. På den ene siden vil en kvalitativ

fortolkning av elevens besvarelser svare til den sosialkonstruktivistiske vitenskapstradisjonen. På den andre siden vil opptelling og sammenligning av skåringsresultater ha positivistiske trekk.

Den faglige bakgrunnen vi har som norsklærere gjør at vi har erfaringer med å tilrettelegge, gjennomføre og skåre nasjonale prøver i lesing. I vår masteroppgave gjennomfører vi vurdering av elevtekster. Vi gjennomgår prøvetekster og vurderingsveiledere for å forstå og fortolke dette materialet. Vår profesjon vil forme vår forståelseshorisont og subjektivitet i forhold til det materialet som undersøkes. Med andre ord kan vi ikke betraktes som nøytrale observatører, men vi er styrt av våre forforståelser og språket som formidler noen verdier (Halvorsen, 2008, ss. 21-23).

Hermeneutikk dreier seg om å forstå og fortolke. «Det hermeneutiske tolkningsarbeidet tar sikte på å avdekke underliggende sammenhenger eller en dypere mening i handlinger, tekster, kunstverk, historiske kilder og lignende kulturuttrykk (Gilje, 2019, s. 11)». I lesing av faglitteratur veksler vi mellom å lese enkeltkilder (delene) og relatere kildene til hverandre for å forstå den akademiske samtalen (helheten). Denne vekslingen mellom del og helhet, kalles for den hermeneutiske sirkelen (Gleiss & Sæther, 2022, s. 69). Den hermeneutiske sirkelen står sentralt i teorien om hermeneutikk: I lesing av faglitteratur veksler vi mellom å lese enkeltkilder (delene) og relatere kildene til hverandre for å forstå den akademiske samtalen (helheten). Denne vekslingen mellom del og helhet, kalles for den hermeneutiske sirkelen (Gleiss & Sæther, 2022, s. 69). Teorien om hermeneutikken vil stå sentralt i flere deler av masterprosjektet; først i arbeidet med tekstgjennomgang av *Stargate* og *Snikfotografen*, deretter i arbeidet med tolkning og vurdering av elevbesvarelser sett opp mot kriteriene for skåring i vurderingsveiledere. Videre arbeidet vi fram en omforent skår i profesjonsfellesskapet vi utgjorde. Disse elementene kan vi forstå som deler og helhet (jf. Gilje, s. 74).

3.2 Metode og begrunnelse for valg av metode

I vårt masterprosjekt har vi valgt å kombinere en kvalitativ og en kvantitativ tilnærming for å besvare våre forskningsspørsmål. For å besvare det første forskningsspørsmålet har vi valgt en kvalitativ tilnærming. Dette begrunnes med at fleksibilitet og åpenhet er viktige styrker ved et slikt metodevalg (Gleiss & Sæther, 2022, s. 30). Den kvalitative delen av vår analyse vil være preget av en hermeneutisk tilnærming, som tillater en åpen og refleksiv utforsking av tekstene med lav grad av forhåndsstrukturering.

Tekstgjennomgangen av *Stargate* og *Snikfotografen* med vurderingsveiledere var viktig for vårt prosjekt. Vi så på tekstlige strukturer, tolkningsrom og særtrekk ved tekstene. Lesing slik det er operasjonalisert i rammeverket måles via utvalget av tekster i prøven. I tekstgjennomgangen undersøker vi validitetsaspektet til de åpne oppgavene gjennom å analysere sammenhengen mellom tekst, oppgave og vurderingskriterier. Hvilke muligheter for måling av lesekompetanse rommer de åpne oppgavene?

Til å besvare det andre forskningsspørsmålet, valgte vi en kvantitativ innholdsanalyse som metode. I følge Gleiss og Sæther er innholdsanalyser ikke en klart definert analysemetode. De skriver at "Innholdsanalysen gir forskerne stor frihet til å velge relevante teoretiske begreper basert på forskningslitteraturen, som kan hjelpe forskeren til å analysere og fortolke materialet. Videre legges det vekt på at innholdsanalyse er en aktuell metode hvis man ønsker å si noe om et større felt, eksempelvis eksamensoppgaver (Gleiss og Sæther 2022, s.137). I vårt masterprosjekt har vi avgrenset oss til to av de åpne oppgavene med tilhørende elevsvar fra 8. og 9. trinn. Årsaken til at vi ønsket å se på begge trinn er at de får den samme prøven. Vi undret oss også over om vi kunne se forskjeller i elevsvarene mellom de ulike trinnene.

Gleiss og Sæther (2022) skriver at det kan være aktuelt å bruke forhåndsdefinerte kategorier i arbeid med et større datasett som tekstene kan kodes ut fra. Kategorisering gjør det mulig å klassifisere systematisk, kvantifisere dataene og identifisere mønstre (Gleiss & Sæther 2022, s. 139). Vi vurderte det som en god fremgangsmåte for å håndtere det store antallet elevsvar fra de åpne oppgavene, slik at vi kunne se om vi fant noen mønstre i elevsvarene.

Halvorsen (2018) omtaler at kvalitative undersøkelser kan brukes som forberedelser til kvantitative undersøkelser, gjerne gjennom forprosjekter (Halvorsen, 2018, s. 149). I vårt prosjekt utviklet vi kategorier til innholdsanalysen av elevsvar basert på tidligere arbeid og erfaring med et forhåndsmateriale. Forhåndsmaterialet besto av et utvalg elevsvar fra en lokal skole i en forberedende undersøkelsesfase. Vi fikk tilgang til å studere, drøfte og skåre

autentiske elevsvar på de åpne oppgavene fra høsten 2023. Dette var ikke en reell skåringssituasjon, men som en del av et forarbeid i en tidlig utviklingsfase av vårt prosjekt.

Ut fra vårt forprosjekt utviklet vi fem forhåndstrukturerte kategorier (A-E). Kategoriene favnet grupperinger av elevsvar opp mot kriteriene i vurderingsveilederen. Dette åpnet opp for opptelling og sammenligning av skåringsresultater og kategoriserte elevbesvarelser. Kategoriene fordelte seg slik; A – riktig svar i henhold til vurderingsveilederen, B – uriktig svar i henhold til vurderingsveilederen, C – tvilstilfelle (svar som er vanskelig å skåre), D – Elevuttrykk (svar som kommenterer at de ikke vet, ikke forstår og lignende), og E – blank besvarelse.

I neste fase undersøkte vi sensorreliabiliteten ved å skåre og kategorisere et utvalg elevbesvarelser hver for oss. I etterkant av individuell kategorisering og skåring, sammenlignet vi resultatene våre. Elevsvarene vi hadde vurdert ulikt, diskuterte vi oss nå frem til en omforent skår. Vi valgte å jobbe frem en omforent skår for å kvalitetssikre elevsvaret en så riktig vurdering som mulig. Denne prosessen ga oss innblikk i hvilke tilfeller skapte diskrepans, og hvorfor ulike vurderinger oppstod.

Deretter sammenlignet vi resultatene fra omforent skår med de faktiske resultatene elevene hadde fått på den nasjonale prøven. Vi omtaler lærere som har gitt sine vurderinger som *Lærer x*. Denne betegnelsen representerer alle lærerne som har utført vurdering og skåring i vårt datasett. Dette ble valgt som metode i et forsøk på å kunne si noe om påliteligheten i vurderingsarbeidet av de åpne oppgavene.

I arbeidet med skåring og kategorisering brukte vi Excel. Videre brukte vi flere funksjoner i dataprogrammet til å telle opp og utføre flere utregninger av data, blant annet gjennom Pivottabeller og andre søkefunksjoner. I analyse av kvantitative skåringsdata ble IMB SPSS *Statistics* versjon 29 anvendt. Programmet ble brukt til statistisk sammenligning av skåringsresultater, av Cohens kappa og Fleiss multirater kappa.

Cohens kappa er et kjent mål på enighet mellom to vurderere (raters) – interrater reliabilitet. Det er generelt antatt å være et mer robust mål på interrater reliabilitet enn prosentvise sammenligninger. Dette begrunnes med at Cohens kappa tar hensyn til sannsynligheten for at overensstemmelse i vurdering eller skåring, oppstår ved tilfeldigheter. Som de fleste korrelasjonsstatistikker kan kappa variere fra -1 til 1. Kappaverdier symboliseres med den greske bokstaven, κ (McHugh, 2012; Landis & Koch, 1977).

McHugh (2012) beskriver at 0 representerer mengden samsvar som kan forventes fra tilfeldig tilfeldighet, og 1 betyr perfekt samsvar mellom vurderere. Kappaverdier under 0 er mulig, men kan anses som usannsynlig i praksis. Cohens kappa kan tolkes likt på tvers av studier og er en standardisert verdi (McHugh, 2012). Fleiss kappa er basert på Cohens kappa, og er et statistisk mål som vurderer overenstemmelse mellom tre eller flere vurderere (McHugh, 2012; Koch & Landis, 1977 s. 163). Tengberg og Skar (2016) skriver at det finnes flere standarder, ofte kalt benchmarks, for hva som regnes som akseptabel grad av enighet i skåring. De viser til at konsistensindikatorer, som korrelasjonskoeffisienter som varierer fra -1,0 til 1,0, helst bør være over ,70. Det er fordi dette regnes som det laveste akseptable nivået av enighet mellom vurderere. Samtidig påpeker de at målet på ,70 i noen kontekster ikke alltid er tilstrekkelig, særlig når testresultater har store konsekvenser for elevene. I noen tilfeller anser man verdier omkring ,60 og ,80 som akseptable, og verdier over ,80 som høye (Tengberg & Skar, 2016, s. 6). McHugh (2012) skriver i sin artikkel «Interrater reliability: The kappa statistics» at tolkning av kappa gjøres ut fra referanseverdier. Koch og Landis (1977) sine referanseverdier er også mye brukt. I vårt masterprosjekt benytter vi modellen fra McHugh (2012), se bilde under:

Tolkning av Cohens kappa.

Verdien av Kappa	Avtalenivå	% av data som er pålitelige
0–.20	Ingen	0–4%
.21–.39	Minimal	4–15%
.40–.59	Svak	15–35%
.60–.79	Moderat	35–63%
.80–.90	Sterk	64–81%
ovenfor.90	Nesten perfekt	82–100%

Figur 2: Tolkning av Cohens Kappa (McHugh, 2012).

3.3 Datainnsamling og utvalg

Datamaterialet i masterprosjektet består av fire deler. Den første delen er et forhåndsmateriale fra en lokal skole. Den andre delen består av et tilfeldig utvalg av 4000 oppgavesvar fra Utdanningsdirektoratet. Den tredje delen består av dokumenter tilknyttet leseprøven høsten 2023. Den fjerde delen består av et mindre utvalg på 1800 elevsvar som ble skåret og kategorisert.

I en innledende fase høsten 2023 fikk vi tillatelse av en lokal skole til å studere elevbesvarelser fra de syv åpne oppgavene på NPL i PAS[4] før skåring. Denne dataen knyttet til forprosjektet utgjorde 1120 elevsvar. Vi fant materialet så spennende at vi bestemte oss for å søke Utdanningsdirektoratet om et tilfeldig utvalg elevsvar av de åpne oppgavene med poengsum/skår. Ved å søke Utdanningsdirektoratet om et tilfeldig utvalg elevsvar ivaretok vi flere hensyn. For det første ønsket vi å sikre anonymiteten til elevene i masterprosjektet. I tillegg ønsket vi at empirien skulle være representativ for populasjonen. Populasjonen var elever på 8. trinn og 9.trinn trinn som hadde deltatt på nasjonale prøver i lesing høsten 2023.

Vi fikk tilgang på 4000 elevbesvarelser fra NPL høsten 2023, fordelt på to Excel-filer. Heretter omtalt som opprinnelig utvalg. Filene inneholdt elevresultater for hele prøvesettet for elever på 8. og 9. trinn, fordelt på 1600 bokmålselever og 400 nynorskelever. 2000 elevresultater per datasett. Kjønnsfordelingen i datasettene var rundt 50/50. Datasettene inkluderte variabler som testvarighet, sum skår per elev og maksimal skår for hele prøven, som var 43 poeng. Videre inkluderte datamaterialet tekster tilknyttet leseprøven 2023. Disse var tilgjengelig gjennom nettsiden «PAS prøver» og Utdanningsdirektoratets hjemmesider. Tekstmaterialet omfattet rammeverket, prøvetekstene, oppgavetekstene med vurderingsveiledere.

Leseprøven for 2023 er tilgjengelig i PAS for ansatte i skolesektoren gjennom Feide-innlogging og et kodeord[4]. Som ansatte i skolen hadde vi tillatelse til å logge inn og studere prøven slik den fremsto for elevene i prøvesituasjonen. Vi var oppmerksomme på at prøvetekstene presenteres for elevene i tilfeldig rekkefølge gjennom det digitale

[4] Kodeordet er tilgjengelig for ansatte i skolen. Sensorkommisjon kan be om kodeordet gjennom veileder Morten Bartnæs. Da får man tilgang til prøven slik den ble fremsatt for elevene høsten 2023.

prøvesystemet. Oppgavene knyttet til prøvetekstene kommer alltid i samme kronologiske rekkefølge. Vi fikk skriftlig tillatelse fra Utdanningsdirektoratet til å bruke og gjengi prøvetekstene og vurderingsveilederne i dette masterprosjektet 5. mars 2024.

I arbeidet med skåring av de åpne oppgavene, vurderte vi et mindre utvalg elevbesvarelser. Vi gjorde beregninger som viste fordelingene av poengsummene i det opprinnelige datasettet på 4000. Pivottabellene viste at *Stargate*, *Snikfotografen* og *Skulearbeid* representerte de åpne oppgavene elevene på både 8. og 9. trinn hadde fått minst uttelling for på prøven.

Ut fra disse funnene studerte vi et tilfeldig utvalg på 300 elevsvar fra de åpne oppgavene til *Stargate*, *Snikfotografen* og *Skulearbeid*, for 8. og 9. trinn, totalt 1800 elevsvar. Begrunnelsen for at vi ikke skåret flere enn 1800 elevsvar, var at vi kom fram til et metningspunkt i å utforske materialet. *Skulearbeid* ble tatt ut av masterprosjektet i februar 2024 grunnet tids- og plasshensyn. Utvalget til gjennomgang i masterprosjektet består derfor av 1200 elevsvar fra *Stargate* og *Snikfotografen*. I skåringsarbeidet kodet vi materialet som skårer 1 Cecilie og skårer 2 Christina. Forkortelser for disse navnene kan ses som S1 og S1 i det kvantitative datamaterialet.

Vår utdanning og yrkesliv som norsklærere, samt tidligere erfaringer knyttet til arbeid med nasjonale prøver i lesing er også en del av datamaterialet. Kunnskap som vi har med oss i vurdering og skåring, kan vi ikke kan legge fra oss, det er med i form av vår profesjon. Dette kan også føre til fordommer i vårt materiale. Vi har som formål å søke etter en dypere forståelse av validitet og reliabilitet i de åpne oppgavene. Dette kan stå i motsetning til mange lærere som kanskje opplever skåring av åpne oppgaver som en pliktoppgave.

3.4 Studiens kvalitet

I dette avsnittet redegjør vi for masterprosjektets validitet og reliabilitet, og deretter studiens grad av overførbarhet og generaliserbarhet. Vi redegjør for vår egen rolle i prosjektet og drøfter hvilke styrker og svakheter det har.

3.4.1 Validitet og reliabilitet

Validitet handler om hvorvidt datamaterialet kan gi svar på våre forskningsspørsmål. I tekstutvalget har vi forsøkt å velge tekster som kan utdype og besvare forskningsspørsmålene. Ifølge Thrane (2018) handler validitet om i hvilken grad man evner å måle det teoretiske begrepet man prøver å måle (Thrane, 2018, s. 47). Begrepsvaliditeten i vårt prosjekt er allerede operasjonalisert gjennom rammeverket til de nasjonale prøvene i lesing. Vårt masterprosjekt undersøker i hvilken grad de åpne oppgavene i NPL, gjennom tekstene *Stargate* og *Snikfotografen*, måler elevenes lesekompetanse som beskrevet i rammeverket.

I den første delen av masterprosjektet, som omhandler de to kvalitative tekstgjennomgangene, vil validitet og reliabilitet omtales som grad av troverdighet og refleksivitet (Gleiss & Sæther, 2022, s. 202). I tekstgjennomgangen studerte vi tekstene *Stargate* og *Snikfotografen* for å kunne beskrive og forklare hvilken lesekompetanse som kreves av elevene for å forstå prøvetekstene. I dette er det viktig for oss å påpeke at vi ønsker å være så åpne om våre tolkninger som mulig, men samtidig være klar over at andre kan tolke de samme tekstene på en annen måte enn oss. Det kan være en svakhet at vi ikke hadde forhåndsstrukturert kriterier for tekstgjennomgangene. Valg av tekster til gjennomgang kan ha påvirket validiteten. Dersom vi hadde fått analysert alle prøvetekstene og sett disse opp mot hverandre, er det en mulighet for at vi kunne fått andre svar. Likevel, mener vi at de valgte prøvetekstene var mest interessante for å besvare forskningsspørsmålene våre.

For å styrke vår troverdighet i arbeidet med tekstgjennomgangene, har målet vært å være så transparent som mulig. Våre tekstgjennomganger har spor av vår subjektivitet og vi har forsøkt å styrke denne ved god refleksivitet. Det vil si at redegjør og begrunner påstander og funn i analysene. I tillegg vil vi etterstrebe å være åpen på at egne verdier og interesser påvirker prosessen (Gleiss & Sæther, 2022, s. 203). Vi tror at andre forskere som måtte ønske å etterprøve våre tekstanalyser av *Stargate* og *Snikfotografen* vil kunne komme fram til en relativ lik forståelse av tekstene. Vi har vært to deltakere med lik utdanning i dette prosjektet, i tillegg har vi tilsvarende lik ansiennitet fra arbeid i skolen med norskfaget som praksisfelt.

Gjennom hele prosessen har vi diskutert hverandres tolkninger av tekstene for å forsøke å belyse alle relevante sider. Vi mener dette styrker troverdigheten til masterprosjektet.

Thrane (2018) beskriver at ekstern validitet handler om studiens konklusjoner kan generaliseres til andre personer og situasjoner. Dersom en studie er basert på et representativt utvalg gjennom tilfeldig trekning, kan en si noe ganske presist om en hel populasjon. Han viser også til at Surveybasert forskning har høy ekstern validitet (Thrane, 2018, s. 170). Empirien i denne masteroppgaven baserer seg på et tilfeldig utvalg elevdata med skåringsresultater fra nasjonale prøver i lesing høsten 2023. Alle landets elever på 8. og 9. trinn skal gjennomføre nasjonale prøver i lesing. Utdanningsdirektoratet var behjelpelig med å trekke et tilfeldig utvalg elevsvar til oss. Det vil si at hele populasjonen hadde lik sjanse til å bli trukket ut (Halvorsen, 2008, s. 155), vi argumenterer for at våre data er representative. Vi mener den eksterne validiteten til masterprosjektet er styrket gjennom vårt datamateriale. Våre kvantitative funn representert gjennom Cohens kappa og Fleiss kappa, kapittel 4, har likheter med resultater fra tidligere studier som viser kappaverdier på (,73) som viser variasjon sensorreliabilitet (Tengberg & Skar 2016; Tengberg et al, 2018). Masterprosjektets kvalitative funn kan være overførbare og vise tendenser når det gjelder elevsvar fra åpne oppgaver fra *Stargate* og *Snikfotografen*, høsten 2023.

3.4.2 Vår rolle som skårere i masterprosjektet

I andre del utførte vi en kvantitativ innholdsanalyse hvor vi undersøkte elevsvar opp mot kriteriene i vurderingsveilederen. I forskning vil høy validitet innebære at man måler det man ønsker å måle (Gleiss & Sæther, 2022, s. 205). Vi ønsket å undersøke hvor presise, dekkende og brukbare vurderingskriteriene for skåring var, og i hvilken grad skårernes vurderingspraksis svarte til disse kriteriene. Ved å studere og skåre et mindre utvalg på 1800 autentiske elevsvar, hadde vi mulighet til å undersøke sensorreliabiliteten i materialet. For å realisere dette, var det nødvendig med lærere som kunne vurdere elevsvarene, slik at resultatene ble sammenlignbare. Da vi fikk utlevert datamaterialet fra Utdanningsdirektoratet skrev vi under på en taushetserklæring. Dette låste muligheten for at andre lærere kunne få tilgang til elevsvarene i vårt datasett. Taushetserklæringen utelukket derfor muligheten for eksterne lærere som skårere i vårt prosjekt. Vi valgte derfor selv å vurdere og skåre 300 elevsvar for hver av tekstene *Stargate*, *Snikfotografen* og *Skulearbeid*, både for 8. og 9. trinn.

For å sikre reliabiliteten i skåring og kategorisering av elevsvarene, arbeidet vi hver for oss. Vi innførte en praksis hvor vi skjulte hverandres vurderinger i dataprogrammet Excel. Elevsvarene til 8. og 9. trinn ble fordelt i to separate Excel-filer. For å unngå å bli påvirket av hverandres vurderinger, arbeidet vi i motsatte Excel- filer om gangen. Når vi byttet fil, sørget vi for å skjule de allerede tildelte skåringsresultatene for å bevare uavhengigheten i skåring. Skåringen ble utført individuelt på våre respektive arbeidsplasser. Videre var de autentiske skåringsresultatene som elevene hadde fått av lærer x, skjult ved hjelp av samme funksjon i Excel. For å vurdere reliabiliteten i skåring, kombinerte vi våre respektive kolonner for å kunne sammenligne tildelte poeng. Totalt ble alle elevsvarene vurdert tre ganger. Den autentiske vurderingen fra lærer x, Skårer 1 Cecilie og Skårer 2 Christina. Til sist fikk alle elevsvarene også en siste vurdering gjennom en omforent skår. Da vi skulle sammenholde våre resultater la vi sammen kolonnene våre, slik at vi kunne sammenligne skåringen. Interrater reliabilitet i skåring ble regnet ut gjennom dataprogrammet SPSS. Det kan være en svakhet for studien at vi ikke hadde et bedre program enn Excel å skåre i. Vi hadde en teoretisk mulighet til å se hverandres skåringer underveis, dersom vi åpnet kolonnene med skåringsresultater. Selv om det var mulig, var vi hele tiden tro mot prosjektet og lot oss ikke friste til å lene oss på hverandre eller lærer x sin skåring underveis.

3.4.3 Faktorer som kan ha påvirket vårt masterprosjekt.

Vårt masterprosjekt handler om fordypning i de åpne oppgavene fra *Stargate* og *Snikfotografen*, har vi opparbeidet oss svært mye erfaring i møte med datamateriale. I masterprosjektet har vi valgt de åpne oppgavene til fordypning, og våre erfaringer og deltakelse gjennom prosjektet, mener vi er en styrke. Vi har nærlest og studert materialet grundig, noe vi ikke har kunnskaper om lærer x også har gjort. Videre kunne vi valgt å anonymisere oss selv i skrivingen, men dette kunne ført til en falsk distanse i teksten. Derfor velger vi å opptre som navngitte enkeltpersoner i møte med materialet.

Selv om vi tok forholdsregler i arbeidet med skåring og kategorisering, kan vi ha gjort feil. Vi kan ha gjort feil ved registrering av poeng og kategori, og vi kan ha gjort beregningsfeil ved bruk av SPSS. I alle prosesser der mennesker er involvert er det alltid en fare for å gjøre menneskelige feil i registrering av data. Dette kan påvirke reliabiliteten. På den andre siden var det en styrke å være to deltakere som har studert det samme materialet og kontrollert alle utregninger sammen jmf. intercoder reliabilitet (Gleiss & Sæther, 2022, s. 203). Vi samarbeidet med forhåndsmaterialet og ble etter hvert godt kjent med hverandres faglige

skjønn og argumentasjonsrekker i skåring. På den andre siden var arbeidet med forhåndsmaterialet knyttet til prøvetekstene *Tasmansk pungulv*, samt *Den korte historien om egget*. I tillegg studerte vi prøveteksten *Et skår i gleden*. Det skal nevnes at Christina også hadde ansvar for skåring av åpne oppgaver for egne elever, og skåret 310 åpne oppgaver gjennom sin funksjon som norsklærer i september 2023. I tillegg hadde vi studert og diskutert vurderingsveilederen flere ganger, men oppdaget også at vi måtte sette oss inn i vurderingsveilederen på nytt da vi skulle skåre datamaterialet utlevert fra Utdanningsdirektoratet i januar 2024.

I møte med datamaterialet i skåring var det viktig å være vår posisjon som lærere bevisst. Ettersom vi jobber som lærere har vi god kjennskap til prøvekonteksten med tanke på gjennomføring av nasjonale prøver i lesing, samt skåring. Vi holder det åpent at vi kan ha hatt forutinntatte tanker og holdninger knyttet til prøvene, i tillegg kjenner vi godt til elever som testdeltakere. Det er en styrke at vi i arbeid med dette materialet hadde avstand til elevene og lærerne som var uvitende deltakere i vårt masterprosjekt. Både elever og lærere er helt anonymisert, da denne informasjonen var skjult for oss i Utdanningsdirektoratets Excel-filer. Ettersom åpne oppgaver på nasjonale prøver i lesing, stort sett blir vurdert av elevenes lærere, kan målstyring være en trussel for reliabiliteten. Lærere *kan* tendere til å lettere "la tvilen komme eleven til gode", *The halo effect*, som omtalt over i 2.4.5. I vårt masterprosjekt var vi ikke under denne formen for påvirkning, da vi ikke hadde et forhold til materialet vi arbeidet med. Det skal også nevnes at vi arbeidet med materialet i en annen kontekst enn lærere flest. Vi hadde ingen nærhet til de vi har forsket på, noe vi mener er en faktor som øker reliabiliteten og generaliserbarheten i vår studie.

