Konstruksjon av flervalgstester for måling av biologikompetanse: Bruk av moderne testteori til analyse og validering av flervalgsoppgaver

Masteroppgave i biologididaktikk

av

Ruben Jørstad og Rakel Notevarp Paulsen

Institutt for biologi

Universitetet i Bergen

Juni 2015

Forord

Denne masteroppgaven har tatt oss med på en spennende, lærerik og tidvis krevende reise. For oss har det mest interessante vært å lære om konstruksjon og analyse av flervalgsoppgaver. Dessuten har det vært lærerikt å repetere biologiske fagkunnskaper. Vi tror vi kommer til å ha stor nytte av kunnskapene vi har fått gjennom arbeidet med denne masteroppgaven.

Det er mange personer som fortjener en takk i forbindelse med denne masteroppgaven. Først og fremst vil vi takke vår veileder, førsteamanuensis Tom Olav Klepaker for at vi fikk gjennomføre dette prosjektet. Takk for all veiledning, pågangsmot og støtte!

Dernest vil vi rette en stor takk til Torbjørn Torsheim, som har vært til stor hjelp i forbindelse med dataoppsettet og analysene i R. Takk for god kunnskap og innsikt!

Vi vil også takke bioCEED for økonomisk støtte og teknisk hjelp i forbindelse med SurveyXact, Brage Førland for god teknisk hjelp, alle institusjonene som hjalp oss med å distribuere flervalgstesten, alle våre pretestere og alle studenter og ansatte på BIO som på en eller annen måte har hatt noe med prosjektet å gjøre.

Så vil vi takke våre familier, venner og kjente. Dere har vært viktige støttespillere for oss gjennom hele prosessen.

Til slutt vil vi takke hverandre. Vi ser begge frem til å avslutte arbeidet og å begynne en ny, enda mer spennende fase i livet: Yrkeslivet!

Våren 2015

Rakel Notevarp Paulsen og Ruben Jørstad

Sammendrag

Utdanningssystemets hovedformål er å legge til rette for gode og effektive utdanningstilbud som optimaliserer studenters læringsmiljø og maksimerer deres læringsutbytte. *Vurdering av læring* gir innsikt i studenters kompetanse og *vurdering for læring* gir innsikt i hvordan opplæringen kan tilpasses enhver studentgruppe. For å gjennomføre kvalitetsfremmende tiltak i utdanningen, kreves det at man har gode metoder for å måle det som er målet med tiltakene, studentenes læringsutbytte. Denne oppgaven har som mål å utvikle og kvalitetssikre et verktøy som kan gi kunnskap om biologistudenters faglige basiskunnskaper. Verktøyet baserer seg på testing med lukkede flervalgsoppgaver. Den teoretiske forankringen baserer seg på litteratur om flervalgsoppgaver med hovedfokus på konstruksjon av gode flervalgsoppgaver. Det ble også fremlagt teori om analyser av flervalgsoppgaver ved hjelp av moderne testteori.

Verktøyet utviklet i denne masteroppgaven var en flervalgstest som besto av 290 innsamlede og egenkonstruerte flervalgsoppgaver som testet grunnleggende kompetanse i biologi. Disse flervalgsoppgavene ble distribuert til biologistudenter ved universiteter og høyskoler rundt om i Norge, og dataene fra 713 respondenter ble analysert ved bruk av moderne testteori ("item response theory"). Modellvalganalyser viste at den endimensjonale Rasch-modellen var best tilpasset dataene. Flervalgsoppgavenes diskrimineringsevne, dimensjonalitet og lokal uavhengighet ble beregnet og evaluert, og oppgaver som diskriminerte dårlig mellom respondenter, viste seg å være utilpasset Rasch-modellen eller var lokalt avhengige ble fjernet fra oppgavebanken. Gjennom analysene ble 51 oppgaver fjernet og det endelige produktet var en oppgavebank med 239 flervalgsoppgaver.

Verktøyet som ble utviklet hadde god innholdsvaliditet, men var mindre dekkende for studenter med svært lav og svært høy dyktighet. Anbefalinger for videre bruk av verktøyet er gitt i oppgaven.

Abstract

The main purpose of the educational system is to provide a good and effective education that optimizes students' teaching environments and maximizes their learning outcome. Summative evaluation gives insight into students' level of expertise while formative evaluation gives insight into how one can optimize the education for one particular group of students. In order to make quality promoting measures in education there is need for a tool that measures the prime goal of education: The students' learning outcome. The main purpose of this master's thesis is to develop and quality check a tool which can bring forth knowledge about biology students' level of expertise in biology. The theoretical framework is based on literature on multiple choice questions, with a main focus on how to construct good multiple choice questions. Literature on how to analyze multiple-choice questions by modern test theory is also presented.

The tool developed in this master's thesis was a multiple-choice test, that consisted of 290 collected and constructed multiple-choice questions that measured students' level of expertise in biology. These multiple-choice questions were distributed to biology students at universities and university collages around Norway, and the data from 713 respondents were analyzed using modern test theory ("item response theory"). Model-choice analyses showed that the one-dimensional Rasch model fitted the data best. The multiple-choice questions' discrimination ability, dimensionality and local independence were calculated and evaluated, and multiple-choice questions that either discriminated poorly, didn't' fit the Rasch model or were locally dependent were moved from the item pool. 51 multiple-choice questions were removed from the item pool and the final product was an item pool that consisted of 239 multiple choice questions.

The tool developed in this master's thesis had a good content validity, but weren't as reliable for students with much higher or lower abilities. Recommendations for further use of this tool are given in the thesis.

Innholdsfortegnelse

Kapittel 1 – Innledning

1.1 Bakgrunn for oppgaven

Biologi er et fagområde som stadig er i utvikling. Utviklinger i vitenskapsfaget biologi har også betydning for studiefaget biologi, noe som skaper nye behov for innholdet i biologiundervisningen og for hvordan utdanningen av fremtidige biologer foregår (bioCEED). På grunn av dette er det viktig med kunnskap om undervisning og læring. Dette gjelder på alle utdanningsnivåer, men kanskje spesielt i høyere utdanning hvor detaljnivået i undervisningen generelt sett er ganske høyt.

I den forbindelse er det viktig å kunne studere studenters kunnskaper for å kunne utvikle undervisningen for å øke studentenes læringsutbytte. Ved å få kunnskap om studentenes læring og kompetanse kan utdanningsinstitusjoner utvikle sitt utdanningstilbud for ulike studentgrupper. Å få kunnskap om studentenes læringskurver og progresjonen underveis i et studium vil være verdifullt for å se hvordan undervisningen på studiet fungerer. Men for å få kunnskap om studentenes faglige kompetanse og læringsprogresjon kreves det at man har et kvalitetssikret verktøy som kan brukes til å måle dette.

Et slikt verktøy vil også kunne komme til gode når institusjoner skal gjennomføre vurdering for læring. I grunnskolen og den videregående skolen har vurdering for læring blitt godt innarbeidet. Vurdering for læring blir definert som all vurdering som gis underveis i opplæringen og som bidrar til å fremme læring. I skolen innebærer dette at vurdering av prestasjoner, arbeidsprosesser eller oppgaver skal brukes til å justere egen læring eller undervisningsopplegg underveis (Utdanningsdirektoratet, 2014). Vurdering for læring kalles derfor også underveisvurdering. I høyere utdanning finner man ofte underveisvurderinger i form av tester eller innleveringer underveis i semesteret, der resultatet gjerne teller inn på sluttkarakteren. Denne masteroppgaven har til formål å utvikle et slikt verktøy som kan brukes til å måle faglig kompetanse innenfor biologi.

1.2 Vitenskapsfaget biologi og biologisk kompetanse

Biologi kan defineres som *læren om liv eller levende organismer* (Marion & Strømme, 2008). Biologi er et fagområde vi har tilknytning til hver eneste dag, siden det omfatter alt fra oss

selv som levende organismer, naturen rundt oss, maten vi spiser og prosessene som foregår i naturen og som er nødvendige for at vi skal leve. Gjennom skoleverket blir biologi først introdusert gjennom faget naturfag, og videre i videregående opplæring kan man ta biologi som valgt programfag. Biologi i videregående opplæring er delt opp i ulike hovedområder, som for eksempel "den unge biologen", "fysiologien til mennesket" og "økologi". Alle hovedområdene har kompetansemål der det er beskrevet hvilke kompetanser elevene skal ha. Verb som for eksempel "å diskutere" og "forklare" presiserer på hvilken måte eleven skal kunne bruke kunnskapen (Utdanningsdirektoratet, 2006). Dette videreføres hos universitetene der ulike verb beskriver kunnskapen og kompetansen studentene skal inneha etter et spesifikt emne er ferdig (UiB).

Siden fagområdet biologi er såpass bredt, vil det også være vanskelig å gi en spesifikk definisjon av hva biologikompetanse er. Dette vil også være helt avhengig av om man snakker om grunnskolen, den videregående skolen eller høyere utdanning. Innenfor grunnskolen og den videregående skolen er det læreplanen som gir retningslinjene for hvilken kompetanse elevene skal sitte med, gjennom bruk av kompetansemålene. De grunnleggende ferdighetene *å kunne uttrykke seg muntlig og skriftlig i biologi*, *å kunne lese i biologi*, *å kunne regne i biologi* og *å kunne bruke digitale verktøy i biologi* er integrert gjennom kompetansemålene (Utdanningsdirektoratet, 2006).

Ved høyere utdanning har man også mål for læringsutbyttet som studenter skal ha etter endt studium. Ved Institutt for biologi, Universitetet i Bergen (Institutt for biologi) har man formulert læringsutbyttet som bachelorprogrammet i biologi skal gi studentene:

Studiet skal gi en bred plattform i naturfag og biologi fra molekylære til evolusjonære prosesser. Gjennom en grunnleggende forståelse av basale prosesser i naturen skal studenten kunne tilegne seg og bruke vitenskapelig kunnskap og innsikt i en rekke samfunnsrelevante utfordringer som omfatter naturmiljøet.

Dette er generelt og lite spesifikt. Derfor konkretiseres dette i følgende kompetansemål:

- *ha god kjennskap til moderne biologi og biologiens relevans i samfunnet*
- *ha en bred og anvendelig naturfaglig bakgrunn*
- *være i stand til å bruke utviklingslæra som en nøkkel til å forstå organismene sine tilpasninger*

- *kunne lese og forstå vitenskapelige arbeid om aktuelle miljøspørsmål*
- *besitte grunnleggende kunnskaper om virkemåten til organismer og økosystem*
- *ha ferdigheter i naturvitenskapelig metode som gjør kandidaten i stand til å sette seg inn i nye problemstillinger, skrive analyserende rapporter og vurdere hvor sikker kunnskap er*
- *evne å løse problemer og oppgaver som krever grunnleggende kunnskap om naturen*

Kompetansemålene over inneholder elementer av både kunnskap og ferdigheter, og ettersom biologisk kunnskap er blitt svært omfattende, vil man først konsentrere seg om grunnprinsipper, for deretter å spesialisere seg innen ulike retninger (Marion & Strømme, 2008).

1.3 Fagområder

Biologi er et omfattende fag, og kan deles inn i ulike fagområder, og på ulike måter. Man kan dele fagområdene opp etter *hvilke* organismegrupper som studeres, og får da områder som botanikk, marinbiologi, mikrobiologi og zoologi som hovedområder. Man kan også dele opp biologifaget etter *hva* som studeres ved disse organismene, og får områder som for eksempel evolusjonsbiologi, genetikk, systematikk og økologi. Noen områder grenser også til andre fag innenfor naturvitenskapen, som for eksempel biokjemi og molekylærbiologi som ligger mellom biologi og kjemi (Wikipedia).

De obligatoriske emnene i bachelorprogrammet i biologi ved UiB er også med på å definere hva som kan betegnes som grunnleggende biologisk kompetanse, og hva som trengs av bakgrunnskunnskap. For eksempel er grunnemner i disiplinene kjemi, fysikk og statistikk en del av grunnkunnskapen som skal være på plass. Derfor er disse fagene lagt inn som obligatoriske emner. De obligatoriske grunnemnene i biologi hos UiB er

- Innføring i evolusjon og økologi
- Organismebiologi 1
- Organismebiologi 2
- Cellebiologi og genetikk
- Komparativ fysiologi
- Innføring i molekylærbiologi

Noen av emnene innebærer laboratorieøvelser og feltarbeid, med føring av rapporter og planlegging og gjennomføring av prosjekter.

1.4 Vurdering

Biologi er et fag med mange underområder og ulike arbeidsformer. Dette åpner opp for et bredt utvalgt av metoder for å vurdere elever eller studenter. Metoden som brukes vil være avhengig av hensikten med vurderingen og innholdet i det som skal vurderes (Imsen, 2009). Vurdering har en sterk tradisjon som summativ vurdering (vurdering *av* læring), og en svakere tradisjon som formativ vurdering (vurdering *for* læring), men de siste årene har det vært et større fokus formativ vurdering, spesielt i skoleverket gjennom programmet Vurdering for læring (2010-2014) (Hopfenbeck & Lillejord, 2013). En skiller også mellom formelle og uformelle metoder for vurdering, der eksempler på formelle metoder i skolen er ulike former for spørreskjemaer, tester, lærerlagde prøver, tentamensprøver og eksamensprøver. Uformelle metoder kan være loggføring, observasjon, samtale og intervju (Imsen, 2009).

Bredden i biologifaget kommer også til uttrykk i eksamensformen i videregående skole, der en i biologi 1 kan bli trukket ut til muntlig-praktisk eksamen, og i biologi 2 kan man bli trukket ut til skriftlig eller muntlig-praktisk eksamen (Utdanningsdirektoratet, 2006). Underveis i skoleåret gjennomføres det prøver, innlevering av lab-journaler, fremføring, prosjekter og uformell vurdering. Prøver kan ha ulike former, for eksempel åpne spørsmål, lukkede spørsmål eller flervalgsoppgaver. På universitetsnivå finnes det også ulike former for vurdering. Underveis i fagene vil det være vurdering av lab-journaler, rapporter, prosjektoppgaver og muntlig fremleggelse av arbeid. Eksamensformen kan variere mellom muntlig og skriftlig, der skriftlig eksamen også her har ulike former for oppgaver, for eksempel flervalgsoppgaver.

1.5 Flervalgsoppgaver som verktøy for å måle kompetanse

Flervalgsoppgaver ble først tatt i bruk på begynnelsen av 1900-tallet som vurderingsverktøy og ble etter hvert svært utbredt i USA, spesielt etter at det på 1950-tallet kom maskiner som kunne utføre automatiske avlesninger og analyser av svarskjemaer. I Norge har ikke bruken

av flervalgsoppgaver vært like utbredt, men i de siste årene har debatten om kvalitet i skolen ført til bruk av nasjonale prøver og komparative undersøkelser som innebærer blant annet flervalgsoppgaver (Sirnes, 2005). Internasjonale undersøkelser som PISA og TIMSS bruker også flervalgsoppgaver som en del av sine undersøkelser (Universitetet i Oslo). Sist, men ikke minst, så er flervalgsoppgaver som en del av eksamen blitt et vanlig innslag i sluttvurderingen i videregående og høyere utdanning. I biologi 2 i den videregående skolen er for eksempel flervalgsoppgaver en del av eksamen, i den delen som skal gjennomføres uten hjelpemidler (Utdanningsdirektoratet, 2015). Websider som It's Learning, Kahoot og Socrative er eksempler på hyppig brukte verktøy som kan brukes til å lage og distribuere flervalgstester på en enkel, morsom og praktisk måte.

