Hvilken sammenheng er det mellom elevers prestasjoner på en test i naturfag og formatet til oppgavene?

En sammenligning av resultater oppnådd på lukkede oppgaver (SR) og tilsvarende oppgaver i åpent format (CR)

Grethe Ravlo

Hvilken sammenheng er det mellom elevers prestasjoner på en test i naturfag og formatet til oppgavene?

En sammenligning av resultater oppnådd på lukkede oppgaver (SR) og tilsvarende oppgaver i åpent format (CR)

Masteroppgave i naturfagdidaktikk

EDU3910

NTNU 2010

Forord

Å skrive en masteroppgave er som å lære et nytt håndverk. Da er det trygt å ha erfarne "håndverkere" som kan øse av sin erfaring, bidra med gode råd, være kritisk venn og slik forhindre at den uerfarne faller i "grøfta" eller utfor "stupet". I særdeleshet har Rolf Vegard Olsen vært en slik veileder gjennom sine bidrag med konstruktive "framovermeldinger", kritiske spørsmål og oppmuntrende kommentarer. Han fortjener en stor takk! En annen som skal takkes, er Peter van Marion. Hans evne til raskt å gripe fatt i hovedlinjene når man stiller et spørsmål, har bidratt til effektive løsninger og god framdrift i prosessen med å få fullført denne masteroppgaven. Takk også til min søster som med stødig blikk har hjulpet til med korrekturlesing!

325 elever har til sammen besvart 12 000 enkeltoppgaver, og velvillige lærere har bidratt i gjennomføringen av testene. Uten disse personene ville denne masteroppgaven ikke vært mulig å gjennomføre. Derfor "tusen takk for hjelpen" til de to skolene og lærerne og elevene i de elleve klassene som har deltatt i studien!

Mine kolleger ved Matematikksenteret og ansatte ved Skolelaboratoriet ved NTNU fortjener også takk for oppmuntrende og støttende kommentarer. Sist, men ikke minst, vil jeg takke min nærmeste familie for all støtte. Min hobby de siste årene har i høy grad påvirket familiens hverdag, og det er nok mange som synes det er greit at jeg nå er i mål.

Trondheim 20. mai 2010

Grethe Ravlo

Innhold

1. Kapittel Innledning

1.1. Bakgrunn for valg av problemstilling

Temaet i denne masteroppgaven er vurdering med fokus på to ulike oppgaveformat, lukkede oppgaver (SR, - Selected Response Items) og åpne oppgaver (CR, - Constructed Response Items). Når vi vurderer, ønsker vi å få kjennskap til elevers reelle kompetanse innenfor bestemte områder. Vurderingen skjer ofte på grunnlag av resultater oppnådd på en skriftlig prøve, og i Norge har vi hatt tradisjon for å bruke åpne oppgaver i disse testene. Det vil si at elevene får et spørsmål og at de med egne ord skal formulere et svar. Andre deler av verden har andre tradisjoner, og etter at norske elever i 1984 begynte å delta i internasjonale komparative undersøkelser (SISS[1]), og fortsatte med TIMSS[2] (1995, 2003, 2007, 2008) og PISA[3] (2000, 2003, 2006), er vi blitt kjent med oppgaver i lukket format (Grønmo mfl. 2004, 2009; Kjærnsli mfl. 1999, 2004, 2007; Lie mfl. 1997, 2001; Sjøberg 1986). I slike oppgaver skal elevene ikke formulere et eget svar, men bare ta stilling til svaralternativer som er oppgitt. Av det lukkede formatet er flervalgsoppgaver (MC, - Multiple Choice) den oppgavetypen som er mest brukt. I disse oppgavene svarer elevene ved å krysse av for det alternativet de mener er det rette eller det beste svaret på oppgaven. Vanligvis er det fire eller fem alternativer å velge mellom, men det kan også være to eller tre. Oppgavetypene innenfor det lukkede formatet (SR), er forklart i kapittel 2.

Jeg har siden 2004 arbeidet ved Nasjonalt senter for matematikk i opplæringen, med utvikling av både åpne oppgaver (CR) og flervalgsoppgaver (MC) til nasjonale prøver i matematikk og regning. I Norge har mange vært opptatt av i hvilken grad det er mulig å få et riktig bilde av elevers kompetanse når elevene i stor grad bare skal velge svar blant alternativer som er oppgitt. Det har derfor vært knyttet mange kritiske spørsmål til innføring av tester med flervalgsoppgaver, og noen har sammenlignet testene med tippekuponger (Sirnes 2007).

En innvending fra kritikerne er at det kan være lettere å gjette seg til rett svar i flervalgsoppgaver (MC) enn i åpne oppgaver. Muligheten for å finne riktig løsning ved å gjette, vil vel alltid være til stede, men det spørs om gjetting kan utgjøre en så stor tilfeldig

[1] SISS – The second International Science Study
[2] TIMSS - Trends in International Mathematics and Science Study
[3] PISA - Programme for International Student Assessment

målefeil at det derfor er grunn til å være skeptisk til resultatene elevene oppnår. Det finnes forskning som har undersøkt dette, og det vil jeg komme nærmere tilbake til i kapittel 2.

Andre spørsmål som er reist, handler om kjønnsforskjeller. Man kan lure på om de ulike oppgaveformatene påvirker resultatene til jentene på en annen måte enn de påvirker resultatene til guttene. Likeså om det er grupper av elever som ut fra kompetansenivå, i større grad enn andre elever, skårer på oppgaver i lukket format. Det er satt fram påstander om at man med lukkede oppgaver bare kan måle faktakunnskaper. Stemmer dette, eller kan man også få målt kunnskaper på et høyt kognitivt nivå med dette oppgaveformatet? Disse spørsmålene behandler jeg i kapittel 2 i forhold til relevant litteratur, og i kapittel 4 og 5 gjennom drøfting av egne resultater.

Til tross for innvendinger fra mange hold er obligatoriske prøver med stor vekt på lukkede oppgaver allerede innført på nasjonalt nivå i Norge. Det skjer ved at ca. 120 000 elever hvert år gjennomfører nasjonale prøver i lesing, regning og engelsk, noe som har pågått siden 2004. De første årene bestod prøvene i stor grad av åpne oppgaver. I evalueringsrapportene som ble utarbeidet ved Institutt for lærerutdanning og skoleutvikling (ILS) ved Universitetet i Oslo etter prøvene i 2004 og 2005 (Lie mfl. 2004 og Lie mfl. 2005), kommenterte imidlertid forskerne den utstrakte bruken av åpne oppgaver i de nasjonale prøvene. Det ble foreslått at flervalgsoppgaver (MC) burde inngå i større grad, blant annet for at prøvene skulle oppnå høyere reliabilitet, - det vil si at resultatet skulle bli mer pålitelig. Disse anbefalingene har ført til at prøvene siden 2007 har inneholdt ca. 70 prosent lukkede oppgaver. Omfanget av åpne og lukkede oppgaver er nedfelt i *Rammeverk for de nasjonale prøvene* (Utdanningsdirektoratet 2006), og vedtatt av Kunnskapsdepartementet.

Elever i Norge blir vanligvis opplært til å begrunne sine svar. Kan denne tradisjonen være en ulempe når elevene i flervalgsoppgaver bare skal krysse av for et alternativ som er oppgitt? I hvilken grad kan dette i så fall påvirke resultatet til enkeltelever? Resultater fra den internasjonale undersøkelsen, TIMSS 1995, viser at i matematikk presterte norske elever like bra på flervalgsoppgaver som på åpne oppgaver, mens de i naturfag presterte spesielt godt på oppgaver som krevde forklaring med egne ord (Lie mfl. 1997). Analysene fra TIMSS 2007 viser at norske lærere i naturfag på 8. trinn, benyttet flervalgsoppgaver i mye mindre grad enn lærere i andre land (Bergem mfl. 2009). Ut fra dette skulle man kanskje forvente at norske elever fremdeles skårer bedre på åpne oppgaver enn på flervalgsoppgaver når de testes i

naturfag? Gjør de egentlig det? Resultatene fra PISA 2003 tyder på at dette har endret seg. Norske elever fikk denne gang bedre resultat på flervalgsoppgaver enn på åpne oppgaver (Kjærnsli mfl. 2004). Er dette også tilfellet hvis vi tester elevene i forhold til læreplanmålene i Kunnskapsløftet (LK06)?

I Norge eksisterer det fremdeles både fordommer mot og mange spørsmål knyttet til vurdering med lukkede oppgaver. Samtidig blir det lukkede oppgaveformatet stadig mer brukt i storskalatester i undervisningssammenheng. Utdanningsdirektoratet i Norge har som mål å få de fleste skriftlige eksamener overført til elektroniske prøver innen få år. Siden disse prøvene skal være selvrettende, vil de i svært stor grad bestå av lukkede oppgaver.

Mange har forsket på oppgaveformatets betydning i tester, og resultatene er ikke entydige (DeMars 2000, Hastedt mfl. 2005, Wester 1995). Dette er noe av bakgrunnen for at jeg med min masteroppgave ønsker å finne ut mer om hvilken betydning formatet til oppgavene har for resultatet elevene oppnår på en prøve.

1.2. Problemstilling

Oppgavens hovedproblemstilling er:
Hvilken sammenheng er det mellom elevers prestasjoner på en test i naturfag og formatet til oppgavene?

For å undersøke dette, har jeg følgende forskningsspørsmål:

1. ***Hvilket samsvar er det mellom resultatet en gruppe elever oppnår på naturfagoppgaver i åpent format (CR, - Constructed Response Items), og resultatet som oppnås av en sammenlignbar elevgruppe når oppgaver med samme opplysninger og spørsmål er i formatet lukket (SR, - Selected Response Items)?***

2. ***Dersom vi tar utgangspunkt i elevenes faglige nivå, er det noen elevgrupper som i større grad enn andre, har overvekt av riktige løsninger innenfor et av oppgaveformatene?***

3. ***Er det kjønnsforskjeller når det gjelder å mestre åpne og lukkede oppgaver?***

Jeg ønsker å undersøke om oppgaveformatet kan ha betydning for resultater som elever oppnår på en prøve. Erfaringer fra arbeidet med de nasjonale prøvene i regning, gjør at jeg har en hypotese om at elevene i gjennomsnitt skårer bedre på de lukkede enn på de åpne oppgavene. Er det slik, og gjelder dette for alle elever? Eller er det noen elevgrupper som når vi ser på elevenes kompetansenivå, har større overvekt enn andre grupper av riktige svar på de lukkede oppgavene i forhold til de åpne?

Først vil jeg se på elevenes totale poengsum på prøven og på hvor stor andel av de riktige løsningene som er på lukkede oppgaver og på åpne oppgaver. Deretter vil jeg sammenligne resultatene i åpent og lukket format for en del av oppgavene. Enkeltoppgaver hvor det er liten forskjell i løsningsprosent mellom oppgaven i åpent og lukket format, og hvor det blir betydelig forskjell i løsningsprosenten når formatet endres, vil bli studert nærmere.

Om kjønnsforskjeller kan knyttes til de ulike oppgaveformatene er interessant å undersøke. Er det forskjell på prestasjonene til jentene og guttene når det gjelder å løse åpne oppgaver og er det kjønnsforskjeller når det gjelder å løse lukkede oppgaver? Har jentene høyere gjennomsnittlig løsningsprosent på de åpne oppgavene enn på de lukkede oppgavene eller omvendt, eller kanskje det ikke er noen forskjell? Hva med guttene når det gjelder å løse oppgaver i de ulike formatene?

For å prøve å finne svar på disse forskningsspørsmålene, har jeg gjennomført to kvantitative undersøkelser med fire måneders mellomrom. Det var de samme elevene som ble testet i begge undersøkelsene. Formålet med den første prøven, Test 1, var på bakgrunn av resultatene å få delt elevene inn i to sammenlignbare grupper. Formålet med den andre prøven, Test 2, var å sammenligne resultatene som elevene i disse to gruppene oppnådde på oppgaver i åpent format med resultater de oppnådde på oppgaver med samme innhold i lukket format.

Test 1 og Test 2 var bygd opp på samme måte og bestod av 10 åpne og 10 lukkede oppgaver. Test 2 bestod av to prøver, prøve 1 og prøve 2. Oppgavene var like, men de 10 oppgavene som var åpne i prøve 1 var lukkede oppgaver i prøve 2 og vice-versa. Analysene på oppgaveformatets betydning for resultatet og om det er kjønnsforskjeller når det gjelder å mestre åpne og lukkede oppgaver, ble gjennomført på tre nivågrupper ut fra resultatene på Test 2.

I tilknytning til Test 1 fikk elevene 8 spørsmål som handlet om holdninger til naturfag, selvtillit i faget og strategier elevene brukte når de løste flervalgsoppgaver. Resultatene er analysert og brukt som justering ved sammensetting av gruppene til Test 2, men er ikke en del av problemstillingen. I utgangspunktet var elevers holdninger til og strategier ved løsing av flervalgsoppgaver ment å være et av forskningsspørsmålene. Dette måtte jeg endre underveis for at oppgaven skulle holde seg innenfor vanlige rammer for en masteroppgave.

1.3. Oversikt over masteroppgaven

Etter dette innledende kapittelet med begrunnelse for og klargjøring av problemstillingene, redegjør jeg i kapittel 2 for relevant teori i tilknytning til forskningsspørsmålene. Dette omfatter blant annet de ulike oppgaveformatene som er brukt i undersøkelsen, og fordeler og ulemper knyttet til bruk av disse i tester. I tillegg forklarer jeg i kapittel 2, ganske detaljert, noen av de psykometriske begrepene som benyttes ved kvantitative analyser. Dette er sentrale begreper for kvalitetssikring av resultater, og jeg har derfor har valgt å gi dem relativt stor oppmerksomhet.

Kapittel 3 handler om forskningsdesign, data og utvalg, operasjonalisering av begreper og metoder som ble brukt, ved innsamling av data.

I kapittel 4 presenteres resultatene. Først analyseres Test 1 med hensyn på kvalitetssikring. Dette har fått relativt stort fokus, siden de to gruppene som sammenlignes i Test 2, er dannet på bakgrunn av resultatene fra Test 1. Andre del av kapittel 4 inneholder en kvalitetssikring av de to prøvene, prøve 1 og prøve 2, som utgjør Test 2. Til slutt i kapittelet kommer de fagrelaterte resultatene fra prøvene i Test 2. Resultatene fra prøve 1 og prøve 2 i Test 2 blir sammenlignet og diskutert i kapittel 5, og det er disse som skal gi svar på problemstillingene i denne masteroppgaven. Diskusjonen i kapittel 5 omfatter forøvrig sammenligning av mine resultater med resultater fra tilsvarende undersøkelser.

Oppsummering, konklusjoner, perspektiver og konsekvenser er å finne i kapittel 6.

2. Kapittel Oppgaveformater, forskning og psykometri

Dette kapittelet handler om hva som kjennetegner oppgaveformatene åpne og lukkede oppgaver, i hvilken grad de ulike oppgaveformatene kan bidra til kartlegging av kompetanse, og hva som er fordelen og ulempen med hvert av formatene. I tillegg forklarer og tydeliggjør jeg en del begreper som jeg stadig kommer tilbake til når enkeltoppgaver og hele prøver kvalitetssikres. Kvalitetssikring av prøvene som jeg har benyttet, var nødvendig for å vite i hvilken grad vi skal kunne stole på resultatet av analysene.

2.1. Lukket oppgave (Selected Response Item, SR)

2.1.1 Hva er en lukket oppgave?..... og litt historikk

I en lukket oppgave (Selected Response Item, SR) skal eleven ikke formulere svaret selv, men velge blant alternativer som er oppgitt. Lukkede oppgaver kan klassifiseres på mange måter både ut fra hvordan opplysningene er gitt i oppgaven, antall svaralternativer som elevene skal ta stilling til og hvordan svaralternativene presenteres. Dette er omtalt senere i dette kapittelet.

I Norge har vi ingen lang tradisjon for bruk av lukkede oppgaver i tester i forbindelse med utdanning, selv om bruken av slike tester har sin opprinnelse fra tidlig på 1900-tallet. Den første publiserte testen av lukkede oppgaver bestod av flervalgsoppgaver (MC) og ble utviklet av den amerikanske læreren Fredrick Kelly i 1914 (The Kansas Silent Reading Test) (Monroe 1918, Sirnes 2007). Bakgrunnen var at man i skolen hadde behov for et enkelt hjelpemiddel til å måle leseferdigheten til elevene. Testen som Kelly utviklet, ble tilgjengelig for offentlig distribusjon i september 1915 (Monroe 1918), og var forløperen for flervalgstestene *Army Alpha* og *Army Betha* som ble brukt til utstrakt testing av rekrutter i den amerikanske hæren under 1. verdenskrig. Tester med lukkede oppgaver er derfor ikke noe nytt fenomen, og helt siden 1950-tallet har det vært mulig å bruke maskiner til automatisk å lese av og analysere resultater.

2.1.2 Formater av lukkede oppgaver

For at vi skal kunne stole på resultatene fra en undersøkelse, er det viktig å følge anbefalte retningslinjer ved utvikling av oppgavene. Haladyna mfl. (2002) har gjennomført en studie hvor de undersøkte validiteten (gyldigheten, se kapittel 2.3.3) til 31 retningslinjer for

konstruksjon av lukkede oppgaver. Retningslinjer for utvikling av oppgaver er ikke et fokus i denne masteroppgaven, men de sju oppgaveformatene som Haladyna mfl. (2002) brukte i undersøkelsen, er et godt utgangspunkt for meg når jeg skal redegjøre for lukkede oppgaver. Jeg vil på denne måten ikke omtale alle formater av lukkede oppgaver, men bare de som er mest relevant for denne masteroppgaven. Jeg har for øvrig kun benyttet lukkede oppgaver med bare ett rett svar.

Det tradisjonelle flervalgsformatet (Multiple Choice, MC)

Den tradisjonelle flervalgsoppgaven blir sett på som prototypen på en lukket oppgave. Det er også den oppgavetypen jeg har benyttet i størst grad i denne masteroppgaven. En flervalgsoppgave (MC) består av en innledende opplysning, *stimulus,* og i tilknytning til denne et spørsmål eller et utsagn som forteller hva eleven skal gjøre i oppgaven. Spørsmålet eller utsagnet kalles *stammen*. Til oppgaven hører svaralternativer, og det vanligste antallet er tre til fem. Bare et av svaralternativene representerer riktig løsning på oppgaven. Den riktige løsningen kalles *nøkkel*, og de gale alternativene som er oppgitt, kalles *distraktorer*. I figur 2.1 er stammen et spørsmål, nøkkelen er alternativ B (100 ml vann) og oppgaven har tre distraktorer A (50 ml vann), C (5 gram salt) og D (10 gram salt).

Figur 2.1 *Et eksempel på en oppgave i formatet flervalg med fire svaralternativer og ett riktig svar. Oppgave i Test 1*

Oppgave B 3

David løser opp 10 gram salt i 100 ml vann.

Hva må han tilsette den opprinnelige løsningen for å få en løsning som er halvparten så konsentrert?

A 50 ml vann
B 100 ml vann
C 5 gram salt
D 10 gram salt

Det er gjort undersøkelser for å finne ut om stammen bør være et spørsmål eller i fortellende form som en setning. Undersøkelsene viser ingen forskjell verken i diskriminering (hvordan oppgavene skiller mellom elevene, se kapittel 2.3.4) eller reliabilitet (pålitelighet, se kapittel

2.3.1) ved bruk av disse to formuleringsmåtene. Men det er spørsmålsformen som anbefales, fordi et spørsmål fører eleven mer direkte til den sentrale ideen i oppgaven (Haladyna mfl. 2002). Flervalgsoppgaver (MC) anbefales til alle typer undersøkelser, og S. Downing (Downing 2006) mener blant annet at alle grader av kognitive nivå kan testes med denne typen oppgave.

Valg mellom to alternativer (Alternate Choice, AC)

Alternate Choice er egentlig en flervalgsoppgave (MC) med bare to svaralternativer. Haladyna mfl (2002) viser til forskere som gjennom artikler på 1970- og 1980-tallet, har argumentert sterkt for dette formatet. Et argument er at formatet er veldig anvendelig både ved at det er lett å lage oppgaver, og at oppgavene er enkle å vurdere. Innvendinger som skyldes frykt for at elevene skal kunne gjette seg til mange riktige svar, møtes med to motargumenter. Det ene er at hvis man har mange oppgaver i en test, vil gjetting ha liten innvirkning (Downing 2003, 2006). Det andre er at man kan kompensere for gjetting ved å kreve minst 50 prosent riktige svar for å bestå testen. Oppgavetypen anbefales både til tester på klassenivå og til stor-skala tester, og forskere mener oppgavetypen burde vært mer benyttet (Haladyna mfl 2002).

Rett eller galt (True False, TF)

I praksis fungerer dette alternativet som *Alternate Choice*, men er ulikt ved at man skal ta stilling til et utsagn ved å svare ja/nei, rett/galt eller lignende. Grosse & Wright (1985) mente at gjetting kunne medføre en stor trussel mot *True False*, og mange forskere var enig i dette. Haladyna (1999) så også problemer med dette formatet. Det er imidlertid Ebel mfl. (1991) som har foretatt den mest omfattende studien av *True False*. De sier at det er viktig å ha ferdigheter og dyktighet til å kunne skille ut det som er essensielt. Et utsagn kan alltid formuleres slik at det enten er sant eller usant, og hvor mye kunnskap en elev har om et emne, kan bestemmes ved antall riktige svar. Ebel mfl. (1991) mener at *True False* kan teste høyere kognitivt nivå, men at dette avhenger av kompetansen til oppgaveutviklerne. En fordel med *True False* er at man kan prøve elevene i svært mange oppgaver på relativt kort tid, og dette øker påliteligheten til resultatene (se kapittel 2.3.1 og 2.3.3). Som i *Alternate Choice* kan det også her vurderes om man skal kreve at elevene må ha mer enn 50 prosent riktige svar for å bestå testen. Oppgaveformatet egner seg til tester på klassenivå (Haladyna 1999).

Multiple True False (MTF) er et lukket format som kan sees på som en sammensmelting av *Multiple Choice* og *True False*, ved at elevene her skal vurdere hvert av svaralternativene med rett eller galt. Dette er det andre lukkede formatet som jeg har benyttet i min masteroppgave (se figur 2.2). Downing mfl. (1995) fant at *Multiple True False* gir mer reliable resultater enn *Multiple-Choice*, men at *Multiple True False* vanligvis tester et lavere kognitivt nivå. Med et omfang på minimum 30 oppgaver blir imidlertid faren for at blind gjetting skal påvirke resultatet, ubetydelig.

Figur 2.2 *Et eksempel på en oppgave i formatet flervalg, type multiple true false. Oppgave i Test 1. Alle fire utsagnene må være riktig besvart for å få poeng*

Oppgave A 29

Nedenfor finner du flere utsagn om magnetisme. Vurder om hvert utsagn er riktig eller galt. **Sett ring rundt "Riktig" eller "Galt" for hvert utsagn.**

Utsagn	**Riktig eller galt?**
Magneter trekker til seg metaller og glass	Riktig / Galt
Jorda er en stor magnet	Riktig / Galt
Magneter kan støte fra seg andre magneter	Riktig / Galt
Ulike poler frastøter hverandre	Riktig / Galt

Hva som hører sammen (Matching, M)

Hvis det er listet opp en del begreper som skal kobles med en figur eller et sett av andre begreper (se figur 2.3), har vi en oppgave i formatet *Matching.* Lite forskning foreligger på denne oppgavetypen, og den anbefales derfor ikke til stor-skalatester (Downing 2006, Haladyna mfl. 2002). Oppgaveformatet er imidlertid mye brukt i gode fremstillinger i lærebøker, er godt egnet i pedagogiske sammenhenger og er svært anvendelig til tester på klassenivå. *Matching* i form av å sette bestemte ord inn på rett plass i en setning eller trekke forbindelse mellom begrep og ord som "hører sammen", er det tredje og siste lukkede formatet som jeg har benyttet i testene til masteroppgaven min.

Figur 2.3 *Et eksempel på en oppgave i formatet flervalg, type matching. Oppgave i Test 2, prøve 1*

Oppgave fA7

Sett strek mellom ordene på sidene og riktig sted på planten i midten.

Stengel

Støvbærer

Kronblad

Blad

Rot

Pollenknapp

Sammensatt flervalg (Complex MC)

Dette er en oppgavetype hvor hvert svaralternativ består av mer enn ett svar (se figur 2.4). Svarene er satt sammen i grupper. Dette medfører at det er tilstrekkelig for en elev å kjenne igjen et av begrepene som står nevnt i en svargruppe, for å avgi riktig svar. Dette gir både lav pålitelighet for resultatene og medfører at oppgaven vil skille dårlig mellom elevenes kompetanse. Formatet krever mer administrering, er mindre effektivt enn andre flervalgsformat, er plasskrevende i en test, og gir mindre pålitelig resultat. Haladyna mfl. (2002) anbefaler derfor ikke oppgaveformatet.

Figur 2.4 *Oppgave i formatet sammensatt flervalg*

Hvilke av de som er listet opp nedenfor, er et eksempel på en kjemisk reaksjon?

1. Vann som koker
2. En eplebit som blir brun
3. Sølvtøy som blir svart

A 1 & 2
B 2 & 3
C 1 & 3
D 1, 2 & 3

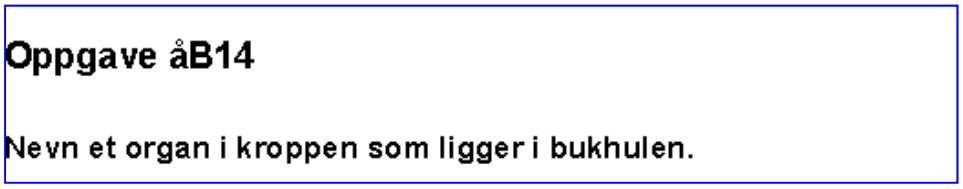
Kontekstavhengig item og sett av item (Context Dependent Item Set)
Oppgavetypen egner seg til å teste evne til å tenke, analysere og vurdere på et høyt kognitivt nivå. Innledende informasjon (stimulus) kan ha ulike uttrykk, som for eksempel en tegning, en tabell, en graf, et kart eller en sammenhengende tekst. Etter innledningen kommer et spørsmål som fører til første trinn i oppgaven. Deretter følger flere oppgaver i tilknytning til teksten, og i alle trinn av oppgaven er det flere valgmuligheter. Haladyna mfl. (2002) anbefaler dette formatet fordi det har en evne til å simulere tankeprosesser i flere trinn og derfor egner seg til problemløsing. Innvendinger mot denne oppgavetypen er at den tar mye plass i en test, og at det tar mye tid både å utvikle og vurdere testen. Et annet problem er at det kan være avhengighet innenfor oppgaven, ved at man ved å velge et svar er bundet i forhold til neste valg som kan gjøres. Da bryter man prinsippet om lokal uavhengighet som er en forutsetning for psykometriske modeller.

2.2. Åpen oppgave (Constructed Response Item, CR)

2.2.1 Formater av åpne oppgaver

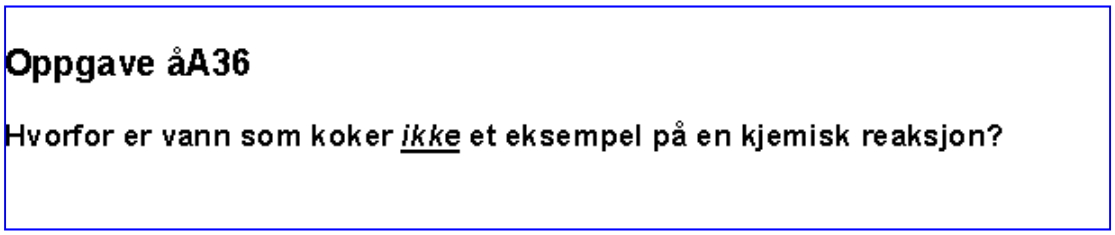
Åpne oppgaver (Constructed Respons Items, - CR) består av stimulus og stamme, har ingen svaralternativer, og personen som testes, skal formulere et svar med egne ord (Lukhele mfl. 1994, Martinez 1999). Svaret kan være alt fra et ord til en fortelling. En av oppgavetypene er kortsvaroppgaver (Brief Constructed-Response, - BCR). En kortsvaroppgave kan bestå i å forklare et diagram eller et begrep, redegjøre for fakta, navngi deler av en figur eller fullføre en setning. Oppgavene kan kreve svar på både lavt og høyt kognitivt nivå (se figur 2.5 og 2.6).

Figur 2.5 Oppgave i Test 2, prøve 1, som krever tenking på et lavt kognitivt nivå



I oppgave åB14 spør man etter faktastoff og detaljer. For å kunne svare på oppgaven, trenger personen som testes, å kjenne begrepene organ og bukhule og plasseringen til ett indre organ. Dette krever ikke forståelse, men at personen kan begrepene.

Figur 2.6 Oppgave som krever tenking på et høyere kognitivt nivå. Oppgave i Test 2, prøve 2



Oppgaven i figur 2.6 krever at personen som skal svare, kan anvende kunnskap. Her kan man ikke bare ramse opp en ferdiglært definisjon. Det kreves refleksjon i forhold til et konkret eksempel som vil vise i hvilken grad personen har forstått begrepet kjemisk reaksjon.

En annen type åpne oppgaver (Extended Constructed Response, - ECR) er mer arbeidskrevende, og krever mer sammensatte kunnskaper og ferdigheter. Oppgavene kan for eksempel være å løse et matematisk problem i flere trinn, lage et diagram eller en graf, gjøre lengre utgreiinger og å løse sammensatte problemer. Videre kan det være å skrive fortellinger og i naturfag å lage rapporter til vitenskapelige eksperimenter. All form for muntlig aktivitet og opptredener hører med blant åpne oppgaver av denne typen. En fordel med ECR er muligheten til å foreta en finere nivåinndeling av svarene, og å gi poeng for grad av måloppnåelse. Det kan for eksempel være å skrive et essay, hvor man vurderer innhold, setningsbygning, ortografi og struktur, og ut fra vurderingskriterier graderer måloppnåelse innenfor alle disse områdene.

Det har tidligere vært en utbredt oppfatning at kunnskaper på et høyt kognitivt nivå, bare kan måles med åpne oppgaver (Guilford 1967, Boodoo 1993, Messick 1993). Forskning de siste ti til femten år har imidlertid endret på denne oppfatningen. Hva som kan måles med lukkede oppgaver, avhenger bare av kompetansen til oppgaveutviklerne (Haladyna 2002, Downing 2006).

2.3. Sentrale statistiske begreper

Undersøkelser som er gjennomført på åpne og lukkede oppgaver, referer til verdier på psykometriske begreper som kommer fram gjennom analyser. Dette er begreper som det stadig blir referert til i denne masteroppgaven. I de neste delkapitlene vil jeg derfor klargjøre disse begrepene før jeg fortsetter teorien om åpne og lukkede oppgaver.

2.3.1 Reliabilitet

Reliabilitet er et uttrykk for kvalitet, betyr egentlig pålitelighet, og er et mål for hvor god en måling er. Skal vi kunne stole på analysene av et resultat, må vi vite i hvor stor grad resultatet gir et riktig bilde av situasjonen, og hvor mye som skyldes tilfeldigheter. Begrepet reliabilitet ble opprinnelig utviklet innenfor den grenen av psykologisk forskning som kalles psykometri (Kleven 2002). Det klassiske reliabilitetsbegrepet var bare knyttet til påliteligheten til en måling på det tidspunktet målingen ble foretatt. I hvilken grad målingen kunne gjentas med samme resultat, eller om man i virkeligheten målte det man mente å måle, var ikke et fokus i begynnelsen. Når vi nå snakker om reliabilitet i undersøkelser, er det imidlertid tre aspekter vi tenker på. Det første kalles *stabilitetsaspektet*. Med det menes om en person som gjennomfører en test på ett tidspunkt, vil kunne få samme resultat dersom han tar samme test på et annet tidspunkt. Dette er det vanskelig å teste ut siden vi hele tiden utsettes for påvirkninger. Det er også grunn til å tro at gjennomføring av en test kan påvirke resultatet ved neste gjennomføring av samme test på de samme personene. Av den grunn er det umulig å si hva som eventuelt er årsaken til at to testresultater ikke stemmer overens. Her kan det for eksempel være at tilfeldige målefeil som en persons dagsform, luft- og lysforholdene i prøvelokalene og andre situasjonsbestemte hendelser er blitt avgjørende for resultatet.

Det andre aspektet som kan påvirke påliteligheten til et resultat, kalles *ekvivalensaspektet* (Kleven 2002). Da spør vi oss om resultatet ville blitt det samme om vi hadde stilt spørsmålene på en annen måte, dvs. om vi hadde erstattet oppgavene i testen med et tilsvarende sett av oppgaver. Dette kan måles ved at vi lar en test på for eksempel 20 oppgaver bestå av 10 par oppgaver hvor spørsmålene i hvert par er så like at oppgavene kan erstatte hverandre. Vi kan da dele prøven i to ”like” halvdeler og finne summen for alle elevene til sammen for hver halvdel. Ved å sammenligne summene kan vi undersøke i hvilken grad delene måler det samme, korrelerer. Denne måten å undersøke påliteligheten av resultatet på en test, kalles halveringsmetoden (Kleven 2002). Tallet som kommer fram kalles korrelasjonskoeffisienten, og er et uttrykk for reliabiliteten til en prøvehalvdel (r_{halv}).

I følge Spearman-Browns formel, kan vi da regne ut reliabiliteten for hele prøven (r_{hel})

$r_{hel} = \frac{2r_{halv}}{1+r_{halv}}$, og slik finne ut i hvilken grad ulike måter å stille spørsmålene på har påvirket resultatet.

I de fleste tilfeller har vi ikke en prøve som er laget slik at den kan deles i to like halvdeler. Da kjører vi en test i SPSS[4] som deler prøven i to halvdeler på alle mulige tenkelige måter, og regner ut den gjennomsnittlige reliabilitetskoeffisienten for alle halveringene. Dette er Cronbachs alfa (r) for prøven, og forteller noe om prøvens indre konsistens. Med det menes i hvilken grad det er sannsynlig at resultatet på prøven blir det samme om vi bytter ut alle oppgavene i prøven med et tilsvarende sett av oppgaver. I dette tilfellet ser vi på de oppgavene som er valgt ut, som et tilfeldig utvalg av alle spørsmål som kunne vært stilt om emnet, og dette er nettopp ekvivalensaspektet ved reliabiliteten. Fordi spørsmålene som er valgt ut til prøvene kan sees på som et tilfeldig utvalg av aktuelle oppgaver, forteller alfa også i hvilken grad man kan generalisere fra resultatene på de oppgavene som er med i prøven. Det er imidlertid flere faktorer enn reliabilitet som må oppfylle visse krav for at en undersøkelse skal være generaliserbar. Ikke minst er det viktig hvordan utvalgene som vi analyserer resultatene fra, er satt sammen. Dette kommer jeg tilbake til senere i oppgaven.

Det tredje aspektet ved reliabiliteten er vurderingen. Hvis en test har åpne oppgaver, kan det i noen tilfeller være mulig å tolke et svar på flere måter, og da er det alltid en mulighet for at subjektiv oppfatning kan påvirke resultatet. For å øke reliabiliteten kan vi ha minst to sensorer som vurderer alle besvarelsene. Da kan sensorreliabiliteten beregnes ved å se på hvordan resultatene til de to som vurderer, samsvarer (korrelerer).

Et tiltak for å øke reliabiliteten i åpne oppgaver, er å utarbeide detaljerte vurderingsinstrukser, det vil si å standardisere datainnsamlingen. Et annet tiltak kan være å øke antall oppgaver. Hvis den gjennomsnittlige korrelasjonen mellom oppgavene i en test for eksempel er r_1 (gjennomsnittlig inter-item korrelasjon) og antall oppgaver i testen er n, kan vi regne ut Cronbachs alfa (r) ved hjelp av Spearman-Browns formel: $r = \frac{nr_1}{1+(n-1)r_1}$. Da ser vi at ved å øke antall oppgaver (n), vil også alfa øke, og vi kan beregne hvor mange oppgaver vi trenger for å få den reliabiliteten vi ønsker. Vi forutsetter da at det er samme type oppgaver som brukes, slik at også testen med de nye oppgavene har samme gjennomsnittlige inter-item korrelasjon. Hvis vi for eksempel har en test med en inter-item korrelasjon på 0,12, kan vi ved hjelp av formelen beregne at en test med 28 oppgaver gir reliabiliteten 0,79. Med 40 oppgaver blir reliabiliteten 0,85.

[4] SPSS: Statistical Package for the Social Sciences

Lukkede oppgaver (SR) ansees for å være mer reliable enn åpne oppgaver. I lukkede oppgaver kreves ingen tolkning når man vurderer. Her skal elevene krysse av for et svar, og svaret blir enten rett eller galt. Man trenger egentlig ikke fagpersoner for å vurdere disse oppgavene. Alt som trengs er en fasit som viser hvilke svaralternativer som gir rett svar. I lukkede oppgaver kan man derfor helt se bort fra tilfeldige målefeil, i forbindelse med vurderingen, og man har derfor grunn til å anta at det er 100 prosent samsvar mellom vurderingen til to sensorer som vurderer de samme besvarelsene, - sensorkorrelasjonen er perfekt positiv.

Korrelasjonen (samsvaret) er et tall mellom -1 (perfekt negativ korrelasjon), via 0 (ingen korrelasjon) og 1 (perfekt positiv korrelasjon). Hvis prøven er fullstendig reliabel, er korrelasjonen lik 1. Det kan imidlertid skyldes at oppgavene er for like, og er ikke nødvendigvis gunstig. Hvis man skal teste et emne, er det viktig å ha oppgaver med ulike innfallsvinkler i forhold til emnet, slik at alle elever har mulighet til å få vist sin kompetanse.

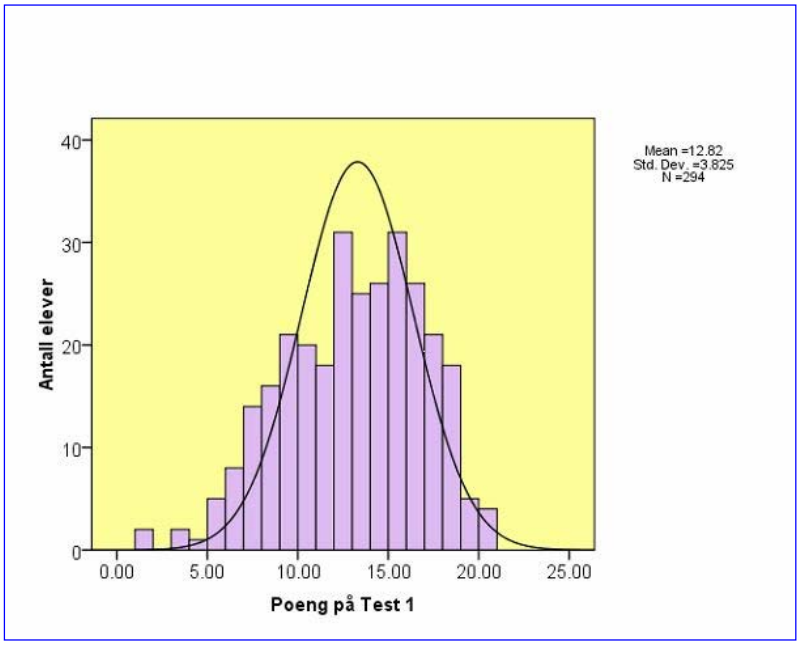
Hvilken korrelasjon som er akseptabel, avhenger av hva som er hensikten med prøven. En prøve som skal si noe om prestasjonene til enkeltelever, skal ha høy reliabilitet. Ved utvikling av de nasjonale prøvene i regning i Norge, setter rammeverket (Utdanningsdirektoratet 2006) et krav om at Cronbachs alfa (r) minst skal være lik 0,85 (Lie mfl 2005). Det betyr at oppgavene i en prøve skal representere det faglige innholdet som skal måles i så stor grad at vi kan stole på at 85 prosent av resultatet skyldes elevenes faglige kompetanse i emnet (er sann skåre), mens 15 prosent skyldes tilfeldigheter (feil skåre).

2.3.2 Standardavvik, standard målefeil og standardfeil til gjennomsnittet

En annen faktor som har betydning i en undersøkelse, er spredningen i resultatene. Standardavviket er et spredningsmål som er mye referert til i statistiske undersøkelser. Hvis vi for eksempel undersøker en gruppe på 1000 elever, vil vi vanligvis finne noen som skårer lavt, mange som svarer riktig på fra 40 til 70 prosent av oppgavene og noen som besvarer mer enn 70 prosent riktig. Spredningen blant resultatene i en undersøkelse, har betydning når vi skal snakke om resultatene til enkeltelever.

Standardavviket (s) regnes ut ved formelen: $s = \sqrt{\frac{\sum x^2}{N}}$. Innenfor et område på middelverdien $\pm 1s$, finner vi ut fra poeng ca. 68 prosent av elevene som har deltatt i en undersøkelse. Innenfor $\pm 2s$ finner vi ca. 95 prosent av alle deltakerne (se figur 2.7).

Figur 2.7 *Et eksempel som viser middelverdi lik 12,8 poeng og standardavvik lik 3,8 poeng. Det betyr at 68,26 % av alle personene har poengsummer som ligger mellom 9 og 16,6 poeng, og at 95 % har poengsummer fra 5,2 til 20 poeng. Eksempelet er resultater fra Test 1*



Spredningen spiller rolle både for hvor godt vi kan estimere poengsummene for de enkelte elevene, og for hvor nøyaktig vi kan bestemme gjennomsnittet til en gruppe elever. For å bestemme poengsummen til hver elev, må vi vite standardfeilen til målingen (s_m). Den er bestemt av standardavviket (s) og reliabiliteten (r) ved formelen

$$s_m = s\sqrt{1-r}$$

Hvis standardavviket er 3,8 poeng og reliabiliteten er 0,79, blir standard målefeil 1,7 poeng. Det betyr at en elev som oppnår 12 poeng på en prøve, med 68 prosent sannsynlighet har en poengsum som ligger mellom 10,3 og 13,7 poeng, og at sannsynligheten er 95 prosent for at poengsummen til eleven ligger mellom 8,6 og 15,4 poeng. Vi sier da at innenfor et 95 % konfidensintervall har eleven en poengsum som går fra 8,6 til 15,4 poeng (Kleven 2002). Det betyr også at sannsynligheten er mindre enn 5 prosent ($p < 0,05$) for at elevens poengsum er utenfor dette intervallet. Dette viser for øvrig hvor viktig det er at en undersøkelse som skal si noe om enkeltresultater, har høy reliabilitet.

Standardfeilen til gjennomsnittet ($s_{\bar{X}}$) henger sammen med standardavviket ved formelen

$$s_{\bar{x}} = \frac{s}{\sqrt{N}}$$

Standardfeilen til gjennomsnittet regnes ut ved å dividere standardavviket på kvadratrota av antall personer (N) som er med i undersøkelsen. Det betyr at jo større standardavvik, jo større blir standardfeilen til gjennomsnittet, men også at dette kan gjøres mindre ved at vi har et høyt antall personer med i utvalget. Statistiske tabeller viser at med 95 prosent sikkerhet kan vi estimere en gjennomsnittsverdi til å være $\pm 2 \cdot s_{\bar{x}}$. Dette er viktig når vi for eksempel skal finne ut om en forskjell i gjennomsnittsskåre mellom resultatene til jenter og gutter er reel eller skyldes tilfeldigheter. Hvis gjennomsnittsverdien til guttene $\pm 2 \cdot s_{\bar{x}1}$, er forskjellig fra gjennomsnittsverdien til jentene $\pm 2 \cdot s_{\bar{x}2}$, er forskjellen statistisk signifikant innenfor et konfidensintervall på 95 prosent. Det betyr at hvis vi undersøker 100 nye tilfeldig trukket utvalg innenfor samme populasjon av jenter og gutter, vil vi teoretisk sett i 95 av tilfellene få en gjennomsnittsverdi som ligger i intervallet $\pm 2 \cdot s_{\bar{x}1}$ for guttene og $\pm 2 \cdot s_{\bar{x}2}$ for jentene.

2.3.3 Validitet

Validitet betyr gyldighet (Kleven 2002), er ikke en egenskap ved en prøve, og kan derfor ikke tallfestes på samme måte som reliabilitet. Hvis en undersøkelse skal være valid, må den handle om det emnet man ønsker å måle. I praksis vil det si at dersom det er godt samsvar mellom et begrep slik det er definert teoretisk og det man spør om i oppgavene, så har undersøkelsen høy validitet. Kleven (2002) kaller dette begrepsvaliditet.

Vi setter oss et mål, men om vi når målet, avhenger av hvordan vi definerer kjennetegnene på at målet er nådd. Her er det spesielt to problemer. Det ene kan være at enkelte sider ved begrepet er underrepresentert i målingen, dvs. at vi lager oppgaver som har den skjevheten at de bare måler en del av begrepet. Det andre oppstår når oppgavene inneholder forstyrrende elementer, dvs. elementer som er irrelevante for målingen, og som kan påvirke testpersonens resultat. Skal vi kunne stole på resultatene, må vi vite hva vi måler med en oppgave.
En måling må sees i forhold til det den skal brukes til. Kleven (2002) ser på begrepsvaliditeten som den overordnede formen for validitet ved måling, men sier at den kan erstattes av underaspektet innholdsvaliditet i tester som måler opp mot klart definerte mål. Med validitet i kunnskapsprøver til skolebruk menes derfor innholdsvaliditet i form av at vi vurderer i hvilken grad det er samsvar mellom det som kreves for å oppnå et godt resultat på en prøve, og læreplanens kompetansemål (LK06) i emnet. Ved å bruke flere ulike

målemetoder, kan vi øke validiteten. Robson (2002) mener imidlertid at å innføre flere måter å måle på, ikke nødvendigvis er noen "mirakelkur" siden alle målemetoder har sine svakheter.

2.3.4 Diskriminering, t-test, effektstørrelse og p-verdi

For at en test skal ha god kvalitet er det viktig at det er et visst samsvar mellom oppgavene i testen. Den enkelte oppgave skal samsvare (korrelere) med hver enkelt av de andre oppgavene. Samtidig skal hver enkelt oppgave korrelere med summen av alle oppgavene. Denne korrelasjonen kalles point-biserial og viser i hvilken grad oppgaven diskriminerer, dvs bidrar på riktig måte i testen. Verdier som viser hvordan en oppgave diskriminerer kan bestemmes ved at vi bruker Pearsons korrelasjonstest, produktmomentkorrelasjon. Oppgavene som måler det samme skal ligne på hverandre, men ikke for mye og ikke for lite. En tommelfingerregel for at en oppgave skal bidra positivt i en test, er at point-biserial (diskrimineringen, d) skal være større enn 0,3 (Lie mfl. 2005).

Man kan imidlertid ikke se bare på point-biserial for å finne ut om en oppgave har god nok kvalitet til å være del av en test. Man må også se på *dyktigheten* til elevene, det vil si se på gjennomsnittlig poengsum til gruppen av elever som har løst de forskjellige oppgavene. En elev som oppnår høy poengsum på en prøve, skal ha større sannsynlighet enn en elev med lav poengsum for å løse oppgaver som få elever har funnet løsningen på. Dette er et krav som må oppfylles for at vi skal kunne stole på resultatene til en oppgave i en test.

En test som kan brukes til å undersøke om forskjellen mellom to gjennomsnittsverdier er signifikante, kan være Dunetts t-test (Kleven 2002, Robson 2002). T-tester skal egentlig bare brukes på normalfordelte populasjoner og de populasjonene som sammenlignes, skal ha tilnærmet samme spredning (Kleven 2002). I store utvalg spiller ikke dette så stor rolle. Effektstørrelse er et annet begrep som forteller om forskjell i gjennomsnittsverdier til to grupper. Effektstørrelse har benevning standardavvik og beregnes ved å dividere differansen til gjennomsnittsverdiene i de to gruppene, med standardavviket. Det er vanlig å bruke "pooled" standardavvik, dvs. kvadratrota av gjennomsnittet til kvadratet av standardavvikene til de to gruppene som undersøkes. Effektstørrelse som er mindre eller lik 0,3 betyr liten effekt ($e \leq 0{,}3$). Hvis verdien ligger mellom 0,3 og 0,8 er effekten middels ($0{,}3 < e < 0{,}8$), og en verdi større eller lik 0,8 betyr at effekten av hvilken gruppe man tilhører er stor ($e \geq 0{,}8$).