I beregningen av Cohens kappa og Fleiss kappa kan det være en svakhet at vi ikke hadde tre uavhengige vurderere som vi kunne sammenlignet med hverandre. Skårer 1 Cecilie og skårer 2 Christina er konstant, men sammenlignes med lærer x. Lærer x består som sagt av mange ulike lærere, og vi vet ikke hvor mange lærere det gjelder. Det kan være en svakhet i beregningen. Vi er bevisst på at lærer x utgjør et tverrsnitt av vurderere som vi sammenligner reliabiliteten opp mot, noe vår omforente skår også er. Ved bruk av Cohens og Fleiss kappa sammenlignes våre skåringer mot hver enkelt lærers vurdering av elevsvaret. Det vil si at hvert elevsvar får tre individuelle vurderinger. Dog, utgjør den omforente skåren et slags gjennomsnitt av skårer 1 og skårer 2. Ettersom vi har valgt de åpne oppgavene som interessefelt, kan vi bare anta at lærerne som skåret elevsvarene høsten 2023 var i en annen

situasjon enn oss i skåring. Det samme vil gjelde for Fleiss kappa beregningene. For beregning av Fleiss kappa har vi målt skårer 1 Cecilie, skårer 2 Christina og lærer x mot hverandre. Her kunne vi ikke bruke vår omforente skår. Det begrunnes med at omforent skår ikke kan sammenlignes med oss selv, da vi utgjør en halvpart av den omforente skåren. Dersom vi skulle hatt et mer uavhengig svar, hadde det mest ideelle vært at de utvalgte elevsvarene ble vurdert av et større antall uavhengige lærere. Dessverre var det ikke mulig å gjennomføre, grunnet rammebetingelsene i masterprosjektet og taushetsavtalen inngått med Utdanningsdirektoratet.

3.5 Forskningsetikk[5]

Den nasjonale og forskningsetiske komite for samfunnsvitenskap og humaniora (NESH) har ansvar for å utarbeide nasjonale forskningsetiske retningslinjer for sitt fagfelt. På deres nettsider[6] skrives det innledningsvis at forskning er en systematisk og kollektiv søken etter ny innsikt. Videre påpekes det at forskningsetikken bidrar til å fremme fri, god og forsvarlig forskning, samt sikre god vitenskapelig praksis (De nasjonale forskningsetiske komiteene, 2023, s. 5). Elevene som har gjennomført nasjonale prøver i lesing er under 15 år. Utdanningsdirektoratet påpeker på sine hjemmesider[7] at de nasjonale prøvene i utgangspunktet er obligatoriske, men skolene kan innvilge fritak til elever som har vedtak om spesialundervisning eller særskilt språkopplæring. Nasjonale prøver er en del av kvalitetsutviklingssystemet i den norske skolen. Det er åpnet for at både stat, skoleeier, skole, lærere, foresatte og elever kan få innsyn i resultater. Informert samtykke er samtidig et grunnprinsipp i all forskning (Gleiss & Sæther, 2022, s. 44). Ettersom nasjonale prøver er obligatoriske prøver for alle landets elever på 8. og 9. trinn og et av formålet med prøvene er kvalitetsutvikling, faller prinsippet om informert samtykke bort.

Formålet med masterprosjektet var å studere elevbesvarelser på de åpne oppgavene, derfor hadde vi ikke behov for å samle inn personopplysninger. Likevel var det viktig å sørge for at all datahåndtering fulgte Universitetet i Tromsøs etablerte prinsipper og retningslinjer for sikker forvaltning av forskningsdata (Universitetet i Tromsø, 2021). Dette tiltaket var essensielt for å beskytte eventuelle sensitive opplysninger som kunne være til stede i datasettet. Vi forventet ikke å finne personsensitive data i elevbesvarelsene, men kunne ikke utelukke denne muligheten. Elevene står i prinsippet fritt til å skrive hva de selv ønsker som svar til de åpne oppgavene. Derfor undertegnet vi en taushetserklæring utstedt av Utdanningsdirektoratet før vi fikk tilgang til forskningsmaterialet. Dersom vi kom over sensitiv informasjon, var prosedyren å ekskludere denne informasjonen fra studien og slette

[5] Deler av underkapittelet «Forskningsetikk» har likheter med et arbeidskrav som vi leverte i forbindelse med emnet LER-3501 Metoden, UiT. Levert 1.12.2023 i Canvas.

[6] www.forskningsetikk.no

[7] https://www.udir.no/eksamen-og-prover/prover/nasjonale-prover/administrere-nasjonale-prover2/#a124941 sist lest 02.04.2024.

personopplysningene. På grunn av våre forholdsregler var det ikke nødvendig å melde prosjektet inn til Sikt. Vi fant ingen slike tilfeller i vårt utvalg.

Forskere skal ikke påvirkes av sitt eget syn, men det er alltid en fare for å påvirke forskningsresultatene når en selv bidrar med empiri. I vår oppgave forstått som at vi bidrar med skåringer av elevsvar. Vi reflekterer over at vi har god kjennskap til å forstå konteksten og sammenhengen elever og lærere befinner seg i, i arbeid med de nasjonale leseprøvene. Likevel skal vi ikke være forutinntatt i møte med datamaterialet. Vi har reflektert rundt våre egne holdninger og forsøkt å stille oss nøytral og objektiv i møte med masterprosjektets empiri.

Fra vårt profesjonsfaglige ståsted som lærere savner vi prioritering og tilrettelegging for formativ bruk av resultatene på de nasjonale prøvene i lesing. I dag er det i stor grad lagt opp til den enkelte skole og lærer hvordan en skal jobbe videre med resultatene for å gi formativ og læringsstøttene vurdering. Vi har blitt oppmerksom på noen begrensninger i hvordan de ulike lesekompetansene måles. Erfaringsvis vet vi at det er mange flere elever som evner å reflektere og trekke slutninger fra tekst i større grad enn de får vist i de nasjonale prøvene i lesing. Resultatene fra dette prosjektet kan kanskje spille inn i en aktuell debatt om videre utvikling av de nasjonale prøvene i lesing, og kanskje spesielt utvikling av vurderingsveiledere og skåring av åpne oppgaver.

4 Presentasjon av data og funn

I det følgende vil vi redegjøre for to analyser: *Stargate* og *Snikfotografen*. De to analysene følger samme struktur. Kapitelet behandler presentasjon av tekstgjennomgang, data og funn. Først foretar vi tekstgjennomgang av prøvetekstene, deretter beskriver vi spørsmål elevene besvarte i forkant av den åpne oppgaven. Vi presenterer data og funn fra kategorisering og skåring, før vurderingssamtalene fram til omforent skår gjennomgås. Videre fremstiller vi kvantitative resultater fra skåring av åpne oppgaver for 8. og 9. trinn. Graden av samsvar i skåring presenteres gjennom statistiske beregninger av Cohens kappa og Fleiss kappa.

4.1 Stargate

4.1.1 Tekstgjennomgang *Stargate*

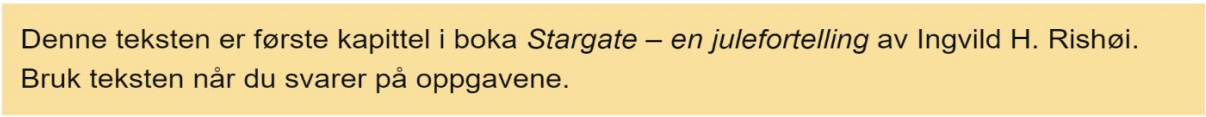
I de nasjonale prøvene i lesing for høsten 2023 er *Stargate*[8] den eneste skjønnlitterære teksten. I boken *God leseopplæring med nasjonale prøver* (2018) argumenterer Roe, Ryen og Weyergang for nødvendigheten av å inkludere minst én skjønnlitterær tekst i disse prøvene. Dette argumentet begrunnes i skjønnlitteraturens sentrale rolle i norskfaget.

Roe, Ryen og Weyergang (2018) trekker frem at skjønnlitterære tekster utfordrer lesernes evne til å tolke og forstå innhold som ikke alltid er eksplisitt uttrykt. Tekstene åpner for et bredt og variert tolkningsrom. *Stargate* er en tekst som krever symbolsk forståelse og tolkning av innhold som ikke er eksplisitt uttrykt. Et eksempel på en slik symbolsk forståelse er når vaktmesteren tar ned hele plakaten og gir den til jenta. Her må leseren selv slutte seg til at dette kan hindre andre i å få jobben. Oppgaver knyttet til skjønnlitterære tekster er derfor godt egnet til å måle elevens evne til å tolke og trekke slutninger fra implisitt informasjon. Elever som oppnår to poeng, kan hevdes å være nærmere en tenkt modell-lesing av teksten (Frønes & Ryen, 2020, s. 80)

I arbeidet med å forstå en tekst kan vi ikke la hvilken som helst assosiasjon styre lesingen, da havner vi en privat sfære. For at en tolkning ikke bare skal være knyttet til den private sfæren, er det nyttig å støtte seg til Eco sitt modell-leserbegrep. I skolen må vi prøve å nærme oss en modell-leserens posisjon i teksten, for samtale og refleksjon rundt innhold og menig i tekst

[8] *Stargate* er digitalt tilgjengelig på Nasjonalbibliotekets hjemmeside: https://urn.nb.no/URN:NBN:no-nb_pliktmonografi_000010733

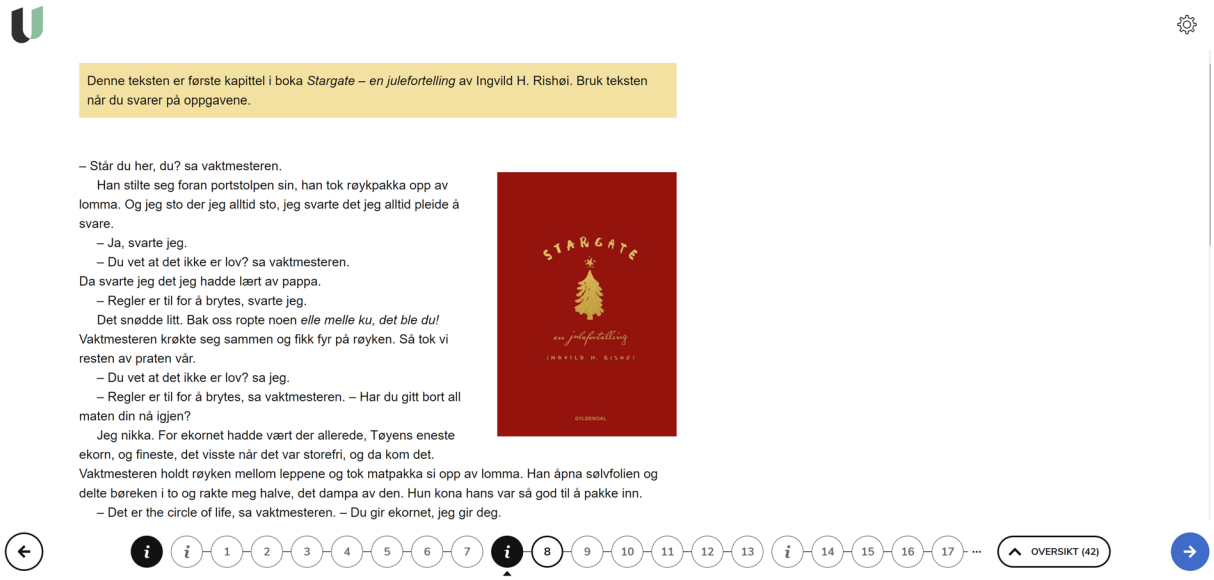
(Roe, Ryen, & Weyergang, 2018, ss. 79-80). Vi reflekterer over rommet mellom de individuelle tolkningene som de skjønnlitterære tekstene åpner for, samtidig må vi enes om hovedinnhold og tematikk. I den videre tekstgjennomgangen vil vi komme inn på mulige tolkninger knyttet til *Stargate*.



Figur 3: Informasjon til elev før tekst - Stargate

I utdraget fra *Stargate* presenteres elevene for en tekstboks, men ingen overskrift, som vist over. Tekstboksen forbereder leserne på at dette er en skjønnlitterær tekst. Et bilde av bokens forside med rød bakgrunn, tittel i gull og et juletre, gir leseren en antydning om at fortellingen er knyttet til julen. Dette kan aktivere forkunnskaper og skape noen forventinger hos leseren i møte med teksten. *Stargate* har også fellestrekk med julefortellinger som Piken med svovelstikkene av H.C. Andersen. I prøveteksten kan vi lese ut tematikken sosiale forskjeller, jul og barn i sårbare situasjoner.

På bildet av bokens forside er ordet *Stargate* skrevet i en typografi som kan minne om håndskrift. Utdraget som er brukt i prøven, er tekstidentisk med første kapittel i boken. Prøveutdraget skiller seg fra originalteksten ved at den sistnevnte har et lite forord som er skrevet av forfatteren. Forordet trekker leseren inn i fortellingen, denne mangler i prøveteksten (Rishøi, 2021). Elevene har bare forsidebilde og teksten til å skape mening og forstå konteksten.



Figur 4:Stargate som vist i prøvesituasjon

Elevene leser digitalt under de nasjonale prøvene i lesing. Teksten var plassert til venstre på skjermen, og spørsmålene til høyre. Elevene har mulighet skrolle opp og ned i teksten, samt navigere fram og tilbake i prøvesettet mens de leser. Teksten begynner in medias res, med en dialog mellom jenta og vaktmesteren. Det gis vage holdepunkter om sted; «ved en portstolpe». Dialogen viser at jenta befinner seg et sted hvor hun ikke har lov til å være, og det beskrives iterativt; «hun står der hun alltid står». Innledningen åpner altså for flere spørsmål; Hvor er jenta? Hva er det hun gjør som ikke er lov? Teksten gir implisitte hint om at de befinner seg i nærheten av en skolegård. Frasen; «Bak oss ropte noen *elle melle ku, det ble du!*» kan tolkes som lyden av barn som leker og sier regler. Vi tror elever med forkunnskaper og erfaring fra lek med regler vil fylle inn dette tomrommet. Ifølge Roe og Blikstad-Balas appellere skjønnlitterære tekst i større grad til fantasien, og tolkningen overlates mer til leseren enn hos fagtekster. Den sosiale og kulturelle bakgrunnen har også innvirkning på tolkningen (Roe & Blikstad-Balas, 2022, s. 62).

Prøveteksten starter in medias res og innledes av en førstepersonforteller som skildrer det som skjer. Her er fortelleren en del av historien. Fokaliseringen skifter mellom å være intern og ekstern. Fokalisering handler om perspektivet leseren får presentert handlinger gjennom (Gaasland, 1999, ss. 28-31). Det observeres utenfra som å se en film, samtidig som jeg-personens tanker kommer fram «Da svarte jeg det jeg hadde lært av pappa». Leseren må orientere seg under lesingen når det gjelder hvem som sier hva i dialogene. I teksten går det ikke fram at jeg-personen er en jente. Elevene blir orientert om dette gjennom den første oppgaveformuleringen, som inneholder ordet "jenta". Et annet element som indikerer tid og

sted, er omtalen av at jenta har gitt bort maten sin til et ekorn. Ekornet vet når det er storefri, noe som ytterligere antyder at handlingen foregår ved en skole og i skoletiden: «Har du gitt bort all maten din nå igjen? Jeg nikka. For ekornet hadde allerede vært der, «Tøyens eneste ekorn, og fineste, det visste når det var storefri, og da kom det».

Lesere med geografisk kjennskap til Tøyen vil forstå at dette er et multikulturelt område i Oslo. Det er rimelig å anta at mange elever i resten av landet, ikke har denne samfunnsgeografiske kunnskapen. Teksten kan få et ukjent preg, fordi den forutsetter ulike elementer av kulturkunnskap. For elever uten denne kunnskapen kan en slik formen for innforståtthet være til hinder for den basale tekstforståelsen, og dermed påvirke deres mulighet til å vise leseferdigheter på høyere nivå. Vaktmesterens matpakke inneholder børek, dette er ikke en spesielt velkjent matrett. Elever som har kjennskap til børek kan komme til å assosiere Tyrkia som vaktmesterens hjemland. Jenta gir maten sin til ekornet, og han deler av sin mat med henne. Vaktmesteren referer til «the circle of life», og i teksten omtaler han seg selv som en «stor tenker». Elevene som har støtte i sin forkunnskap, kan forstå at store tenkere ofte assosieres med filosofer. Vi reflekterer over hvorvidt eleven trenger å fylle inn alle tomme rommene for å kunne besvare oppgavene, men det kan bidra til å bygge dypere forståelse av hvem jenta og vaktmesteren er. Deres handlinger og personligheter blir forstått bedre ved å tolke det som ikke står eksplisitt i teksten. I følge Ryen og Frønes legger de skjønnlitterære tekstene opp til at leseren selv skal prøve å forstå hva som foregår, og hvorfor karakterene gjør som de gjør. Leseren gjør det med bakgrunn i kunnskap om verden og gjennom forestillingen som de bygger opp under lesingen (Ryen & Frønes, 2020, s. 149).

Begge karakterene i teksten utfordrer grenser. Jenta utfordrer grenser ved å være utenfor skolegården. Vaktmesteren, på sin side, tenner seg en røyk og tøyer sannheten for en god sak. Vaktmesteren forteller at han hadde en annen jobb i hjemlandet sitt og snakker om fordelene ved å komme til Norge. Han peker på en plakat som er oppslag for jobb som juletreselger. Gjennom dialogen blir det klart at vaktmesteren ønsker at jenta skal vise lappen med telefonnummeret til faren sin. Først sier han:

- Kanskje det kunne ha vært noe? sa vaktmesteren.
- Jeg tror ikke man får jobb når man er ti år, sa jeg.
- Jeg tenkte ikke på deg, sa vaktmesteren.

Han gikk bort til stolpen og rev av en lapp og kom tilbake og la den i hånda mi.

- Vis den til faren din, sa han.

For å forstå dialogen mellom vaktmesteren og jenta må elevene tolke at vaktmesteren ser for seg at jentas far kan søke på jobben. Når vaktmesteren foreslår at jenta skal vise plakaten til faren sin, er det en implisitt oppfordring om at han bør søke på jobben. Videre sier vaktmesteren til jenta at hun skal be sin far, og si at han kjenner Alfred. Dette kan leseren forstå som en referanse, som igjen kan øke sjansen for at faren får jobben. I den avsluttende delen av dialogen sier vaktmesteren noe som er avgjørende for forståelsen av teksten: «Du kan like gjerne ta med hele». Deretter beskrives det at han tar ned plakaten, ruller den sammen og gir den til jenta. Jenta responderer med: «(…) hva hvis noen andre vil søke jobben»? Siste setning i utdraget er: «Nettopp, sa han. -Du står her og ser på en stor tenker». Vaktmesteren har en spøkefull tone når han omtaler seg selv på denne måten. Elevene må selv forstå at vaktmesteren ved å hindre andre å se plakaten øker sjansen for at jentas far får jobben. Plakater som oppslag på butikker og lyktestolper er noe vi ser sjeldnere i dag enn tidligere. Det grunn til å tenke at mange elever ikke ser for seg denne sjangeren, da de er mer vant til å se digitale annonser. Utdraget inneholder ingen elementer som stikker seg ut tekstlig sett. Vi forventer at elevene leser utdraget lineært, linje for linje. Samtidig vil det være mulig å søke-lese for å lete etter svar på de aktuelle spørsmålene. Som lærere vet vi at elever ofte er gode til å lese tekst med motivasjon for å finne svar på spørsmål, blant annet fordi slike oppgaver ofte gis i skolen. En slik lesing av teksten vil kunne være en trussel for helhetsforståelsen av tekst. Dette støttes av Astrid Roe og Marte Bilkstad- Balas. Elever som bare søkeleser for å finne svar, risikerer at lesingen blir mekanisk informasjonsuthenting. Dette kan føre til at de ikke går dypere inn i teksten og mister dybdeforståelsen (Roe & Blikstad-Balas, 2022, s. 191).

4.1.2 Gjennomgang av spørsmålene til *Stargate*

Vi studerte alle spørsmålene til teksten, både den åpne oppgaven og flervalgsoppgavene. For Stargate kommer spørsmålene i hovedsak i samme rekkefølge som handlingen i teksten. Leseren må forstå at vaktmesteren ikke mener at jenta skal søke på jobben. Dersom eleven bare leser de ulike delene av teksten mekanisk for å søke svar på de enkelte spørsmålene adskilt, vil de kunne miste den helhetlige forståelsen. I noen av spørsmålene må eleven ha en helhetsforståelse for å løse oppgavene, mens andre bare krever at det hentes ut informasjon fra deler av teksten. Skaftun (2006) påpeker at en strategi hvor elevene kun henter ut enkeltelementer fra teksten, kan noen ganger være en styrke for eleven. Spørsmålsstillingen

og svaralternativene i flervalgsoppgavene kan lede elevene til å forholde seg til teksten på en ny måte.

Flervalgsoppgave nummer to til Stargate krever at eleven forholder seg til helheten i teksten og summen av ulike handlinger og replikker. Nettopp denne helhetsforståelsen gjør elevene i stand til å besvare denne. Oppgaven ber elevene ta stilling til om vaktmesteren er omsorgsfull, streng, snill, smart og masete ved avkryssingsboks for «passer» eller «passer ikke». Vi vurderer at eleven må lese hele teksten for å forstå disse personlige egenskapene til vaktmesteren.

Et annet moment som vi har reflektert over, er om elevene kan bli ledet av spørsmålene som stilles før den åpne oppgaven. Dette kan påvirke oppfattelsen eller tolkningen av teksten både negativt og positivt. På den ene siden kan det støtte elevene og hjelpe dem til å tolke teksten riktig. På den andre siden er det også en fare for at spørsmålsformulering og svaralternativer kan danne grunnlag for feiltolkninger. For eksempel vil spørsmål 1, svaralternativ 2 "Hun henger opp plakaten" kunne være med på å gjøre usikre elever enda mer usikker, noe som kan føre til at de trekker feilslutninger.

Oppgave 1. I begynnelsen av teksten gjør jenta noe som egentlig ikke er lov? Hva er det?

Dette er den første oppgaven til teksten *Stargate* oppgaveformatet er flervalgsoppgave. Tabell 2 viser fordelingen av de ulike svaralternativene fra vårt opprinnelige datasett på 2000 elevsvar for 8. og 9. trinn.

Tabell 2: Fordeling av elevsvar fra opprinnelig utvalg

Spørsmål 1 Svaralternativ	I begynnelsen av teksten gjør jenta noe som egentlig ikke er lov. Hva er det?	8. trinn elevbesvarelser antall	prosent	9. trinn elevbesvarelser antall	prosent
1	Hun røyker.	450	22,5 %	440	22 %
2	Hun henger opp plakater.	154	7,7 %	127	6,3%
3	Hun mater et ekorn.	642	32,1 %	637	31,8 %
4	Hun står utenfor skolegården	749	37,45 %	784	39,2 %
Blank	Elever som ikke har valgt noen av alternativene	5	0,25 %	12	0,6 %

Det første spørsmålet gir leseren svar på hvem jeg-personen i denne teksten er. Det er en jente. Om eleven sammenholder dette med informasjon som presentert implisitt tidligere, forstår de at det er en ung jente på ti år. Dersom noen elever velger alternativ 2, kan det føre til videre feiltolkninger. Vi kan se etter sammenhenger ved å studere hvor mange prosent av elevene i vårt opprinnelige datasett som har svart alternativ 2. Resultatet viser at det er 154 av elevene på 8.trinn og 127 elever på 9.trinn, som har svart at jenta henger opp plakaten. Av elevene på 8.trinn er det 154 som har svart alternativ 2. Disse har de fordelt seg med 92 elever som fikk 0 poeng, 51 elever som fikk 1 poeng, og 11 elever som fikk 2 poeng på den åpne oppgaven. Resultatet tyder på at mange har denne feiltolkningen med videre i sin lesing. Bare 11 av disse elevene har klart å få 2 poeng på den åpne oppgaven.

Antall av Stargate_1_RESPONSE	Kolonneetiketter					
Radetiketter	choice_1	choice_2	choice_3	choice_4	NA	Totalsum
0	245	92	243	227	4	811
1	142	51	267	264	1	725
2	63	11	132	258		464
Totalsum	**450**	**154**	**642**	**749**	**5**	**2000**

Figur 5: Korrelasjon mellom svaralternativ 2 og poengfordeling åpen oppgave – 8. trinn

Oppgave 2.

Hvilke ord passer for å beskrive vaktmesteren? Sett kryss for passer eller passer ikke.

Det andre spørsmålet til teksten er også et flervalgspørsmål, men med et annet design. Her må elevene velge hvilke påstander som stemmer. Vi legger merke til at det beskrives som «koblinger» og ikke avkrysninger. Kan dette misforståes av elevene siden de bare skal krysse av for et alternativ og går raskt videre?

Oppgave 2 av 6 til teksten

Hvilke ord passer for å beskrive vaktmesteren? Sett kryss for «Passer» eller «Passer ikke».

Du må lage opptil 5 koblinger

	Passer	Passer ikke
omsorgsfull	○	○
streng	○	○
smart	○	○
masete	○	○
slem	○	○

Figur 6: Bilde av oppgave 2 - flervalgsoppgave

Vi lyktes ikke i å foreta en analyse av hvordan elevsvarene fordelte seg på hvert av alternativene, da vi ikke klarte å skille diss variablene i kolonnen i vår rå-data. Vi har foretatt den beregning av hvor mange elever som har fått poeng på oppgaven.

Tabell 3: Poengfordeling oppgave 2. 8 og 9. trinn

Poeng	8. trinn Antall	Prosent %	9. trinn Antall	Prosent %
0	1064	53,2	897	44,85
1	936	46,8	1103	55,15
totalt	2000	100	2000	100

Fordelingen viser at det var høy andel av elevene som hadde fått poeng på denne oppgaven. Vi legger merke til at antallet riktige svar går noe ned for 9.tinn.

Oppgave 3.

For å besvare denne flervalgsoppgaven må elevene har kjennskap til at tekst i kursiv er et virkemiddel som benyttes for å vise at dette er direkte sitat/ avskrift fra plakaten, slik den så ut da den hang på stolpen. Elevene må ha kunnskap om typografiske konvensjoner i tekst. Vi kan argumentere for at dette er forkunnskaper og kanskje ikke noe de nødvendigvis leser ut av teksten. På den andre siden kan en finne støtte i teksten til å forstå dette, da utdraget er formulert slik: «Søker juletreselger, sto det. *Du er: pliktoppfyllende. Ansvarsfull. Glad i å*

være ute.». Dette sitatet er til hjelp for å forstå at utdraget er tekst fra plakaten. Elevene kan finne tekstlig støtte i at det står «Søker juletreselger, sto det». Vi finner holdepunkt for at elevene kan få støtte i både i teksten, men enda lettere forstå spørsmålet med forkunnskap om tekstlige strukturer. Vi ser sammenheng mellom vår vurdering og Roe, Ryen og Weyergang sin (Roe, Ryen, & Weyergang, 2018, s. 75).