Det å utvikle en flervalgstest innebærer flere faser: Første fase er *testdesign*, det å lage spørsmål med høy kvalitet og bestemme designen på testen, for eksempel antall oppgaver og tilbakemelding respondentene får underveis. Andre fase er *distribuering* av testen, for eksempel hvordan respondentene kan ta testen, antall forsøk og tidsperspektiv. Tredje fase involverer det å lage *score* til testen, og eventuell karaktersetting. Siste fase er *forbedring* av testen. Resultatene må analyseres, der dårlige oppgaver og uegnede svaralternativ kan identifiseres og revideres, eller fjernes fra testen (Horgen, 2007).

1.6 Problemstilling og rammer for oppgaven

Målet for denne masteroppgaven er å utvikle et verktøy som kan teste studenters kompetanse innenfor biologi, på et nivå tilsvarende et bachelorstudium i biologi. Verktøyet skal bestå av en bank med flervalgsoppgaver, der de som skal bruke dette verktøyet kan sette sammen delsett med oppgaver som er egnet for det som skal testes. Ettersom biologi er et omfattende fag var det nødvendig å sette noen rammer for oppgaven. Innholdsmessig så vil flervalgstesten være begrenset til lærestoff som ligger på bachelornivå. Ikke alle kompetanser innenfor biologi kan testes ved hjelp av flervalgsoppgaver, og kompetanser som innebærer praktisk arbeid som laboratoriearbeid og feltarbeid, rapportskriving og lesing og skriving av oppgaver eller artikler, vil ikke bli testet gjennom flervalgstesten. Verktøyet som skal bli utviklet skal kunne brukes i forskning på læring, utbytte av undervisning og studenters progresjon gjennom studiet. Denne masteroppgaven vil ta for seg hele prosessen med innsamling og konstruksjon av oppgaver, innsamling av data gjennom å distribuere

oppgavene til studenter og analyse av dataene. Til slutt vil vi vurdere reliabiliteten og validiteten til flervalgstesten og gi anbefalinger om hvordan disse oppgavene kan brukes videre.

Hovedmålet med oppgaven er å utvikle verktøyet. Men som et lite eksempel på hvordan et slikt verktøy kan brukes, ønsker vi å teste hypotesen om at den biologiske basiskunnskapen øker gjennom studiet. Dette gjør vi ved å se på sammenhengen mellom hvor mange studiepoeng studenten har oppgitt å ha avlagt mot hvordan han/hun presterer på testen. . Dette vil også kunne gi et bilde på hvordan verktøyet fungerer, ettersom vi har en antakelse om «studier lønner seg»; at andelen korrekte svar vil stige med økt antall studiepoeng.

Kapittel 2 – Teori

2.1 Flervalgsoppgaver

2.1.1 Elementene i en flervalgsoppgave

Flervalgsoppgaver består tradisjonelt sett av en lukket oppgave etterfulgt av flere svaralternativer. Oppgaveteksten kalles *stammen* og inneholder selve problemstillingen i en flervalgsoppgave. Det riktige svaralternativet kalles *nøkkelen* mens de gale svaralternativene kalles *distraktorer*. Distraktorene er ment å distrahere respondenter som er usikre på hvilket svaralternativ som er nøkkelen. Før stammen kan flervalgsoppgaver inneholde en *stimulus*, visuell eller auditiv informasjon som er nødvendig for å løse oppgaven, og en *orientering* om hvor informasjonen er hentet fra. Hverken stimulus eller orientering er obligatoriske deler av flervalgsoppgaver (Sirnes, 2005).

Boks 2.1: *Elementene i en flervalgsoppgave*

Fra en lærebok i evolusjonsbiologi: *(Orientering)*

En populasjon med villblomster som er i Hardy-Weinberg-likevekt har ett gen med to alleler, A1 og A2. Tester viser at 70 % av pollen som blir produsert i populasjonen inneholder A1-allelet. *(Stimulus)*

Hvor stor andel av villblomstene i denne populasjonen er heterozygoter? *(Stamme)*

A. 0.21 *(Distraktor)*
B. 0.42 *(Nøkkel)*
C. 0.49 *(Distraktor)*
D. 0.70 *(Distraktor)*

Flervalgsoppgaver kalles også *testledd*. Det engelske ordet item er også *vanlig*.

2.1.2 Ulike typer flervalgsoppgaver

Flervalgsoppgaver kan ta mange former. Haladyna *et al.* (2002) beskrev 6 typer flervalgsoppgaver som her er omkategorisert til 4 kategorier.

Konvensjonelle flervalgsoppgaver

Vanligvis er flervalgsoppgaver formulert enten som et spørsmål eller som en ufullstendig setning. Begge oppgaveformene kan innebære instrukser som går ut på å finne det *beste* svaret (Sirnes, 2005). Konvensjonelle flervalgsoppgaver har som regel mellom tre og fem svaralternativer. Fordi et spørsmål fremstiller den sentrale ideen i en oppgave på en mer

direkte måte enn en ufullstendig setning er spørsmålsvarianten å foretrekke (Haladyna *et al.*, 2002):

Boks 2.2: *Eksempel på konvensjonell flervalgsoppgave formulert som et spørsmål*

Hva er sluttproduktet i glykolysen?

A. NADH
B. pyruvat
C. acetyl
D. laktat
E. ATP

Ufullstendige setninger er likevel vanlige og bidrar til mer variasjon i en flervalgstest:

Boks 2.3: *Eksempel på konvensjonell flervalgsoppgave formulert som en ufullstendig setning*

En nonsense-mutasjon fører som oftest til

A. et forbedret protein.
B. et unormalt protein som medfører sykdom.
C. et ikke-funksjonelt protein.
D. ingen endring i proteiner.

Sant-usant-flervalgsoppgaver

I denne oppgavetypen skal respondenter vurdere om en eller flere påstander er sanne eller usanne:

Boks 2.4: *Eksempel på sant-usant-flervalgsoppgave*

Avgjør om påstandene nedenfor er sanne eller usanne:

		Sant	Usant
1.	Furutrær er tofrøbladete planter.	☐	☐
2.	Pattedyrs ekskresjonsprodukt er urinsyre.	☐	☐
3.	H_2O er IKKE en drivhusgass.	☐	☐

Sant-usant-oppgaver er vanlige og populære på lavere nivåer, men ikke fullt så vanlige i psykometriske tester (Haladyna *et al.*, 2002). Til tross for at den er godt likt så finnes det noen ulemper ved denne oppgaveformen. For det første så gjør to svaralternativer at respondentene har 50 % sjanse for å tippe riktig svar. For det andre, selv om sant-usant-oppgaver kan være lettere å lage enn konvensjonelle flervalgsoppgaver, så kan det være vanskelig å lage sant-usant-oppgaver som tester kunnskap på høyere kognitive nivåer. For det tredje så gir sant-

usant-oppgaver ingen diagnostisk informasjon (Sirnes, 2005). Besvarelsen "usant" på påstand nummer 2 over sier for eksempel lite om respondenter vet hva ekskresjonsproduktet til pattedyr faktisk er. Sant-usant-oppgaver kan likevel bidra til å skape variasjon i flervalgstester.

Kombinasjonsflervalgsoppgaver

Dette oppgaveformatet innebærer en stamme etterfulgt av ett sett med enheter, gjerne ord eller setninger, hvor målet er å koble (*matche*) to og to enheter, gjerne ved å trekke streker mellom dem:

Boks 2.5: *Eksempel på kombinasjonsflervalgsoppgave*

Kombiner hvert hormon til høyre med riktig produksjonskjertel til venstre:

bukspyttkjertel	adrenalin
magesekk	insulin
eggstokk	antidiuretisk hormon
hypofyse	gastrin
binyremarg	progesteron

Kombinasjonsoppgaver er vanlige på lavere nivåer, men brukes ikke spesielt mye i psykometriske tester. Kombinasjonsoppgaver er relativt lette å lage, besvare og vurdere (Haladyna *et al.*, 2002; Sirnes, 2005).

Kombinert respons-flervalgsoppgaver

I en kombinert-respons-oppgave blir stammen etterfulgt av flere ord eller setninger, hvor svaralternativene er satt opp som ulike kombinasjoner av disse ordene/setningene:

Boks 2.6: *Eksempel på kombinert-respons-flervalgsoppgave*

Hvilke av følgende anatomiske strukturer finnes i marine strålefinnefisker (for eksempel ørret)?
1. kloakk
2. spirakel
3. svømmeblære
4. sidelinjesystem
5. operculum

A. 1, 2, 4
B. 1, 2, 5
C. 2, 3, 4
D. 3, 4, 5

Kombinert-respons-oppgaver krever ofte mer tid å produsere, besvare og administrere enn vanlige flervalgsoppgaver. På den måten er de mindre effektive enn andre typer flervalgsoppgaver og bør benyttes deretter. Haladyna *et al*. (2002) frarådet derfor bruk av slike oppgaver. De kan allikevel bidra til å skape variasjon.

2.1.3 Konstruksjon av flervalgsoppgaver

Innhold

Når man konstruerer flervalgsoppgaver bør man bør ha klart for seg hva hver oppgave skal vurdere. Hver oppgave bør reflektere bestemte og viktige læringsmål, og trivielle fakta bør unngås. Innholdsmessig så bør flervalgsoppgaver hverken være for detaljerte eller for generelle, og innholdet i hver oppgave bør holdes uavhengig fra innholdet i de andre oppgavene i en og samme test. Språket bør være enkelt og forståelig, og man bør unngå lurespørsmål og formuleringer som på noen som helst måte kan betraktes som subjektive synspunkter (Haladyna *et al.*, 2002; Sirnes, 2005).

Sirnes (2005) beskriver noen punktumregler for flervalgsoppgaver: Det er vanlig å sette punktum etter svaralternativer formulert som setninger eller perioder, men ikke etter svaralternativer som består av enkeltord eller ordgrupper; I tilfeller hvor svaralternativene er ikke-avsluttede utsagn settes det punktum etter alle svaralternativene. Dersom svaralternativene består av ufullstendige setninger som svarer på et spørsmål settes det som oftest ikke punktum etter svaralternativene.

Stammen

Stammen i en flervalgsoppgave konstrueres gjerne etter at formålet med oppgaven (det oppgaven skal teste) er identifisert (Kehoe, 1995). Oppgaven bør være konstruert på en slik måte at den som leser stammen kan svare på oppgaven uten å lese svaralternativene. Det betyr at stammen bør inneholde nok tekstinformasjon til at den danner en konkret problemstilling (Brame, 2015; Haladyna *et al.*, 2002; Sirnes, 2005). Samtidig som at stammen skal være så presist og klart formulert som mulig, skal den ikke inneholde mer informasjon enn hva som er nødvendig for å løse oppgaven. Mesteparten av teksten i en oppgave bør ligge i stammen for å redusere lesetiden på hver oppgave (Brame, 2015; Haladyna *et al.*, 2002; Kehoe, 1995; Sirnes, 2005).

I konstruksjon av stammer bør man begrense bruken av negative formuleringer slik som *ikke*, *aldri*, *unntatt* og *med unntak av*, ettersom disse uttrykkene lett kan overses av respondenter. Slike uttrykk kan være med dersom de er nødvendige for å måle relevante læringsmål (Haladyna *et al.*, 2002), men dersom slike uttrykk er med bør de utheves: Ordene kan skrives med store bokstaver, være understreket eller skrevet i kursiv eller fet skrift (Brame, 2015; Haladyna *et al.*, 2002; Kehoe, 1995; Sirnes, 2005).

Svaralternativene

Det er ulike syn på hvor mange svaralternativer flervalgsoppgaver skal ha (Haladyna *et al.*, 2002; Kehoe, 1995). Å utvikle en tredje eller fjerde distraktor er ofte ikke verdt arbeidet dette innebærer. Det viser seg nemlig at ca. 67 % av alle flervalgsoppgaver i tester kun har én eller to effektive distraktorer og at bare 5 % av alle flervalgsoppgaver tre effektive distraktorer (Haladyna & Downing, 1993). Dyktige respondenter vil ofte enkelt snevre antall sannsynlige svaralternativer ned (Lord, 1977). Generelt sett så stiger tiden det tar å produsere, besvare og vurdere flervalgsoppgaver i takt med et økt antall svaralternativer (Ebel, 1982; Haladyna *et al.*, 2002; Rodriguez, 2005). Dermed kan to eller tre svaralternativer være vel så effektivt som fire eller fem. I tillegg så har flere studier vist at diskrimineringen mellom respondenter økes i takt med et redusert antall svaralternativer (Haladyna & Downing, 1993; Haladyna *et al.*, 2002). På den annen side så har andre studier vist at både vanskelighetsgraden og diskrimineringen reduseres når antall svaralternativer reduseres (Cizek *et al.*, 1988; Haladyna *et al.*, 2002). Dessuten innebærer færre svaralternativer en større sjanse for å gjette riktig (Sirnes, 2005). Når man konstruerer flervalgsoppgaver anbefales det derfor å holde seg til tre eller fire svaralternativer (Haladyna et al., 2002; Lord, 1977; Rodriguez, 2005).

Det er viktig at svaralternativene i en flervalgsoppgave samsvarer med stammen, både grammatisk og strukturelt. Svaralternativene bør være like i form og omtrent av samme omfang (Brame, 2015; Frary, 1995; Haladyna *et al.*, 2002; Kehoe, 1995; Sirnes, 2005). Svaralternativene bør være så korte og konsise som mulig, og i mange tilfeller kan svaralternativene forkortes dersom felleselementer i svaralternativene kan overføres til stammen (Brame, 2015; Kehoe, 1995; Sirnes, 2005):

Boks 2.7: *Dårlig konstruert flervalgsoppgave I*

Sykdommer overføres ofte fra en organisme til en annen via en tredjepart.[1] En organisme som overfører sykdommer fra en vert til en annen[2]

A. kalles[3] en transmitter
B. kalles[3] prioner[4]
C. kalles[3] en vektor
D. kalles[3] et virus

Kommentarer til boks 2.7:
1: Denne setningen er ikke nødvendig for å løse oppgaven.
2: Stammen byr ikke på et konkret problem.
3: "kalles" bør flyttes opp til stammen.
4: Dette svaralternativet byr på et grammatisk ulogisk problem, siden *en* organisme ikke kan være *flere* prioner.