Løsningsprosenter oppgis som p-verdier (p). Hvis p-verdien til en oppgave er 0,59, har

59 prosent av elevene funnet riktig løsning på oppgaven. Symbolet p brukes i tillegg for å angi sannsynligheter. Dette er omtalt i forbindelse med standard målefeil i kapittel 2.3.2.

2.4. Tester med lukkede oppgaver

Mange har forsket på oppgaveformat og hvilke ferdigheter man kan få testet ved de ulike formatene (Haladyna 2004, Lukhele mfl. 1994, Martinez 1999, Wester 1995). Til vanlig er kanskje lukkede oppgaver mest brukt i sammenhenger hvor man måler kognitive prosesser på et lavt nivå, mens åpne oppgaver forekommer mest når det er snakk om å måle sammensatte tankeprosesser. Martinez (1999) mener dette imidlertid er mer et spørsmål om hva som er mest vanlig å gjøre enn hvilke begrensninger hvert format har, og det finnes forskning som bekrefter dette (Downing 2006). Messick (1993) mener at selv om flervalgsoppgaver kan måle kunnskaper på et høyt kognitivt nivå, så kan åpne oppgaver måle bredere, slik at alle typer åpne oppgaver derfor ikke kan erstattes av flervalgsoppgaver.

Det har vært mye skepsis knyttet til tester med oppgaver i lukket format. Martinez (1999) omtaler mange undersøkelser som ble gjennomført i perioden 1967 til 1998, som viser at det imidlertid etter hvert har skjedd en utvikling i forhold til anerkjennelse av det lukkede oppgaveformatet. Guilford (1967) mente at flervalgsoppgaver ikke kunne brukes for å teste alle områder som burde testes i forbindelse med undervisning og utdanning. Han mente at flervalgsoppgaver ikke oppmuntret til kreativ tenkning, og uttrykte dyp bekymring med tanke på hva testing utelukkende med flervalgsoppgaver ville kunne gjøre med intellektet til en nasjon. Selv om åpne oppgaver ikke ble utelukket fra testene, uttrykte 25 år seinere G. M. Boodoo (1993) også bekymring over at flervalgsoppgaver mer og mer ble det dominerende oppgaveformat i alle typer tester (Martinez 1999).

Andre konkluderte motsatt ut fra forskning de hadde gjennomført, og sa at lukkede oppgaver er det best egnete formatet til å måle kognitive ferdigheter og evner, og spesielt høyere orden av kognitive ferdigheter som problemløsning, å kunne se sammenhenger, vurdere, tolke og analysere (Downing 2006, Haladyna 2004). Med ferdigheter menes i denne sammenheng begreper, fakta, naturlover og anvendelse av disse i nye situasjoner. Når det gjelder begrensningene i det lukkede formatet, handler det mer om svakheter ved utvikling av oppgavene, både når det gjelder formulering av stimulus, stamme og distraktorer, enn selve oppgaveformatet. Hvis den som lager oppgavene selv er faglig dyktig innen området som skal

testes, følger retningslinjene for utvikling av oppgaver, og tester oppgavene i forhold til psykometriske krav, unngår man oppgaver som ikke fungerer i en test (Downing 2006, Haladya mfl. 2002).

Spesielt i storskala-undersøkelser er tester med lukkede oppgaver godt egnet, og også når man skal trekke slutninger om store emneområder. Med lukkede oppgaver har man muligheten til å knytte bilder og figurer opp mot spørsmålene, og slik oppnå større bredde i oppgavene. Dette gir testen høyere validitet. I tillegg kan en test med lukkede oppgaver inneholde flere oppgaver enn en tilsvarende test med åpne oppgaver når man har en bestemt tid til disposisjon (Kleven 2002, Robson 2002, Wester 1995). Å løse en lukket oppgave kan for eksempel ta fra et halvt til 10 minutter, mens å skrive et essay kan ta opp til flere timer. Med mange oppgaver vil en test med lukkede oppgaver bli både mer reliabel og få høyere validitet, og vi får mer pålitelige resultater enn for en test med åpne oppgaver.

Dette støttes av en undersøkelse Lukhele mfl. utførte i 1994. Testen inneholdt like mange åpne som lukkede oppgaver. Resultatene viser at det tok lenger tid for elevene å løse de åpne enn de lukkede oppgavene, samtidig som det var mindre informasjon som kunne hentes ut av de åpne oppgavene. På den tiden en elev brukte for å løse en åpen oppgave, kunne han løse 16 lukkede oppgaver. En test med lukkede oppgaver som tok 75 minutter, var like reliabel som en test på 185 minutter med åpne oppgaver, og det kostet 300 ganger mer å vurdere en test med åpne oppgaver. Det tar lenger tid å vurdere en åpen enn en lukket oppgave, og åpne oppgaver blir derfor dyrere å benytte. Undersøkelsen var en del av *The Advanced Placement testing program* og ble gjennomført med *high school* elever.
En annen fordel med oppgaver i lukket format er at den som vurderer ikke har mulighet til å påvirke resultatet. Objektivitet er en viktig faktor i all testing. Downing mfl. (2006) gjennomførte en undersøkelse som viste høy sensorkorrelasjon og derfor høy reliabilitet for flervalgsoppgaver. Hvis testene i tillegg gjøres computerbaserte, får man automatisk retting, og et lettere utgangspunkt for å gjøre analyser. At reliabiliteten i en undersøkelse er høy nok, er avgjørende for om vi skal kunne stole på et resultat. Downing (2006) mener at dette er så avgjørende, at åpne oppgaver i skriftlige prøver bare skal brukes til å teste de områdene av et fagfelt som ikke kan testes med lukkede oppgaver. Han mener dette gjelder svært få områder, men kan være testing som krever sammensatt dyktighet på et høyt nivå, som for eksempel evne til skriftlig kommunikasjon, eller at det er praktisk dyktighet som skal måles.

Nettopp for å få høyere reliabilitet og validitet i de nasjonale prøvene i Norge i lesing på norsk, i regning og engelsk, ble det bestemt at prøvene fra 2007 skulle bestå av minst 70 prosent lukkede oppgaver (Utdanningsdirektoratet 2006). Selv om elevene bruker kortere tid på å løse en lukket enn en åpen oppgave, må man imidlertid ikke glemme at elevene må ha tid til både å kladde og å tenke seg om også når de løser lukkede oppgaver. Dette må elevene minnes på og oppmuntres til, særlig hvis testen er computerbasert.

2.5. Tester med åpne oppgaver

Oppgaver i åpent format, kan som vist i figur 2.5 og 2.6, brukes til å teste fakta, forståelse og anvendelse. I tillegg er åpne oppgaver den eneste måten å få testet elevenes evne til å skrive sammenhengende fortellinger. Grunnleggende motivasjon for å benytte åpne oppgaver, kommer av ideen om at de kan måle kunnskaper som ikke kan måles med lukkede oppgaver. Da tenker man for eksempel på dynamiske kognitive prosesser og å identifisere elevenes misoppfatninger i diagnostisk testing (Guilford 1967, Boodoo 1993, Messick 1993). Et annet argument er "eksempelets makt". Mange mener at å bruke åpne oppgaver er mer realistisk når det gjelder å forberede for situasjoner som elevene kan møte i hverdagen (Lukhele mfl. 1994). Storskala-tester har status og påvirkningskraft på undervisning. Hvis lærere og elever visste at elevene ville bli prøvd i problemløsing, i å tolke grafer og å uttrykke seg muntlig og skriftlig, ville dette kunne føre til at slike aktiviteter ble vektlagt i undervisningen, noe som er ønskelig. Det er også ønskelig at testene er effektive og minst mulig kostnadskrevende. Lukhele mfl. (1994) konkluderte imidlertid til slutt i sin artikkel med at det var ingenting som tydet på at åpne og lukkede oppgaver målte ulike ting.

I tillegg til at åpne oppgaver er tids- og kostnadskrevende er det en ulempe at personen som vurderer, i stor grad har mulighet til å la sin subjektive oppfatning påvirke vurderingen (Kleven 2002, Robson 2002). Bruk av vurderingskriterier kan ikke forhindre dette. Forskning har vist at sensorreliabiliteten i åpne oppgaver ofte er svært lav. Dette betyr at det kan være lite samsvar mellom resultatet som to ulike sensorer kommer fram til når de vurderer de samme besvarelsene (Kleven 2002, Robson 2002).

Både åpne og lukkede oppgaver kan inneholde mye tekst. I slike tilfeller kan elevene i lukkede oppgaver muligens få hjelp av svaralternativene. Denne muligheten finnes ikke i åpne oppgaver, og av den grunn kan det være at åpne oppgaver krever mer av elevene enn lukkede

når det gjelder evne til lesing og skriving. Dette kan være et forstyrrende element i forhold til å få en reliabel vurdering. Er det fagstoffet som er problemet, eller er det lese- og skriveferdigheten?

2.6. Åpne oppgaver i forhold til lukkede oppgaver

Det finnes rikelig med eksempler på at elevers forståelse av begreper er kontekst- og situasjonsavhengige (Hennessey 1993, Angell 1996). Hvordan elevene bruker begreper avhenger av sammenhengen de er satt inn i, og dette er ting som må tas hensyn til når man formulerer et spørsmål i en oppgave. Dette kan være årsaken til at elever svarer svært forskjellig på to oppgaver med samme innhold, men formulert på ulik måte. Lukkede oppgaver kan i slike tilfeller være en fordel ved at de presiserer spørsmålet gjennom svaralternativene.

Åpne og lukkede oppgaver kan virke ulikt på elevene ut fra kunnskapsnivå. Lukkede oppgaver kan være en fordel for lese- og skrivesvake elever, mens flinke elever som har god formuleringsevne, kan ha fordel av åpne oppgaver. Hvis en faglig flink elev mangler noe kunnskap for å besvare en oppgave og oppgaven er åpen, kan eleven tolke og avgrense oppgaven slik at han får vist det han kan, men også skjult manglende kunnskap. Den som vurderer besvarelsen blir usikker på elevens nivå, lar tvilen komme eleven til gode, og eleven får fordel av det åpne formatet (Martinez 1999).

Eksamensangst og andre tilfeldige målefeil er en trussel for validiteten. Forskning har vist at elever kan føle mer eksamensangst når oppgavene er åpne enn når de er lukkede (Crocker mfl. 1987). Mange elever føler trygghet når de kan velge blant svaralternativer.

Et annet område som er forsket på, handler om i hvilken grad elevenes læringsutbytte påvirkes av om de forventer seg en prøve med åpne eller lukkede oppgaver. Martinez (1999) har undersøkt forskning som er gjort på dette i perioden 1932 til 1998, og funnet ulike resultater. En undersøkelse som for eksempel ble gjennomført av Terry (1933) viste at elevene leste mest på detaljer hvis prøven skulle bestå av lukkede oppgaver, mens de så mer på hovedpunkter og oversikter hvis de forventet åpne oppgaver. Gay (1980) rapporterte ingen forskjell i resultatene på lukkede oppgaver i forhold til om studentene var opplært til å løse åpne eller lukkede oppgaver. Derimot var de studentene som var opplært til å løse åpne

oppgaver, signifikant bedre på de åpne oppgavene enn de studentene som tidligere kun hadde gjennomført tester med lukkede oppgaver. Kumar (1979) hadde annen erfaring. Han så ingen forskjell i studentenes resultater på prøver om de forventet et bestemt format eller ikke. Allikevel ser det ut til at oppfatningen heller i den retning at elevene forbereder seg ulikt om de forventer seg en test med åpne oppgaver i forhold til om de forventer en test med lukkede oppgaver, og at læringsutbyttet er større hvis de forventer åpne oppgaver enn om de forventer lukkede. Martinez (1999) konkluderer imidlertid med at dette er ikke nok dokumentert, og videre forskning er nødvendig.

Olsen mfl. (2001) spør om hva som er kriteriene for at vi skal kunne si at elever ”kan” eller ”forstår” noe. Betyr det at elevene må ha evne til å formulere et selvkomponert svar, eller er det så at elevene også ”kan” når de er i stand til å velge riktig alternativ i en lukket oppgave? Er disse to formatene like når det gjelder å avdekke elevenes kunnskaper, eller gir de tilgang til ulike dimensjoner av kunnskaper hos elevene?

Martinez (1999) refererer til studier som har sammenlignet oppgaver med samme innhold i åpent og lukket format. I noen tilfeller er det funnet så store ulikheter i svarene at det er konkludert med at oppgavene er forskjellige. I andre tilfeller er korrelasjonen mellom responsformatene så høy at man ikke kan si at oppgavene er ulike. I et tredje tilfelle ser man en antydning til at ulikt oppgaveformat gir ulike resultater. Traub (1993) undersøkte ni relevante studier og konkludert med at det ikke ble funnet nok bevis til å kunne trekke noen generelle konklusjoner om oppgaveformatets betydning. Det ble imidlertid konkludert med at formatets effekt var mer tydelig når det var lavt kognitivt nivå som ble målt (Martinez 1999). Dette kan skyldes at det i oppgaver med lav vanskelighetsgrad ofte er spørsmål om å ”gjenkjenne” i lukkede oppgaver og å ”gjengi” i åpne oppgaver. Motsatt gjelder for måling av mer sammensatt kunnskap. Jo mer krevende oppgavene er, jo bedre samsvar er det mellom resultatene til en åpen og en lukket oppgave (Martinez 1999).

Andre undersøkelser på oppgaveformat viser nesten perfekt korrelasjon mellom oppgaver med samme ordlyd i åpent og lukket format. Korrelasjonene ble korrigert for tilfeldige målefeil. Dette var funn fra 67 studier som Rodriguez (2003) sammenfattet. 29 av studiene inneholdt 56 korrelasjoner av oppgaver med samme innhold i åpent og lukket format. De 56 korrelasjonene hadde svært ulike verdier. Oppgaver som hadde samme formulering (lik stamme) og bare var ulike pga svaralternativene i det lukkede formatet, hadde høy

gjennomsnittlig korrelasjon. Denne var signifikant høyere enn når oppgavene ikke hadde samme formulering. Størst var forskjellen når den åpne oppgaven var i essay-format. Rodriguez (2003) konkluderte med at innholdsekvivalens delvis avhenger av oppgaveformatet og delvis av hvordan oppgaven er formulert.

2.7. Oppgaveformatets betydning i forhold til kjønn

Det foreligger mange studier på kjønnsforskjeller. En hypotese er at gutter har fordel framfor jenter når det gjelder flervalgsoppgaver, mens jenter skårer bedre enn gutter i åpne oppgaver. Dette er bekreftet i undersøkelser gjort av Murphy (1982), Bolger mfl. (1990) og Bridgeman mfl. (1994), men det har ikke vært enighet om hva resultatet skyldes. En forklaring som har vært antydet, er at jenter har større språklig evne og derfor er flinkere til å formulere seg i åpne oppgaver, mens gutter er flinkere til problemløsing.

DeMars (2000) har gjennomført en studie hvor hun så på sammenhengen mellom statusen til en test og formatet til oppgavene, og på sammenhengen oppgaveformat og kjønn. Hun fant at både for åpne og lukkede oppgaver skåret elevene i gjennomsnitt best på testen med høy status, og forskjellen i skåre på testene var størst i de åpne oppgavene. Resultatene i forhold til kjønn og oppgaveformat var helt i tråd med tidligere hypoteser ved at gjennomsnittlig p-verdi for guttene var høyere enn for jentene på lukkede oppgaver, mens jentene skåret bedre enn guttene på de åpne oppgavene.

En annen studie på oppgaveformat ble gjennomført av Bell mfl. (1987) i engelsk. De gjennomførte en tredelt test som bestod av en del med lukkede oppgaver og to deler med åpne oppgaver, hvor den ene testet forståelse og den andre friskriving. Jentene gjorde det bedre enn guttene på alle tre områdene. Kjønnsforskjellen var minst på de lukkede oppgavene. Guttene gjorde det imidlertid bedre relativt sett på de lukkede oppgavene enn på de åpne oppgavene. Ut fra dette er resultatet i tråd med tidligere oppfatninger om at gutter skårer bedre på lukkede enn åpne oppgaver, og at det er motsatt for jentene, selv om undersøkelsen viser at lukkede oppgaver ikke er til hinder for jentene.

Studier gjennomført av Wester-Wedman (1992c) bekrefter ikke hypotesen om at det er oppgaveformatet åpne oppgaver som favoriserer jentene i forhold til guttene. En studie viste at variasjoner i innhold og tema hadde større betydning for kjønnsforskjellene enn

oppgaveformatet. Holland mfl. (1986) har samme konklusjon ut fra egne undersøkelser. Da ble det konkludert med at temaet i oppgavene og hvilke løsningsstrategier som kan brukes, spiller en større rolle enn formatet.

Beller mfl. (2000) konkluderte med at i matematikk var det ingen generell regel som sa at det var kjønnsforskjeller relatert til oppgaveformat, men at vanskelighetsgrad og innhold var avgjørende for kjønnsforskjeller. Guttene gjorde det bedre enn jentene i problemløsningsoppgaver og sammensatte oppgaver, mens jentene gjorde det bedre enn guttene på lettere oppgaver som for eksempel handlet om beregninger. Det var viktigere å se på hva som ble målt enn hvordan målingen ble gjort.

Wester (1995) gjennomførte en studie på elever som gikk siste året på videregående skole, hvor man så på forskjell i resultater på åpne oppgaver og flervalgsoppgaver (MC). I denne undersøkelsen besvarte elevene de samme spørsmålene i to ulike formater. Oppgavene ble først prøvd ut som flervalgsoppgaver, og deretter endret til åpne oppgaver og prøvd ut på nytt på andre elever. Resultatene viste at guttene i gjennomsnitt gjorde det bedre enn jentene både da oppgavene var åpne og da de var flervalgsoppgaver. Forskjellen var imidlertid minst da oppgavene var flervalgsoppgaver, så dette tyder ikke på at jenter er best på åpne oppgaver, snarere tvert i mot. Testen bestod av 20 oppgaver, og to oppgaver som guttene hadde skåret best på som flervalgsoppgaver, skåret jentene best på da de var åpne oppgaver. Dette var problemløsningsoppgaver innen lesing og måleenheter, mens sju av de ni oppgavene som guttene gjorde det best på, var problemløsning som krevde en viss grad av beregning, omgjøring av brøker eller egenproduksjon. Egenproduksjon var begrenset til et tall, et ord eller en setning, og dette gjorde at endring fra flervalgsoppgave til åpen oppgave ikke burde ført til økning i forskjellen mellom kjønn. Konklusjonen kan være at kjønnsforskjellen på åpne og lukkede oppgaver er liten når de åpne oppgavene er av en slik art at de lett kan omformes til flervalgsoppgaver. Det ble konkludert med at forskjeller som allerede fantes mellom kjønn i flervalgsoppgaver, ble forsterket da oppgavene ble åpne (Wester 1995). Denne undersøkelsen viste også at jentene i større grad enn guttene lot oppgaver stå ubesvart, uansett hvilket format oppgavene var i.

Resultatene for ca. 20 000 av de 60 000 elevene som gjennomførte nasjonal prøve i regning for 8. trinn i 2009, viser at guttene gjorde det signifikant bedre enn jentene på de lukkede oppgavene som alle var flervalgsoppgaver med fire svaralternativer. Differansen var

ca. 5 prosentpoeng i guttenes favør. På de åpne oppgavene gjorde guttene det også best, men her var forskjellen bare 1 prosentpoeng i forhold til jentene. I tillegg gjorde guttene det ca. 11 prosentpoeng bedre på flervalgsoppgavene enn på de åpne oppgavene, mens tilsvarende differanse for jentene var ca. 7 prosentpoeng (Ravlo mfl. 2010). I denne undersøkelsen var forholdet mellom åpne og lukkede oppgaver i prøvesettet 3:7, og det var ulike oppgaver som var i åpent og lukket format. Prøven ble gjennomført elektronisk. 59 oppgaver skulle besvares på 90 minutter.

Hastedt mfl. (2005), konkluderer i en undersøkelse av resultater fra TIMSS 1995 og 1999, at det ikke var signifikant, men en klar tendens til at matematikkoppgaver i åpent format favoriserte jentene. Stort sett gjorde guttene det bedre enn jentene på både lukkede og åpne oppgaver, men forskjellen var minst i de åpne oppgavene. Resultatene viser i 17 av de 41 landene hvor TIMSS 1995 ble gjennomført, at guttene skåret signifikant bedre enn jentene på de lukkede oppgavene. Totalt på prøven skåret guttene bedre enn jentene i 13 land. I 5 land gjorde guttene det bedre enn jentene på de åpne oppgavene i matematikk. Det var også fem land hvor jentene gjorde det bedre enn guttene på åpne oppgaver i naturfag, og i 8 land skåret jentene i gjennomsnitt 10 poeng mer på de åpne enn på de lukkede oppgavene. I alle land unntatt Russland var differansen mellom gutters og jenters poengsum både på flervalgsdelen og den totale poengsummen, større enn differansen gutt minus jente i de åpne oppgavene i testen. I Norge var denne forskjellen over 20 poeng i guttenes favør i matematikk i TIMSS 1995, regnet i forhold til et internasjonalt gjennomsnitt på 500 poeng på hele testen (Hastedt mfl. 2005).

I naturfag i TIMSS 1995 gjorde de norske elevene det bedre på de åpne oppgavene i forhold til de andre landene enn hva de gjorde på de lukkede oppgavene. Tendensen var for øvrig klar ved at i de fleste land, og spesielt i de øst-europeiske landene, gjorde elevene det bedre på lukkede enn på åpne oppgaver. Resultatene for øvrig var at guttene gjorde det bedre enn jentene og størst var forskjellen på de lukkede oppgavene. Andre undersøkelser bekrefter tendensen til at det er større kjønnsforskjell i guttenes favør i flervalgsoppgaver enn i åpne oppgaver (Ben-Shakhar mfl. 1991, DeMars 2000, Kjærnsli mfl. 2004). I naturfagoppgavene til TIMSS 1999, gjorde guttene det bedre enn jentene på de lukkede oppgavene i 31 av 38 land. I de åpne oppgavene gjorde guttene det bedre i 17 av 31 land, mens jentene gjorde det bedre enn guttene i 3 land. Tendensen er at personer som skårer godt på lukkede oppgaver i en test, også ofte skårer godt på lukkede oppgaver i andre tester. Bridgeman mfl. (1996)

konkluderer med at dette ofte gjelder for menn og hvite studenter (!) (Martinez 1999). For norske elever viser resultatene i matematikk fra PISA 2006 at en del flervalgsoppgaver ser ut til fremdeles å favorisere guttene. Kjærnsli mfl. (2007) konkluderte med at de ikke fant annet mønster i dette, enn at resultatet var oppgavebetinget.

Martinez (1999) viser til en studie som Gallanger mfl. (1994), gjennomførte i matematikk på high-schoolelever. Studien viste at jenter vanligvis var mindre utholdende og hadde mindre selvtillit enn gutter i matematikk. I tillegg brukte jentene mer tradisjonelle metoder på utradisjonelle matematiske problemer, og dette viste seg å være en ulempe.

Faktorer som ikke er av faglig art, vil kunne påvirke en prøvesituasjon. Snow (1993) argumenterte av den grunn for at man også skulle tenke på elevers følelser og motivasjon når man valgte oppgaveformat. Studier kan ikke gi noen forklaring, men tilbakevendende resultater viser at gutter relativt sett skårer bedre på lukkede enn åpne oppgaver, og at jenter relativt sett skårer bedre på åpne oppgaver enn flervalgsoppgaver. Oppgaveformatet kan derfor ha betydning selv om utfordringene i oppgavene i utgangspunktet ikke er knyttet til oppgaveformatet, men oppgavenes faglige innhold.

2.8. Gjetting

Tilfeldig gjetting hevdes å være en svakhet og en trussel for resultatet ved lukkede oppgaver (Downing 2003, Downing 2006). Hvis testen er utviklet av profesjonelle utviklere og oppgavene er prøvd ut og analysert før de blir benyttet i en test, er oppgaver som i utprøvingen ikke ble besvart av de riktige elevene, luket bort. Kvalifisert (kunnskapsbasert) gjetting forekommer derimot. At elevene får anledning til å løse oppgaver ved å bruke deler av kunnskapen de har, er reelt i forhold til situasjoner i dagliglivet (Downing 2003). Sjansen for at tilfeldig gjetting skal føre til høy skår på en test er svært liten, og den avtar hvis antall oppgaver i testen øker. Tar vi som eksempel en test med 10 flervalgsoppgaver med fire svaralternativer i hver oppgave, er sjansen for kun å gjette seg til rett svar i alle oppgavene, $(1/4)^{10}$, dvs. ca. 0,000000954. De som forsvarer lukkede oppgaver kan derfor blant annet argumentere med at "gjettefaktoren" kan gjøres ubetydelig ved at man øker antall oppgaver i testen. En større fare ligger i at oppgavene ikke har høy nok kvalitet ved at de for eksempel er dårlig formulert eller inneholder feil, og at testen er for kort.

En studie gjennomført av Ben-Shakar mfl. (1991) støtter at det er kjønnsforskjeller når det gjelder å løse oppgaver. Da ble kjønnsforskjeller på flervalgsoppgaver i matematikk sett i forhold til tendens til gjetting og til ikke å svare på oppgaver. Resultatene viste at guttene gjettet mer enn jentene, og at jentene hadde flere ubesvarte oppgaver enn guttene. Det kan se ut som at guttene både har en tendens til å ta flere sjanser enn jentene og at de bruker andre strategier for å finne svaret i lukkede oppgaver. Beller mfl. (2000) viser også til lignende resultat når det gjelder undersøkelse gjennomført av Gafni mfl. (1994). Det ble imidlertid konkludert med at dette bare kunne være en liten del av forklaringen på at guttene fikk bedre resultat enn jentene på testene (Beller mfl. 2000).

2.9. Distraktorenes betydning og diagnostisk verdi

Det kan være forskjellige årsaker til at elever velger en bestemt distraktor. Gjetting og eliminasjon kan for eksempel være måter som er brukt for å velge ut et svar. Kritikere mener derfor at siden det finnes studier som viser at elever ikke velger distraktorer ut fra det oppgavene er ment å måle, gir bruk av lukkede oppgaver i tester ikke tilgang til elevers kunnskaper i faget. Schoultz (2000) har gjennomført studier som viser at elevenes svar kan skyldes språkproblemer. Det kan for eksempel være problemer med et ord eller et uttrykk i spørsmålet eller formulering av en distraktor. Resultatet av intervjuer analysert av Clerk mfl. (2000), viser at det ikke var elevenes misoppfatninger, men misforståelse i forbindelse med språklige formuleringer som var årsak til valg av distraktorer. Alle oppgavene som ble studert i denne undersøkelsen hadde imidlertid alvorlige mangler og feil som lett kunne blitt rettet opp. Dette viser hvor viktig det er at både åpne og lukkede oppgaver prøves ut før de blir brukt i en test, og at oppgavene blir kvalitetssikret gjennom analyser. I tillegg må oppgavene være utviklet i tråd med anbefalte retningslinjer.

Kazemi (2002) gjennomførte en kvalitativ studie av åpne oppgaver og flervalgsoppgaver (MC) i matematikk. I studien ble 90 elever på 4. trinn intervjuet for å høre hvordan de tenker og resonnerer i prøvesituasjoner. Dette gjaldt både åpne og lukkede oppgaver. Resultatene viser at elevene i lukkede oppgaver ofte er mer opptatt av svaralternativene enn å tenke igjennom oppgaveteksten og prøve å løse oppgavene. De generaliserer ut fra metoder de kan om problemløsning og velger ofte feil svar. Hvis konteksten er kjent for dem kan de trekke konklusjoner ut fra egne erfaringer og slik også få feil svar. Noen regnet feil og ble forledet til å velge galt svar fordi de fant sin egen misoppfatning som et svaralternativ. I denne

undersøkelsen var det flere tilfeller av høyere løsningsprosent blant de åpne oppgavene enn blant de tilsvarende lukkede oppgavene, fordi elevene i de åpne oppgavene hadde mulighet til å forklare svaret sitt. Om eleven like før hadde løst en åpen eller en lukket oppgave, påvirket både utførelsen og resonnementet i neste oppgave.

Martinez (1999) omtaler en undersøkelse Birenbaum mfl. (1987) gjennomførte for å finne ut hvilket av formatene åpent og lukket som hadde størst diagnostisk verdi. Selv om svaralternativene som ble brukt i de lukkede oppgavene var elevsvar fra åpen utprøving, var den diagnostiske verdien signifikant større i de åpne enn i de lukkede oppgavene. Temaet i undersøkelsen var aritmetikk og en konklusjon var at mange elever i stedet for å løse en oppgave trolig bare velger et alternativ. Da er det ikke sikkert at de velger det alternativet som er i overensstemmelse med egen forståelse, og læreren får feil signal. De konkluderte med at det samme trolig også gjelder for andre fag enn matematikk.

Som referert i de forrige avsnittene, er mange negative til distraktorenes diagnostiske betydning, og skeptiske til i hvilken grad man kan stole på at elevene velger svar i tråd med egen overbevisning. Ikke alle deler denne skepsisen, men en forutsetning må være på plass for at distraktorene skal ha diagnostisk verdi, og det er at de er gode distraktorer. Med dette menes at de er mulige elevsvar og ved utprøving er valgt av minst 5 prosent av elevene (Downing 2006). Ved utvikling av oppgaver til de nasjonale prøvene i regning i Norge, prøves oppgavene først ut i åpent format, og elevsvarene brukes til distraktorer når oppgavene gjøres om til lukkede oppgaver (Ravlo mfl. 2010). Det erfaringsbaserte inntrykket til faggruppa som utvikler disse oppgavene, er at det er god overensstemmelse mellom p-verdien til distraktorene og p-verdiene til de ulike elevsvarene da oppgaven var åpen. På denne måten kan vi få indikasjoner på bestemte misoppfatninger hos grupper, klasser og på nasjonalt nivå gjennom gode distraktorer.

I en studie av oppgaveformater på resultater fra internasjonale undersøkelser, fremhever Olsen mfl. (2001) nettopp den diagnostiske verdien som kan ligge i gode distraktorer. På den annen side omtaler de i samme artikkel hvordan distraktorene både kan veilede og villede elevene når de skal velge et svar ut fra alternativer. Dette igjen understreker viktigheten av at det i utviklingsfasen gjøres grundige analyser på oppgaver som skal brukes i tester hvor resultatet vil kunne få stor betydning, både når det gjelder for enkeltpersoner og for grupper.

Olsen mfl. (2001) mener elevene bruker ulike strategier når de skal løse oppgaver med svaralternativer. Ut fra distraktorenes funksjon, snakker de om seks hovedstrategier, som jeg nå vil omtale nærmere. Oppgaver fra min studie blir brukt som eksempler.

Distraktorer som sjekkliste

I dette tilfellet kan alle alternativene ved første øyekast virke som mulige svar på oppgaven. For å finne løsningen må elevene gjøre beregninger og sammenligne eget svar med alternativene som er oppgitt. I slike oppgaver er stimulus og stamme lik i åpen og lukket utgave, og beregningene som skal gjøres er de samme uansett oppgaveformat. En fordel i lukket format er at hvis eleven gjør en feil, vil han få respons på dette ved at han ikke finner svaret sitt blant alternativene og dermed får en mulighet til å rette feilen. Svarprosenten kan derfor teoretisk bli høyere i lukket enn åpen versjon av oppgaven. Et annet alternativ til lukket oppgave i dette formatet er at alle mulige alternativ for svar er oppgitt. Oppgave fB41a i figur 2.8 er et eksempel på dette. Alle mulige dreieretninger innbyrdes for tannhjulene er listet opp, og elevenes oppgave blir å etterprøve disse for å finne riktig løsning.

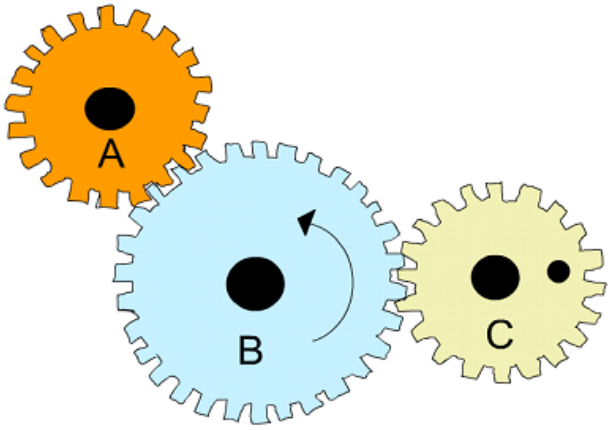
Figur 2.8 kan også være et eksempel på språklige formuleringer som kan villede elevene til å svare annerledes enn de egentlig mener (Schoultz 2000, Clerk mfl. 2000). Det er en mulighet for at lesesvake og minoritetsspråklige elever kan ha problemer med tolkningen når bruken av begrepene *samme* og *motsatt* er det eneste som skiller alternativene fra hverandre. Dette kan føre til at denne typen oppgave i noen tilfeller får høyere p-verdi i åpent enn lukket format. Oppgaven i figur 2.8 ble gjennomført på 151 elever i åpent og 155 elever i lukket format. I åpent format avga 79 prosent av elevene riktig svar på oppgaven. Som lukket oppgave var løsningsprosenten 68 prosent.

Figur 2.8 Distraktorer som sjekkliste. Oppgave i Test 2 prøve 2

Oppgave fB41a)

Tannhjul B dreier i pilens retning.

I hvilken retning dreier tannhjulene A og C?



1 A og C dreier begge i samme retning som B

2 A dreier i samme retning som B og C dreier i motsatt retning

3 A og C dreier begge i motsatt retning som B

4 C dreier i samme retning som B og A dreier i motsatt retning

Eliminering av distraktor

I noen tilfeller kan elever raskt se at en eller flere distraktorer umulig kan være riktig svar på oppgaven. Det kan være at distraktoren er ulogisk som svar i den oppgitte sammenhengen, direkte feil ut fra en naturfaglig synsvinkel, eller riktig naturfaglig, men ikke som svar på det aktuelle spørsmålet. Det kan for eksempel være spurt om forklaring på et fenomen som i figur 2.9. Fordi elevene i slike oppgaver i realiteten får færre svaralternativer å ta stilling til, har denne typen lukket oppgave ofte høyere p-verdi enn tilsvarende åpen oppgave. Oppgave fA22 ble gjennomført på 151 elever. Alternativ C ble valgt av en elev. Det er grunn til å tro at mange elever eliminerte bort alternativet fordi det virket lite troverdig.

Figur 2.9 Eliminering av distraktor. Oppgave i Test 2 prøve 1

Oppgave fA22

Hvorfor lyser månen?

A Den er glødende

B Solen skinner på den

C Den blir synlig i mørket

D Den består av hvite gasser

Svaralternativer som presiserer spørsmålet

Det kan være at oppgaven inneholder noen vanskelige begreper eller at den fortoner seg uklar for elevene. Ved å se på svaralternativene kan eleven få oppklarende hint om i hvilken retning det er ment at man skal tenke, og eleven unngår både å misforstå oppgaven og å bruke unødig tid på tolkning. Svaralternativene til oppgave fA32 i figur 2.10 avgrenser oppgaven i forhold til åpent format, og senker derfor vanskelighetsgraden fra åpen til lukket oppgave. 78 prosent av elevene fant riktig løsning da oppgaven var lukket, mens løsningsprosenten var 59 prosent som åpen oppgave.

Figur 2.10 Svaralternativer som presiserer spørsmålet. Oppgave i Test 2 prøve 1

Oppgave fA32

Tabellen viser temperaturen på et sted til forskjellige tider på dagen i tre dager.

	6.00	9.00	12.00	15.00	18.00
Mandag	15°C	17°C	20°C	21°C	19°C
Tirsdag	15°C	15°C	15°C	5°C	4°C
Onsdag	8°C	10°C	14°C	14°C	13°C

Når begynte det å blåse en kaldere vind?

A Mandag ettermiddag

B Tirsdag morgen

C Tirsdag ettermiddag

D Onsdag morgen

Distraktor mangler

Mennesker stoler på sine etablerte hverdagsforestillinger (Sjøberg 1998), og hvis elever finner sin overbeviste misoppfatning blant svaralternativene, så viser det seg at de ofte velger denne. Motsatt vil det derfor både kunne skape forvirring og bidra til færre feil på en oppgave hvis en typisk misoppfatning mangler blant svaralternativene. Denne typen oppgave har vanligvis høyere p-verdi som lukket enn åpen oppgave.

Kognitiv felle

Hvis man legger inn svaralternativer som kan være sannsynlige, men som elever ikke ville tenkt på å svare om oppgaven var åpen, kan vi kalle en slik distraktor for en kognitiv felle. Man villeder elevene og skaper usikkerhet. I slike tilfeller får den åpne oppgaven ofte høyere løsningsprosent enn oppgaven i lukket format.

Svaralternativene definerer spørsmålet

I denne typen oppgave skal elevene sammenligne svaralternativene og plukke ut det alternativet som er *best* svar på oppgaven. Siden svaralternativene er sentrale og avgjørende for oppgaven, er det ikke mulig å gjennomføre denne typen oppgave som åpen (se figur 2.11).

Figur 2.11 Oppgave hvor svaralternativene definerer spørsmålet. Oppgave fra TIMSS 1995

Hva beskriver BEST Jordas overflate gjennom milliarder av år?

A En plan overflate blir gradvis løftet opp til høyere og høyere fjell helt til Jorda er dekket av fjell.

B Høye fjell brytes gradvis ned helt til det meste av Jorda er på havnivå.

C Høye fjell brytes gradvis ned, mens nye fjell hele tiden blir dannet, om og om igjen.

D Høye fjell og sletter har vært slik de er nå uten noen vesentlig forandring gjennom milliarder av år.

Olsen mfl. (2001) presiserer at det ikke er klare skillelinjer mellom disse seks funksjonene som distraktorene kan ha. I en oppgave kan en distraktor samtidig ha flere ulike funksjoner, og det kan også være ulike typer distraktorer i en oppgave.

2.10. Betydningen av antall distraktorer i et svar

En distraktor som velges av færre enn 5 prosent av elevene, kalles en ikke-fungerende distraktor (Downing 2006). Alle distraktorer skal alltid være et sannsynlig alternativ til svar for elevene. Hvis en distraktor ikke er god, kan det være lett å eliminere den bort, og slik redusere antall reelle valgmuligheter.

Studier viser at synet er delt når det gjelder hvor mange distraktorer det bør være i svaralternativene i en lukket oppgave. Noen liker ideen med så mange distraktorer som mulig, mens et fåtall foretrekker en standard på fire. I perioden 1989 til 2002 ble det rapportert sju studier på antall distraktorer (Haladyna mfl. 2002). Resultatene varierer i forhold til om en oppgave blir lettere eller vanskeligere når antall svarmuligheter øker. Fem av studiene viser at en økning i antall distraktorer førte til økt vanskelighetsgrad i oppgavene, mens i to andre undersøkelser førte det til at oppgaven ble lettere (Haladyna mfl. 2002). I et tilfelle fant man

at diskrimineringen avtok da antall distraktorer økte, og i et annet tilfelle ingen forskjell i diskriminering. En undersøkelse rapporterte om lavere reliabilitet da antall svaralternativer ble redusert, mens en annen konkluderte med ingen endring i reliabiliteten. En evaluering Haladyna og Downing gjennomførte i 1993 på fire standardiserte tester med lukkede oppgaver, viste at det eksisterte ingen klar sammenheng mellom antall virkningsfulle distraktorer og vanskelighetsgraden i en oppgave, men jo mer virkningsfull distraktorene var, jo bedre diskriminerte en oppgave. I denne undersøkelsen ble det konkludert med at antall effektive distraktorer per oppgave ikke var mer enn en.

I en undersøkelse av Tarrant mfl. (2009) viste resultatene at bare i 13,8 prosent av oppgavene fungerte tre distraktorer. Antall distraktorer i gjennomsnitt som fungerte ($p > 0,05$) i en oppgave, var bare 1,5, og hvis man reduserte antall svaralternativer fra fire eller fem til tre i en test med 100 oppgaver, førte dette til en økning i gjennomsnitt på kun 1,22 poeng. Det var heller ingen signifikant forskjell verken i diskriminering eller vanskelighetsgrad når man sammenlignet oppgaver med tre og fire svaralternativer. Derimot økte både diskriminering og reliabilitet da man reduserte antall alternativer fra fem til tre (Rodriguez 2005). Vel å merke gjelder disse undersøkelsene oppgaver som var konstruert ut fra anbefalte retningslinjer (Haladyna mfl 2002) og prøvd ut slik at man visste at de skilte godt mellom elever som var på forskjellig faglig nivå. Ut fra dette ble det konkludert med at to distraktorer i tillegg til nøkkelen var det optimale, og dette er helt i tråd med anbefalinger fra Haladyna mfl (2002) og Downing (2006). Ved å bruke få distraktorer kan elevene også prøves i flere oppgaver innenfor et gitt tidsrom. Dette kan også være et tiltak for å øke reliabiliteten i en undersøkelse.

2.11. Oppsummering av hovedpunkter i kapittel 2

Undersøkelser viser at lukkede oppgaver er den mest effektive måten å teste kunnskaper på både når det gjelder klasseromstesting og stor-skala-testing. Oppgavene må være utviklet i tråd med anbefalte retningslinjer for at reliabilitet og validitet skal være ivaretatt. Oppgaver med lik stamme i åpent og lukket format har vist høy korrelasjon, og korrelerer signifikant høyere enn oppgaver med samme innhold, men ulik formulering i åpent og lukket format.

For å teste skrivekompetanse må man bruke åpne oppgaver. Åpne oppgaver har ofte lavere reliabilitet enn lukkede oppgaver, fordi vurderingen kan bli preget av subjektiv oppfatning hos den som vurderer. Muntlig kompetanse i språk og praktiske ferdigheter kan ikke testes med

lukkede oppgaver. Åpne oppgaver egner seg bedre til å kartlegge elevers misoppfatninger enn lukkede oppgaver (Haladyna 1999). Først da vet man med sikkerhet at svaret bare skyldes elevenes egne tanker i øyeblikket. Gode distraktorer i lukkede oppgaver kan også gi verdifull diagnostisk informasjon om elevers misoppfatninger. Det er en tendens til at gutter skårer bedre enn jenter på lukkede oppgaver.

3. Kapittel Metode og gjennomføring

3.1. Forskningsdesign

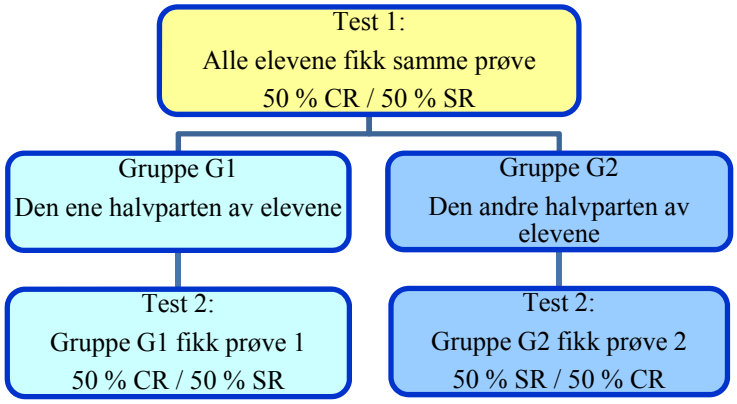
3.1.1 Test 1 og Test 2

Jeg gjennomførte en kvantitativ undersøkelse ved hjelp av tre skriftlige prøver i naturfag, Test 1 (vedlegg 1) og Test 2 (vedlegg 2 og 3). Test 1 bestod av *en* prøve mens Test 2 bestod av to prøver, prøve 1 og prøve 2. Alle prøvene inneholdt 10 lukkede og 10 åpne oppgaver, og var sammenlignbare ved at oppgavene var laget ut fra kompetansemålene i naturfag etter 7. trinn i LK06. I Test 1 var oppgavene i de to formatene tilfeldig plassert i oppgavesettet, og alle elevene fikk samme prøve.

I Test 2 fikk den ene halvparten av elevene prøve 1 og den andre halvparten prøve 2. Åpne og lukkede oppgaver var som i Test 1, i utgangspunktet plassert tilfeldig rundt i oppgavesettene, men prøve 1 og prøve 2 var innbyrdes avhengige. Oppgavene i prøve 1 hadde samme innhold som oppgavene i prøve 2, og var plassert på samme sted i prøvene. Formatet til oppgavene var imidlertid motsatt i de to prøvene. Med det menes at de 10 oppgavene som var i åpent format i prøve 1, hadde formatet lukket i prøve 2, og vice-versa.

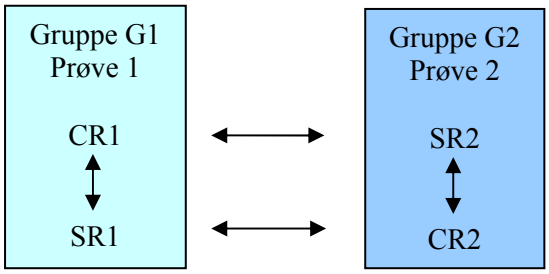
Målet med Test 1 var ut fra resultatene å få etablert to grupper av elever, gruppe G1 og gruppe G2. Gruppene skulle være tilnærmet like og sammenlignbare når det gjelder antall jenter og gutter og fordeling av elevene på tre faglige nivåer. I Test 2 fikk elevene i gruppe G1 prøve 1 og elevene i gruppe G2 prøve 2. Elevene i de to gruppene ble på denne måten prøvd i det samme fagstoffet, men med oppgaver i motsatt format (se figur 3.1).

Figur 3.1 *Viser sammenhengen mellom tester og grupper. Alle elevene fikk Test 1. G1 og G2 er sammenlignbare grupper, slik at det til hver elev i gruppe G1 svarer minst en elev i gruppe G2. I Test 2 fikk gruppe G1 prøve 1 og gruppe G2 prøve 2. Prøve 1 og prøve 2 består av de "samme" oppgavene, men i motsatt format*



Målet med Test 2 var å sammenligne elevenes resultater med fokus på oppgaveformat. Dette innebærer å sammenligne resultatene elevene fikk på oppgavene i åpent format med resultatene som ble oppnådd da de samme oppgavene var i formatet lukket, - gruppe G1 med gruppe G2. Det var også et mål å se på i hvilken grad de ulike elevene løste oppgaver i lukket eller åpent format. Dette ble undersøkt ved å sammenligne resultatene internt i gruppe G1 og internt i gruppe G2 med hensyn på oppgaveformatet (se figur 3.2).

Figur 3.2 *Sammenligne resultater på åpne (CR1) i gruppe G1 med resultater på lukkede (SR2) i gruppe G2 og motsatt. Tilsvarende sammenligning internt i gruppe G1 og i gruppe G2*



Test 1 og Test 2 var uavhengige i den forstand at resultatene på Test 2 ikke ble sammenlignet med resultatene på Test 1. Det var imidlertid viktig at Test 1 og Test 2 hadde samme oppbygning og innhold fordi resultatene på Test 1 var avgjørende for hvordan elevene ble fordelt på gruppene G1 og G2 (se figur 3.1).

Integrert i Test 1 var et spørreskjema med fire utsagn om holdninger til og interesse for naturfag og tre utsagn om egenvurdering i forbindelse med testen. Utsagnene var i formatet flervalg (MC) og skulle besvares ved hjelp av en firedelt *likert-skala* (Robson 2002) med alternativene enig, litt enig, litt uenig og uenig. Til slutt fikk elevene et spørsmål om hvilke strategier de bruker når de løser flervalgsoppgaver (MC). Dette spørsmålet var i åpent format.

3.1.2 Design for Test 1

Hovedproblemstillingen i denne masteroppgaven er å sammenligne resultater i naturfag på lukkede oppgaver med resultater på åpne oppgaver innenfor samme faglige tema og kompetansemål. Det ideelle hadde vært å kunne teste de samme elevene først i åpne oppgaver og deretter prøve dem på de samme oppgavene i lukket format. Dette er dessverre ikke mulig. Den første testen ville i seg selv kunnet påvirke elevene og slik virke inn på prøveresultatet. Første trinn i min undersøkelse ble derfor å etablere to sammenlignbare grupper av elever, gruppe G1 og gruppe G2. Gruppene skulle være på 8. trinn og bestå av elever som representerte tre ulike faglige nivåer, - faglig sterke, middels flinke og faglig under middels.