I vårt datasett er svarene fordelt som vist i tabellen under. Den viser at dette er en oppgave relativt mange har fått til. Prosentvis fordeling av rett svar er for 8. trinn 70,6 % og 74,7 % for 9.trinn. Det er også relativt få blanke svar.

Tabell 4: Prosentvis fordeling av elevsvar oppgave 3 - Stargate

Svaralternativ	Hvorfor er «Du er: pliktoppfyllende. Ansvarsfull. Glad i å være ute» skrevet med kursiv (skrå skrift)?	8.trinn Antall - prosenttall		9. trinn Antall - prosenttall	
1	For å vise at det er noe vaktmesteren sier	265	13,25%	208	10,4 %
2	For å vise at det er noe jenta tenker	156	7,8 %	134	6,7 %
3	For å få fram at det er noe som står på plakaten	1412	70,6 %	1494	74,7 %
4	For å få frem at det er ironisk ment	153	7,65 %	146	7,3 %
Blank	Elever som ikke har valgt noen av alternativene	14	13,3 %	18	0,9 %

Oppgave 4. Hvem kan passe som juletreselger, ifølge vaktmesteren?

Hva elevene svarer på dette spørsmålet anser vi som viktig for hvilken forståelse de har for teksten så langt. Det er mange elever som har svart riktig alternativ, «faren til jenta». Denne forståelsen hjelper elevene med å svare på den åpne oppgaven.

Tabell 5:Prosentvis fordeling av elevsvar oppgave 4 - Stargate

Svaralternativ	Hvem kan passe som juletreselger, ifølge vaktmesteren?	8. trinn Antall - prosenttall		9. trinn Antall - prosenttall	
1	Jenta	178	8,9 %	138	6,9 %
2	Faren til jenta	1514	75,7 %	1590	79,5 %
3	Vaktmesteren selv	85	4,25 %	67	3,35 %
4	Alfred	214	10,7 %	192	9,6 %
Blank	Elever som ikke har valgt noen av alternativene	9	0,45 %	13	0,65 %

Oppgave 5 Åpen oppgave: Hvorfor tar vaktmesteren ned hele plakaten og gir den til jenta?

Oppgave fem er den eneste åpne oppgaven til teksten Stargate, og elevene formulerer sine egne svar. Elevene møter ingen beskrivelse for forventet lengde. Elevene besvarer oppgaven med egne ord i en tekstboks, som utvider seg hvis teksten blir lengre. Dette er også den eneste åpne oppgaven i den nasjonale leseprøven som det er mulig oppnå 2 poeng på. Poengskalaen går fra 0, 1 og 2 poeng, også kalt en polytom oppgave (Fevold, Thoresen, & Eriksen, 2019, s. 9). Oppgaven har som formål å vurdere leseformålet: "Å kunne tolke og trekke slutninger på bakgrunn av informasjon i teksten" og er plassert inn under mestringsnivå 4. I rammeverket beskrives delskalaen slik: «Forstå hvordan ikke tydelige informasjonselementer i en eller flere tekster henger sammen, og/eller hvordan disse henger sammen med tekstene som helhet. Forstå meningsinnhold som står i motsetning til det forventede» (Utdanningsdirektoratet, 2022, s. 22).

Oppgaveteksten er ganske tydelig og kort, det er lite informasjon å holde rede på. Den nevner også eksplisitt poenget om at vaktmesteren tar ned hele plakaten. Eleven må ha denne informasjon med seg når de besvarer oppgaven med egne ord. Først tar vaktmesteren ned nummeret fra lappen, og på slutten tar han ned hele plakaten. Det er nettopp denne siste delen som det er viktig at elevene har med i besvarelsen for å få høyeste poengsum. Dette synliggjør hvilken forståelse elevene må ha for å få full skår på denne åpne oppgaven. Elevene må forstå teksten på et dypere nivå og ikke bare bokstavelig. Elevene som har klart å få 2 poeng på denne oppgaven, har klart å tolke og trekke slutning om at vaktmesteren tar ned plakaten for å hindre andre enn jentas far å søke på jobben og/eller øke sjansen for at han får jobben. Disse elevene har klart å gå ut over tekstens eksplisitte formulering. Tekstutdraget slutter med at vaktmesteren stikker rullen med plakaten i jakkelomma til jenta, og kommer med sluttreplikken; «Nettopp, sa han – du står og ser på en stor tenker.». I tillegg krever den åpne oppgaven at eleven uttrykker sin forståelse med egne ord. Dette er i seg selv noen som kan være en hindring for elever som har svake skriftlig uttrykksevne. Oppgaveteksten har et spørreord som er viktig, *hvorfor*. Utover dette blir ikke elevene presentert for noen presisering om hva som forventes av svaret. De blir ikke bedt om å vise til eksempler fra teksten. Vi tenker det er mulig å tolke oppgaven både på et overfladisk og bokstavelig nivå, eller et dypere og mer symbolsk nivå.

I tillegg krever den åpne oppgaven at eleven uttrykker sin forståelse med egne ord. Dette er i seg selv noen som kan være en hindring for elever som har svake skriftlig uttrykksevne. Oppgaveteksten har et spørreord som er viktig, *hvorfor*. Utover dette blir ikke elevene presentert for noen presisering om hva som forventes av svaret. De blir ikke bedt om å vise til eksempler fra teksten. Vi tenker det er mulig å tolke oppgaven både på et overfladisk og bokstavelig nivå, eller et dypere og mer symbolsk nivå. Roe og Blikstad-Balas omtaler forskjellen på modenhet hos gutter og jenter. Hvor guttene i noen tilfeller svarte bokstavelig og konkret, mens jentene i større grad begrunnet svarene med støtte i symbolske tolkningshandlinger. I tillegg nevner de at symbolsk tolkning krever større grad av modenhet (Roe & Blikstad-Balas, 2022, s. 62). I vår oppgave har vi ikke rette fokuset spesielt mot kjønnsforskjeller, men dette poenget anses som aktuelt å diskutere opp mot validitetsaspektet for oppgavene. Hvordan forstår elevene hva de blir bedt om å beskrive. Kan det være slik at modenheten påvirker forståelsen av både spørsmålet som stilles også elevsvaret som blir gitt? Tabell 5 viser oversikt over hvor mange elever som har klart å oppnå 1, 2 eller 0 poengfordeling i vårt opprinnelige datasett på 4000 elevsvar for 8. og 9. trinn.

Tabell 6: Poengfordeling åpen oppgave - Stargate

Poeng i skåring	Hvorfor tar vaktmesteren ned hele plakaten og gir den til jenta?	8.trinn Resultater		9.trinn Resultater	
2		811	40,5 %	644	32,2 %
1		725	36,25 %	629	31,45 %
0		464	23,2 %	722	36,1 %

4.1.3 Vurderingsveilederen til skåring av åpen oppgave - *Stargate*

Tekst: STARGATE

Oppgave 5: Hvorfor tar vaktmesteren ned hele plakaten og gir den til jenta?

Oppgavens formål: *Å kunne tolke og trekke slutninger på bakgrunn av informasjon i teksten.*

2 poeng
Svar som viser til at vaktmesteren vil hindre andre enn jentas far å søke, eller at han vil øke sjansen for at faren får jobben.

- *Han gjør det fordi han vil ikke at andre skal ha muligheten til å søke på jobben. Han vil at jenta sin far skal få jobben.*
- *Vaktmesteren ville liksom at faren skulle være den eneste som visste om jobben og kunne søke på den.*
- *Sånn at det ikke er noen andre som søker enn faren til jenta.*
- *For at ingen andre skal prøve å få jobben.* [Underforstått: ingen andre enn faren til jenta.]
- *Han tar ned plakaten fordi det ikke er lov å henge opp plakater der. Jeg tror også at han fjerner den så faren til jenta skal være den eneste som ser den og søker og dermed får jobben.* [Se bort fra første del, siste del av svaret er riktig.]

1 poeng
Svar som enten viser til at vaktmesteren vil at jentas far skal søke på jobben, eller at jenta skal vise plakaten til faren sin.

- *Han ville at jenta skulle vise plakaten til faren.*
- *Han ville at faren skulle søke på jobben.*
- *Sånn at pappaen kan få jobb.*
- *Han vil at faren skal søke jobben, og det var ikke lov å henge plakater der.* [Se bort fra den siste delen av svaret, første del er riktig.]

0 poeng
Svar som ikke viser til at vaktmesteren vil hindre andre å søke på jobben, eller at jenta skal vise plakaten til faren sin. Andre vage, irrelevante eller gale svar.

- *Fordi det ikke er lov å henge opp plakater der.*
- *For å gi den til jenta sånn at hun kan få seg en jobb eller vise den til faren sånn at han kan hjelpe henne med å få jobben.*
- *Vaktmesteren tar ned plakaten for å være sikker på at jenta får jobben.*
- *Fordi han mente at man kan ikke få jobb når man er 10 år.*

Figur 7: Vurderingsveilederen Stargate

Vurderingsveilederen til Stargate trekker tre hovedskiller mellom svar som skal skåres til 0, 1 eller 2 poeng. Hver av disse inneholder en overordnet beskrivelse av kriteriene og kulepunkt med eksempler på elevsvar som kan godkjennes eller ikke. Noen av eksemplene har også en

klammeform som har en tilleggskommentar. Disse beskrives som følger: “Svar som viser at eleven har forstått at vaktmesteren vil hindre andre enn jentas far å søke, eller at han vil øke sjansen for at faren får jobben”. Dette forståes som at verbene *hindre* og, eller *øke sjansen* er viktig. I vårt arbeid med materialet og tolkning av veilederen forstår vi det som at det var tilstrekkelig med ett av de to alternativene, enten hindre eller øke sjansen, for å få 2 poeng. Det begrunner vi med at det står «hindre og, eller», i tillegg underbygges det av eksemplene på elevsvar.

Vurderingsveilederen kommer med eksempler på svar som skal vurderes som irrelevant, vage og gale. Disse skal gis 0 poeng. Denne vurderingen kommer ikke med noen beskrivelse om hvorfor disse svarene skal tolkes slik. Den enkelte lærer som skal skåre må støtte seg til eksemplene for å vurdere hva som er for vagt, galt eller irrelevant. Vi har reflektert over verdien av tilleggsopplysninger bak eksemplene. Med det mener vi eksempler på feiltolkning som: elever som tror at jenta skal søke på jobben. Misoppfatninger: At Alfred er faren til jenta. Vurderingsveilederen inneholder autentiske elevbesvarelser fra pilotering av prøven (Roe, Ryen, & Weyergang, 2018, s. 185). I arbeidet med eksemplene finner vi noen uklarheter. I fjerde kulepunkt under eksempler på elevsvar som kan få to 2 poeng finner vi følgende eksempel: «For at ingen andre skal prøve å få jobben. (Underforstått: ingen andre enn faren til jenta.)» Dette kulepunktet vurderer vi som ikke samsvarende med det overordnede kriteriet som er beskrevet for svar som skal få 2 poeng. I denne beskrivelse står det at svar som viser til at vaktmesteren vil *hindre* eller *øke sjansen* for at faren får jobben. Svaret i kulepunktet viser ikke til hvem jobben gjelder.

4.1.4 Resultater og funn fra kategorisering av åpen oppgave

Vi skåret elevbesvarelsene i henhold til vurderingsveilederen, og grupperte svarene i fem ulike kategoriene som vi hadde utarbeidet jmf. metodekapittelet.

Tabell 7: Kategoribeskrivelse

Kategori	Beskrivelse
A	Elevbesvarelsen vurderes som riktig i henhold til vurderingsveilederen
B	Elevbesvarelsen vurderes som uriktig i henhold til vurderingsveilederen
C	Elevbesvarelse som er vanskelig å skåre /Tvilstilfelle
D	Elevuttrykk
E	Blankt svar

Kategoriene skulle indikere i hvilken grad vi mente elevsvaret ble fanget opp av vurderingsveilederen eller ikke. Hvor stor andel vurderte vi som lett å skåre i henhold til veilederen og hvor oppstod tvilen? Kategoriene våre rommer også elevuttrykk som «idk[9]» og blanke svar. Hovedtanken med å kategorisere elevbesvarelsene under skåring var å markere hvordan vår subjektive opplevelse av skåringen var. Dette var til hjelp da vi skulle komme fram til et omforent svar som vi kunne sammenligne med de virkelige skåringene som elevenes lærere (lærer x) hadde gitt. Den kvantitative kategoriseringen tallfestet antallet elevsvar fordelt på kategorier. Tabellene viser resultatene fra kategorisering av elevsvar Stargate.

Tabell 8: Resultat fra kategorisering 8. trinn Stargate

Stargate 8. trinn	S1 Cecilie		S2 Christina	
Kategori	Antall elevsvar	Prosent %	Antall elevsvar	Prosent %
A	168	56 %	163	54,33 %
B	106	35,45 %	118	39,33 %
C	11	3,68 %	5	1,67 %
D	6	2,01 %	6	2 %
E	8	2,68 %	8	2,67 %
Totalt	300	100	300	100

[9] Engelsk forkortelse for *I don't know.*

Tabell 9: Resultat fra kategorisering 9. trinn Stargate

Stargate 9. trinn	S1 Cecilie		S2 Christina	
Kategori	Antall elevsvar	Prosent %	Antall elevsvar	Prosent %
A	178	59,33 %	172	57,33 %
B	70	23,33 %	86	28,67 %
C	13	4,33 %	4	1,33 %
D	8	2,67 %	7	2,33 %
E	31	10,33 %	31	10,33 %
Totalt	300	100	300	100

Tabell 7 viser differansen mellom antall elevsvar vi har kategorisert som A på 8.trinn. Skårer 1 Cecilie har plassert 168 (56,00 %) elevsvar i kategori A, og skårer 2 Christina har kategorisert 163 (54,33 %) elevsvar som A. Resultatene for 9.trinn har likheter i fordelingen, Cecilie med 59,33 % og Christina med 57,33 % av elevsvarene vurdert i kategori A. Det som er unikt for *Stargate* er at kategori A, rett svar i henhold til vurderingsveilederen, rommer variablene 1 og 2 poeng. Det vil si at elevsvarene vi har kategorisert som A har variasjon i poengsummen fra 1 til 2 poeng. Med andre ord, har det vært større variasjon i poengsummene mellom skåringene våre enn differansen mellom 178 og 172 viser for eksempelvis 9. trinn. Vi vil presentere resultater for forholdet mellom 1 og 2 poeng senere.

Elevsvar vurdert som uriktige svar i henhold til veilederen på 8. trinn utgjør 106 (35,33 %) for skårer 1 Cecilie og 118 (39,33 %) for skårer 2 Christina. For elevsvarene på 9. trinn har Cecilie vurdert 70 (23,33 %) elevsvar som uriktige og Christina 86 (28,67 %) som uriktig. Det vil si en differanse på 12 og 16 elevsvar, mellom Cecilie og Christina sin kategorisering for 8. og 9. trinn. Det er viktig være oppmerksom på at det ulike elevsvar som er vurdert til 0 poeng. Dette er ikke et mål på interrater reliabilitet, noe vi foretar i kap. 4.1.6.
Skårer 1, Cecilie har under skåring og kategorisering valgt kategori C oftere. Resultatet viser at Cecilie har definert 12 (4,00 %) elevsvar i kategori C og Christina har 5 (1,67 %) på 8.trinn. For 9. trinn fordelte det seg med 13 (4,33 %) elevsvar for Cecilie, og 4 (1,33 %) elevsvar for Christina. Det er et uttrykk for at Cecilie har vært usikker på flere tilfeller enn Christina. Denne fremstilling sier ikke noe om hvilke elevsvar, hver av oss har vurdert som kategori C. Det er mulig at vi har kategorisert både samme og ulike elevsvar som tvilstilfeller.

Kategori D, elevuttrykk utgjorde 6 elevsvar 2,67 % for både skårere 1 og 2 på 8.trinn. I kategori E – som var ubesvarte svar/blanke svar var det 8 svar som utgjorde 2,67 % av

svarene, også her likt for oss begge. For 9. trinn utgjorde kategori D 8 elevsvar (2,67 %) for Cecilie og 7 elevsvar (2,33 %) Christina. Kategori E er en god del større på 9. trinn med 31 (10,33%) elevbesvarelser som er blanke, ubesvart. Vi undres oss over dette funnet.

Resultatene fra disse to kategoriene vurderer vi som nyttig informasjon når vi forsøker å oppnå en oversikt over hvordan elevsvarene fordelte seg. Kategori D rommer ulike elevuttrykk som for eksempel elevsvaret fra rad 258 - «hver ikke», rad 219- «vet ikke», rad 90- «69». Elevsvarene kan også ses på som uttrykk for at elevene ikke mestrer oppgaven, eller bare skriver noe for å fylle tomrommet. Disse kan også ses opp mot antallet blanke svar, kategori E og indikere vanskelighetsgraden på teksten og oppgaven. Eventuelt at elevene ikke mestrer oppgaven.

4.1.5 Kvalitative funn i diskusjon fram til omforent skår

I utvalget på 600 elevbesvarelser knyttet til Stargate fra 8. trinn og 9. trinn diskuterte vi oss frem til en omforent vurdering og skår. Da vi gjennomgikk elevsvarene med diskrepans, ble elevsvarene vurdert på nytt. I utgangspunktet hadde vi samsvar i 270/300 elevsvar på den åpne oppgaven i Stargate 8. trinn og 269/300 9. trinn. Vi hadde en uenighet på 30 elevsvar for 8. trinn og 31 på 9.trinn.

Vi vil i dette avsnittet beskrive fagsamtalene vi hadde i arbeidet med å komme fram til en omforent skår. Avsnittet under dokumenterer prosessen og reelle samtaler vi hadde. Synspunktene som kommer fram, vil hver for seg representere skritt i retning av en omforent forståelse- og dermed av en omforent skåring. I omtale av oss som uavhengige vurderere har vi i datasettet brukt begrepene skårer 1 Cecilie og skårer 2 Christina. Vi velger å bruke våre navn for å forenkle språket, men også for å vise at vi har nærhet til prosessen. Elevsitatene fra datasettet er markert i kursiv, dette for å synliggjøre sitatene i diskusjonen.

En fellesnevner for elevbesvarelsene vi hadde vurdert ulikt, var at de verken var sammenfallende med den overordnede beskrivelsen, eller eksemplene på riktige og gale elevsvar. Noen av elevbesvarelsene kunne også treffe på flere av kriteriene i vurderingsveilederen, enten den generelle beskrivelsen eller eleveksemplene i kulepunktene.

En tydelig utfordring var, som nevnt, elevsvar som ikke falt godt inn under noen av vurderingsveilederens punkter, men som likevel kunne indikere at eleven hadde forstått deler av teksten. Disse utmerket seg som vanskelig å vurdere, og ble grunnlag for faglig diskusjon.

Et eksempel på dette var elevsvar nr. 216, 8. trinn: *«Fordi han sa at det ikke er lov å henge plakater der, og fordi han ville at faren hennes skulle ringe flere ganger.»*

Dette elevsvaret åpnet opp for mye tolkningsarbeid. Cecilie hadde markert elevsvaret som 1C - 1 poeng og tvilstilfelle, mens Christina hadde markert besvarelsen som 0B- 0 poeng og uriktig svar i henhold til veilederen. Den første leddsetningen *«Fordi han sa at det ikke er lov å henge plakater der, (...)»*. Vurderingsveilederen åpner opp for at læreren kan se bort fra første leddsetning, så lenge andre leddsetning kan vurderes som riktig og kvalifisere til 1 eller 2 poeng. Den andre leddsetningen i elevsvaret *«(...) og fordi han ville at faren hennes skulle ringe flere ganger»* kan indikere at eleven har forstått at faren skulle ringe for å få jobben. Underforstått: *å ringe flere ganger* skulle øke sjansene hans for å få jobben. I tekstutdraget *Stargate* får vi vite at plakaten har lapper med telefonnummer på. Cecilie argumenterte for at eleven med sin besvarelse viste en forståelse av formatet på slike plakater/jobbannonser. I prøveutdraget kommer det frem at vaktmesteren river av en lapp, men mot slutten tar han ned hele plakaten. Har eleven forstått at også ved å ringe flere ganger viser du stor interesse for jobben? Cecilie undrer seg også over om denne til og med kan vurderes som 2 poeng, nettopp på grunn av at å ringe flere ganger kan forståes som: det øker sjansen for å få jobben.

Christina vurderte at den første leddsetningen til eleven «*Fordi han sa at det ikke er lov å henge plakater der* (…) indikerer at eleven har misforstått hensikten med at vaktmesteren tok ned plakaten. Likevel påpeker vurderingsveilederen at vi skal se bort fra denne delen av besvarelsen som er gal. Den andre leddsetningen «(…) *og fordi han ville at faren hennes skulle ringe flere ganger.»* sammenfaller ikke med noen av eksemplene fra elevsvar som kvalifiserer til 1 eller 2 poeng. Likevel kom Christina fram til at elevsvaret kanskje kunne vurderes inn under 1 poeng «Svar som viser til at jentas far skal søke på jobben (…)», men at hun da støtter seg til en sterk tvil som skal komme eleven til gode. Den omforente vurderingen ble derfor satt til 1 poeng.

Et annet funn var flere elevsvar som inneholdt «å ringe på jobben». For eksempel rad 285, 8. trinn: «For at faren skal få ringt nummeret», og elevsvar nr.19*: "Vaktmesteren tok plakaten til jenta fordi man kan ringe et tlf nr som sto på plakaten og det tlf nr er til et juletre jobb".* Disse elevsvarene va utfordrende å skåre da veilederen ikke hadde noen eksempler på slike elevsvar. Ordet *ringe* nevnes ikke i vurderingsveilederen. I vårt datasett nevnes ordet ringe til sammen 6 ganger. I arbeidet med vurdering og skåring måtte vi flere ganger vurdere, hvor

langt kunne vi gå i å tolke elevsvarene som handlet om *ringing.* Dette vil vi komme tilbake til i drøftingsdelen.

Teksten formidler en viss tvetydighet, som ikke alle elevsvarene uttrykker forståelse for. Til tross for dette har mange av disse svarene fått med seg de andre delene av forståelsen som gir 1 og 2 poeng. Et eksempel på dette er fra Rad 227, 9. trinn: «*Fordi han vil at faren får jobben ved at ingen flere har tilgang til plakaten. Og at da ingen søker utenom han. Vaktmesteren tar den også ned fordi det ikke er lov å henge plakater der.»* I dette eksemplet er det siste setningen som er uriktig svar, mens første del kvalifiserer til to poeng i henhold til vurderingsveilederen. Dette svaret har vi begge skåret til 2A hver for oss og omforent skår ble satt til 2 poeng. Vi diskuterte behovet for å skille elever som bare uttrykker forståelse fra de som uttrykker misforståelse i tillegg. Hvordan kan vi skille mellom mer korrekt lesing eller forståelse opp mot tydelige misforståelser. Vurderingsveilederen påpeker at første eller andre leddsetning kan inneholde uriktige opplysninger, men at så lenge en av leddsetningene er riktig, skal eleven få poeng. Videre kan vi diskutere hvorvidt «fordi det ikke er lov å henge å plakater der» er en feil oppfatning? Derimot viser et slikt svar ikke dypere forståelse av hvorfor vaktmesteren tok ned plakaten. Når vi studerte elevsvarene gjorde vi observasjoner av leddsetningen, «fordi det ikke er lov å henge plakater der» ofte.Vi tror mange elever som har med dette i svaret kan ha søkelest og funnet «svaret» og kanskje stoppet der.

Elevsvaret på rad 85, 8. trinn er et tilsvarende eksempel på at den første leddsetningen ikke skal vurderes; «*Fordi de er ikke lov å henge plakaten og for at faren til jenta skal få jobben først»*. Cecilie hadde vurdert elevsvaret til 2A, (2 poeng og rett svar i henhold til vurderingsveilederen) og Christina 1A, (1 poeng og rett svar i henhold til vurderingsveilederen). Likevel ble det gode diskusjoner om den andre leddsetningens ordlyd og hvordan denne kunne fanges opp av veilederens eksempler. Her mente Cecilie at faren skal få jobben «først» kunne forståes som "fremfor" noen andre jobbsøkere. Med andre ord argumenterte Cecilie for at eleven har eleven forstått teksten, men uttrykker seg ikke så nøyaktig. Vurderingsveilederens kriterier for 2 poeng sier at elevsvaret må være et «Svar som viser at vaktmesteren vil hindre andre enn jentas far å søke, eller at han vil øke sjansen for at faren får jobben».

Christina mente at elevsvaret på rad 85 egentlig ikke traff så godt til denne beskrivelsen, og mente at veilederens kriterier for 1 poeng «Svar som viser til at vaktmesteren vil at jentas far skal søke jobben, eller at jenta skal vise plakaten til faren sin» passet noe bedre. Christina

argumenterte videre for at leddsetningen «(…) *for at faren til jenta skal få jobben først»* passer bedre under veilederens kulepunkt 8 «Sånn at pappaen kan få jobb». Elevsvaret sier ikke noe om «(…) å hindre at andre skal få jobben». Etter diskusjon ble vi enig om at å "få jobben først" ble noe upresist, og elevsvaret ble til slutt vurdert til 1 poeng i omforent skår.

En annen kilde til diskrepans var elevsvar som ble vurdert til ulike poeng. Dette til tross for at vi ikke nødvendigvis hadde markert elevsvaret som kategori C, tvilstilfelle. Et eksempel på et slikt funn er rad 95, 8. trinn: *«Fordi han ikke ville at noen andre skulle sende inn jobbsøknad.»* Cecilie har vurderte dette til 0 poeng og kategori B, begrunnet i at svaret er upresist og det uttrykker ikke forståelse for hvem som skulle søke på jobben. Christina har vurdert dette svaret til 2A- 2 poeng og riktig i henhold til veilederen. Hun viser til at elevsvaret kan sidestilles med kulepunkt 4 i veilederen: «For at ingen andre skal prøve å få jobben [Underforstått: ingen andre enn faren til jenta.]». Vi diskuterte hva vi «underforstått» kunne lese av elevsvaret og hvordan vi kan tolke eleven her. Den omforente skåringen ble satt til 2 poeng.