Bedre:

Boks 2.8: *Dårlig konstruert flervalgsoppgave I forbedret*

En organisme som overfører sykdommer fra en vert til en annen kalles

A. en transmitter.
B. et prion.
C. en vektor.
D. et virus.

Ord og uttrykk som er felles for en stamme og dens tilhørende nøkkel kan gi respondenter hint om hvilket svaralternativ som er nøkkelen (Kehoe, 1995). Slike felleselementer bør derfor unngås. Av samme grunn bør man også unngå absolutte utsagn -svaralternativer som inneholder ord som "alltid", "aldri" og "absolutt" e.l. (Haladyna *et al.*, 2002), nøkler med for mye detaljer og svaralternativer som spiller på humor eller stereotypisk hverdagsspråk (Kehoe, 1995). Det er viktig at alle svaralternativene framstår som like sannsynlige og attraktive, og man bør ikke benytte seg av absurde svaralternativer eller svaralternativer som på andre måter fremstår som ulogiske. Man bør heller ikke benytte seg av svaralternativer som spiller på nesten like ord og uttrykk (Sirnes, 2005). Alle svaralternativene bør være uavhengig av hverandre og de må ikke være overlappende (Brame, 2015; Frary, 1995; Haladyna *et al.*, 2002):

Boks 2.9: *Dårlig konstruert flervalgsoppgave II*

Hvilket av følgende karaktertrekk er unikt for- og finnes i alle ryggstrengdyr[1]?
A. dorsal, hul nervestreng[1]
B. kjever og mineralisert skjelett[2]
C. ryggrad[2]
D. gjeller[3]

Kommentarer til boks 2.9:
1: "Ryggstrengdyr" og "nervestreng" har "streng" til felles.
2: Alternativene overlapper siden ryggrad går innunder mineralisert skjelett.
3: Siden gjeller er respirasjonsorganer for dyr som lever i vann så blir dette svaralternativet mer usannsynlig dersom respondentene vet at mange chordater lever på land.

Bedre:

Boks 2.10: *Dårlig konstruert flervalgsoppgave II forbedret*

Hvilket av følgende karaktertrekk er unikt for- og finnes i alle chordater?
A. dorsal, hul nervestreng
B. kjever og mineralisert skjelett
C. amniotiske egg
D. lunger eller lungeavstamninger

Dersom flere svaralternativer i en oppgave inneholder felleselementer, så bør svaralternativene i oppgaven kunne kobles på en homogen måte. Det vil si at svaralternativene bør være parallelle og at alle alternativene ligner på *ett* annet alternativ (Brame, 2015; Kehoe, 1995; Haladyna *et al.*, 2002; Sirnes, 2005):

Boks 2.11: *Dårlig konstruert flervalgsoppgave III*

Funksjonen til hormonet glukagon er å
A. [senke glukosenivået i blodet] ved å [hemme hydrolysen av glykogen].
B. [øke glukosenivået i blodet] ved å [fremme hydrolysen av glykogen].
C. [øke glukosenivået i blodet] ved å [hemme hydrolysen av glykogen].
D. [øke glukosenivået i blodet] ved å [stimulere insulinproduksjonen].

I oppgaven over inneholder svaralternativene elementer ([…]) fordelt på en tilsynelatende tilfeldig måte, uten noe form for system. Enkeltstående elementer eller elementer fordelt i par eller tripletter vil ofte utmerke seg og vil dermed kunne påvirke respondenters valg av svar (Brame, 2015; Kehoe, 1995; Haladyna *et al.*, 2002; Sirnes, 2005).

Bedre:

Boks 2.12: *Dårlig konstruert flervalgsoppgave III forbedret*

Funksjonen til hormonet glukagon er å
A. senke glukosenivået i blodet ved å fremme hydrolysen av glykogen.
B. øke glukosenivået i blodet ved å fremme hydrolysen av glykogen.
C. senke glukosenivået i blodet ved å stimulere insulinproduksjonen.
D. øke glukosenivået i blodet ved å stimulere insulinproduksjonen.

Den forbedrede versjonen inneholder en blanding av nesten de samme elementene fordelt på en homogen måte. I denne versjonen er elementene fordelt på en systematisk måte, og alle elementene går igjen to ganger. Dette gjør alle svaralternativene mer sannsynlige.

Man bør være kritisk til bruk av svaralternativer som omhandler flere av eller alle de andre svaralternativene, da respondenter kan benytte partiell kunnskap for å besvare disse spørsmålene (Brame, 2015). Partiell kunnskap vil i denne sammenhengen si at man ved å ha kunnskap om én ting kan ha en indirekte kunnskap om en annen ting. *Alle alternativene er riktige* (AAER) er dermed aldri et ønskelig svaralternativ, siden gjenkjenning av en distraktor eliminerer AAER som riktig svar og gjenkjenning av to korrekte svar identifiserer AAER som nøkkelen (Brame, 2015; Frary, 1995; Kehoe, 1995; Haladyna et al., 2002). Det samme gjelder for *ingen av alternativene er riktige* (IAAER), men dette svaralternativet kan benyttes så lenge det vurderes nøye. Studier viser at bruk av IAAER gjør oppgaver vanskeligere (Haladyna *et al.*, 2002), og IAAER kan egne seg godt som svaralternativ dersom en oppgave krever beregninger. IAAER bør selvsagt aldri benyttes etter en negativ stamme eller dersom respondentene blir bedt om å velge det beste svaret (Frary, 1995; Kehoe, 1995; Haladyna *et al.*, 2002).

Når stammen og svaralternativene er konstruert, er det viktig å sikre seg at det tenkte svaralternativet er korrekt eller klart best (Haladyna *et al.*, 2002). Det er fort gjort å gjøre feil, og det er ikke uvanlig at det oppstår uenigheter om hvilket svaralternativ som er nøkkelen (Sirnes, 2005).

2.1.4 Fra flervalgsoppgaver til flervalgstest

En flervalgstest er en samling av flervalgsoppgaver. Når man skal sette sammen en flervalgstest er det viktig å variere nøkkelens posisjon i de ulike flervalgsoppgavene på en tilfeldig måte. Det gjelder ikke dersom det finnes en hierarkisk struktur i svaralternativene (f.

eks. tall eller årstall); i så fall bør svaralternativene stå i den mest logiske rekkefølgen (f. eks. i numerisk eller kronologisk rekkefølge) (Sirnes, 2005; Brame, 2015; Haladyna *et al.*, 2002, Kehoe, 1995).

Ved produksjon av flervalgstester bør eksperter på området vurdere utforming av og formuleringer i oppgavene, dette øker validiteten til testene (DeVellis, 1991; Kehoe, 1995; Messick, 1989). Vurderingen bør baseres på både fagkunnskap og kunnskap om flervalgsoppgaver. Viktige spørsmål i denne sammenhengen kan for eksempel være om stammen er klar og konsis, om nøkkelen er et godt svar på oppgaven og om distraktorene framstår som like sannsynlige (Sirnes, 2005).

En del av utviklingen av flervalgstester kalles pilotering eller pre-testing. Piloteringen kan ha som formål å kartlegge tidsbruk, feil eller mangler ved testen, eller å avdekke om oppgavene ikke oppfyller de statistiske kravene som stilles. Ideelt sett bør man ha så mange pre-testere som mulig (Sirnes, 2005).

Det er vanlig at flervalgstester innledes med generell informasjon og veiledende instrukser til respondentene om hvordan testen skal besvares. Det er viktig at slike instruksjoner er nøye gjennomtenkt, konkrete og uttrykt i et presist språk (Sirnes, 2005).

2.1.5 En taksonomi for produksjon av flervalgsoppgaver

Boks 2.13 oppsummerer taksonomien for konstruksjon av flervalgstester (fritt etter Haladyna *et al*., 2002):

Boks 2.13: *Taksonomi for konstruksjon av flervalgsoppgaver og flervalgstester*

Innhold

1. Hver oppgave bør reflektere én bestemt kompetanse og én kognitiv prosess.
2. Baser hver oppgave på viktige læringsmål, unngå trivielt innhold.
3. Hold innholdet i hver oppgave uavhengig fra andre oppgaver i én og samme test.
4. Unngå overdetaljerte og overgeneraliserte oppgaver.
5. Unngå formuleringer som kan betraktes som objektive synspunkter.
6. Unngå lurespørsmål.
7. Hold språket på et enkelt nivå.

Format og stil

8. Bruk konvensjonelle flervalgsoppgaver, sant-usant-oppgaver og kombineringsoppgaver. Unngå kombinert-respons-oppgaver.
9. Bruk korrekte regler for gramatikk, stavelse, tegnsetting og små- og store bokstaver.
10. Minimer lesemengden i hver oppgave.

Stammen

11. Kontrollér at instruksene i stammen er entydig og klar.
12. Fremstill oppgavens sentrale idé i stammen i stedet for i svaralternativene.
13. Unngå overflødig informasjon i stammen.
14. Unngå negasjoner i stammen (slik som IKKE eller UNNTATT). Slike ord bør benyttes med forsiktighet, og de bør utheves dersom de blir benyttet.

Svaralternativene

15. Benytt mellom to og fem svaralternativer. Studier viser at tre svaralternativer er tilstrekkelig.
16. Bruk gjerne vanlige hverdagsforestillinger eller misoppfattelser som ditraktører.
17. Unngå å gi hint om hvilket svaralternativ som er nøkkelen, som for eksempel
    a. Absolutte utsagn (setninger med ord som "alltid", "aldri", "absolutt" e.l.)
    b. Svaralternativer som ligner på- eller er identiske med stammen.
    c. Grammatiske og strukturelle uoverensstemmelser.
    d. Overdetaljerte eller på annet vis iøyenfallende nøkler.
    e. Svaralternativer i par eller tripletter (heterogenitet).
    f. Absurde og latterlige svaralternativer.
18. Sørg for at alle svaralternativene er sannsynlige.
19. Bruk *ingen av alternativene er riktig* med forsiktighet.
20. Unngå *alle alternativene er riktig*.
21. Unngå negative ord slik som IKKE.
22. Hold lengden på svaralternativene nokså lik.
23. Hold svaralternativene uavhengige av hverandre. De bør ikke overlappe.
24. Hold svaralternativene homogene i innhold og grammatisk struktur.
25. Kontrollér at kun ett svar er riktig, eller at det tenkte svaret er klart best.
26. Varier nøkkelens posisjon.
27. Plasser svaralternativer i logiske eller nummeriske rekkefølger.

2.2 Matriseinnsamling av flervalgsoppgaver

Generelt kan man si at jo flere testledd en flervalgstest består av, jo mer innhold vil testen kunne dekke. Samtidig vil det ta lengre tid å gjennomføre testen og sannsynligheten for at alle respondentene gjennomfører alle testleddene reduseres (Childs & Jaciw, 2003). For å dekke et bredt innhold uten at testen blir for omfattende, er en metode å dele oppgavebanken inn i flere, mindre sett, hvor hvert sett distribueres til en andel av respondentene. En slik inndeling av testledd i mer enn ett sett kalles en matriseinnsamling (eng.: Matrix sampling) og har til hensikt å oppnå en bred innholdsdekning samtidig som tidsbruken per respondent minimeres (Aningbo, 2011; Childs & Jaciw, 2003). Selv om hver respondent kun besvarer et utvalg testledd, vil man ifølge Shoemaker & Shoemaker (1981) kunne bruke resultatene fra hvert sett til å beregne universale statistikker, som om man hadde distribuert alle testleddene til alle respondentene. På denne måten vil testleddsparametre kunne estimeres på en nøyaktig måte selv om alle testledd kun besvares av et utvalg respondenter.

Popham (1993) identifiserte to måter å dele testledd inn i sett på. Den første typen henviser til situasjoner hvor hver respondent får tildelt *ett sett* med testledd. Dette kan skje på to måter; med og uten *felleselementer* (Aningbo, 2011; Childs & Jaciw, 2003).

I en inndeling uten felleselementer (eng.: *Item sampling*), vil testleddene i et gitt sett kun bli besvart av respondentene som får det gitte settet tildelt:

Tabell 2.1: *Item-sampling-modellen for matrisesamling av flervalgsoppgaver*

Respondent	Sett 1	Sett 2	Sett 3	Sett 4
1	X			
2		X		
3			X	
4				X

I følge Aningbo (2011) er en slik inndeling tidsbesparende i tillegg til at den sikrer en god innholdsdekning. Ulempen er at hvert testledd bare besvares av et fåtall respondenter, og at sammenligning mellom de ulike settene blir svært krevende. En fordel med at hvert testledd kun besvares av et fåtall respondenter er at flere testledd kan testes ut i samme tidsrom.

En inndeling *med* felleselementer (eng.: *Partial matrix sampling*) er tilsvarende, men i tillegg vil alle respondentene få tildelt et sett som er felles, et sett som alle respondentene besvarer:

Tabell 2.2: *Partial-matrix-modellen for matrisesamling av flervalgsoppgaver*

Respondent	Felles	Sett 1	Sett 2	Sett 3	Sett 4
1	X	X			
2	X		X		
3	X			X	
4	X				X

En ulempe med dette oppsettet er også her at hvert testledd kun besvares av et fåtall respondenter. Også her er en fordel at flere testledd kan testes ut. I tillegg så fungerer fellesoppgaver i inndelinger lik den over som referansepunkter mellom de ulike settene og øker sammenlignbarheten mellom dem. (Dings, Childs & Kingston, 2002).

Pophams andre måte å dele testledd inn i sett på refererer til en måte hvor gitte respondenter blir tildelt *flere*, gitte sett (eng.: *Genuine matrix sampling*). Her besvares testleddene i et gitt sett av flere respondenter (Aningbo, 2011; Childs & Jaciw, 2003):

Tabell 2.3: *Genuine-matrix-modellen for matrisesamling av flervalgsoppgaver*

Respondent	Sett 1	Sett 2	Sett 3	Sett 4	Sett 5	Sett 6
1	X	X	X			

2			X	X	X		
3				X	X	X	
4					X	X	X
5	X					X	X
6	X	X					X

Fordelen med oppsettet over er at hver enkeltoppgave besvares av flere respondenter. Ulempen med oppsettet over et at færre testledd kan testes ut. Som ved all annen testing er det er viktig å ta stilling til hvor mye innhold en matrisesamling skal dekke samt hvor lang tid det skal ta å gjennomføre ett sett (Aningbo, 2011; Childs & Jaciw, 2003). Childs & Jaciw (2003) nevner også viktigheten av å planlegge hvor mange sett en matrisesamling skal bestå av, hvor mange oppgaver hvert sett skal inneholde samt hvor mange respondenter som skal gjennomføre hvert sett.

2.3 Fordeler og ulemper med flervalgsoppgaver

Det er mange fordeler med å bruke flervalgsoppgaver. Flervalgsoppgaver er vanligvis lukkede oppgaver som ikke krever at respondenter uttrykker seg skriftlig, organiserer ideer eller drøfter synspunkter (Sirnes, 2005, s. 27). Det betyr at respondentenes skriveevne og formuleringsevne ikke spiller inn i vurderingen. Flervalgsoppgaver er som regel mindre tidkrevende å løse, og vurderingen i etterkant går som regel raskere og er helt objektiv. Kortere tidsbruk per oppgave og en redusert skrivemengde kan også bidra til å gi respondenter en økt motivasjon og mestringsfølelse (Lykknes & Smidt, 2009). Et viktig poeng er at flervalgsoppgaver gjennom analyse kan gi verdifull diagnostisk informasjon, samt at oppgavene kan brukes om igjen over flere år. Flervalgstester kan benyttes i alle fagfelt, kan dekke store deler av pensum på en effektiv måte og kan brukes til å teste kompetanse på alle nivåer (Sirnes, 2005; Zimmaro, 2004).