Undersøkelsen har fast design og bruker *survey* metode. I dette ligger en standardisert og strukturert utspørring av et utvalg personer. Robson (2002) sier at med en survey metode samler man vanligvis inn små datamengder fra et relativt stort antall individer. Individene skal være representative for populasjonen de skal representere. Prøven jeg gav elevene var relativ kort med bare 20 oppgaver, og utvalget var relativt stort, på ca. 300 elever. Utvalget bestod av alle elevene på 8. trinn ved to store ungdomsskoler i en av Norges største byer. Utvalget var ikke representativt (se kapittel 3.2) med tanke på alle de 60 000 elevene som tilhører 8. trinn et skoleår, og skolene var ikke trukket tilfeldig. Allikevel har utvalget en styrke ved at det inneholdt alle de aktuelle elevene på disse to store skolene.

Siden undersøkelsen med Test 1 ble gjennomført praktisk talt på samme tidspunkt for alle elevene, kan man si at designet er en ”Cross-sectional study” (Robson 2002). Designet er ikke-eksperimentelt siden jeg ikke skulle måle endring i kunnskapene til elevene fra Test 1 til Test 2.

De begrepene som studeres i en forskningssammenheng, kalles variabler (Kleven 2002). I ”Cross-sectional studies” har man kun en forsøksgruppe, og fokuset er på variabler innenfor denne gruppen. Robson (2002) sier at ”Cross-sectional studies” trolig er det mest brukte

designet i samfunnsundersøkelser, og at det ofte innebærer bruk av survey metode for innsamling av data. I eksperimentell metode snakker man om uavhengige og avhengige variabler. I "Cross-sectionale studier" kalles de uavhengige variablene forklarende *(explanatory)* variabler og de andre resultatvariabler *(outcome)* (Robson 2002). Fordi variabler er sentrale begreper i forskningssammenheng, velger jeg allerede på dette tidspunkt å forklare begrepene nærmere.

Forklarende (uavhengige) variabler er hva den som gjennomfører en undersøkelse, bestemmer seg for å undersøke betydningen av, for eksempel kjønn, holdninger og motivasjon (Robson 2002, Kleven 2002). Resultatvariabler (avhengige) er det som undersøkes for å se etter forskjeller, for eksempel om jenter og gutter har ulikt resultat i gjennomsnitt etter at et undervisningsopplegg er gjennomført. Da er skåreverdien på en test resultatvariabel, fordi skåreverdien sees i forhold til det valgte området for fokus, - kjønn -, som er den forklarende variabelen i dette eksempelet. I tillegg er det alltid noe usikkerhet forbundet med et resultat. Det er derfor vanlig å si at et resultat (y) kan skrives som en funksjon, $y = f(x) + e$, der x for eksempel er den forklarende variabelen kjønn, f(x) er gjennomsnittlig resultat som skyldes faglige kunnskaper for guttene, og e symboliserer usikkerheten i form av alt annet som kan påvirke resultatet, men som vi ikke har kontroll på hva er. Y er i dette eksemplet det som fremkommer som gjennomsnittlig poengsum for guttene og er resultatvariabelen. Nærmere beskrivelse av hvordan usikkerheten i resultater kan kontrolleres og bestemmes er behandlet i kapittel 2.3. (sentrale statistiske begreper).

Etter at elevene hadde gjennomført Test 1 og en kort holdningsundersøkelse, ble de ut fra resultatene, vurdert med hensyn på faglig nivå. Deretter ble elevene satt sammen i par. To elever som utgjorde et par, hadde omtrent samme poengsum og var av samme kjønn. I tillegg hadde de tilnærmet lik poengsum på de 7 utsagnene i Test 1 om holdninger til og interesse for naturfag, og på spørsmålet om strategier de bruker når de løser flervalgsoppgaver. Poengsummen på holdningsoppgavene ble imidlertid tillagt liten vekt, og fikk ingen avgjørende betydning. I utgangspunktet var holdningsoppgavene heller ikke ment å være av avgjørende betydning, men ble tatt med for at jeg skulle få litt føling med elevenes tanker om faget naturfag.

Fra hvert par ble en elev plassert i gruppe G1 og en i gruppe G2. For noen elever var det vanskelig å finne en tilsvarende elev for å danne et par. Disse elevene ble tilfeldig plassert i

gruppe G1 eller G2. Det var også tilfeller hvor tre elever ble vurdert som sammenlignbare. Disse fikk utgjøre en treergruppe og den tredje eleven ble plassert i gruppe G1 eller G2 ut fra antallet for øvrig i hver av gruppene. Det skulle være omtrent samme antall elever i gruppe G1 som i gruppe G2 (se tabell 3.9 – 3.12).

3.1.3 Design for Test 2

Etter fire måneder fikk elevene en ny test, Test 2, som utgjør hoveddelen av masteroppgaven, og som skulle gi svar på hvilken betydning oppgaveformatet har for elevenes resultater på en prøve. Test 2 bestod av to prøver, prøve 1 og prøve 2. Gruppe G1 fikk prøve 1 og gruppe G2 fikk prøve 2. Resultatene på de åpne oppgavene til gruppe G1, ble sammenlignet med resultatene på flervalgsoppgavene til gruppe G2 og motsatt. Fase 2 har derfor sammenlignende design (comparative design) (Robson 2002). Man kan si at det i testfase 2, er gjennomført en *survey undersøkelse* med *comparative design.* Fordi gruppene G1 og G2 var satt sammen på grunnlag av resultatene fra Test 1, fungerte de som kontroll for hverandre (se figur 3.1 og 3.2).

Robson (2002) sier at en fordel med sammenlignende design (comparative design) i forhold til eksperimentell design er muligheten for å inkludere mange forklarende variabler. Man må imidlertid holde seg til en ”tommelfingerregel” som sier at det må være minst 15 deltakere per variabel.

3.2. Data

3.2.1 Utvalget

Samtlige elever på 8. trinn ved to store ungdomskoler i en av Norges største byer, utgjør i utgangspunktet utvalget for undersøkelsen. I fortsettelsen vil skolen med det høyeste elevtallet bli kalt *skole 1* og den andre *skole 2*. Skole 1 hadde da undersøkelsen ble gjennomført, 182 elever på 8. trinn fordelt på 6 klasser. Tilsvarende for skole 2 var 151 elever fordelt på 5 klasser. Det totale antall elever ved skolene var henholdsvis ca. 540 og 450. Begge er ”rene” ungdomsskoler med elever på 8., 9. og 10. trinn. Ungdomsskolene har tilnærmet like tradisjoner når det gjelder undervisningspraksis, og ved begge skolene var man da Test 1 ble gjennomført høsten 2007, i ferd med å høste erfaringer med undervisning i skolelokaler uten avdelte klasserom. Å velge to sammenlignbare ungdomsskoler, var et bevisst valg. Slik ønsket jeg å få størst mulig gruppe av mest mulig sammenlignbare elever, og dermed stort nok antall til å kunne dele elevene inn i tre nivåer. Ved å velge de skolene jeg gjorde, unngikk jeg i

særlig stor grad å måtte ta hensyn til minoritetsspråklige elever hvor kompetanse i norsk kan utgjøre en ekstra feilkilde.

Hvis alle elevene hadde vært på skolen på prøvedagene, ville jeg hatt et mulig utvalg på 333 personer. I første testrunde var imidlertid 39 elever fraværende. Utgangspunktet for undersøkelsen ble derfor et utvalg på 294 elever som gjennomførte Test 1. Det var 30 flere elever fra skole 1 enn fra skole 2, og totalt 12 flere gutter enn jenter som deltok (se tabell 3.1).

Tabell 3.1

Fordeling av elever på skoler, og antall jenter og gutter som gjennomførte Test 1 og Test 2

	Elever på 8. trinn	**Elever Test 1**	**Jenter Test 1**	**Gutter Test 1**	**Elever Test 2**	**Jenter Test 2**	**Gutter Test 2**
Skole 1	182	**162**	83	79	**166**	83	83
Skole 2	151	**132**	58	74	**140**	65	75
Totalt	333	**294**	**141**	**153**	**306**	**148**	**158**

I andre testrunde, - gjennomføring av Test 2, var 306 elever til stede (se tabell 3.1). Noen elever som deltok på Test 1, var fraværende, mens andre var kommet til. Skole 1 deltok med 26 flere elever enn skole 2, og denne gangen var det totalt 10 flere gutter enn jenter til stede. Hvis vi ser på fordeling på skoler og antall jenter og gutter, var utvalget ved gjennomføring av Test 2 ikke svært forskjellig fra utvalget ved Test 1.

Test 2 bestod av to prøver. Gruppe G1 fikk prøve 1 og gruppe G2 fikk prøve 2. Prøvene var jevnt fordelt på skolene slik at det var omtrent like mange elever av hvert kjønn på prøve 1 som på prøve 2 på hver skole. Det var 72 jenter som deltok på prøve 1 og 76 jenter på prøve 2. Tilsvarende tall for guttene var 79 både på prøve 1 og prøve 2 (se tabell 3.2).

Tabell 3.2

Fordeling av elever på skoler, og antall jenter og gutter som gjennomførte henholdsvis prøve 1 og prøve 2 i Test 2

	Elever Test 2 prøve 1	**Jenter Test 2 prøve 1**	**Gutter Test 2 prøve 1**	**Elever Test 2 prøve 2**	**Jenter Test 2 prøve 2**	**Gutter Test 2 prøve 2**
Skole 1	**82**	41	41	**84**	42	42
Skole 2	**69**	31	38	**71**	34	37
Totalt	**151**	**72**	**79**	**155**	**76**	**79**

Tabell 3.3 viser antall elever som var til stede både på Test 1 og Test 2. Av de 294 elevene i Test 1 gjennomførte 275 elever Test 2. Resultatene til disse 275 elevene er grunnlaget for analysene som skal gi svar på problemstillingene i denne masteroppgaven. Kvalitetssikringen av testene er imidlertid gjort på grunnlag av resultatene til alle elevene som deltok på hver av testene. For Test 1 gjelder dette 294 elever (se tabell 3.1). For Test 2 er antallet 151 elever på prøve 1 og 155 elever på prøve 2. Dette tilsvarer fordelingen av de 306 elevene på gruppe G1 og gruppe G2 i Test 2 (se tabell 3.2).

Tabell 3.3

Elever som deltok på begge testene, - fordeling på skole og antall jenter og gutter

	Elever både Test 1 og 2	**Jenter både Test 1 og 2**	**Gutter både Test 1 og 2**
Skole 1	150	75	75
Skole 2	125	55	70
Totalt	**275**	130	145

Tabell 3.4

Elever som deltok på begge testene, - fordeling på skole og antall jenter og gutter.
Viser hvor mange som fikk prøve 1 og prøve 2 i Test 2

	Elever Test 2 prøve 1	**Jenter Test 2 prøve 1**	**Gutter Test 2 prøve 1**	**Elever Test 2 prøve 2**	**Jenter Test 2 prøve 2**	**Gutter Test 2 prøve 2**
Skole 1	**76**	39	37	**74**	36	38
Skole 2	**64**	27	37	**61**	28	33
Totalt	**140**	**66**	**74**	**135**	**64**	**71**

Tabell 3.4 viser at av de elevene som deltok både på Test 1 og Test 2, var det 5 flere som deltok på prøve 1 enn på prøve 2. For skole 1 var fordelingen av jenter og gutter jevn for begge prøvene, mens det for skole 2 var noen færre jenter enn gutter både på prøve 1 og prøve 2. De jentene fra skole 2 som deltok, var imidlertid jevnt fordelt på begge prøvene.

3.3. Operasjonalisering av begrepene

3.3.1 Oppgavene til Test 1 og Test 2

Målet med Test 1 var å få etablert to sammenlignbare grupper G1 og G2 som skulle undersøkes videre i Test 2. Målet med Test 2 var å finne svar på problemstillingen om hvilken betydning oppgaveformatet kan ha for resultatene elever får på en naturfagprøve. Etablering av gruppene G1 og G2 og arbeidet for å finne svar på hovedproblemstillingene i oppgaven, skulle skje på grunnlag av kunnskapstester i naturfag hvor kompetansemål i Læreplanen Kunnskapsløftet (LK06) etter 7. trinn var det faglige nivået.

For å ivareta begreps- og innholdsvaliditeten best mulig, det vil si at det skulle være størst mulig samsvar mellom det jeg ønsket å måle, og det som i virkeligheten ble målt, tok jeg utgangspunkt i oppgaver som var utviklet til Osloprøven i naturfag for 8. trinn 2007. Oppgavene som jeg fikk tillatelse til å bruke, var utviklet ved Institutt for lærerutdanning og skoleutvikling (ILS) ved Universitet i Oslo og tilpasset kompetansemålene i naturfag etter 7. trinn. Sju av oppgavene som jeg benyttet, hadde sin opprinnelse fra TIMSS 1995. Hver oppgave var validert ved at den var relatert til et bestemt kompetansemål i LK06. Jeg kunne velge blant 80 oppgaver. Disse var kvalitetssikret gjennom tidligere å ha blitt testet ut på et representativt utvalg elever. At et utvalg er representativt betyr at alle grupper som det skal trekkes konklusjoner for, er representert i utvalget. Resultatene fra den tidligere utprøvingen, viste at oppgavene var velfungerende på den måten at de tilfredsstilte psykometriske krav til blant annet reliabilitet, diskriminering og validitet, - et kvalitetsstempel som medfører at vi med rimelig sikkerhet vil kunne stole på resultatene. På denne måten unngikk jeg fasen med utvikling og utprøving av oppgaver, og kunne komme raskere i gang med egen studie.

I den opprinnelige utprøvingen til Osloprøven, var hver test på 40 oppgaver. Oppgavene ble prøvd ut i to hefter A og B, og jeg har av praktiske årsaker valgt å beholde oppgavenes opprinnelige numre. Alle prøvene i Test 1 og Test 2 i min undersøkelse bestod av 20 oppgaver. Jeg måtte derfor gjøre grundige analyser for å finne ut om de oppgavene jeg hadde valgt ut, også oppfylte de psykometriske kravene til å ha god nok kvalitet som samlet prøve i en test med 20 oppgaver. Analysene for Test 1 og Test 2 er i kapittel 4.

Naturfag i grunnskolen er et sammensatt fag som består av fagene biologi, kjemi, fysikk, geofag og teknologi. Jeg kunne valgt å lage en test som bare tok for seg et av fagene, eller

kanskje bare et emne, for eksempel økologi. Spørsmålene ville da ha blitt mer fokusering på detaljer. Jeg valgte ikke den løsningen. Elevene på begge ungdomsskolene kommer fra flere ulike barneskoler som erfaringsmessig vektlegger de ulike emnene i naturfag ulikt. Dette kan blant annet henge sammen med barneskolelærerens fagbakgrunn. Hva som er behandlet på barneskolene og på hvilken måte stoffet er gjennomgått, er en feilkilde man må være oppmerksom på. Detaljorienterte spørsmål med smalt fokus kunne derfor muligens ført til flere ubesvarte åpne oppgavene, og kanskje mer gjetting på flervalgsoppgavene. Siden målet med Test 1 ikke var å få en oversikt over elevenes kunnskapsnivå, men å få etablert to sammenlignbare grupper, valgte jeg å bruke mange emner og ikke fokusere smalt. På denne måten regnet jeg med å ha større sjanse til å treffe alle elevene ”hjemme” på noen områder. Det viktigste var tross alt at de fleste elevene svarte på så mange oppgaver som mulig.

Siden gruppene G1 og G2 skulle bli etablert på grunnlag av resultatene fra Test 1, og dette var det eneste faglige kriteriet for utvelgelse, var det nødvendig at neste prøve, Test 2, var så lik den første testen som mulig. Bare slik kunne jeg forsvare inndelingen i par, og bruke dette til å sammenligne elevenes resultater på åpne og lukkede oppgaver i prøve 1 og prøve 2 i Test 2.

Test 1 og Test 2 ble derfor utarbeidet på samme tidspunkt, og fordelingen av oppgaver på tema, vanskelighetsgrad og format, var tilstrebet å være så lik som mulig. Prøvene korresponderte på mange områder, og fungerte tilfredsstillende i forhold til målet om at elevene skulle bli prøvd i det samme fagstoffet i Test 1 og Test 2, selv om man ikke kan si at Test 1 og Test 2 var parallelle tester per definisjon. I parallelle tester er oppgavene parvis like (Kleven 2002), i den betydning at til enhver oppgave i den ene testen svarer en oppgave i den andre. Dette ble ikke ivaretatt i utarbeidelsen av Test 1 og Test 2, da det ikke var mulig ut fra de oppgavene jeg hadde til rådighet, men oppgavene var typelike. Innholdet i oppgavene var relatert til kompetansemål innenfor kropp og helse, fenomener og stoffer, forskerspiren, verdensrommet og mangfold i naturen. Det var både lukkede og åpne oppgaver innenfor alle emnene. For å gi alle elevene mulighet til å få vist hva de kunne, var det viktig å ha oppgaver av ulik vanskelighetsgrad. Løsningsprosenten (p-verdien) på oppgavene fra utprøvingen til Osloprøven, ble lagt til grunn da jeg valgte oppgaver. Friheten til valg av oppgaver var størst i Test 1, siden alle de 294 elevene skulle ha samme prøve. Ved utvelgelsen måtte jeg kontrollere at det ble 10 åpne og 10 lukkede oppgaver, og at det til hver oppgave som ble valgt til Test 1, eksisterte en noenlunde tilsvarende når det gjelder p-verdi og tema som kunne brukes i Test 2. Test 1 kom til å bestå av fem oppgaver som flere enn ca. 64 prosent av

elevene hadde svart riktig på i den opprinnelige utprøvingen til Osloprøven, ni oppgaver som var riktig besvart av fra 47- 62 prosent, og seks som mindre enn ca. 41 prosent av elevene hadde besvart riktig (se tabell 3.5).

Tabell 3.5

Innhold, løsningsprosent (p-verdi) og diskriminering (d-verdi) ved utprøving til Osloprøven 2007, område (biologi, kjemi, fysikk), format (SR=lukket, CR=åpen), vanskelighetsgrad (L=lav, M=middels, V=vanskelig), gjennomsnittlig p-verdi for oppgavene som ble Test 1 og kompetansemål i LK06 relatert til de seks hovedområdene i fagplanen for naturfag

Oppg.nr. Test 1	**Oppg.nr. Osloprøven.**	**Innhold Test 1**	**p – verdi**	**d – verdi**	**Område**			**Format**		**Kategori**			**Mål LK06 7.trinn**
					Bio	**Kje**	**Fys**	**SR**	**CR**	**L**	**M**	**V**	
1	A23	Planeter	0,81	0,53			1		1	1			4.1
2	A25	Månen / sola	0,75	0,63			1		1	1			5.1
3	A2	Begrep I kjemi	0,62	0,45		1		1			1		1.2
4	B29a	Faseovergang	0,55	0,55		1			1		1		5.6
5	B29b	Faseovergang	0,64	0,58		1			1	1			5.6
6	B8	Befruktning	0,52	0,38	1			1			1		2.5
7	A18	Skjelettet	0,50	0,55	1				1		1		3.2
8	A35a	Kraftverk	0,47	0,49	1				1		1		5.4
9	A35b	Miljøpåvirkning	0,17	0,42	1				1			1	5.4
10	A37	Kjem. reaksjon	0,55	0,45		1		1			1		5.8
11	A16	Fordøyelse	0,41	0,31	1			1				1	3.1
12	A9	Dyregrupper	0,62	0,62	1			1			1		2.3
13	A14a	Næringsnett	0,38	0,20	1			1				1	2.5
14	A14b	Næringsnett	0,31	0,52	1				1			1	2.5
15	A29	Magnetisme	0,47	0,31			1	1			1		5.3
16	B31	Kvele flamme	0,81	0,46		1			1	1			5.6
17	A31	Vannmolekyler	0,41	0,49		1		1				1	5.6
18	B34	Kjem. reaksjon	0,73	0,60		1		1		1			5.6
19	B39	Energi	0,61	0,62			1		1		1		6.3
20	B3	Løsninger	0,29	0,39		1		1				1	1.3
		Gj.sn. p-verdi	**0,53**		**8**	**8**	**4**	**10**	**10**	**5**	**9**	**6**	

Av praktiske årsaker var oppgavene sortert som biologi, kjemi og fysikk (se tabell 3.5 og 3.6), selv om fagene er integrert i hverandre. Det var i utgangspunktet ønskelig med like mange fra hvert av de tre områdene og like mange vanskelige, middels vanskelige og lette oppgaver. Dette var det ikke mulig å få til med det oppgaveutvalget som var tilgjengelig. Test 1 kom til å bestå av 8 oppgaver med hovedvekt på biologi, 8 kjemirelaterte oppgaver og 4 fra fysikk.

Tabell 3.6 viser oppgavene som ble valgt som utgangspunkt for Test 2. Oppgavene 1 og 8 var prøvd ut både som åpne og lukkede oppgaver til Oslopröven. Røde tall viser verdiene da oppgavene var i åpent format. I summen for kolonnene og gjennomsnittlig p-verdi brukes verdiene for oppgave 1 og 8 i lukket format.

Tabell 3.6

Oppgavene som var grunnlaget for prøve 1 og prøve 2 i Test 2. Innhold, løsningsprosent (p-verdi) og diskriminering (d-verdi) fra utprøving til Oslopröven 2007, format (SR=lukket, CR=åpen) og vanskelighetsgrad (L=lav, M=middels, V=vanskelig) er for de opprinnelige oppgavene til Test 2 før de ble fordelt på prøve 1 og prøve 2. Oppgave 1 og 8 med røde tall for formatet åpen. Mål relatert til fagplanen i naturfag etter 7. trinn i LK06

Oppg.nr. Test 2	Oppg.nr. Oslopröven	Innhold Test 2	p-verdi	d-verdi	Område			Format		Kategori			Mål LK06
					Bio	Kje	Fys	SR	CR	L	M	V	7. trinn
1	A40 (B27)	Torden og lyn	0,74 (0,37)	0,46 (0,58)			1	1	(1)	1		(1)	5.2
2	A22	Månen	0,73	0,42		1		1		1			4.2
3	B14	Organer	0,51	0,36	1			1			1		3.1
4	B41a	Tannhjul	0,61	0,45			1		1		1		6.1
5	B41b	Tannhjul	0,73	0,42			1		1	1			6.1
6	A3	Fordamping	0,76	0,37		1		1		1			1.1
7	A33	Gass	0,53	0,36		1		1			1		5.6
8	A15 (B12)	Vekselvarm	0,53 (0,32)	0,44 (0,61)	1			1	(1)		1	(1)	2.5
9	A34	Atomer	0,61	0,37		1		1			1		5.7
10	A32	Tabell grader	0,69	0,46			1	1		1			5.5
11	B25	Korrolering	0,17	0,39	1			1				1	5.1
12	B33	Kjemisk reaksjon	0,27	0,48		1		1				1	5.8
13	A41	temp og atomer	0,53	0,33		1		1			1		5.6
14	B10	Lunger og gjeller	0,58	0,49	1			1			1		2.3
15	B30	Gasser	0,33	0,36		1		1				1	5.6
16	A19	Kropp	0,32	0,45	1			1				1	3.3
17	A28	Brensel	0,67	0,47	1			1		1			5.4
18	A7	Blomsterdeler	0,33	0,22	1			1				1	2.2
19	A36	Kjemisk reaksjon	0,36	0,36		1		1				1	5.8
20	B17	Kropp	0,46	0,21	1			1			1		3.1
		Gj.sn. p-verdi	**0,52**		**8**	**8**	**4**	**18**	**2**	**6**	**8**	**6**	

Til Test 2 ble det valgt 8 oppgaver innenfor biologi, 8 innenfor kjemi og 4 fysikkoppgaver. Av disse var det 6 oppgaver som flere enn 67 prosent av elevene hadde løst riktig i utprøvingen til Oslopröven, 8 oppgaver som 46 - 66 prosent av elevene hadde løst, og 6 som var løst av færre enn 41 prosent av elevene.

Test 2 bestod av to prøver, prøve 1 og prøve 2. Fordelt på disse to prøvene var samme oppgave i åpent og lukket format, og begge prøvene bestod av 10 åpne og 10 lukkede oppgaver. En åpen oppgave i prøve 1 hadde formatet lukket i prøve 2 og vice-versa. Her er det viktig å påpeke at man aldri kan si at to oppgaver er like med mindre de er identiske. En åpen oppgave kan aldri bli lik en oppgave i lukket format, men vi kan tilstrebe at oppgavene er slik formulert at de etterprøver de samme kunnskapene. To tilsvarende oppgaver, men i ulikt format, stod med samme oppgavenummer i prøve 1 og prøve 2. Slik ble oppgaverekkefølgen lik, men oppgaveformatet motsatt når det gjelder plassering i prøve 1 og prøve 2. For å synliggjøre formatet til en oppgave, valgte jeg på dette tidspunktet å tilføye *å* foran navnet til de åpne oppgavene og *f* foran navnet til de lukkede oppgavene. Det betyr at åB41a og fB41a er to formater av samme oppgave, og at den ene oppgaven hører til i prøve 1 og den andre i prøve 2.

Tabell 3.7 viser hvilke oppgaver i Test 2, prøve 1 og prøve 2, som ble endret i forhold til opprinnelige form og format i Osloprøven. I prøve 1 gjennomgikk 7 av 20 oppgaver små endringer. I prøve 2 var det nødvendig å endre litt på 11 av 20 oppgaver.

Prøve 1 og prøve 2 ble utviklet ved at jeg først valgte ut 16 lukkede oppgaver. Ti av disse kunne direkte gjøres om til åpne oppgaver uten at det var nødvendig å endre teksten for at oppgaven skulle ha samme innhold. Disse oppgavene har X i kolonnen for p-verdi i tabell 3.7. I de seks andre oppgavene måtte teksten skrives om for at oppgavene tilnærmelsesvis skulle kunne teste tilsvarende i åpent format. Disse oppgavene har E i kolonnen for p-verdi i tabell 3.7. I tillegg ble det valgt to oppgaver, oppgave 1 og 8, som var testet ut i Osloprøven både som åpen oppgave og i formatet lukket (se figur 3.3 og 3.4).

Da manglet det to oppgaver som begge var åpne og måtte gjøres om til lukkede oppgaver. Dette var oppgave B41a og B41b, som stod som oppgave 4 og 5 i Test 2 (se tabell 3.7 og kode EF). Disse to oppgavene hadde imidlertid et innhold som gjorde at svaralternativene nærmest ga seg selv (se figur 3.5).

Tabell 3.7

Fordeling av lukkede oppgaver (SR) og åpne oppgaver (CR) i prøve 1 og prøve 2 for Test 2. Løsningsprosenter (p-verdier) og diskriminering (d-verdier) fra Osloprøven som var utgangspunkt for utvelging av oppgaver. Oppgaver som ble endret fra sin opprinnelige form, er merket med X, E og EF i kolonnen for p-verdi. Eks: B14 var oppgave 14 i pilot B for Osloprøven, og opprinnelig i lukket format. A3 var tilsvarende oppgave 3 i pilot A for Osloprøven. I prøve 1 er begge oppgavene endret fra lukket til åpen oppgave, og teksten endret noe for at oppgavene i åpent format skulle måle det samme som i lukket format. Oppgavene har symbolet E i kolonnen for p-verdi. Oppgavene merket X er uendret fra Osloprøven, og EF betyr at jeg laget svaralternativene i det lukkede formatet

Oppgavens plassering i Test 2	**Oppgavenr. i Osloprøven**	**Innhold**	**Test 2 Prøve 1 Verdier fra utprøving til Osloprøven**			**Test 2 Prøve 2 Verdier fra utprøving til Osloprøven**		
			Format	**p - verdi**	**d - verdi**	**Format**	**p - verdi**	**d - verdi**
1	A40 / B27	Torden og lyn	SR	0,74	0,46	CR	0,37	0,58
2	A22	Månen	SR	0,73	0,42	CR	X	
3	B14	Organer	CR	E		SR	0,51	0,36
4	B41a	Tannhjul	CR	0,61	0,45	SR	EF	
5	B41b	Tannhjul	CR	0,73	0,42	SR	EF	
6	A3	Fordamping	CR	E		SR	0,76	0,37
7	A33	Gass	SR	0,53	0,36	CR	X	
8	B12 / A15	Vekselvarm	CR	0,32	0,61	SR	0,53	0,44
9	A34	Atomer	CR	X		SR	0,61	0,37
10	A32	Tabell grader	SR	0,69	0,46	CR	X	
11	B25	Korrolering	CR	X		SR	0,17	0,39
12	B33	Kjemisk reaksjon	CR	X		SR	0,27	0,48
13	A41	temp og atomer	SR	0,53	0,33	CR	X	
14	B10	Lunger og gjeller	SR	0,58	0,49	CR	X	
15	B30	Gasser	CR	E		SR	0,33	0,36
16	A19	Kropp	SR	0,32	0,45	CR	X	
17	A28	Brensel	CR	E		SR	0,67	0,47
18	A7	Blomsterdeler	SR	0,33	0,22	CR	E	
19	A36	Kjemisk reaksjon	SR	0,36	0,36	CR	E	
20	B17	Kropp	SR	0,46	0,21	CR	X	

Figur 3.3 *Oppgave 1 i Test 2 i åpent og lukket format, uendret slik de ble testet ut i utprøvingen til Osloprøven. Originalnavn med f og å som angir formatet*

Oppgave fA40

Hvorfor ser vi lyn før vi hører torden?

A Fordi øynene oppfatter lys lettere enn ørene oppfatter lyd
B Fordi synssansen er sterkere enn hørselssansen
C Fordi lynet skjer nærmere oss enn torden
D Fordi lys beveger seg fortere enn lyd

Oppgave åB27

Hvorfor ser vi lyn før vi hører torden?

Oppgave 8 var ikke identisk i åpent og lukket format. Begge oppgaveformatene krever imidlertid kjennskap til begrepet vekselvarme dyr. Fordi begge oppgavene var prøvd ut og kvalitetssikret, valgte jeg å la dem være med som to sammenlignbare oppgaver i motsatt format. Alternativt kunne jeg brukt oppgaveteksten til fA15. I så fall ville jeg i stedet ha brukt en oppgave som ikke var kvalitetssikret, og det prøvde jeg å unngå så fremt det var mulig (se figur 3.4).

Figur 3.4 *Oppgave 8 i Test 2 i åpent og lukket format, uendret slik de ble testet ut i utprøvingen til Osloprøven. Originalnavn med f og å som angir formatet*

Oppgave fA15

På hvilken måte er varmblodige dyr forskjellige fra vekselvarme dyr?

A Hos varmblodige dyr øker stoffskiftet i varmt vær
B Varmblodige dyr er mer fiendtlige i fangenskap
C Varmblodige dyr har alltid høyere temperatur i blodet
D Varmblodige dyr har konstant kroppstemperatur uavhengig av temperaturen i omgivelsene

Oppgave åB12

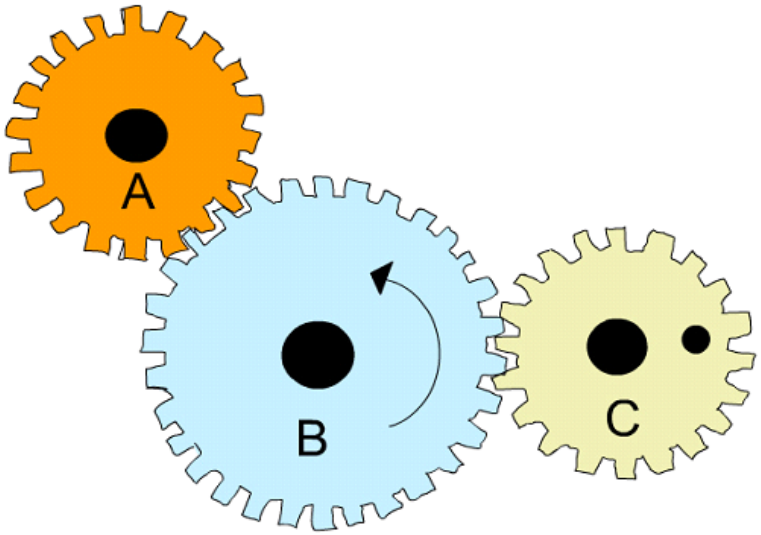
Hva betyr det at et dyr er vekselvarmt?

Figur 3.5 *Oppgave 4 og 5 i prøve 1 i Test 2 i åpent format slik de ble testet ut til Osloprøven 2007*

Oppgave åB41a)

Hvilken retning dreier tannhjulene?

Tegn inn piler på tannhjul A og tannhjul C som viser hvilken retning de dreier.



Oppgave åB41b)

Hvilket av tannhjulene B og C bruker kortest tid på en hel runde?

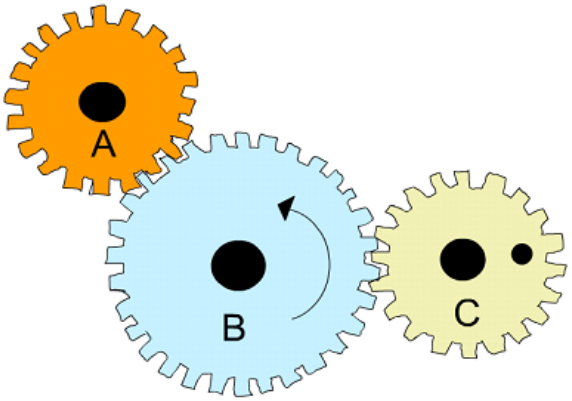
I figur 3.6 er oppgavene i lukket format. En liten omformulering var nødvendig ved endring av format, men jeg mener svaralternativene var uproblematiske å lage ut fra innholdet i oppgavene.

Figur 3.6 *Oppgave 4 og 5 i prøve 2 til Test 2 i lukket format. Endret fra åpent som var formatet ved utprøvingen til Osloprøven*

Oppgave fB41a)

Tannhjul B dreier i pilens retning.

I hvilken retning dreier tannhjulene A og C?



1 A og C dreier begge i samme retning som B

2 A dreier i samme retning som B og C dreier i motsatt retning

3 A og C dreier begge i motsatt retning som B

4 C dreier i samme retning som B og A dreier i motsatt retning

Oppgave fB41b)

Hvilket av utsagnene er riktige?

1 B bruker kortere tid enn C på en hel runde

2 C bruker kortere tid enn B på en hel runde

3 B og C bruker like lang tid på en hel runde

4 Det er umulig å avgjøre hvem av B og C som bruker kortest tid på en hel runde

Av praktiske årsaker ble ingen av oppgavene som det ble gjort endringer i, prøvd ut i sitt nye format.

3.3.2 Holdninger og interesse for naturfag

Test 1 inneholdt fire utsagn som handlet om holdninger, interesse for faget og egenvurdering av faglig kompetanse. Utsagnene som skulle vurderes på en firedelt likert-skala som vist i figur 3.7, kom før de faglige spørsmålene og var ment å virke motiverende og skjerpende.

Figur 3.7 *Utsagn som elevene måtte ta stilling til før de gjennomførte Test 1*

Først skal du svare på disse spørsmålene:

Du skal bare sette ett kryss for hvert utsagn:

		Enig	Litt Enig	Litt uenig	Uenig
1.	Naturfag er et viktig fag	☐	☐	☐	☐
2.	Jeg liker naturfag	☐	☐	☐	☐
3.	Jeg kan mye naturfag	☐	☐	☐	☐
4.	Jeg liker å se på "Newton" eller "Schrødingers katt"	☐	☐	☐	☐

Da er det bare å starte med prøven! ☺

Jeg valgte å bruke kun to positive og to negative svaralternativer for å tvinge elevene til å ta et valg.

Figur 3.8 *Oppgaver som ble besvart umiddelbart etter at Test 1 var gjennomført*

Dette skal du svare på etter at du har levert prøven:

Du skal bare sette ett kryss for hvert utsagn:

		Enig	Litt enig	Litt uenig	Uenig
5.	Temaene i oppgavene var kjent for meg	☐	☐	☐	☐
6.	Jeg fikk vist hva jeg kan på denne prøven	☐	☐	☐	☐
7.	Jeg har i hvert fall gjort mitt beste! ☺	☐	☐	☐	☐

8. Anta at du skulle svare på en oppgave ved å krysse av for rett svar, og at du ikke visste svaret. Hva ville du gjøre da?

Beskriv så nøyaktig som mulig hva du ville gjøre! ☺

Etter at Test 1 var levert inn, besvarte elevene et spørreskjema hvor de måtte ta stilling til i hvilken grad temaene i oppgavene var kjent for dem, om de gjennom prøven hadde fått vist kunnskapene sine, og til slutt måtte de vurdere sin egen innsats ut fra en firedelt likert-skala. Et siste åpent spørsmål bad elevene fortelle hvilken strategi de bruker når de skal besvare flervalgsoppgaver og ikke vet hva som er rett svar på oppgaven (se figur 3.8).

Det var bevisst at ingen av utsagnene inneholdt negasjoner. Ved å ha bare en type utsagn, antok jeg at det var mindre sjanse for at elevene misforstod hva de skulle svare på. Elevenes svar på holdningsoppgavene ble vektlagt i liten grad, men brukt til justering, da jeg skulle fordele elevene på gruppene G1 og G2. Elevene skulle ikke forberede seg spesielt til denne prøven, men de ble oppfordret til å gjøre sitt beste i prøvesituasjonen.

3.4. Metode for innsamling av data

For å få svar på de valgte problemstillingene, gjennomførte jeg to undersøkelser.
Alle elevene på 8. trinn på samme skole gjennomførte prøvene på samme tidspunkt. Prøvene var skriftlige, uten hjelpemidler og elevene hadde 45 minutter per prøve til å løse oppgavene. Jeg var til stede, informerte lærerne og sørget for at gjennomføringen ble så lik som mulig på begge skolene. Elevene på de to skolene visste ikke om hverandre.

Den første undersøkelsen var Test 1 (se vedlegg 1) som ble gjennomført 3. og 6. november 2007 på henholdsvis skole 1 og skole 2. 294 elever deltok på Test 1 (se tabell 3.1). Test 2 med prøve 1 og prøve 2 foregikk på samme dag på begge skolene, 25. mars 2008. Skole 2 hadde testen på morgenen, og skole 1 hadde samme test etter spisepausen. Totalt 306 elever deltok på prøve 1 og prøve 2 i Test 2 (se tabell 3.1 og 3.2).

Test 1 var lik for alle elevene. Ut fra antall poeng på prøven og resultatet på de åtte oppgavene som var et tillegg til prøven (se figur 3.7 og 3.8), ble elevene delt i to sammenlignbare grupper, gruppe G1 og gruppe G2. Med sammenlignbar menes at til hver elev i gruppe G1 kan vi ut fra poeng, finne en tilsvarende elev i gruppe G2.

I Test 2 fikk elevene i gruppe G1 og gruppe G2 oppgaver med så likt innhold som mulig, men i ulikt format (se kapittel 3.3.1.). Gruppe G1 og G2 er omtalt ved flere anledninger, men har

foreløpig bare fått en overfladisk behandling. Jeg vil i neste delkapittel redegjøre for kriteriene som ble brukt ved inndeling av gruppene.

3.5. Utvelging av elever til gruppe G1 og gruppe G2

Ut fra poengsummene som elevene fikk på Test 1, ble de delt inn i tre nivågrupper for jenter og tilsvarende for gutter. Nivå 1 bestod av de 25 prosent av elevene som fikk lavest poengsum på Test 1, dvs. prosentil 25. Dette var gutter og jenter som hadde oppnådd 10,0 poeng eller lavere på testen. Av tabell 3.8 ser vi at dette var 40 gutter og 40 jenter. På prosentil 75 går nedre grense for det høyeste nivået (nivå 3), som bestod av de 25 prosent dyktigste elevene. Dette var gutter med mer enn 16,5 poeng og jenter med mer enn 15,25 poeng. De 11 guttene som hadde skåret til akkurat 16,5 poeng, ble også valgt til å være med i denne gruppen. Det høyeste nivået kom derfor til å bestå av 41 gutter og 35 jenter (se tabell 3.8).

Tabell 3.8

Gutter og jenter hver for seg plassert på tre nivåer. Poeng og antall elever for de ca. 25 % med lavest skår, antall elever i den midterste gruppen og poenggrense og antall elever for de ca. 25 % av elevene som fikk høyest poengsum på Test 1

Gutter	Antall elever: 153			Antall elever på hvert nivå
Nivå 1	Prosentiler	**25**	**10,0 p**	**40**
Nivå 2				**72**
Nivå 3		**75**	**16,5 p**	**41**
Jenter	Antall elever: 141			
Nivå 1	Prosentiler	**25**	**10,0 p**	**40**
Nivå 2				**66**
Nivå 3		**75**	**15,25 p**	**35**

Den midterste gruppen bestod av 72 gutter som hadde fra 10,5 til 16,0 poeng, og 66 jenter med poeng fra 10,5 til 15,0. Alle elevene ble plassert på et nivå, og det medførte at det ble glidende overganger mellom nivåene. Med en standard målefeil på 1,7 poeng (se kapittel 2.3.2 og tabell 4.1) vil det alltid være elever som kunne vært på to av nivåene.

Etter å ha kjørt analyser i SPSS, ble alle de 294 elevsvarene på prøven og resultatene fra holdningsspørsmålene analysert i Excel. Fordi det gikk omtrent fire måneder mellom Test 1 og Test 2, ble resultatene fra hver skole behandlet for seg. Det ble prøvd å finne elever på samme nivå innenfor en klasse for å danne par innenfor klassen, men det var ikke mulig for

alle elevene. Den nest beste løsningen ble å danne par innenfor samme skole, og på denne måten prøve å redusere feilkildene mest mulig ut fra antagelse (vedlegg 4).

Hver skole ble analysert for seg. Først ble elevene sortert etter kjønn. Så ble de sortert etter oppnådde poeng på prøven. Ved hjelp av fargekoder var det relativt enkelt å få oversikt over hvilke elever innenfor hvert kjønn som hadde tilnærmet lik poengsum (vedlegg CD, rådata for 294 elever). Det ble ikke sett på hvilke oppgaver elevene hadde besvart riktig. Analysene skulle basere seg på prøven som helhet, og ikke på enkeltoppgaver eller deler av prøven. Etter å ha sammenlignet poengsummene på holdningsspørsmålene til for eksempel de jentene som hadde samme poengsum, fikk to av samme kjønn som tilsynelatende hadde mest sammenfallende holdning, lav eller høy verdi, utgjøre et par. Resultatet ble par av elever som hadde skåret høyt på prøven, og med verdier som viste positive holdninger, elever som hadde skåret lavt, og med positive holdninger, elever med høy poengsum, men med mindre positive holdninger og lavtskårende likeså. Poenggivingen for holdninger er valgt slik at jo større interesse en elev sa hun/han hadde for faget, jo større tro på egen kompetanse, egen innsats og selvtillit i utøving av faget, jo lavere poengsum oppnådde eleven på holdningsoppgavene. Tabell 3.9 – 3.12 viser resultatet for fordeling av elevene på tre nivå og for jenter og gutter.

Tabell 3.9

Gutter på skole 1, fordeling på nivå, i par og treergrupper etter resultater på Test 1

Gutter skole 1 Prosentil	Poeng	Antall gutter	Antall par	Antall trippel
Nivå 1 (25 %)	$X \leq 10$	27	9	3
Nivå 2 (50 %)	$10 < X \leq 16$	34	14	2
Nivå 3 (25 %)	$16 < X \leq 20$	18	9	
Sum		**79**		

Tabell 3.10

Jenter på skole 1, fordeling på nivå, i par, treergrupper og alene etter resultater på Test 1

Jenter skole 1 Prosentil	Poeng	Antall jenter	Antall par	Antall trippel	Antall single
Nivå 1 (25 %)	$X \leq 10$	29	8	3	4
Nivå 2 (50 %)	$10 < X \leq 15{,}25$	39	17	1	2
Nivå 3 (25 %)	$15{,}25 < X \leq 20$	15	4	2	1
Sum		**83**			

Tabell 3.11

Gutter på skole 2, fordeling på nivå, i par, treergrupper og single etter resultater på Test 1

Prosentil	Poeng	Antall Gutter	Antall par	Antall trippel	Antall single
Nivå 1 (25 %)	X ≤ 10	13	3	2	1
Nivå 2 (50 %)	10 < X ≤ 16	38	16	2	
Nivå 3 (25 %)	16 < X ≤ 20	23	10	1	
Sum		**74**			

Tabell 3.12

Jenter på skole 2, fordeling på nivå, i par og treergrupper etter resultater på Test 1

Prosentil	Poeng	Antall Jenter	Antall par	Antall trippel	Antall single
Nivå 1 (25 %)	X ≤ 10	11	3	1	2
Nivå 2 (50 %)	10< X ≤ 15,25	27	12	1	
Nivå 3 (25 %)	15,25 < X ≤ 20	20	10		
Sum		**58**			

Resultatet ble gruppene G1og G2, hvor 38 elever var på høyeste nivå i hver gruppe, 69 på midterste nivå og 40 elever utgjorde nivå 1 med lavest poengsum i hver gruppe. Gruppene ble kontrollgrupper for hverandre i Test 2 som ble gjennomført fire måneder seinere, våren 2008, og som utgjør hovedundersøkelsen i denne masteroppgaven.

Forsøksgruppene G1 og G2 ble ikke tilfeldig valgt, men dannet på grunnlag av poeng på en prøve, kombinert med resultatet fra en holdningsundersøkelse. Utvalget var nøye planlagt for at det skulle være representativt for en stor ungdomsskole. Jeg har kontrollert at elever fra begge skolene var tilnærmet likt representert i begge gruppene, at begge kjønn var representert på alle nivåer, og at det var sammenlignbare elever i begge gruppene og på alle nivåene. For å prøve å begrense effekten av ulik påvirkning i perioden mellom Test 1 og Test 2, ble sammenligningen avgrenset til å gjelde elever på samme skole og helst i samme klasse. Ved at jeg på denne måten hadde en viss kontroll på variabler som tidligere forskning har vist kan påvirke effekten av oppgaveformat, hadde jeg et godt utgangspunkt for å kunne trekke slutninger om at eventuelle effekter kunne knyttes til formatet til oppgavene og ikke til skjevheter i utvalgets sammensetning.

3.6. Bakgrunn for valg av utvalg og begrensninger

De internasjonale komparative undersøkelsene TIMSS (Grønmo mfl. 2004, 2009), PISA (Lie mfl. 2001, Kjærnsli mfl. 2004, 2007), Nasjonale prøver 2007 (Bonesrønning mfl. 2008) og Nasjonale prøver 2008 (Bonesrønning mfl. 2010), konkluderer med at hjemmeforhold har betydning for barns skoleprestasjoner. Det foreligger blant annet resultater som viser at barn som har god tilgang på bøker i hjemmet og har foreldre med høy utdannelse og høyt lønnsnivå, i gjennomsnitt presterer bedre enn elever som ikke har samme ressurstilgang. Jeg har tidligere i denne rapporten (se kapittel 3.2.1.) begrunnet mitt utvalg av skoler ut fra elevers kompetanse i norsk språk. En annen årsak var at jeg i utgangspunktet hadde planer om å sammenligne elevene på tvers av skolene, ved å velge en elev fra hver skole da jeg skulle sette sammen parene. Derfor valgte jeg to skoler som ligger i områder med sammenlignbare bolig- og hjemmeforhold, og som gjennom mange år har hatt ganske sammenfallende resultater ved offentlig eksamen.

Fordi det kom til å gå fire måneder mellom Test 1 og Test 2, besluttet jeg at det var mer fornuftig å danne par av elever innenfor hver skole og helst innenfor hver klasse. Det var større sjanse for at elever innenfor en klasse eller en skole var blitt utsatt for lik påvirkning i løpet av fire måneder, enn om de gikk på hver sin skole. Slik prøvde jeg å minske feilkilden fra Test 1 til Test 2 siden gruppene G1 og G2 skulle dannes på grunnlag av resultatene fra Test 1.

Et representativt utvalg inneholder representanter for alle grupper som en undersøkelse skal gjelde for. Er utvalget i tillegg trukket tilfeldig blant alle aktuelle grupper og enkeltpersoner, er det ikke noe i veien for å trekke generelle slutninger. I den grad utvalget ikke er trukket tilfeldig eller er representativt, må man være oppmerksom på begrensninger som systematiske feilkilder gir (se kapittel 2.3.2). I denne masteroppgaven har jeg brukt et formålsutvalg (Kleven 2002, Robson 2002) som er et ikke-sannsynlighetsutvalg. Av praktiske og bekvemmelighetsmessige årsaker, valgte jeg å forholde meg til elevene på to ungdomsskoler. Elevene var fra to relativt veletablerte boligstrøk, og tilnærmet uten representasjon av minoritetsspråklige elever. Styrken med utvalget var imidlertid at det i utgangspunktet bestod av alle elever på 8. trinn på begge skolene og til sammen utgjorde 333 elever. Selv om noen falt fra var det etter Test 2 resultater på 275 elever.