En ny gruppering av elevbesvarelser hadde fellestrekket at de var for vage eller utydelige. Disse kunne være utfordrende å skåre, skrivefeil, ufullstendige setninger kan tenkes å påvirke skåringen i noen tilfeller. Da ble tolkningsrommet og bruken av vårt faglige skjønn svært viktig. Samtidig gav det oss en innsikt i hvor ulikt skjønn kan utøves. Et eksempel på dette er elevbesvarelsen på rad 148, 8. trinn: *«For ay bare Faren seal få jobb.»* Skårer 1 Cecilie har gitt 2A, vurdert svarte til å falle inn under veilederens krav for to poeng. Dette begrunnes med at ordet «bare» kan forståes som det samme som å hindre andre i å få jobben. Christina har gitt 1 poeng med begrunnelsen at elevsvaret har likheter med kulepunkt 8 som viser til: «Sånn at pappaen kan få jobbe». Begge er enig i at eleven skal ha poeng og vi ender med en omforent skår på 1 poeng. Som vurderere må vi foreta en skjønnsmessig vurdering i hvordan vi skal tolke elevsvaret, når eleven uttrykker seg vagt eller utydelig. Nivået av elevenes skriftkyndighet kan påvirke hvilken vurdering eleven får i skåring. Dette kan problematiseres når vurderingsveilederen påpeker at skrivefeil skal ses bort fra i vurdering. Vi kom også over elevsvar som hadde blitt feilvurdert i henhold til vurderingsveilederen.

Som vurderere oppdaget vi feilskåringer som vi hadde gjort. Et eksempel på slike feilskåringer var elevbesvarelsen fra rad 38: «fordi det ikke er lov med plakater der». Her har Cecilie skåret 1A – 1 poeng og riktig svar som er i henhold til vurderingsveilederen og Christina skåret 0B, 0 poeng og galt svar som er i henhold til vurderingsveilederen. Dette er

en åpenbar feilskåring. Vi lander på å gi eleven 0 poeng. Feilskåring av åpne oppgaver er noe som kan skje i et stort datasett i slike prosjekter, men også i en reell vurderingssituasjon kan feilskåringer forekomme. Det er viktig å påpeke at vi hadde låst våre endelige vurderinger av elevsvaret før felles gjennomgang fram til omforent skår. Der vi oppdaget feilskåringer, gikk vi ikke tilbake i datasettet for å korrigere dette, men elevsvaret ble vurdert på nytt i omforent skår.

Vi hadde også noen tydelige uenigheter, det handlet i stort om hvor langt vi skulle ta hensyn til hva eleven kan ha ment i sin besvarelse. Her var vi uenig flere ganger og det viser seg i flere av eksemplene under. Elevsvaret på rad 23 var interessant: «Fordi henne skal ta den med hjem og vise faren hennes (Alfred)».

Her har vi skåret ulikt. Cecilie hadde merket elevsvaret som 0 poeng og kategori B- feilsvar i henhold til vurderingsveilederen. Christina hadde merket elevsvaret som 1 poeng og kategori A – rett svar i henhold til vurderingsveilederen. Diskusjonen som fulgte handlet om at Cecilie mente at eleven hadde en tydelig misoppfatning av hvem Alfred faktisk er i fortellingen. Alfred er vaktmesteren, og ikke faren til jenta. Dermed vurderte Cecilie elevsvaret til 0 poeng fordi eleven i elevsvaret sitt viser en grunnleggende misforståelse av teksten. Christina argumenterte for at eleven skriver at det er plakaten som skal vises til pappaen, og mener at misforståelsen om navnet ikke er så relevant. Christina mente også at elevsvaret i seg selv faller godt inn under vurderingsveilederens eksempler på riktige elevsvar som skal ha poeng, og velger å se bort fra at eleven har feil navn med i besvarelsen. Vi enes om å gi eleven 1 poeng.

Kulepunkt 4 og 8 i vurderingsveilederen var en tilbakevendende hovedgruppe i diskusjonen fram til omforent skår. Det kan se ut som at disse kulepunktene gjorde tolkningsrommet uklart og vi syntes de var noe uklar. Dette viste seg gjennom flere elevsvar som endte opp med å bli diskutert opp mot kulepunktene. I det følgende trekker vi fram noen slike eksempler. Elevsvaret på rad 15, 8. trinn: «*For at ikke andre skal søke jobben.»*.

Her var vi begge enig om å gi eleven 2 poeng da svaret traff på kriteriet for overordnet beskrivelse for 2 poeng: *«Svar som viser til at vaktmesteren vil hindre andre enn jentas far å søke jobben, (...)»*. Christina hadde markert elevsvaret som 2C, 2 poeng og tvilstilfelle på bakgrunn om besvarelsen var presis nok til å gi eleven 2 poeng. Cecilie argumenterte for at

eleven ved å svare som den gjorde, implisitt hadde forstått at faren til jenta trengte en jobb og at vaktmesteren ville hindre andre i å få jobben ved å ta ned plakaten.

Kulepunkt 4 beskriver: «For at ingen andre skal prøve å få jobben. [Underforstått: ingen andre enn faren til jenta.]. Klammeformens tilleggskommentar åpner opp for en bredere tolkning og større tvetydighet i vurderingsveilederen. Sammenlignet med kulepunkt 8: «Sånn at pappaen kan få jobb», vurderer vi som et mer presist kriterium i veiledningen. Mens kulepunkt 4 bare viser til at *ingen andre* skal få jobben, men sier ikke noe om elevens forståelse for hvem som skal søke jobb. Dette åpner opp for at læreren som skal vurdere elevsvarene kan bruke sitt faglige skjønn. Likevel kan denne formuleringen åpne opp for ulike vurderinger.

Elevsvar som inneholdt ulike verb i sitt forsøk på å beskrive at faren skulle få jobben, ble også grunnlag for noe diskusjon. Vi fant flere eksempler på diskrepans i skåring da elevene hadde brukt ulike verb i sin beskrivelse av at faren skulle få/ta/ha/søke jobben. For eksempel rad 105, 8. trinn: «*Vaktmesteren tar ned plakaten og gir den til jenta fordi han mener at faren til jenta* ***kunne fått*** *den jobben».* Et annet eksempel er rad 123: *«Det var ikke lov å henge plakater der, og sånn at ingen andre skulle* ***ta*** *jobben».* Diskrepansen i dette handlet om hvordan vi vurderte elevenes bruk av verb opp mot eksemplene i veilederen. I fagsamtalene ble vi enig om at disse elevsvarene skulle ha poeng i skåring. Vi innså at vi hadde blitt for opptatt av kulepunktene med eksempler, og kanskje la vi mindre vekt på den overordnede beskrivelsen av hva som kjennetegnet riktig eller uriktig svar. Dette er et element vi vil gjennomgå nærmere i kapittel 5 «Tolkningsrommet og lærerens skjønn».

4.1.6 Kvantitativt resultat fra skåring 8. og 9. trinn

Tabellen under viser resultatene i skåring sammenlignet med vår omforente skår sett opp mot de autentiske lærernes vurderinger. Oversikt over skåringsresultater skårer 1 Cecilie, skårer 2 Christina, vår omforent skår og lærer x. «Lærer x» representeres av autentiske lærere som har gitt sine skåringer til de ulike elevsvarene i vårt datasett.

Tabell 10: Poengfordeling 8. trinn - Stargate

Poengskår 8. trinn	S1 Cecilie	S2 Christina	Omforent skår	Lærer x
0	122	132	124	121
1	123	108	118	117
2	55	60	58	62
Sum antall	300	300	300	300
Total poengsum	233	228	234	241

Tabell 11: Poengfordeling 9. trinn Stargate

Poengskår 9. trinn	S1 Cecilie	S2 Christina	Omforent skår	Lærer x
0	108	125	115	122
1	103	88	90	78
2	89	87	95	100
Sum antall	300	300	300	300
Total poengsum	281	262	280	278

Tabell 2 og 3 viser at den omforente skåren i større grad er samsvarende med lærer x. Det vil si hvor mange av de 300 elevsvarene fra skåringsutvalget for 8.trinn som har blitt vurdert til 0 poeng. Resultatet som viser antallet elevsvar som har fått 1 poeng, har i omforent skår blitt 118 av de 300. Differansen mellom oss og lærer x er da 1 elevsvar. Den omforente skåringen ligger relativt nærme lærer x, men litt færre 2-poengsskåringer og litt flere 0- og 1- poeng skåringer. Differansen mellom Cecilie og Christina, utgjør en forskjell på 15 elevsvar som vi har gitt 1 poeng. Christina har skåret flest elevsvar med 0 poeng med et antall på 132. Lærer x skåret noen flere svar til 2 poeng. Resultatet indikerer at det er forskjeller i hvordan skårerne tolker og vurderer elevsvarene.

4.1.6.1 Cohens kappa 8. trinn - *Stargate*

Cohens kappa er en statistikk som brukes til å måle interrater reliabilitet for kvalitative eller kategoriske elementer, som beskrevet i metodekapittelet. Cohens kappa kontrollerer både den enighet som oppstår som følge av tilfeldigheter og tar hensyn til hvor stor avstanden er mellom to observatører/vurderer (Tengberg & Skar, 2016, ss. 7-8).

I de følgende avsnittene vil vi vise våre beregninger av Cohens kappa og Fleiss kappa. Vi sammenligner våre individuelle skåringsresultater opp mot hverandre, Cecilie (S1) og Christina (S2) for begge trinn. Deretter beregner vi vår omforente skår opp mot lærer x for begge trinn. Til sist foretar vi en Fleiss multirater kappa med Cecilie, Christina og Lærer x.

Tabell 12: Krysstabell -Skårer 1 Cecilie og skårer 2 Christina

S1 * S2 Crosstabulation

			S2			
			0	1	2	Total
S1	0	Count	119	2	1	122
		% within S1	97,5%	1,6%	0,8%	100,0%
		% within S2	90,2%	1,9%	1,7%	40,7%
		% of Total	39,7%	0,7%	0,3%	40,7%
	1	Count	11	102	10	123
		% within S1	8,9%	82,9%	8,1%	100,0%
		% within S2	8,3%	94,4%	16,7%	41,0%
		% of Total	3,7%	34,0%	3,3%	41,0%
	2	Count	2	4	49	55
		% within S1	3,6%	7,3%	89,1%	100,0%
		% within S2	1,5%	3,7%	81,7%	18,3%
		% of Total	0,7%	1,3%	16,3%	18,3%
Total		Count	132	108	60	300
		% within S1	44,0%	36,0%	20,0%	100,0%
		% within S2	100,0%	100,0%	100,0%	100,0%
		% of Total	44,0%	36,0%	20,0%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,843	,027	20,010	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Krysstabellen viser at Cecilie og Christina har vurdert med stor grad av samsvar. Den største differansen finner vi mellom utdeling av 1 og 2 poeng, til sammen 14 tilfeller. Cecilie har gitt 2 elevsvar 2 poeng som Christina har gitt 0 poeng. Kappa verdien for Cecilie og Christina er, ,843, som er tilfredsstillende jmf. Tengberg (2016) konsistensindikatorer, som helst bør være over ,70. Videre beregnet vi kappaverdien for vår omforente skår opp mot lærer x for 8. trinn.

Tabell 13: Krysstabell lærer x og omforent skår

Lx * Omforent Crosstabulation

			Omforent			
			0	1	2	Total
Lx	0	Count	116	4	1	121
		% within Lx	95,9%	3,3%	0,8%	100,0%
		% within Omforent	93,5%	3,4%	1,7%	40,3%
		% of Total	38,7%	1,3%	0,3%	40,3%
	1	Count	5	97	15	117
		% within Lx	4,3%	82,9%	12,8%	100,0%
		% within Omforent	4,0%	82,2%	25,9%	39,0%
		% of Total	1,7%	32,3%	5,0%	39,0%
	2	Count	3	17	42	62
		% within Lx	4,8%	27,4%	67,7%	100,0%
		% within Omforent	2,4%	14,4%	72,4%	20,7%
		% of Total	1,0%	5,7%	14,0%	20,7%
Total		Count	124	118	58	300
		% within Lx	41,3%	39,3%	19,3%	100,0%
		% within Omforent	100,0%	100,0%	100,0%	100,0%
		% of Total	41,3%	39,3%	19,3%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,766	,032	18,199	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

I krysstabellen er det flest tilfeller av differanse mellom utdeling av 1 og 2 poeng. Det er 17 tilfeller hvor lærer x har gitt 2 poeng og omforent har gitt 1 poeng. Det er også 15 tilfeller hvor omforent skår har gitt 2 poeng og lærer x har gitt 2. Cohens kappa for interrater reliabilitet av vår omforente skår sammenlignet med lærer x viste ,766. Sammenlignet med

reliabilitetsmålet for Christina målt mot Cecilie på , 843, viser at reliabilitetsmålet lærer x mot omforent skår at kappaverdien ble noe lavere. I *Stargate* viser tallene at vi som skårere alene og sammen utfører skåring med konsistens. Det kan ha sammenheng med vår særlige interesse for de åpne oppgavene og at vi valgte å fordype oss i disse.

4.1.6.2 Cohens kappa 9. trinn – *Stargate*

Tabell 14: Krysstabell Skårer 1 Cecilie og skårer 2 Christina

S1 * S2 Crosstabulation

			S2			
			0	1	2	Total
S1	0	Count	108	0	0	108
		% within S1	100,0%	0,0%	0,0%	100,0%
		% within S2	86,4%	0,0%	0,0%	36,0%
		% of Total	36,0%	0,0%	0,0%	36,0%
	1	Count	13	79	11	103
		% within S1	12,6%	76,7%	10,7%	100,0%
		% within S2	10,4%	89,8%	12,6%	34,3%
		% of Total	4,3%	26,3%	3,7%	34,3%
	2	Count	4	9	76	89
		% within S1	4,5%	10,1%	85,4%	100,0%
		% within S2	3,2%	10,2%	87,4%	29,7%
		% of Total	1,3%	3,0%	25,3%	29,7%
Total		Count	125	88	87	300
		% within S1	41,7%	29,3%	29,0%	100,0%
		% within S2	100,0%	100,0%	100,0%	100,0%
		% of Total	41,7%	29,3%	29,0%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T [b]	Approximate Significance
Measure of Agreement	Kappa	,814	,028	19,981	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Vi utførte de samme beregningene for 9. trinn for å undersøke om vi så de samme tendensene i datamaterialet. Tabell X viser Cecilie opp mot Christina.

Krysstabellen viser også for 9. trinn at diskrepans i skåring handler mest om hvilke elevsvar som skal ha 1 eller 2 poeng, 18/300 elevsvar. Vi finner også fire tilfeller hvor Cecilie har gitt 2 poeng og Christina 0 poeng. Kappaverdien for 9. trinn mellom Cecilie og Christina går litt ned fra 8.trinn med verdien , 843 mot , 814. Kappaverdien kan fortsatt anses som høy. Den neste tabellen viser beregninger av Cohens kappa for omforent skår opp mot lærer x. Se tabellen under:

Tabell 15: Krysstabell Lærer x og omforent skår

Lx * Omforent Crosstabulation

			Omforent			
			0	1	2	Total
Lx	0	Count	110	7	5	122
		% within Lx	90,2%	5,7%	4,1%	100,0%
		% within Omforent	95,7%	7,8%	5,3%	40,7%
		% of Total	36,7%	2,3%	1,7%	40,7%
	1	Count	5	66	7	78
		% within Lx	6,4%	84,6%	9,0%	100,0%
		% within Omforent	4,3%	73,3%	7,4%	26,0%
		% of Total	1,7%	22,0%	2,3%	26,0%
	2	Count	0	17	83	100
		% within Lx	0,0%	17,0%	83,0%	100,0%
		% within Omforent	0,0%	18,9%	87,4%	33,3%
		% of Total	0,0%	5,7%	27,7%	33,3%
Total		Count	115	90	95	300
		% within Lx	38,3%	30,0%	31,7%	100,0%
		% within Omforent	100,0%	100,0%	100,0%	100,0%
		% of Total	38,3%	30,0%	31,7%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,793	,030	19,380	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Krysstabellen for Cohens kappa mot lærer x viser også her at diskrepans i skåring handler mest om hvilke elevsvar som skal ha 1 eller 2 poeng. Vi legger merke til at den omforente

skåren har gitt 5 elevsvar 2 poeng, som lærer x opprinnelig hadde skåre til 0 poeng. Kappaverdien for omforent skår opp mot lærer x viser et mål på ,793. Videre valgte vi å beregne Fleiss kappa som måler interrater reliabilitet mellom flere enn to vurderere, som vist i metodekapittelet 3.2. I SPSS la vi inn Cecilie, Christina og Lærer x målt opp mot hverandre.

4.1.6.3 Fleiss kappa 8. trinn - *Stargate*

Tabell 16: Fleiss kappa 8. trinn - Stargate

Overall Agreement[a]

		Asymptotic			Asymptotic 95% Confidence Interval	
	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
Overall Agreement	,767	,024	31,498	<,001	,719	,814

a. Sample data contains 300 effective subjects and 3 raters.

Agreement on Individual Categories[a]

			Asymptotic			Asymptotic 95% Confidence Interval	
Rating Category	Conditional Probability	Kappa	Standard Error	z	Sig.	Lower Bound	Upper Bound
0	,920	,863	,033	25,886	<,001	,798	,928
1	,830	,724	,033	21,707	<,001	,658	,789
2	,746	,684	,033	20,506	<,001	,618	,749

a. Sample data contains 300 effective subjects and 3 raters.

Fleiss kappa viser en verdi på ,767 som konsistensmål for alle tre skårerne. Vi ser også at kappaverdien «synker» i takt med den polytome poengskalaen. Det er høyere overenstemmelse i hvilke elevsvar som skulle ha 0 poeng, enn 1 og 2 poeng. Det viser at interrater reliabiliteten har dårligst samsvar i å fastsette 2 poeng, kappa ,684. Det kan muligens forklares med at når kravene for å få poeng etter vurderingsveilederens kriterier økte, økte også diskrepansen i skåring.

4.1.6.4 Fleiss kappa 9. trinn - *Stargate*

Tabell 17: Fleiss kappa 9. trinn - Stargate

Overall Agreement[a]

	Kappa	Asymptotic Standard Error	Asymptotic z	Asymptotic Sig.	Asymptotic 95% Confidence Interval Lower Bound	Asymptotic 95% Confidence Interval Upper Bound
Overall Agreement	,765	,024	32,328	<,001	,718	,811

a. Sample data contains 300 effective subjects and 3 raters.

Agreement on Individual Categories[a]

Rating Category	Conditional Probability	Kappa	Asymptotic Standard Error	Asymptotic z	Asymptotic Sig.	Asymptotic 95% Interv Lower Bound
0	,915	,860	,033	25,813	<,001	,795
1	,766	,666	,033	19,979	<,001	,601
2	,830	,754	,033	22,632	<,001	,689

a. Sample data contains 300 effective subjects and 3 raters.

For 9. trinn ble Fleiss kappa ,765, nesten samme som på 8. trinn med ,767. Det er interessant å trekke fram at det også her er akseptabel grad av enighet omkring elevsvar som er vurdert til 0 poeng med ,860. Samtidig ser vi at reliabilitetsmålet «synker» i takt med poenggivingen, men skiller seg ut i tildeling av 1 poeng. Vi ser at graden av enighet i elevsvar som er tildelt 1 poeng er lavere enn for 2 poeng med en kappaverdi på ,666.

4.1.7 Drøfting av hovedfunn fra *Stargate*

I arbeidet med teksten *Stargate*, oppgavene, vurderingsveilederen og skåringen har vi forsøkt å oppnå oversikt og innsikt i hvordan forholdet mellom disse delene virker inn på validiteten og å reliabiliteten. I det følgende vil vi drøfte funn som peker seg spesielt ut. Det vil si sentrale faktorer for validiteten i målingen av lesekompetansen, og hva som kan påvirke reliabiliteten i skåringen av den åpne oppgaven.

For besvare den åpne oppgaven til *Stargate* krevde det at elevene kunne tolke og trekke slutninger på bakgrunn av informasjon i teksten. I tekstgjennomgangen så vi at elevenes forforståelse også kan ha betydning for å trekke de rette slutningene. Vi reflekterer over at det til en viss grad alltid vil være slik at forkunnskaper påvirker lesingen. Til tross for dette, kan det problematisere hva oppgaven måler. Er forkunnskaper avgjørende for å kunne besvare oppgaven godt nok? En slik forforståelse er det flerkulturelle bakteppet i *Stargate*, hvor handlingen skjer på Tøyen. Er det slik at det er veldig ufordelaktig for en elev fra Finnmark og forstå, eller stiller alle elvene ganske likt? Forforståelse av det kulturelle preget i teksten, kan støtte leseren i å trekke de rette slutningene. *Stargate* er nok ikke en tekst som mange 8. og 9. klassinger hadde valgt å lese på egenhånd, men den har tematikk som noe kan kjenne seg igjen i. Teksten har en jente som hovedperson, og den er lokalisert til østkanten i Oslo. I tillegg er denne teksten med på å gjenspeile et tekstmangfold i samfunnet slik som rammeverket beskriver (Utdanningsdirektoratet, 2022, s. 8).

Det bringer oss videre til tolkningsrommet som utfordret elevene i den skjønnlitterære teksten *Stargate*. Dette krevde at leseren måtte lese mellom linjene. Det er ikke vanskelig å argumentere for at de skjønnlitterære tekstene har en naturlig plass i de nasjonale prøvene, som redskap for å vurdere grunnleggende leseferdigheter og leseforståelse (Roe, Ryen, & Weyergang, 2018, s. 79). I rammeverket beskrives leseformålet «å tolke og sammenholde informasjon» som en leseprosess som spenner imellom det å trekke enkle slutninger til mer kompliserte. I rammeverket beskrives validitet i prøven på oppgavenivå:

> Oppgavene skal spenne fra enkle til det komplekse. De skal åpne for at elevene kan vise sine ferdigheter på både høyt og lavt nivå. Prøvens målinger skal ha så høy presisjon som mulig, slik at resultatene fra prøven blir sikre på alle deler av ferdighetsskalaen (Utdanningsdirektoratet, 2022, s. 15).

Vi forstår det slik at Utdanningsdirektoratet sier at nasjonale prøver i lesing er utviklet for å måle dette spennet i leseferdigheter med høy presisjon langs hele ferdighetsskalaen. Våre diskusjoner dreide seg om hva som utfordret denne presisjonen i målingen. Presisjon i målingen av leseferdigheter hos elevene, og for lærere i vurderingsarbeidet.

Den åpne oppgaven i *Stargate* er plassert inn på mestringsnivå 4 og er derfor en litt krevende oppgave. For at elevene skulle få 2 poeng måtte tolke og trekke slutninger som viste at de hadde forstått at vaktmesteren tok ned plakaten for å hindre andre å få jobben. Nettopp denne forskjellen i innsikt skulle vurderes med en ploytom skala, 0, 1 eller 2 poeng. Her drar Stargate opp et skille mellom de elevene som har størst innsikt, og som klarer å forstå og trekke slutninger som går ut over tekstens eksplisitte uttrykk. Dette har vi forstått slik som Roe, Ryen og Weyergang, at disse elevene har reflektert selvstendig (Roe, Ryen, & Weyergang, 2018, s. 88).

Poengskalaen i vurderingsveiledningen gir rom for å skille mellom grad av selvstendighet og refleksjon i møte med teksten. Vi har reflekter over at det kan være utfordrende å plassere elevsvarene inn i en slik polytom skala. Våre funn fra Fleiss kappa målet fra tabell 9 og 10 viste at graden av interrater reliabiliteten sank i takt med økende krav til elevsvarene og mulighet for 2 poeng. Tengberg og Skar (2016) gjengir Eckes (2012) som beskriver at forskning på feltet om «rater cognition», viser at strenghetsgraden øker i takt med vanskegrad; «Jo viktigere kriterium, desto strengere vurdering» (Tengberg & Skar, 2016, s. 15). For vår del er dette gjenkjennbart fra skåringsdiskusjonene hvor vi noen ganger brukte argumentasjonsrekker som «Skal dette svaret virkelig få 2 poeng? Innfrir eleven kriteriene for å få høyest poengsum?». Vil det komplekse i dette true presisjon og pålitelighetene i målingene?

Vurderingsveiledningen skal være en rettesnor for hva som kan falle innenfor modellesingen av teksten (Frønes & Ryen, 2020, s. 139). Tengberg, Roe og Skar (2018) anbefalte i sin artikkel at vurderingsveilerede bør inneholde flere eksempler på elevsvar som kan godkjennes og ikke godkjennes, slik at rommet for tvil hos vurdereren blir mest mulig begrenset (Tengberg, Roe, & Skar, 2018). Vi diskuterte hvorvidt vurderingsveilederen i *Stargate* rommet mange nok eksempler på elevsvar som kunne godkjennes. Spesielt oppstod behovet for flere eksempler på elevbesvarelser når vi skulle skåre elevsvar som inneholdt nye momenter som vurderingsveilederen ikke rommet. Et av våre funn var flere elevsvar som inneholdt ordet *å ringe*, eksempelvis, «å ringe på jobben». I vurderingsveilederen er det ingen

eksempler på elevsvar som inneholder ordet ringing. Vårt skåringssett inneholdt 6 slike svar av de 300 for 9.trinn. Kanskje er det slik at eksemplene på elevbesvarelser ikke rommer godt nok alle autentiske variasjoner? Dette til tross for at prøvene piloteres på forhånd. Våre funn på svar som innholdet ringing er kanskje eksempler på nettopp dette. Som presentert i teoridelen viser Skaftun (2006) til at vurderingsveilederen aldri være entydig, og det vil alltid vil være noe rom for tolkning. Om vi begynner å stole på vurderingsguiden som en autoritativ sannhet har vi, ifølge Skaftun, gått i den «testteoretiske rottefella» (Skaftun, 2006, s. 40).