Det finnes også ulemper med flervalgsoppgaver. Å lage gode flervalgsoppgaver er ofte tidkrevende, og å formulere gode distraktorer kan være svært krevende (Sirnes, 2005). I tillegg til at respondenters leseevner kan virke inn på resultatene deres, så egner

flervalgsoppgaver seg dårlig til å måle evne til å uttrykke seg (muntlig eller skriftlig), og det hindrer også respondenter i å uttrykke kreativitet, originalitet og fantasi (Zimmaro, 2004). Flervalgsoppgaver kan lett bli detaljorienterte, og det er ikke alltid oppgavene tester høyere kognitiv tenkning (Sirnes, 2005; Zimmaro, 2004).

2.4 Klassifisering av oppgaver:

For å klassifisere oppgaver som brukes i undervisning eller tester er det nyttig å bruke en taksonomi, og Blooms kognitive taksonomi er ofte brukt til dette formålet (Sirnes, 2005). Blooms taksonomi består av seks hierarkiske kunnskapsnivåer: kunnskap, forståelse, anvendelse, analyse, syntese og vurdering (Bloom, 1956). Noen av nøkkelverbene som kjennetegner de ulike nivåene er gitt i tabell 2.4 (Sirnes, 2005).

Tabell 2.4: *Blooms kognitive taksonomi*

Kategori	**Nøkkelverb**
kunnskap	beskrive, definere, gjengi
forståelse	forklare, skjelne, tolke
anvendelse	bruke, demonstrere, måle
analyse	identifisere, klassifisere, skille ut, sammenligne
syntese	generalisere, organisere, trekke slutninger
vurdering	avgjøre, skille mellom, velge

Kunnskap blir satt som det laveste nivået og *vurdering* som det høyeste, og de ulike nivåene blir ofte visualisert som en pyramide eller en trapp der man starter på det laveste nivået. Ved bruk av flervalgsoppgaver blir det ofte hevdet at man bare tester kompetanse på de lavere nivåene i Blooms taksonomi, men det er mulig å teste elevene på høyere kunnskapsnivåer (Haladyna, 2004).

2.5 Reliabilitet

Reliabilitet kan sies å være et synonym for pålitelighet, konsistens og replikerbarhet over tid, over instrumenter og gjennom grupper av respondenter. Forskning som utføres på en annen

gruppe respondenter ved lik kontekst, vil være reliabel dersom resultatene blir de samme i den andre gruppen (Bortolotti *et al.*, 2012; Cohen *et al.*, 2011). Resultater er reliable hvis feilene som forekommer er helt tilfeldige uten noen stor feilmargin. Gjentatte målinger kan veie opp for avvik i målinger. Reliabilitet har ikke like stor betydning som validitet, men reliabilitet blir ansett som det viktigste kriteriet som målinger og resultater blir målt opp mot (Geisinger, 2013).

Innenfor kvantitativ forskning skiller man mellom tre typer reliabilitet; *reliabilitet som stabilitet, reliabilitet som likhet* og *reliabilitet som intern konsistens*. Reliabilitet som stabilitet dreier seg om at resultater skal være konsistente over tid og over ulike grupper av respondenter. Når man undersøker om respondentenes svar er konsistente over tid, bruker man en test og deretter en re-test etter et gitt tidsrom. Tidsrommet som velges må være passende, og man kan da undersøke korrelasjon koeffisienten mellom svarene i test og re-test. Hvis man skal undersøke om resultater er konsistente over ulike grupper respondenter, bør man bruke grupper med respondenter som er mest mulig like i karakteristikker som er relevant, som for eksempel alder, kjønn og dyktighetsnivå. Resultater eller responser skal da være tilsvarende for de ulike gruppene dersom reliabilitet skal oppnås (Cohen *et al.*, 2011).

Reliabilitet som likhet kan deles inn i to komponenter. Den første komponenten går på omt man får konsistente responser ved bruk av alternative former innenfor et undersøkelsesinstrument, for eksempel gjennom en pre- og post-test i et eksperiment. Den andre komponenten går på at det skal være konsistens mellom ulike forskere som deltar i et forskningsprosjekt. Spesielt i undersøkelser der man bruker observasjon eller semi-strukturerte intervjuer vil dette være viktig (Cohen *et al.*, 2011).

Reliabilitet som indre konsistens måler hvor godt det er samsvar mellom ulike testledd, der disse testleddene til sammen skal gjenspeile individuell variasjon i en egenskap, for eksempel dyktighet (Nasjonalt kunnskapssenter for helsetjenesten). Indre konsistens kan måles ved ”split-half”-metoden der man deler en test inn i to halvdeler, der hver halvdel skal være mest mulig lik i vanskelighetsgrad og type innhold. Man kan deretter undersøke korrelasjonen mellom de to gruppene av respondenter. Spearman-Brown formelen brukes da til å undersøke reliabiliteten til hele testen, basert på reliabiliteten til de to deltestene (Traub & Rowley, 1991 i Resaland, 2013):

Formel 1: *Spearman-Brown-formelen*

$$P_{nn} = \frac{kP_{xx}}{1+(k-1)P_{xx}}$$

der

P_{nn} er estimatet av reliabiliteten til den forlengede testen
P_{xx} er estimatet av reliabiliteten til den originale testen
k er antall oppgaver i den forlengede testen

En annen måte å sjekke indre konsistens er å bruke Cronbachs alpha. Dette er den hyppigst brukte reliabilitetskoeffisienten (DeVellis, 1991). Den er populær å bruke fordi man da slipper å ta for seg problemet med hvordan man skal dele opp en test i to slik man må ved ”split-half” metoden (Falk & Savalei, 2011). Cronbachs alpha gir en koeffisient for korrelasjonen av hver oppgave med summen av alle de andre oppgavene. Det er derfor den indre konsistensen mellom oppgaver og ikke mellom personer som blir målt. Cronbachs alpha er gitt av formelen (Cohen *et al.*, 2011):

Formel 2: *Cronbachs alpha*

$$alpha = \frac{nr_{ii}}{1+(n-1)r_{ii}}$$

der:

n er antallet oppgaver i en test
r_{ii} er gjennomsnittet av alle korrelasjonene mellom oppgaver

Følgende retningslinjer kan bli brukt for både metoden ”split-half” og Cronbachs alpha (Cohen *et al.*, 2011):

Tabell 2.5: *Reliabilitetskategorier for "split-half"-metoden og Cronbachs alpha (Cohen et al., 2011)*

Korrelasjon	**Reliabilitet**
> 0,90	svært høy reliabilitet
0,80 – 0,90	høy reliabilitet
0,70 – 0,80	middel reliabilitet
0,60 – 0,70	minimal reliabilitet
< 0,60	uakseptabel reliabilitet

2.6 Validitet

I en vitenskapelig undersøkelse betraktes validitet som et mål på hvor godt resultatene samsvarer med det undersøkelsen er ment å skulle beskrive, forklare eller teoretisere. Validiteten til en undersøkelse antyder hvor godt undersøkelsen måler det den hevder å måle og hvor godt den gjenspeiler den virkelige verden (Bortolotti *et al.*, 2012; Cohen *et al.*, 2011; DeVellis, 1991). Validitet har forskjellige betydninger i kvantitativ og kvalitativ forskning (Winter, 2000). I kvantitativ forskning forbindes validitet med variabler som blant annet observasjoner, kontroll, replikasjon, forutsigbarhet, generalisering og statistiske analyser. Det er i kvantitativ forskning at begrepet validitet har sitt opphav, og da som et mål på hvor godt variablene, enkeltvis eller sammen, måler det de er ment å måle. I kvalitativ forskning derimot forbindes validitet med sammensatte begreper som meninger og holdninger, sosiokulturelle betingelser, uforutsigbarhet, ærlighet og samtalekvalitet (Cohen *et al.*, 2011, Winter, 2000).

Det finnes ulike dimensjoner av validitet, som *deskriptiv validitet*, *indre-* og *ytre validitet*, *innholdsvaliditet*, *konstruktvaliditet*, *kriterievaliditet*, *teoretisk validitet* osv. De ulike dimensjonene har ulike anvendelsesområder, og forskjellige forskningsområder vektlegger gjerne ulike dimensjoner av validitet i sin forskning. I denne sammenhengen er det viktig at validiteten blir vurdert etter de kriterier som gjelder i ethvert forskningsområde (Cohen *et al.*, 2011). I de neste avsnittene vil innholdsvaliditet, konstruktvaliditet og kriterievaliditet bli nøyere gjennomgått, med et spesielt fokus på hvordan man kan vise til disse typene validitet når man konstruerer flervalgstester.

Innholdsvaliditet defineres som graden av hvorvidt et målingsinstrument måler det det er designet for å måle (Cohen *et al.*, 2011; DeVellis,1991). I flervalgstester handler innholdsvaliditet først og fremst om graden av innholdsdekning; Hvorvidt innholdet i testen er representativt for fagområdet testen er ment å dekke. Det vil si at enhver oppgave bør reflektere den underliggende kompetansen som oppgaven er ment å teste (DeVellis, 1991; Messick, 1989; Sirnes, 2005). Den innholdsrelaterte validiteten svekkes i følge Messick (1989) når oppgaver i en flervalgstest ikke måler alle underdimensjoner av kompetansen oppgavene hevder å måle, eller når oppgavene måler noe annet enn de underliggende kompetansene. I utviklingen av flervalgsoppgaver er det også vanlig at eksperter på området vurderer oppgavene og at oppgavene knyttes til teoretiske rammeverk. Dette øker validiteten i en flervalgstest (Cohen et al., 2011; DeVellis, 1991; Haladyna, 2004; Messick, 1989).

Kriterievaliditet defineres som graden av korrelasjon mellom resultatene fra undersøkelser som er ment å måle det samme. Kriterievaliditet kan oppnås ved at resultater korrelerer med tilsvarende målinger som er gjort på tidligere eller fremtidige tidspunkter, eller dersom resultatene korrelerer med resultater fra målinger gjort med andre målingsinstrumenter (Cohen *et al.*, 2011; DeVillis, 1991; Messick, 1989). Kriterievaliditet i flervalgstester tar ofte utgangspunkt i spørsmålet om en test kan si noe om fremtidige prestasjoner, hvor hovedelementet i denne problemstillingen er forholdet mellom testscorene og kriteriet man vil vurdere. Dette forholdet er vanlig å uttrykke gjennom en korrelasjonskoeffisient. En høy korrelasjonskoeffisient antyder at to tester rangerer personer på tilsvarende vis (Sirnes, 2005).

Konstruktvaliditet defineres som de empiriske holdepunktene man har for å hevde at en forskningsdesign måler det den er designet for å måle. Det innebærer at utformingen av designet bør være forankret i tilstrekkelig relevant litteratur, at designet bør korrelere positivt med tilsvarende designer og at potensielle moteksempler som kan falsifisere designet er presentert. Man er først i stand til å vurdere konstruktvaliditet når bekreftende og avkreftende argumenter er vurdert og balansert (Cohen *et al*., 2011; DeVellis, 1991; Grimm & Widaman, 2012). Som med innholdsvaliditeten svekkes også den konstruktrelaterte validiteten når oppgaver måler noe annet enn de var ment å måle. Fordi konstruktvaliditet også dreier seg om hvor godt resultatene på en test samsvarer med det testen skulle teoretisere kan konstruktvaliditet sees på som en blanding av både kriterievaliditet og innholdsvaliditet. Derfor kan konstruktvaliditet sees på som den viktigste typen validitet når flervalgsoppgaver skal valideres (Loevinger, 1957).

2.7 Introduksjon av Item Response Theory

Den grunnleggende utviklingen av psykometriske ble gjort i et teoretisk rammeverk som kalles klassisk testteori (KTT). KTT ble introdusert allerede på tidlig på 1900-tallet og gir en gir et rammeverk for hvordan konstruksjon og analyse utføres, og hvordan poengscore på tester settes (Embretson & Reise, 2000). KTT bygger på matematiske modeller og ideen om at andre faktorer enn den testen skal måle også kan spille inn på resultatene. KTT kan være nyttig å bruke innenfor for eksempel psykologi fordi egenskaper ofte ikke er mulig å måle direkte. Egenskaper må da måles indirekte gjennom måling av andre observerbare trekk (Lord & Novick, 1968 i Ostini & Nering, 2006). Problemer med å teste antagelsene og bruke

resultater i virkeligheten har ført til at nye modeller har blitt utviklet som en utvidelse og liberalisering av klassisk testteori (Brennan, 1998). Både tester innenfor utdanning og psykologi har i klassisk testteori en utfordring i at resultatene for enkeltpersoner er avhengig av den spesielle respondentgruppen de er en del av, og at nivået på respondenters egenskaper er avhengig av det spesifikke utvalget av oppgaver/spørsmål som er valgt for testene. Item Response Theory (IRT) ble utviklet som svar på disse utfordringene (Hambleton et al., 1991).

Konseptet IRT dukket opp allerede før 1950, og pionerarbeidet rundt IRT som modell skjedde på 1950- og 1960-tallet, men på grunn av krav til datamaskiner med større kapasitet har det ikke blitt mer utbredt før i den siste tiden. Teorien ble først utviklet for å kunne analysere flervalgstester med korrekte/ikke korrekte svaralternativer, men har etter hvert også utviklet seg til å omhandle andre typer tester (Harvey & Hammer, 1999).

IRT bygger på at det er en gitt sannsynlighet for at en respondent svarer på et spesifikt spørsmål på en gitt måte. Metoden beskriver hvordan en person med høyere nivå av en egenskap vil svare i en annen kategori enn en person med lavere nivå av den samme egenskapen. Funksjonen som da brukes kalles en oppgavens responsfunksjon (eng.: *Item response function* – IRF) (Ostini & Nering, 2006). I IRT kalles egenskapen som skal testes en latent egenskap (eng.: *Latent trait*), fordi det antas at det er denne egenskapen som bestemmer responsen på spørsmålene som blir gitt i en test eller en undersøkelse. For eksempel kan denne ”egenskapen” være om en person er deprimert. Dette kan ikke måles direkte, men indirekte gjennom at personen svarer på spørsmål. IRT gir da en sannsynlighet for at personen svarer på en viss måte på et spesifikt spørsmål. En person som har en høyere grad av depresjon har større sannsynlighet for å svare på en gitt måte på et spørsmål, enn en person som har en lavere grad av depresjon. Både personens grad av egenskapen og ”vanskelighetsgraden” på spørsmålet ligger på samme skala, og man kan derfor bestemme sannsynligheten for at en person svarer på en gitt måte, på ethvert spørsmål (Reise *et al.*, 2005).

2.7.1 Antagelser for bruk av IRT

Dimensjonalitet

Når man skal bruke IRT som analyseverktøy er det visse antagelser som ligger i bunn. En av antagelsene ved noen av de mest brukte IRT-modellene er at bare én egenskap eller ett trekk

blir målt av et sett av spørsmål/oppgaver. Dette kalles endimensjonalitet (Hambleton *et al.*, 1991). For eksempel kan man være ute etter å måle matematisk kompetanse på en matematikktest, og da bør det være matematisk kompetanse alene som ligger til grunn for variasjonen i svarene respondentene gir på testen. Men det vil alltid være flere faktorer som spiller inn, så antagelsen vil aldri gjelde fullstendig. Andre slike faktorer kan være kognitive egenskaper, personlighet, motivasjon, nervøsitet for tester, evnen til å arbeide raskt og tendensen til å gjette på svar når man er usikker på hva som er riktig. For at antagelsen om endimensjonalitet skal bli møtt, må det være én dominant komponent som påvirker testresultatene (Hambleton *et al.*, 1991).