Ved å la alle elevene på trinnet delta, og med et såpass stort utvalg, har jeg grunn til å anta at mange elevtyper var representert i utvalget og at den faglige spredningen blant elevene var relativt typisk. Jeg må kunne anta at jeg har gode representanter for tre faglige nivåer, - faglig sterke elever, elever som er middels flinke og de faglig svakeste elevene, både for jenter og gutter. Det er hva jeg behøver for å kunne sammenligne resultatene til to elevgrupper og finne svar på de formulerte problemstillingene.

Selv om jeg ikke vil kunne trekke generelle slutninger ut fra et slikt utvalg, vil jeg tro at noen skjønnsmessige betraktninger vil kunne gjøres på bakgrunn av resultatene.

3.7. Reliabilitet og validitet for undersøkelsen

Siden forskerne ved ILS hadde prøvd ut oppgavene på en gruppe elever som var representative for de elevene som skulle gjennomføre Osloprøven, og kvalitetssikret oppgavene ved at de oppfylte psykometriske krav til tester, er det grunn til å anta at testene mine kunne oppnå høy reliabilitet. At prøvene i Test 1 og Test 2 bestod av 50 prosent lukkede oppgaver, bidro også til å øke reliabiliteten. Det gjorde at jeg i halvparten av oppgavene kunne se bort fra truslene om tilfeldige målefeil på grunn av at vurderingen var subjektiv og avhengig av den som vurderte.

Ved selv å være til stede ved gjennomføring av prøvene, jevnlig i bevegelse i landskapene hvor prøvene ble gjennomført og tilgjengelig hvis noen lurte på noe, prøvde jeg å legge forholdene til rette for like rammer i prøvesituasjonen. At prøvene ble gjennomført på tilnærmet samme tidspunkt og at lærerne som gjennomførte testen, hadde fått samme instruksjon muntlig av meg, bidro også til høyere reliabilitet.

Jeg vurderte selv alle besvarelsene. I de lukkede oppgavene skulle elevene bare krysse av for å avgi et svar, og det ble derfor enten rett eller galt. Det var ikke mulig for meg som vurderte, å påvirke resultatet. I de lukkede oppgavene må jeg derfor kunne anta at sensorreliabiliteten var 100 prosent.

De åpne oppgavene ble vurdert ut fra nøyaktig ført fasit (vedlegg CD). Allikevel kan det i noen tilfeller være rom for tolkning av den som retter, og det kan være mulig å misforstå det

som eleven har skrevet. Ved at jeg selv vurderte alle besvarelsene ut fra en skriftlig fasit, prøvde jeg å ivareta reliabiliteten i forhold til vurdering av de åpne oppgavene.

Begrepsvaliditeten ble ivaretatt ved at oppgavene var laget ut fra kompetansemålene i naturfag slik de er formulert i LK06 etter 7. trinn.

3.8. Metode for analyse

Både holdningsspørsmålene og resultatene på de faglige spørsmålene ble hver for seg analysert i programmet SPSS (vedlegg CD). I Test 1 ble riktig besvart oppgave honorert med ett poeng, og oppnåelig poengsum var 20 poeng. I de åpne oppgavene ble det gitt et halvt poeng for delvis rett svar ut fra bestemte kriterier (se fasit, vedlegg CD). I de lukkede oppgavene ble det gitt 1 poeng for riktig svar og ellers 0 poeng.

I Test 2, prøve 1 og prøve 2, ble det ikke gitt delpoeng. Riktig svar ble honorert med 1 poeng og alt annet gav 0 poeng. Det ble ikke gitt delpoeng fordi alle oppgavene både åpne og lukkede skulle telle likt, og hvert oppgaveformat utgjøre nøyaktig 50 prosent av testen (vedlegg CD).

Holdningsoppgavene ble i begge testene vurdert med 1 poeng for enig, 2 poeng for litt enig, 3 poeng for litt uenig og 4 poeng for uenig.

Analysen består for det første i å vise at Test 1 er valid og reliabel nok til at den kan brukes til å etablere to sammenlignbare grupper av elever. Her blir Cronbachs alfa, oppgavenes løsningsprosenter, oppgavenes korrelasjon med hverandre og med summen av oppgaver, og elevenes dyktighet sett i forhold til hvilke oppgaver som er løst, viktige analysedeler. Deretter må prøve 1 og prøve 2 i Test 2 kvalitetssikres på tilsvarende måte, før man studerer elevenes svar på oppgavene. Løsningsprosentene på oppgavene i åpent og lukket format, om det er kjønnsforskjeller og om elevenes kompetanse spiller inn når det gjelder om de har løst de åpne eller de lukkede oppgavene, blir analysert. Analysen vil også ta for seg om oppgavens innhold spiller rolle for p-verdier i hvert av formatene.

4. Kapittel Resultater

I dette kapittelet behandles resultatene fra Test 1 og Test 2 i kronologisk rekkefølge. Test 1 hadde ett mål, - å få etablert to sammenlignbare grupper. Test 2 hadde mål knyttet opp mot problemstillingene i masteroppgaven. I første del av kapittelet blir validitet og reliabilitet, ut fra psykometriske krav, dokumentert for Test 1 og for prøve 1 og prøve 2 i Test 2.

4.1. Resultater for Test 1

4.1.1 Resultater for å undersøke kvaliteten på Test 1

I kapittel 3.3.1 har jeg redegjort for kriterier som ble brukt ved utvelging av oppgavene til Test 1. De tjue oppgavene dekket til sammen minst et kompetansemål innen hvert av de seks hovedområdene i fagplanen for naturfag i LK06 (se tabell 3.5). Dette var viktig for å tilrettelegge for å nå alle elevene, siden ungdomsskolene til sammen var mottakere av elever fra 8 barneskoler, og elevene derfor kunne ha noe ulik bakgrunn. Siden jeg skulle bruke resultatene til å velge ut elever til tre kompetansenivåer, var det viktig å ha oppgaver av ulik vanskelighetsgrad. Dette har i to tilfeller medført at til tross for lav diskriminering, ble oppgaven godkjent for testen, fordi oppgaven skilte godt mellom elevene på det kompetansenivået den var ment å si noe om.

Analysen av resultatene fra Test 1, viser en reliabilitet i form av Cronbachs Alfa lik 0,79 (se tabell 4.1). Det betyr at 79 prosent av et resultat er sann skåre, mens 21 prosent skyldes tilfeldigheter (er feil skåre, se kapittel 2.3.1).

Tabell 4.1

Tekniske data for Test 1

Antall elever	Gjennom-snittlig poengsum p-verdi	Standard-feil til gjennom-snittet	Standard-avvik (s)	Standard målefeil	Cronbachs alfa	Antall oppgaver
294	**12,8 p P = 0,64**	**0,2 p**	**3,8 p**	**1,7 p**	**,79**	**20**

Ved hjelp av standardavviket og Cronbachs alfa kan vi beregne standardfeilen til målingen og finne ut hvor nøyaktig vi kan bestemme resultatet til hver enkelt elev. Dette har betydning i min masterstudie siden jeg skulle bruke resultatene til å sette sammen elevene i par på grunnlag av poengene de oppnådde på prøven. For Test 1 er standardfeilen til målingen 1,7 p

(se kapittel 2.3.2). Det betyr at en elev har en sann skåre som med 68 prosent sannsynlighet ligger i intervallet elevens poengsum ± 1,7 poeng, og med 95 prosent sannsynlighet ligger i intervallet oppnådd poengsum ± 3,4 poeng (Kleven 2002).

294 elever deltok på Test 1 (se tabell 3.1 og 4.1). Det var 162 elever fra skole 1 (83 jenter og 79 gutter), og 132 elever fra skole 2 (58 jenter og 74 gutter). Resultatene viser at elevene i gjennomsnitt løste 64 prosent av oppgavene (se tabell 4.1). Standardfeil til gjennomsnittet var 0,2 poeng, og dette medførte at gjennomsnittlig poengsum lå i intervallet 12,4 til 13,2 poeng (12,8 p ± 2*0,2 p). Standardavviket var 3,8 poeng og gir at 95 prosent av elevene hadde poeng i intervallet 5,2 til 20 poeng. Dette viser god spredning i resultatene og det var viktig da jeg skulle ha elever til tre nivågrupper i Test 2.

Tabell 4.2

Oppgaver i Test 1, Cronbachs Alfa hvis oppgave fjernes, korrigert diskriminering, p-verdi for oppgaver, p- verdi ubesvart, diskriminering og signifikans. Rødt er åpne oppgaver.294 elever

Opp-gaver Test 1	Cronbachs Alfa hvis oppgaven fjernes	Korrigert diskriminering (Corrected item-total correlation)	Løsnings-prosent, full skår (p-verdi)	Ubesvarte Oppgaver (p-verdi)	Diskriminering (Pearsons korr.) for Test 1 med 20 oppgaver	a) sign. forskjell på 1p og 0 p i CR b) SR er korrekt i forhold til dyktighet c) lav diskriminering
A23	,781	,359	0,70	0,01	,457	a)
A25	,782	,342	0,80	0,05	,433	a)
A2	,777	,425	0,65	0,01	,525	b)
B29a	,774	,463	0,68	0,11	,551	a)
B29b	,769	,542	0,71	0,12	,622	a)
B8	,784	,313	0,61	0,02	,426	b)
A18	,781	,371	0,45	0,06	,452	a)
A35a	,790	,165	0,90	0,04	,242	c)
A35b	,793	,188	0,43	0,16	,310	a)
A37	,780	,372	0,47	0,01	,482	b)
A16	,789	,214	0,31	0,02	,307	
A9	,779	,411	0,68	0,01	,486	b)
A14a	,776	,438	0,74	0,00	,529	b)
A14b	,778	,410	0,47	0,10	,515	a)
A29	,784	,318	0,37	0,00	,416	
B31	,789	,199	0,84	0,00	,259	a) c)
A31	,784	,326	0,49	0,00	,440	b)
B34	,777	,425	0,79	0,02	,511	b)
B39	,775	,473	0,49	0,15	,552	a)
B3	,785	,295	0,29	0,04	,401	b)

Kolonne nummer to fra høyre i tabell 4.2 viser hvordan hver oppgave korrelerer (diskriminerer) med summen av alle oppgavene. Alle signifikante korrelasjoner er innenfor en sannsynlighet på 99 prosent. Oppgave A35a, A35b, A16 og B31 har imidlertid noe lav diskriminering (d), og de ligger i grenseland i forhold til anbefalinger fra Lie mfl. (2005). En tommelfingerregel er at diskrimineringen skal være større enn 0,3. I tillegg viser kolonne nummer to fra venstre i tabell 4.2 at bare oppgave A35b påvirker reliabiliteten negativt, og det i ubetydelig grad. Jeg vil i fortsettelsen kommentere noen resultater knyttet til disse fire oppgavene og også til den lukkede oppgaven som har lavest løsningsprosent (oppgave B3). Slik vil jeg belyse i hvilken grad disse oppgavene svekker reliabiliteten og validiteten til Test 1.

4.1.2 Elevenes dyktighet og prosent riktige svar

En god prøve skal være konstruert slik at alle elever får vist sin kompetanse. Den skal inneholde oppgaver på alle nivåer, og en oppgave ansees ikke for god hvis en stor del av elevene lar oppgaven stå ubesvart. Vi må ha svar fra elevene for at vi skal kunne vite noe om kunnskapene de har (Olsen mfl. 2001). Resultatene fra Test 1 viser løsningsprosenter fra 29 til 90 prosent på riktige svar (se tabell 4.2). 29 prosent er svært lavt i en flervalgsoppgave (MC) med fire svaralternativer. Hvis en elev gjetter og har maksimal flaks, kan han/hun oppnå 25 prosent rett. Sannsynligheten for dette er imidlertid liten, bare $9,5 \cdot 10^{-7}$ for 10 flervalgsoppgaver (MC) med fire alternativer. Det er allikevel viktig å undersøke hvor god kompetanse de elevene har som har svart riktig på oppgaver med lav løsningsprosent.

Tabell 4.3 viser data for oppgave B3 som var den lukkede oppgaven med lavest løsningsprosent i Test 1. Analysen av B3 viser at elevene som løste oppgaven, i gjennomsnitt var betydelig dyktigere (3,4 prosentpoeng) enn de som fikk 0 poeng på oppgaven. Bare 4 prosent lot oppgaven stå ubesvart (se tabell 4.2). Oppgave B3 er derfor et positivt bidrag i Test 1. Det samme gjelder oppgave A16 som var en lukket oppgave av type *matching.* Resultatene viser at de 92 elevene som svarte rett og fikk 1 poeng på oppgave A16, hadde en gjennomsnittlig poengsum på prøven på 14,3 poeng (se tabell 4.3). De som fikk delvis rett (0,5 p) eller feil (0 p) hadde lavere gjennomsnittspoengsum (12,8 p og 11,1 p). Resultatet på t-tester viser signifikant forskjell ($p = 0,02 < 0,05$) mellom elevene som fikk 0 poeng og 0,5 poeng på oppgaven, og mellom elevene som fikk 0,5 poeng og 1 poeng ($p = 0,02 < 0,05$).

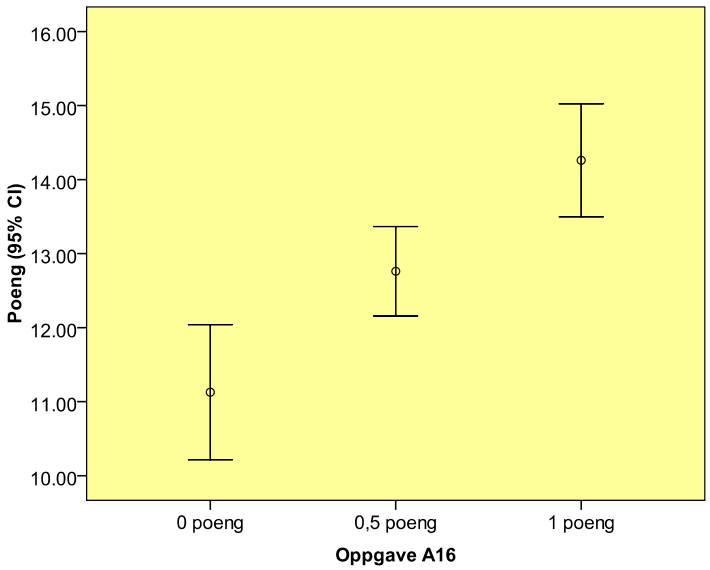
Tabell 4.3

Gjennomsnittlig poengsum på Test 1 for elever som fikk 0 poeng, 0,5 poeng og riktig svar på oppgave A16 og A35b, og 0 og 1 poeng på oppgave B3. 294 elever

Oppgave A16			**Oppgave B3**			**Oppgave A35b**		
Poeng på oppgave A16	Gj.snitt poeng på hele prøven	Antall elever	Poeng på oppgave B3	Gj.snitt poeng på hele prøven	Antall elever	Poeng på oppgave A35b	Gj.snitt poeng på hele prøven	Antall elever
0	11,1	74	0	11,8	208	0	11,8	160
0,5	12,8	128				0,5	12,1	8
1	**14,3**	**92**	**1**	**15,2**	**86**	**1**	**14,2**	**126**

Figur 4.1 viser signifikans mellom alle svarmulighetene, ved at det ikke er overlapp mellom områdene til gjennomsnittsverdiene for de tre gruppene i oppgave A16. Det var de riktige elevene ut fra dyktighet som hadde løst oppgaven, og oppgaven fungerer godt selv om det er en lukket oppgave med lav p-verdi. At bare 2 prosent av elevene lot være å svare på oppgaven, styrker oppgaven ytterligere.

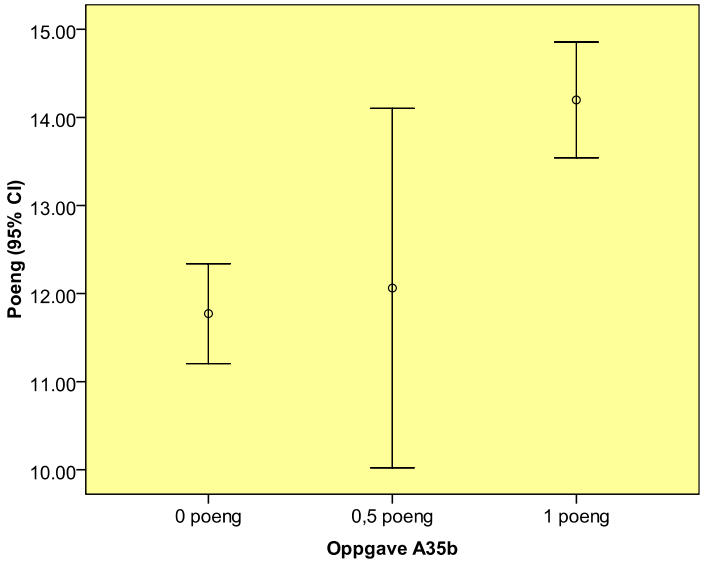
Figur 4.1 *Gjennomsnittlig poengsum for elever som fikk 0 p, 0,5 p og 1 p på oppgave A16. 294 elever*



A35b er den eneste oppgaven som trekker ned Cronbachs alfa, men bare i svært liten grad (0,003). T-tester viser signifikante forskjeller mellom de som svarte feil på oppgaven og de som svarte riktig. De 8 elevene som fikk delvis riktig, har imidlertid poengsummer fra 9 til 15,5 poeng og er ikke signifikant forskjellig verken fra de som fikk 0 poeng eller de som fant riktig løsning. Figur 4.2 viser tydelig overlapp mellom områdene til de som fikk 0,5 poeng på

oppgaven og hver av de andre to gruppene. I tillegg svekkes oppgaven ved at 16 prosent av elevene lot oppgaven stå ubesvart. Siden uregelmessighetene i oppgave A35b bare gjelder 8 av 294 elever og oppgaven har en diskriminering over godkjent grense (d = 0,310 > 0,3) fungerer oppgaven allikevel godt nok til å kunne være en del av Test 1 og telle med i resultatene.

Figur 4.2 *Gjennomsnittlig poengsum på hele prøven for elever som har 0 p, 0,5 p og 1 p på oppgave A35b. Resultatet av T-test innenfor et konfidensintervall på 95 %. 294 elever*



A35a og B31 er åpne oppgaver med høy p-verdi (p = 0,90 og p = 0,84) og vil naturlig nok skille dårlig mellom elevene fordi de fleste elevene greide å løse oppgavene. For A35a hadde de som løste oppgaven, i gjennomsnitt 3 prosentpoeng mer på hele prøven enn de som fikk 0 poeng på oppgaven, og prosent ubesvart var 4 prosent. Oppgave A35a skiller derfor i virkeligheten godt mellom elevene. Tilsvarende forskjell for oppgave B31 var 5 prosentpoeng (13 p – 8 p). De som fikk 0,5 poeng på oppgave B31 hadde imidlertid i gjennomsnitt 12 poeng på prøven, og var nesten like flink som de som fikk 1 poeng på oppgaven (gjennomsnitt 13 poeng). Årsaken ligger i upresist svar, ved at de som fikk 0,5 poeng svarte ”....luft” i stedet for ”....oksygen” som var riktig svar. Presis bruk av begreper er viktig å fokusere på, og oppgaven ble derfor godkjent til å være en del av Test 1. Ingen elever lot oppgaven stå ubesvart.

4.1.3 Vanskelighetsgrad og kjønnsforskjeller for Test 1

Test 1 hadde gjennomsnittlig p-verdi 0,64 og svarprosent 95,4. Høy løsnings- og svarprosent var en fordel med tanke på at resultatene skulle brukes til å etablere to sammenlignbare grupper av elever. Det var viktig å få svar fra flest mulig på alle oppgavene slik at grunnlaget for Test 2, skulle bli så riktig som mulig.

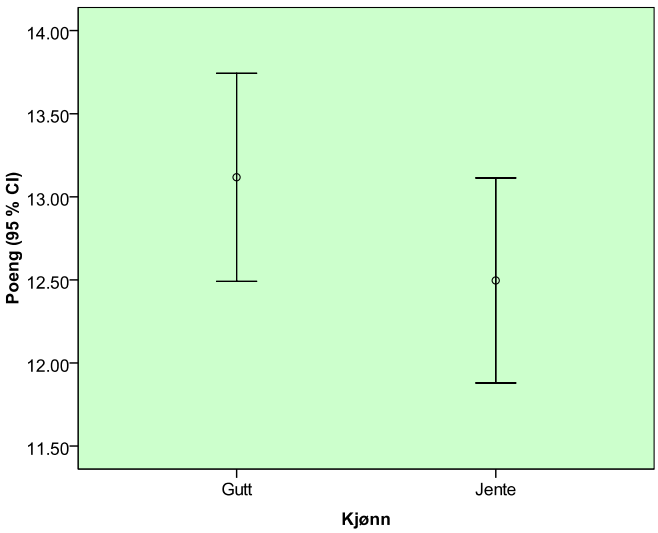
Tabell 4.4

Gjennomsnittlig poengsum, standardavvik, standardfeil til gjennomsnittet og effektstørrelser for jenter og gutter på Test 1

Test 1	Gutter	Jenter	Effektstørrelse
Antall	153	141	0,16
Gjennomsnitt	13,1 p (65,5 %)	12,5 p (62,5 %)	
Standard feil til gj.snitt	,3 p	,3 p	
Standard avvik	3,9 p	3,7 p	

Et ledd i kvalitetssikringen av en test, er å undersøke i hvilken grad testen faller likt ut for jenter og gutter. Analysene baserer seg på resultatene til 141 jenter og 153 gutter. Resultatene viser at guttene, med en standardfeil til gjennomsnittet på ca. 0,3 poeng, har en middelverdi som ligger mellom 12,5 og 13,7 poeng (se tabell 4.4). Jentenes middelverdi ligger i intervallet 11,9 til 13,1 poeng med samme standardfeil på 0,3 poeng. Det betyr at middelverdien for både gutter og jenter kan ligge i intervallet 12,5 til 13,1 poeng, og at det er ingen signifikant kjønnsforskjell i Test 1. Grafisk fremstilling i figur 4.3 viser hvordan områdene for middelverdiene overlapper med en sannsynlighet på 95 prosent. Effektstørrelsen utregnet med ”pooled” standardavvik med hensyn på kjønn, er 0,16, som betyr 16 prosent av standardavviket og er svært lav.

Figur 4.3 Gjennomsnittlig poengsum på Test 1 for jenter og gutter. Konfidensintervall på 95 %



4.1.4 Konklusjon for kvalitetssikring av Test 1

Resultatene fra utvalget på 294 elevbesvarelser, viser at Test 1 har reliabilitet i form av Cronbachs alfa lik 0,79 og oppgavene svarprosenter fra 29 til 90 prosent. Det var de "dyktigste" elevene som svarte riktig også på oppgaver med lav løsningsprosent, og en p-verdi på 0,64 viser at testen ikke var for vanskelig. Det er ingen signifikant forskjell på middelverdien for jenter og gutter. Samvariasjonen (diskrimineringen) mellom hver oppgave og total sum er undersøkt ved å bruke Pearsons korrelasjonstest. Analysene viser at oppgavene hver for seg og samlet som prøve, har god nok kvalitet til å kunne være utgangspunkt for å velge ut elever til to tilnærmet like testgrupper.

I tillegg til resultatene som er kommentert i de forrige punktene, har jeg undersøkt gjennomsnittlig poengsum til de elevene som valgte de ulike svaralternativene i de lukkede oppgavene (se eks i tabell 4.5). Et kriterium for at en oppgave fungerer godt, er at de flinkeste elevene velger det riktige svaralternativet. I oppgave A37 er C riktig svar på oppgaven. De elevene som svarte C, hadde i gjennomsnitt 15 poeng på Test 1, og var klart bedre enn de elevene som svarte et av de andre alternativene. Dette ser vi også av korrelasjonen mellom svaralternativene og summen. Alternativene A, B og D korrelerer negativt, mens C korrelerer positivt og viser samsvar med summen av oppgaver. Tilsvarende resultat ser vi i oppgave A14a, men ikke i A31. I A31 var de elevene som svarte B, i gjennomsnitt praktisk talt like dyktige som de som svarte riktig (D). Dette gjenspeiler seg ved at også alternativ B korrelerer svakt positivt. Se figur 4.4. for oppgaven det gjelder. Siden distraktor B representerer en velkjent misoppfatning, er resultatet ikke overraskende. Av pedagogiske grunner er det viktig at slike oppgaver er med i en test. Alle elevene svarte på oppgave A31.

Tabell 4.5

Sammenhengen mellom elevenes dyktighet og valg av svaralternativer i tre lukkede oppgaver. Løsningsprosenter, kjønnsforskjeller, oppgavenes diskriminering og svaralternativenes korrelering med summen av oppgavene i Test 1. 294 elever. Gult for riktig alternativ

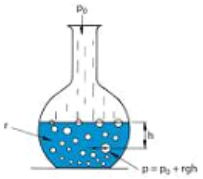
Opp-gave	Gjennomsnittlig poengsum på Test 1					Diskri-minering	Andel som har 1p på oppgaven	Forskjell i p-verdi kjønn	SR - Korrelasjon svaralternativ og sum			
	A	B	C	D	Blank	d-verdi	p-verdi	j-g	A	B	C	D
A37	9	12	15	6	9	0,48	0,47	-0,05	-0,233	-0,293	0,482	-0,245
A14a	14	10	8	10	4	0,53	0,74	-0,16	0,529	-0,218	-0,282	-0,306
A31	10	14	10	15		0,44	0,49	-0,02	-0,191	0,100	-0,425	0,440

Figur 4.4 *Oppgave fra Test 1. Distraktor B er en utbredt misoppfatning. 16 % av guttene og 9 % av jentene svarte B. Disse elevene var i gjennomsnitt like dyktige som de som valgte riktig svar, D. 294 elever*

Oppgave A 31

Hva skjer med vannmolekylene når vann fordamper?

A De forsvinner
B De blir større og lettere
C De blir til luft
D De fjerner seg fra hverandre



4.2. Resultater fra holdningsoppgavene til Test 1

Holdningsoppgavene (se figur 3.7 og 3.8) ble analysert i SPSS på samme måte som prøveresultatene. Utsagnene ble vurdert på en skala som gjorde at jo lavere poengsum en elev fikk, jo mer positiv var elevens holdninger til faget, egen kompetanse og egen prestasjon. 8 spørsmål er et tynt grunnlag å kjøre analyser på, men kan allikevel gi noen signaler. 289 elever svarte på alle spørsmålene og reliabiliteten i form av Cronbachs alfa var 0,63 (se tabell 4.6). Det betyr at det er en viss usikkerhet knyttet til målefeil for dette resultatet. Resultatene på holdningsundersøkelsen ble kun brukt som justering da elevene ble satt sammen i par.

Tabell 4.6

Cronbachs alfa for åtte holdningsspørsmål relatert til Test 1 høsten 2007. 289 elever

Cronbachs alfa	Antall spørsmål
,63	8

Resultatene viser at det er forskjell på holdningene til gutter og jenter i det utvalget som ble analysert (se tabell 4.7). Med middelverdi 14,7 poeng i forhold til jentenes 16,1 poeng, gav guttene i større grad enn jentene, uttrykk for positiv innstilling til naturfag, og hadde bedre selvtillit og tro på egen kompetanse i faget. T-test viser at forskjellen er signifikant innenfor konfidensintervall på 99 prosent. Effektstørrelsen er 0,38, og viser middels effekt av kjønn.

Tabell 4.7

Resultater for jenter og gutter på 8 spørsmål som handler om holdninger til og interesse for naturfag, og egenvurdering av kompetanse i faget. Jo lavere verdi, jo mer positiv holdning

Gutt	Antall	153	Effektstørrelse
	Middelverdi	**14,7 p**	
	Standardfeil til middelverdi	,3 p	
	Standard avvik	3,6 p	
Jente	Antall	141	0,38
	Middelverdi	**16,1 p**	
	Standardfeil til middelverdi	,3 p	
	Standard avvik	3,8 p	

Pearsons korrelasjonstest viser at alle oppgavene korrelerer med summen (diskriminerer) med verdier mellom 0,42 og 0,63. Med så lav Cronbachs alfa ble imidlertid konklusjonen at resultatene på holdningsspørsmålene måtte behandles med varsomhet, men at de kunne vektlegges til en viss grad da elevene ble satt sammen i par for å bestemme hvem som skulle få prøve 1 og prøve 2 i Test 2. Ved at elevene ble grovsortert ut fra svært høye og svært lave verdier, overtolket jeg ikke holdningsresultatene, samtidig som de fikk telle med (se vedlegg 4).

4.3. Oppsummering Test 1

I denne delen av oppgaven har jeg lagt stor vekt på å vise hvordan jeg gikk fram for å kvalitetssikre Test 1. Svar på utsagn relatert til holdninger ble analysert for å begrunne i hvilken grad disse kunne vektlegges ved utvelgingen av elevene. Det er umulig å etablere to helt like grupper av mennesker. Her har jeg prøvd å gjøre det på grunnlag av en test som i stor grad oppfyller vanlige psykometriske krav innen testteori. En undersøkelse inneholder imidlertid både systematiske og tilfeldige målefeil som kan gi varierende utslag. Hovedfokus i analysene av den andre testen (Test 2), er derfor ikke lagt på enkeltelever, men på tre nivågrupper. Før jeg redegjør for resultatene relatert til forskningsspørsmålene, vil jeg vise at prøve 1 og prøve 2 fungerer etter intensjonene og bestemme i hvilken grad vi kan stole på resultatene til disse prøvene.

4.4. Resultater for kvalitetssikring av Test 2

Test 2 bestod av prøve 1 og prøve 2. Gruppe G1 fikk prøve 1 og gruppe G2 fikk prøve 2 (se figur 3.1, kapittel 3.1). Prøvene var like i betydningen av at begge inneholdt 10 åpne og 10 lukkede oppgaver. Når det gjelder innhold, var det 20 ulike oppgaver, men alle oppgavene forelå både i åpent og lukket format (se kapittel 3.3.1., tabell 3.6 og 3.7).

Til sammen 306 elever deltok på Test 2, med fordelingen 151 elever på prøve 1 og 155 elever på prøve 2. Av disse var det 79 gutter og 72 jenter på prøve 1 og 79 gutter og 76 jenter på prøve 2. Fordelingen av jenter og gutter var tilnærmet lik innenfor hver skole (se tabell 3.2). Dette er gunstig når man skal sammenligne resultater med hensyn på kjønn. Før jeg kunne sammenligne resultatene fra prøve 1 med resultatene fra prøve 2, måtte jeg undersøke om prøvene hver for seg tilfredsstilte psykometriske krav til reliabilitet og validitet. Dette var nødvendig på samme måte som for Test 1, for å få vite i hvilken grad vi kan stole på resultatene.

4.4.1 Resultater for å undersøke kvaliteten på prøve 1 i Test 2

Tabell 4.8

Tekniske data for prøve 1 i Test 2

Antall elever Test 2 Prøve 1	Gjennom-snittlig poengsum og p-verdi	Standard-feil til gjennom-snittet	Standard-avvik (s)	Standard målefeil	Cronbachs alfa	Inter-item korrela-sjon	Antall oppgaver
151	**10,3 p p = 0,52**	**0,3 p**	**3,2 p**	**1,7 p**	**0,703**	**0,10**	**20**

Resultatene viser at elevene i gjennomsnitt løste 52 prosent av oppgavene, at gjennomsnittlig poengsum med 95 prosent sannsynlighet lå mellom 9,7 og 10,9 poeng (± 2*0,3 p), og at 95 prosent av elevene hadde poeng i intervallet 3,9 til 16,7 poeng (± 2*3,2 p). Reliabiliteten (Cronbachs alfa) var 0,70.

Resultatene viser at oppgavene i prøve 1 i Test 2 hadde p-verdier fra 0,03 til 0,96. Oppgaven med lavest p-verdi, fA7, var en lukket oppgave av type *matching,* hvor seks begreper skulle knyttes opp mot delene på en blomst. Bare 5 elever (3 %) greide alle seks, mens 89 prosent greide fem riktige. Det ble kun gitt poeng for helt riktig svar. ”Pollenknapp” var ukjent begrep for 96 prosent av elevene. Oppgaven hadde diskriminering 0,03 som henger sammen både med svært lav p-verdi, og at oppgaven ikke skilte mellom elever som fikk 1 poeng og 0 poeng

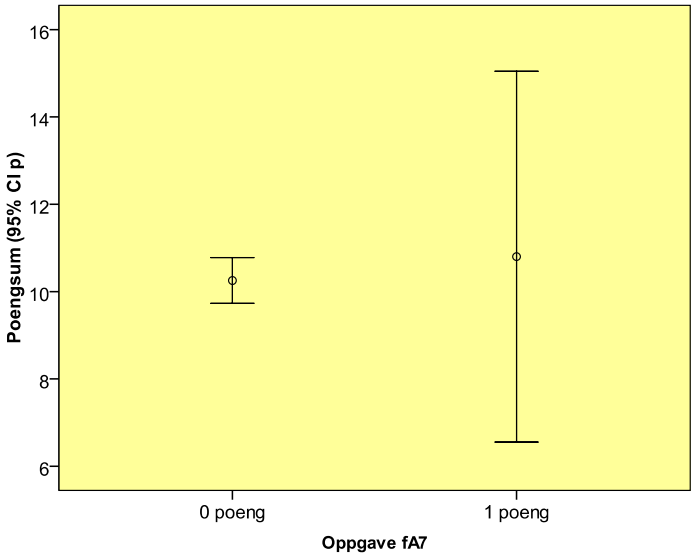
på oppgaven. De fem elevene som fikk 1 poeng på oppgaven hadde 6, 10, 10, 13 og 15 poeng på prøven. Figur 4.5 viser at det er tilfeldig hvem som har fått riktig svar på oppgaven, og at det ikke er samsvar mellom elevenes totale poengsum på prøven og poeng på oppgaven. Resultatene til oppgave fA7, kan vi ikke stole på (se figur 2.3).

Tabell 4.9

Oppgavene i Test 2 prøve 1, Cronbachs alfa hvis en oppgave fjernes, korrigert diskriminering, p-verdi for hver oppgave, p- verdi for ubesvarte oppgaver, diskriminering og signifikans. Åpne oppgaver er merket med rød skrift. 151 elever

Opp-gaver Test 2 Prøve 1	**Cronbachs alfa hvis oppgaven fjernes**	**Korrigert diskriminering (Corrected item-total correlation)**	**Løsnings-prosent, full skår (p-verdi)**	**Ubesvarte Oppgaver (p-verdi)**	**Diskriminering (Pearsons korr.) for Test 1 med 20 oppgaver**	**a) sign. forskjell på 1p og 0 p i CR b) SR er korrekt i forhold til dyktighet c) lav diskriminering**
fA40	,695	,253	0,92	0,01	,332	b)
fA22	,702	,136	0,93	0,00	,215	c)
åB14	,698	,219	0,30	0,55	,353	
åB41a	,682	,385	0,79	0,03	,491	a)
åB41b	,695	,250	0,72	0,01	,380	a)
åA3	,679	,392	0,65	0,11	,515	a)
fA33	,694	,267	0,52	0,01	,409	b)
åB12	,684	,354	0,23	0,38	,466	a)
åA34	,682	,369	0,53	0,23	,500	a)
fA32	,692	,280	0,78	0,01	,399	b)
åB25	,697	,220	0,13	0,09	,318	a)
åB33	,668	,489	0,33	0,16	,597	a)
fA41	,697	,235	0,68	0,01	,370	
fB10	,680	,384	0,66	0,01	,507	b)
åB30	,701	,152	0,06	0,04	,224	a) c)
fA19	,699	,212	0,34	0,05	,352	b)
åA28	,677	,418	0,33	0,51	,535	a)
fA7	,708	-,025	0,03	0,01	,031	c)
fA36	,707	,145	0,42	0,01	,294	c)
fB17	,703	,106	0,96	0,02	,166	c)

Figur 4.5 *Gjennomsnittlig poengsum på prøve 1 Test 2 for elever som har 0 p og 1 p på oppgave fA7. Resultatet av T-test innenfor et konfidensintervall på 95 %*



Oppgavene fA22, åB30, fA36 og fB17 hadde også diskriminering lavere enn 0,3. fA22 var en lukket oppgave av formatet flervalg (MC) med p-verdi 0,93. Distraktoren ”månen lyser fordi den er glødende” ble valgt av fem elever som hadde fra 5 til 12 poeng på prøven, mens de som fikk et poeng på oppgaven, hadde 11 poeng i gjennomsnitt. Hvis vi derimot sammenligner elevene som fikk 1 poeng med alle som fikk 0 poeng på oppgaven, er forskjellen signifikant. Oppgaven senker heller ikke prøvens reliabilitet (se tabell 4.9). Alle elevene svarte på oppgave fA22.

Oppgave åB30 var en åpen oppgave som har p-verdi 0,06 og lav diskriminering. Det var imidlertid de flinkeste elevene som fikk 1 poeng på oppgaven, og forskjellen mellom de som fikk 0 poeng og 1 poeng er signifikant. Gjennom oppgave åB30 får jeg bekreftet en misoppfatning om at de fleste elevene (p = 0,64) tror at oksygen er den gassen det er mest av i luft. Slike oppgaver er det viktig å ha med for å få avdekket misoppfatninger.

Oppgave fA36 hadde diskriminering 0,29, var en flervalgsoppgave (MC) og inneholdt en negasjon. Temaet i oppgaven var kjemisk reaksjon og elevsvarene fordelte seg med 23 og 30 prosent på to av distraktorene (C og D) i tillegg til at 42 prosent av elevene fikk riktig svar på oppgaven. To elever lot oppgaven stå ubesvart. Disse to elevene hadde høyere gjennomsnitt enn elevene som svarte riktig. Hvis vi sammenligner gjennomsnittspoengsummene til elevene som fikk 1 poeng på oppgaven med de som fikk 0 poeng, er forskjellen imidlertid signifikant

innenfor et konfidensintervall på 95 prosent. Oppgaven hadde derfor god nok kvalitet for formålet den skulle brukes til (se figur 4.6).

Figur 4.6 Oppgave fra Test 2 prøve 1. Flervalgsoppgave med lav p-verdi

Oppgave fA36

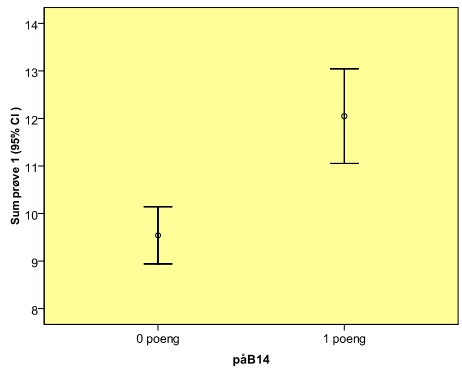
Hva av dette er *ikke* et eksempel på en kjemisk reaksjon?

A Vann som koker
B Jern som ruster
C Ved som brenner
D En eplebit som blir brun

FB17 er en lukket oppgave av type *matching*. Oppgaven har p-verdi 0,96 og lav diskriminering. Når p-verdien er så høy, - 145 av 151 elever har fått 1 poeng på oppgaven -, vil den naturlig nok ikke skille mellom elevenes kompetanse. Når i tillegg en elev som fikk 15 poeng på prøven, ikke svarte på oppgave fB17, vil dette også påvirke diskrimineringen negativt. De fem andre som fikk 0 poeng på oppgaven, hadde fra 5 til 8 poeng på hele prøven. Oppgaven påvirket ikke prøven negativt, men den fortalte ingenting om forskjellen i kompetanse mellom elevene. Den har imidlertid godt nok resultat til å kunne sammenlignes med tilsvarende oppgave i åpent format.

Oppgave åB14 er en åpen oppgave som må kommenteres fordi 55 prosent av elevene lot oppgaven stå ubesvart. Bukhulen er tydeligvis et ukjent begrep for mange elever. Elevene som fikk 1 poeng på oppgaven, var imidlertid signifikant bedre enn de som fikk 0 poeng (se figur 4.7), og oppgaven hadde diskriminering 0,353.

Figur 4.7 Gjennomsnittlig poengsum på prøve 1 Test 2 for elever som har 0 p og 1 p på oppgave åB14. Resultatet av T-test innenfor et konfidensintervall på 95 %



Tilsvarende resultater viser oppgave åA28. 51 prosent av elevene lot oppgaven stå ubesvart, men oppgaven skilte godt mellom elevene ved at de som løste oppgaven, i gjennomsnitt hadde 4 poeng mer på prøven enn de som ikke fikk poeng på oppgaven. Analysen viser ingen signifikant forskjell på resultatene til jenter og gutter på prøve 1 (se tabell 4.10). Effektstørrelsen ved sammenligning av jenter og gutter utregnet med "pooled" standardavvik er 0,09 og svært lav.

Tabell 4.10

Antall gutter og jenter, gjennomsnittlig poengsum, standardavvik, standardfeil til gjennomsnittet og effektstørrelse med hensyn på kjønn for prøve 1 i Test 2

Test 2 Prøve 1	**Gutter**	**Jenter**	**Effektstørrelse**
Antall	79	72	0,09
Gjennomsnitt	10,4 p (52 %)	10,1 p (51 %)	
Standard feil til gj.snitt	,4p	,4p	
Standard avvik	3,2 p	3,2p	

4.4.2 Resultater for å undersøke kvaliteten på prøve 2 i Test 2.

Tabell 4.11

Tekniske data for prøve 2 i Test 2

Antall elever Test 2 Prøve 2	**Gjennom-snittlig poengsum og p-verdi**	**Standard-feil til gjennom-snittet**	**Standard-avvik**	**Standard målefeil**	**Cronbachs alfa**	**Inter-item Korrela-sjon**	**Antall oppgaver**
155	**9,8 p P = 0,49**	**0,3 p**	**3,5 p**	**1,9 p**	**0,713**	**0,11**	**20**

Resultatene viser at elevene i gjennomsnitt løste 49 prosent av oppgavene, og at gjennomsnittlig poengsum med 95 prosent sannsynlighet lå mellom 9,2 og 10,4 poeng (± 2*0,3 p). 95 prosent av elevene hadde poeng i intervallet 2,8 til 16,8 poeng (± 2*3,2 p). Reliabiliteten til prøven var 0,71 (se tabell 4.11), og oppgavene hadde p-verdier fra 0,05 til 0,81.

fB14, fA3, fB25, fB30 og åA7, var oppgaver som måtte undersøkes nærmere på grunn av lav diskriminering, lav p-verdi eller høy prosent ubesvarte (se tabell 4.12). Oppgave fB14 var en flervalgsoppgave (MC) med p-verdi 0,52 og diskriminering 0,30. Litt lav diskriminering kan skyldes at 8 av elevene som svarte feil og fikk 0 poeng på oppgaven, var like dyktige i

gjennomsnitt som elevene som fikk 1 poeng. Hvis vi sammenligner alle elevene og lar de som lot oppgaven stå ubesvart også telle med, var de som fikk 1 poeng på oppgaven signifikant bedre enn de som fikk 0 poeng. Begrepet ”bukhule” ser ut til å være problematisk for mange elever uansett om oppgaven var åpen eller lukket. Prosent ubesvarte i flervalg var imidlertid bare 5 prosent i motsetning til 55 prosent som var andel ubesvarte da oppgaven var åpen i prøve 1.

fA3 var en flervalgsoppgave (MC) hvor 4 elever som fikk 0 poeng, var like flinke i gjennomsnitt som de som fikk rett på oppgaven. Oppgaven hadde p-verdi 0,79, og det var signifikant forskjell mellom gjennomsnittlig løsningsprosent på hele prøven for de elevene som fikk 0 poeng og de som fikk 1 poeng på oppgaven. Dette godkjenner oppgaven.

Selv om oppgave fB25 hadde lav diskriminering, var forskjellen mellom elevene som fikk 0 og 1 poeng, signifikant. To av distraktorene i flervalgsoppgaven (MC) ble valgt av henholdsvis 14 og 35 prosent av elevene, og dette kan ha påvirket diskrimineringen. Oppgaven er omtalt i figur 4.17.

Flervalgsoppgaven (MC) fB30 hadde p-verdi 0,08. Til tross for lav p-verdi fungerte oppgaven godt både når det gjelder dyktigheten til elever som hadde løst oppgaven, sett i forhold til de som fikk 0 poeng på oppgaven, og diskrimineringen. Resultatet bekrefter misoppfatningen om at oksygen er gassen det er mest av i luft. Dette er viktig informasjon til læreren, og er samme resultat som da oppgaven var åpen i prøve 1.

Tabell 4.12

Oppgavene i Test 2 prøve 2, Cronbachs alfa hvis en bestemt oppgave fjernes, korrigert diskriminering, p-verdi for hver oppgave, p- verdi for ubesvarte oppgaver, diskriminering og signifikans. Åpne oppgaver med rød skrift

Opp-gaver Test 2 Prøve 2	**Cronbachs Alfa hvis oppgaven fjernes**	**Korrigert diskriminering (Corrected item-total correlation)**	**Løsnings-prosent, full skår (p-verdi)**	**Ubesvarte Oppgaver (p-verdi)**	**Diskriminering (Pearsons korr.) for Test 1 med 20 oppgaver**	**a) sign. forskjell på 1p og 0 p i CR b) SR er korrekt i forhold til dyktighet c) lav diskriminering**
åB27	,691	,392	0,57	0,21	0,509	a)
åA22	,701	,304	0,77	0,16	0,413	a)
fB14	,714	,165	0,52	0,05	0,304	
fB41a	,703	,271	0,68	0,01	0,393	b)
fB41b	,697	,342	0,71	0,01	0,455	b)
fA3	,709	,203	0,79	0,01	0,314	
åA33	,701	,301	0,26	0,27	0,414	a)
fA15	,703	,279	0,61	0,05	0,406	b)
fA34	,692	,383	0,61	0,02	0,500	b)
åA32	,692	,382	0,59	0,03	0,499	a)
fB25	,718	,133	0,48	0,00	0,274	c)
fB33	,701	,297	0,30	0,03	0,416	b)
åA41	,706	,249	0,52	0,10	0,382	a)
åB10	,699	,325	0,77	0,11	0,431	a)
fB30	,696	,429	0,08	0,00	0,493	b)
åA19	,707	,230	0,25	0,33	0,346	b)
fA28	,707	,221	0,81	0,04	0,328	b)
åA7	,714	,092	0,05	0,03	0,155	c)
åA36	,700	,307	0,27	0,43	0,420	a)
åB17	,699	,329	0,18	0,07	0,427	a)

Oppgave åA7 hadde p-verdi 0,05 og diskriminering 0,16. Det var 8 elever som hadde rett svar og fikk 1 poeng. Disse elevene hadde 12 poeng i gjennomsnitt på prøven, og var 2 poeng dyktigere enn elevene som fikk 0 poeng. Spredningen blant de som fikk rett svar, var imidlertid fra 6 til 18 poeng på prøven totalt, og viser at oppgaven gav ingen tilbakemelding om kompetanse. Oppgave A7 fungerte dårlig både som åpen og som lukket oppgave.

Resultatene viser at det var ingen signifikant forskjell på gjennomsnittlig poengsum for jenter og gutter på prøve 2, selv om guttenes resultat var 4 prosentpoeng høyere i gjennomsnitt enn jentenes. Effektstørrelsen med ”pooled” standardavvik var 0,23, som bekrefter lav effekt av kjønn (se tabell 4.13).

Tabell 4.13

Antall gutter og jenter, gjennomsnittlig poengsum, standardavvik og standardfeil til gjennomsnittet for Test 2, prøve 2

Test 2 Prøve 2	Gutter	Jenter	Effektstørrelse
Antall	79	76	0,23
Gjennomsnitt	10,2 p (51 %)	9,4 p (47 %)	
Standard feil til gj.snitt	,4 p	,4 p	
Standard avvik	3,7 p	3,2 p	

4.4.3 Sammenligning av tekniske data for prøve 1 og prøve 2

Tabell 4.14

Tekniske data for prøve 1 (P1) og prøve 2 (P2). Antall elever (N)

Test 2	N	Ant. oppg.	Gj.sn. poeng	P-verdi	Cronbachs alfa	Std. avvik	Std. målefeil	Std.feil til gj.sn.	Inter item korr.	Effekt
Prøve 1	151	20	10,3 p	0,52	0,703	3,2 p	1,7 p	0,3 p	0,10	0,15
Prøve 2	155	20	9,8 p	0,49	0,713	3,5 p	1,9 p	0,3 p	0,11	
Jente P1	72	20	10,1 p	0,51		3,2 p		0,3 p		0,22
Jente P2	76	20	9,4 p	0,47		3,2 p		0,4 p		
Gutt P1	79	20	10,4 p	0,52		3,2 p		0,3 p		0,06
Gutt P2	79	20	10,2 p	0,51		3,7 p		0,4 p		

Resultatene viser at gjennomsnittlig løsningsprosent for prøve 2 var omtrent tre prosentpoeng lavere enn for prøve 1. Forskjellen var ikke statistisk signifikant (t-test), og prøvene hadde tilnærmet lik reliabilitet. Standardavvikene viser litt større spredning av resultatene i prøve 2 enn i prøve 1 (se tabell 4.14), men prøvene synes å oppføre seg som tilnærmet parallelle tester. En effektstørrelse på 0,15 bekrefter lav effekt på total poengsum avhengig av prøve. For guttene var effekten av hvilken prøve en elev deltok på 0,06 og for jentene 0,22. Begge verdiene betyr lav effekt, selv om effekten for jentene var større enn for guttene. Lave effektstørrelser og ingen signifikante forskjeller gav et godt utgangspunkt for å sammenligne resultatene på åpne og lukkede oppgaver.