I den andre ytterkanten kan en tenke seg vurderingsveiledere med bare en overordnet beskrivelse av hva som skal gi de ulike poengsummene. Gitt dette, hadde vi stått fritt til å bruke vårt profesjonsfaglig skjønn i større grad. Vi diskuterte dette flere ganger. Spesielt oppsto problemet når vi opplevde kulepunkt 4 «For at ingen andre skal prøve å få jobben (Underforstått: ingen andre enn faren til jenta) og kulepunkt 8 «Sånn at pappaen kan få jobb» som uklare og i noen tilfeller ikke i samsvar med den overordnende beskrivelsen av kriteriene. Diskusjoner som gjenspeilet frustrasjon rundt at den overordnede beskrivelsen for 1 og 2 poeng ikke alltid var i samsvar med eksemplene for godkjente elevsvar. De små kommentarene til eleveksemplene kunne virke selvmotsigende og til forvirring. Kommentaren presentert i klammeform til kulepunkt 4; "Underforstått: ingen andre enn faren til jenta" legger opp til at læreren kan gå svært langt i å tolke elevsvarene. «Intensjonen med skåringsguiden er å fylle tolkningsgapet, som åpner seg mellom spørsmålet og teksten (Solheim & Skaftun, 2009, s. 156)». Uklarheten i kulepunkt 4 og 8 som for oss skapte forvirring, brøt med intensjonen om at vurderingsveilederen skal fylle tolkningsgapet, men tvert imot opplevde vi på dette punktet at veilederen gjorde tolkningsgapet større. Solheim og Skaftun beskriver disse ytterpunktene for vurdering av åpne oppgaver. På den ene siden maksimalt uttømmende beskrivelser og i andre enden åpnere retningslinjer for skåring og tolkning av elevsvar (Solheim & Skaftun, 2009, ss. 156-157). I drøftingen har vi bevegd oss mellom disse ytterpunktene og har vurdert det ene opp mot det andre. Dette tar oss videre til oppgaveformatet – de åpne oppgavene.

I arbeidet med de åpne oppgavene som oppgaveformat, har vi fått en større innsikt i mulighetene og begrensinger de kan inneha. Solheim og Skaftun (2009) omtaler i sin artikkel at åpne oppgaver formatet, de åpne oppgavene ofte brukes tilknyttet tolkningsoppgaver. Tolkning som leseformål som gir elevene mulighet til å vise dybde i sin forståelse. Da må oppgaven og teksten åpne opp med muligheter for dette. Vi har tenkt over i hvilket omfang

eleven står fritt til å tolke og trekke slutninger i Stargate, kanskje er ikke tolkningsrommet så stort. Denne problemstillingen vil vi drøfte nærmere.

Vi mener at det finnes et ubenyttet potensial i Stargate. Teksten har mange elementer som gir muligheter til å stille gode åpne spørsmål, der elevene kan vise grad selvstendighet i å tolke og trekke slutninger. I faglige diskusjoner rundt materialet utforsket vi ideen om andre mulige spørsmålsstillinger. Forslag til spørsmål som hadde åpnet opp for enda større tolkningsrom: Hva tror du vaktmesteren mener med «the circle of life»? Eller «Hva mener vaktmesteren når han sier; -du står her og ser på en stor tenker?» Ved å åpne tolkningsrommet lurere vi på om flere sterke lesere hadde fått vist sitt potensiale. Vi forstår som Solheim og Skaftun at bruk av åpne oppgaver er viktig for å kunne måle elevens grad av forståelse og oppgavetypen gir rom for å utnytte tekstens sin semantiske potensiale (Solheim & Skaftun, 2009, s. 155). Det krever også en god åpen oppgave. Som igjen gir oss mulighet til å få innblikk i elevenes tolkinger og refleksjoner. Fra lærerperspektivet gir det oss informasjon både på individ- og gruppenivå. Her er det en mulighet for å bruke dette formativt i videre undervisning. En større åpenhet i tekst og oppgaveformulering vil også medføre andre krav til en vurderingsveileder og skåringen.

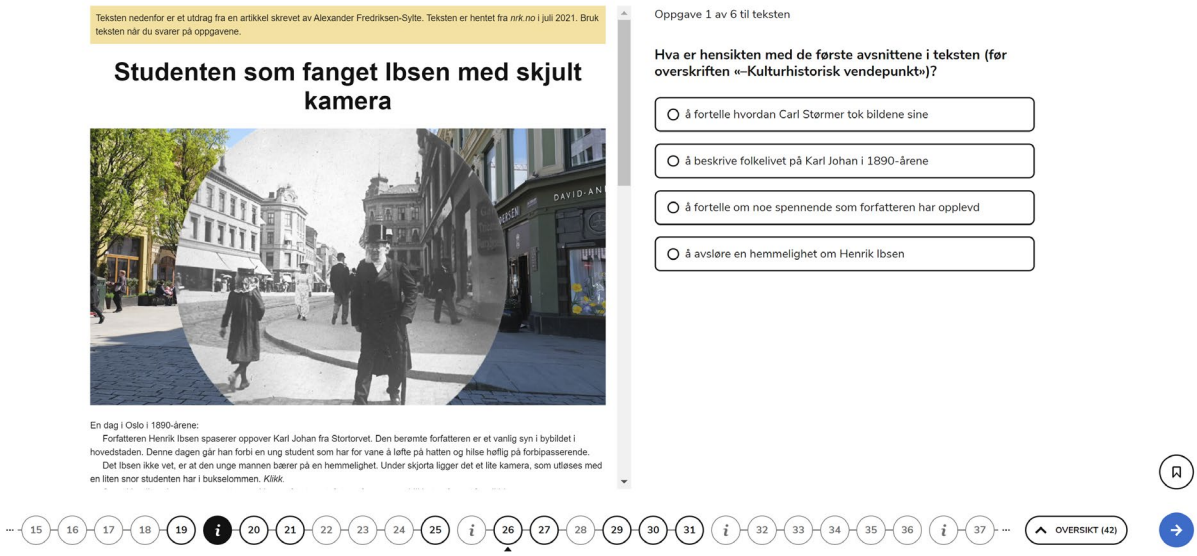
I det følgende vil vi drøfte noen begrensninger som ble synlig for oss i arbeidet med de åpne oppgavene til *Stargate*. I arbeidet med skåringen gjorde vi funn av elevsvar som inneholdt både riktig og uriktig forståelse av teksten. Vi reflekterte over hvorvidt det er problematisk at mange elever ikke uttrykker bare riktig forståelse, men i tillegg tar med feiltolkninger og irrelevant informasjon. Eller er det uproblematisk så lenge de har forstått resten? Vi synes dette er vanskelig å foren med at disse svarene skal skåres til 2 poeng som er den høyeste poengsummen de kan få. Bør ikke den additive skalaen skille ut de med høyest grad av riktig forståelse? Vurderingsveilederen til Stargate viser i kulepunkt 5 til et eksempel på elevsvar som skåres til 2 poeng. Veilederen kommenterer at læreren skal se bort fra den første leddsetningen “Han tar ned plakaten fordi det ikke er lov til å henge plakater der (...)” fordi siste del av svaret er riktig. Dette kan problematiseres ved at det åpner opp for et brudd i rekken av stigende krav til elevsvarene. Vi synes det er problematisk at svar av veldig ulik kvalitet kan skåres til 2 poeng, som hos Solheim og Skaftun (Solheim & Skaftun, 2009, s. 160). Dette anser vi som en trussel både på validitet i målingen av grad av lesekompetanse og i andre rekke reliabiliteten i skåringen.

4.2 Snikfotografen

4.2.1 Tekstgjennomgang *Snikfotografen*

Teksten nedenfor er et utdrag fra en artikkel skrevet av Alexander Fredriksen-Sylte. Teksten er hentet fra *nrk.no* i juli 2021. Bruk teksten når du svarer på oppgavene.

Tekstutdraget i leseprøven er hentet fra en nettartikkel fra NRK- kultur. Prøveteksten er knyttet opp mot samfunnsfag. *Snikfotografen* er en sammensatt tekst som inkluderer en tekstboks som vist over, et fremtredende bilde og en tydelig overskrift. I tillegg inneholder artikkelen tre mindre bilder, hver med tilhørende bildetekst. Den er strukturert med tre underoverskrifter og vi vurderer det til å være en sammenhengende tekst som leses fra begynnelse til slutt. Elevene kan skrolle opp og ned som i *Stargate* for å lese teksten. Etter dette kommer oppgavene til spørsmålene til høyre på skjermen.



Figur 8: Snikfotografen som vist i prøvesituasjon

Prøveteksten er ikke lik den autentiske nettartikkelen, siden noen bilder og avsnitt er utelatt. I tillegg er det gjort noen mindre endringer i verbalteksten. *Dengang* er endret til den gang, og 1890- tallet er endret til 1890-årene. Endringer i ordet kan ha blitt gjort for å modernisere språket, slik at elevene letter forstå hva som menes.

Den originale nettartikkelen på nrk.no er lengre enn utdraget brukt i den nasjonale leseprøven høsten 2023[10]. I tillegg inneholder den flere elementer som bildekarusell, video og større bilder. Nettartikkelen er multimodal med funksjoner der leseren kan bevege bildene til høyre eller venstre for å skifte fra nåtid til 1890-tallet. I tillegg er det to tekstbokser, den første med fakta om spionkameraet og de runde bildene, figur 10. Denne boksen består av tre kulepunkt og to avsnitt. Det siste avsnittet er flyttet og omgjort til ny bildetekst i leseprøveutgaven, som vist i bilde til høyre i figur 10. Det betyr at det er gjort endringer også i strukturen til teksten. Figur 1 viser teksten i den autentiske nettartikkelen. Vi vurderer det som tydelig for leseren og forstå at dette er relatert til kameraets tekniske aspekter. Det er også andre deler av den originale nettartikkelen som er utelatt. Informasjonen som underbygger hvorfor personene på portrettbildene virket veldig oppstilte. Elevene får derfor ingen visuell støtte i bildene for å forstå at menneskene måtte sitte helt stille og se alvorlig ut, ved fotografering. Den opprinnelige bildeteksten til bilde vist til venstre i figur 10 er: «I tillegg til de 500 bildene som Norsk Teknisk Museum sitter på i dag, laget Carl Størmer et album med overskuddsbilder og kopier som han fordelte til egen familie.».

Figur 9: Utklipp fra den originale nettartikkelen - Snikfotografen

[10] https://www.nrk.no/kultur/xl/carl-stormer_-studenten-som-snikfotograferte-ibsen-1.15501709 Sist lest 23.04.2024

Tiltros for at opprinnelige nettartikkelen er lengre og inneholder flere elementer, tror vi at dens detaljrikdom og gjentakelse av sentrale poeng, bidrar til en dypere forståelse for leseren. De ulike modalitetene – som illustrasjoner, faktabokser og grafiske fremstillinger – gir leseren mer hjelp og støtte sammenlignet med den forenklede tekstversjonen som benyttes i leseprøven. Roe, Ryen og Weyergang påpeker at multimodale tekster kan ha flere leseveier, men det kreves en aktiv leser som kobler ulike uttrykk i teksten sammen (Roe, Ryen, & Weyergang, 2018, s. 125).

I tekstutdraget fra leseprøven er teksten stilt til venstre og bildene plassert til høyre. Roe, Ryen og Weyergang deler multimodale tekster inn i ulike strukturer og *Snikfotografen* kan plasseres som en blanding av vertikal og sentrert struktur (Roe, Ryen, & Weyergang, 2018, s. 125). Er det slik at disse detaljene kan påvirke hvordan elevene navigerer i teksten, noe som igjen påvirker leseforståelsen? Hvordan elevene navigerer i tekststrukturen som foreligger i *Snikfotografen*, kan igjen påvirke leseforståelsen. Roe et al påpeker at komposisjon gir en ide om verdien til de ulike elementene i teksten.

Verbaltekster formidles gjennom både ord, setninger, men også ved valg av skrifttype, uthevinger og layout. Samspillet mellom verbalteksten og de andre elementene har ulike intensjoner. I prøveteksten er bildeteksten og bildene med på å tydeliggjøre og visualisere innholdet (Roe, Ryen, & Weyergang, 2018, ss. 53-54). Ifølge Anne Løvland er visuelle kilder viktig for å formidle historiske sannheter, det kan være vanskelig for elevene å se for seg hvordan ting var langt tilbake i tid (Løvland, 2015, s. 126). Vi har diskutert i hvilken grad bildene som er valgt ut fra den autentiske teksten støtter eleven mot den åpne oppgaven eller ikke. Av de tre mindre bildene som er valgt ut, er det bilde av spionkameraet med bildetekst; «… slik så det ut ...». Det er også et bilde av Carl Størmer, der bildeteksten presenterer han som en livsglad og utadvendt student. Det opplyses også om at Størmer studerte matematikk ved universitetet i datidens Kristiania. Det siste bilde er allerede kommentert, der de også har endret teksten som satte søkelys på det teknisk ved bildet. Elevene må tolke og sammenholde denne informasjon for å forstå at dette var med på å gi bilder med mennesker som var uformelle og ikke oppstilte. Det er ingen tydelig eksplisitt informasjon i bildeteksten som hjelper elevene til å besvare den åpne oppgaven som spør etter «På hvilken måte så folk annerledes ut på Carl Størmers bilder sammenlignet med typiske bilder fra 1890-årene?». Elevene må forstå at størrelsen på kameraet lot fotografen Størmer ta bilder i naturlige settinger, ettersom kameraet var et lite spionkamera. Videre må de ha forstått at dette var

årsaken til at menneskene som ble tatt bilde av, fremsto mindre alvorlig, i motsetning til de mer formelle og oppstilte bildene som var vanlig.

Hovedformålet med denne teksten er å vurdere elevenes evne til å finne informasjon i teksten. Det er en åpen oppgave til denne teksten, oppgave 3 som er plassert i mestringsnivå 4. Det vil si at elevene skal: «Lokalisere og kombinere informasjon fra ulike steder i en eller flere tekster og vurdere hvilken informasjon som er relevant» (Utdanningsdirektoratet, 2022, s. 22).

Overskriften er: «Studenten som fanget Ibsen med skjult kamera». Nettartikler har ofte fengende overskrifter som skal skape interesse. Elever som vet hvem Henrik Ibsen er vil få en forventning om at teksten handler om han og tiden han levde. Ibsen er ikke så sentral for det videre innholdet, men det unike var at det ble tatt gode bilder av han, uten at han var klar over det. Ibsen blir omtalt som en berømt forfatter og som en superkjendis.

Artikkelen innledes med et avsnitt som setter leseren inn i en kontekst for tid og sted. Første setning er: «En dag i Oslo i 1890-årene». Deretter beskrives at Henrik Ibsen spaserte oppover Karl Johan fra Stortorvet, som var et vanlig syn i bybildet. Bybildet er et ikke et direkte uttrykk for hvordan byen så ut, men en implisitt beskrivelse av at Ibsen var vanlig å se i byen på den tiden. Vi tror dette kan utfordre noen elever sin forståelse. Språkbruken har kanskje litt avstand fra ungdomsskole elevers primærdiskurs, og muligens også enda vanskeligere å forstå for elever med norsk som andrespråk. I samme avsnitt brukes også uttrykket «- og øyeblikket er fanget for alltid …». I boken *Lesedidaktikk – etter den første leseopplæringen* referer Roe og Blikstad-Balas til doktoravhandlingen av Anne Golden. Hun gjorde funn på at minoritetselever i ungdomsskolen hadde problemer med å forstå språklige metaforer (Roe & Blikstad-Balas, 2022, s. 59).

«Kulturhistorisk vendepunkt» er den første underoverskriften, og representerer terminologi som er vanlig i samfunnsfaglige tekster. Slike sammensatte ord kan være vanskeligere å lese og å forstå ifølge Roe og Blikstad-Balas (Roe & Blikstad-Balas, 2022, s. 58). *Snikfotografen* inneholder mange sammensatte ord som, bybildet gatefotograf, turgåeren, snikfotograf, kulturhistorisk, vitenskapsmenn, overskuddsbilder og spionkamera. Ordene består av et forledd og etterledd som får ny betydning når de blir satt sammen. I det norske språket spesifiserer forleddet etterleddet, eksempelvis spionkamera. For språk som eksempelvis somalisk er det motsatt rekkefølge på sammensatte substantiv (Golden & Kulbrandstad, 2007, s. 55).

Elevenes evne til å forstå teksten vil påvirkes av i hvor stor grad de har kjennskap til ordene og begrepene. Elever med forkunnskaper og forståelse for ord som *kulturhistorisk,* kan vi anta leser med større sammenheng. Dette gjør det enklere å oppfatte at måten bildene ble tatt på, gjorde de uformelle og ikke oppstilt. Disse elevene forstår at det ikke bare handler om de tekniske nyvinningene, men konteksten for hvordan bildene ble tatt. I det samme avsnitte forklares det at det endret «fotografiets natur». Dette er også en ny metafor som kan være vanskelig å forstå.

4.2.2 Gjennomgang av spørsmålene til *Snikfotografen*

Vi undersøkte alle spørsmålene elevene møtte i forkant av det åpne spørsmålet til teksten *Snikfotografen*, som i *Stargate*. Spørsmålsstillingen og svaralternativene i flervalgsoppgavene kan lede elevene til å forholde seg til teksten på en ny måte, og vise «vei» fram til det åpne spørsmålet. Tabellen under viser oversikt over hvordan elevsvarene fordelte seg prosentvis fra det opprinnelige utvalget på 2000 elever fra hvert av trinnene.

Tabell 18: Fordeling av elevsvar – Snikfotografen - oppgave 1

Oppgave 1 Svaralternativ	Hva er hensikten med de første avsnittene i teksten (før overskriften «–Kulturhistorisk vendepunkt»)?	8. trinn Antall	 Prosent	9. trinn Antall	 Prosent
1	Å fortelle hvordan Carl Størmer tok bildene.	1238	61,9 %	1347	67,3 %
2	Å beskrive folkelivet på Karl Johan i 1890- årene	323	16,2 %	276	13,8 %
3	Å fortelle om noe spennende som forfatteren har opplevd	208	10,4 %	186	9,3 %
4	Å avsløre en hemmelighet om Henrik Ibsen	177	8,9 %	134	6,7 %
5	Blank- Elever som ikke har valgt noen av alternativene	54	2,7 %	57	2,9 %

For å gi riktig svar må elevene forstå hovedtematikken som er innleder resten av artikkelen. Det er hele 61,9 % og 67,3 % elever på 8. og 9 trinn som har valgt riktig alternativ på spørsmålet, som handler om hvordan Carl Størmer tok sine bilder. Det er grunn til å anta at elevene som har svart riktig her, er på rett vei i lesingen siden de har forstått det første avsnittet. Nå vil det kunne være en forventing hos leseren om å få mer informasjon om disse bildene i den videre lesingen.

Tabell 19: Fordeling av elevsvar – Snikfotografen- oppgave 2

Oppgave 2 Svaralternativ	Hvem mener at Størmers bilder representerer et «kulturhistorisk vendepunkt»?	8. trinn Antall	Prosent	9. trinn Antall	Prosent
1	Robert Meyer	1260	63 %	1399	70 %
2	Alv Egeland	167	8,3 %	129	6,4 %
3	Georg Størmer	341	17 %	263	13,2 %
4	Alexander Fredriksen-Sylte	167	8,3 %	137	6,9 %
5	Blank- Elever som ikke har valgt noen av alternativene	65	3,2 %	72	3,6 %

For spørsmål 2 til *Snikfotografen* fordelte svarene seg slik som vist i tabellen. En relativ høy prosentandel med 63 % og 70 % har svart riktig, Robert Meyer. Det fremkommer ganske tydelig hvem som sier dette i teksten: «Dette er et kulturhistorisk vendepunkt, mener Robert Meyer.» Til tross for at det også er nevnt en del ulike navn som leseren må forholde seg til, klarer mange å finne frem til riktig informasjon i teksten.

Den åpne oppgaven til *Snikfotografen* er det tredje spørsmålet og det er kun en åpen oppgave til denne prøveteksten. Det mulig å oppnå 1 poeng for riktig svar eller 0 poeng for galt svar, en såkalt dikotom skala. Oppgavens formål er teste elevenes evne til å finne informasjon i teksten. I rammeverket beskrives denne lesekompetansen: «Lokalisere og kombinere informasjon fra ulike steder i en eller flere tekster og vurdere hvilken informasjon som er relevant.» (Utdanningsdirektoratet, 2022, s. 22). Oppgaveteksten spør: «På hvilken måte så folk annerledes ut på Carl Størmers bilder sammenlignet med typiske bilder fra 1890-årene?»

Her må elevene først tolke hva som menes med «å se annerledes ut». Annerledes i forhold til klær og tiden de levde i, eller hvordan de fremstod optisk. Forståelsen av ordets betydning kan være ulikt for en 13- 14 åring og voksne lesere. Referansen og relasjonen til ordet kan endre seg over tid. Elevens svar vil formes av forståelsen de har tatt med seg fra teksten, oppgaveteksten og egne forkunnskaper.

For å svare riktig på den åpne oppgaven må elevene lokalisere hvor i teksten det er beskrevet at folk så annerledes ut. I første avsnitt blir Carl Størmer omtalt som en gatefotograf som fanger øyeblikk. Dette er det første i teksten som viser til at det var endringer i hvordan bilder

ble tatt: «Nå ble folk på gata tatt bilde av uten at de var klar over det». Elevene må identifisere en av tekstens hovedhensikter og forstå spørsmålsstillingen.

Elevene må videre i teksten for å få mer informasjon, i andre avsnitt står det: «Kameraet ga ny form for bilder, og endret selve fotografiets natur.» Dette er en ny ledetråd på endringen, men fortsatt ikke beskrevet eksplisitt. Like etterpå er det gjengitt fra et intervju med Størmer fra 1942, hvor han forteller om gatefotograferingen: «Jeg var en ung student på 19 år den gang og hadde fått et morsomt detektivkamera. Jeg slentret nedover Carl Johan, utså mig et offer, hilste, fikk et blidt smil og trakk av». Videre henviser teksten til uttalelse gjort av Robert Meyer hvor han sier at det ble mulig å ta bilder uten at det handlet om selve fotograferingen. Bildene var nå av mennesker i situasjoner hvor de ikke var så alvorlige, og han mente at det kunne en se i bildene. I tillegg blir det brukt forklaring om en mer subjektiv beskrivelse. Disse opplysningene må elevene ha registrert for å besvare den åpne oppgaven på en god måte. I vårt opprinnelige datasett, som besto av et tilfeldig utvalg av totalt 4000 elevsvar fra 8. og 9. trinn, er elevsvarene skåret med følgende fordeling:

Tabell 20: Oppgave 3 - Åpen oppgave - Snikfotografen

Oppgave 3 Skåring	På hvilken måte så folk annerledes ut på Carl Størmer bilder sammenlignet med typiske bilder fra 1890-årene?	8. trinn Antall	 Prosent	9. trinn Antall	 Prosent
1	Svar som er skåret til 1 poeng	548	27,4 %	736	36,8 %
0	Svar som er skåret til 0 poeng	1452	72,6 %	1264	63,2 %
Totalt		2000	100 %	2000	100 %

Analysen viser at 27, 4 % av elevene på 8. trinn og 36,8 % av elevene på 9. trinn har fått 1 poeng. Det vil si at det er henholdsvis 72,6 % på 8. trinn og 63,2 % 9. trinn som ikke har fått poeng på denne oppgaven. Vi vurderer denne oppgaven som en vanskelig oppgave, noe som gjenspeiles i den lave prosenten av elever som har fått poeng. Antallet som har klart oppgaven øker som forventet fra 27,4 % på 8. trinn til 36, 8 % på 9. trinn. Likevel er det 2716/4000 som ikke fikk uttelling i den åpne oppgaven.

4.2.3 Vurderingsveilederen til skåring av åpen oppgave – *Snikfotografen*

Tekst: SNIKFOTOGRAFEN

Oppgave 3: På hvilken måte så folk annerledes ut på Carl Størmers bilder sammenlignet med typiske bilder fra 1890-årene?

Oppgavens formål: *Å finne informasjon i teksten.*

1 poeng
Svar som viser til at folk ikke var så alvorlige på bildene, ELLER at bildene viste mennesker i mer naturlige situasjoner enn det som var vanlig.

- *Han kunne ta bilder av folk i situasjoner der de ikke var alvorlige.*
- *Det er mer glade mennesker på bildene, fordi de ikke vet at de blir tatt bilde av.*
- *Da kunne man se hvordan de hadde det og at dem ikke var alvorlige.*
- *De som det ble tatt bilde av, gjorde seg ikke til.*
- *Folk ser mer avslappet og glade ut siden de ikke er klar over at de blir fotografert.*
- *Smil og naturlig væremåte.* [Underforstått at dette ikke var vanlig.]
- *Mer avslappet* [Minimumssvar]

0 poeng
Svar som ikke viser til at folk var mindre alvorlige, eller at bildene viser mennesker i mer naturlige situasjoner. Andre vage, irrelevante eller gale svar.

- *De som ble tatt bilde av, visste det ikke så de ble annerledes.* [Gjentar deler av spørsmålet.]
- *De som blir tatt bilde av, vet det ikke selv.* [Sier ikke hvordan dette synes på bildene.]
- *Man kunne ta bilder uten at det handlet om selve fotograferingen.* [Sier ikke hvordan dette påvirket folk på bildene.]
- *Bildene fra kameraet gir inntrykk av at man ser gjennom et lite tittehull.*
- *Bildene var i svart og hvitt.*
- *Bildene er unike.*

Figur 10: Vurderingsveilederen - Snikfotografen

Vurderingsveilederen for *Snikfotografen* skiller mellom eksempler på svar som kan få 1 poeng, og svar som skal vurderes til 0 poeng. Elevsvar som viser til at folk ikke er så alvorlig eller at menneskene er i en mer naturlig situasjon skal skåres til 1 poeng. Vi har forstått vurderingsveilederen slik at svarene må inneholde en av eller begge disse momentene. Veilederen inneholder syv kulepunkter med eksempler på elevsvar som skal vurderes som riktige. Av de syv kulepunktene inneholder to av eksemplene ytterligere utdyping, i form av tilføyelser i klammer. Siste kulepunkt: «Mer avslappet» er minimumssvar.