Det er også utviklet flerdimensjonale modeller der to eller flere parametere bestemmer sannsynligheten for at en person svarer riktig på et element. Kravet om endimensjonalitet gjelder naturlig nok ikke her. Når personer varierer systematisk i hvordan de svarer på spørsmål/elementer vil en flerdimensjonal modell ha en bedre "fit" for elementresponsdataene (Embretson & Reise, 2000). Flerdimensjonale modeller vil ikke bli behandlet i denne masteroppgaven.

Lokal uavhengighet

En annen forutsetning er forutsetningen om lokal uavhengighet. Lokal uavhengighet betyr at når evner som påvirker testresultater blir holdt konstant, vil respondenters svar på ethvert par av testelementer være statistisk uavhengig av hverandre. Hvis antagelsen om endimensjonalitet holder, vil det også være lokal uavhengighet. Lokal uavhengighet kan likevel også oppnås ved flerdimensjonale modeller, så lenge alle evner og egenskaper som påvirker resultatene er tatt med i beregningene. Et brudd på lokal uavhengighet kan for eksempel være at det i en matematikktest kreves gode leseferdigheter for å gjøre det bra. Dersom ikke alle respondentene har gode leseferdigheter, vil dette kunne føre til at ikke alle egenskapene som påvirker testen blir tatt med i beregningene (Hambleton *et al.*, 1991). Antagelsen blir også brutt hvis et svar på et spørsmål er avhengig av et annet svar (Marais & Andrich, 2008).

2.7.2 Modeller innenfor IRT

Innenfor IRT skiller man mellom *dikotome* modeller og *polytome* modeller. Dikotome modeller har data med to svaralternativer som for eksempel ja-nei, korrekt–ikke korrekt,

riktig-galt. Polytome modeller tar for seg graderte svar som for eksempel Likert-skalaen hvor man svarer på en skala fra for eksempel 1-5 der man ved 1 er helt enig og ved 5 er helt uenig i en påstand. Også svarkategorier hvor man får delvis poengscore for delvis rette svar eller hvor flere svar er korrekt og man kan få poeng for alle de korrekte svarene vil bli behandlet som polytome data (Ostini & Nering, 2006). Polytome modeller vil ikke bli behandlet videre i denne oppgaven.

Dikotome modeller er ikke begrenset til spørsmål som har 2 svaralternativer. Det kan være flervalgsoppgaver med flere svaralternativer, og også åpne oppgaver der selve svaret kan kategoriseres som dikotomt. Spørsmålet trenger heller ikke være formulert på en "riktig-galt" måte, men svarene må kunne kategoriseres innenfor de to kategoriene riktig og galt (Harvey & Hammer, 1999). Innenfor dikotome modeller blir de logistiske modellene med én, to og tre parametere (1PLM, 2PLM, 3PLM) som oftest brukt.

Logistisk modell med én parameter

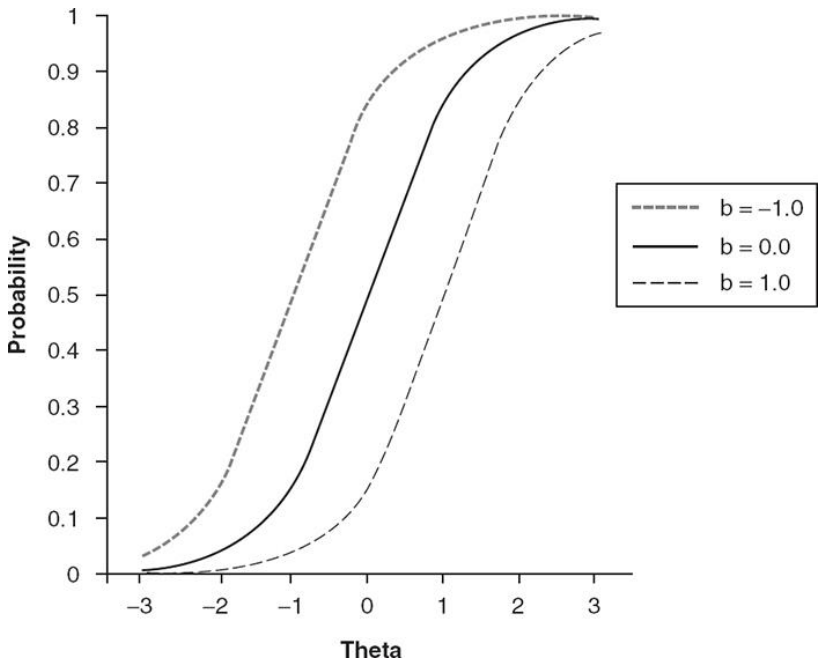
I den logistiske modellen med én parameter (1PLM) som også blir kalt *endimensjonal Rash-modell for dikotome data* er mellom hva som kreves (dyktighet) og sannsynlighet for å løse oppgaven gitt som *oppgavens karakteristiske kurve* (eng.: Item characteristic curve – ICC) gitt av ligningen:

Formel 3: *Oppgavekarakteristisk kurve i den logistiske modellen med én parameter*

$$P_i(\theta) = \frac{e^{(\theta - b_i)}}{1 + e^{(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots, n$$

der
θ er definert som respondenters dyktighet
P_i (θ) er sannsynligheten for at en tilfeldig valgt respondent med dyktighet θ svarer korrekt på oppgave i
b_i er oppgave i sin vanskelighetsgrad
n er antall oppgaver i testen
e er en matematisk konstant med tilnærmet verdi 2.718
P_i (θ) er en S-formet kurve med verdier mellom 0 og 1 på dyktighetsskalaen

Parameteren b_i for en oppgave er plasseringen på dyktighetsskalaen hvor sannsynligheten for å få korrekt svar er 0,5. Jo høyere verdi av b_i, jo høyere dyktighet kreves av respondenten for å ha 50 % sjanse for å svare korrekt (Hambleton *et al.*, 1991).



Figur 2.1: *Oppgavers karakteristiske kurve for den logistiske modellen med én parameter (1PLM)*

I figur 2.1 vises tre kurver der parameteren b varierer bra -1,0 til 1,0. Figuren viser hvordan dette gir utslag i kurven, der det er en forskyvning langs x-aksen. Parameteren b i modellen 1PLM viser det punktet der en person med gjennomsnittlig dyktighet θ har 50 % sjanse å svare korrekt på oppgaven (Kline, 2005).

Logistisk modell med to parametre

Den logistiske modellen med to parametere (2PLM) har en ekstra parameter, en diskrimineringsparameter. Parameteren gir informasjon om hvor godt en oppgave kan skille mellom respondenter.

2PLM-modellen er gitt av ligningen:

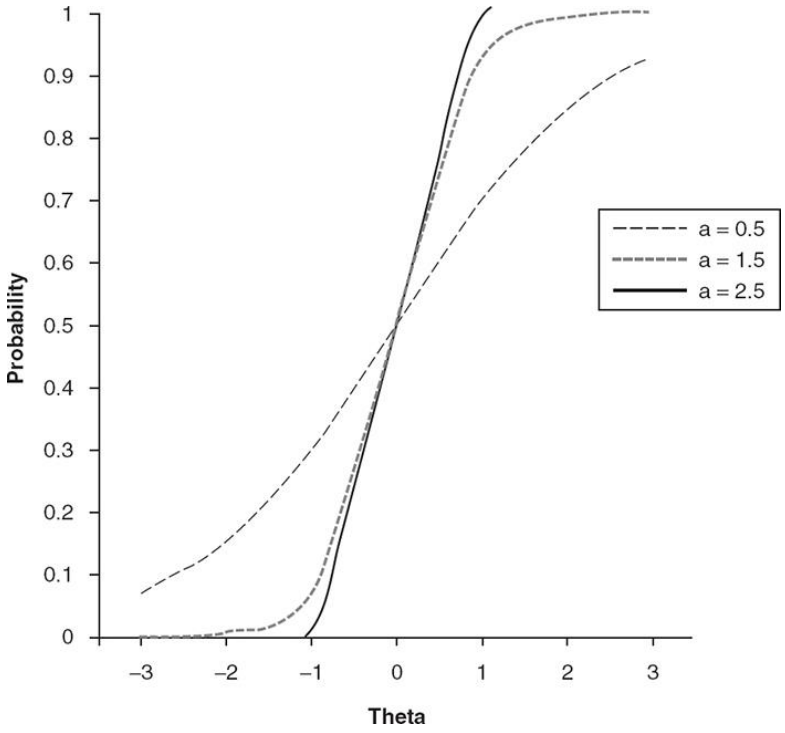
Formel 4: *Oppgavekarakteristisk kurve i den logistiske modellen med to parametre*

$$P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots, n$$

der

D er en skaleringsfaktor med verdi 1,7
a_i er diskrimineringsparameteren for oppgave i

Skaleringsfaktoren *D* er tatt med for at den logistiske funksjonen skal være så nær som mulig den normalfordelte funksjonen. Parameteren a_i er proporsjonal til stigningen til oppgavens karakteristiske kurve i punktet b_i på skalaen over vanskelighetsgrad. Kurver som er brattere, skiller bedre mellom respondenter enn kurver som er slakkere. Diskrimineringsparameteren ligger på en skala mellom -∞ og ∞, men oppgaver med negative verdier bør fjernes fra tester som måler evner ettersom respondenter med lavere dyktighet da har høyere sannsynlighet for å svare korrekt på en oppgave, enn en person med høyere dyktighet (Hambleton *et al.*, 1991).



Figur 2.2: *Oppgavers karakteristiske kurve for den logistiske modellen med to parametere (2PLM)*

Figur 2.2 viser oppgavens karakteristiske kurve for den logistiske modellen med to parametere (2PLM). Her er *b* satt til 0 for alle de tre kurvene, mens parameteren a varierer fra 0,5 til 2,5. Høyere verdier for *a* gir en brattere kurve og diskriminerer bedre mellom respondenter som har gjennomsnittlig dyktighetsnivå. Lavere verdier av parameteren a gir slakkere kurver og diskriminerer bedre på lavere og høyere nivåer av dyktighet (Kline, 2005).

Logistisk modell med tre parametre

Den logistiske modellen med tre parametere (3PLM) tar også med en parameter som tar i betraktning at respondenter med lav dyktighet kan svare korrekt på et spørsmål ved gjetting. Respondenter ved flervalgstester har mulighet til å svare korrekt ved gjetting, men

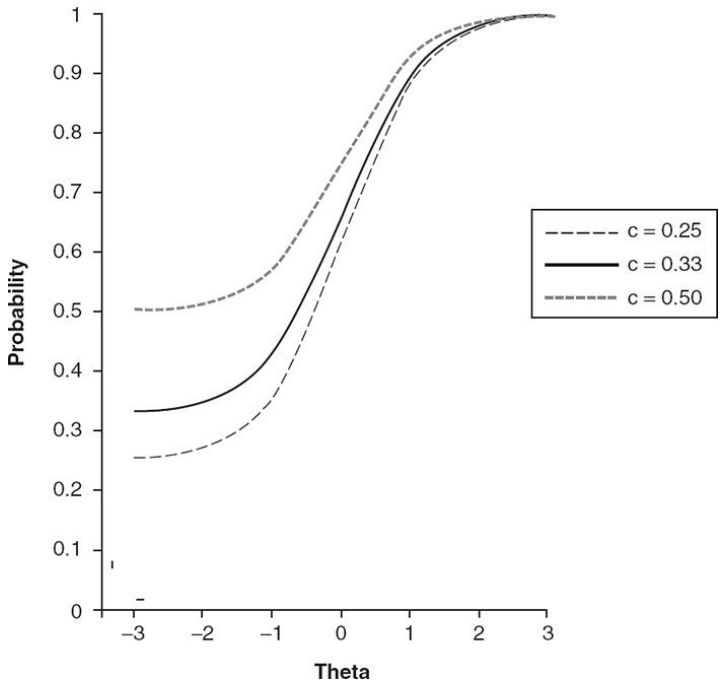
parameteren som tas med er lavere enn den ville vært hvis man gjetter tilfeldig på oppgavene. Dette er forklart ved at noen av svaralternativene kan være attraktive, men ukorrekte. Parameteren som kalles c_i, pseudo-sjanse-parameter, lager en nedre asymptote for oppgavens karakteristiske kurve (Hambleton et al., 1991).

3PLM-modellen er gitt av ligningen:

Formel 5: *Oppgavekarakteristisk kurve i den logistiske modellen med tre parametre*

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}} \qquad i = 1, 2, 3, \ldots, n$$

der
c_i er parameteren pseudo-sjanse



Figur 2.3: *Oppgavers karakteristiske kurve for den logistiske modellen med tre parametere (3PLM)*

Figur 2.3 viser hvordan parameteren c spiller inn på kurvene. De tre kurvene har $b = 0$, $a = 1.0$ mens c har verdiene 0,25, 0,33 og 0,5. Det er verdt å legge merke til at oppgavens vanskelighetsgrad er endret i forhold til modellene 1PLM og 2PLM. Vanskelighetsgraden er forskjøvet ut i fra den nedre asymptoten. Ved *c*-verdier over 0 vil dette ha negativ

innvirkning for respondenter med høy dyktighet θ. Det tar lengre tid å nå den øvre asymptoten sammenlignet med 1PLM og 2PLM modeller med like verdier for *a* og *b* (Kline, 2005).

2.8 Størrelse på testutvalg

Selv om det ikke finnes noen bestemte regler for hvor stort utvalget i en undersøkelse bør være er det allikevel enkelte retningslinjer som bør følges. Generelt sett så kan man si at de ulike modellene innenfor IRT stiller ulike krav til størrelsen på utvalget basert på modellenes kompleksitet (Edelen & Reeve, 2007; Tanaka, 1987). For at modellparametrene skal estimeres så nøyaktig så mulig anbefales det å ha en minimum utvalgsstørrelse på mellom 50 og 200 for 1PLM (Hula *et al.*, 2012, Linacre, 1994) og 350 for 2PLM (Embretson & Reise, 2000). 3PLM er så kompleks at den gjerne krever over 1000 respondenter for at estimeringer skal bli nøyaktige (Lord, 1968). Grunnet oppgavens omfang som tilsier at 1000 respondenter ikke er realistisk, vil ikke 3PLM bli videre behandlet.

Kapittel 3 – Material og metode

I den første delen av dette kapittelet er metoden for innsamling og konstruksjon av flervalgsoppgaver beskrevet, i tillegg til en beskrivelse av bruken av det nettbaserte systemet SurveyXact® som ble brukt til å distribuere flervalgstestene. Deretter er innsamlingen av svar og utvalget av respondenter gjort rede for. I den andre delen av kapittelet er analysene som ble utført på enkeltoppgaver for å utvikle testen beskrevet.