4.5. Resultater til forskningsspørsmål 1

Hvilket samsvar er det mellom resultatet en gruppe elever oppnår på naturfagoppgaver i åpent format (CR, - Constructed Response Items), og resultatet som oppnås av en sammenlignbar elevgruppe når oppgaver med samme opplysninger og spørsmål er i formatet lukket (SR, - Selected Response Items)?

4.5.1 Sammenligning av resultatene på 10 åpne og 10 lukkede oppgaver

Kvalitetssikringen av Test 1 og Test 2 (prøve 1 og prøve 2) har skjedd på bakgrunn av resultatene til alle elevene som deltok på Test 1 og/eller Test 2. Analysen videre baserer seg på resultatene til de 275 elevene som deltok både på Test 1 høsten 2007 og Test 2 våren 2008. Det var resultatene til disse 140 elevene (prøve 1) og 135 elevene (prøve 2) som ble analysert for å gi svar på problemstillingen og forskningsspørsmålene. Se tabell 3.4. for antall elever fra hver skole og antall jenter og gutter, og tabell 4.15 som viser tekniske data for prøvene i Test 2, ut fra resultatene til de 275 elevene som deltok både høsten 2007 og våren 2008. Begge prøvene fikk høyere reliabilitet (0,715 og 0,728) i dette utvalget av elever, men forskjellene for øvrig var ubetydelig endret, og grunnlaget for sammenligning var på ingen måte svekket ved at utvalget ble redusert til 275 elever.

Tabell 4.15

Tekniske data for prøve 1 og prøve 2 for utvalget på 275 elever (N)

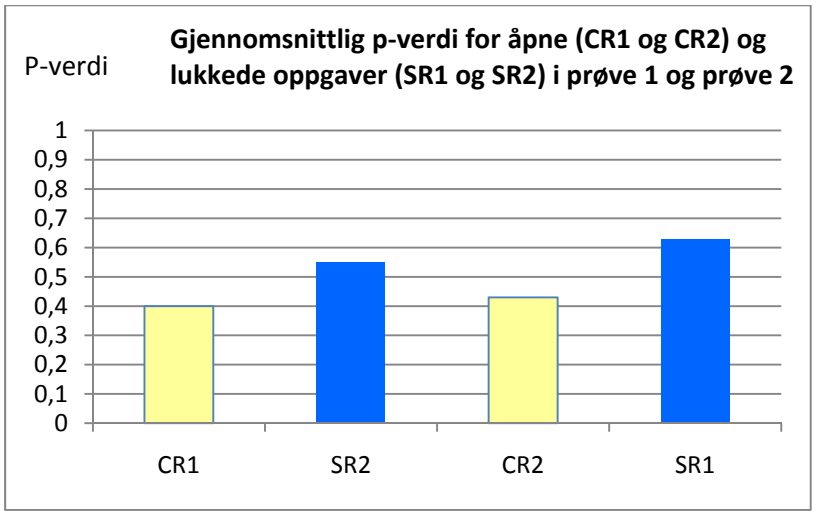
Test 2	N	Ant. oppg.	Gj.sn. poeng	P-verdi	Cronbachs alfa	Std. avvik	Std. målefeil	Std.feil til gj.sn.	Inter item korr.	Effekt
Prøve 1	140	20	10,3 p	0,51	0,715	3,3 p	1,8 p	0,28 p	0,10	0,14
Prøve 2	135	20	9,8 p	0,49	0,728	3,6 p	1,9 p	0,31 p	0,12	

Inter-item korrelasjonen i begge prøvene var lav, 0,10 og 0,12, og gir signal om at oppgavene i prøve 1 og i prøve 2 var mer ulike enn hva som kunne vært ønskelig. Å ha stor spredning i tema i oppgavene var et bevisst valg fra min side. Valget er begrunnet i kapittel 3.3.1.

Resultatene for gruppe G1 som fikk prøve 1, viser at elevene i gjennomsnitt løste 40 prosent av de åpne oppgavene (CR1, p = 0,40) og 63 prosent av de lukkede (SR1, p = 0,63). For elevene i gruppe G2 som fikk prøve 2, var løsningsprosenten i de åpne oppgavene i gjennomsnitt 43 prosent (CR2, p = 0,43) og for de lukkede 55 prosent (SR2, p = 0,55). Hvis

vi sammenligner resultatene til gruppe G1 med resultatene til gruppe G2, ser vi at de åpne oppgavene med p-verdi 0,40 i gruppe G1 hadde p-verdi 0,55 som lukkede oppgaver i gruppe G2. Tilsvarende for de andre 10 oppgavene var p-verdi 0,43 som åpne oppgaver og 0,63 i lukket format (se figur 4.8).

Figur 4.8 *Gul søyle viser gjennomsnittlig p-verdi for løste åpne oppgaver og blå søyle for løste oppgaver i lukket format. CR1 (åpne prøve 1)og SR2 (lukkede prøve 2) er oppgaver med samme innhold og motsatt format. Tilsvarende for CR2 og SR1. Resultatene er for 275 elever som gjennomførte både Test 1 og Test 2*



Dette viser at begge prøvene hadde lavere gjennomsnittlig løsningsprosent for de åpne enn for de lukkede oppgavene, samtidig som grupper av oppgaver hadde lavere gjennomsnittlig løsningsprosent da oppgavene med samme innhold var åpne enn da de var lukkede. Effektstørrelsene viser at da vi sammenlignet åpent og lukket oppgaveformat både i hver gruppe G1 og G2 og på tvers av gruppene, hadde oppgaveformatet fra middels til stor effekt (se tabell 4.16).

Tabell 4.16

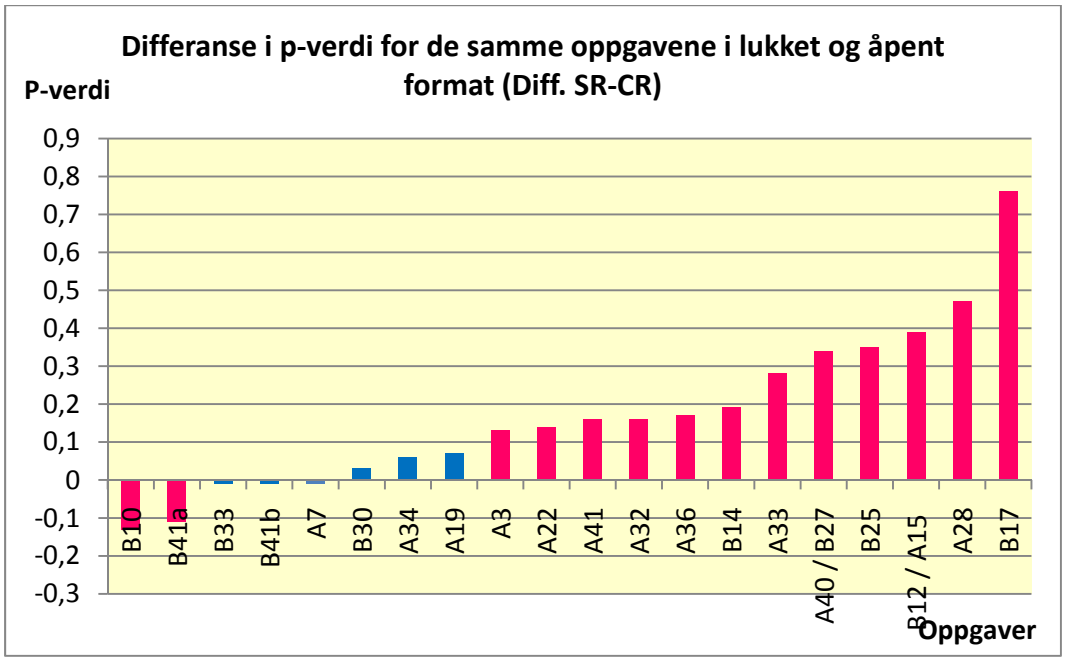
Effektstørrelser for sammenligning av åpne (CR) og lukkede oppgaver (SR) i Test 2. 275 elever

CR sammenlignet med SR	Effektstørrelse (e)	Grad av effekt
Gruppe G1 (CR1 og SR1 i prøve 1)	e = 1,17	Stor effekt (> 0,8)
Gruppe G2 (CR2 og SR2 i prøve 2)	e = 0,6	Middels effekt (0,3 < e < 0,8)
CR1 og SR2	e = 0,7	Middels effekt (0,3 < e < 0,8)
CR2 og SR1	e = 0,9	Stor effekt (> 0,8)

4.5.2 Enkeltoppgaver som ikke var signifikant forskjellig i åpent og lukket format

T-tester på alle oppgavene opp mot prøve 1 og prøve 2 viser at 14 oppgaver var signifikant forskjellig i åpent og lukket format. Oppgavene B33, B41b, A7, B30, A34 og A19 var ikke signifikant forskjellig som åpne og lukkede oppgaver, og hadde en differanse i p-verdi mellom åpent og lukket format fra 1 til 7 prosentpoeng. I figur 4.9 er disse oppgavene markert med blå søyler.

Figur 4.9 *Forskjell i p-verdi for samme oppgave i lukket og åpent format.140 elever på prøve 1 og 135 elever på prøve 2*



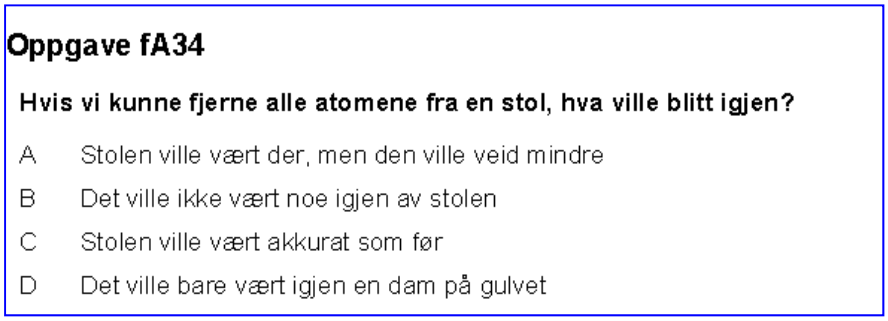
Oppgavene B33, B41b, A34, og A19 hadde diskriminering høyere enn 0,3 i begge formatene, og skilte godt mellom elevene. B30 skilte godt i lukket format, men ikke som åpen oppgave, og A7 hadde lav diskriminering i begge formatene. Av disse seks var det en fysikkoppgave, tre kjemi- og to biologioppgaver. I diskusjonen i kapittel 5 vil jeg beskrive nærmere hva som kjennetegner disse oppgavene.

Oppgave B41b (se figur 3.5 og 3.6) var en fysikkoppgave som hadde p-verdi 0,71 i åpent format og 0,70 i lukket. Som svaralternativ i lukket format, var listet opp de mulige betraktningene som det er naturlig å gjøre uansett om oppgaven er åpen eller lukket. Da oppgaven var åpen, svarte 13,6 prosent av elevene at det største hjulet ville bruke kortest tid

på en omdreining. I lukket format var det 5,2 prosent som mente dette, mens 15,6 prosent mente hjulene ville bruke like lang tid. Bare 1 prosent valgte ikke å svare på oppgavene.

Oppgave A34 var en oppgave som krevde at elevene kjente til og forstod konsekvensene av at alle ting er bygd opp av atomer. Oppgaven hadde samme ordlyd i åpent og lukket format (se figur 4.10), og p-verdien var 0,53 som åpen og 0,59 som lukket. 23 prosent svarte ikke på oppgaven da den var åpen, men som lukket var ubesvart bare 1,5 prosent. Det var også 23 prosent som valgte alternativ A, og trodde stolen fortsatt ville være der, men veie mindre om atomene ble fjernet. Tilnærmet like mange jenter som gutter valgte alternativ A.

Figur 4.10 *Kjemioppgave som ikke var signifikant forskjellig i åpent (p = 0,53) og lukket format (p = 0,59). Tester forståelse*



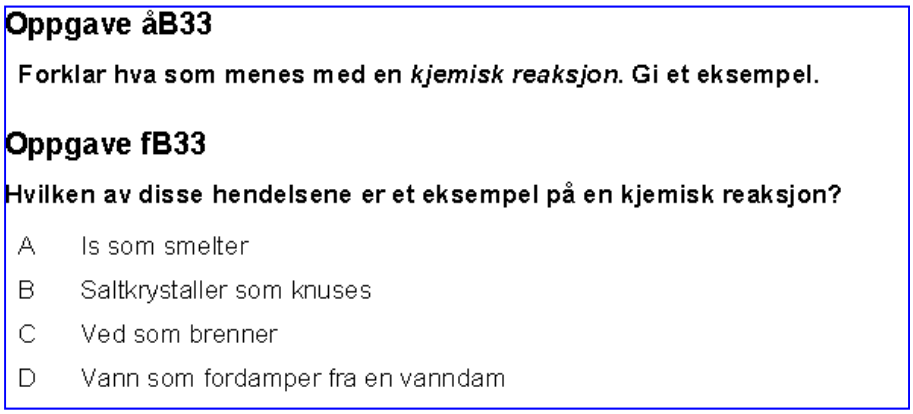
Oppgave fA34

Hvis vi kunne fjerne alle atomene fra en stol, hva ville blitt igjen?

A Stolen ville vært der, men den ville veid mindre
B Det ville ikke vært noe igjen av stolen
C Stolen ville vært akkurat som før
D Det ville bare vært igjen en dam på gulvet

I B33 valgte 30 prosent av elevene alternativ D i lukket format, og dette var ett prosentpoeng mindre enn de som fikk riktig svar på oppgaven. Alle alternativene var aktuelle svar. Hver femte elev svarte alternativ A, og 13 % svarte B (se figur 4.11).16 prosent av elevene svarte ikke da oppgaven var åpen og som lukket var ubesvart 1,5 prosent.

Figur 4.11 *Oppgave som ikke var signifikant forskjellig i åpent (p = 0,32) og lukket format (p = 0,31). Tester fakta som åpen og forståelse som lukket*



Oppgave åB33

Forklar hva som menes med en *kjemisk reaksjon*. Gi et eksempel.

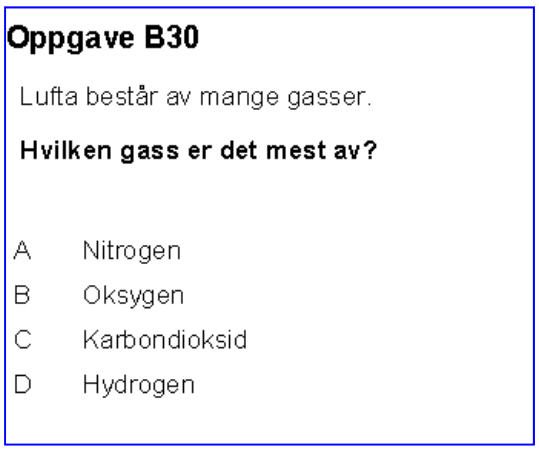
Oppgave fB33

Hvilken av disse hendelsene er et eksempel på en kjemisk reaksjon?

A Is som smelter
B Saltkrystaller som knuses
C Ved som brenner
D Vann som fordamper fra en vanndam

I oppgave B30 var 73 prosent av elevene overbevist om at oksygen var riktig svar på oppgaven, og bare 9 prosent valgte A som er riktig svar (se figur 4.12). Ingen lot oppgaven stå ubesvart som lukket, og 4,3 prosent svarte ikke da den var åpen. Som åpen oppgave var løsningsprosenten 6 prosent, og 63 prosent av elevene svarte oksygen.

Figur 4.12 *Oppgave som ikke var signifikant forskjellig i åpent (p = 0,06) og lukket format (p = 0,09). Kjemi og faktastoff. Lav diskriminering som åpen oppgave*



Oppgave B30

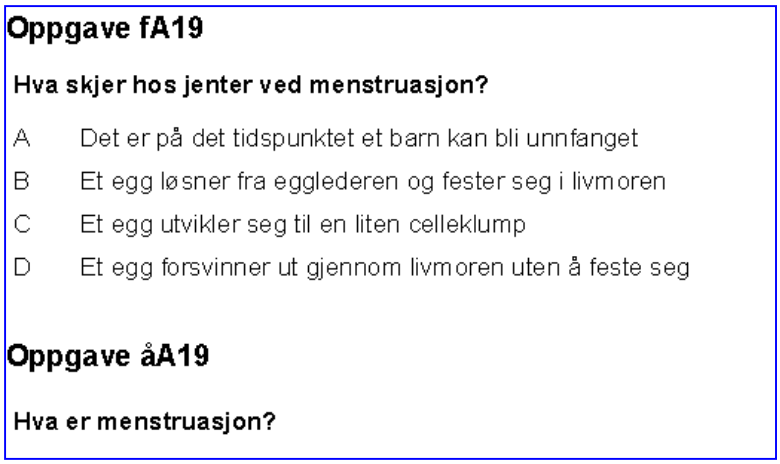
Lufta består av mange gasser.

Hvilken gass er det mest av?

A Nitrogen
B Oksygen
C Karbondioksid
D Hydrogen

I oppgave A19 (se figur 4.13) svarte 45 prosent av elevene alternativ B, 10 prosent svarte A og 6 prosent svarte C. 5,7 prosent lot oppgaven stå ubesvart i lukket format, mens 32 prosent lot være å svare da den var åpen. Oppgaven viste store kjønnsforskjeller i jentefavør i begge formatene.

Figur 4.13 *Oppgave som ikke var signifikant forskjellig i åpent (p = 0,26) og lukket format (p = 0,33). Emne biologi*



Oppgave fA19

Hva skjer hos jenter ved menstruasjon?

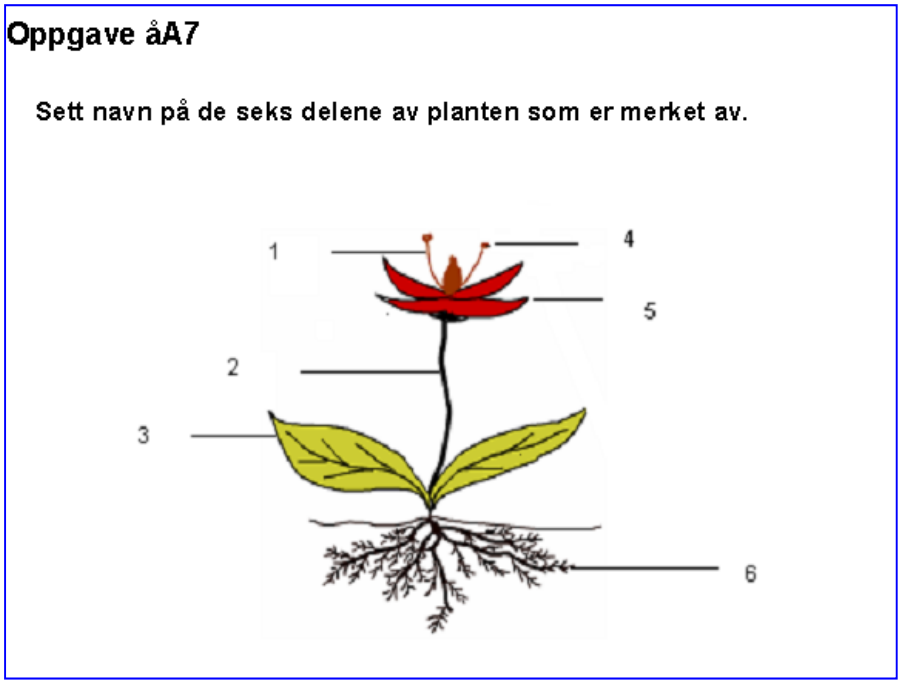
A Det er på det tidspunktet et barn kan bli unnfanget
B Et egg løsner fra eggleder en og fester seg i livmoren
C Et egg utvikler seg til en liten celleklump
D Et egg forsvinner ut gjennom livmoren uten å feste seg

Oppgave åA19

Hva er menstruasjon?

Oppgave A7 hadde svært lav løsningsprosent og lav diskriminering (se figur 4.14). Oppgaven er omtalt både i kapittel 4.4.1.og 4.4.2, og var en oppgave som ikke bidro positivt til prøvene. Det var tilfeldig hvilke elever som fikk riktig på oppgaven. Både i åpent og lukket format var utfordringen at elevene ikke kjente begrepet pollenknapp. Bare 1 prosent av elevene lot oppgaven stå ubesvart. For oppgave A7 i lukket format, se figur 2.3.

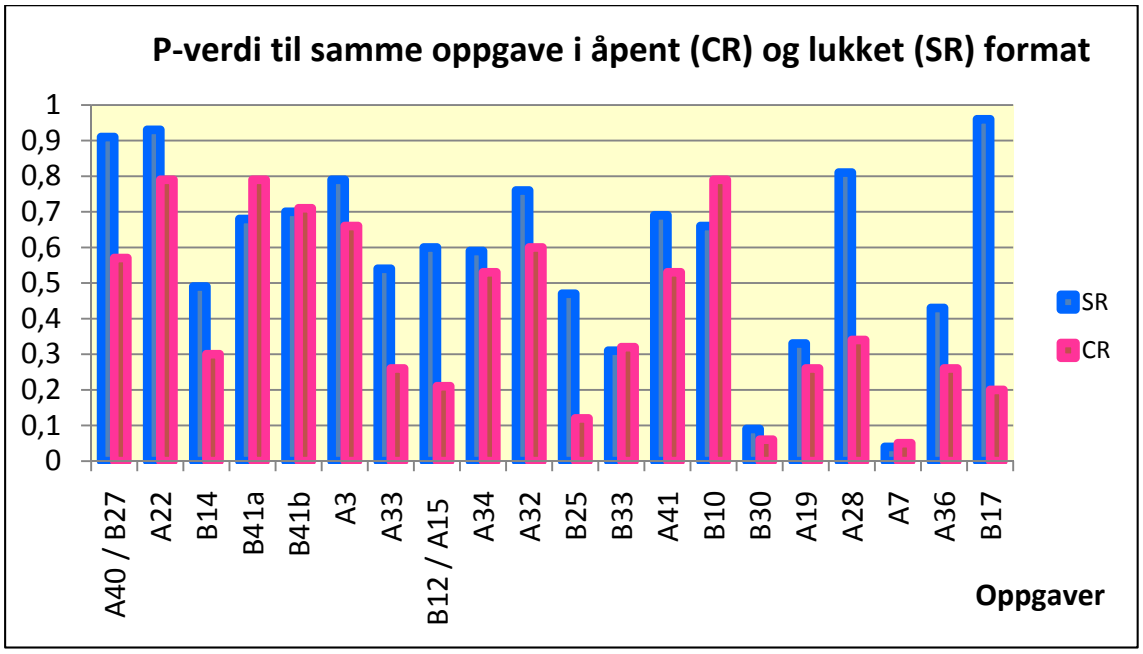
Figur 4.14 *Oppgave som ikke var signifikant forskjellig i åpent (p = 0,05) og lukket format (p = 0,04). Skilte ikke mellom elevene*



4.5.3 Oppgaver som var signifikant forskjellig i åpent og lukket format, og hadde betydelig differanse i p-verdi

Av de fjorten oppgavene med statistisk signifikant forskjellige p-verdier i åpent og lukket format, hadde oppgavene A40/B27, B12/A15, B25, A28 og B17 størst forskjell i p-verdi med differanse fra 34 til 76 prosentpoeng (se figur 4.15). Dette var en fysikkoppgave og fire biologioppgaver.

Figur 4.15 *P-verdier til oppgaver i åpent (CR) og lukket (SR) format. Resultater til 140 elever på prøve 1 og 135 elever på prøve 2*



For alle disse fem oppgaveparene hadde de lukkede formatene høyest p-verdi. B17 hadde lav diskriminering i åpent format (d = 0,160) og var den eneste av disse fem parene som ikke diskriminerte godt både som åpen og lukket oppgave. A40/B27 handler om lyn og torden (se figur 3.3) og hadde 34 prosentpoeng høyere p-verdi i lukket enn i åpent format. I lukket format var det ingen elever som valgte alternativ B, og bare en elev som valgte A. I realiteten var det derfor bare to alternativer å velge mellom for elevene. 20 prosent svarte ikke da oppgaven var åpen.

B12/A15 handler om forskjellen på varmblodige og vekselvarme dyr. Oppgavene var det paret som var mest ulike innholdsmessig i åpent og lukket format. Som åpen oppgave fokuseres det på begrepet vekselvarmt, og 40 prosent svarte ikke på oppgaven. I lukket format er det hovedfokus på varmblodighet i sammenligning med vekselvarme, og her var det bare 4 prosent ubesvart. En distraktor ble valgt av bare 4 elever. Forskjellen i prosentpoeng mellom åpent og lukket oppgave var 39 (se figur 4.16).

Figur 4.16 *Oppgave med 39 prosentpoeng forskjell i p-verdi mellom lukket og åpent format. Oppgaven krever faktakunnskap. For ulik til sammenligning*

Oppgave åB12(A15)

Hva betyr det at et dyr er vekselvarmt?

Oppgave fA15

På hvilken måte er varmblodige dyr forskjellige fra vekselvarme dyr?

A Hos varmblodige dyr øker stoffskiftet i varmt vær

B Varmblodige dyr er mer fiendtlige i fangenskap

C Varmblodige dyr har alltid høyere temperatur i blodet

D Varmblodige dyr har konstant kroppstemperatur uavhengig av temperaturen i omgivelsene

Oppgave B25 (se figur 4.17) hadde samme ordlyd i begge formatene, men oppgavene fortonte seg tydeligvis ulikt for elevene. Alternativ C ble valgt av bare 5 elever. Allikevel fungerte oppgaven godt i settet ved at det var elevene med høyest gjennomsnittlig poengsum som løste oppgavene uansett format. Åpen utgave hadde 9 prosent ubesvarte, og i lukket format var det ingen.

Figur 4.17 *Oppgave med 35 prosentpoeng forskjell mellom lukket og åpent format. Faktakunnskap*

Oppgave fB25

Tegningen viser to forskjellige fjellområder. Fjellene i område A er ujevne og taggete. Fjellene i område B er glatte og avrundede

Hva er trolig grunnen til at fjellene ser så forskjellige ut?

A Fjellområde A er eldst

B Fjellområde B er eldst

C Fjellområdene er like gamle, men B er nedslitt av fotturister

D Fjellområdene er like gamle, men A har opprinnelig vært en vulkan

De to siste oppgavene hadde henholdsvis 47 og 76 prosentpoeng forskjell mellom lukket og åpen oppgave. A28 spør om hvordan fossilt brensel dannes. Oppgaven hadde ulik formulering i åpent og lukket format. 51 prosent av elevene lot oppgaven stå ubesvart i åpent format, og i lukket format var tilsvarende prosent 3. Alternativ B var valgt av bare tre elever, og de to andre distraktorene hadde også lav svarprosent.

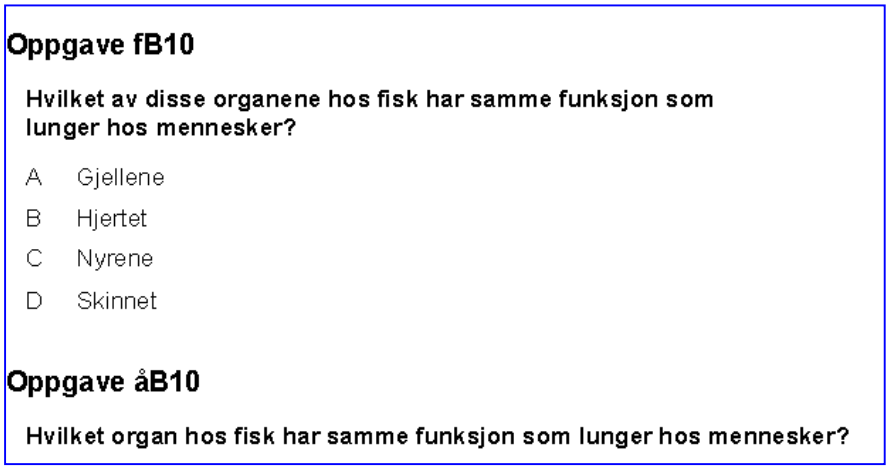
B17 var i lukket format av type *matching* med fire alternativer hvor hjertet, lungene, huden og skjelettet skulle kobles til tekstbokser som handler om organenes oppgave i kroppen. Som åpen oppgave måtte elevene selv fortelle om oppgaven til disse fire organene. Bare helt riktig svar for alle fire gav ett poeng. Andel ubesvart var lavt i begge formatene, med 1 prosent i lukket format og 4 prosent som åpen. B17 skilte svært dårlig mellom elevene i lukket format fordi 96 prosent av elevene løste oppgaven riktig.

4.5.4 Oppgaver med høyere p-verdi som åpen enn som lukket oppgave

To oppgaver hadde signifikant høyere løsningsprosent i åpent enn i lukket format. Det var en fysikk- og en biologioppgave. Oppgave B41a er omtalt på side 30 og 31. Oppgaven krever logisk resonnement, mer enn kunnskaper innenfor noen deler av naturfaget. Som lukket oppgave var det to elever som ikke svarte på oppgaven, og distraktor 1 ble valgt av bare tre elever. Da oppgaven var åpen var det fire elever som ikke svarte på oppgaven.

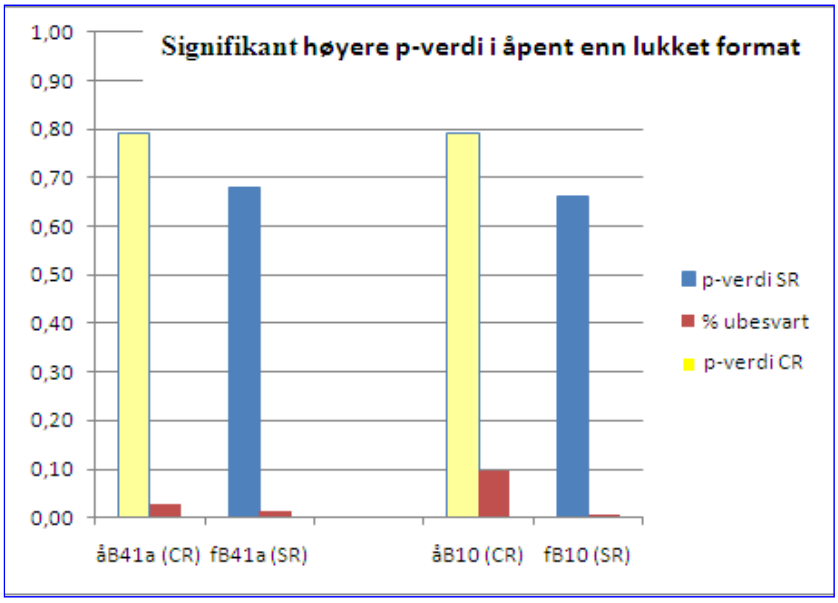
Oppgave B10 (se figur 4.18) tester faktakunnskaper. 18 prosent av elevene svarte alternativ B (hjertet) og 14 prosent C (nyrene). Prosent ubesvart var 10 prosent da oppgaven var åpen og 1 prosent som lukket oppgave.

Figur 4.18 Oppgave med høyere p-verdi som åpen enn som lukket oppgave



Effektstørrelsen for oppgave B41a i åpent og lukket format var 0,25 og for B10 var den 0,27. Verdiene er innenfor grensen til lav effekt. Det er ikke grunn til å tro at andel ubesvarte har påvirket resultatet, siden oppgaven med høyest løsningsprosent hadde høyest andel ubesvarte av disse fire oppgavene (se figur 4.19).

Figur 4.19 P-verdi som åpen (CR) og lukket oppgave (SR), og andel ubesvarte i hvert format



4.5.5 Oppsummering av forskningsspørsmål 1

Resultatene viser at elevene i gjennomsnitt fikk flere riktige svar når oppgavene var i lukket enn i åpent format. Samme resultat fikk vi også innen gruppe G1 og gruppe G2 på prøve 1 og prøve 2. Resultatene viser også at det er grunn til å anta at både distraktorenes utforming, om det er fakta eller forståelse som testes, og elevenes forkunnskaper og etablerte misoppfatninger, har betydning for resultatet. Dette blir behandlet under drøfting av resultater i kapittel 5.

4.6. Resultater til forskningsspørsmål 2

Dersom vi tar utgangspunkt i elevenes faglige nivå, er det noen elevgrupper som i større grad enn andre, har overvekt av riktige løsninger innenfor et av oppgaveformatene?

Ut fra resultatene på Test 1 ble elevene delt inn i tre nivågrupper, nivå 1, 2 og 3, hvor 1 var laveste og 3 høyeste nivå. Dette ble gjennomført for hver skole, og for jenter og gutter hver for seg.

Nivå 1 bestod av de 25 prosent av elevene som hadde fått færrest poeng på Test 1, og nivå 3 av de 25 prosentene som hadde fått flest poeng på testen. Inndelingen er detaljert beskrevet i forklaringene til tabell 3.9, 3.10, 3.11 og 3.12. Inndelingen var nødvendig for å være sikker på at det ble elever på alle nivåene i begge testgruppene, gruppe G1 og gruppe G2, til Test 2.

I analysene av resultatene fra Test 2 har jeg ikke fokusert på om enkeltelever presterte på samme nivå i Test 2 som i Test 1. Jeg har på grunnlag av resultatene fra Test 2 delt elevene inn i nye nivågrupper, med tilnærmet samme prosentfordeling som etter Test 1. Resultater på hvert nivå er analysert, og det er gjort sammenligninger både innen nivåene og nivåene i mellom på oppgaveformat.

4.6.1 Resultater på oppgaveformat sett i forhold til faglig nivå for elevene

Elevene som fikk prøve 1 i Test 2, oppnådde fra 3 til 18 poeng, og de som gjennomførte prøve 2, fikk fra 1 til 18 poeng. Ingen elever fikk full skår på noen av prøvene. Ut fra poengene elevene oppnådde, ble de delt inn i tre nivågrupper for hver prøve. Inndelingen i nivågrupper ble prøvd lagt så nært opp til prosentfordelingen 25 - 50 - 25 som mulig, tilsvarende inndelingen etter Test 1 (se tabell 4.17 og 4.18).

Tabell 4.17

Nivågrupper ut fra resultater på prøve 1. P-verdier for lukkede (SR1) og åpne oppgaver (CR1) på hvert nivå, forskjell i prosentpoeng (Diff. pp SR1-CR1), prosent ubesvarte totalt og for hvert format og effektstørrelser

Prøve 1	**% elever**	**Antall elever**	**Poeng**	**p-verdi SR1**	**p-verdi CR1**	**Diff. pp SR1-CR1**	**p-verdi ubesvart prøve 1**	**p-verdi SR1 ubesvart**	**p-verdi CR1 ubesvart**	**Effekt av SR1/ CR1**
Nivå 1	24	30	$3 \leq X \leq 7$	0,44	0,17	0,27	0,16	0,02	0,31	2,7
Nivå 2	55	77	$8 \leq X \leq 13$	0,64	0,40	0,24	0,13	0,01	0,24	1,8
Nivå 3	21	33	$14 \leq X \leq 18$	0,81	0,68	0,13	0,04	0,01	0,07	1,5
Sum		**140**								

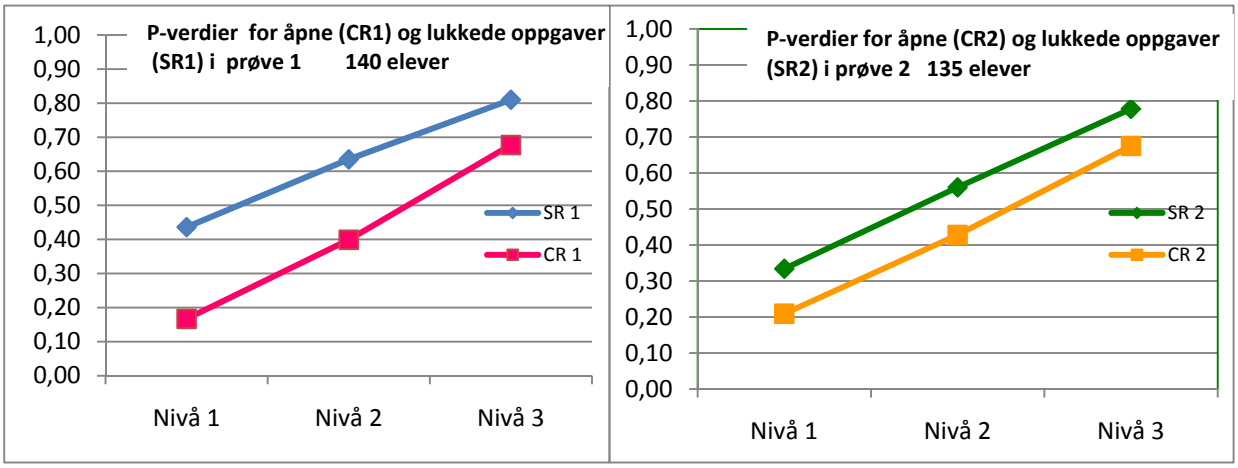
Tabell 4.18

Nivågrupper ut fra resultater på prøve 2. P-verdier for lukkede (SR2) og åpne oppgaver (CR2) på hvert nivå, forskjell i prosentpoeng (Diff. pp SR2-CR2), prosent ubesvarte totalt og for hvert format og effektstørrelser

Prøve 2	% elever	Antall elever	Poeng	p-verdi SR2	p-verdi CR2	Diff. pp SR2-CR2	p-verdi ubesvart prøve 2	p-verdi SR2 ubesvart	p-verdi CR2 ubesvart	Effekt av CR2/ SR2
Nivå 1	26	35	$1 \leq X \leq 7$	0,33	0,21	0,12	0,14	0,03	0,25	1,1
Nivå 2	50	68	$8 \leq X \leq 12$	0,56	0,43	0,13	0,09	0,02	0,16	1,1
Nivå 3	24	32	$13 \leq X \leq 18$	0,78	0,68	0,10	0,04	0,00	0,07	1,0
Sum		**135**								

For begge prøvene gjelder at på alle nivå har elevene i gjennomsnitt fått flere poeng på de lukkede enn på de åpne oppgavene. Effektstørrelsen for oppgaveformat er stor på alle nivåene for begge prøvene, og effekten er størst på nivå 1 og minst på nivå 3. Dette er mest tydelig for prøve 1, hvor effektstørrelsen er 2,7 på nivå 1 og 1,5 på nivå 3. Figur 4.20 viser fordelingen av riktige svar på åpne og lukkede oppgaver på hvert nivå for prøve 1 og prøve 2.

Figur 4.20 P-verdier på nivå for åpne (CR) og lukkede oppgaver (SR) i prøve 1 og prøve 2



Resultatene på prøve 1 og prøve 2 var grunnlaget for nivågruppene, og er viktig å kommentere av den grunn. I forhold til problemstillingen er det imidlertid resultatene på oppgavene som er ment å samsvare i åpent og lukket format, som må studeres (se tabell 4.19, 4.20 og figur 4.21).

Tabell 4.19

Nivågrupper for sammenligning av oppgavene som svarer til hverandre, i åpent (CR2) og lukket format (SR1)

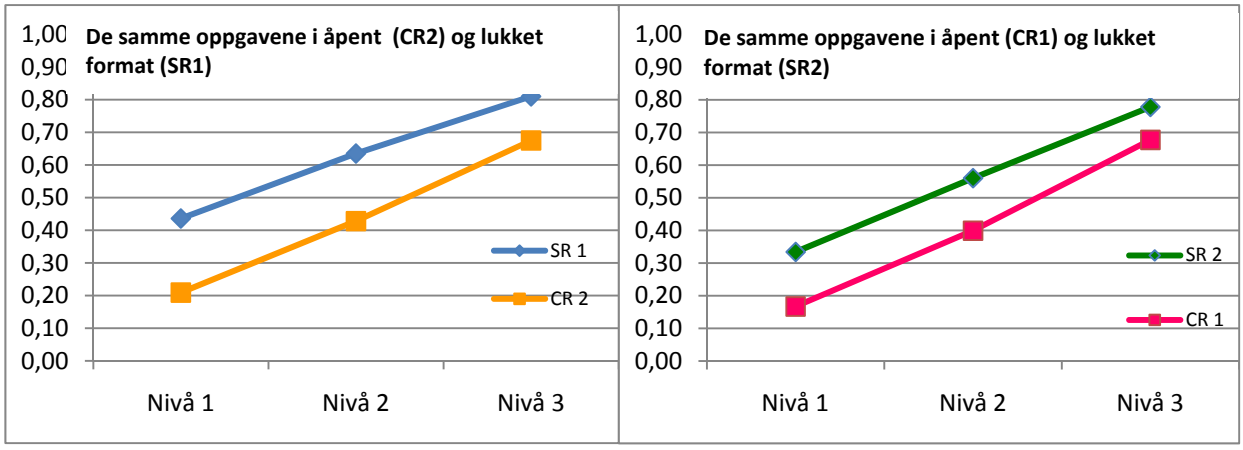
	Antall prøve 1	Antall prøve 2	p-verdi SR1	p-verdi CR2	Diff. pp SR1-CR2	p-verdi SR1 ubesvart	p-verdi CR2 ubesvart	Effekt av SR1/ CR2
Nivå 1	30	35	0,44	0,21	0,23	0,02	0,25	2,3
Nivå 2	77	68	0,64	0,43	0,21	0,01	0,16	1,9
Nivå 3	33	32	0,81	0,68	0,13	0,01	0,07	1,5
Sum	**140**	**135**						

Tabell 4.20

Nivågrupper for sammenligning av oppgavene som svarer til hverandre, i åpent (CR1) og lukket format (SR2)

	Antall prøve 1	Antall prøve 2	p-verdi SR2	p-verdi CR1	Diff. pp SR2-CR1	p-verdi SR2 ubesvart	p-verdi CR1 ubesvart	Effekt av SR2/ CR1
Nivå 1	30	35	0,33	0,17	0,16	0,03	0,31	1,5
Nivå 2	77	68	0,56	0,40	0,16	0,02	0,24	1,4
Nivå 3	33	32	0,78	0,68	0,10	0,00	0,07	1,0
Sum	**140**	**135**						

Figur 4.21 P-verdier på nivå for samme gruppe av oppgaver i åpent (CR) og lukket format (SR)



Resultatene viser at jo mer faglig sterke elevene var, jo mindre betydning hadde oppgaveformatet. For de faglig svakeste elevene, ser det ut til at oppgaveformatet hadde svært stor betydning. Elevene på nivå 1 som fikk de åpne oppgavene CR2, greide i gjennomsnitt å løse 21 prosent av disse 10 oppgavene. Elevene som fikk de samme oppgavene i lukket format (SR1), løste i gjennomsnitt mer enn dobbelt så mange oppgaver (dvs. mer enn 100 %

økning). For elevene på nivå 2 økte løsningsprosenten fra CR2 (p = 0,43) til SR1 (p = 0,64) med 50 prosent, og på nivå 3 var økningen fra p = 0,68 til p = 0,81, dvs. 19 prosent. Effektstørrelsene viser at oppgaveformatet har betydning på alle nivåer, men størst er betydningen for elevene på nivå 1. Effektstørrelsen for CR2 og SR1 var 2,3 på nivå 1, 1,9 på nivå 2, og på nivå 3 var verdien 1,5. For CR1 og SR2 var de tilsvarende tallene 1,5, 1,4 og 1,0 (se tabell 4.19 og 4.20).

4.6.2 Sammenheng ubesvarte oppgaver og nivå

Resultatene viser at på alle nivåer er andel ubesvarte oppgaver i gjennomsnitt høyere for de åpne enn for de lukkede oppgavene. Prosent ubesvarte er høyest på nivå 1 og lavest på nivå 3, og dette gjelder både for åpne og lukkede oppgaver (se tabell 4.19 og 4.20). Prosent ubesvarte er imidlertid svært lav på de lukkede oppgavene på alle nivåer, mens ubesvart på de åpne er betydelig på nivå 1 og avtar til nivå 3 hvor gjennomsnittet er 7 prosent.

4.6.3 Åpne og lukkede oppgaver for elever på nivå 1

Tabell 4.21 viser sju oppgaver med stor forskjell i p-verdi mellom lukket og åpent format, for elever på nivå 1. Dette gjelder fem faktaoppgaver og to forståelsesoppgaver. Fire av oppgavene har tema knyttet til biologi, to til kjemi og en til fysikk.

Tabell 4.21

Oppgaver med stor forskjell i p-verdi i åpent og lukket format på nivå 1. P-verdi 0,70 betyr at 70 prosent av elevene på nivå 1 løste oppgaven. Oppgaver markert med gult er lukkede oppgaver (SR)

Oppgave	Innhold	p-verdi prøve 1	p-verdi prøve 2		Diff. SR-CR	CR p-verdi ubesvart	SR p-verdi ubesvart
B17	Kropp (B)	0,88	0,06	Fakta	0,82	0,09	0,03
A28	Brensel (B)	0,09	0,66	Fakta	0,57	0,67	0,06
A40 / B27	Torden og lyn (F)	0,70	0,29	Fakta	0,41	0,37	0,03
B12 / A15	Vekselvarm (B)	0,00	0,31	Fakta	0,31	0,49	0,11
A3	Fordamping (K)	0,33	0,63	Forståelse	0,30	0,21	0,00
A33	Gass (K)	0,30	0,06	Fakta	0,24	0,34	0,00
B25	Korrolering (B)	0,00	0,23	Forståelse	0,23	0,18	0,00

Resultatene kan tyde på at elevene hadde større hjelp av svaralternativer i faktaoppgaver enn i forståelsesoppgaver. Det kan også være interessant å undersøke om differanse i p-verdi ut fra oppgaveformat hadde sammenheng med at elevene lot være å svare på oppgaver.
Resultatene gav meg ingen grunn til å trekke konklusjoner om at det var noen sammenheng mellom p-verdi og andel ubesvarte, selv om det på nivå 1 var stor forskjell i både p-verdier og andel ubesvarte i åpent og lukket format. På nivå 1 hadde for eksempel bare 6 prosent av

elevene løst oppgave B17 i åpent format, men det var bare 8,6 prosent av elevene som ikke hadde svart på oppgaven. Til sammenligning hadde 33 prosent løst oppgave A3 som åpen oppgave, og da var andel ubesvarte 21 prosent. Ingen elever hadde riktig løsning på oppgave B25, men bare 18 prosent lot være å svare på oppgaven.

4.6.4 Oppsummering av forskningsspørsmål 2

På alle nivå var gjennomsnittlig p-verdi høyere på de lukkede enn på de åpne oppgavene. Dette gjaldt både i prøve 1, i prøve 2 og da vi sammenlignet åpne oppgaver i prøve 1 med lukkede i prøve 2 og vice-versa. Forskjellen i p-verdier for åpne og lukkede oppgaver var størst på nivå 1 og minst på nivå 3. Det så derfor ut til at oppgaveformatet hadde større betydning jo lavere kompetansenivå elevene var på.

For alle gruppene var det høyere andel ubesvarte oppgaver på de åpne enn på de lukkede oppgavene, og andel ubesvarte var størst på det laveste og avtok opp til det høyeste nivået. Resultatene viser ingen sammenheng mellom p-verdi og andel ubesvarte. På nivå 1 hadde oppgaver med p-verdi 0,00, ubesvart andel på 7 prosent i en oppgave og 46 prosent i en annen. En oppgave med p-verdi 0,29 hadde 37 prosent ubesvart, og i en oppgave som var løst av kun 5 prosent av elevene, var andel ubesvart bare 3 prosent (se tabell 4.9 og 4.12).

4.7. Resultater til forskningsspørsmål 3

Er det kjønnsforskjeller når det gjelder å mestre åpne og lukkede oppgaver?