For elevsvarene som skal skåres til 0 poeng er kriteriene at svarene ikke viser til at folk var mindre alvorlig eller at menneskene var i en mer naturlig situasjon, eller andre vage, irrelevante eller gale svar. Denne veilederen skiller seg fra *Stargate*-oppgaven ved at vi finner flere tilfeller av tilleggsinformasjon, presentert i klammeformer. Det er tre tilleggskommentarer for kriteriene for 0 poeng. Disse bidrar til å underbygge hvorfor elevsvaret ikke kan vurderes som tilfredsstillende. Vurderingsveilederen for *Snikfotografen* viser seks eksempler på elevsvar som ikke er tilfredsstillende, dette er to flere enn i veilederen til *Stargate*.

I hovedbeskrivelsen for hvilke svar som kan oppnå 1 poeng står det: «Svar som viser til at folk ikke var så alvorlig på bildene, ELLER at bildene viste mennesker i mer naturlige situasjoner enn det som er vanlig.» Til tross for dette er kulepunkt 7 «Mer avslappet [Minimumssvar]» godkjent som svar. Dette svaret er kun to ord og viser ikke til i hvilken grad folk fremstod som annerledes.

4.2.4 Resultater og funn fra kategorisering av åpen oppgave

I arbeidet med skåring av den åpne oppgaven til *Snikfotografen*, benyttet vi oss av samme metode som ved skåring av *Stargate*. Fra det opprinnelige utvalget på 4000 elevsvar, foretok vi et tilfeldig utvalg på 300 elevsvar for hvert av trinnene. Disse 600 elevsvarene ble skåret, kategorisert, og resultatene analysert.

Tabell 21: Beskrivelse av kategoriene

Kategori	Beskrivelse
A	Elevbesvarelsen vurderes som riktig i henhold til vurderingsveilederen
B	Elevbesvarelsen vurderes som uriktig i henhold til vurderingsveilederen
C	Elevbesvarelse som er vanskelig å skåre /Tvilstilfelle
D	Elevuttrykk
E	Blankt svar

Tabell 22: Fordeling av kategorisering Snikfotografen 8. trinn

Snikfotografen 8. trinn	S1 Cecilie		S2 Christina	
Kategori	Antall elevsvar	Prosent	Antall elevsvar	Prosent
A	63	21 %	82	27,33 %
B	147	49 %	143	47,67 %
C	24	8 %	9	3 %
D	22	7,33 %	23	7,67 %
E	44	14,67 %	43	14,33 %
Totalt	300	100 %	300	100 %

Tabell 23: Kategorisering Snikfotografen 9. trinn

Snikfotografen 9. trinn	S1 Cecilie		S2 Christina	
Kategori	Antall elevsvar	Prosent %	Antall elevsvar	Prosent %
A	96	32 %	102	34 %
B	98	32,67 %	112	37,33 %
C	24	8 %	4	1,33 %
D	13	4,33 %	13	4,33 %
E	69	23 %	69	23 %
Totalt	300	100	300	100

Tabellene 22 og 23 viser en oversikt over hvordan kategoriseringen fordelte seg totalt på utvalget for 8. trinn og 9. trinn. Resultatene etter skåring og kategorisering på 8. trinn viser at Cecilie har skåret 21 % av svarene til kategori A, i dette tilfellet 1 poeng siden oppgaven var rangert med en dikotom skala, 1 eller 0 poeng. Christina har vurdert 27 % til kategori A. Noe av denne differansen kan kanskje forklares med at Cecilie har et høyre antall av svarene på kategori C, tvilstilfelle. For 9.trinn er resultatet noe ulikt, Cecilie har kategorisert 32 % i kategori A og Christina 34 % til A. Vi ser også at antall elever som har fått 1 poeng øker fra 8. trinn til 9. trinn, fra 21-27 % på 8. trinn øker prosenten elever som har fått poeng til 32-34% på 9. trinn.

For 8. trinn hadde Cecilie kategorisert flere elevsvar til kategori B – galt svar i henhold til veilederen med 49 %, enn skårer 2, Christina med 47,67 %. Dette endret seg for 9.trinn. Det viste seg at Christina har kategorisert flere elevsvar til B, galt svar i henhold til vurderingskriteriene i veilederen med 37,33 % mot Cecilie som hadde plassert 32,67 % av elevsvarene i kategori B. Dette viser oss at vi kan ha byttet på å være «streng» i skåring. Cecilie viser seg å være den «strenge» for 8. trinn, mens Christina muligens øker kravene til 9. trinn og skårer flere elevsvar til 0 poeng. Det kan også være tilfeldig at elevsvarene fordelte seg slik, for det er ikke usannsynlig at 8. trinn i mindre grad mestret å svare godt på denne oppgaven enn 9. trinn. 9. trinn har et helt års skolegang mer enn 8. trinn. I tillegg kan en forvente at elevens modningsnivå kan ha hatt sin innvirkning på elevene. Likevel, er kategori B er det størst kategorien for begge årstrinn.

Skårer 1, Cecilie vært mest i tvil under skåringen og kategorisert av elevsvar på *Snikfotografen*. For 8. trinn er det 24 elevsvar, som utgjør 8 % og for 9. trinn samme antall og prosent. Skårer 2, Christina har et mindre antall i kategori C. For 8. trinn hadde Christina markert 3 % som tvilstilfeller og for 9.trinn kun 1,33 %. Vi ser at Cecilie har vært mest i tvil i skåringsprosessen. Elevsvarene fra tvilstilfellene fikk likevel poeng i skåring, og disse fordelte seg inn under både 1 og 0 poeng.

Kategori D viser oss hvor stor andel av elevsvarene som uttrykker noe, men ikke svarer på oppgaven. Kategori D har vi kategorisert med sammenfallende antall, 13 elevsvar, for 9. trinn. For 8. trinn er det nesten helt likt med henholdsvis 22 og 23 elevsvar i kategori D. Vi ser at andelen elever som skriver fritt i svarrubrikken til den åpne oppgaven er størst på 8. trinn med en prosent fra 7,33-7,67 % av 300. På 9. trinn er denne mindre med 4,33 %. Det kan indikere at det er flere elever som forsøker å svare på oppgaven på 8. trinn. Vi ser også at kategori E

som representerer blanke svar, er større på 9. trinn med 23 % av elevsvarene. For 8. trinn er det om lag 14 % av elevsvarene som er blanke – ubesvarte. Det forteller oss at det er flere elever på 9. trinn som har valgt å ikke svare på den åpne oppgaven til *Snikfotografen*, enn på 8. trinn. Etter en nærmere gjennomgang av skåringssettet fant vi et svar som bare besto av et punktum. Denne var på rad 158 8. trinn. Der har Cecilie valgt kategori D, elevuttrykk og Christina har kategorisert elevsvaret (punktumet) det til B, feilsvar. Det forklarer hvorfor det er en differanse mellom oss i kategori E 8. trinn.

4.2.5 Kvalitative funn i diskusjon fram til omforent skår

Fra utvalget på 600 elevbesvarelser for *Snikfotografen* fra 8. trinn og 9. trinn diskuterte vi oss frem til en omforent vurdering og skår. Vi sammenlignet skåringsresultatene og diskuterte oss frem til en omforent vurdering for elevsvarene med diskrepans. For den åpne oppgaven fra *Snikfotografen* var det relativt høy grad av samsvar i skåringen mellom oss som skårere. Vi samsvarte på 280/300 på 8. trinn og 289/300 på 9. trinn. Da vi hadde arbeidet fram et omforent svar og sammenlignet dette med de andre lærerne kom vi fram til at vi hadde diskrepans i 20/300 elevsvar på 8. trinn som utgjorde 6,3 %. På 9. trinn var den mindre, med 11/300 elevsvar, noe som utgjorde 4,6 % diskrepans. Det klare inntrykket etter denne prosessen var at vi opplevde færre uenigheter på 9. trinn. Vår omforente vurdering på 9. trinn la seg også tett opp mot lærer x sine vurderinger.

Som i *Stargate* vil vi i dette avsnittet beskrive fagsamtalene vi hadde i arbeidet med å komme fram til en omforent skår. Avsnittet under dokumenterer prosessen og reelle samtaler vi hadde. Synspunktene som kommer fram, vil hver for seg representere skitt i retning av en omforent forståelse- og dermed av en omforent skåring. I omtale av oss som uavhengige vurderere har vi også i dette datasettet brukt våre navn. Elevsitatene fra datasettet er markert i kursiv, dette for å synliggjøre sitatene i diskusjonen.

Det første vi la merke til i diskusjon fram til omforent skår, var elevsvar med ulike beskrivelser av hvordan folk så annerledes ut. Vi la merke til flere uttrykksmåter elevene brukte når de skulle beskrive at folk ikke var så alvorlige på bildene. Elevene brukte ord som «*realistisk, ikke så seriøs, alvorlig, livlig, så litt mer glad ut, ikke sur, mindre alvorlig og stresset».* Videre fant vi beskrivelser som «*De så normal ut, bildene var uforventet - så de så annerledes ut, morsomme og interessant, Ekte og følelsesfull».*
Første eksempel er elevsvaret på 8. trinn, rad 15: «*De var mer realistiske».* Cecilie hadde markert svaret som 0B, 0 poeng og galt i henhold til veilederen, kategori B. Cecilie

argumenterte med utgangspunkt i kulepunkt 7 «Mer avslappet (Minimumssvar) og kulepunkt 13 «Bildene var unike.». I «Det norske akademis ordbok» beskrives *realistisk* innenfor kunsthistorie og litteratur slik: «virkelighetstro, virkelighetsnær (i kunstnerisk fremstilling)». Christina hadde skåret elevsvaret til 1A, 1 poeng og riktig i henhold til vurderingsveilederen. Dette ble begrunnet med at Christina mente at ordet *realistisk* var det samme som autentisk eller virkelighetsnær. Vurderingsveilederens overordnede beskrivelse viser til «(…) mennesker i mer naturlige situasjoner enn det som var vanlig.». Cecilie argumenterte for om det å være realistisk virkelig er det samme som å være naturlig og avslappet? Vi valgte å gi elevsvaret 1 poeng, fordi vi ble enig om at det å være realistisk kan tolkes som å være virkelighetsnær.

På samme vis var elevsvarene på rad 154, 8. trinn: «*de var ikke så seriøse»* og rad 157, 8.trinn «*De var ikkje seriøse»* elevsvar vi hadde skåret ulikt. Disse elevsvarene ble tatt opp til ny vurdering. Her diskuterte vi om elevsvar som kunne ligne på vurderingsveilederens beskrivelse som nevner at «(…) folk ikke var så alvorlig på bildene (…)» kvalifiserte til 1 poeng. Kunne vi sidestille disse elevsvarene som likestilt med den overordnede beskrivelsen for 1 poeng? Cecilie hadde markert elevsvarene fra rad nr. 154 og 157 som 1C, 1 poeng og tvilstilfeller. Christina hadde markert begge elevsvarene som 1A- 1 poeng og riktig svar i henhold til vurderingsveilederen. I diskusjon fram til omforent skår la vi vekt på at ordet *seriøs* også kan bety alvorlig, høytidelig, oppriktig eller bokstavelig. Når eleven da svarer *«ikke så seriøs»* kan vi forstå det som at eleven tenker på det motsatte av seriøs eller alvorlig. Hvis vi tenker oss at det motsatte av alvorlig er glad, lett, naturlig og lignende, vil elevsvarene kunne falle inn under vurderingsveiledningens overordnede kriterier. Vurderingsveilederens overordene kriterier var «Svar som viser til at folk ikke var så alvorlig på bildene, ELLER at bildene viste mennesker i mer naturlige situasjoner enn det som var vanlig». Vi landet på å gi begge disse elevsvarene 1 poeng, men undres om vi kanskje gikk langt i å tilgodese eleven her. På den andre siden skal tvilen komme eleven til gode, så vi velger å stå godt i vår vurdering.

Et annet elevsvar som bruker verbet *å smile* for å forklare på hvilken måte folk så annerledes ut, er elevsvaret på rad 181, 8.trinn: *«Det som var annerledes var at det var sånn de så ut til vanlig, kanskje smilte de.»*

Dette elevsvaret hadde Cecilie skåret til 0C- 0 poeng, tvilstilfelle og Christina 1A – 1 poeng og riktig svar i henhold til veilederen. Tvilen for Cecilie dreide seg om at elevsvaret kunne oppleves som for vagt og litt *utydelig*. Samtidig kunne den første leddsetningen av svaret tolkes som at eleven har forstått at *«de»* [menneskene] befinner seg i en mer naturlig situasjon. Eleven formidler *at det var sånn de så ut vil vanlig,* dette tolket vi som ikke oppstilt og unaturlig.

I tillegg nevner eleven at de *kanskje smilte*. Spørsmålet blir om dette svaret samlet sett er nok sett opp mot kulepunkt 6 «Smil og naturlig væremåte. [Underforstått at dette ikke var vanlig.]» og 7. «Mer avslappet [Minimumssvar]». Den første delen av svaret sier ikke noen om hvordan bildene var sammenlignet med typiske bilder på 1890-tallet. Da må det være siste delen av svaret som omtaler at de *kanskje smilte* som kvalifisere til poeng. I tillegg er det i henhold til veilederen problematisk at eleven gjentar deler av spørsmålet. Kulepunkt 8: «*De som ble tatt bilde av, visste det ikke så de ble annerledes*. [Gjentar deler av spørsmålet.] Vurderingsveilederen bruker dette elevsvaret som eksempel på svar som skal skåres til 0 poeng. Christina la vekt på *kanskje smilte* som avgjørende for poeng. I tillegg sier retningslinjene for skåring i PAS[11] i forkant av skåringen at lærerne skal la tvilen komme elevene til gode. Omforent skår ble satt til 1 poeng.

Den andre grupperingen av elevsvar som hadde likheter, handlet om fotograferingssituasjoner. Elevsvaret på rad 119: «*ingen av de så på kameraet eller poserte».* Dette elevsvaret hadde Christina markert som 1C, 1 poeng, men tvilstilfelle, mens Cecilie hadde markert elevsvaret 1A -1 poeng og riktig svar i henhold til veilederen. Tvilen for Christina sin del handlet om det å ikke se på kameraet, kan være det samme som å være i en naturlig situasjon. Vi kom fram til at eleven kan ha tenkt, at når du ikke poserer – ser du naturlig ut. Eleven fikk 1 poeng. Cecilie trakk frem at dette elevsvaret uttrykker en forståelse og beskrivelse av hvordan folk opptrådte under fotografering. Et annet elevsvar vi var i tvil om, var elevsvaret på rad 60, 9. trinn: «*De poserte ikke*». Cecilie hadde markert dette svaret som 1C og Christina som 1A. Ordet *posere* var ikke nevnt som eksempel i vurderingsveilederen. Vi var usikre på om vi gikk for langt i å tolke dette elevsvaret i elevens

[11] Se figur 1, s. 16

favør. Vi diskuterte oss fram til at det å *ikke* posere, kan forståes som mer avslappet og naturlig. I omforent skår ble vi derfor enig om å gi elevsvaret 1 poeng.

Ut fra fotograferingskategorien fant vi en gruppe elevsvar som la vekt på det fototekniske, som hvordan kameraet til Carl Størmer så ut, til eksempel; «*ikke profesjonelt kamera, spionkamera, han så gjennom et lite tittehull».* Felles for denne gruppen elevsvar, var at de har hentet ut feil informasjon i teksten. Det viser flere av elevsvarene fra 8. trinn, for eksempel besvarelsen fra rad 29: *«Det så ut som man så gjennom et tittehull»*, rad 30: *«ser ut som om man titter fra et kikkehull og spionerer»* og rad 48*: «at han han hadde et lite kamera som kunne skjules isteden for et stort kamera».* Som vi trakk frem i gjennomgangen av teksten er det gjort endringer i prøveteksten fra originalen. Elevsvarene over viser til informasjon i teksten som vist i bilde til høyre figur 10. En av bildetekstene referer til hvordan spionkameraet til Carl Størmer så ut. Den andre bildeteksten er relativ lang og viser til: «*Bildene fra kameraet gir inntrykk av at man ser gjennom et lite tittehull. Fordi kameraet er skjult og bildene har svarte kanter rundt, føles det som om man spionere på virkeligheten den gangen (...)*» Det var relativt mange elever som har forstått dette som riktig uthenting av informasjon. Andre elevsvar av samme type var: «*De så annerledes ut fordi på 1890 tallet var det svarthvit bilder som ikke hadde noe farge.»* og *«De så annerledes ut fordi bildene han to var rund og ikke firkantete*.»

Avslutningsvis vil vi trekke frem eksempler på elevsvar der vi endret mening fra 1 poeng til 0 poeng etter diskusjon fram til omforent skår. Et eksempel på elevsvar som fikk ny vurdering var rad 225, 8. trinn: «*I bildene Stormers tok var ikke menneskene klare over at de ble fotografert og derfor fikk vi et tydligere bilde av hvordan folkene gikk rundt i oslo for flere år siden».*

Her var elevsvaret skåret til 1C – 1 poeng og tvilstilfelle og 1A- 1 poeng og rett i henhold til vurderingsveilederen. Vi fant dette elevsvaret vanskelig å skåre fordi ingen av eksemplene i vurderingsveilederen var sammenfallende med svaret. I tillegg mener vi at det ikke kommer frem noen om hvordan folk så annerledes ut. Det kan også argumenteres for at eleven har riktig forståelse for den uformelle lokaliseringen i Oslos gater. Artikkelen fremhever at det ikke var mulig i en slik uformell setting som Carls Størmer fanget i sine bilder. Til tross for dette ble omforent skår satt til 0 poeng, da svaret ikke tydelig uttrykker at denne lokasjon var mer naturlig eller noen om hvordan det påvirket de som ble tatt bilde av.

På samme måte var elevsvaret på rad 289, 8. trinn et elevsvar vi endret mening om i diskusjon: *«de var ikke forbrett på å bli tatt bildet av».* Det samme gjaldt elevsvaret på rad 97, 9. trinn: *«På Carl Stormers bilder var folk ikke forbrett på å bli fotografert og de ble tatt bilde av mens de gikk på gata. På bilder fra 1890 var folk forbrett på å bli tatt bilde av».*

For begge disse elevsvarene hadde Cecilie skåret 1C og Christina 1A. Felles for alle de tre elevsvarene vi har fremhevet er at de legger vekt på omgivelsene i større grad enn hvordan folk så ut. Er det mulig at noen av disse elevsvarene utrykker seg innforstått, men at det er utfordrende å vurdere. Spørsmålet legger vekt på at elevene skal svare «På hvilken måte så folk annerledes ut (…)», og ber elevene sammenligne Carl Størmers bilder med typiske bilder fra 1890-årene. Det å ikke være forberedt på å bli tatt bilde av, sier ikke noe om på hvilken måte de så annerledes ut. Dette samsvarer derfor ikke med veilederes krav om 1 poeng. Derfor endret vi mening i vår omforente skår og vurderte elevsvarene til 0 poeng. I det neste avsnittet vil vi se på reliabilitetsmålene for skåring og poenggiving.

4.2.6 Kvantitativt resultat fra skåring 8. trinn og 9. trinn

For å besvare forskningsspørsmålet som omhandler reliabilitet i skåring av de åpne oppgavene, har vi valgt å sammenligne skåringsresultatene i antall, prosent og beregnet Cohens kappa og Fleiss kappa. Tabellen under viser hvor mange poeng vi delte ut i skåring. Lærer x representerer den opprinnelige skåringen lærerne har tildelt elevsvarene i vårt datasett.

Tabell 24: Oversikt skåringsresultater 8. trinn

Snikfotografen 8. trinn	S1 Cecilie	S2 Christina	Omforent skår	Lærer x
0	219	211	222	215
1	81	89	78	85
Antall skåringer	300	300	300	300
Total poengsum	81	89	78	85

Tabell 25: Oversikt skåringsresultater 9. trinn

Snikfotografen 9. trinn	S1 Cecilie	S2 Christina	Omforent skår	Lærer x
0	192	195	199	197
1	108	105	101	103
Antall skåringer	300	300	300	300
Total poengsum	108	105	101	103

Tabellene viser en kvantitativ fremstilling, gjennom opptelling av hvor mange elevsvar vi har skåret til 0 eller 1 poeng. Vi er oppmerksomme på at denne fremstillingen ikke sier noe om hvorvidt det er de samme elevsvarene som er skåret til 1 eller 0. Den har derfor sine begrensninger som reliabilitetsmål, og vi foretar en analyse av dette i neste steg. Vi velger å ha oversikten med for å diskutere antall og hovedtendenser mellom 8 og 9. trinn og oss som vurderere.

For 8. trinn viser resultatet at vi i omforent skår har vurdert elevsvarene noe «strengere», enn hver for oss. Det er altså færre elever som får tildelt 1 poeng etter fagsamtalene vi hadde. Omforent skår har skåret 222 elevsvar til 0 poeng, og lærer x har skåret 215 elevsvar til 0 poeng. Det viser at våre individuelle skåringer er nærmere lærer x i vurdering og skåring enn

vår omforente skår. Resultatene for 9.trinn viser mindre differanser mellom den omforente skåren og lærer x. For elevsvar som er skåret til 0 poeng er antallet relativt likt. Vi ser at vi også på 9. trinn har gått tilbake på utdeling av poeng i omforent vurdering, sammenlignet med våre individuelle skåringer. For omforent skår på 8. trinn er det bare 78/300 elevsvar som får poeng. På 9. trinn er det 101/300 (33,67 %) elevsvar som får poeng i skåring av *Snikfotografens* åpne oppgave. Vi reflekterer rundt hvilke faktorer dette kan skyldes og vil drøfte dette i kap. 5.

4.2.6.1 Cohens kappa 8. trinn - *Snikfotografen*

Tabellen under viser Cohens Kappa beregning for interrater reliabilitet i vurdering av den åpne oppgaven til *Snikfotografen* på 8.trinn og 9. trinn. Som omtalt i metodekapittelet kontrollerer Cohens kappa mål på enighet mellom to vurderere (raters) – interrater reliabilitet (McHugh, 2012). De to følgende tabellene viser utregningene for Cohens kappa for 8. trinn skårer 1 Cecilie opp mot skårer 2 Christina, og omforent skår opp mot lærer x.

Tabell 26: Skårer 1 Cecilie og skårer 2 Christina

S1 * S2 Crosstabulation

			S2		
			0	1	Total
S1	0	Count	205	14	219
		% within S1	93,6%	6,4%	100,0%
		% within S2	97,2%	15,7%	73,0%
		% of Total	68,3%	4,7%	73,0%
	1	Count	6	75	81
		% within S1	7,4%	92,6%	100,0%
		% within S2	2,8%	84,3%	27,0%
		% of Total	2,0%	25,0%	27,0%
Total		Count	211	89	300
		% within S1	70,3%	29,7%	100,0%
		% within S2	100,0%	100,0%	100,0%
		% of Total	70,3%	29,7%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T [b]	Approximate Significance
Measure of Agreement	Kappa	,836	,035	14,511	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Krysstabell nr. 22 viser hvordan vi samsvarer i skåring av Snikfotografen 8. trinn. Vi er enig i 205 elevsvarene skal ha 0 poeng, og 75 elevsvar skal ha 1 poeng. I 20 elevsvar har vi vurdert ulikt. Kappaverdien for dette beregningen gir et mål på interrater reliabilitet ,836. Dette kan anses som høyt. Neste beregning var omforent skår opp mot lærer x.

Tabell 27: Lærer x og omforent skår

Lx * Omforent Crosstabulation

			Omforent		
			0	1	Total
Lx	0	Count	210	5	215
		% within Lx	97,7%	2,3%	100,0%
		% within Omforent	94,6%	6,4%	71,7%
		% of Total	70,0%	1,7%	71,7%
	1	Count	12	73	85
		% within Lx	14,1%	85,9%	100,0%
		% within Omforent	5,4%	93,6%	28,3%
		% of Total	4,0%	24,3%	28,3%
Total		Count	222	78	300
		% within Lx	74,0%	26,0%	100,0%
		% within Omforent	100,0%	100,0%	100,0%
		% of Total	74,0%	26,0%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T [b]	Approximate Significance
Measure of Agreement	Kappa	,857	,034	14,868	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Krysstabellen for omforent skår opp mot lærer x, sammenlignet med Cecilie (S1) og Christina (S2) viser at den omforente skåren og lærer x har bedre samsvar i skåring av de samme

elevsvarene. Det kan skyldes at flere elevsvar har fått 0 poeng. Cecilie og Christina har gitt 205 elevsvar 0 poeng uavhengig av hverandre, mens den omforente skåren opp mot lærer x øker andelen elevsvar som har fått 0 poeng til 210. Det er også et lavere antall elevbesvarelser i den siste krysstabellen som viser diskrepans i skåring av 1 eller 0 poeng, 17 mot 20 for S1 og S2.

Målet for samlet interrater-reliabilitet mellom Cecilie (S1) og Christina (S2) gir kappaverdien ,836 på 8. trinn og dette indikerer høy interrater reliabilitet eller samsvar i skåring mellom Cecilie og Christina for elevsvar på den åpne oppgaven i *Snikfotografen*, 8. trinn (Landis & Koch, 1977, s. 165). Likevel viser kappaverdien for omforent skår opp mot lærer x er bedre kappaverdi på ,857.

4.2.6.2 Cohens kappa 9. trinn - *Snikfotografen*

Tabell 28: Skårer 1 Cecilie og skårer 2 Christina

S1 * S2 Crosstabulation

			S2		
			0	1	Total
S1	0	Count	188	4	192
		% within S1	97,9%	2,1%	100,0%
		% within S2	96,4%	3,8%	64,0%
		% of Total	62,7%	1,3%	64,0%
	1	Count	7	101	108
		% within S1	6,5%	93,5%	100,0%
		% within S2	3,6%	96,2%	36,0%
		% of Total	2,3%	33,7%	36,0%
Total		Count	195	105	300
		% within S1	65,0%	35,0%	100,0%
		% within S2	100,0%	100,0%	100,0%
		% of Total	65,0%	35,0%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T[b]	Approximate Significance
Measure of Agreement	Kappa	,920	,024	15,938	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Cohens kappa for 9. trinn Cecilie og Christina viser en verdi på ,920. Dette viser sterk grad av samsvar jmf. McHuch (2012) og Koch og Landis (1977). Vi observerer at 188/300 elevsvar er skåret til 0 poeng samsvarende, mens 101 elevsvar er skåret samsvarende til 1 poeng.