3.1 Innsamling og konstruksjon av flervalgsoppgaver

3.1.1 Læreverk

Etter en sammenligning av innholdet i de tre lærebøkene *Biology - How Life Works*, *Campbell Biology* og *Life - The Science of Biology* ble konklusjonen at bøkene var svært like, både tematisk og innholdsmessig. Basert på denne sammenligningen anså vi det som sannsynlig at boken Campbell Biology dekket det som kunne regnes for "grunnleggende basiskunnskaper i biologi", og være den faglige rammen for hva testen vi skal utvikle skal måle. Denne boken benyttes forøvrig i flere av grunnemnene i bachelorutdanningen i biologi ved Universitetet i Bergen, og vi syntes derfor at dette var en egnet lærebok å ta utgangspunkt i når vi skulle samle inn og skrive flervalgsoppgaver.

Campbell Biology er delt inn i 8 hovedområder der hvert hovedområde er delt inn i mellom 4 og 12 kapitler:

Tabell 3.1: *Oversikt over temaer og kapittelfordeling i Campbell Biology*

Tema	Antall kapitler
1: Livets kjemi	4
2: Cellen	7
3: Genetikk	9
4: Evolusjonære mekanismer	4
5: Den biologiske diversitetens evolusjonære historie	9
6: Planters oppbygning og funksjon	5
7: Dyrs oppbygning og funksjon	12
8: Økologi	5

3.1.2 Innsamlig og konstruksjon av oppgaver

Noen av flervalgsoppgavene som ble brukt i denne studien ble konstruert av oss selv etter gitte retningslinjer (Boks 2.13). Resten av flervalgsoppgavene ble samlet inn fra ulike nettsider[1]. Disse ble kritisk vurdert og som oftest redigert i forhold til gitte retningslinjer (Boks 2.13). Redigeringen kunne bestå i endring av stammen, endring av svaralternativene, endring av både stammen og svaralternativene eller å legge til eller fjerne svaralternativer. Noen av nettsidene tilhørte forlagene til diverse lærebøker, blant annet McGraw-Hill Education® og Pearson Education® Som et utgangspunkt var målet å samle inn og lage tilsammen ca. 350- 400 flervalgsoppgaver som dekket de ulike temaene i Campell Biology.

Ettersom det var mulighet for at respondentene ikke kunne norsk, ble alle oppgavene laget på både norsk og engelsk. De fleste oppgavene som ble hentet fra nettet var allerede på engelsk, og når vi lagde egne oppgaver ble disse stort sett lagd på engelsk og deretter oversatt til norsk. Noen av de biologiske fagordene og begrepene er i større grad kjent for studenter på engelsk, og i disse tilfellene ble det engelske ordet skrevet i parentes.

3.1.3 Kvalitetssikring

Det ble deretter gjennomført flere former for kvalitetssikring av oppgavene. Den første gjennomgangen ble gjort av den personen som skrev oppgavene. Her ble oppgavene sjekket opp mot kriterier for gode flervalgsoppgaver (Boks 2.13) og språk. Deretter gikk vi igjennom hverandres oppgaver, og i denne gjennomgangen ble oppgavene vurdert for innhold, språk samt overensstemmelse med teori. I denne prosessen delte vi også oppgavene inn i kategoriene "*kunnskap*" (K), "*anvendelse*" (A) og "*vurdering*" (V) slik at fordelingen av oppgaver med de forskjellige kategoriene kunne bli jevn mellom ulike settene. Kategorien K, A og V er laget med utgangspunkt i Blooms taksonomi (Tabell 2.4), men vi har en inndeling i tre kategorier istedenfor seks. Kategorien kunnskap blir kjennetegnet ved å kunne beskrive, definere og reprodusere og er derfor ganske lik kunnskapsnivået *kunnskap* i Blooms taksonomi. Vår kategori anvendelse har vi brukt kjennetegnene: bruke, demonstrere, identifisere, skille ut og er dermed en blanding mellom kategoriene *anvendelse* og *analyse* fra Blooms taksonomi. Kategorien *vurdering* er en blanding mellom kategoriene syntese og vurdering fra Blooms taksonomi med kjennetegnene: trekke slutninger, avgjøre og skille mellom. En del endringer ble utført etter vi fikk tilbakemeldinger fra hverandre, og noen

[1] Educational Testing Service, Garland Science, HCC Southeast Commons, IndiaBIX, Oxford University Press, Pearson Higher Education, Tutor Vista & Ulysses S. Grant High School

oppgaver ble i denne prosessen fjernet fra databasen. I løpet av denne prosessen ble også det faglige innholdet i noen av oppgavene gjennomgått av en professor ved Institutt for biologi, UiB.

3.1.4 Matrisesamling

For å teste ut oppgaver som dekker alle temaene i Campbell Biology (Tabell 3.1) var det nødvendig med et stort utvalg av oppgaver. Men det var ikke realistisk å tenke at alle respondentene skulle svare på alle oppgavene. Vi delte derfor oppgavene inn i syv sett, og oppgavene ble jevnt fordelt etter tema og kategori (K, A og V). For å kunne sammenligne settene valgte vi ut ti oppgaver som var felles for alle settene, etter partial-matrix-modellen (Tabell 2.2). Fellesoppgavene ble valgt ut fra et ønske om at de skulle dekke flest mulig temaer og alle kategoriene K, A og V. Vi bestemte oss for å ha 50 oppgaver totalt i hvert sett. En del oppgaver fra temaet livets kjemi ble fjernet ettersom disse oppgavene gikk mer inn på kjemi enn biologi. Andre oppgaver som ble fjernet var noen av oppgavene som gikk direkte på definisjoner av begreper.

3.1.5 Pretest

Vi gjennomførte pretester av settene med utvalgte studenter. Vi fikk en student til å gjennomføre hvert sett. Vi ba studentene å skrive ned hvor lang tid de brukte på testene, samt notere seg dersom noe var uklart i forhold til språk eller formuleringer. I tillegg ba vi dem skrive ned dersom de hadde noen generelle kommentarer. Utfra tilbakemeldingene ble noen av oppgavene revidert, og vi bestemte oss for å beholde 50 oppgaver i hvert sett.

3.2 Konstruksjon av flervalgstester ved hjelp av Survey-Xact

For å gjøre testene våre nettbaserte, brukte vi systemet SurveyXact®. SurveyXact (SX) er et nettbasert system som er utviklet for å gjennomføre og analysere spørreskjemaundersøkelser. Vi fikk tilgang til SX gjennom Universitetet i Bergen.

Flervalgsoppgavene ble lagt inn i SX som oftest ved hjelp av en funksjon hvor systemet omdannet innlimt tekst til tilnærmet fullverdige oppgaver. Oversettelser til engelsk ble lagt inn manuelt på hver oppgave og på all annen tekst. Svaralternativene ble randomisert av programmet for hver respondent. Ved oppgaver hvor svaralternativene bestod av tall som naturlig bør stå i rekkefølge, ble ikke funksjonen "randomisering" brukt. Oppgavene ble også

randomisert slik at rekkefølgen ikke var gitt etter tema, og slik at fellesoppgavene ikke hadde lik plassering i de ulike settene. Dette var på grunn av at respondentene kom til å få mulighet til å gjennomføre flere av settene. Noen av oppgavene som vi vurderte som lettere ble flyttet frem til den første siden med oppgaver slik at respondentene ikke skulle "miste motet" og falle av i starten.

I starten av hver test ble det lagt inn to sider med veiledende informasjon og hvor respondentene måtte velge hvilket språk de ønsket å ta testen på, oppgi hvor mange studiepoeng de hadde i biologi og hvilket studiested de tilhørte. Disse spørsmålene ble lagt inn med en "validering", altså at respondentene måtte fylle ut disse for å gå videre. Deretter la vi inn sideskift etter hvert femte spørsmål for å øke brukervennligheten for PC, nettbrett og mobiltelefon. Farger, bakgrunn, skrifttype og skriftstørrelse ble valgt i forhold til brukervennlighet.

På slutten av testen la vi inn en videresending til en annen side, der fasiten på oppgavene ble gitt. Fasiten på fellesoppgavene ble ikke gitt på grunn av at respondentene kunne gjennomføre flere sett. På fasitsiden la vi også inn en mulighet for å skrive inn epostadresse slik at respondentene hadde mulighet til å bli med i trekningen av en iPad som gevinst for deltakelse. På denne måten unngikk vi at epostadressen ble koblet til respondenten, og at de kunne gå tilbake og endre svar etter at de hadde fått oppgitt fasiten.

3.3 Markedsføring og innsamling av data

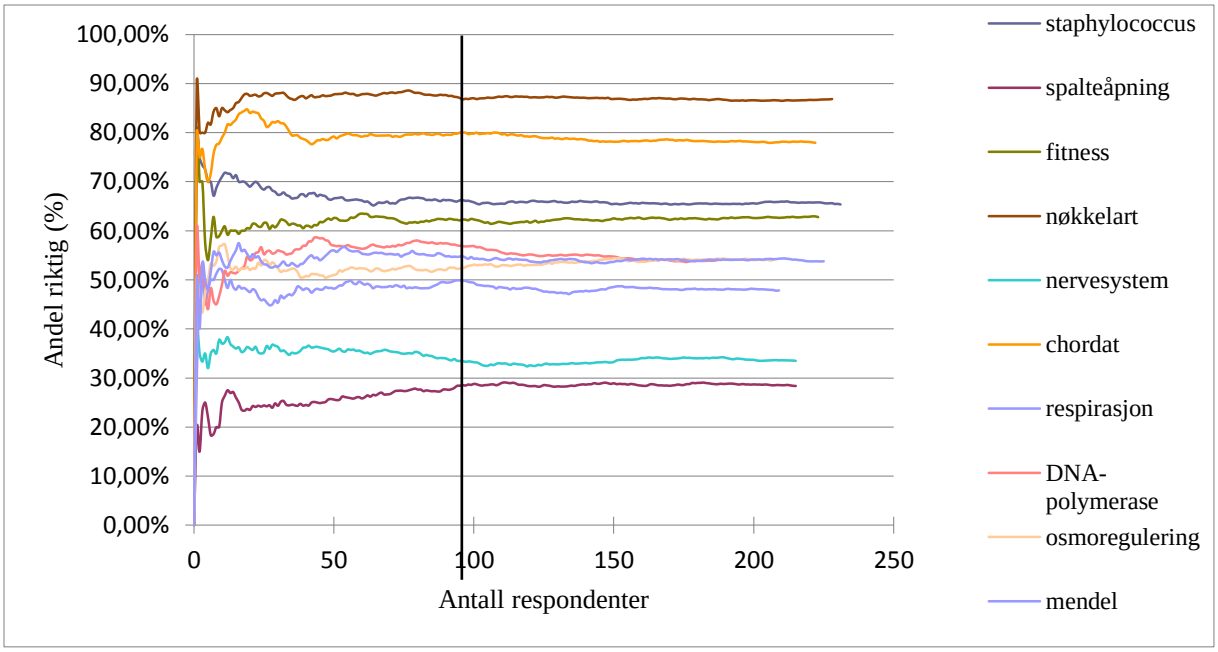
For å distribuere testene våre, ble det opprettet en link for hver av de syv testene. For å gjøre distribusjonen lettere, fikk vi hjelp til å lage én link som videresendte respondentene tilfeldig til en av de syv testene. For å distribuere linken til biologistudenter ved UiB ble mediene Webmail, Mi side og Facebook brukt. Gjennom Webmail og Mi side ble linken distribuert til alle studenter som var tilknyttet Institutt for biologi. På Facebook ble linken distribuert på kullgruppen til studenter på 1. og 2. året på bachelor i biologi, i tillegg til en gruppe som en god del av masterstudentene i biologi er med i. Ved å ha linken liggende på Mi side, var den i hele perioden lett tilgjengelig for studentene. Linken ble også videresendt til 8 andre

studiesteder i Norge[2] som distribuerte linken blant sine biologistudenter. For å øke svarprosenten på flervalgstestene våre besøkte vi forelesninger i to av de største grunnkursene i biologi ved UiB, der vi fikk forklart hvorfor det var viktig at studentene gjennomførte testen, og studentene fikk også knyttet ansikter opp mot testen. I tillegg fikk vi reklamert for testene våre gjennom en PowerPoint-Slide som sirkulerte på flere TV-skjermer i ganger og oppholdsarealer rundt om på Høyteknologisenteret, Bergen, og Biologen, Bergen, der man gjerne treffer flest biologistudenter. Vi reklamerte også for testene gjennom samtaler med andre studenter. Etter to uker var gått sendte vi ut påminnelse om undersøkelsen på Webmail og Facebook, og dette ble gjentatt fem dager før testen ble stengt for besvarelser. I tillegg sendte vi ut forespørsler til de andre studiestedene om å sende ut påminnelser til sine studenter. Testene lå ute totalt én måned og fem dager. For å øke motivasjonen til respondentene for å gjennomføre undersøkelsen fikk vi innvilget støtte til kjøp av en iPad til en tilfeldig uttrekt respondent.

3.4 Utvalgseffektivitet

Før undersøkelsen ble stengt for besvarelser, var det ønskelig å vurdere om det totale antall respondenter var nådd et nivå hvor tillegging av ytterligere respondenter ikke lenger ville ha stor betydning for vurderingen av oppgavene. Dette ble estimert ut i fra de 10 fellesoppgavene, som på det tidspunktet hadde mellom 187 og 231 besvarelser. For å gjøre estimeringen mer statistisk representativ ble det konstruert 10 paralleller for hver fellesoppgave, hvor hver parallell besto i en tilfeldig rekkefølge av de faktiske respondentene. For hver slik parallell ble endringene i oppgavenes vanskelighetsgrad regnet ut etterhvert som tilfeldige responder ble lagt til i beregningene. Gjennomsnittet av alle parallellene ble deretter beregnet for alle fellesoppgavene. Endringen av dette gjennomsnittet er vist for alle fellesoppgavene Figur 3.1.

[2] Høgskulen i Telemark, Norges miljø- og biovitenskapelige universitet, Norges teknisk-naturvitenskapelige universitet, Universitetet i Agder, Universitetet i Bergen, Universitet i Norland, Universitetet i Oslo, Universitetet i Tromsø & Universitetssenteret på Svalbard.



Figur 3.1: *Endring i vanskelighetsgrad som funksjon av antall respondenter for de 10 fellesoppgavene. Når de 10 fellesoppgavene blir sett under ett, endrer ikke vanskelighetsgraden seg mer enn 0,5 % etter 105 respondenter (loddrett linje).*

Etter at endringene i vanskelighetsgrad som funksjon av antall respondenter ble beregnet for alle fellesoppgavene, ble det estimert hvordan selve endringene forandret seg for hver respondent. For hver fellesoppgave ble den grensen identifisert hvor alle de påfølgende respondentene ikke medførte noen nevneverdig endring (< 0,5 %) i vanskelighetsgrad. Gjennomsnittsgrensen for de 10 fellesoppgavene ble beregnet til 105 respondenter (med et standardavvik på 48). For fellesoppgavene sett under ett betydde det at ytterligere respondenter utover 105 ikke tilførte nevneverdige endringer i vanskelighetsgrad. Gjennomsnittsgrensen er symbolisert med en svart, loddrett linje i figur 3.1.