4.7.1 Resultatene til jentene og guttene på åpne og lukkede oppgaver

Av de 140 elevene som deltok på Test 2 prøve 1 høsten 2007 og våren 2008, var 66 jenter og 74 gutter. Fordelingen var 39 jenter og 37 gutter fra skole 1 og 27 jenter og 37 gutter fra skole 2. På prøve 2 deltok 64 jenter og 71 gutter og det var 36 jenter og 38 gutter fra skole 1 og 28 jenter og 33 gutter fra skole 2 (se tabell 3.4).

Tabell 4.22

Tekniske data for prøve 1 og prøve 2 for jenter og gutter i utvalget på 275 elever

Test 2	Antall elever	Ant. oppg.	Gj.sn. poeng	P-verdi	Std. avvik	Std.feil til gj.sn.	Inter item korr.	Effekt
Prøve 1 jenter	66	20	10,1 p	0,51	3,3 p	0,41 p	0,10	0,12
Prøve 1 gutter	74	20	10,5 p	0,52	3,2 p	0,38 p	0,10	
Prøve 2 jenter	64	20	9,3 p	0,46	3,2 p	0,40 p	0,08	0,28
Prøve 2 gutter	71	20	10,3 p	0,52	3,8 p	0,45 p	0,15	

Tabell 4.22 inneholder tekniske data fra resultatene til jenter og gutter for prøvene i Test 2. Resultatene viser at jentene i gjennomsnitt presterte 6 prosentpoeng dårligere enn guttene på prøve 2, og at guttene hadde større spredning i resultatene på prøve 2 enn hva noen av de andre gruppene hadde på noen av prøvene. Effekten av å være gutt eller jente var lav i begge prøvene (0,12 og 0,28), men kjønn hadde større betydning i prøve 2 enn i prøve 1. T-tester viste ingen signifikant forskjell på gjennomsnittlig resultat for jenter og gutter, verken for prøve 1 eller prøve 2.

Tabell 4.23

Løsningsprosent for åpne (CR) og lukkede oppgaver (SR) for jenter og gutter på Test 2

Test 2	**Antall**	**Ant. oppg.**	**P-verdi SR**	**P-verdi CR**	**Effekt j/g SR**	**Effekt j/g CR**
Prøve 1 jenter	66	10	0,62	0,39	0,03	0,13
Prøve 1 gutter	74	10	0,63	0,42		
Prøve 2 jenter	64	10	0,52	0,41	0,33	0,21
Prøve 2 gutter	71	10	0,58	0,45		

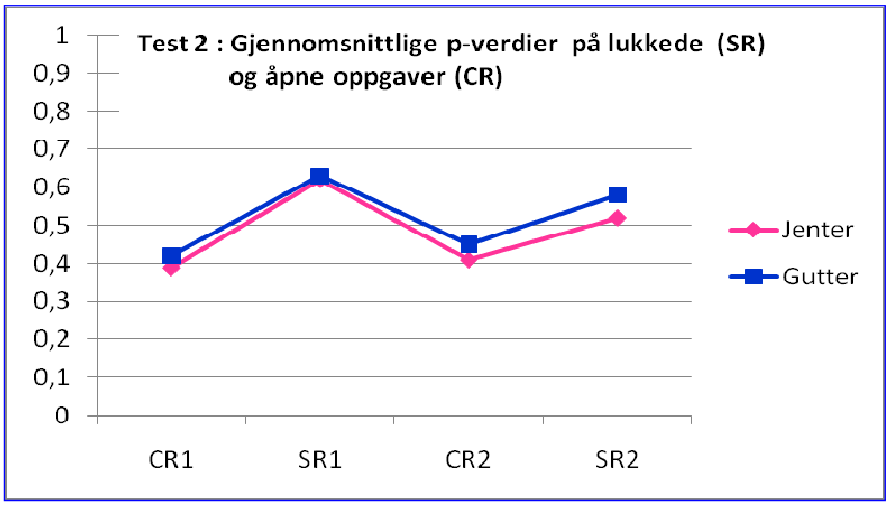
I prøve 1 løste jentene i gjennomsnitt 62 prosent av de lukkede oppgavene og 39 prosent av de åpne. Tilsvarende tall for guttene var 63 prosent og 42 prosent, og t-tester viste ingen signifikant forskjell på resultatene til guttene og jentene. Dette bekreftes av lav effektstørrelse på 0,03 og 0,13 (se tabell 4.23).

I prøve 2 var gjennomsnittlig løsningsprosent for jentene 52 prosent i de lukkede oppgavene og 41 prosent i de åpne, mens guttene løste 58 prosent av de lukkede og 45 prosent av de åpne oppgavene (se tabell 4.23). På de lukkede oppgavene gjorde guttene det 6 prosentpoeng bedre enn jentene, og en effektstørrelse på 0,33 signaliserer at kjønn hadde middels betydning for resultatet. T-test på prøve 2 i forhold til kjønn viser at resultatet til guttene på de lukkede oppgavene i prøve 2 var på grensen til å være signifikant bedre enn resultatene til jentene innenfor en sannsynlighet på 95 prosent. Resultatet på de åpne oppgavene var ikke signifikant forskjellig.

Guttene fikk i gjennomsnitt høyere poengsum enn jentene både på de åpne og de lukkede oppgavene, men forskjellen var ikke signifikant. Forskjellen mellom prestasjonene til gutter og jenter var størst i de lukkede oppgavene i prøve 2 (SR2).

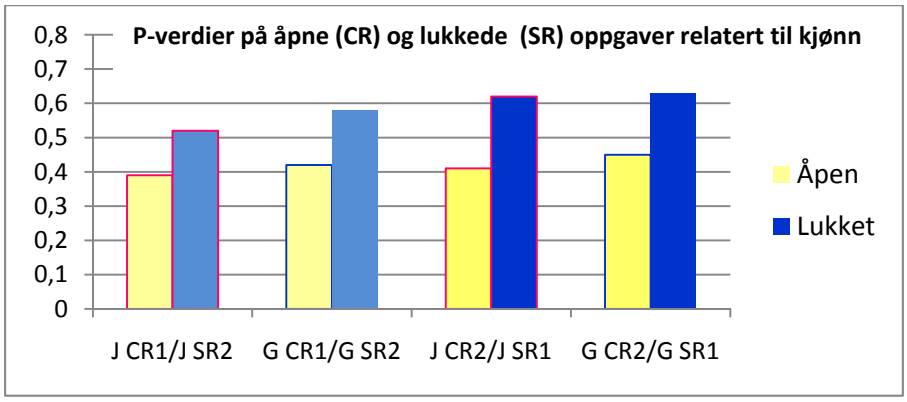
Figur 4.22 viser en sammenligning av gjennomsnittlig p-verdi for jenter og gutter på gruppene med åpne og lukkede oppgaver.

Figur 4.22 *Gjennomsnittlige p-verdier for jenter og gutter på åpne (CR1 og CR2) og lukkede oppgaver (SR1 og SR2) i Test 2*



I figur 4.23 vises to sammenligninger. Den ene er hvert av kjønnenes prestasjoner da oppgavene var åpne sammenlignet med resultatene da de samme oppgavene var i formatet lukket. Gul søyle representerer åpne oppgaver og blå søyle lukkede oppgaver. JCR1 betyr "Jenter og åpne oppgaver i prøve 1" og JSR2 "jenter og lukkede oppgaver i prøve 2" (oppgaver med samme innhold). Tilsvarende for de andre søylene når G betyr gutter. Den andre sammenligningen er jentenes resultater sammenlignet med resultatet til guttene på de samme oppgavene, dvs. JCR1/JSR2 med GCR1/GSR2 og JCR2/JSR1 med GCR2/GSR1. Analysen viser at både jentene og guttene hadde høyere gjennomsnittlig løsningsprosent i de lukkede enn i de åpne oppgavene, uansett om oppgavene tilhørte prøve 1 eller prøve 2, eller om vi sammenlignet tilsvarende oppgaver på tvers av prøvene. I tillegg viser resultatet at guttene hadde høyere gjennomsnittlig løsningsprosent enn jentene både i de åpne og de lukkede oppgavene, men resultatene var ikke signifikante (se figur 4.23).

Figur 4.23 *Gjennomsnittlige p-verdier for jenter og gutter på de samme oppgavene i to formater*



4.7.2 Kjønnsforskjeller på oppgavenivå

I 20 oppgaver var prosent gutter som hadde løst oppgavene, mer enn to prosentpoeng høyere enn prosent jenter. Det var åtte fysikk-, sju kjemi- og fem biologioppgaver, og 11 av oppgavene var lukkede og ni var åpne. Resultater fra t-tester viser at guttene var signifikant bedre enn jentene i fire av disse oppgavene. Det var to fysikkoppgaver i prøve 1 (fA40, t-test: $p = 0,04 < 0,05$ og åB41a, t-test: $p = 0,03 < 0,05$) og to kjemioppgaver i prøve 2 (fA34, t-test: $p = 0,02 < 0,05$ og fB30, t-test: $p = 0,03 < 0,05$). Tre av disse oppgavene var lukkede oppgaver og en var åpen (se tabell 4.24 og figur 4.24 og 4.25).

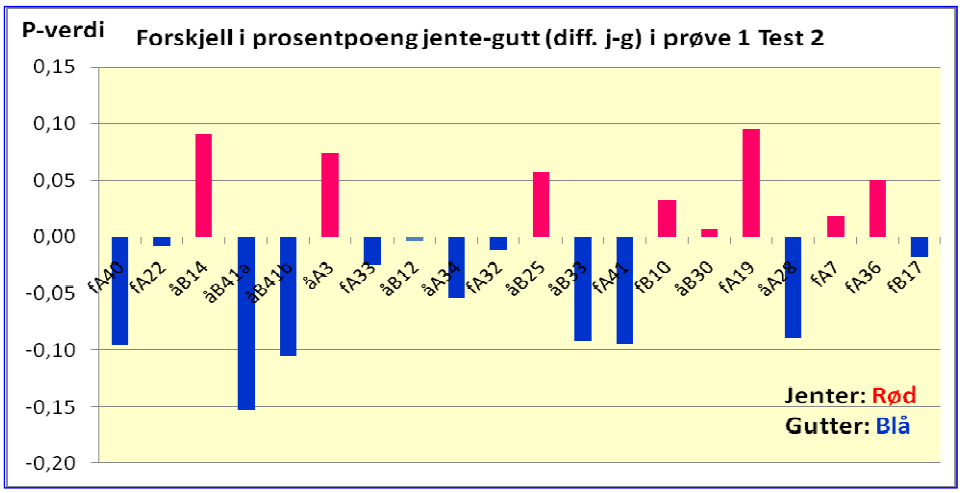
Tabell 4.24

Fordeling av lukkede oppgaver (SR) og åpne oppgaver (CR) i prøve 1 og prøve 2 for Test 2. P-verdier for elever som deltok både høst 2007 og vår 2008, innhold og fag (fysikk (F), kjemi (K), biologi (B)) og differanse jente – gutt (diff j-g). Lukkede oppgaver er markert med gult. 140 elever på prøve 1 og 135 elever på prøve 2. Markerte verdier viser oppgaver med signifikant forskjell for jenter og gutter. Negative verdier for oppgaver i guttefavør

Oppgave	Innhold og fag	Test 2 Prøve 1			Test 2 Prøve 2		
		Format	Diff j-g	p-verdi	p-verdi	Diff j-g	Format
A40 / B27	**Torden og lyn (F)**	SR	**-0,10**	0,91	0,57	-0,16	CR
A22	Månen (F)	SR	-0,01	0,93	0,79	-0,07	CR
B14	Organer (B)	CR	0,09	0,30	0,49	0,02	SR
B41a	**Tannhjul (F)**	CR	**-0,15**	0,79	0,68	-0,05	SR
B41b	Tannhjul (F)	CR	-0,11	0,71	0,70	-0,12	SR
A3	Fordamping (K)	CR	0,08	0,66	0,79	-0,10	SR
A33	Gass (K)	SR	-0,02	0,54	0,26	-0,02	CR
B12 / A15	Vekselvarm (B)	CR	0,00	0,21	0,60	-0,07	SR
A34	**Atomer (K)**	CR	-0,05	0,53	0,59	**-0,21**	SR
A32	Tabell grader (F)	SR	-0,01	0,76	0,60	-0,16	CR
B25	Korrolering (B)	CR	0,06	0,12	0,47	-0,04	SR
B33	Kjemisk reaksjon (K)	CR	-0,09	0,32	0,31	-0,03	SR
A41	Temp. og atomer (K)	SR	-0,09	0,69	0,53	-0,02	CR
B10	Lunger og gjeller (B)	SR	0,03	0,66	0,79	-0,04	CR
B30	**Gasser (K)**	CR	0,01	0,06	0,09	**-0,11**	SR
A19	Kropp (B)	SR	0,10	0,33	0,26	0,13	CR
A28	Brensel (B)	CR	-0,09	0,34	0,81	0,04	SR
A7	Blomsterdeler (B)	SR	0,02	0,04	0,05	0,02	CR
A36	Kjemisk reaksjon (K)	SR	0,05	0,43	0,26	-0,02	CR
B17	Kropp (B)	SR	-0,02	0,96	0,20	-0,08	CR
	Gjennomsnitt		-0,02	0,51	0,49	-0,05	

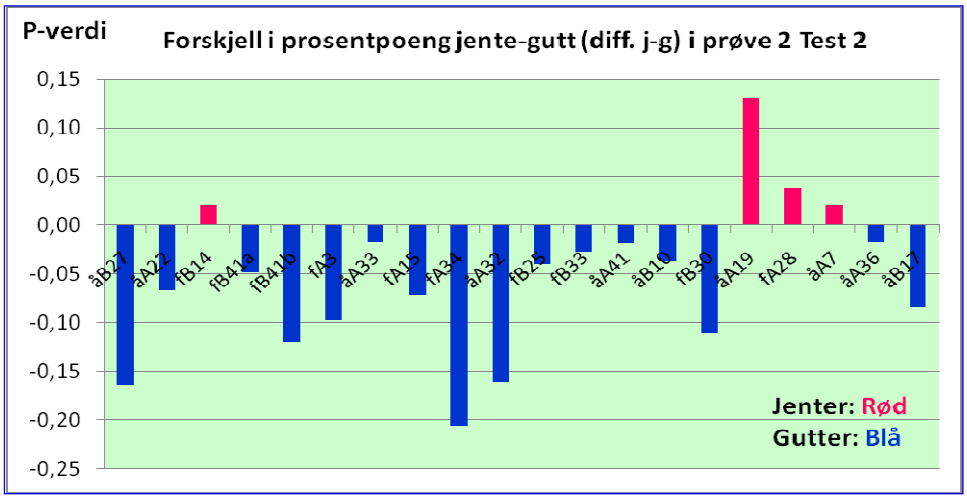
Jentene var mer enn to prosentpoeng bedre enn guttene i seks biologi- og to kjemioppgaver. Dette var fire åpne og fire lukkede oppgaver. T-tester viser at forskjellene ikke var signifikante (se figur 4.24 og 4.25).

Figur 4.24 *Forskjeller i p-verdier for jenter og gutter på prøve 1. Røde søyler viser forskjell i jentefavør og blå viser forskjell i guttefavør*



På prøve 1 hadde jentene og guttene tilnærmet samme gjennomsnittlige p-verdi, men på prøve 2 hadde guttene i gjennomsnitt 6 prosentpoeng bedre resultat enn jentene. I figur 4.24 ser vi at det er god balanse mellom antall oppgaver som jentene eller guttene har gjort det best på, mens figur 4.25 viser stor overvekt av blå søyler som er oppgaver hvor guttene gjorde det bedre enn jentene.

Figur 4.25 *Forskjeller i p-verdier for jenter og gutter på prøve 2. Røde søyler viser forskjell i jentefavør og blå viser forskjell i guttefavør*



På oppgave fA40 (se figur 3.3) gjorde guttene det signifikant bedre enn jentene. Dette er en tradisjonell flervalgsoppgave hvor elevene skulle velge blant fire svaralternativer. Oppgaven handler om lyn og torden og hadde p-verdi 0,96 for guttene og 0,86 for jentene. Som åpen

oppgave var kjønnsforskjellen i guttenes favør 16 prosentpoeng (se figur 4.25), og ikke signifikant forskjellig, men helt på grensen til å være det (t-test: p = 0,056 > 0,05). ÅB41var den andre oppgaven i prøve 1 som skilte signifikant mellom guttene og jentene. Oppgaven er åpen og handler om tannhjul som griper inn i hverandre (se figur 3.5). Guttene hadde p-verdi 0,87 og jentene 0,71 på denne oppgaven. I lukket format hadde både jentene og guttene lavere p-verdi enn da oppgaven var åpen, og forskjellen var fem prosentpoeng i guttenes favør.

I prøve 2 var det i oppgavene fA34 (se figur 4.10) og fB30 (se figur 4.12) at guttene gjorde det signifikant bedre enn jentene. Begge er tradisjonelle flervalgsoppgaver med fire svaralternativer. Den største forskjellen i prosentpoeng var i oppgave fA34.

Oppgaven tester forståelse av atombegrepet. P-verdier for lukket oppgave var 0,69 for guttene og 0,48 for jentene Som åpen oppgave var p-verdiene 0,55 for guttene og 0,50 for jentene. Det betyr at jentene som løste oppgaven som åpen, i gjennomsnitt fikk to prosentpoeng bedre resultat enn de jentene som løste oppgaven som lukket, mens guttene fikk 14 prosentpoeng dårligere resultat da oppgaven var åpen enn da den var lukket.

Oppgave fB30 inneholder i lukket format distraktoren ”oksygen” som er lett å velge når det er snakk om hvilken gass det er mest av i luft. Da oppgaven var åpen var det ett prosentpoeng flere jenter enn gutter som fikk rett på oppgaven, men som lukket oppgave endret forskjellen seg til å bli 11 prosentpoeng i guttenes favør. Oppgaven er omtalt flere steder i denne masteroppgaven på grunn av spesielle resultater.

Andre oppgaver med store kjønnsforskjeller var en biologioppgave som tester begrepet ”bukhule” (B14). Jentene gjorde det best uansett oppgaveformat, og forskjellen var 9 prosentpoeng da oppgaven var åpen og 2 prosentpoeng som lukket (se tabell 4.24). Både guttene og jentene gjorde det best da oppgaven var lukket. A19 er en biologioppgave hvor resultatene viser jentefavør med 10 prosentpoeng i lukket oppgave og 13 prosentpoeng i åpen. Oppgaven hadde lave p-verdier og tester begrepet menstruasjon.

B33 og A41 er to kjemioppgaver hvor guttene gjorde det bedre enn jentene i begge formatene. B33 handler om begrepet kjemisk reaksjon, og A41 om sammenhengen mellom temperatur og atombevegelse. Guttene fikk i gjennomsnitt 9 prosentpoeng bedre resultat enn jentene da

oppgaven var åpen, og 2 prosentpoeng bedre resultat da oppgaven var lukket. Jentene hadde høyest p-verdi i lukket oppgave, og guttene var best i åpen.

4.7.3 Oppgaver som endret kjønnsfordel etter format

Tabell 4.25

Oppgaver som jentene gjorde det best på i det ene formatet, og guttene best på i det andre

Oppgave	Innhold og fag	Test 2 prøve 1			Test 2 prøve 2		
		p-verdi jente	p-verdi gutt	Diff j-g	Diff j-g	p-verdi jente	p-verdi gutt
A3	Fordamping (K)	0,70	0,62	**0,08**	**-0,10**	0,73	0,83
B25	Korrolering (B)	0,15	0,10	**0,06**	**-0,04**	0,45	0,49
B10	Lunger og gjeller (B)	0,68	0,65	**0,03**	**-0,04**	0,77	0,80
B30	Gasser (K)	0,06	0,05	**0,01**	**-0,11**	0,03	0,14
A28	Brensel (B)	0,29	0,38	**-0,09**	**0,04**	0,83	0,79
A36	Kjemisk reaksjon (K)	0,46	0,41	**0,05**	**-0,02**	0,25	0,27

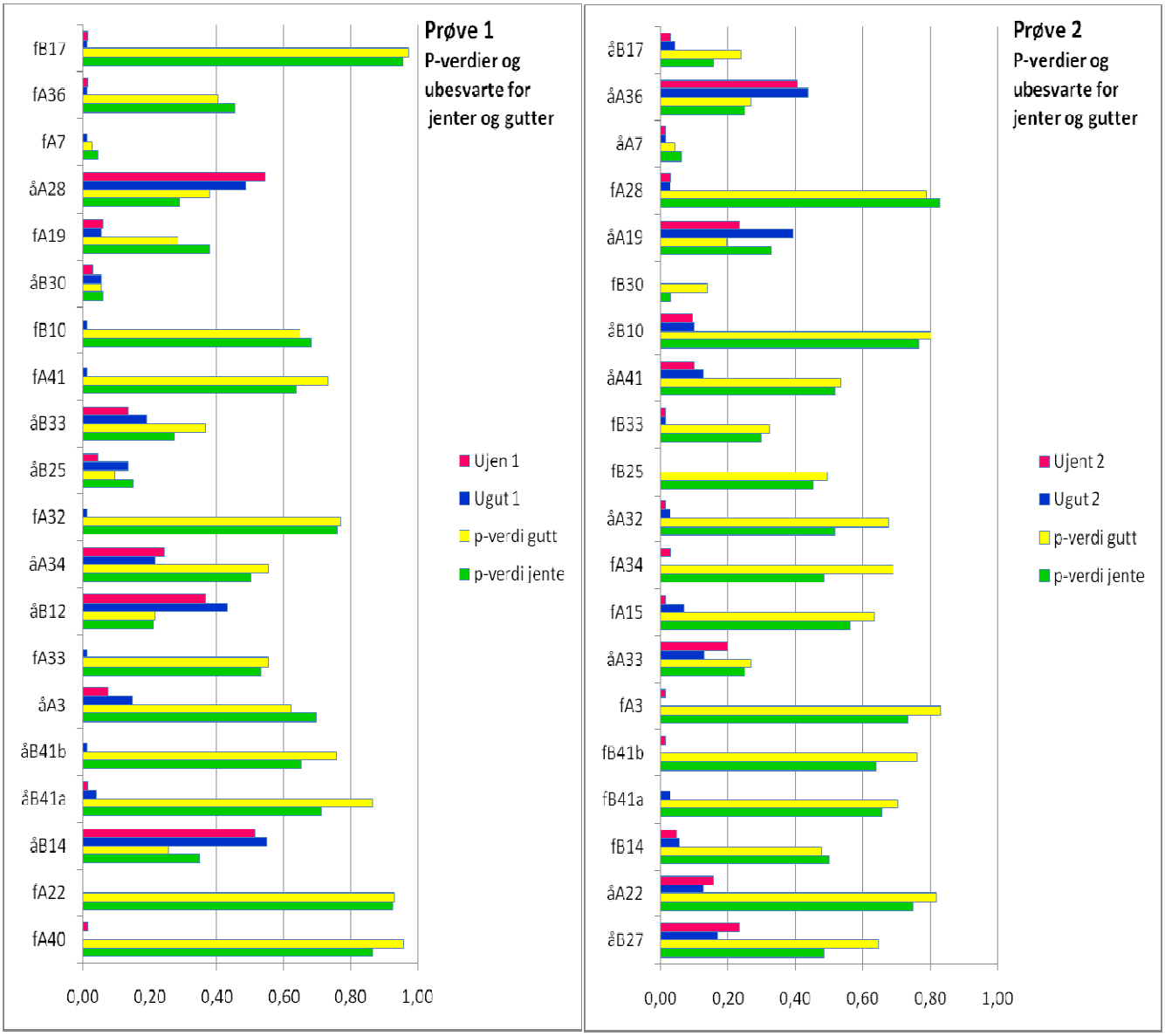
Rader markert med gult i tabell 4.25, gjelder oppgaver i lukket format. Hvite rader gjelder åpne oppgaver. Rød skrift viser p-verdier for differanse i jentefavør, og blå skrift for differanse i guttefavør. Til sammen for prøve 1 og prøve 2 var det tre åpne og tre lukkede oppgaver hvor jentene fikk bedre resultat enn guttene, og tre åpne og tre lukkede oppgaver hvor guttene skåret bedre enn jentene. Jentene hadde to oppgaver med høyere p-verdi i åpen enn lukket utgave, oppgave B10 og B30. Guttene fikk også høyere p-verdi da B10 var åpen enn lukket (se tabell 4.25). Innholdet i tre av oppgavene var kjemirelatert og de tre andre hadde biologirelatert tema.

4.7.4 Kjønnsforskjeller, ubesvarte oppgaver og oppgaveformat

Figur 4.26 viser at prosent ubesvart for både gutter og jenter var høyere da oppgaven var åpen enn da den var lukket. I prøve 1 var det sju åpne oppgaver som hadde mer enn 10 prosent ubesvarte, og i prøve 2 gjaldt dette 6 åpne oppgaver.

Gjennomsnittlig p-verdi for ubesvarte oppgaver i prøve 1 var 0,10 for jentene og 0,12 for guttene, og i prøve 2 var p-verdien 0,08 for jentene og 0,09 for guttene. Forskjellen mellom jenter og gutter for ikke å svare på oppgaver i prøvene, var ikke statistisk signifikant.

Figur 4.26 *P-verdier og ubesvarte oppgaver for jenter og gutter på prøve 1 og prøve 2 i Test 2. Ujen og Ugut betyr ubesvarte for jenter og gutter. Oppgave fB17 er lukket oppgave i prøve 1 som tilsvarer åpen oppgave åB17 i prøve 2, osv for alle oppgavene*



Resultatene viser også at det ikke var forskjell på jente- og guttegruppene verken når det gjelder å la åpne oppgaver eller å la lukkede oppgaver stå ubesvart (se tabell 4.26). Gjennomsnittlig p-verdi for ubesvarte lukkede oppgaver (SR1 og SR2) var ubetydelig for begge kjønn, og ubesvart for åpne oppgaver (CR1 og CR2) var relativt høyt for begge. Resultatene viser heller ingen sammenheng mellom lav p-verdi på oppgavene og høy andel ubesvarte oppgaver (se figur 4.26).

Tabell 4.26

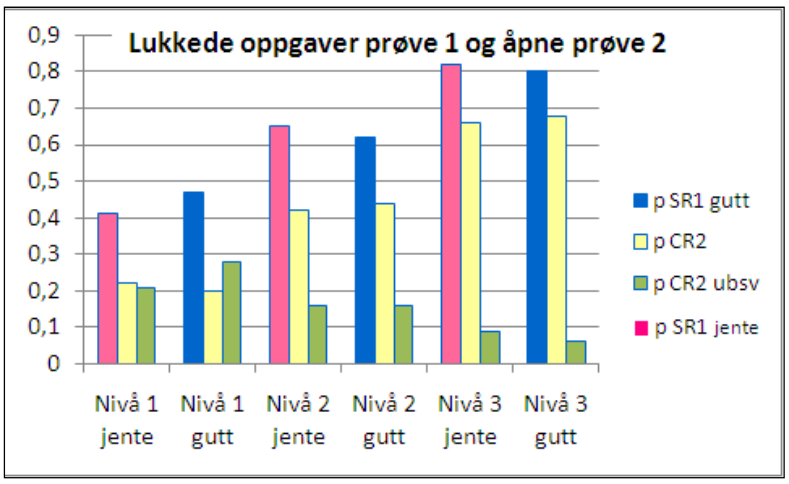
P-verdier for ubesvarte åpne (CR1 og CR2) og lukkede oppgaver (SR1 og SR2) for jenter og gutter. CR1 og SR2 er samme oppgaver i ulikt format og tilsvarende for CR2 og SR1

	Gjennomsnittlig p-verdi ubesvart			
	CR1	**SR2**	**CR2**	**SR1**
Jenter	0,20	0,02	0,15	0,01
Gutter	0,23	0,02	0,16	0,02

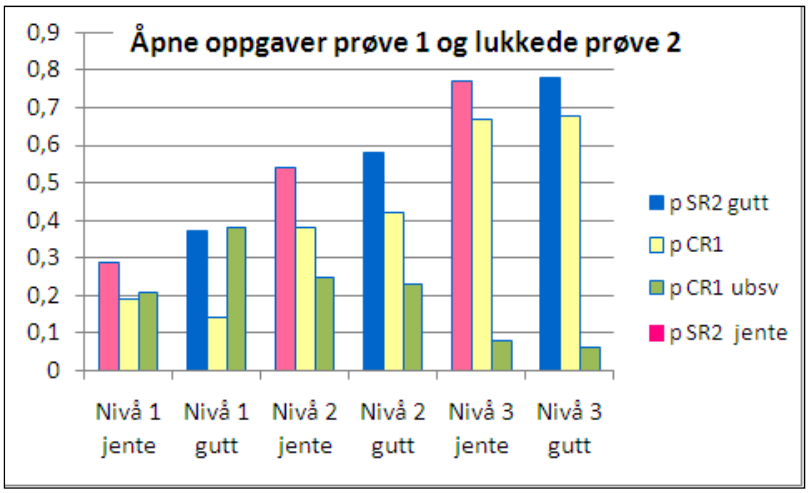
4.7.5 Kjønnsforskjeller på nivå for åpne og lukkede oppgaver

Ved hjelp av figur 4.27 og figur 4.28 kan vi gjøre to sammenligninger på nivå. Den ene er hva hver jentegruppe og hver guttegruppe presterte da oppgavene med samme innhold var i åpent eller lukket format. Den andre er å sammenligne resultatene til jentene med resultatene til guttene på hvert nivå. I tillegg viser grønn søyle p-verdi for ubesvarte åpne oppgaver. Ubesvarte lukkede oppgaver var svært lavt (se tabell 4.27), og er derfor ikke tatt med i sammenligningen. På alle nivåer og for alle grupper av elever var gjennomsnittlige p-verdier for de lukkede oppgavene høyere enn da oppgavene var åpne. Effektstørrelsene viste stor effekt for oppgaveformat og var fra 0,8 til 3,3 standardavvik.

Figur 4.27 *P-verdier på nivå for lukkede oppgaver i prøve 1 (SR1) og åpne oppgaver i prøve 2 (CR2), for jenter og gutter. P-verdier for ubesvarte åpne oppgaver i prøve 2. 140 elever på prøve 1 og 135 elever på prøve 2*



Figur 4.28 *P-verdier på nivå for åpne oppgaver i prøve 1 (CR1) og lukkede oppgaver i prøve 2 (SR2), for jenter og gutter. P-verdier for ubesvarte åpne oppgaver i prøve 1. 140 elever på prøve 1 og 135 elever på prøve 2*



Størst var effektstørrelsen for guttene på nivå 1, hvor effekten var 3,3 for lukkede oppgaver i prøve 1 sammenlignet med åpne i prøve 2, og 2,6 for åpne oppgaver i prøve 1 og lukkede i prøve 2. Se figur 4.27 og 4.28 som viser størst differanse i blå og gul søyle for "Nivå 1 gutt". Gutter på nivå 1 hadde også størst andel ubesvarte åpne oppgaver. I prøve 1 hadde de i gjennomsnitt høyere andel ubesvarte åpne oppgaver enn gjennomsnittlig p-verdi til de samme oppgavene i lukket format (se figur 4.28).

Jentene på nivå 2 og 3 i prøve 1, hadde henholdsvis 3 og 2 prosentpoeng høyere gjennomsnitt enn guttene på flervalgsoppgavene (se tabell 4.28 og figur 4.27). Effektstørrelsen var 0,29 i begge tilfellene og lav. Jentene på nivå 1 hadde høyere gjennomsnittlig p-verdi enn guttene på de åpne oppgavene i begge prøvene. Effektstørrelsen var middels på prøve 1 (e = 0,6) og liten på prøve 2 (e = 0,2).

Tabell 4.27

Effektstørrelser for resultater på åpne (CR) og lukkede oppgaver (SR), på nivå og for kjønn. 140 elever på prøve 1 og 135 elever på prøve 2

Effekt av oppgaveformat	**Nivå 1**	**Nivå 2**	**Nivå 3**
Stor effekt	Jenter for SR og CR av samme oppgaver Gutter for SR og CR av samme oppgaver	Jenter for SR og CR av samme oppgaver Gutter for SR og CR av samme oppgaver	Jenter for SR og CR av samme oppgaver Gutter for SR og CR av samme oppgaver
Middels effekt	Jente/gutt for SR i begge prøvene og CR i prøve 1	Jente/gutt for SR i begge prøvene	
Lav effekt	Jente/gutt for CR i prøve 2	Jente/gutt for CR i begge prøvene	Jente/gutt for CR og SR i begge prøvene

Oppgaveformatet hadde minst betydning for resultatene til jentene og guttene på nivå 3. Effekten av oppgaveformat på kjønn var også lav på nivå 1 for de åpne oppgavene i prøve 2, og for de åpne oppgavene i begge settene på nivå 2. De åpne oppgavene i prøve 2 (CR2) hadde lavest løsningsprosent, og her var effekten like lav på alle nivåer ved sammenligning av jenter og gutter. Effekten av kjønn var større i de lukkede enn i de åpne oppgavene (se tabell 4.27).

Med inndeling av elevene i jenter og gutter på tre nivåer, ble det få elever i hver gruppe. Fargekodene i tabell 4.28 viser hvilke oppgaver som hadde samme innhold, men ulikt format. Av tabellen kan vi se at gutter på nivå 1 lot være å svare på åpne oppgaver i større grad enn jentene på samme nivå. Dette ser vi også av figur 4.27 og 4.28.

Tabell 4.28

Inndeling i nivågrupper for jenter og gutter som deltok på prøve 1 eller prøve 2. P-verdier for åpne (CR) og lukkede oppgaver (SR) og ubesvarte for åpne og lukkede oppgaver

Prøve 1 JENTER	**%**	**Antall**	**Poeng**	**p-verdi SR1**	**p-verdi CR1**	**p-verdi ubsvart CR1**	**p-verdi ubsvart SR1**
Nivå 1	26	17	$3 \leq X \leq 7$	0,41	0,19	0,21	0,01
Nivå 2	55	36	$8 \leq X \leq 13$	0,65	0,38	0,25	0,01
Nivå 3	20	13	$14 \leq X \leq 18$	0,82	0,67	0,08	0,00
Sum		**66**					
Prøve 1 GUTTER				**p-verdi SR1**	**p-verdi CR1**		
Nivå 1	22	16	$3 \leq X \leq 7$	0,47	0,14	0,38	0,03
Nivå 2	55	41	$8 \leq X \leq 13$	0,62	0,42	0,23	0,01
Nivå 3	23	17	$14 \leq X \leq 18$	0,80	0,68	0,06	0,02
Sum		**74**					
Prøve 2 JENTER				**p-verdi SR2**	**p-verdi CR2**	**p-verdi ubsvart CR2**	**p-verdi ubsvart SR2**
Nivå 1	23	15	$1 \leq X \leq 7$	0,29	0,22	0,21	0,05
Nivå 2	61	39	$8 \leq X \leq 12$	0,54	0,42	0,16	0,01
Nivå 3	16	10	$13 \leq X \leq 18$	0,77	0,66	0,09	0,00
Sum		**64**					
Prøve 2 GUTTER				**p-verdi SR2**	**p-verdi CR2**		
Nivå 1	28	20	$1 \leq X \leq 7$	0,37	0,20	0,28	0,03
Nivå 2	41	29	$8 \leq X \leq 12$	0,58	0,44	0,16	0,03
Nivå 3	31	22	$13 \leq X \leq 18$	0,78	0,68	0,06	0,00
Sum		**71**					

4.7.6 Oppsummering forskningsspørsmål 3

Guttene presterte i gjennomsnitt 6 prosentpoeng bedre på prøve 2 enn jentene, mens resultatet på prøve 1 var 1 prosentpoeng i guttenes favør. T-tester viser at forskjellen på resultatene til jentene og guttene ikke var signifikant i noen av prøvene. Undersøkelse av alle grupper og på alle nivåer viser at gjennomsnittlig løsningsprosent for oppgavene i lukket format alltid var høyere enn da oppgavene var åpne. Det betyr at jentene og guttene i gruppe G1 og jentene og guttene i gruppe G2 hadde høyere gjennomsnittlig p-verdi på de lukkede enn på de åpne oppgavene generelt sett. Videre betyr det at jentene og guttene i G1 som løste de lukkede oppgavene i prøve 1, gjorde det bedre enn jentene og guttene i G2 som løste de samme oppgavene i åpent format og vice-versa.

Resultater på oppgavenivå viser at guttene gjorde det signifikant bedre enn jentene i fire av 40 oppgaver. Dette var en lukket og en åpen fysikkoppgave og to lukkede kjemioppgaver. Det var ingen oppgaver hvor jentene gjorde det bedre enn guttene. Seks oppgaver endret kjønnsfordel da de endret format. Av disse var det tre lukkede oppgaver som gikk til jentefordel, og tre lukkede som gikk til guttefordel da de ble åpne.

Andel ubesvarte var høyest i de åpne oppgavene, og det var ingen signifikant forskjell på gjennomsnitt ubesvarte oppgaver for jenter og gutter. Nivå 1 for gutter skilte seg ut med størst effektstørrelse på oppgaveformat. I begge guttegruppene på nivå 1 var det betydelig forskjell i gjennomsnittlige p-verdier for de samme oppgavene i lukket og åpent format. Det var også guttene på nivå 1 som hadde høyest andel ubesvarte åpne oppgaver i gjennomsnitt. Ingen resultater gir imidlertid signaler om at oppgaveformatet alene er avgjørende for resultatene til verken gutter eller jenter.

5. Kapittel Diskusjon

5.1. Kritisk blikk på metode

5.1.1 Positivt blikk på utvalg og metode

Utvalget var ikke representativt per definisjon, men omfattet alle elevene på 8. trinn ved to store ungdomsskoler og var i utgangspunktet på 333 elever. Ungdomsskolene hadde tilnærmet like tradisjoner når det gjelder undervisningspraksis, og var sammenlignbare både når det gjelder skolelokaler og eksamensresultater for elevene. Å velge to sammenlignbare ungdomsskoler, var et bevisst valg for å få størst mulig gruppe av mest mulig sammenlignbare elever, og dermed stort nok antall til å kunne dele elevene inn i tre nivåer. Ved å velge de skolene jeg gjorde, unngikk jeg i særlig stor grad å måtte ta hensyn til minoritetsspråklige elever hvor kompetanse i norsk kunne utgjøre en ekstra feilkilde.

Fordi det i utgangspunktet var mulig å delta for alle elevene på 8. trinn ved skolene, kunne jeg forvente å ha en heterogen gruppe av elever og representanter for alle faglige nivåer. Tilfeldig frafall gjorde at det ble 275 elever som deltok i begge undersøkelsene. At frafallet var tilfeldig, bare enkeltelever fra forskjellige klasser og fordelt på skolene, var en fordel med anke på å unngå flest mulig forstyrrende faktorer som kunne føre til ekstra skeivheter i utvalget. Metoden for å velge ut elever til gruppene i Test 2, vil jeg betegne som et godt forsøk på å etablere to sammenlignbare grupper. Ved at elevene ble plassert i par med elever fra egen klasse eller på egen skole, kunne jeg i noe grad kontrollere at elevene i et par ikke ble utsatt for totalt forskjellig påvirkning i de fire månedene det gikk fra Test 1 til Test 2. I begge testgruppene for Test 2, var det elever fra begge skolene, og elever av begge kjønn på alle nivåer.

Test 1 og Test 2 ble komponert på samme tid og ut fra samme kriterier. Oppgavene var pilotert og kvalitetssikret og tilpasset faglig nivå for elever med både høy og lav kompetanse. Resultater fra en utprøving styrte utvelgelsen av oppgaver for å gjøre Test 1 og Test 2 mest mulig sammenlignbar. Dette bidro til at elevene i Test 2 ble prøvd i samme fagstoff som i Test 1, noe som var svært viktig med tanke på at elevene som deltok i Test 2, var plassert i gruppe på grunnlag av resultatene på Test 1. Resultatene fra en liten holdningsundersøkelse ble i tillegg brukt til justering da elevene ble fordelt på de to gruppene.

Kriterier ble fulgt ved gjennomføring og evaluering, og jeg var fysisk til stede i alle klassene ved gjennomføringen av alle prøvene. Resultatene viser relativt høy reliabilitet i form av Cronbachs alfa. Det var relativ høy sensorreliabilitet ved at 50 prosent av oppgavene var lukkede og ved at jeg rettet alle besvarelsene selv ut fra en retteguide. Resultatene viser at effekten av å være elev på prøve 1 eller prøve 2 i Test 2 var lav, og det samme viser effekten av å være jente i prøve 1 eller prøve 2, eller gutt i prøve 1 eller prøve 2.

5.1.2 Negativt blikk på utvalg og metode

Utvalget var ikke representativt og ikke tilfeldig trukket. Test 1 og Test 2 var per definisjon ikke parallelle prøver. For å få alle de 20 oppgavene i Test 2 i både åpent og lukket format, måtte jeg foreta noen justeringer. Disse oppgavene ble ikke prøvd ut på nytt før Test 2 ble gjennomført, og dette kan svekke både reliabiliteten og validiteten til prøvene. I Test 2 var fire av de 20 oppgaveparene så ulike at vi ikke kan si at de var samme oppgave i åpent og lukket format. Lang tekst kan gi problemer for lesesvake elever, og ingen elever fikk i særlig grad spesiell tilrettelegging ved gjennomføring av prøvene. Utvelging til gruppene G1 og G2 skjedde på grunnlag av bare en prøve. Resultatene på åtte holdningsspørsmål med Cronbachs alfa 0,63, ble brukt til justering av gruppene. 9 av 60 oppgaver hadde diskriminering (point-biserial) lavere enn 0,3, og av disse hadde 3 oppgaver lavere enn 0,2. Inndeling i nivågrupper for jenter og gutter gjorde at noen av gruppene som ble analysert, hadde få elever.

5.1.3 Konklusjon angående utvalg og metode

En sammenligning av positive og negative effekter viser at det er noe usikkerhet forbundet både med utvalg og metode. Begge deler er imidlertid godt begrunnet, og effektanalyser viser at prøve 1 og prøve 2 har fungert psykometrisk sammenlignbart i forhold til gruppene G1 og G2, og i forhold til kjønn. Selv om ikke alle oppgavene diskriminerte tilfredsstillende, taler det til undersøkelsenes fordel at jeg har kontroll på disse oppgavene gjennom psykometriske resultater. Jeg kan ikke generalisere, men jeg mener det er mulig å trekke noen konklusjoner for dette utvalget av elever, som kan være representativt for lignende skoler.

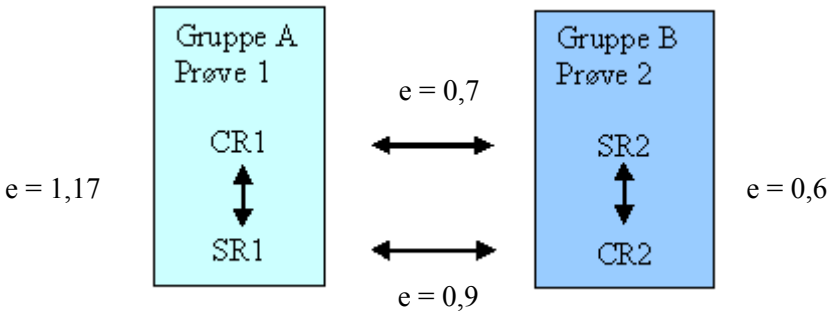
5.2. Diskusjon av resultater til forskningsspørsmål 1

Hvilket samsvar er det mellom resultatet en gruppe elever oppnår på naturfagoppgaver i åpent format (CR, - Constructed Response Items), og resultatet som oppnås av en sammenlignbar elevgruppe når oppgaver med samme opplysninger og spørsmål er i formatet lukket (SR, - Selected Response Items)?

5.2.1 Sammenligning av gjennomsnittsresultater på 10 åpne og 10 lukkede oppgaver

Jeg vil først kommentere resultatene til gruppene av elever som løste de samme oppgavene i åpent eller lukket format. For begge gruppene viser resultatet at gjennomsnittlig p-verdi for de lukkede oppgavene var høyere enn gjennomsnittlig p-verdi for de åpne oppgavene. Effektstørrelsen for oppgaveformat var middels (e = 0,7) i den ene gruppen og stor (e = 0,9) i den andre (se figur 5.1, figur 4.8 og tabell 4.16). Dette har sammenheng med at det var større effektstørrelse på format i gruppe G1 (e = 1,17) enn i gruppe G2 (e = 0,6).

Figur 5.1 *Effektstørrelser på oppgaveformat for gruppe G1, gruppe G2 og de åpne (CR1) i prøve 1 og lukkede (SR2) i prøve 2, og de åpne (CR2) i prøve 2 og lukkede (SR1) i prøve 1*



At elevene gjorde det bedre på lukkede enn åpne oppgaver, stemmer med norske elevers prestasjoner i TIMSS og PISA siden 2001 (Kjærnsli mfl. 2004, 2007). I TIMSS 1995 var tendensen at norske elever presterte bedre på de åpne enn på de lukkede oppgavene i naturfag (Lie mfl. 2001). Dette var helt motsatt av resultatene i de fleste landene i Europa, og har endret seg i Norge til tross for at norske lærere på 8. trinn i TIMSS 2006, sa at de sjelden brukte lukkede oppgaver i undervisningen (Bergem mfl. 2009). Elever som er kjent med strategier for å nyttiggjøre seg den hjelpen det kan være i å ha svaralternativer (Olsen mfl. 2001), kan oppnå høyere løsningsprosent på lukkede enn åpne oppgaver.

Forskning på åpne og lukkede oppgaver har imidlertid konkludert med at oppgavenes innhold ser ut til å spille større rolle enn oppgaveformatet (Beller mfl. 2000, Kjærnsli mfl. 2007, Lukhele mfl. 1994, Rodriguez 2003). I prøve 1 og prøve 2 var seks av de 20 oppgaveparene ikke signifikant forskjellig i åpent og lukket format. Det kan være interessant å undersøke om disse seks oppgavene har noen felles egenskaper.

5.2.2 Enkeltoppgaver som ikke var signifikant forskjellig i åpent og lukket format

Av de seks oppgavene som ikke var signifikant forskjellig i åpent og lukket format (se tabell 5.1), var det en fysikkoppgave og to kjemioppgaver som krever at eleven viser forståelse, og en kjemi- og to biologioppgaver som krever faktakunnskaper. Felles for disse seks oppgavene var først og fremst at elever som i utgangspunktet ikke visste hva de skulle svare, ikke fikk særlig hjelp av distraktorene. I realiteten var det bare fem av disse oppgaveparene som hadde relevans til problemstillingen. Oppgave B33 var for ulik i åpent og lukket format til å kunne kalles samme oppgave. Jeg har allikevel valgt å ta oppgavene med i drøftingen da de kan være et eksempel på hvordan både formulering, innhold og format påvirker et resultat.

Tabell 5.1

Oppgaver som ikke var signifikant forskjellig i åpent og lukket format

Oppgave	Lik stamme	Fakta	For-ståelse	Mulig mis-oppfatning	Kognitivt nivå	p-verdi CR	p-verdi SR	Nytte av distraktorer
A34	Ja		Ja	Ja	Middels	0,53	0,59	Nei
B30	Ja	Ja		Ja	Høyt	0,06	0,09	Nei
A19	Ja	Ja			Høyt	0,26	0,33	Nei
B41b	Nei		Ja		Middels	0,71	0,70	Nei
B33	Nei	Ja	Ja		Høyt	0,32	0,31	Nei
A7	Nei	Ja			Høyt	0,05	0,04	Nei

Oppgavene A34 (figur 4.10), B30 (figur 4.12) og A19 (figur 4.13) hadde lik formulering (lik stamme) som åpen og lukket oppgave (se tabell 5.1). Rodriguez (2003) viser også til undersøkelser hvor han konkluderte med at lik stamme kunne være en faktor som bidro til at resultatene på åpen og lukket versjon av en oppgave fikk høy korrelasjon. I tillegg til lik stamme, inneholdt to av oppgavene i lukket utgave distraktorer som er typiske misoppfatninger.

Oppgave A34 testet forståelse og anvendelse av atombegrepet. Konsekvensene av atombegrepet er vanskelig å forstå for mange elever. Det bekreftes ved at omtrent hver fjerde elev trodde at ”stolen vil være der, men veie mindre” hvis alle atomene ble fjernet. Dette kan være en misoppfatning, men også et resultat av en distraktor som villedet elevene, kognitiv

felle (Olsen mfl. 2001). Da oppgaven var åpen, var det 23 prosent av elevene som ikke svarte på oppgaven, mens bare 1,5 prosent lot være å svare da oppgaven var lukket. Siden løsningsprosenten i lukket format bare var 9 prosentpoeng høyere enn i åpent, kan dette tyde på at noen elever har stolt på sin misoppfatning og at andre kanskje var blitt villedet til å velge en distraktor som de ikke ville ha valgt hvis oppgaven var åpen.