Tabell 29: Lærer x og omforent skår

lærerx * omforent Crosstabulation

			omforent		
			0	1	Total
lærerx	0	Count	191	6	197
		% within lærerx	97,0%	3,0%	100,0%
		% within omforent	96,0%	5,9%	65,7%
		% of Total	63,7%	2,0%	65,7%
	1	Count	8	95	103
		% within lærerx	7,8%	92,2%	100,0%
		% within omforent	4,0%	94,1%	34,3%
		% of Total	2,7%	31,7%	34,3%
Total		Count	199	101	300
		% within lærerx	66,3%	33,7%	100,0%
		% within omforent	100,0%	100,0%	100,0%
		% of Total	66,3%	33,7%	100,0%

Symmetric Measures

		Value	Asymptotic Standard Error [a]	Approximate T [b]	Approximate Significance
Measure of Agreement	Kappa	,896	,027	15,521	<,001
N of Valid Cases		300			

a. Not assuming the null hypothesis.

b. Using the asymptotic standard error assuming the null hypothesis.

Cohens kappa for 9. trinn lærer x og omforent skår viser en verdi på ,896. Dette viser sterk grad av samsvar jmf. McHuch (2012) og Koch og Landis (1977). Vi observerer også her at 191/300 elevsvar er skåret til 0 poeng samsvarende, mens 95 elevsvar er skåret samsvarende til 1 poeng.

4.2.6.3 Fleiss kappa 8. trinn – *Snikfotografen*

Tabell 30: Fleiss kappa 8. trinn

Overall Agreement[a]

	Kappa	Asymptotic Standard Error	Asymptotic z	Asymptotic Sig.	Asymptotic 95% Confidence Interval Lower Bound	Asymptotic 95% Confidence Interval Upper Bound
Overall Agreement	,825	,033	24,747	<,001	,760	,890

a. Sample data contains 300 effective subjects and 3 raters.

Agreement on Individual Categories[a]

Rating Category	Conditional Probability	Kappa	Asymptotic Standard Error	Asymptotic z	Asymptotic Sig.	Asymptotic 95% Confidence Interval Lower Bound	Asymptotic 95% Confidence Interval Upper Bound
0	,950	,825	,033	24,747	<,001	,760	,890
1	,875	,825	,033	24,747	<,001	,760	,890

a. Sample data contains 300 effective subjects and 3 raters.

Fleiss kappa viser grad av enighet mellom Cecilie, Christina og lærer x. Samlet viser Fleiss kappa verdiene på ,825 at det er sterk grad av samsvar i skåring av de 300 elevbesvarelsene fra den åpne oppgaven i Snikfotografen jmf. McHuch (2012) og Koch og Landis (1977).

4.2.6.4 Fleiss kappa for 9.trinn – *Snikfotografen*

Tabell 31: Fleiss kappa for 9. trinn

Overall Agreement[a]

	Kappa	Asymptotic Standard Error	Asymptotic z	Asymptotic Sig.	Asymptotic 95% Confidence Interval Lower Bound	Asymptotic 95% Confidence Interval Upper Bound
Overall Agreement	,902	,033	27,074	<,001	,837	,968

a. Sample data contains 300 effective subjects and 3 raters.

Agreement on Individual Categories[a]

Rating Category	Conditional Probability	Kappa	Asymptotic Standard Error	Asymptotic z	Asymptotic Sig.	Asymptotic 95% Confidence Interval Lower Bound	Asymptotic 95% Confidence Interval Upper Bound
0	,966	,902	,033	27,074	<,001	,837	,968
1	,937	,902	,033	27,074	<,001	,837	,968

a. Sample data contains 300 effective subjects and 3 raters.

Resultatene for Fleiss kappa beregningen viser «nesten perfekt» grad av samsvar i skåring med et kappamål på ,902 som ligger innenfor «almost perfect» ifølge kategorien til McHuge (2012).

4.2.7 Drøfting av hovedfunn - *Snikfotografen*

I arbeidet med *Snikfotografen*, den åpne oppgaven, vurderingsveilederen og skåring har vi forsøkt å oppnå oversikt og innsikt. Hvordan påvirker de ulike delene inn på validiteten og reliabiliteten i arbeidet med å skåre i de åpne oppgavene? I det følgende vil vi drøfte funn som peker seg spesielt ut. Det vil si sentrale faktorer for validiteten i målingen av lesekompetansen, og hva som kan påvirke reliabiliteten i skåringen av den åpne oppgaven.

Vi undrer oss over hvorfor så få elever har fått 1 poeng på denne oppgaven, ettersom vi observerte at flere elever har klart å besvare flervalgsoppgavene 1 og 2, tabell 18 og 19. På flervalgsoppgavene 1 og 2 er det mellom 61,9 % til 70 % som har valgt rett svaralternativ. Roe og Blikstad-Balas referer til at et aspekt ved god validitet i leseprøver er at den skiller mellom gode og dårlige lesere (Roe & Blikstad-Balas, 2022, s. 221). Vi undres over hvorvidt den åpne oppgaven til *Snikfotografen* er konstruert vanskelig nettopp av denne grunn (Fevolden, Thorsen, & Eriksen, 2019, s. 9). I likhet med *Stargate* er den åpne oppgaven til *Snikfotografen* også plassert på mestringsnivå 4. Mestringsnivå 4 er det nest høyeste i prøvesettet. Gjennom vår tilgang til PAS fant vi en oversikt over den nasjonale løsningsprosenten på den åpne oppgaven til *Snikfotografen*. Den viste at 25 % av elevene på 8.trinn hadde fått poeng på den åpne oppgaven, og for 9. trinn var det 38 % av elevene. Dette samsvarer omtrentlig med resultatene fra vårt skåringsutvalg. I vår omforente skår har 22% av elevene på 8. trinn og 33,6 % av elevene på 9.trinn, fått poeng på oppgaven. Vi har diskutert hvorvidt disse funnene kan indikere at mange elever leser og forstår *Snikfotografen*, men at det kan være faktorer ved det åpne spørsmålet som gjør at mange ikke har klart å besvare oppgaven tilfredsstillende.

Et aspekt ved *Snikfotografen* som vi har trukket frem i tekstgjennomgang er, hvor tydelig kommer det frem at menneskene var annerledes? Oppgaven har som formål å måle leseprosessen «å finne informasjon» knyttet opp mot mestringsnivå 4. I rammeverket er denne beskrevet som: «Lokalisere og kombinere informasjon fra ulike steder i en eller flere tekster og vurdere hvilken informasjon som er relevant». Den samme leseprosessen er på mestringsnivå fem beskrevet som: «Lokalisere og kombinere informasjon – ofte implisitt eller

godt skjult – fra ulike steder i én eller flere tekster. Å skille relevant informasjon fra sterkt konkurrerende informasjon». Vi vurderer at den åpne oppgaven til *Snikfotografen* kunne ha vært plassert under mestringsnivå fem. Dette begrunnes med at elevene må kombinere informasjon fra ulike steder i teksten for å finne svar på oppgaven. I den første delen av teksten blir de introdusert for skjult kamera og gatefotografen som fanget øyeblikk. Neste avsnitt med overskriften «Kulturhistorisk vendepunkt» er sentralt, men informasjonen er også noe implisitt, som setningen: «Kameraet ga en ny form for bilder, og endret selve fotografiets natur». I neste del av teksten handlet det om settingen bildene ble tatt i, og at Størmer tok bilder uten at det handlet om selve fotograferingen. Elevene må både lokalisere og kombinere dette, i tillegg til at de må forstå tekst som implisitt uttrykker hva som gjorde folk annerledes.

Roe og Blikstad-Balas påpeker at de ulike lesemåtene ikke er mulig å måle atskilt. Som leser er en avhengig av et sett med ulike strategier for å lese tekst med forståelse (Roe & Blikstad-Balas, 2022, s. 211). I startfasen av masterprosjektet hadde vi en antakelse om at leseformålet «å finne informasjon» skulle være relativt greit, og tenkte nok at det gikk på ren innhenting av tydelig og eksplisitt uttrykt meningsinnhold. I etterkant ser vi at den åpne oppgaven til *Snikfotografen* er et eksempel på at det innhenting av informasjon kan være mer sammensatt.

En av årsakene kan være formuleringen av spørsmålet til den åpne oppgaven. Under gjennomgangen av det åpne spørsmålet har vi trukket frem hvordan forståelsen og tolkningen av ordet annerledes anses et viktig aspekt. Elevene må forstå at «annerledeshet» handler om væremåten og sinnsstemningen til de som ble fotografert. Vi har undret oss over hvorvidt alternative formuleringer av det åpne spørsmålet kunne ha gjort det tydeligere for eleven hva spørsmålet søker svar på. Et av våre forslag er: «Hvorfor så folk annerledes ut på bildene Carl Størmers tok sammenlignet med typiske bilder fra 1890-årene», eller: «Kan du forklare hvorfor menneskene i Carl Størmer bilder ser annerledes ut enn det bildene som var typisk for bilder som ble tatt på 1890-tallet?». På den andre siden, kan det være tilsiktet fra prøveutviklerne, nettopp for å skille mellom lesenivåer. Vi har fundert over om en tilleggsopplysning som beskriver fremgangsmåte og omfang, kunne fått opp svarprosent og kvaliteten på svarene. En slik tekst kunne har vært: «I denne oppgaven skal du lese nettartikkelen *Snikfotografen* og finne informasjon fra et eller flere steder i teksten. Svaret ditt bør være fra 1 til 4 setninger». For noen elever kunne det vært avklarende i prøvesituasjonen, og kanskje fått flere til å forsøke?

Roe og Blikstad-Balas omtaler validitet som et mål for i hvilken grad prøven måler det den sier at den måler, lesing (Roe & Blikstad-Balas, 2022, s. 221). Kan manglende motivasjon, og at oppgaven ikke oppleves som meningsfull eller relevant, vises som blanke svar eller elevuttrykk. Dette kan være en trussel mot validiteten i den grad elevene ikke svarer på den åpne oppgaven, som resultat blir elevene utilgjengelig for måling og vurdering (Cohen, Manion, & Morrison, 2007, s. 159). Resultatene fra skåringen og kategoriseringen viste at mange elever hadde svart blankt, og ulike elevutrykk. For å undersøke hva elevene som fikk 0 poeng hadde svart feil, gjorde vi et søk i det opprinnelige datasettet. For 8. trinn inneholdt 70 elevsvar ordet *tittehull* og *titehull*. Av disse var det 2 svar som kvalifisere til 1 poeng, for 9. trinn var det 95 svar som innehold det samme ordet. Felles for svarene var beskrivelser som; «De så ut som man ser gjennom et tittehull og at de var tatt med spionkamera». Svarene fokuserte også på det fototekniske som, skarpere bilder og sort kant. Vi tror at elevene har funnet dette svaret under bildeteksten til figur 10 omtalt i tekstgjennomgangen.

Vurderingsveilederen til *Snikfotografen* skilte seg fra *Stargate* ved at den graderte svarene i en dikotom skala. Tatt dette i betrakting kunne vi kanskje forvente at tolkningsrommet var mindre og at det var lettere å vurdere elevsvarene, også med tanke på at oppgavene sitt formål «finn informasjon». De kvantitative resultatene viser at det var mindre diskrepans i vurdering og skåring av den åpne oppgaven for *Snikfotografen*, sammenlignet med *Stargate*. Resultatene fra Cohens og Fleiss kappa viste høy grad av enighet om hvilke elevsvar som skulle ha 0 poeng. For 8. trinn oppdaget vi at vår omforente skår faktisk hadde blitt «strengere» elevsvar som fikk 1 poeng var færre. Kan vi selv ha gått i den testteoretiske rottefella, og forholdt oss stramt til vurderingsveilederen? (Skaftun, 2006, s. 40). Dette vil vi komme tilbake til i avsluttende drøfting, «Tolkningsrommet og lærerens skjønn».

I fagsamtalene til omforent skår dreide diskusjonen seg i stor grad om minimumssvaret «Mer avslappet [minimumssvar]» egentlig åpnet opp for at flere elevsvar kunne fått svaret sitt godkjent. Klammeformene som kommenterte eksemplene på elevsvar som skulle vurderes til 0 poeng, var noe mer oppklarende. Som vist i våre funn fra omforent diskusjon hadde elevene mange ulike måter å forklare på hvilken måte folk så annerledes ut. Vurderingsveilederen til *Snikfotografen* var noe enklere å forholde seg til som et resultat av at det var en dikotom skala. Likevel opplevde vi at minimumsvaret «mer avslappet» åpnet opp for at flere elevsvar kunne få poeng. Dermed ble vår erfaring at denne veilederen heller ikke var helt entydig og det oppstod rom for tolkning, som også i Skaftun (Skaftun, 2006). Vi er ydmyke for at det er

vanskelig å finne eksakte svar på hvorfor denne oppgaven har så lav svarprosent, men gjennom vår drøfting har vi belyst flere momenter som kan være årsak.

4.2.8 Hovedfunn for kategoriene - *Stargate* og *Snikfotografen*

Formålet med kategoriene var å favne grupperinger av elevsvar opp mot kriteriene i vurderingsveilederen, samt sortere ut tvilstilfellene. I utgangspunktet hadde vi nok trodd at kategori C – tvilstilfelle skulle være en større kategori enn den viste seg å være. I forkant hadde vi forventet at diskrepansen mellom oss i hovedsak skulle være knyttet til denne kategorien. Cecilie kategoriserte flere elevsvar til C, tvilstilfeller med 72 av 1200 elevsvar. Christina kategoriserte 22 av 1200 som tvilstilfeller. Det gjenspeiler vår subjektive opplevelse i kategorisering og skåring. I fasen hvor vi sammenlignet skåringene oppdaget vi at ulikhetene oftest omhandlet elevsvar kategorisert som A og B, og diskusjonen handlet om 1, 2 eller 0 poeng. Dette kan forklares med ulik tolkning av elevsvar opp mot vurderingsveilederen.

Et resultat vi ønsker å trekke frem er den registrerte økningen i antall blanke svar, markert som kategori E. Ved hjelp av kategorisering av alle elevsvarene vi skåret, oppdaget vi at andelen blanke elevsvar på de åpne oppgavene økte fra 8. trinn til på 9. trinn. For *Stargate* økte det fra 2,66 % til 10,33 %, og for *Snikfotografen* fra 14,33 % til 23 %. Dette var litt overraskende, for vi hadde nok forventet at antallet blanke svar skulle synke på 9.trinn. Tallene viser at flere elever på 9. trinn gir akseptable svar på den åpne oppgaven, men flere svarer ikke. I utgangspunktet, sett opp mot at de er mer modne og har gått ett år lengre på skole, skulle flere ha mestret å svare på dette. Vi undrer oss over hvilke faktorer det kan skyldes. Handler det om oppgavens validitet – gitt at oppgavens ordlyd ikke «fenger» elevene, eller reliabilitet – at elevene av ulike årsaker ikke svarer på oppgaven som motivasjon, datatrøbbel eller andre faktorer? Vi kan ikke påvise noe av dette gjennom vårt prosjekt, men velger å trekke det frem.

5 Tolkningsrommet og lærerens skjønn

I dette kapittelet vil vi drøfte hvordan tolkningsrommet og lærerens skjønn står fram som viktige faktorer i forholdet mellom validitet og reliabilitet i de åpne oppgavene. Står vi ovenfor et testteoretisk dilemma? Kapittelet løfter hovedmomenter i våre samlede funn.

5.1.1 Problemet med tolkningsrommet

I en autentisk skåringssituasjon skal lærerne lese retningslinjene[12] for skåring, alle prøvetekstene, og i tillegg studere alle vurderingsveilederne. I sum utgjør dette mange dokumenter. Før lærerne setter i gang med skåring presenteres det i retningslinjene at: «Er du i sterk tvil, skal tvilen komme eleven til gode (…)». I retningslinjene informeres det også: «Dersom det gale står i motsetning til det riktige, skal svaret ikke godkjennes». På neste punkt åpnes det opp for at elevsvar kan godkjennes, dersom de inneholder en ukorrekt tilleggsinformasjon. Det kan oppleves forvirrende når retningslinjene åpner opp for lærerens faglige skjønn, men vurderingsveilederen lukker det noe. I presentasjon av kvalitative funn og tidligere drøfting viser vi til kulepunkt 4 og 8 i *Stargate*. I vår iver etter å skåre «så riktig som mulig» ble vi kanskje noe opphengt i kulepunktene i vurderingsveilederen.

I arbeidet med dette masterprosjektet kom vi til en fase hvor vi så tilbake på egen skåring, og stilte spørsmål til noen av våre skåringsavgjørelser. Vi løftet et kritisk blikk på oss selv som skårere. Vi reflekterer over om eleveksemplene i vurderingsveilederne la sterke føringer på skåringen vår. En diskusjon som gjenspeiler dette, vurderingen av de ulike verbene elevene brukte for å beskrive at faren til jenta hadde fått jobben. Førte dette til at skåringen av elevsvarene stod i fare for å bli veldig instrumentell, og ikke så fleksibel og funksjonell (Matre, et al., 2021, s. 269)? Fagsamtalen til omforent skår burde i større grad dreid seg om leseformålet. Hadde elevene tolket og trukket slutninger på bakgrunn av informasjon i teksten? Ensidig fokusering på detaljer kan hindre anbefalingen fra retningslinjene om «å la tvilen komme elevene til gode».

Våre kvantitative funn kan også indikere at denne såkalte instrumentelle vurderingen kom til uttrykk. Ble vi strengere i omforent skår? Et eksempel var S*nikfotografen* 8. trinn hvor vi

[12] Se figur 1, s. 16 - kap. 2.3.1

tildelte færre poeng enn vi hadde gjort hver for oss. Resultatene fra tabell 13 viser at Cecilie hadde skåret 81 svar til 1 poeng, Christina 89 til 1 poeng, mens i omforente skår tildelte vi 78 elevsvar 1 poeng. Skåringen i *Stargate* stod i motsetning til dette, hvor vi endte med å tildele flere poeng omforent. Som vist i teorikapittelet kan vurderere befinne seg på ulike stadier i vurderingsarbeid. Sensorer kan være strenge i starten av vurderingsarbeidet, og mildere i senere stadier av vurderingsarbeidet (Cohen, Manion, & Morrison, 2007, s. 159). Når vi over peker på at vi ble noen strengere, tror vi den omforente vurderingen kan ha avdekket noen svakheter i individuell vurdering. I fagsamtalene våre ble fokuset rettet mot diskrepansen, og det er viktig å påpeke at de ikke utgjorde så stor andel. Resultatene for interrater reliabilitet viste kappaverdier fra laveste verdi på ,76 til ,92. Disse verdiene strekker seg fra «moderat» til «nesten perfekt» (McHugh, 2012). De åpne oppgavene fra Stargate og Snikfotografen vi har studert i vårt masterprosjekt hadde lav løsningsprosent nasjonalt[13]. I våre diskusjoner har vi reflektert over hva som kan ha gjort disse to oppgavene utfordrende, og i hvilken grad de målte leseforståelsen som var tiltenkt.

5.1.2 Er de åpne oppgavene åpne nok?

Underveis i prosjektet diskuterte vi hvorvidt spørsmålsformuleringene til de åpne oppgavene kunne ha vært åpnet mer opp? Spesielt synes vi at *Stargate* var en tekst med gode kvaliteter, og med mulighetsrom for dette. Med gode konstruerte oppgaver og veiledere som ikke lukker den semantiske åpenheten, gir det elevene mulighet til å vise større grad av selvstendighet og refleksjon. Oppgavene kan åpne opp for å ikke bare ha ett riktig svar, og elevene kan komme med sine egne meninger og selvstendige refleksjoner (Solheim & Skaftun, 2009: Roe, Ryen og Weyergang, 2018). Flere peker på mulighetsrommet i de åpne oppgavene. I sin «Rapport fra pilotering og gjennomføring av den nasjonale prøven i lesing på 8. og 9. trinn 2018» påpeker Fevolden et al at «Ved å ta med en åpen oppgave knyttet til hver tekst, inviteres elevene til å uttrykke sin egen mening, og lærerne får anledning til å bli kjent med tekstenes innhold når de vurderer elevsvarene (Fevold, Thoresen, & Eriksen, 2019, s. 5)». Vi undres over om dette egentlig stemmer. Våre kvalitative funn viser at den tiltenkte modelleseren, uttrykt gjennom vurderingsveilederen, egentlig ikke tillater så mange ulike lesemåter og tolkninger. Vi tror at det kan skyldes at prøvekonstruktet ønsker å ha høy reliabilitet i vurdering av elevsvarene. Hvis man legger stor vekt å oppnå svært høy reliabilitet, kan dette

[13] Nasjonal løsningsprosent er tilgjengelig i PAS.

være en trussel mot validiteten. Da tester man en smalere del av lesekompetansen i prøven (Roe, Ryen, & Weyergang, 2018, s. 178). Kan dette ha skjedd her, eller var det tilsiktet fra prøveutviklerne for å skille mellom svake og sterke lesere?
I datasettet vi fikk utlevert fra Utdanningsdirektoratet la vi merke den åpne oppgaven til «Den korte historien om egget». Formålet var også her å hente ut informasjon fra teksten,

Nasjonale prøvene skal være utviklet for å måle lesing med høy validitet og reliabilitet. For at prøvene skal ha høy grad av validitet måles leseferdighetene gjennom teoretiske komplekse beskrivelser, og operasjonaliseringer. I tillegg til å ivareta det testteoretiske idealene gjennom entydighet og enkelhet (Skaftun, 2006, s. 92). Vi forstår at det er avveininger mellom hvor enkelt prøvene skal være konstruert for å ivareta både validitet og reliabilitet. Er det slik at for å oppnå de testteoretiske kravene i form av kvantitative verdier, mister vi noen muligheter?

I en hektisk skolehverdag får et utvalg lærere tilleggsoppgaven som skåring av åpne oppgaver er. Å foreta skåringer kommer ofte på toppen av andre arbeidsoppgaver, og kan belaste lærerne dersom det ikke settes av tid til arbeidet. Våre erfaringer er at læreren ofte er alene i skåringsprosessen og primært avhengig av vurderingsveilederens retningslinjer og eksempler. Dette kan utfordre reliabiliteten i skåringen. Tengberg (2018) beskriver at opplæring av vurderere, såkalt «rater training», samt å bruke flere lærere til å skåre de samme elevsvarene kan redusere variasjon og diskrepans i skåring (Tengberg, Roe, & Skar, 2018). Roe, Ryen og Weyergang (2018) har også diskutert verdien i et vurderingsfellesskap i form av kursing eller samskåring i vurdering av åpne oppgaver. Som pekt på tidligere opplevde vi de faglige diskusjonene som verdifulle, og det hadde vært spennende å ha slike fagsamtaler i forbindelse med skåring av åpne oppgaver som en fast rutine i arbeidet med nasjonale prøver. Diskusjoner i profesjonsfellesskapet om tolkning av prøvetekster og vurderingsveiledere før skåring, kan være med på å styrke reliabiliteten. Et tilleggsaspekt er verdien av lærere med større innsikt i prøvekonstruktet og operasjonaliseringen. Dette kan i sin tur vise seg i skolen på flere nivåer, vi som lærer er de som har ansvaret for leseopplæringen i skolen. Vi har tenkt på muligheten for et prøvekonstrukt som ivaretar både mål på kvaliteten i skolen, og som større grad kan ha en læringsstøttende funksjon. Kunne dette potensielt åpnet opp for å måle lesing med større fokus på validitet uten å måtte inngå avgrensede kompromiss mot reliabiliteten.

Ulike syn og tolkninger av både elevsvar og vurderingsveiledere vil fortsatt forekomme, samtidig må vi godta at reliabiliteten i skåring av åpne oppgaver aldri vil bli 100 %, da lærerens subjektive og skjønnsmessige vurdering vil være avgjørende i tvilstilfeller (Skaftun, 2006, s. 40).

På den andre siden «man unngå at høy reliabilitet står i direkte motsetning til høy validitet» (Roe, Ryen, & Weyergang, 2018, s. 178). Avslutningsvis vil vi påpeke at det kan være langt flere faktorer som har sammenheng med de åpne oppgavenes validitet og reliabilitet enn det som kommer fram i vårt masterprosjekt.

6 Avslutning

6.1 Svar på forskningsspørsmål

Formålet med denne masteroppgaven har vært få større innsikt i de åpne oppgavenes validitet og reliabilitet. Vi søkte etter en dypere forståelse av hva prøvetekstene og de åpne oppgavene krever av elevene sett opp mot rammeverket, samt hvordan dette ble vurdert i to åpne oppgaver – *Stargate* og *Snikfotografen*. Problemstillingen for denne masteroppgaven har vært: I hvilken grad gir åpne oppgaver på nasjonale prøver i lesing valide målinger og reliable resultater? I det følgende svarer vi ut våre to forskningsspørsmål.

1) I hvilken grad måler åpne oppgaver elevenes lesekompetanse i å tolke og sammenholde informasjon, og å finne informasjon i tekster?

Våre funn fra tekstgjennomgang av *Stargate* viser at; Oppgaven måler og skiller mellom grad av «å tolke og sammenholde informasjon» gjennom elever som gjengir tekstforståelse på implisitt eller eksplisitt nivå. *Stargate* har tekstkvaliteter som kunne ha åpnet for mer selvstendig tolkning, slik at den kunne målt elevenes kompetanse i større grad.

I *Snikfotografen* fant vi at nettartiklenes ulike elementer måtte sammenholdes. Oppgaven skilte mellom elever som hentet ut riktig informasjon som var noe implisitt uttrykt. Elevene måtte forstå språklige metaforer og fagterminologi for å vise grad av leseforståelse knyttet til leseformålet «å finne informasjon». Et annet funn var grupperingen av feilsvar, sentret rundt fototeknisk informasjon ifra teksten. I tillegg fant vi at elevens forforståelse også vil kunne påvirke grad av leseforståelse. Det kan være en utfordring for validiteten at elevsvar med relativt store innholdsmessige ulikheter, kan skåres til høyeste poengsum.

2) Hvor presise, dekkende og anvendelig er kriteriene for skåring som gis i vurderingsveilederen, og i hvilken grad svarer skårernes vurderingspraksis til disse kriteriene?