I figur 3.1 ser vi at variasjonen i andel riktige svar minker jo flere respondenter som legges til. Konfidensintervallet for gjennomsnittsgrensen er ganske stort ([10,200]), men hele konfidensintervallet ligger innenfor det maksimale respondentantallet, og siden respondentantallet på hvert testledd kom til å bli maksimert ved imputering ble det antatt at utjevningen kom til å gjelde størsteparten av de resterende testleddene. Siden kriteriet for gjennomsnittsgrensen var en svært lav endring (< 0,5 %) og siden hele konfidensintervallet var innenfor det maksimale respondentantallet ble det antatt at grensen ble nådd for størsteparten av de resterende testleddene, hvor ytterligere respondenter ikke lenger ville være statistisk merkbart. På grunnlag av dette vurderte vi at antallet respondenter var tilstrekkelig og datainnsamlingen kunne avsluttes.

3.5 Analyse

Statistiske analyser ble gjennomført i programmet R, gjennom R Studio versjon 0.98.1056. Data fra alle de syv settene i SurveyXact ble eksportert til Excel og deretter gjort om til dikotome datasett. Korrekte svar ble kodet 1 og ukorrekt svar ble kodet 0. For å øke svarprosenten på fellesoppgavene og for å kunne linke settene til hverandre, ble de syv settene satt sammen til ett datasett.

3.5.1 Klassisk oppgaveanalyse:

Første steg i analysen av oppgavene var å sjekke parameteren for vanskelighetsgrad (p-verdi). P-verdiene 0,00 og 1,00 gir ingen informasjon siden da ingen av respondentene eller alle respondentene, respektivt, har svart riktig på en spesifikk oppgave. Slike oppgaver diskriminerer ikke mellom respondenter og vil derfor bli fjernet fra settet hvis de eksisterer.

For å kunne bruke datasettet videre i analyser, var det nødvendig å sette inn forventede verdier for manglende verdier. Dette ble gjort gjennom pakken "mirt" i R, der programmet bruker verdiene for dyktighet hos respondentene, i tillegg til vanskelighetsgraden til oppgavene til å estimere verdier der data mangler (Chalmers, 2012).

3.5.2 Evaluering av imputeringsteknikk

Når man benytter seg av matrisesamlinger (hvor respondentene ikke besvarer et identisk oppgavesett), vil de samlede dataene naturlig nok inneholde en god del manglende verdier siden respondentene kun besvarer et utvalg testledd. En vanlig måte å håndtere slike manglende verdier på er å erstatte de manglende verdiene med estimater (Wolkowitz & Skorupski, 2013). Slike teknikker kalles imputeringer (eng.: Impute missing values), og det finnes en god del slike metoder, nyttige for ulike formål.

Å imputere manglende verdier er ikke uten fare, og ved bruk av imputeringsteknikker bør man vise aktsomhet, hovedsakelig fordi man lett kan bli lurt til å tro at et ukomplett datasett kan bli komplett (Little & Rubin, 2006). En fare er systematiske feil forårsaket av unøyaktigheter ved systemet. Systematiske feil vil kunne føre til feilaktige analyser og

resultater og konklusjoner vil dermed kunne trekkes på galt grunnlag, noe som vil redusere systemets reliabilitet og validitet (Huisman, 2000; Wolkowitz & Skorupski, 2013).

Den vanligste måten å imputere manglende verdier på i analyser av flervalgsoppgaver innebærer at de manglende verdiene blir erstattet med vanskelighetsgraden på et bestemt testledd sammenlignet med respondentens beregnede dyktighet (θ) (Chalmers, 2012). Denne metoden antar at årsakene til de manglende verdiene er helt tilfeldig, og at verdiene ikke har noen relasjon til modellparametrene (eng.: Missing at random - MAR) (Rubin, 1975).

En imputeringsteknikk kan sies å være god dersom den frembringer korrekte estimater av de manglende verdiene (Huisman, 2000). Kromey & Hines (1994) påpeker at effektiviteten av en imputeringsteknikk bør vurderes opp mot bestemte kriterier som vanligvis benyttes i vitenskapelige analyser. I analyser av flervalgsoppgaver vil det derfor være naturlig å vurdere effektiviteten av en imputeringsteknikk ved å sammenligne verdier på dyktighetsskalaen (θ) før og etter imputeringen (Huismann, 2000). Men like viktig er en imputeringsteknikks evne til å bevare relasjoner mellom testleddene og redusere systematiske feil forårsaket av de manglende verdiene (Sande, 1982).

For å vurdere effektiviteten av imputeringsteknikken *t* ble verdier på dyktighetsskalaen før imputeringen sammenlignet med tilsvarende verdier etter imputeringen. Dette var for å evaluere en eventuell endring i respondentenes dyktighet som følge av imputeringen. I tillegg ble verdiene for Cronbach's alpha før imputeringen sammenlignet med de tilsvarende verdiene etter imputeringen. Dette var for å evaluere om imputeringsteknikken bevarte relasjonene mellom testleddene på en god måte. I begge disse tilfellene ble kvadratisk gjennomsnittsavvik beregnet og brukt for å evaluere verdien av imputeringsteknikken. Det kvadratiske gjennomsnittsavviket (eng.: Root mean square deviation - RMSD) er gitt ved formelen:

Formel 6: *Kvadratisk gjennomsnittsavvik*

$$RMSD = \sqrt{\frac{1}{n}\sum_{v=1}^{n} d_v(t)^2}$$

der

n = antall personer med manglende data i datamatrisen
$d_v(t) = r_v(t) - v(t)$
$r_v(t)$ = verdi etter imputeringen
v(t) = verdi før imputeringen

3.5.3 Point-biserialkorrelasjoner

Det fullstendige datasettet ble undersøkt videre ved å se på point-biserialkorrelasjon. Point-biserialkorrelasjon er Pearson-korrelasjonen mellom "scoren" på hver oppgave (0 eller 1) og "scoren" på den totale testen. Verdiene til point-biserialkorrelasjon varierer mellom -1,0 og +1,0. En høy verdi av point-biserialkorrelasjon sier at respondenter med høy dyktighet generelt svarer korrekt på oppgaven, mens respondenter med lav dyktighet svarer feil på oppgaven. En negativ korrelasjonsverdi sier at personer med lav dyktighet har høyere grad av korrekt svar på en gitt oppgave, enn personer med høy dyktighet (Varma, 2006). Point-biserialkorrelasjon er med andre ord et mål for hvor godt en oppgave diskriminerer mellom sterke og svake respondenter, men må ikke forveksles med diskrimineringsfaktoren i 2PLM (Kavitha *et al.*, 2012). Oppgaver med negativ verdi må fjernes fra settet. En minimumsverdi på 0,15 er anbefalt (Varma, 2006), og oppgaver med point-biserialverdi under 0,15 ble fjernet fra oppgavesettet.

3.5.4 Dimensjonalitet

En av antakelsene for å bruke modellene 1PLM, 2PLM og 3PLM er at dataene er tilstrekkelig endimensjonale, altså at bare en egenskap blir testet. For å teste dimensjonaliteten ved modellen 1PLM ble en *modifisert parallellanalyse* utført. Denne prosedyren er foreslått av Drasgow and Lissak (1983) for å teste dimensjonaliteten til oppgavesett der respondentenes svar er dikotomt scoret. En modifisert parallell analyse bruker den andre egenverdien[3] fra "tetrachoric"-korrelasjonsmatrisen av de dikotome dataene, og det blir testet om den egenverdien fra de observerte dataene er høyere enn den egenverdien for de syntetiserte dataene. Dersom egenverdien fra de observerte dataene er vesentlig høyere enn egenverdien for de syntetiserte dataene indikerer dette flerdimensjonalitet (Drasgow and Lissak, 1983).

For å teste dimensjonaliteten til modellen 2PLM, ble det utført en likelihood ratio test mellom to 2PLM modeller, den ene med én latent variabel og den andre med to latente variabler.

[3] "tetrachoric"-korrelasjonsmatrisen oppgir flere egenverdier, men den modifiserte parallellanalysen benytter "egenverdi nummer to".

Akaike information criterion (AIC) og Bayesian information criterion (BIC) verdier ble sammenlignet. Modellen med lavest AIC (Wagenmakers & Farell, 2004) og BIC verdier (Cavanaugh, 2012) er best tilpasset dataene.

3.5.5 Valg av modell

For å teste mellom modellene 1PLM og 2PLM ble det utført en likelihood-ratio-test mellom modellen PLM1 og PLM2. Akaikisk informasjonskriterium (eng.: Akaike information criterion – AIC) og bayesisk informasjonskriterium (eng.: Bayesian information criterion – BIC) ble sammenlignet. Modellen med lavest AIC (Wagenmakers & Farell, 2004) og BIC verdier (Cavanaugh, 2012) er best tilpasset dataene.

En annen måte å vurdere hvilken modell som passer best til dataene, er å se på hvor mange oppgaver som har en dårlig tilpasning til modellene. Man kan da se på hvilken modell som har færrest oppgaver som er dårlig tilpasset, og denne modellen vil da være best egnet. Det blir testet gjennom oppgavenes tilpasningsstatistikk (*item fit statistics*). Her blir ett nytt datasett simulert ved maximum-likelihood-estimater, og en Monte-Carlo-prosedyre blir brukt til å lage en tilnærming til distribueringen av oppgavetilpasningsstatistikken under null-hypotesen. Oppgavenes tilpasningsstatistikk gir et bilde på hvor alvorlig observasjoner avviker fra modellen, og kan identifisere problematisk konstruksjon av oppgaver, for eksempel oppgaver som er scoret feil, og kan også indikere error som har skjedd i kalibreringsfasen av testutviklingen (Reise, 1990). Modellen med færrest avvik ble valgt. Oppgaver som var dårlig tilpasset denne modellen ble fjernet før videre analyse.

3.5.6 Lokal uavhengighet

Antagelsen om lokal uavhengighet ble testet ved bruk av Yen´s Q_3-statistikk, som ofte er brukt til å sjekke lokal uavhengighet mellom oppgaver. En parvis indeks av oppgavenes avhengighet blir estimert gjennom korrelasjon mellom residualer fra IRT modellen. Når man bruker Q_3 for å sjekke for avhengighet er det vanlig å bruke en uniformt kritisk verdi med absoluttverdi 0,2 (Chen & Thissen, 1997). Det vil si at par av oppgaver som får verdi $< -0,2$ eller $> 0,2$ er betegnet som lokalt avhengige. En av metodene for å håndtere lokalt avhengige oppgaver på er å fjerne en av oppgavene fra oppgavesettet. En annen metode er å merke disse oppgavene som "fiender" og sørge for ved bruk av delsett av testen, at en respondent ikke får

begge oppgavene. Disse må da kontrolleres for under kalibreringen for å unngå påvirkning av høyt korrelerte oppgaver (Reise & Revicki, 2015). En av oppgavene fra hvert par ble fjernet. Ved stort avvik i p-verdi mellom de parvise oppgavene ble de oppgavene med p-verdi nær 0,5 beholdt i testen, ettersom det var ønskelig å ha en overvekt av oppgaver med p-verdi på rundt 0,5 (Varma, 2006). Ved p-verdi i samme område, ble oppgaven med høyest verdi av point-biserial korrelasjon beholdt i testen.

3.5.7 Vurdering av uegnede oppgaver

Det å undersøke fordelingen av svar på de ulike svaralternativene kan være nyttig ved analysering av flervalgsoppgaver som er brukt i en test. Ved å se svarprosenten på de ulike svaralternativene i en oppgave kan dette gi nyttig informasjon om svaralternativene er egnet. Det at alle svaralternativene er valgt av respondentene er et godt tegn, men hvis svarprosenten på et alternativ er svært lav eller 0, bør dette alternativet undersøkes nærmere. Dersom et av de ukorrekte svaralternativene har svært høy svarprosent kan dette tyde på at ordvalgene er dårlige, og kan feiltolkes. Det hender også at oppgaver er scoret feil, og et av de andre alternativene er det korrekte svaralternativet (Osterlind, 2002).

3.5.8 Reliabilitet

En testinformasjonsfunksjon (TIF) kan gi en indikasjon på reliabiliteten til en test på ulike nivåer av dyktighet hos respondentene. En testinformasjonsfunksjon er summen av oppgaveinformasjonsfunksjonene på en gitt plassering langs skalaen dyktighet. Oppgaveinformasjonsfunksjonen ved et gitt dyktighetsnivå er gitt ved ligningen (Hambleton *et al.*, 1991):

Formel 7: *Oppgaveinformasjon ved dyktighetsnivå θ*

$$Ii(\theta) = \frac{[P'i(\theta)]^2}{Pi(\theta)Qi(\theta)}$$

der

$I_i(\theta)$ er informasjonen gitt av oppgave *i* ved dyktighetsnivå θ
$P'_i(\theta)$ er den deriverte av $P_i(\theta)$ med hensyn på θ
$P_i(\theta)$ er oppgavens responsfunksjon

$Q_i(\theta) = 1\text{-}P_i(\theta)$

Testinformasjonfunksjonen er dermed gitt ved (Hambleton *et al.*, 1991):

Formel 8: *Testinformasjon ved dyktighetsnivå θ*

$$I(\theta) = \sum_{i=1}^{n} I_i(\theta)$$

Dersom kurven til testinformasjonsfunksjonen når et høydepunkt langs dyktighetsaksen vil testen skalere dyktighet med ulik presisjon langs dyktighetsskalaen. En kurve som er horisontal vil skalere dyktighet med lik presisjon langs hele dyktighetsskalaen. Noen kurver er relativt flat over deler av skalaene, og vil da skalere dyktighet med noenlunde lik presisjon langs dette intervallet på dyktighetsskalaen. Testen vil da være best egnet for personer som har dyktighetsnivå innenfor dette intervallet (Baker, 2001). Innenfor IRT vil en testinformasjon på omtrent 10 tilsvare en reliabilitet på 0,90. Nivået 10 er derfor ofte valgt som en kritisk verdi for å evaluere tester innenfor IRT (Hambleton & Lam, 2009).
Presisjonen en test måler dyktighet på et gitt punkt på skalaen er invers relatert til testinformasjonen i dette punktet og kan måles ved å estimere standardfeil (Hambleton et al., 1991):

Formel 9: *Standardfeil ved dyktighetsnivå θ*

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

En standardfeil på 0,33 på dyktighetsscoren vil tilsvare en test informasjon på 10.

Cronbachs koefisient alpha ble også brukt til å måle indre konsistens mellom oppgaver. Cronbachs koefisient alpha gir korrelasjonen av hver oppgave med summen av alle oppgavene. Ligger Cronbachs alpha på > 0,90 indikerer dette svært høy reliabilitet, mens ligger den på < 0,60 indikerer dette uakseptert lav reliabilitet på testen (Tabell 2.5) (Cohen *et al.*, 2011).

3.6 Antall studiepoeng mot andel riktige svar

Når respondentene skulle svare på hvor mange studiepoeng de hadde kunne de velge mellom kategoriene: *mindre enn 60*, *60-120*, *121-180*, *181-240* og *mer enn 240*. For å teste hypotesen om at studentenes dyktighet øker med økende antall studiepoeng utførte vi en enveis ANOVA (eng.: One-way Analysis of Variance), en analyse som tester nullhypotesen (H0) at dyktigheten er likt fordelt mellom gruppene av respondenter med ulik antall studiepoeng. Den alternative hypotesen (H1) blir da at minst en av gruppene har en annen fordeling av dyktighet.

Dersom H0 er sann vil de standardiserte variansene være lik hverandre. Dersom man får en signifikant forskjell vet vi at det er en forskjell i minst to av gruppene, men vi vet ikke hvilke to grupper. Derfor blir en post-hoc Tukey-HSD-test brukt til å se på mellom hvilke grupper det fins en signifikant forskjell.