Misoppfatning som forsterkes gjennom distraktor, er også tydelig i oppgave B30, hvor elevene trodde at oksygen var gassen det er mest av i luft. Det har tydeligvis hatt liten betydning hvilke alternative svar elevene ble fristet med. I lukket format valgte 73 prosent av elevene distraktoren oksygen, og i åpent format var oksygen svaret fra 63 prosent. Elever velger ut fra egen overbevisning, og det er grunn til tro at oppgave B30 var en god representant for denne oppfatningen (se figur 4.12). Bekreftet misoppfatning og distraktorer som villeder samsvarer både med undersøkelser som Kazemi (2002) gjorde med 4. klassinger, hvordan Olsen mfl. (2001) beskriver distraktorers virkemåte i oppgaver, og resultater Ravlo mfl. (2010) har funnet gjennom analyser av et utvalg på 20 000 elever fra de nasjonale prøvene i regning.

Oppgave A19 handlet om menstruasjon. I den lukkede versjonen var det bare 6 prosent som ikke svarte på oppgaven, men som åpen oppgave var andel ubesvart 32 prosent. Det typiske for distraktorene i denne oppgaven, var at de inneholdt mange begreper som kunne by på språkproblemer. Det kunne føre til at elevene ikke svarte det de mente å svare. Dette stemmer med undersøkelser gjennomført av Schoultz (2000) og Clerk mfl. (2000). I Clerks (2000) undersøkelse inneholdt riktignok oppgavene mange feil og mangler, men lesesvake elever kan ha språkproblemer selv om oppgavene oppfyller psykometriske krav. Det er viktig å få svar fra elevene. At 45 prosent trodde egget fester seg til livmoren, er en grei tilbakemelding for læreren å kunne ta tak i.

De tre andre oppgavene som ikke var signifikant forskjellig i åpent og lukket format, var B41b (se figur 3.5 og 3.6), B33 (se figur 4.11) og A7 (se figur 2.3 og 4.14) som ikke hadde samme stamme. Oppgave B41b hadde i lukket format fått listet opp alle alternativene som kunne være svar på oppgaven. Dette passer med *distraktor som sjekkliste* (Olsen mfl. 2001), selv om oppgave B41b ikke hadde samme stamme i oppgaveformatene, og det ikke førte til at p-verdien i lukket oppgave var høyere enn i åpen. En annen mulighet er at språkproblemer kan ha påvirket løsningsfrekvensen (Schoultz 2000, Clerk mfl. 2000). Å liste opp denne typen

alternativer ble sjonglering med de samme ordene, og kunne bli vanskelig å holde styr på for lesesvake elever. I slike oppgaver er det begrenset hvor mye hjelp elevene får fra alternativene. Oppgavens tema var fysikkrelatert og tester forståelse av begrep.

B33 målte fakta i åpent format og forståelse i lukket. Dette gjorde i utgangspunktet den lukkede utgaven vanskeligere enn den åpne. Rodriguez (2003) har i undersøkelser konkludert med at innholdsekvivalens delvis avhenger av oppgaveformatet og delvis av hvordan oppgaven er formulert. Her tyder resultatene på at elever som ikke hadde forståelsen på plass, heller ikke hadde særlig hjelp av svaralternativene. 30 prosent av elevene svarte alternativ D (se figur 4.11), og dette var omtrent like mange som de som fikk riktig svar på oppgaven. Som åpen oppgave ble det spurt om fakta (kjemisk reaksjon). Det var fullt mulig for elever å skrive ned et pugget svar da oppgaven var åpen, og det kan ha vært en medvirkende årsak til lik p-verdi i åpent og lukket format på denne oppgaven. Dette oppgaveparet var imidlertid for ulikt til at vi kan si at oppgavene testet samme kompetanse, men kan være et eksempel på at innfallsvinkelen til et tema er avgjørende for hvilken kompetanse elever får vist.

Oppgave A7 (se figur 4.14) hadde svært lav løsningsprosent og lav diskriminering. Lav diskriminering er en bekreftelse på at en oppgave skiller dårlig mellom elevene. Oppgaven er omtalt både i kapittel 4.4.1.og 4.4.2, og var en oppgave som ikke bidro positivt til prøvene. Begrepet pollenknapp var ukjent. Dette gjaldt ca. 80 prosent av elevene, uansett gjennomsnittlig p-verdi på prøven. Det hjelper ikke med alternativer hvis begrepet er helt ukjent. Oppgaven hadde ikke god nok kvalitet til at vi kan ta hensyn til resultatet.
For oppgave A7 i lukket format, se figur 2.3.

Jeg har bestemt kognitivt nivå ut fra kompetansemål i fag etter 7. trinn i LK06, kombinert med i hvilken grad oppgaven krever å gjengi fakta (lavt nivå), eller krever forståelse, anvendelse og vurdering (høyt nivå). Alle disse seks oppgavene er vurdert til å være på et høyt eller middels høyt kognitivt nivå. Vi kan si at i dette tilfellet samsvarer mine resultater bra med uttalelser fra Martinez (1999), om at jo mer krevende oppgavene er, jo bedre samsvarer åpen og lukket versjon av en oppgave. Dette må imidlertid sees i lys av at reliabiliteten var ca. 0,7 for begge prøvene i Test 2, og derfor gir resultatene en viss grad av usikkerhet.

5.2.3 Oppgaver som var signifikant forskjellig i åpent og lukket format, og hadde betydelig differanse i p-verdi

De fem oppgaveparene som hadde størst forskjell i p-verdi mellom åpent og lukket format, hadde høyest p-verdi som lukkede oppgaver. To hadde identisk formulering i stammen, fire testet fakta og en testet forståelse. Tre av faktaoppgavene var på et lavt kognitivt nivå, og de to andre krevde forståelse og kan sies å være av middels vanskelighetsgrad (se tabell 5.2). Oppgaveparet B12/A15 tester ulik kompetanse og er for ulik til å kunne kalles samme oppgave og ha relevans for problemstillinga. Jeg har allikevel valgt å diskutere oppgavene med hensyn på distraktorene.

Felles for oppgavene var at elevene i lukket versjon av oppgavene hadde nytte av distraktorene. Martinez (1999) har sagt at i oppgaver på et lavt kognitivt nivå, var det gjerne snakk om å gjengi i åpent og å gjenkjenne i lukket format. I slike oppgaver vil elevene kunne få hjelp av distraktorene. Dette kan være en årsak til at oppgaver på et lavt kognitivt nivå ofte gir ulikt resultat i åpent og lukket format.

I oppgave A40 (se figur 3.3) valgte ingen elever distraktor B, og bare 1,3 prosent av elevene distraktor C. Det gir grunn til å tro at mange elever har kjent igjen det riktige svaret (nøkkelen), eller at elevene har eliminert bort to av alternativene (Olsen mfl. 2001). Oppgaven krevde faktakunnskap. Lyn og torden er et begrep som det til tider fokuseres på gjennom media. Hvis man kjenner begrepene overfladisk, er det grunn til å tro at man kan få hjelp av svaralternativene.

Tabell 5.2

Oppgaver som var signifikant forskjellig i åpent og lukket format, med stor forskjell i p-verdi

Oppgave	Lik stamme	Fakta	For-ståelse	Mulig betydning av distraktor	Kogni-tivt nivå	p-verdi CR	p-verdi SR	Nytte av distraktor
A40/B27	Ja	Ja		Gjenkjenne Eliminere	Lavt	0,57	0,91	Ja
B12/A15	Nei	Ja	Ja	Eliminere	Middels	0,21	0,60	Ja
B25	Ja		Ja	Definerer spørsmålet Eliminere	Middels	0,12	0,47	Ja
A28	Nei	Ja		Gjenkjenne Eliminere	Lavt	0,34	0,81	Ja
B17	Nei	Ja		Gjenkjenne	Lavt	0,20	0,96	Ja

At svaralternativene har hatt betydning, kan vi også anta i oppgave A28 (se figur 5.2), siden bare 15,5 prosent av elevene til sammen valgte distraktorene A, B eller C. A28 spurte om fakta, og man har grunn til å tro at mange elever har gjenkjent det riktige svaret. I tillegg lot 51 prosent av elevene oppgaven stå ubesvart i åpent format, noe som kan tyde på at mange ikke husket svaret. Dette er et eksempel på en oppgave på lavt kognitivt nivå som måler faktakunnskaper, og at elevene da kan ha stort utbytte av svaralternativene (Martinez 1999).

Figur 5.2 Oppgave A28 i åpent og lukket format

Oppgave fA28

Fossilt brennstoff er dannet av

A uran
B sjøvann
C sand og grus
D døde planter og dyr

Oppgave åA28

Hvordan dannes fossilt brennstoff?

Oppgave B12/A15 (se figur 4.16) hadde ulik stamme og målte fakta i åpen og forståelse som lukket oppgave. Svaralternativene var imidlertid av en slik art at de trolig har vært til hjelp. Distraktor A og B ble valgt av til sammen 11,5 prosent av elevene. Her er det grunn til å tro at elevene både har eliminert bort en distraktor og kjent igjen nøkkelen (Haladyna 2002, Olsen mfl. 2001). Oppgavene kan ikke sees på som par siden de er for ulike.

Oppgave B25 (se figur 4.17) testet forståelse. Oppgaven var formulert på en slik måte at svaralternativene presiserte spørsmålet (Olsen mfl. 2001). Dette gav trolig de elevene som løste oppgaven i lukket format en fordel ved at de slapp å spekulere på hva som mentes med oppgaven. Verre var det da oppgaven var helt åpen. Oppgaven hadde både lav løsningsprosent ($p = 0,12$) og lav prosent ubesvart ($p = 0,10$). Dette kan tyde på at mange elever trodde de visste svaret. Lav diskriminering i lukket format kan tyde på at oppgaven ikke skilte godt mellom elevene. Analyser bekrefter dette. Elevene som fikk riktig på oppgaven, hadde fra 4 til 18 poeng på prøven.

Fire begreper skulle forklares i den siste oppgaven, B17. I lukket format skulle hvert begrep knyttes til riktig forklaring. Oppgaven hadde en form for avhengighet ved at hvis eleven kjente noen av begrepene, ville de andre falle på plass av seg selv. Dette kan trolig forklare den store forskjellen på 76 prosentpoeng mellom lukket og åpent format. Dette stemmer med uttalelse fra Lukhele (1994), Olsen mfl. (2001) og Rodriguez (2003) om hvilken betydning innhold og formulering av en oppgave har for et resultat, og hvilken betydning distraktorene kan ha når vi skal sammenligne åpne og lukkede oppgaver. B17 er tidligere omtalt i lukket versjon i kapittel 4.4.1.

For alle disse oppgavene er det grunn til å tro at distraktorene har hatt betydning for p-verdien. Eliminering, sjekkliste eller gjenkjenning og å presisere spørsmålet, er strategier som trolig er brukt i disse oppgavene hvor forskjellen i p-verdi for åpen og lukket oppgave er stor. Eliminering er lett å bruke når ikke alle distraktorene har god nok kvalitet. Forskning viser at det er vanskelig å lage mange, gode distraktorer. Et høyt antall distraktorer er derfor ikke ensbetydende med at en oppgave blir mer reliabel, eller at den skiller bedre mellom elevene (Haladyna mfl. 2002, Tarrant mfl. 2009). Tvert i mot kan mange distraktorer føre til at oppgaven blir mindre reliabel. Forskning har vist at i gjennomsnitt er det 1,5 distraktorer som fungerer i en oppgave. Anbefalt antall distraktorer er derfor to i tillegg til nøkkelen.

5.2.4 Oppgaver med signifikant høyere p-verdi som åpen enn lukket oppgave

To oppgaver hadde signifikant høyere løsningsprosent i åpent enn i lukket format (se tabell 5.3). Resultatene gir grunn til å tro at elevene ikke har hatt nytte av distraktorene, men at de heller har virket forstyrrende.

Tabell 5.3

Oppgaver med signifikant høyere p-verdi i åpent enn i lukket format

Oppgave	Lik stamme	Fakta	For-ståelse	Mulig betydning av distraktor	Kognitivt nivå	p-verdi CR	p-verdi SR	Nytte av distraktor
B41a	Nei		Ja	Distraherer Eliminering	Middels	0,79	0,68	Nei
B10	Nei	Ja		Distraherer Eliminering	Lavt	0,79	0,66	Nei

Oppgave B41a er omtalt på side 36 og 37. Oppgaven tester forståelse og er en fysikkoppgave. Dette kan være et eksempel på at tekst skaper forvirring for elevene (Schoultz 2000) eller kognitiv felle, ved at elevene møter alternativer som de ellers ikke ville vurdert å svare (Olsen

mfl. 2001). Det var enklere for elevene bare å sette piler på tannhjulene enn å ta stilling til relativt ordrike beskrivelser. Alternativ 1 var i tillegg bare valgt av 3 elever. Dette gir grunn til å tro at eliminering av distraktor er benyttet av noen elever.

Oppgave B10 (se figur 4.18) hadde ikke samme formulering som åpen og lukket oppgave, men var par i ulikt format. I lukket versjon svarte 18 prosent av elevene at hjertet hos fisk har samme oppgave som lunger hos mennesket, 14 prosent svarte nyrene, og en elev svarte skinnet. Som åpen oppgave svarte ingen av elevene verken nyrer eller skinnet. Dette gir grunn til å tro at noen av distraktorene har virket forstyrrende på elevene, og at eliminering også har vært benyttet i denne oppgaven.

Effekten av oppgaveformat på B41a var 0,25 og for B10 var effekten 0,27. Dette viser at oppgaveformatet ikke har hatt stor betydning, selv om forskjellene var statistisk signifikante. Begge oppgavene hadde større standardavvik som lukket enn som åpen oppgave, og altså større spredning blant resultatene i de lukkede enn de åpne utgavene.

5.3. Diskusjon av resultater til forskningsspørsmål 2

Dersom vi tar utgangspunkt i elevenes faglige nivå, er det noen elevgrupper som i større grad enn andre, har overvekt av riktige løsninger innenfor et av oppgaveformatene?

5.3.1 Resultater på oppgaveformat sett i forhold til faglig nivå for elevene

I gruppene G1 og G2 for Test 2, ble elevene delt inn i tre nivågrupper ut fra poengsummene de oppnådde på prøve 1 og prøve 2. I alle grupper og på alle nivåer var gjennomsnittlig p-verdi høyere for de lukkede oppgavene (SR) enn for de åpne (CR) (se tabell 4.17 – 4.20 og figur 4.20 og 4.21). Effekten av oppgaveformat da de samme gruppene av oppgaver var åpne og lukkede, var størst på det laveste nivået, e = 2,3 og e = 1,5, og lavest på det høyeste nivået hvor effekten for de tilsvarende gruppene var 1,5 og 1,0. Andel ubesvarte lukkede oppgaver, var ubetydelig på alle nivåene, men allikevel størst på nivå 1 og lavest på nivå 3. Ubesvarte åpne var betydelig på nivå 1 (25 % og 31 %), avtok til nivå 2 (16 % og 24 %) og bare 7 prosent i gjennomsnitt på nivå 3 for begge de sammenlignbare gruppene av oppgaver i ulikt format.

Oppgaveformatet viser stor effekt for alle nivåene, men det er de faglig svakeste elevene som har hatt størst utbytte av at oppgavene var i formatet lukket. Elevene på nivå 1 både svarte på oppgavene i større grad og løste mer enn dobbelt så mange oppgaver i forhold til da oppgavene var åpne.

Kan det være at faglig svake elever blir mer motiverte og føler at de har større sjanse til å lykkes i lukkede oppgaver? Crocker & Smitt (1987) har forsket på eksamensangst og mener svaralternativene skaper trygghet. Det kan også ha noe å gjøre med hvilke oppgaver elevene på nivå 1 løser. Martinez (1999) mener oppgaver på et lavt nivå som tester faktakunnskaper handler om å gjenkjenne og derfor av den grunn får høyere p-verdi i lukket enn åpent format. Dette er vist gjennom diskusjoner for forskningsspørsmål 1 i denne masteroppgaven. Gjetting er også en mulig årsak. Det er nærliggende å tenke på dette siden forskjellen på ubesvarte åpne og lukkede på nivå 1 er 23 prosentpoeng for den ene gruppen av oppgaver (SR1/CR2) og 28 prosentpoeng for den andre (SR2/CR1). Spesielt gutter på nivå 1 har vist tendenser til å gjette (Kazemi 2000).

5.3.2 Åpne og lukkede oppgaver for elever på nivå 1

Tabell 4.21 viser oppgaver hvor elever på nivå 1 hadde størst forskjell i p-verdi for lukket og åpent format. Fem av de sju oppgavene testet faktakunnskaper. I det lukkede formatet til to faktaoppgaver og en forståelsesoppgave, var minst en distraktor valgt i svært liten grad av elevene, og det er grunn til å tro at eliminering og gjenkjennelse har bidratt til høy løsningsprosent (Martinez 1999, Olsen mfl. 2001). Alle disse oppgavene hadde høy prosentandel ubesvarte i åpent format (21 %, 37 % og 67 %) og lav prosentandel ubesvarte i lukket format (0 %, 3 % og 6 %).

Den fjerde oppgaven var en matching-oppgave med forskjell i p-verdi på 82 prosentpoeng fra lukket til åpent format (se kapittel 4.5.3 oppgave B17). Her er det lav andel ubesvarte i begge formatene. Dette skyldes trolig at det var enkelt å sette piler mellom begrep og begrepets oppgave i den ene versjonen, og at noen av begrepene, men ikke alle, var greie å forklare i den åpne versjon.

I de lukkede versjonene av de siste tre oppgavene, ble alle distraktorene valgt i omtrent like stor grad av elevene. To av disse oppgavene hadde p-verdi 0 som åpen oppgave, og den siste hadde p-verdi 0,06. Andel ubesvart på de åpne oppgavene var 18, 34 og 49 prosent, mens de

lukkede hadde andel ubesvart fra 0 til 11 prosent. Siden alle distraktorene var valgt i relativt stor grad, kan det tyde på at svarene er avgitt tilfeldig, og at elever har gjettet. Analyser av nasjonale prøver i regning for 2007 konkluderte med at elever på det laveste nivået gjettet mer enn andre elever (Ravlo mfl. 2008). Det kan virke som om vi kan konkludere på tilsvarende måte her.

5.4. Diskusjon av resultater til forskningsspørsmål 3

Er det kjønnsforskjeller når det gjelder å mestre åpne og lukkede oppgaver?

Internasjonale undersøkelser refererer til resultater som viser at gutter besvarer lukkede oppgaver i større grad enn jenter (Hastedt mfl. 2005, Kjærnsli mfl. 2004, Martinez 1999). Kan en av årsakene til dette være at gutter som gruppe, i mange situasjoner ser ut til å ha både mer selvtillit enn jentene og tør å ta flere sjanser? I så fall kan det være at de både bruker svaralternativene mer konstruktivt og strategien kvalitativ gjetting oftere enn jentene når de i utgangspunktet ikke vet svaret (Beller mfl. 2000, Ben-Shakhar mfl. 1991).

5.4.1 Resultatene til jentene og guttene på åpne og lukkede oppgaver

Guttene oppnådde høyere gjennomsnittlig p-verdi enn jentene både på de åpne og de lukkede oppgavene, uansett hvilke grupper vi sammenlignet, men forskjellen var ikke signifikant (se figur 4.22). For de lukkede oppgavene i prøve 2 (SR2) var imidlertid effekten av å være gutt eller jente middels (e = 0,33) og større enn for de andre gruppene, hvor den var lav (se tabell 4.23). Hvis vi ikke skiller på jenter og gutter, viser t-tester som sammenligner gjennomsnittlig p-verdi for gruppe G1 på lukkede oppgaver i prøve 1 (SR1) med de lukkede oppgavene i prøve 2 (SR2) en signifikant forskjell ($p = 0{,}01 < 0{,}05$). At SR2 var vanskeligere enn SR1 har fått større konsekvenser for jentene enn for guttene, men resultatene viser ingen signifikant forskjell i gjennomsnittlige p-verdier for jenter og gutter når det gjelder å løse åpne eller lukkede oppgaver.

Resultatene fra min undersøkelse, stemmer ikke med tidligere forskning. At guttene gjør det bedre enn jentene på lukkede oppgaver, har vært en trend i undersøkelser så lenge dette er blitt forsket på, og tidligere var det også en oppfatning om at jentene gjorde det bedre enn guttene på de åpne oppgavene (Bolger mfl. 1990, Bridgeman mfl. 1994, Murphy 1982). Resultatene var imidlertid ikke entydige, og Wester (1995) viser til resultater hvor guttene i matematikk gjorde det best på begge områder, og hvor forskjellen var minst i de lukkede

oppgavene. I resultatene til undersøkelser som er gjort siden år 2000, er guttene fortsatt bedre enn jentene på de lukkede oppgavene, men jentene er ikke bedre enn guttene på de åpne og spesielt ikke i naturfag (DeMars 2000, Hastedt mfl. 2005, Kjærnsli mfl. 2004).

5.4.2 Kjønnsforskjeller på oppgavenivå

Guttene gjorde det bedre enn jentene i 10 åpne og 10 lukkede oppgaver, og jentene gjorde det bedre enn guttene i 4 åpne og 4 lukkede. Forskjellen var bare signifikant i guttefavør i fire av oppgavene, og disse resultatene på enkeltoppgaver, viser i tråd med annen forskning (Kjærnsli mfl. 2004) en tendens til at guttene gjør det bedre enn jentene både på enkelte åpne og enkelte lukkede oppgaver i naturfag. Senere års forskning mener dette mer skyldes innhold og formulering av oppgavene enn oppgaveformatet, men evidensen for dette er foreløpig svak (Clerk mfl. 2000, Rodriguez 2003, Schoultz 2000).

Tabell 5.4

Fagrelevans, antall oppgaver hvor det er mer enn 2 prosentpoeng forskjell på prestasjonene til jentene og guttene, og hvilket kjønn som har høyest p-verdi på oppgaven

Fagemne	**Antall lukkede oppgaver (SR)**		**Antall åpne oppgaver (CR)**		**Antall oppgaver med signifikant forskjell**	
	G > J	**J > G**	**G > J**	**J > G**	**G > J**	**J > G**
Fysikk	3		5		2	
Kjemi	5	1	2	1	2	
Biologi	2	3	3	3		
Sum	10	4	10	4	4	0

Ut fra antall oppgaver ser det ikke ut til at oppgaveformatet har hatt betydning for kjønnsforskjellene (se tabell 5.4). Både guttene og jentene gjør det bedre enn motsatt kjønn i like mange åpne som lukkede oppgaver. Bare antallet er ulikt og i guttefavør med 10 oppgaver for guttene og 4 for jentene. For guttene gjelder dette i størst grad oppgaver innen fysikk og kjemi med signifikant forskjell i fire oppgaver. Jentene gjør det bedre enn guttene først og fremst i biologi.

Tabell 5.5

Oppgaver som er signifikant forskjellig for jenter og gutter. Best resultat for guttene

Oppgave	**Fakta**	**Forståelse**	**Kognitivt nivå**	**p-verdi jente**	**p-verdi gutt**	**Nytte av distraktorer**
fA40	Ja		Lavt	0,86	0,96	Ja Gjenkjenne
åB41a		Ja	Middels	0,71	0,87	Trolig ikke
fA34		Ja	Middels	0,48	0,69	Nei Kognitiv felle
fB30	Ja		Høyt	0,03	0,14	Nei Kognitiv felle

Forskjellen på jenter og gutter i oppgave fA40 (se figur 3.3) var at flere jenter enn gutter trodde at ”lynet var nærmere oss enn torden”. Andel ubesvart var tilnærmet lik for jentene og guttene (se figur 4.26). Her er det grunn til å tro at det fysikkfaglige temaet har vært avgjørende for resultatet. I åpent format gjorde guttene det 16 prosentpoeng bedre enn jentene på denne oppgaven. At tema og innhold spiller større rolle for kjønnsforskjeller enn oppgaveformatet stemmer med undersøkelser gjennomført av Wester-Wedman (1992c), Holland (1986) og Martinez (1999).

Oppgave åB41a er en åpen fysikkoppgave med høyere p-verdi for både gutter og jenter i åpent enn i lukket format. Guttene oppnådde 17 prosentpoeng og jentene 5 prosentpoeng bedre gjennomsnitt som åpen oppgave. Her er det grunn til å tro at noen elever er blitt språklig forvirret eller at leseforståelsen har påvirket resultatet (Clerk mfl. 2000, Schoultz 2000).

I oppgave fA34 skyldes trolig noe av forskjellen at guttene eliminerte bort en distraktor, mens jentene fordelte svarene på alle distraktorene. Det virker som om spesielt jentene har valgt distraktorer som de normalt ikke ville tenkt på om oppgaven hadde vært åpen. Dette kan være en grunn til at forskjellen i lukket format var 21 prosentpoeng i guttefavør, mens forskjellen i guttefavør bare var 5 prosentpoeng da oppgaven var åpen.

Oppgave fB30 (se figur 4.12) kan være et eksempel på at distraktorer virker ulikt på jenter og gutter. To distraktorer er valgt i like stor grad av jentene og guttene. Den tredje er en typisk misoppfatning som er valgt av 15 prosentpoeng flere jenter enn gutter. Dette har ført til at omtrent tre ganger så mange gutter har svart riktig i lukket enn i åpent format (fra 5 % til 14 %), mens prosent jenter som har svart riktig er redusert til halvparten (fra 6 % til 3 %) fra åpen til lukket oppgave.

Noen undersøkelser har konkludert med at forskjellen mellom gutter og jenter er størst i åpne oppgaver (Bell mfl. 1987, Wester 1995) og andre har konkludert med at det er i flervalgsoppgaver (Ben-Shakhar mfl. 1991, DeMars 2000, Hasted mfl. 2005, Kjærnsli mfl. 2004). I mine resultater finner vi begge deler. I 11 oppgaver er forskjellen mellom jenter og gutter minst i de lukkede oppgavene, og i 7 oppgaver minst når elevene løser de åpne oppgavene. Dette er uavhengig av om det er guttene eller jentene som har høyest p-verdi på oppgavene. Det hersker imidlertid ganske stor enighet om at guttene vanligvis gjør det bedre

enn jentene på de lukkede oppgavene. Ut fra mine resultater kan jeg ikke trekke konklusjoner om at oppgaveformatet virker ulikt på jenter og gutter. Guttene presterte riktignok bedre enn jentene i 14 av 20 lukkede oppgaver, men også i 13 av 20 åpne oppgaver. Signifikant bedre enn jentene var guttene bare i tre lukkede og en åpen oppgave. Dette var to fysikk- og en kjemioppgave som krevde forståelse og en kjemioppgave som krevde faktakunnskap, og gir vel heller signaler om at tema og innhold kan ha betydning for prestasjonene til guttene og jentene.

5.4.3 Oppgaver som endret kjønnsfordel etter format

I fire oppgaver gjorde både jentene og guttene det bedre i lukket enn åpent format. I to oppgaver resulterte dette i at guttene fikk høyere gjennomsnittlig p-verdi enn jentene og i de to andre var det jentene som fikk bedre resultat enn guttene. I en annen oppgave gjorde begge kjønn det bedre i den åpne enn den lukkede oppgaven, og da ble differansen størst i guttefavør. I den siste oppgaven gikk jentene ned i p-verdi fra åpen til lukket, mens guttene gikk opp og økte differansen. Det var i den mye omtalte oppgaven om gasser i luft. Det er vanskelig å finne noe mønster i dette, og det er nærliggende å konkludere med at innhold og formulering har større betydning for kjønn enn oppgaveformat.

Andel ubesvarte var i gjennomsnitt 23 prosent for guttene og 20 prosent for jentene i de åpne utgavene av disse seks oppgavene. I de lukkede utgavene var gjennomsnittet 1 prosent for både jentene og guttene. Det var ingen forskjell i forhold til kjønn, og dette har derfor ikke påvirket resultatet guttene og jentene i mellom.

5.4.4 Kjønnsforskjeller på nivå for åpne og lukkede oppgaver

På alle nivå har både gutter og jenter i gjennomsnitt fordel av lukkede oppgaver, ved at de oppnår høyere løsningsprosent i de lukkede enn i de åpne oppgavene. Oppgaveformatet har minst betydning for elevene på nivå 3 og mest betydning for elevene på nivå 1. På alle nivå er det større forskjell på resultatene til gutter og jenter på de lukkede enn på de åpne oppgavene.

Gutter på nivå 1 skiller seg ut ved å ha større fordel av de lukkede oppgavene enn de andre nivåene. I tillegg er guttene på nivå 1 i prøve 2 signifikant bedre enn jentene når det gjelder å løse lukkede oppgaver (t-test: $p = 0{,}047 < 0{,}05$). Guttene på nivå 1 i prøve 1 har signifikant større andel ubesvarte enn jentene på samme nivå (t-test: $p=0{,}032<0{,}05$).

6. Kapittel Konklusjon

6.1. Hvilken sammenheng er det mellom elevers prestasjoner på en test i naturfag og formatet til oppgavene?

Utgangspunktet var to sammenlignbare grupper av elever (G1 og G2) som gjennomførte en test på 20 oppgaver i naturfag. Begge gruppene fikk 10 åpne og 10 lukkede oppgaver, men de oppgavene som den ene gruppen fikk som åpne, fikk den andre gruppen som lukkede. Åtte av oppgavene hadde parvis samme stimulus og stamme i åpent og lukket format. I 12 oppgaver var stammen endret noe for at oppgavene parvis skulle ha samme innhold som åpne og lukkede oppgaver. Den ene gruppen (G1) bestod av 140 elever og den andre (G2) av 135 elever. Analysene viser at prøven som gruppe G1 gjennomførte ikke var signifikant forskjellig fra prøven til gruppe G2, innenfor et konfidensintervall på 95 prosent. Dette var en viktig kvalitetssikring av grunnlagsmaterialet mitt for å undersøke om det var forskjell på resultatet elever oppnådde da de samme oppgavene var åpne og lukkede. For å kontrollere for feilkilden som ligger i at mennesker er ulike, lot jeg begge gruppene få 10 åpne og 10 lukkede oppgaver. Mine to elevgrupper utgjorde derfor i realiteten fire forsøksgrupper som var kontroll for hverandre ved at hver elev var involvert i tre av gruppene (se figur 3.1 og 3.2, s. 38 og 39).
I to av gruppene undersøkte jeg sammenlignbare elevers resultater på de samme oppgavene i to formater. I de to andre gruppene undersøkte jeg disse elevenes evne til å besvare åpne og lukkede oppgaver.

Funn 1

Med oppsummeringen i forrige avsnitt som bakgrunn, er resultatene jeg har funnet entydige i retning av at oppgaveformatet har betydning. Elevene hadde i gjennomsnitt fordel av det lukkede formatet i forhold til da de samme oppgavene var åpne. Begge gruppene av elever oppnådde i gjennomsnitt høyere p-verdi for de samme ti oppgavene da oppgavene var i lukket enn i åpent format. Effektstørrelsen på oppgaveformat var middels i den ene gruppen og stor i den andre. Også innen gruppe G1 og gruppe G2 oppnådde elevene høyere gjennomsnittlig p-verdi på de 10 lukkede enn på de 10 åpne oppgavene.

Jeg har ingen resultater som entydig peker i retning av hvorfor oppgaveformatet har betydning. Drøftingene av resultatene har imidlertid gitt signaler som kan tyde på at både tekstlige formuleringer, innhold, hvilke distraktorer som er listet opp i oppgavene og elevenes

førforestillinger, kan ha påvirket resultatet. Jeg har heller ingen resultater som kan gi svar på om poengsummene elevene fikk ved å besvare lukkede oppgaver, er mer eller mindre korrekte i forhold til elevenes reelle kompetanse, enn svarene elevene ga på de åpne oppgavene.

Funn 2

Elever med lav kompetanse hadde i gjennomsnitt forholdsvis større utbytte av oppgaver i lukket format enn elever med høy kompetanse. Elevene ble plassert på tre nivåer ut fra resultatene de oppnådde på prøvene. Alle nivågruppene fikk i gjennomsnitt høyere p-verdi da oppgavene var i lukket enn i åpent format. Størst var forskjellen på det laveste nivået og minst på det høyeste. Effektstørrelsen var imidlertid stor på alle nivåene i favør av lukket oppgaveformat.

Funn 3

Det var ingen signifikant forskjell på hele gruppen av gutter og jenter når det gjelder å løse åpne oppgaver eller å løse lukkede oppgaver. Både guttene og jentene oppnådde i gjennomsnitt høyere p-verdi på de lukkede enn på de åpne oppgavene. Det var heller ingen signifikant forskjell når det gjelder å la oppgaver stå ubesvart. Begge kjønn hadde størst andel ubesvarte på åpne oppgaver, ubetydelig andel ubesvarte på lukkede, og det var ingen signifikant forskjell i gjennomsnitt ubesvarte oppgaver for jenter og gutter i utvalget på 275 elever.

Hvis vi ser på kjønn og kompetansenivå, finner vi imidlertid noe interessant. Gutter på nivå 1 skiller seg ut både ved å ha størst differanse i gjennomsnittlig p-verdi mellom de samme oppgavene i lukket og åpent format, og i større grad enn andre grupper å la åpne oppgaver stå ubesvart. Her må vi ikke glemme at utvalget bare var på 16 elever i den ene gruppen og 20 elever i den andre.

Resultatene viser i tillegg en tendens som ikke er knyttet til oppgaveformat, men til tema og innhold. Guttene gjorde det signifikant bedre enn jentene i to fysikk- og en kjemioppgave som testet forståelse, og i en kjemioppgave som testet faktakunnskaper. Dette var tre lukkede og en åpen oppgave. I de tre oppgavene som krevde forståelse, var guttenes gjennomsnittlige

p-verdi fra 5 til 16 prosentpoeng høyere enn p-verdien til jentene også i det andre formatet av oppgavene. En annen tendens er at jentene fikk bedre resultat enn guttene på biologirelaterte oppgaver med lav p-verdi, men disse resultatene er ikke signifikante.

I hvilken grad kan mine resultater ha overføringsverdi? Utvalget mitt bestod av alle elevene på 8. trinn ved to store ungdomsskoler, men manglet i stor grad representanter fra minoritetsspråklige miljøer. Dette var et bevisst valg for å kunne eliminere bort faktoren som angår språkproblemer. Jeg hadde et relativt stort utvalg, og alle elevene på trinnet kunne i utgangspunktet delta. At utvalget ble testet før elevene ble delt inn i to grupper, og måten inndelingen ble foretatt på, gjorde at jeg hadde kontroll på en del faktorer som kan ha innvirkning på et resultat. Oppgavene jeg benyttet, var i utgangspunktet kvalitetssikret og oppfylte psykometriske krav. Jeg kan ikke generalisere, men på bakgrunn av oppsummeringen vil jeg tro at resultatene kan være ganske representative for tilsvarende elevgrupper.

6.2. Konsekvenser av funnene

Jeg har funnet at for mitt utvalg gjelder at elever i gjennomsnitt får bedre resultater på lukkede enn på åpne oppgaver, og spesielt gjelder dette for gutter med lav kompetanse. Andel ubesvarte er høyere på åpne enn lukkede oppgaver, og gutter med lav kompetanse lar åpne oppgaver stå ubesvart i større grad enn andre elever. Gutter på nivå 1 ser altså ut til å ha større fordel enn andre elever av oppgaver i lukket format. For at vi skal få tilbakemelding om elevers kompetanse er vi avhengig av elevsvar. At elever svarer i større grad, mener jeg er et godt argument for å benytte lukkede oppgaver. La så være om elever ikke alltid har valgt svar ut fra egen overbevisning. Det er en grunn til at et svar er avgitt. Årsaken kan i tillegg til faglig kompetanse skyldes distraktorene, kvalifisert gjetting, språkproblemer eller annet, men det er en årsak, og man får et svar, og det er nok til at læreren får signaler å gripe fatt i.

I hvilken grad trenger vi egentlig tester med åpne oppgaver i naturfag? For elever med høy kompetanse er det liten forskjell i gjennomsnittlig p-verdi for lukkede og åpne oppgaver. Elever i mellomgruppen presterer bedre på lukkede oppgaver enn åpne og har større andel ubesvarte på åpne enn lukkede. Elevene med lavest kompetanse har høy prosent ubesvarte åpne oppgaver, og får derfor ikke vist hva de kan med dette formatet. I tillegg har lukkede oppgaver høyere reliabilitet enn åpne oppgaver. Det som er avgjørende synes for meg å være

kvaliteten på de lukkede oppgavene. Hvis oppgavene blir utviklet i tråd med anbefalte retningslinjer og elevsvar brukt diagnostisk som distraktorer, kan lukkede oppgaver være et godt hjelpemiddel for læreren og elevene både i vurdering av og for læring.

6.3. Veien videre?

Vurderingsformer på nasjonalt nivå påvirker skolehverdagen i alle klasserom og slik også elevenes læringsutbytte. Læreplanen Kunnskapsløftet (LK06) legger stor vekt på at vurdering skal fremme læring og utvikling og være et hjelpemiddel i tilpassing av undervisningen til den enkelte elev. Med det menes at elevvurdering først og fremst skal være et motiveringsmiddel og redskap for å hjelpe eleven videre i læringsprosessen (Matthiesen 2007). Hvilken betydning oppgaveformatet har *for* læring, er derfor et viktig spørsmål.

I tilknytning til Test 1 og Test 2 besvarte elevene et spørreskjema som handlet om holdninger til åpne og lukkede oppgaver, og hvilke strategier elevene bruker, når de skal løse lukkede oppgaver. Dette ble for mye å ta med i denne oppgaven, men er svært viktig å vite noe om. Elevenes motivasjon og følelse av mestring har betydning i en prøvesituasjon. Det er ting som tyder på at en del elever blir mer motiverte og føler større trygghet når det finnes svaralternativer i oppgavene. Det er også grunn til å tro at spesielt en del faglig flinke elever, ikke føler at de får vist hva de har av kunnskaper, når de ikke får utfolde seg fritt i en oppgave. En studie som går ut på å få mer kunnskap om elevers forventninger og holdninger til åpne og lukkede oppgaver, hvilke strategier elevene bruker når de står overfor valgene i en oppgave, og ikke minst om oppgaveformatet påvirker måten elevene forbereder seg til tester på, kan derfor være interessant å få gjennomført i fremtiden.

Det er interessant å se at i min masterundersøkelse skiller guttene på nivå 1 i begge gruppene seg ut ved å ha stort positivt utbytte av oppgaver i lukket format. Det kunne også vært interessant å forske mer på denne gruppen av elever. Hvorfor er det slik?

Referanser

Angell, C. (1996). *Elevers fysikkforståelse. En studie basert på utvalgte fysikkoppgaver i TIMSS.* Doktorgradsavhandling, Universitetet i Oslo.

Bell, R. C. & Hay, J. A. (1987). Differences and biases in English language examination formats. *British Journal of Educational Psychology, 57,* 212-220.

Beller, M. & Gafni, Naomi (2000). Can Item Format (Multiple Choice vs. Open-Ended) Account for Gender Differences in Mathematics Achievement? *Sex Roles, 42.* *http://search.ebscohost.com/login.aspx?direct=true&db=fmh&AN=3165660&loginpage=Login.asp&site=ehost-live* 7. april 2010, kl.14.10.

Ben-Shakar, G. & Sinai, Y. (1991). Gender differences in multiple-choice tests: the role of differential guessing tendencies. *Journal of Educational Measuremen*t, *28,* 23-35.

Bergem, O. K., Nylèhn, J. & Grønmo, L. S. (2009). Undervisning i naturfag. I L. S. Grønmo & T. Onstad (red.): *Tegn til bedring. Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2007.* Kap.7. Oslo: Unipub.

Birenbaum, M. & og Tatsuoka, K. K. (1987). Open-ended versus multiple-choice response formats – It does make a difference for diagnostic purposes. *Applied Psychological Measurement, 11*, 385-395.

Boodoo, G. M. (1993). Performance assesments or multiple choice? *Educational Horizons, 72,* 50-56.

Bolger, N. & Kellaghan,T. (1990). Method of measurement and gender differences in cholastic aptitude test. *Journal of Educational Measurement, 27*, 165-174.

Bonesrønning, H. & Iversen, J. M. V. (2008). *Suksessfaktorer i grunnskolen: Analyse av Nasjonale prøver 2007.* SØF-rapport, 5. Senter for økonomisk forskning AS.

Bonesrønning, H. & Iversen, J. M. W. (2010). *Prestasjonsforskjeller mellom skoler og kommuner: Analyse av nasjonale prøver 2008.* SØF-rapport, 1. Senter for økonomisk forskning AS.

Bridgeman, B. & Lewis, C. (1994). The relationship of essay and multiple-choice scores with grades in college courses. *Journal of Educational Measurement, 31,* 37-50.

Bridgeman, B. & Morgan, R. (1996). Success in college for students with discrepancies between performance on multiple-choice and essay tests. *Journal of Educational Psychology, 88*, 333-340.

Clerk, D. & Rutherford, M. (2000). Language as confounding variable in the diagnosis of misconceptions. *International Journal of Science Education*, *22,* 703-717.

Crocker, L. & Smitt, A. (1987). Improving multiple choice test performance for examinees with different levels of test anxiety. *The Journal of Experimental Education, 55,* 201-205.

DeMars, C. E. (2000). Test Stakes and Item Format Interactions. *Applied Measurement in Education, 13* (1), 55-77.

Downing, S. (2003). Guessing on selected response examinations. *Medical Education, 37,* 670–671.

Downing, S. (2006). Selected Response Item Formats in Test Development. In S. Downing & T. M. Haladyna: *Handbook of Test Development,* kap. 12. London: Lawrence Erlbaum Associates, Inc., Publishers.

Ebel, R. L. & Frisbie, D. A. (1991). *Essentials of educational measurement*. Englewood Cliffs: Prentice Hall.

Gafni, N. & Melamed, E. N. (1994). Differential tendencies to guess as a function of gender and lingual-cultural reference group. *Studies in Educational Evaluation, 20,* 309-319.

Gallagher, A. M. & De Lisi, R. (1994). Gender differences in scholastic aptitude test-Matehematics problem solving among high-ability students. *Journal of Educational Psychology, 86,* 204-211.

Gay, L. R. (1980). The comparative effects of multiple choice versus short answer tests on retention. *Journal of Educational Measurement, 17,* 45-50.

Grosse, M. & Wright, B. D. (1985).Validity and reliability of true false tests. *Educational and Psychological Measurement, 45,* 1-13.

Grønmo, L. S., Bergem, O. K., Kjærnsli, M., Lie, S. & Turmo, A. (2004). *Hva i all verden har skjedd i realfagene? Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2003*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Grønmo, L. S. (2009). Hovedfunn og trender i TIMSS 2007. I L. S. Grønmo & T. Onstad (red.): *Tegn til bedring. Norske elevers prestasjoner i matematikk og naturfag i TIMSS 2007.* Kap. 1. Oslo: Unipub.

Grønmo, L. S., Onstad, T., Pedersen, I. F., Lie, S., Angell, C. & Rohatgi, A. (2009). *Matematikk og fysikk i videregående skole, "Et skritt tilbake".* Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo. Lastet ned fra: *http://www.timss.no/rapporter%202008/Kortrapport%20TIMSS%20Advanced%202008%20norsk.pdf* 14. desember 2009, kl. 17.23.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

Haladyna, T. M. (1999). *Developing and Validating Multiple-Choice Test Items, 2. utgave.* Mahwha: Lawrence Erlbaum Associates, Inc.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002). A Review of Multiple Choice Item Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15(3),* 309-334.

Haladyna,T. M. (2004). *Developing and validating multiple-choice test item.* Mahwah: Lawrence Erlbaum Associates.

Hastedt, D. & Sibberns, H. (2005). Differences between multiple choice items and constructed response items in the IEA TIMSS surveys*. Studies in Educational Evaluation 31,* 145 – 161.

Hennessey, S. (1993). Situated cognition and cognitive apprenticeship: Implications for classroom learning*. Studies in Science Education, 22,* 1-41.

Holland, P. W. & Thayer, D. T. (1986). *Differential item performance and the Mantel-Haenszel procedure.* Paper presented at the Annual Meeting of the American Education Research Association (AERA), San Francisco.

Kazemi, E. (2002). Exploring test performance in mathematics: the questions children`s answers raise*. Journal of Mathematical Behavior, 21,* 203 – 224*.*

Kjærnsli, M., Lie, S., Stokke, K. H. & Turmo, A. (1999): *Hva i all verden kan elevene i naturfag? Oppgaver med resultater og kommentarer.* Oslo: Universitetsforlaget.

Kjærnsli, M., Lie, S., Olsen, R. V., Roe, A. & Turmo, A. (2004). *Rett spor eller ville veier? Norske elevers prestasjoner i matematikk, naturfag og lesing i PISA 2003.* OSLO: Universitetsforlaget.

Kjærnsli, M., Lie, S., Olsen, R. V. & Roe, A. (2007). *Tid for tunge løft. Norske elevers kompetanse i naturfag, lesing og matematikk i PISA 2006.* Oslo: Universitetsforlaget.

Kleven, T. A. (2002). Statistikk. I T. A. Kleven (red): *Innføring i pedagogisk forskningsmetode. En hjelp til kritisk tolking og vurdering*. Kap. 4. Oslo: Unipub forlag.

Kleven, T. A. (2002). Hvordan er begrepene operasjonalisert? – Spørsmålet om begrepsvaliditet. I T. A. Kleven (red): *Innføring i pedagogisk forskningsmetode. En hjelp til kritisk tolking og vurdering*. Kap. 5. Oslo: Unipub forlag.

Kleven, T. A. (2002). Hvilke alternative forklaringer er mulige? - Spørsmålet om indre validitet. I T. A. Kleven (red): *Innføring i pedagogisk forskningsmetode. En hjelp til kritisk tolking og vurdering*. Kap. 6. Oslo: Unipub forlag.

Kumar, V. K., Rabinsky, L. & Pandey, T. N. (1979). Test mode, test instructions, and retention. *Contemporary Educational Psychology, 4,* 211-218.

Lie, S., Kjærnsli, M. & Brekke, G. (1997). *Hva i all verden skjer i realfagene? Internasjonalt lys på trettenåringers kunnskaper, holdninger og undervisning i norsk skole*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitet i Oslo.

Lie, S., Kjærnsli, M., Roe, A. & Turmo, A. (2001). *Godt rustet for framtida? Norske 15-åringers kompetanse i lesing og realfag i et internasjonalt perspektiv*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Lie, S., Caspersen, M. & Bjørnson, J. K. (2004). *Nasjonale prøver på prøve. Rapport fra utvalgsundersøkelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prøver våren 2004*. Oslo: Institutt for lærerutdanning og skoleutvikling, Universitetet i Oslo.

Lie, S., Hopfenbach, T. N., Ibsen, E. & Turmo, A. (2005). *Nasjonale prøver på ny prøve. Rapport fra 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.

Lukhele, R., Thissen, D. & Wainer, H. (1994). On the Relative Value of Multiple-Choice, Constructed Response, and Examinee-Selected Items on Two Achievement Tests. *Journal of Educational Measurement, 31,* 234-250.

Martinez, M. E. (1999). Cognition and the Question of Test Item Format. *Educational, Psychologist, 34(4),* 207–218.

Matthiesen, G. (2007). Vurdering i skole og opplæring – problemstillinger og erfaringer. I S. Tveit (red.): *Elevvurdering i skolen. Grunnlag for kulturendring,* 47-55. Oslo: Universitetsforlaget.

Messick. S. (1993). Trait equivalence as construct validity of score interpretation across multiple metods of measurements. I R. E. Bennet & W. C. Ward (Eds.): *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment,* 61-73. Hillsdale: Lawrence Erlbaum Assosiates,Inc.

Monroe, W. S. (1918). Monroe's Standardized Silent Reading Tests. *Journal of Educational Psychology, 9* (6). Lastet ned fra http://psycnet.apa.org/journals/edu/9/6/303.pdf 30. mars 2010, kl. 11.30.

Murphy, R. J. L. (1982). Sex differences in objective test performance. *British Journal of Educational Psychology, 52,* 213-219.

Olsen, R. V., Turmo, A. & Lie, S. (2001). Learning about students` knowledge and thinking in science through large-scale quantitative studies. *European Journal of Psychology of Education, 16* (3), 403-420.

Ravlo, G. (2008). *Rapport, nasjonal prøve i regning 8.trinn 2007.* Trondheim: Nasjonalt senter for matematikk i opplæringen, Norges teknisk naturvitenskapelige universitet. Lastet ned fra http://www.udir.no/upload/Nasjonale_prover/Nasjonale_prover_regning_8_trinn_rapport.pdf 19. oktober 2008, kl. 19.40.