Vurderingsveilederne er i varierende grad presis, dekkende og anvendelig. Lærerens vurdering og skåring av elevsvarene opp mot vurderingsveilederens beskrivelser er i jevnt over samsvarende. Kvantitative funn viser at sensorreliabiliteten utfordres mest på polytom skala for den åpne oppgaven til *Stargate*. *Snikfotografen*, skåret på en dikotom skala, indikerte svært høye verdier for grad av samsvar i skåring. Reliabilitetsmålene bør ses opp mot lav svarprosent og høy andel blanke elevsvar og irrelevante svar.

Lærerens profesjonsfaglige skjønn avgjør skåring i tvilstilfeller. Vi erfarer at arbeid med prøvetekster, vurderingsveiledere og et utvalg elevbesvarelser i profesjonsfellesskapet kan bidra til økt innsikt i prøvekonstruktet. Felles diskusjon av elevsvar i forkant av skåring kan redusere sjansene for diskrepans.

Samtidig har det blitt klart for oss at utfordringen ligger i å konstruere gode oppgaver som ivaretar operasjonaliseringen av leseformålene på en god måte, og som samtidig ivaretar en reliabel vurdering av elevenes svar på disse. Prioritering av validitet og reliabilitet får konsekvenser for målingen av leseformålet. Det overordnede formålet for de åpne oppgavene er med på å skape retning for hva som i størst grad vektlegges av validitet og reliabilitet.

6.2 Videre arbeid med de åpne oppgavene

Et spørsmål vi har stilt oss flere ganger underveis i prosjektet er - hvor godt kjenner vi operasjonaliseringen av de ulike leseformålene? Klarer vi som lærere å svare enkelt ut på hva som kjennetegner de tre leseformålene? Har vi et felles metaspråk om hva elevene blir målt i på de nasjonale prøvene i lesing? I vår jobb som lærere forholder vi oss til læreplanverket og dens ulike beskrivelser hvor lesing inngår. Lesing omtales innenfor flere områder og nivåer, både som grunnleggende ferdigheter, kjerneelementer, tverrfaglige temaer og i kompetansemålene. Vi ønsker å peke på behovet for en tydeligere beskrivelse av lesing som grunnleggende ferdighet knyttet opp mot alle fag. Vil det være slik at et konstrukt som dannet oversikt over det store spennet i lesing som kognitiv ferdighet hadde styrket validitetsaspektet? Kunne leseforskere ha sett til utviklingen av skrivehjulet og laget et rammeverk som var enda mer oversiktlig, og som var med på å skape en felles forståelse for de ulike leseprosessene knyttet opp mot ulike nivåer? Kunne dette ha styrket begrepsvaliditeten? Et slikt tiltak mener vi ville ha gitt mer trygghet og fleksibel tilnærming til skåring av åpne oppgaver, og mulig være et bidrag til å styrke validiteten og reliabiliteten til åpne oppgaver. Dette hadde vært et spennende tema for videre forskning.

7 Referanser APA 6th

Berge, K. L. (2019). Det (nye) nye norskfaget. I M. Blikstad-Balas, K. R. Solbu, & K. R. Solbu (Red.). Bergen: Fagbokforlaget.

Bjørnsson, J. K. (2022). *Prøver og internasjonale storskalaundersøkelser i Norge. Hva kjennetegner disse målingene og hvordan har de virket i det nasjonale kvalitetsvurderingssystemet de siste 20 årene?* Hentet fra https://www.regjeringen.no: https://files.nettsteder.regjeringen.no/wpuploads01/sites/508/2023/02/Forskningsoppsummering-prover-og-ILSA-studier.pdf

Blömeke, S., & Rolf, O. V. (2018, 12 20). På vei mot et sammenhengende nasjonalt kvalitetsvurderingssystem. *Acta Didactica Norge*.

Cohen, L., Manion, L., & Morrison, K. (2007). *Research Metods in Education* (6. utgave. utg.). Abingdon, Oxon, England: Routhledge.

De nasjonale forskningsetiske komiteene. (2023). *Forskningsetiske retningslinjer for samfunnsvitenskap og humaniora.* Hentet april 2, 2024 fra Forskningsetikk.no: https://www.forskningsetikk.no/globalassets/dokumenter/4-publikasjoner-som-pdf/forskningsetiske-retningslinjer-for-samfunnsvitenskap-og-humaniora.pdf

Fevolden, B. H., Thorsen, T., & Eriksen, A. (2019). *Rapport fra pilotering og gjennomføring av den nasjonale prøven i lesing på 8. og 9. trinn 2018.* Oslo: Enhet for kvantitative utdanningsanalyser. Institutt for lærerutdanning og skoleforskning.

Frønes, T. S., & Ryen, J. A. (2020). Introduksjon: Like muligheter til god leseforståelse? 20 år med lesing i PISA. I T. S. Frønes, F. Jensen, T. Frønes Stjern, & F. Jensen (Red.), *20 år med lesing i PISA* (ss. 10-40). Universitetsforlaget.

Frønes, T. S., & Ryen, J. A. (2020). Å forstå det man leser - å trekke slutninger i skjønnlitteratur og sakprosa. I T. S. Frønes, F. Jensen, T. Frønes Stjern, & F. Jensen (Red.), *Introduksjon: Like muligheter til god leseforståelse? 20 år med lesing i PISA* (ss. 135-164). Universitetsforlaget.

Gilje, N. (2019). *Hermeneutikk som metode - ein historisk introduksjon.* Oslo: Det norske samlaget.

Gleiss, M. S., & Sæther, E. (2022). *Forskningsmetode for lærerstudenter.* Oslo: Cappelen Damm Akademisk.

Golden, A., & Kulbrandstad, L. I. (2007, 02). Teksten som utgangspunkt for arbeidet med lesing og ordforråd. Andrespråksdidaktiske utfordringer i videregående opplæring. *NOA- norsk som andrespåk, 23*, ss. 33-66.

Gaasland, R. (1999). *Fortellerens hemmeligheter innføring i litterær analyse.* Oslo: Universitetsforlaget.

Halvorsen, K. (2008). *Å forske på samfunnet - en innføring i samfunnsvitenskapelig metode.* Oslo: J.W. Cappelens Forlag as.

Helland, T., & Tresse, A. (2023, 11 09). Går inn for å avvikle de nasjonale prøvene. *Utdanningsforbundet.no*. Hentet 05 07, 24 fra https://www.utdanningsforbundet.no/nyheter/2023/gar-inn-for-a-avvikle-nasjonale-prover

Jensen, F., Frønes, T. S., Kjærnsli, M., & Roe, A. (2020). Lesing i PISA 2000-2018: Norske elevers lesekompetanse i et internasjonalt perspektiv. I T. S. Frønes, & F. Jensen, *Introduksjon: Like muligheter til god leseforståelse? 20 år med lesing i PISA* (ss. 21-39). Oslo: Universitetsforlaget.

Jensen, F., Pettersen, A., Frønes, T. S., Eriksen, A., Løvgren, M., & Narvhus, E. (2023). *PISA 2022 Norske elevers kompetanse i matematikk, naturfag og lesing.* Cappeln Damm Akademisk. Hentet fra https://doi.org/10.23865/noasp.205

Kjølås, C., & Hansen, C. (2021, 6). Reliabilitet i skåring av åpne oppgaver i nasjonale prøver i lesing. *Upublisert*. Tromsø.

Kulbrandstad, L. I. (2022). *Lesing i utvikling. Teoretiske og didaktiske perspektiver* (3. utgave. utg.). Bergen: Fagbokforlaget.

o. (2003-2004). *St.meld. nr. 30.* Utdannings- og kunnskapsdepartementet. Hentet fra https://www.regjeringen.no/contentassets/988cdb018ac24eb0a0cf95943e6cdb61/no/pdfs/stm200320040030000dddpdfs.pdf

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*(1), ss. 159174. doi:10.2307/2529310

Lie, S., Hopfebeck, T. N., Ibsen, E., & Turmo, A. (2005). *Rapport fra en utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2005.* Oslo: Institutt for lærerutdanning og skoleutvikling Universitetet i Oslo.

Løvland, A. (2015). Sammensatte fagtekster - en multimodal utfordring? I E. Maagerø, E. S. Tønnessen, & E. S. Tønnessen (Red.), *Å lese i alle fag.* oslo: Universitetsforlaget.

Markussen, S., Ræder, H. G., Røgeberg, O., & Raaum, O. (2024, Januar 10). i endring: Utviklingen over tid målt ved nasjonale prøver. *Acta Didactica Norden, 18*, ss. 1-33.

Matre, S., Solheim, R., Hildegunn, O., Berge, K. L., Evensen, L. S., & Thygesen, R. (2021). *Nye grep om skriveopplæringa.* (S. Matre, R. Solheim, & O. Hildegunn, Red.) Oslo: Universitetsforlaget AS.

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. *Biochemia medica, 22*, ss. 276-282. Hentet fra https://pubmed.ncbi.nlm.nih.gov/23092060/

NOU 2023 :27 (2023). *Et nytt system for kvalitetsutvikling - for elevenes læring og trivsel.* Oslo: Norges offentlige utredninger. Hentet fra https://www.regjeringen.no/contentassets/85cc09d37f604868b31cd0da68aaa200/no/pdfs/nou202320230027000dddpdfs.pdf

Rishøi, I. H. (2021). *Stargate en julefortelling.* Oslo: Gyldendal.

Roe, A., & Blikstad-Balas, M. (2022). *Lesedidaktikk -etter den første leseopplæringen* (4. utgave. utg.). Oslo: Universitetsforlaget.

Roe, A., Ryen, J. A., & Weyergang, C. (2018). *God leseopplæring med nasjonale prøver.* Oslo: Universitetsforlaget.

Ruud, M. (2023, 11 11). Nasjonale prøver er omdiskutert i hele Norden. *Utdanningsnytt*.

Ryen, J. A., & Frønes, T. S. (2020). Å forstå det man leser-å trekke slutninger i skjønnlitteratur og sakprosa. I T. S. Frønes, & F. Jensen, *Like muligheter til god leseforståelse?* (ss. 135-165). Oslo: Universitetsforlaget.

Selan, I., Vibe, N., & Hovdhaugen, E. (2013). *Evaluering av nasjonale prøver som system.* Oslo: NIFU Nordisk institutt for studier av innovasjon, forskning og utdanning.

Skaftun, A. (2006). *Å kunne lese grunnleggende ferdigheter og nasjonale prøver.* Bergen: Fagbokforlaget.

Skaftun, A., Roe, A., Narvhus, E., & Solheim, O. S. (2006). Tilnærminger til et teoretisk rammeverk for de nasjonale prøvene i lesing. *Norsk Pedagogisk Tidsskrift*, ss. 360-371.

Skar, G. B., & Aasen, A. J. (2018, 12 17). Å måle skriving som grunnleggende ferdighet. *Acta Didactica Norge*, s. 29. doi: https://doi.org/10.5617/adno.6280

Skrivesenteret. (2016). *Skriveprøven for 5. og 8. trinn.* Hentet fra Skrivesenteret: https://skrivesenteret.no/prosjekt/skriveproven-for-5-og-8-trinn/

Solheim, O. J., & Skaftun, A. (2009, Juli 23). The problem of semantic openness and constructed respons. *Assessment in Education Prinsiples Polisy and Practice, 16*, ss. 149-164. doi:10.1080/09695940903075909

Tengberg, M. (2017). National reading tests in Denmark, Norway, and Sweden: A comparison of construct definitions, cognitive targets, and response formats. *Language Testing, Volum 34*, ss. 83-100. Hentet mars 9, 2024 fra https://journals.sagepub.com/doi/epub/10.1177/0265532215609392

Tengberg, M., & Skar, G. (2017). Hur tillförlitligt är det i nationella provet i läsning i åk 9? *Utbildning & demokrati, 26*, ss. 113-137.

Tengberg, M., & Skar, G. B. (2016). Samstämmighet i lärares bedömning av nationella prov i läsforståelse. *Nordic Journal of Literacy Research, Vol. 2*, ss. 1-18. doi:10.17585

Tengberg, M., Roe, A., & Skar, G. B. (2018, 05 30). Interrater reliability of constructed response items in standardized tests of reading. (H. Sæverot, Red.) *Nordic studies in Education, 2018*, ss. 118-137. doi:10.18261/issn.1891-5949

Thrane, C. (2018). *Kvantitativ metode - en praktisk tilnærming.* Oslo: Cappelen Damm Akademisk.

Universitetet i Tromsø. (2021). *PRINSIPPER OG RETNINGSLINJER FOR FORVALTNING AV FORSKNINGSDATA VED UIT.* Forskningsstrategisk utvalg, Tromsø. Hentet fra https://uit.no/regelverk/sentraleregler#v-pills-742423

Utdanningsdirektoratet. (2020, 08 01). *www.udir.no.* Hentet fra Læreplan i norsk: https://www.udir.no/lk20/nor01-06/om-faget/grunnleggende-ferdigheter?lang=nob

Utdanningsdirektoratet. (2020). *Grunnleggende ferdigheter.* Hentet fra Utdanningsdirektoratet: https://www.udir.no/lk20/nor01-06/om-faget/grunnleggende-ferdigheter

Utdanningsdirektoratet. (2022, august 19). *Rammeverk for nasjonale prøver* Hentet fra Utdanningsdirektoratet.no: https://www.udir.no/eksamen-og-prover/prover/rammeverk-for-nasjonale-prover2/tekniske-krav-til-provene/#provens-validitet

Utdanningsdirektoratet. (2023, november 11). *Resultater nasjonale prøver ungdomstrinnet.* Hentet november 26, 2023 fra Utdanningsdirektoratet: https://www.udir.no/tall-og-forskning/statistikk/statistikk-grunnskole/analyser/2023/analyse-av-nasjonale-prover-for-ungdomstrinnet-2023

Utdanningsdirektoratet. (2023, november 9). *Analyse av nasjonale prøver for ungdomstrinnet*. Hentet fra Utdanningsdirektoratet: https://www.udir.no/tall-og-forskning/statistikk/statistikk-grunnskole/analyser/2023/analyse-av-nasjonale-prover-for-ungdomstrinnet-2023/

Vestheim, O., & Sem, L. (2019, mai). Nasjonale prøver i norske riks- og regionaviser - en historisk av framstillinger av nasjonale prøver. *Nordisk tidsskrift for pedagogikk og kritikk*, ss. 27-45.

Weyergang, C., Siljan, H. H., & Frønes, T. S. (2023). Hvordan vurdere elevers kritiske lesing? Forslag til et praktisk-didkatisk rammeverk. *Nordic of Literacy Research*, ss. 113-140.

Vedlegg 1: Søknad til Utdanningsdirektoratet

fredag 3. november 2023

"Validitet og reliabilitet i åpne oppgaver, i nasjonale prøver i lesing på 8. og 9. trinn."

Vi er Cecilie Hansen og Christina Kjølås, to norsklærere med lang erfaring som lærere i Tromsø kommune. Denne høsten er vi i gang med å ferdigstille vår fagdidaktiske master i norsk ved UIT Norges arktiske universitet. Vår veileder er førsteamanuensis Morten Bartnæs.

Tema for vår masteroppgave springer ut fra egne erfaringer med skåring av åpne oppgaver i arbeidet med nasjonale prøver i lesing for ungdomstrinnet. Når vi som lærere skal skåre elevbesvarelser som faller utenfor eksemplene i vurderingsveiledningen, kan det være krevende å vurdere hvordan tvilstilfellene skal skåres. Vurderingsveilederen påpeker at tvilen skal komme eleven til gode, men er det så enkelt?

Emnet for masterprosjektet vårt er de åpne oppgavenes validitet og reliabilitet. Vi tar utgangspunkt i følgende forskingsspørsmål:

Forskerspørsmål 1: I hvilken grad måler de åpne spørsmålene elevenes evne til å finne, reflektere, tolke og vurdere tekster?

Forskerspørsmål 2: Hvordan sammenfaller elevbesvarelsene med vurderingskriteriene og eksemplene i vurderingsveilederen og faktisk vurderingspraksis?

Forskerspørsmål 3: Hvordan og i hvilken grad fungerer vurderingsveiledningen til de åpne oppgavene i praksis, slik at den gir grunnlag for høy sensorreliabilitet i skåring?

I masterprosjektet ønsker vi å gjøre en kvantitativt orientert innholdsanalyse av et utvalg autentiske elevbesvarelser som er skåret, ved hjelp av kategorier som vi har etablert på forhånd. I denne delen av prosjektet ønsker vi å sammenholde faktiske skåringer av elevsvar med våre vurderinger av svarene.

I tillegg ønsker vi å gjøre en survey hvor vi basert på et utvalg elevbesvarelser, tester ut grad av samsvar i skåring, sensorreliabiliteten.

Målet med vår masterprosjektet er at vi kanskje kan tilføre utdanningsforskningen et nytt perspektiv på vurdering av de åpne oppgavene. Prosjektet kan bidra til å øke kvaliteten på nasjonale prøver i lesing.

Vi søker med dette om tilgang til et utvalg anonymiserte elevbesvarelser fra årets åpne oppgaver i lesing med poengsum/skår.

Til prosjektet trenger vi om lag 12 klassesett fra 8. og 9. trinn, gjerne spredt geografisk fra Nord-Norge, Midt-Norge, Sør-Norge og Vest- Norge. Hvis UDIR kan velge ut dette antallet

fredag 3. november 2023

med en slik geografisk spredning, trenger vi ikke flere opplysninger om datasettene, annet enn trinn, by/land og fylke.

Vi skal lagre materialet i henhold til UiT sine retningslinjer for personvern i forsknings- og studentprosjekter ved UiT. Etikkportalen ved UiT følges. Vi vil destruere materialet ved fullført master høsten 2024.

Ta gjerne kontakt dersom dere har spørsmål vedrørende masterprosjektet.

Med vennlig hilsen
Cecilie Hansen og Christina M. Kjølås (47 64 62 54)

Veileder: Morten Bartnæs førsteamanuensis i norsk, Institutt for lærerutdanning og pedagogikk, Universitetet i Tromsø

Vedlegg 2: Vedtak 2023: 11470

Utdannings-
direktoratet

Saksbehandler:
Hilde Hultin

Vår dato:

Vår referanse:
2023/11470

Deres dato:

Deres referanse:

Morten Bartnæs
UiT Norges arktiske universitet
Universitetsvegen 39
9019 Tromsø

Svar på søknad om dispensasjon fra taushetsplikt i forbindelse med prosjektet Validitet og reliabilitet i de åpne oppgavene i de nasjonale prøvene i lesing på 8. og 9. trinn

Utdanningsdirektoratet viser til brev fra Morten Bartnæs datert 03.11.2023 vedrørende søknad om dispensasjon fra taushetsplikt i forbindelse med utlevering av data til «Validitet og reliabilitet i de åpne oppgavene i de nasjonale prøvene i lesing på 8. og 9. trinn».

Vedtak

Utdanningsdirektoratet kan utlevere de omsøkte opplysningene til Morten Bartnæs til det ovennevnte forskningsprosjektet

Av hensyn til de registrertes personvern og for å sikre at utlevering av data ikke medfører uforholdsmessig ulempe for andre interesser, er det knyttet følgende vilkår til utleveringen:

- Det utleverte datamaterialet kan kun benyttes til forskning i samsvar med det oppgitte formålet i prosjektsøknaden.
- Utlevert datamateriale kan bare gjøres tilgjengelig for Christina Kjølås og Cecilie Hansen. Ved bytte av personer underveis i prosjektet skal prosjektleder kontakte Utdanningsdirektoratet.
- Personer som får tilgang til det utleverte datamaterialet, må underskrive taushetserklæring.
- Eventuelle rapporter og publikasjoner må utgis i en slik form at enkeltpersoner ikke kan identifiseres, verken direkte eller indirekte.
- Dokumentasjon om internkontroll og sikkerhet ved behandlingen av personopplysninger etter personvernforordningen art. 24 og art. 32 skal på forespørsel utleveres til Utdanningsdirektoratet.
- Opplysningene skal behandles i tråd med personopplysningslovens bestemmelser for behandling av personopplysninger, samt EUs

personvernforordning som trådte i kraft mai 2018, og virksomhetens (eventuelle) konsesjon fra Datatilsynet.
- Det utleverte datamaterialet slettes straks det ikke er behov for dem og senest ved prosjektets avslutning den 31.12.2024.
- Sletting skal bekreftes av prosjektleder Morten Bartnæs på vedlagte skjema. Dersom det er behov for å beholde data utover slettedato, må det søkes om ny dispensasjon fra Utdanningsdirektoratet.

Om prosjektet og dataene som ønskes utlevert

Nærmere om prosjektet
Målet med prosjektet er å tilføre utdanningsforskningen et nytt perspektiv på vurdering av de åpne oppgavene i nasjonale prøver i lesing. Prosjektet kan bidra til å øke kvaliteten på nasjonale prøver i lesing. For å kunne undersøke validiteten og reliabiliteten i de åpne oppgavene på nasjonale prøver i lesing, opp mot vurderingsveiledningen og vurderingspraksis trenger forskerne tilgang til disse dataene.

Data som ønskes utlevert
Variabel 1: Elevbesvarelser fra høstens nasjonale prøver i lesing 8. og 9. trinn, alle syv åpne oppgavene. Snikfotografen, Tasmansk Pungulv, Stargate, Skulearbeid, den korte historien omegget og et skår i gleden (2 åpne oppgaver knyttet til denne).

Variabel 2: Skåringen av de samme besvarelsene med poengsum.

Utdanningsdirektoratets vurdering

Rettslig utgangspunkt
Etter forvaltningslovens § 13 d første ledd kan fagdepartementet, når det finnes rimelig og ikke medfører uforholdsmessig ulempe for andre interesser, bestemme at opplysninger kan utleveres til bruk for forskning og at dette skal skje uten hinder av organets taushetsplikt. Med hjemmel i forskrift til forvaltningsloven § 8 første ledd er dispensasjonsmyndigheten i denne saken delegert til Utdanningsdirektoratet.

Utdanningsdirektoratet vurderer det som rimelig at Morten Bartnæs får tilgang til datamaterialet de søker om til bruk i forskningsprosjektet «Validitet og reliabilitet i de åpne oppgavene i de nasjonale prøvene i lesing på 8. og 9. trinn». I denne vurderingen har Utdanningsdirektoratet lagt vekt på at den samfunnsmessige nytteverdien av forskningsprosjektet er stor. Videre har vi vurdert at utleveringen av dataene vil skje på en slik måte at det ikke medfører uforholdsmessig ulempe for andre interesser, siden det er bare de enkelte på prosjektet som har tilgang og utlevert data blir lagret på et sikkert område. Det har blitt også gjort personkonsekvensvurdering for det aktuelle prosjektet. Selve prosjektet blir gjennomført av erfarne forskere.

Klagerett og videre saksgang
Vedtaket kan påklages i henhold til forvaltningslovens (fvl.) bestemmelser om klage på enkeltvedtak, se fvl. § 28. Klagefristen er tre uker fra dette brevet mottas, jf. fvl. § 29. Klageinstansen er Kunnskapsdepartementet, men en eventuell klage skal rettes til Utdanningsdirektoratet.

Vennlig hilsen

navn
avdelingsdirektør

navn
rådgiver

Vedlegg 3: Taushetserklæring - Ekstern - Forsker

Utdannings-
direktoratet

Taushetserklæring – Ekstern - Forsker

Validitet og reliabilitet i de åpne oppgavene i nasjonale prøver i lesing

Sak: 2023/11470

Jeg forstår at

- jeg i mitt arbeid kan få kjennskap til forhold som er taushetsbelagte i medhold av lov eller forskrift
- jeg har taushetsplikt når det gjelder informasjon om noens personlige forhold og/eller forretningshemmeligheter som jeg får kjennskap til gjennom mitt arbeid, jf. forvaltningsloven § 13 første ledd nr. 1 og 2
- forvaltningsloven § 13 e om forskeres taushetsplikt gjelder for de opplysninger jeg får utlevert etter utleveringsavtalen
- jeg i mitt arbeid kan få kjennskap til forhold som må behandles strengt konfidensielt, for eksempel eksamensoppgaver, nasjonale prøver, budsjettopplysninger og datasikkerhet
- taushetsplikten også gjelder etter at mitt arbeid tilknyttet UiT Norges arktiske universitet, institutt for lærerutdanning og pedagogikk avsluttet jf. forvaltningsloven § 13 tredje ledd
- brudd på taushetsplikten og misbruk av informasjon jeg får kunnskap om kan medføre straffeansvar jf. straffeloven kapittel 21
- brudd på plikten til konfidensialitet kan medføre erstatningsansvar for Utdanningsdirektoratets økonomiske tap som følge av bruddet
- brudd på taushetsplikt eller plikt til konfidensialitet kan medføre at kontrakten med Utdanningsdirektoratet opphører med øyeblikkelig virkning

Jeg forplikter meg til å

- overholde den taushetsplikten som følger av lov eller forskrift
- overholde plikten til konfidensialitet for opplysninger som kan påføre Utdanningsdirektoratet økonomisk tap og/eller sikkerhetsbrudd
- opptre i tråd med lojalitetsplikten til Utdanningsdirektoratet og utvise varsomhet med behandlingen av opplysninger om direktoratet mv, herunder ikke gjøre slike opplysninger tilgjengelig for utenforstående uten samtykke fra direktoratet
- vise aktsomhet i behandlingen av alle opplysninger, og arbeide i samsvar med eventuelle vilkår fastsatt av Utdanningsdirektoratet
- ikke å gi opplysninger videre til personer i eller utenfor, og som ikke er nevnt i utleveringsavtalen

Denne taushetserklæring skal underskrives av de personer i virksomheten som er navngitt i "Avtale om utlevering av taushetsbelagte opplysninger til bruk i forskning" og som skal ha tilgang til opplysningene. Erklæringen er undertegnet i to eksemplarer, hvorav underskriver og Utdanningsdirektoratet beholder hver sitt eksemplar.
Jeg har satt meg inn i de lov- og forskriftsbestemmelsene som det er vist til over. Taushetserklæring er lest og akseptert:

Ved digital signering fjerner du feltet under og setter inn teksten: Dokumentet er signert elektronisk

Navn: Christina M. Kjølås Cecilie Hansen Morten Bartnæs	UiT Norges arktiske universitet, institutt for lærerutdanning og pedagogikk
Underskrift: Christina M. Kjølås Cecilie Hansen	Tromsø, den 17.01.2024.