Ravlo, G., Johansen, O. H., Tokle, O. D., Andersen, T. & Vinje, B. (2010). *Rapport, nasjonal prøve i regning 8.trinn 2009.* Trondheim: Nasjonalt senter for matematikk i opplæringen, Norges teknisk naturvitenskapelige universitet. Kan lastes ned fra http://www.udir.no etter 1. juni 2010.

Robson, C. (2002): *Real World Reseach. A Resource for Social Scientists and Practitioner-Researchers. Second edition.* Oxford: Blackwell Publishers Ltd.

Rodriguez, M. C. (2003). Construct Equivalence of Multiple-Choice and Constructed-Response Items: A Random Effects Synthesis of Correlations. *Journal of Educational Measurement, 40* (2), 163-184. Lastet ned fra http://www.jstor.org/pss/1435344 31. mars 2010, kl.13.12.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Journal of Educational Measurement Issues Pract 24* (2), 3-13.

Schoultz , J. (2000). Conceptual knowledge in talk and text: What does it take to understand a science question? *Att samtala om /i naturvetenskap. Kommunikation, context och artefakt*. Doktorgradsavhandling, Universitet i Linkøping.

Sirnes, S. M. (2007). Flervalgsoppgaver, Tippekuponger eller seriøs vurderingsform? *Lektorbladet, 2,* 12-15. Lastet ned fra http://www.norsklektorlag.no/getfile.php/Filer/Lektorbladet%20%28filmappe%29/Lektorbladet_0207.pdf 10. mai 2010, kl. 10.07.

Sjøberg, S. (1986): Elever og lærere sier sin mening. *Rapport nr. 1 fra SISS-prosjektet: The second International Science Study i 1984.* Oslo: Universitetsforlaget.

Sjøberg, S. (1998): *Naturfag som allmenndannelse, en kritisk fagdidaktikk*. Oslo: Ad Notam Gyldendal.

Snow, R. E. & Lohman D. F. (1989). Implications of cognitive psychology for educational measurement. I R. I. Linn (Ed.): *Educational measurement,* 263 – 331. New York: Macmillian.

Tarrant, M., Ware J. & Mohammed, A. M. (2009). An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. *Biomedcentral Medical Education, 9* (40). Lastet ned fra *http://www.biomedcentral.com/1472-6920/9/40*, 8. april 2010, kl. 23.40.

Terry, P. W. (1933). How students review for objective and essay tests. *The Elementary School Journal, 33,* 592-603.

Traub, R. E. & MacRury, K. (1990). Antwort-auswahl vs freie-antwort-aufgaben beilernerfolgs-tests (Multiple-choice vs free-response in the testing of scholastic achievement). I K. Ingenkamp & R. S. Jager (Eds.)*: Tests unds trends 8: Jahrbuch der pedagogischen diagnostic,* 128–159. Weinheim: Beltz Verlag.

Utdanningsdirektoratet (2006). *Rammeverk for nasjonale prøver*. Lastet ned fra *http://www.udir.no*, 10. april 2010, kl. 23.00.

Wester, A. (1995). The Importance of the Item Format with Respect to Gender Differences in Test Performance: a study of open-format items in the DTM test. *Scandinavian Journal of Educational Research, 39* (4).

Wester-Wedman, A. (1992c). Item bias with respect to gender interpreted in the light of problem solving. Paper presenterat at the *IAEA 18th Annual Conference*, 14-15 September, Dublin.

Zimmerman, D. W. & Williams R. H. (2003). A New Look at the Influence of Guessing on the Reliability of Multiple-Choice Tests. *Applied Psychological Measurement, 27*, 357. *Lastet ned fra* http://apm.sagepub.com/cgi/content/abstract/27/5/357, 10. april 2010, kl. 10.20.

Liste over vedleggene

Vedlegg 1 Test 1

Skole:_____ **Klasse:**_____

Navn / Elev nr:_____**Jente:** ☐ **Gutt:** ☐

NATURFAG
8. trinn

Dette er en prøve for å finne ut hva elever husker av det de har holdt på med i naturfag på barneskolen. En annen skole i Trondheim og mange skoler sør i Norge prøver også elevene i de samme oppgavene.

Prøven består av to typer oppgaver. I flervalgsoppgavene skal du bare sette en ring ved det svaret du mener er rett. I de andre oppgavene må du skrive svaret med egne ord.

Det er viktig at alle gjør sitt beste og tenker seg nøye om før de svarer på en oppgave. Lette og vanskelige oppgaver er spredt rundt i settet, så ikke gi opp om du møter på en oppgave du synes er vanskelig! Gå videre, - kanskje er det en lett oppgave på neste side! ☺

Alle elever vil få tilbakemelding om resultatet sitt.

Lykke til!

Først skal du svare på disse spørsmålene:

Du skal bare sette <u>ett kryss</u> for hvert utsagn:

	Enig	Litt enig	Litt uenig	Uenig
1. Naturfag er et viktig fag	☐	☐	☐	☐
2. Jeg liker naturfag	☐	☐	☐	☐
3. Jeg kan mye naturfag	☐	☐	☐	☐
4. Jeg liker å se på "Newton" eller "Schrødingers katt"	☐	☐	☐	☐

Da er det bare å starte med prøven! ☺

Oppgave A 23

Sett inn ordene på riktig plass i teksten:

akser, baner, planeter, solsystemer.

Vårt planetsystem er et av flere ……..... Her beveger åtte

……. seg rundt Sola i ellipseformede ……..

Samtidig roterer de alle rundt sine egne ……..

Oppgave A 25

Sola er større enn månen, men når du ser på den fra Jorda, ser de ut til å være omtrent like store. **Forklar dette**.

……

……

……

Oppgave A2

Maria samlet opp gass som ble avgitt fra glødende kullbiter. Gassen ble så ledet ned i en glasskolbe med klart kalkvann. En setning i Marias arbeidsbok sa:

”Etter at gassen ble ledet ned i glasskolben, fikk kalkvannet gradvis en melkehvit farge.”

Dette utsagnet er

A En observasjon
B En konklusjon
C En teori
D En hypotese

Oppgave B 29

Vann er en væske. **Hva må du gjøre for at**

a) vann skal omdannes til en gass?

……

b) vann skal omdannes til et fast stoff?

……

Oppgave B 8

Hva skjer ved befruktning hos pattedyr?

A Det produseres sædceller og eggceller

B En sædcelle og en eggcelle smelter sammen

C En eggcelle deler seg

D Et foster utvikler seg

Oppgave A 18

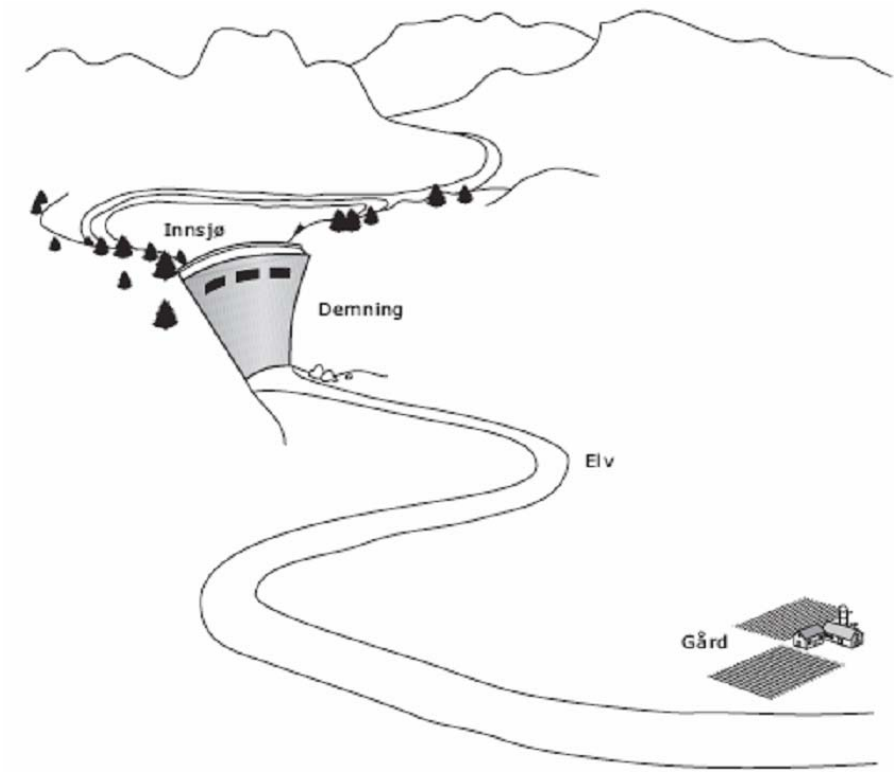
Nevn to viktige oppgaver skjelettet har.

1: .....

2: .....

Oppgave A35

Figuren viser et område hvor en demning nettopp er bygd.



Oppgave a)

Hva kan være en grunn til at denne demningen ble bygd?

……

……

Oppgave b)

Beskriv en negativ virkning demningen kan ha for miljøet i området.

……

……

Oppgave A37

Hva av dette er en kjemisk reaksjon?

A Et metall bankes ut til en tynn plate

B Et metall varmes opp så det smelter

C Et metall får en grønn farge ved å være i luft

D Et metall males opp til et fint og glatt pulver

Oppgave A16

Mye skjer med maten vi spiser fra vi putter den i munnen og til næringsstoffene blir tatt opp i kroppen.

Tegn en strek mellom hver kroppsdel og den funksjonen som denne kroppsdelen har i fordøyelsen.

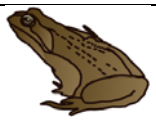
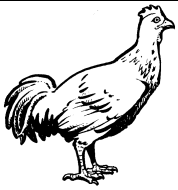
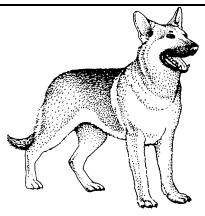
Munn	Mat blandes godt med magesyre og løses opp enda mer
Magesekk	Sukker, fett og andre næringsstoffer blir sugd opp fra maten som nå er tyntflytende
Tolvfingertarm	Mat knuses og blandes med spytt
Tykktarm	Mat tilsettes galle og bukspytt
Tynntarm	Vann blir sugd opp fra maten slik at den blir fastere

Oppgave A9

Virveldyrene har skjelett inni seg og deles i fem forskjellige grupper som vist i tabellen nedenfor.

Gruppe	**Eksempel**
Fisk	Gjedde
Amfibier	Padde
Reptiler	Slange
Fugl	Kjøttmeis
Pattedyr	Ku

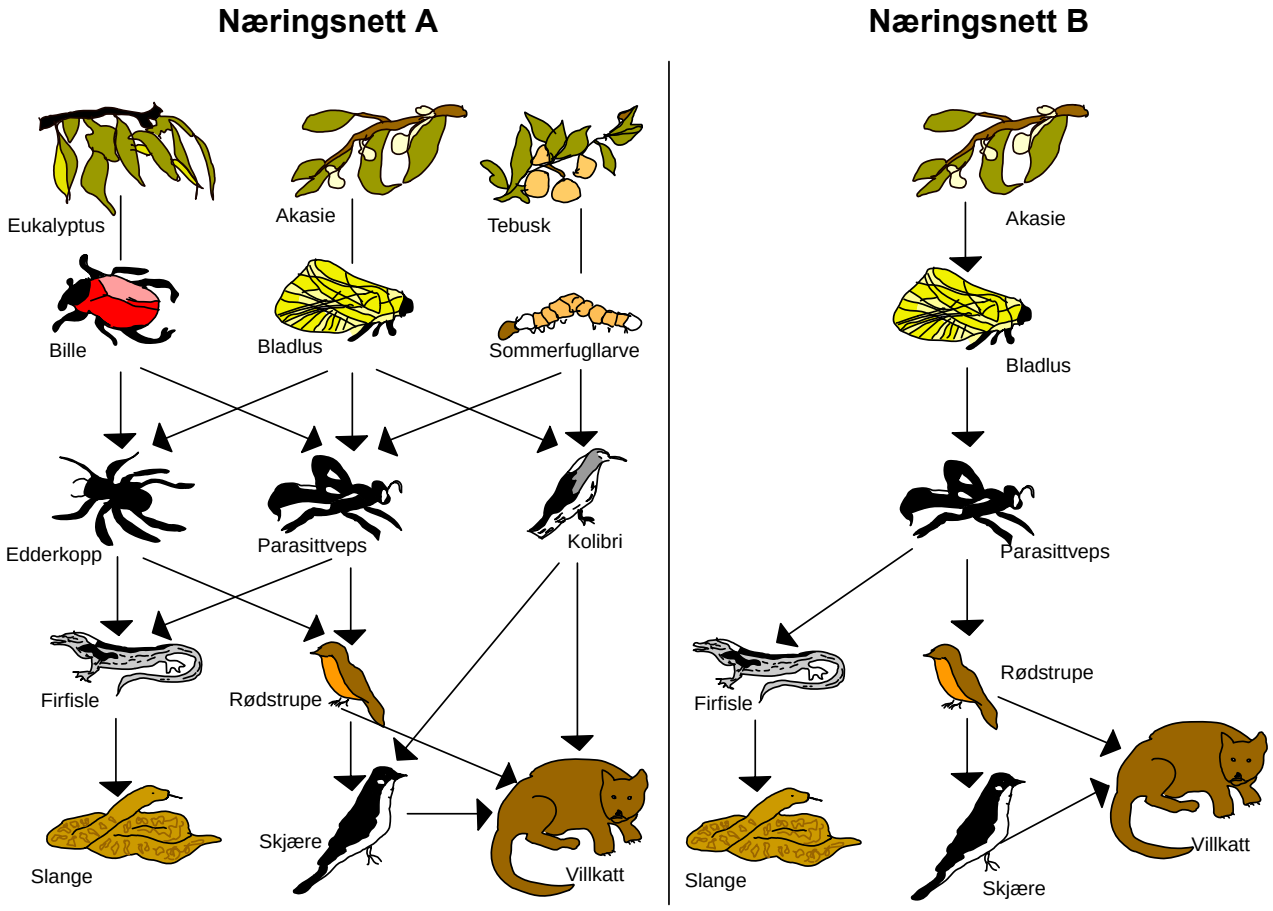
Se på dyrene i tabellen og skriv navn på den gruppen de tilhører:

Gruppe	**Dyr**
	
	
	
	

Oppgave A14

Alle dyr og planter er en del av et økosystem. Ved hjelp av næringsnett kan vi vise hvordan dyr og planter er avhengige av hverandre som matkilde. Økosystemer med stor variasjon av levende organismer har større mulighet for å tilpasse seg miljøforandringer enn økosystemer med færre arter.

Figuren nedenfor viser to ulike næringsnett (A og B).



Oppgave a)

Studer næringsnett **A.** Bare to dyr i dette næringsnettet har tre matkilder.

Hvilke to dyr er dette?

A Villkatt og parasittveps

B Villkatt og skjære

C Villkatt og kolibri

D Parasittveps og bladlus

Oppgave b)

Hvilket av disse næringsnettene blir minst påvirket hvis bladlusene forsvinner?
Begrunn svaret ditt.

……

……

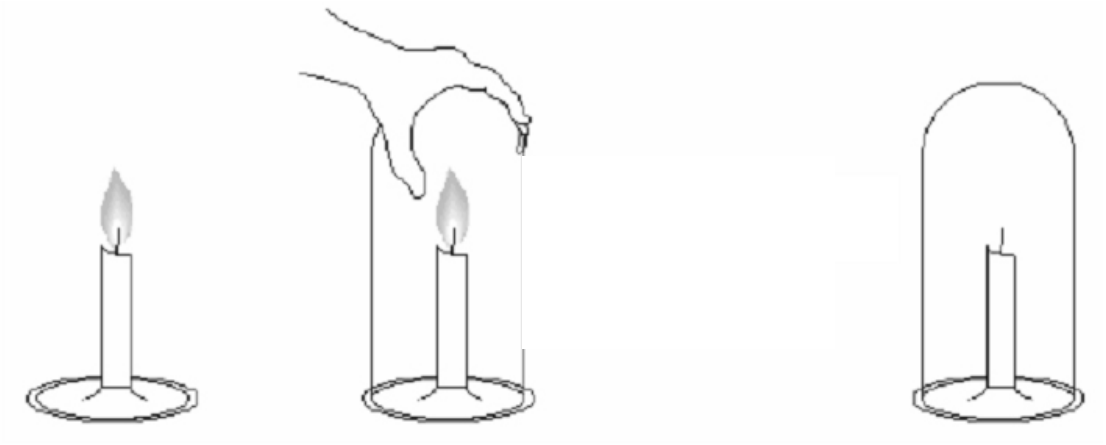
Oppgave A 29

Nedenfor finner du flere utsagn om magnetisme. Vurder om hvert utsagn er riktig eller galt. **Sett ring rundt "Riktig" eller "Galt" for hvert utsagn.**

Utsagn	**Riktig eller galt?**
Magneter trekker til seg metaller og glass	Riktig / Galt
Jorda er en stor magnet	Riktig / Galt
Magneter kan støte fra seg andre magneter	Riktig / Galt
Ulike poler frastøter hverandre	Riktig / Galt

Oppgave B31

Hvis man setter en glassklokke over et tent lys, vil lyset slokne.



Forklar hvorfor dette skjer:

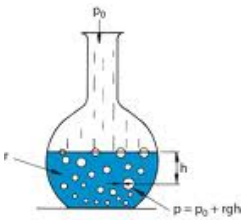
……

……

Oppgave A 31

Hva skjer med vannmolekylene når vann fordamper?

A	De forsvinner
B	De blir større og lettere
C	De blir til luft
D	De fjerner seg fra hverandre



Oppgave B 34

Egenskapene til 3 materialer er sammenlignet i tabellen nedenfor. Materialene er: tre, stein og jern.

Egenskap	Materiale 1	Materiale 2	Materiale 3
Synker i vann?	Ja	Nei	Ja
Brenner lett?	Nei	Ja	Nei
Tiltrekkes av en magnet?	Ja	Nei	Nei

Fyll inn riktig nummer på materialene:

Tre er materiale nummer …….

Stein er materiale nummer ……..

Jern er materiale nummer ……..

Oppgave B 39

I gamle dager, før man hadde elektrisk strøm, utnyttet man energien i vind og vann til praktiske formål.

Gi ett eksempel på hvordan energien i vann ble utnyttet og ett eksempel på hvordan energien i vind ble utnyttet.

Vann: .....

.....……

……..

Vind: .....

.....

……

Oppgave B 3

David løser opp 10 gram salt i 100 ml vann.

Hva må han tilsette den opprinnelige løsningen for å få en løsning som er halvparten så konsentrert?

A 50 ml vann
B 100 ml vann
C 5 gram salt
D 10 gram salt

Jente: ☐ **Gutt:** ☐

Dette skal du svare på etter at du har levert prøven:

Du skal bare sette <u>ett kryss</u> for hvert utsagn:

	Enig	**Litt enig**	**Litt uenig**	**Uenig**
5. Temaene i oppgavene var kjent for meg	☐	☐	☐	☐
6. Jeg fikk vist hva jeg kan på denne prøven	☐	☐	☐	☐
7. Jeg har i hvert fall gjort mitt beste!☺	☐	☐	☐	☐

8. Anta at du skulle svare på en oppgave ved å krysse av for rett svar, og at du ikke visste svaret. Hva gjorde du da?

Beskriv hva du gjorde så nøyaktig som mulig! ☺

_____

_____

Takk for fin innsats! ☺

Vedlegg 2 Test 2 prøve 1

Skole:_____ **Klasse:**_____

Navn / Elev nr:_____ Jente: ☐ Gutt: ☐

Tidspunkt start:_____ **Tidspunkt slutt:**_____

NATURFAG 8. trinn
Test 2 Prøve 1

Dette er en prøve for å finne ut hva elever husker av det de har holdt på med i naturfag på barneskolen. En skole til i Trondheim og mange skoler sør i Norge prøver også elevene i de samme oppgavene.

Prøven består av to typer oppgaver. I flervalgsoppgavene skal du bare sette en ring ved det svaret du mener er rett eller sette streker mellom det som hører sammen. I de andre oppgavene må du skrive svaret med egne ord.

Det er viktig at alle gjør sitt beste og tenker seg nøye om før de svarer på en oppgave. Lette og vanskelige oppgaver er spredt rundt i settet, så ikke gi opp om du møter på en oppgave du synes er vanskelig! ☺ Gå videre, - kanskje er det en lett oppgave på neste side!

Alle elever vil få tilbakemelding om resultatet sitt.

Lykke til!

Oppgave fA40 (B27)

Hvorfor ser vi lyn før vi hører torden?

A Fordi øynene oppfatter lys lettere enn ørene oppfatter lyd

B Fordi synssansen er sterkere enn hørselssansen

C Fordi lynet skjer nærmere oss enn torden

D Fordi lys beveger seg fortere enn lyd

_____

Oppgave fA22

Hvorfor lyser månen?

A Den er glødende

B Solen skinner på den

C Den blir synlig i mørket

D Den består av hvite gasser

Oppgave åB14

Nevn et organ i kroppen som ligger i bukhulen.

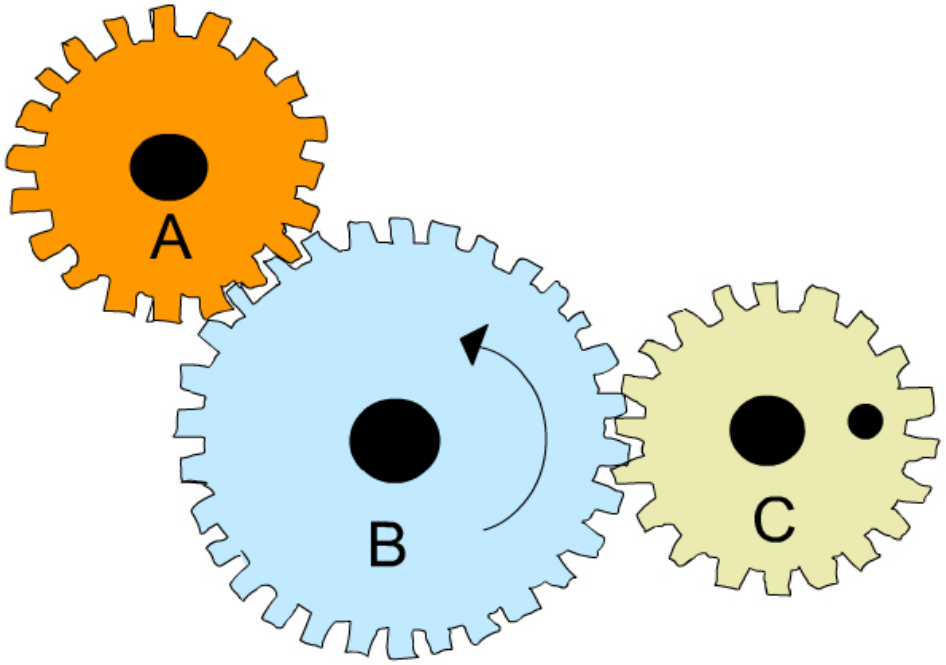
_____

_____

Oppgave åB41a)

Hvilken retning dreier tannhjulene?

Tegn inn piler på tannhjul A og tannhjul C som viser hvilken retning de dreier.



Oppgave åB41b)

Hvilket av tannhjulene B og C bruker kortest tid på en hel runde?

_____

Oppgave åA3

Kaja satte ut en skål med vann og en skål med bensin på et bord en varm solskinnsdag. Etter noen timer var det mindre bensin enn vann igjen i skålene.

Hva kan vi lære av dette eksperimentet?

_____

_____

_____

Oppgave fA33

Hvilken gass kan få en glødende treflis til å brenne?

A Neon

B Oksygen

C Nitrogen

D Karbondioksid

_____

Oppgave åB12(A15)

Hva betyr det at et dyr er vekselvarmt?

_____

_____

_____

Oppgave åA34

Hvis vi kunne fjerne alle atomene fra en stol, hva ville blitt igjen?

_____

_____

_____

Oppgave fA32

Tabellen viser temperaturen på et sted til forskjellige tider på dagen i tre dager.

	6.00	**9.00**	**12.00**	**15.00**	**18.00**
Mandag	15°C	17°C	20°C	21°C	19°C
Tirsdag	15°C	15°C	15°C	5°C	4°C
Onsdag	8°C	10°C	14°C	14°C	13°C

Når begynte det å blåse en kaldere vind?

A Mandag ettermiddag

B Tirsdag morgen

C Tirsdag ettermiddag

D Onsdag morgen

Oppgave åB25

Tegningen viser to forskjellige fjellområder. Fjellene i område A er ujevne og taggete. Fjellene i område B er glatte og avrundede

Fjellområde A

Fjellområde B

Hva er trolig grunnen til at fjellene ser så forskjellige ut?

_____

_____

_____

Oppgave åB33

Forklar hva som menes med en *kjemisk reaksjon*. Gi et eksempel.

_____

_____

_____

Oppgave fA41

En ballong fylles med luft og legges i kjøleskapet over natta.

Hva observerer du når du tar ballongen ut av kjøleskapet neste dag?

A Ballongen har utvidet seg

B Ballongen har krympet

C Det har ikke skjedd noe

D Ballongen har sprukket

Oppgave fB10

Hvilket av disse organene hos fisk har samme funksjon som lunger hos mennesker?

A Gjellene

B Hjertet

C Nyrene

D Skinnet

Oppgave åB30

Lufta består av mange gasser.

Hvilken gass er det mest av?

_____

Oppgave fA19

Hva skjer hos jenter ved menstruasjon?

A Det er på det tidspunktet et barn kan bli unnfanget

B Et egg løsner fra eggleder en og fester seg i livmoren

C Et egg utvikler seg til en liten celleklump

D Et egg forsvinner ut gjennom livmoren uten å feste seg

Oppgave åA28

Hvordan dannes fossilt brennstoff?

_____

_____

_____

Oppgave fA7

Sett strek mellom ordene på sidene og riktig sted på planten i midten.

Stengel

Støvbærer

Kronblad

Blad

Rot

Pollenknapp

Oppgave fA36

Hva av dette er *ikke* et eksempel på en kjemisk reaksjon?

A Vann som koker

B Jern som ruster

C Ved som brenner

D En eplebit som blir brun

Oppgave fB17

Hvilke oppgaver har de ulike organene i kroppen?

Sett strek mellom organene og riktig oppgave.

Hjerte	Beskytter kroppen og hindrer at den tørker ut
Lunger	Holder kroppen oppreist og bevegelig
Hud	Skaffer oksygen til cellene
Skjelett	Pumper blod ut i kroppen

Vedlegg 3 Test 2 - prøve 2

Skole:_____ **Klasse:**_____

Navn / Elev nr:_____ Jente: ☐ Gutt: ☐

Tidspunkt start: _____ **Tidspunkt slutt:** _____

NATURFAG 8. trinn
Test 2 Prøve 2

Dette er en prøve for å finne ut hva elever husker av det de har holdt på med i naturfag på barneskolen. En skole til i Trondheim og mange skoler sør i Norge prøver også elevene i de samme oppgavene.

Prøven består av to typer oppgaver. I flervalgsoppgavene skal du bare sette en ring ved det svaret du mener er rett eller sette streker mellom det som hører sammen. I de andre oppgavene må du skrive svaret med egne ord.

Det er viktig at alle gjør sitt beste og tenker seg nøye om før de svarer på en oppgave. Lette og vanskelige oppgaver er spredt rundt i settet, så ikke gi opp om du møter på en oppgave du synes er vanskelig! ☺ Gå videre, - kanskje er det en lett oppgave på neste side!

Alle elever vil få tilbakemelding om resultatet sitt.

Lykke til!

Oppgave åB27(A40)

Hvorfor ser vi lyn før vi hører torden?

_____

_____

_____

_____

Oppgave åA22

Hvorfor lyser månen?

_____

_____

_____

Oppgave fB14

Hvilket av følgende organer finner vi *ikke* i bukhulen?

A Hjerte

B Magesekk

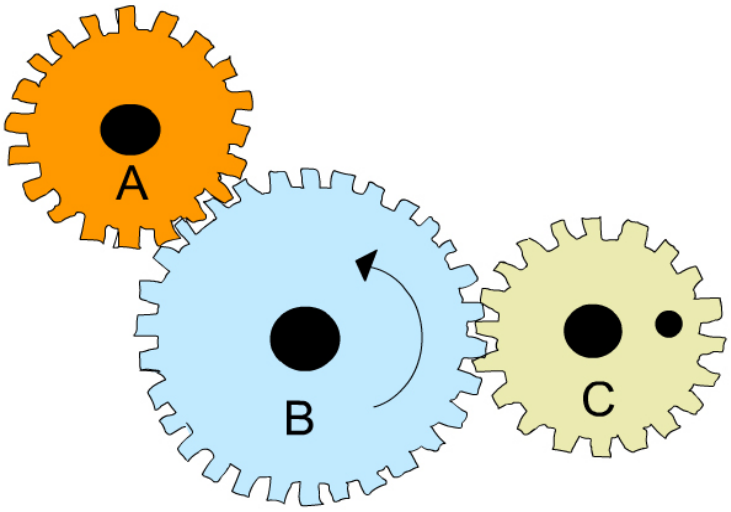
C Urinblære

D Lever

Oppgave fB41a)

Tannhjul B dreier i pilens retning.

I hvilken retning dreier tannhjulene A og C?



1 A og C dreier begge i samme retning som B

2 A dreier i samme retning som B og C dreier i motsatt retning

3 A og C dreier begge i motsatt retning som B

4 C dreier i samme retning som B og A dreier i motsatt retning

Oppgave fB41b)

Hvilket av utsagnene er riktige?

1 B bruker kortere tid enn C på en hel runde

2 C bruker kortere tid enn B på en hel runde

3 B og C bruker like lang tid på en hel runde

4 Det er umulig å avgjøre hvem av B og C som bruker kortest tid på en hel runde

Oppgave fA3

Kaja satte ut en skål med vann og en skål med bensin på et bord en varm solskinnsdag. Etter noen timer var det mindre bensin enn vann igjen i skålene. **Dette eksperimentet viser at**

A alle væsker fordamper

B bensin blir varmere enn vann

C væsker fordamper bare i solskinn

D noen væsker fordamper fortere enn andre

Oppgave åA33

Hva heter gassen som kan få en glødende treflis til å brenne?

_____

Oppgave fA15

På hvilken måte er varmblodige dyr forskjellige fra vekselvarme dyr?

A Hos varmblodige dyr øker stoffskiftet i varmt vær

B Varmblodige dyr er mer fiendtlige i fangenskap

C Varmblodige dyr har alltid høyere temperatur i blodet

D Varmblodige dyr har konstant kroppstemperatur uavhengig av temperaturen i omgivelsene

Oppgave fA34

Hvis vi kunne fjerne alle atomene fra en stol, hva ville blitt igjen?

A Stolen ville vært der, men den ville veid mindre

B Det ville ikke vært noe igjen av stolen

C Stolen ville vært akkurat som før

D Det ville bare vært igjen en dam på gulvet

_____

Oppgave åA32

Tabellen viser temperaturen på et sted til forskjellige tider på dagen i tre dager.

	6.00	**9.00**	**12.00**	**15.00**	**18.00**
Mandag	15°C	17°C	20°C	21°C	19°C
Tirsdag	15°C	15°C	15°C	5°C	4°C
Onsdag	8°C	10°C	14°C	14°C	13°C

Studer tabellen.

Når begynte det å blåse en kaldere vind?

_____

_____

Oppgave fB25

Tegningen viser to forskjellige fjellområder. Fjellene i område A er ujevne og taggete. Fjellene i område B er glatte og avrundede

Fjellområde A

Fjellområde B

Hva er trolig grunnen til at fjellene ser så forskjellige ut?

A Fjellområde A er eldst

B Fjellområde B er eldst

C Fjellområdene er like gamle, men B er nedslitt av fotturister

D Fjellområdene er like gamle, men A har opprinnelig vært en vulkan

---

Oppgave fB33

Hvilken av disse hendelsene er et eksempel på en kjemisk reaksjon?

A Is som smelter

B Saltkrystaller som knuses

C Ved som brenner

D Vann som fordamper fra en vanndam

Oppgave åA41

En ballong fylles med luft og legges i kjøleskapet over natta.

Hvilken forandring skjer med ballongen i løpet av natta?

_____

_____

Oppgave åB10

Hvilket organ hos fisk har samme funksjon som lunger hos mennesker?

_____

Oppgave B30

Lufta består av mange gasser.

Hvilken gass er det mest av?

A Nitrogen

B Oksygen

C Karbondioksid

D Hydrogen

Oppgave åA19

Hva er menstruasjon?

_____

_____

_____

Oppgave fA28

Fossilt brennstoff er dannet av

A uran

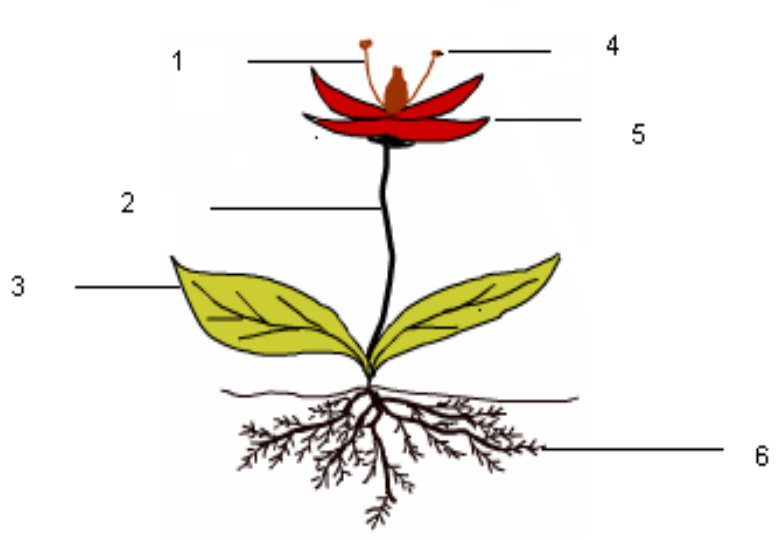
B sjøvann

C sand og grus

D døde planter og dyr

Oppgave åA7

Sett navn på de seks delene av planten som er merket av.



Oppgave åA36

Hvorfor er vann som koker ***ikke* et eksempel på en kjemisk reaksjon?**

_____

_____

Oppgave åB17

Hvilke oppgaver har følgende organer i kroppen vår?

Hjertet: _____

Lungene:_____

Huden:_____

Skjelettet: _____

Vedlegg 4 Etablering av gruppe G1 og gruppe G2 etter Test 1

Gutter:

	Etablering av par av gutter på skole 1 for Test 2				Etablering av par av gutter på skole 2 for Test 2				
	Gutter Skole 1	Poengsum på Test 1	Poengsum holdninger	Endelige par i Test 2	Gutter Skole 2	Poengsum på Test 1	Poengsum holdninger	Endelige par i Test 2	
NIVÅ 3	E	20	10	G1 1g1	A	19	11	G1 2g1	NIVÅ 3
	E	19,5	13	G2 1g1	B	19	15	G1 2g1*	
	D	18	13	G1 1g2	A	20	11	G2 2g1	
	F	18	11	G2 1g2	B	18,5	10	G1 2g2	
	E	17,5	10	G1 1g3	A	18,5	14	G2 2g2	
	F	17	9	G2 1g3	C	18,5	11	G1 2g3	
	E	17,5	13	G1 1g4	C	18,5	12	G2 2g3	
	F	17,5	12	G2 1g4	C	18,5	11	G1 2g4	
	A	17,5	7	G1 1g5	C	18	15	G2 2g4	
	E	17,5	18	G2 1g5	C	18	11	G1 2g5	
	c	17	5	G1 1g6	C	18	13	G2 2g5	
	D	17	17	G2 1g6	A	17,5	14	G1 2g6	
	D	17	17	G1 1g7	A	16,5	14	G2 2g6	
	D	17	18	G2 1g7	B	17,5	12	G1 2g7	
	B	16,5	16	G1 1g8	A	17	11	G2 2g7	
	B	16,5	16	G2 1g8	C	17	12	G1 2g8	
	B	16,5	11	G1 1g9	C	17	12	G2 2g8	
	B	16,5	10	G2 1g9	A	16,5	15	G1 2g9	
					B	16,5	15	G2 2g9	
NIVÅ 2	A	16	10	G1 1g10	B	16,5	18	G1 2g10	NIVÅ 2
	c	16	11	G2 1g10	A	16,5	18	G2 2g10	
	D	16	15	G1 1g11	C	16,5	10	G1 2g11	
	D	15,5	16	G2 1g11	C	16,5	14	G2 2g11	
	B	15,5	12	G1 1g12					
	D	15,5	12	G2 1g12	C	16	13	G1 2g12	
	A	15	11	G1 1g13	C	15,5	12	G2 2g12	
	B	14,5	10	G2 1g13	A	15,5	11	G1 2g13	
	B	15	14	G1 1g14	B	15	9	G2 2g13	
	E	15	13	G2 1g14	B	15,5	16	G1 2g14	
	B	15	16	G1 1g15	B	15,5	15	G2 2g14	
	F	15	17	G2 1g15	B	15,5	13	G1 2g15	
	D	15	22	G2 1g15*	B	15,5	13	G2 2g15	
	A	14	10	G1 1g16	A	15	16	G1 2g16	
	A	14	10	G2 1g16	A	15	15	G2 2g16	
	c	14	13	G1 1g17	A	14,5	13	G1 2g17	
	F	13,5	12	G2 1g17	A	15	12	G2 2g17	
	B	13,5	12	G1 1g18	C	15	19	G1 2g18	
	A	13	15	G2 1g18	C	15	19	G2 2g18	
	D	13,5	19	G1 1g19	C	14,5	15	G1 2g19	
	D	13	22	G2 1g19	B	14,5	16	G2 2g19	
	F	13	13	G1 1g20	C	14	18	G1 2g20	
	F	13	13	G2 1g20	C	14,5	13	G2 2g20	

N	A	12,5	12	G1 1g21	B	14	14	G2 2g20*	**N**
	A	12,5	9	G2 1g21	B	12,5	14	G1 2g21	
I	A	12,5	12	G1 1g22	B	13,5	15	G2 2g21	**I**
	E	12,5	17	G2 1g22	C	13,5	21	G1 2g22	
V	A	12	13	G1 1g23	C	13,5	21	G2 2g22	**V**
	c	11,5	13	G2 1g23	B	12,5	22	G1 2g23	
Å	c	11,5	17	G1 1g24	A	13	21	G2 2g23	**Å**
	E	11	20	G2 1g24	A	13	19	G1 2g24	
	A	10,5	13	G1 1g25	A	13	18	G2 2g24	
	c	10,5	12	G2 1g25	A	12,5	14	G1 2g25	
2	F	10,5	14	G2 1g25*	C	12,5	14	G2 2g25	**2**
					A	12	15	G1 2g26	
	c	10	13	G1 1g26	A	12	15	G2 2g26	
	c	10	30	G2 1g26	C	11,5	17	G1 2g27	
	E	9	15	G1 1g27	B	11,5	15	G2 2g27	
	F	9,5	15	G2 1g27	**A**	**11**	**13**	**G1 2g28**	
N	A	9	16	G1 1g28	**B**	**11**	**12**	**G2 2g28**	
	A	9	17	G2 1g28	C	10,5	14	G1 2g29	
I	D	9	12	G1 1g29	C	10,5	19	G2 2g29	
	D	9	16	G2 1g29	B	11	21	G2 2g29*	
V	**E**	**9**	**21**	**G1 1g30**					
	E	**8**	**20**	**G2 1g30**	A	10	13	G1 2g30	**N**
Å	E	8,5	12	G1 1g31	C	10	13	G2 2g30*	
	c	8	11	G2 1g31	A	9	14	G2 2g30	**I**
	A	8	19	G1 1g32	A	8	19	G1 2g31	
1	A	8,5	15	G2 1g32	B	7	20	G2 2g31	**V**
	c	8,5	20	G1 1g33	B	7,5	18	G1 2g32	
	c	8,5	19	G2 1g33*	A	7,5	15	G1 2g32*	**Å**
	F	8,5	18	G2 1g33	B	8	16	G2 2g32	
	B	7,5	10	G1 1g34	A	6,5	16	G1 2g33	
	F	7	10	G1 1g34*	C	6,5	13	G2 2g33	**1**
	c	7	12	G2 1g34	B	6	20	G1 2g34	
	B	7,5	17	G1 1g35	B	5	20	G2 2g34	
	E	7,5	17	G2 1g35	A	1,5	17	G1 2g35	
	B	6,5	15	G1 1g36					
	F	6	14	G1 1g36*					
	B	5	17	G2 1g36					
	A	6	22	G1 1g37					
	F	5,5	19	G2 1g37					

Figur 1

G1 og G2 betyr ”gruppe 1” og ”gruppe 2”. Elevene i G1 (blå farge) fikk prøve 1 i Test 2 og elevene i G2 (gul farge) fikk prøve 2. Elevene G1 **1**g30 og G2 **1**g30 er gutter som går på **skole 1** og utgjør par nummer 30. Den ene gutten var i gruppe G1 og den andre i gruppe G2, og fikk ulike prøver. Ut fra resultatene på Test 1 var de på nivå 1. G1 **2**g28 og G2 **2**g28 er to gutter som utgjør par nummer 28 og går på **skole 2**. Disse elevene var på nivå 2 i Test 1.

Jenter:

	Etablering av par av jenter på skole 1 for Test 2				Etablering av par av jenter på skole 2 for Test 2				
	Jenter Skole 1	Poengsum på Test 1	Poengsum holdninger	Endelige par i Test 2	Jenter Skole 2	Poengsum på Test 1	Poengsum holdninger	Endelige par i Test 2	
NIVÅ 3	C	18	16	G1 1j1	C	20	15	G1 2j1	NIVÅ 3
	D	20	13	G1 1j1	A	18,5	17	G2 2j1	
	B	18,5	12	G2 1j1	A	19	7	G1 2j2	
	D	18,5	20	G1 1j2	C	19,5	9	G2 2j2	
	E	18	22	G2 1j2	C	18	13	G1 2j3	
	D	17	15	G1 1j3	B	18,5	16	G2 2j3	
	D	17	17	G2 1j3	B	17,5	11	G1 2j4	
	B	16	10	G1 1j4	C	18	12	G2 2j4	
	F	16	14	G1 1j4	A	17	16	G1 2j5	
	F	17	14	G2 1j4	A	16,5	16	G2 2j5	
	E	16,5	21	**G1 1j5**	C	16	10	G1 2j6	
	F	16,5	19	**G2 1j5**	C	16,5	9	G2 2j6	
	D	16	15	G1 1j6	B	16	16	G1 2j7	
	D	16	15	G2 1j6	A	15,5	14	G2 2j7	
	D	15,5	17	G1 1j7	A	16	15	G1 2j8	
					C	16	13	G2 2j8	
NIVÅ 2	E	15	18	G2 1j7	A	15,5	19	G1 2j9	
	F	14,5	12	G1 1j8	A	15,5	17	G2 2j9	
	C	15	14	G2 1j8	A	15,5	14	G1 2j10	
	B	14,5	11	G1 1j9	C	15,5	10	G2 2j10	
	B	14,5	13	G2 1j9					
	A	14,5	17	G1 1j10	A	14,5	14	G1 2j11	NIVÅ 2
	A	14	17	G2 1j10	A	15	15	G2 2j11	
	C	14	15	G1 1j11	C	15	22	G2 2j11	
	B	14	13	G2 1j11	A	14	9	G1 2j12	
	C	13,5	18	G1 1j12	A	14,5	12	G2 2j12	
	E	14	17	G2 1j12	C	14,5	14	G1 2j13	
	E	14	21	G1 1j13	C	14,5	13	G2 2j13	
	C	13,5	21	G2 1j13	B	14	17	**G1 2j14**	
	F	13,5	13	G1 1j14	B	14	10	**G2 2j14**	
	D	13	12	G2 1j14	A	13,5	13	G1 2j15	
	A	12,5	20	G1 1j15	A	13,5	11	G2 2j15	
	E	13	19	G2 1j15	B	13,5	14	G1 2j16	
	B	12,5	12	G1 1j16	A	13	16	G2 2j16	
	C	12,5	12	G2 1j16	B	13	15	G1 2j17	
	D	12,5	15	G1 1j17	B	13	16	G2 2j17	
	E	12,5	18	G2 1j17	A	12,5	18	G1 2j18	
	B	12,5	18	G1 1j18	A	12,5	19	G2 2j18	
	D	12,5	19	G2 1j18	B	12,5	19	G1 2j19	
	B	11,5	17	G1 1j19	B	12,5	15	G2 2j19	
	D	12	17	G2 1j19	C	12,5	16	G1 2j20	
	E	11,5	15	G2 1j19*	B	12	13	G2 2j20	
	E	12	22	G1 1j20	B	12	18	G1 2j21	

NIVÅ 2	F	12	21	G2 1j20	B	12	16	G2 2j21	
	B	11,5	21	G1 1j21	B	11,5	19	G1 2j22	
	F	12	20	G2 1j21	C	12	18	G2 2j22	
	A	11	20	G1 1j22	A	11,5	14	G1 2j23	
	A	11	11	G2 1j22	A	10,5	15	G2 2j23	
	B	10,5	19	G1 1j23					
	A	11	21	G2 1j23	A	10	17	G1 2j24	NIVÅ 1
	B	11	16	G1 1j24	B	9,5	18	G2 2j24	
	D	11	14	G2 1j24	C	9,5	11	G1 2j25	
	D	10,5	8	G1 1j25	C	9,5	14	G2 2j25	
	F	10,5	20	G1 1j26	C	9,5	16	G2 2j25*	
	F	10,5	22	G2 1j26	C	9	19	G1 2j26	
					A	9	19	G2 2j26	
NIVÅ 1	F	10	15	G2 1j25	B	9	10	G1 2j27	
	B	10	16	G1 1j27	C	9	13	G2 2j27	
	D	10	19	G2 1j27	C	7	21	G1 2j28	
	A	10	21	G1 1j28	B	4	19	G2 2j29	
	C	10	22	G2 1j28					
	C	9,5	18	G1 1j29					
	E	9,5	17	G2 1j29					
	E	9,5	22	G1 1j30					
	C	8,5	21	G2 1j30					
	A	9	15	G1 1j31					
	F	8,5	15	G2 1j31					
	D	8,5	17	G1 1j31					
	E	9	20	G1 1j33					
	C	8	20	G1 1j33					
	A	8,5	20	G2 1j33					
	B	7,5	15	G1 1j34					
	C	8	20	G2 1j34					
	E	7	23	G1 1j35					
	F	8	22	G2 1j35					
	E	7	10	G1 1j36					
	B	7	19	G2 1j36					
	A	6,5	16	G1 1j37					
	C	6,5	17	G2 1j37					
	B	6,5	21	G2 1j37					
	A	5	19	G2 1j38					
	C	5,5	27	G2 1j39					
	A	3,5	21	G1 1j40					
	F	3,5	17	G2 1j40					
	D	1,5	5	G1 1j41					

Figur 2

G1 og G2 betyr ”gruppe 1” og ”gruppe 2”. Elevene i G1 (blå farge) fikk prøve 1 i Test 2 og elevene i G2 (gul farge) fikk prøve 2. Elevene G1 **1j**5 og G2 **1j**5 er jenter som går på **skole 1** og utgjør par nummer 5. Den ene jenta var i gruppe G1 og den andre i gruppe G2, og de fikk ulike prøver. Ut fra resultatene på Test 1 var de på nivå 3. G1 **2j14** og G2 **2j14** er to jenter som utgjør par nummer 14 og går på **skole 2**. Disse elevene var på nivå 2 i Test 1.

Tabell 1

Inndeling av guttene på nivå

Nivå	Poeng	Gutter skole 1			Gutter skole 2			
		Antall	Antall par	Antall trippel	Antall	Antall par	Antall trippel	Antall singel
Nivå 3	20 ≥ X > 16	18	9		23	10	1	
Nivå 2	16 ≥ X > 10	34	14	2	38	16	2	
Nivå 1	X ≤ 10	27	9	3	13	3	2	1
Sum		79			74			

Tabell 2

Inndeling av jentene på nivå

Nivå	Poeng	Jenter skole 1				Jenter skole 2			
		Antall	Antall par	Antall trippel	Antall singel	Antall	Antall par	Antall trippel	Antall singel
Nivå 3	20 ≥ X > 15,25	15	4	2	1	20	10		
Nivå 2	15,25 ≥ X > 10	39	17	1	2	27	12	1	
Nivå 1	X ≤ 10	29	8	3	4	11	3	1	2
Sum		83				58			