Kapittel 4 – Resultater

4.1 Innsamling og konstruksjon av flervalgsoppgaver

Resultatet av innsamlingen og konstrueringen av flervalgsoppgaver var en oppgavebank med tilsammen 290 flervalgsoppgaver. 280 av disse oppgavene ble fordelt mellom 7 sett med 40 oppgaver i hvert sett, og de resterende 10 ble lagt til alle 7 settene, slik at alle settene hadde 50 oppgaver tilsammen.

Innsamlingen av flervalgsoppgaver fra internett ble gjort med et kritisk øye; ikke alle oppgavene vi hentet fra nettet oppfylte alle kravene til gode oppgaver (Boks 2.13). Disse oppgavene måtte dermed ofte revideres.

En oppgave vi så som nødvendig å revidere var følgende oppgave:

Boks 4.1: *Oppgave S94 før revidering*

Which of the following statements about species and speciation is true?

A. Hybrids are always selected against in nature.
B. Polyploidy is very rare in plant.
C. Reproductive isolating mechanisms are usually selected against in nature.
D. A single species can undergo adaptive radiation and produce a cluster of species.
E. Species usually have only one type of reproductive isolating mechanism.

Siden vi bare skulle ha 4 svaralternativer ble det siste svaralternativet i oppgaven over vurdert som minst relevant og derfor kuttet ut. I stammen ble "species and" kuttet ut siden vi ønsket en oppgave kun om artsdannelse. "always" ble i det første svaralternativet byttet med "usually", siden hybridisering er ganske vanlig i enkelte planteslekter. Til sist ble "Speciation by" lagt til i begynnelsen av det andre svaralternativet siden "polyploidy is very rare in plants" ikke var et utsagn om artsdannelse slik det opprinnelig var skrevet. Dermed endte vi opp med følgende oppgave etter oversettelse:

Boks 4.2: *Oppgave S94 etter revidering*

Hvilket av følgende utsagn om artsdannelse er korrekt?

A. Hybrider selekteres som oftest mot i naturen.
B. Artsdannelse ved polyploidisering er sjeldent i planter.
C. Reproduktivt isolerende mekanismer selekteres vanligvis mot i naturen.
D. En enkelt art kan gjennomgå adaptiv radiasjon og produsere en mengde nye arter.

En annen oppgave som ble revidert er følgende oppgave:

Boks 4.3: *Oppgave S226 før revidering*

Humid weather makes you feel warmer because humid air, which is saturated with water molecules, _____.

A. interferes with heat loss by evaporation
B. holds warm water vapor
C. interferes with heat by conduction
D. prevents countercurrent heat exchange from occurring
E. increases metabolic heat production

Alternativene A og C ble endret slik at det kom tydelig fram at det er redusert fordampning som fører til et redusert varmetap, og tilsvarende ved konduksjon. Svaralternativet E ble regnet som lite sannsynlig at noen svarer og ble derfor ikke tatt med ettersom vi skulle ha fire svaralternativer på hver oppgave:

Boks 4.4: *Oppgave S226 etter revidering*

Fuktig vær får deg til å føle deg varmere fordi luft med høy luftfuktighet (som er mettet med vann molekyler)

A. reduserer fordampning og dermed varmetapet.
B. inneholder varm vanndamp.
C. reduserer konduksjon (ledning) og dermed varmetapet.
D. forhindrer motstrøms varmeutveksling.

Da vi konstruerte egne flervalgsoppgaver, ble dette gjort ut i fra gitte kriterier (Boks 2.13). Nedenfor er noen av oppgavene vi har laget selv presentert og kategorisert i kategoriene *kunnskap*, *anvendelse* og *vurdering* (Tabell 2.4).

Følgende oppgave er en typisk kunnskapsoppgave, siden oppgaven kun krever at respondenter kan gjenkjenne svaret på spørsmålet:

Boks 4.5: *Egenkonstruert oppgave (S88) i kategorien kunnskap*

Hvilken type bløtdyr (Mollusca) gjennomgår torsjon i utviklingsstadiet?

A. chitoner (Polyplacopoda)
B. snegler (Gastropoda) *
C. blekkspruter (Cephalopoda)
D. sjøtenner (Scaphopoda)

Den neste oppgaven krever at kunnskaper anvendes i en ny kontekst:

Boks 4.6: *Egenkonstruert oppgave (S5) i kategorien anvendelse*

To av funksjonene til den sympatiske divisjonen av det autonome nervesystemet er
A. akselerering av hjertet og hemming av bukspyttkjertelens aktivitet. *
B. stimulering av aktiviteten i magen og sammentrekning av bronkiene i lungene.
C. senkning av hjerterytmen og utviding av pupillene i øynene.
D. stimulering til frigjøring av glukose fra leveren og stimulering av galleblæren.

Oppgaven ovenfor krever først at respondenter har kunnskap om hva den sympatiske divisjonen av nervesystemet er og deretter anvender denne kunnskapen til å velge svaralternativ. Så lenge respondentene vet hva den sympatiske divisjonen av nervesystemet er kan man velge svaralternativ uten å huske funksjonene. Man skal da kunne velge vekk de alternativene som ikke stemmer, slik som *stimulering av aktiviteten i magen* og *senkning av hjerterytmen*.

En annen oppgave som krever at respondentene anvender kunnskap i en ny sammenheng er følgende:

Boks 4.7: *Egenkonstruert oppgave (S77) i kategorien anvendelse*

Hvilken av atferdene kan være et resultat av preging (imprinting)?
A. en hest som løper vekk fra en ukjent lyd
B. en ape som åpner en dør
C. et barn som unngår å ta på en brennesle
D. en sau som oppfører seg som en hund *

I oppgaven over kreves det at respondentene vet hva preging er, og kan bruke dette til å velge den atferden som er et resultat av preging. Dette er ikke svaralternativer som er hentet direkte fra læreboken eller er typiske eksempler, så respondentene må knytte kunnskapen til en ny situasjon.

Oppgaven under krever ikke bare at respondenter skal stille seg positiv eller negativ til en påstand, oppgaven krever også at respondentene gjennom å velge svaralternativ begrunner valget sitt. Oppgaven ble dermed klassifisert som en vurderingsoppgave:

Boks 4.8: *Egenkonstruert oppgave (S214) i kategorien vurdering*

Er alle angiospermer tokjønnede (monoicous)?

A. Nei, fordi blomster kan være ukomplette og mangle fruktemner og/eller pollenbærere. *
B. Ja, fordi alle angiospermer har livssykluser med ett haploid stadium og ett diploid stadium.
C. Ja, fordi alle blomstrende planter produserer både mikrosporer og makrosporer.
D. Nei, fordi ikke alle planter har dobbel befruktning.

En annen oppgave som ble klassifisert som en vurderingsoppgave er følgende oppgave:

Boks 4.9: *Egenkonstruert oppgave (S200) i kategorien vurdering*

Hvilket av følgende utsagn om klimaendringer er riktig?

A. Hull i ozonlaget er hovedårsaken til den nåværende globale oppvarmingen fordi hullene slipper inn mer stråling.
B. Menneskers tilførsel av CO2 til omgivelsene øker den totale mengden av karbon i atmosfæren, noe som kan føre til at karbonsyklusen kommer ut av balanse.
C. En viktig kilde til bevis for klimaendringer kommer fra observasjoner om at gjennomsnittsværet har endret seg for mange regioner. *
D. Klimaet har endret seg mange ganger i den fjerne fortiden, så menneskenes forbrenning av fossilt brensel kan ikke forårsake den nåværende globale oppvarmingen.

4.2 Pre-test

I pre-testen fikk vi syv studenter som studerer biologi ved universitetet til å ta hvert sitt sett med flervalgsoppgaver. Tidsbruken varierte fra 20 til 45 minutter, med et gjennomsnitt på 30 minutter. Tilbakemeldinger fra studenter som tok testen på engelsk, som ikke var deres morsmål, fortalte at disse brukte lenger tid grunnet oversetting av oppgaver, og at en del ord og begreper som er spesifikke innenfor biologi var vanskelig å forstå og ikke mulig å oversette. Studentene skrev også ned kommentarer på oppgavene der de mente noe var uklart eller forslag til endringer. Et eksempel på en kommentar var at svaralternativene i oppgaven nedenfor var for like:

Boks 4.10: *Oppgave S283 i pretest*

Hvordan kan hårstrå i huden gi informasjon om miljøet utenfor kroppen?

A. Bevegelse av håret genererer en spenning som påvirker permeabiliteten til ionekanalene i dendrittene som omringer basen av håret.
B. Bevegelse av håret blir oppdaget av nærtliggende reseptorer, dette påvirker membranpotensialet slik at det resulterer i et aksjonspotensiale.
C. Bevegelse av håret påvirker permeabiliteten til ionekanalene i basen av håret, og dette resulterer i et aksjonspotensiale.
D. Håret fungerer som en mekanoreseptor med ionekanaler som åpnes når håret beveger seg. Dette resulterer i forskjell i membranpotensialet, som igjen kan utløse et aksjonspotensiale.

Svaralternativene ble noe endret etter tilbakemeldingen, og vi endte opp med følgende oppgave:

Boks 4.11: *Oppgave S283 etter revidering*

Hvordan kan hårstrå i huden gi informasjon om miljøet utenfor kroppen?

A. Bevegelse av håret genererer en spenning som påvirker ionekanalene i dendrittene som omringer hårbasen.
B. Bevegelse av håret blir oppdaget av nærliggende reseptorer, dette påvirker membranpotensialet.
C. Bevegelse av håret påvirker permeabiliteten til ionekanalene i hårbasen, og dette resulterer i et aksjonspotensiale.
D. Håret fungerer som en mekanoreseptor med ionekanaler som åpnes når håret beveger seg. Dette resulterer i at membranpotensialet endres.

Ordet *aksjonspotensiale* gikk igjen tre ganger i de originale svaralternativene, mens i de redigerte svarene brukes det kun én gang. Også *permeabiliteten til ionekanalene* går fra å bli brukt to ganger til én gang. Svaralternativene er også kortet ned slik at de er lettere å lese.

På en annen oppgave kom det kommentar på at stammen var uklar:

Boks 4.12: *Oppgave S10 i pretest*

I Mendels' F2-generasjon, hvorfor hadde én av fire planter hvite blomster?

A. Begge foreldrene var heterozygot lilla.
B. Egenskapen er kjønnsbundet.
C. En av foreldrene var recessiv homozygot
D. Begge foreldrene var heterozygot hvit.

Det er ikke sikkert alle kjenner betydningen av *F2-generasjonen*, og oppgaven ville da heller testet om man visste hva dette betydde. Det er ikke poenget med oppgaven og stammen ble derfor endret til:

Boks 4.13: *Oppgave S10 etter revidering*

I ett av Mendels krysningsforsøk med erteplanter ble resultatet at én av fire avkom hadde hvite blomster. Hvilken av de følgende forklarer dette?

A. Begge foreldrene var heterozygot lilla.
B. Egenskapen er kjønnsbundet.
C. En av foreldrene var recessiv homozygot
D. Begge foreldrene var heterozygot hvit.

Vi fikk også tilbakemeldinger på skrivefeil, at enkelte nøkler ikke var klart korrekte og at noen oppgaver inneholdt kompliserte/ukjente ord og begreper og for lange svaralternativer. For ord og begreper som var rapportert som kompliserte, vurderte vi i hvert tilfelle om forståelse av dette ordet inngikk i forventet kompetanse for å løse oppgaven. I de tilfellene hvor dette ikke var tilfelle, gjorde vi endringer slik at det ikke skulle være ukjente ord som spilte en rolle for om respondentene svarte korrekt på oppgavene.

4.3 Datainnsamling

Mellom 9. mars og 14. april, som var perioden flervalgstesten lå ute, var det 713 personer som "klikket" på linken. Av disse var det 173 som ikke svarte på noen av spørsmålene, 187 som kun svarte på de to første spørsmålene om studiepoeng og studiested, 120 som ikke gjennomførte en test 100 %, og 232 respondenter som gjennomførte en test 100 %.

4.4 Analyse

4.4.1 Evaluering av imputeringsteknikk

For å vurdere effektiviteten av imputeringsteknikken ble verdier på dyktighetsskalaen før imputeringen ($andel_{før}$) sammenlignet med tilsvarende verdier etter imputeringen ($andel_{etter}$). Tilsvarende ble verdiene for Cronbach's alpha før imputeringen ($alpha_{før}$) sammenlignet med de tilsvarende verdiene etter imputeringen ($alpha_{etter}$). Resultatet av evalueringen av imputeringsteknikken kan sees i tabell 4.1. Kvadratisk gjennomsnittsavvik (eng.: Root mean square deviation – RMDS) ble beregnet etter Formel 6. Gjennomsnittsverdiene brukes ikke i beregningen av RMSD, men vises for sammenligning.

Tabell 4.1: *Gjennomsnittlige verdier for andelen riktige svar (andel) og Cronsbachs alpha (alpha) før og etter imputeringen. Kvadratisk gjennomsnittsavvik (RMSD) beregnet for differansene $andel_{etter}$-$andel_{før}$ og $alpha_{etter}$-$alpha_{før}$.*

	Gjennomsnitt	RMSD
$andel_{før}$	0,505	**0,114**
$andel_{etter}$	0,489	
$alpha_{før}$	0,756	**0,034**
$alpha_{etter}$	0,744	

Det kvadratiske gjennomsnittsavviket for differansen $aplha_{etter}$-$alpha_{før}$ var 0,034. Det kvadratiske gjennomsnittsavviket for differansen $andel_{etter}$-$andel_{før}$ var 0,114.

Den gjennomsnittlige dyktighetsverdien minket som en følge av imputeringen fra 0,505 til 0,489. Den gjennomsnittlige alpha-verdien minket som følge av imputeringen fra 0,756 til 0,744.

4.4.2 Klassisk oppgaveanalyse

Vi gjennomførte først en klassisk oppgaveanalyse for å undersøke oppgavenes vanskelighetsgrad og diskrimineringsparametre. P-verdiene lå i intervallet 0,051 - 0,971 (Tabell 4.2). Gjennomsnittet var 0,488 med et standardavvik på 0,217. Oppgavenes diskrimineringsevne ble undersøkt ved hjelp av point-biserial korrelasjon, der gjennomsnittet var 0,211 med et standardavvik på 0,061. Oppgaver med point-biserial korrelasjon < 0,15 vil ikke diskriminere tilstrekkelig mellom respondenter og ble fjernet før flere analyser ble gjennomført. 37 av totalt 290 oppgaver hadde point-biserial korrelasjon < 0,15. Disse er vist i Tabell 4.2.

Tabell 4.2: *p-verdi og point-biserialkorrelasjon for oppgaver med point-biserialkorrelasjon < 0,15*

Oppgave	P(x=1|z=0)	Point-biserialkorrelasjon	Oppgave	P(x=1|z=0)	Point-biserialkorrelasjon
S18	0,461	0,133	S140	0,159	0,138
S46	0,858	0,044	S166	0,784	0,120
S49	0,260	0,148	S169	0,805	0,126
S53	0,968	0,052	S172	0,535	0,105

S60	0,901	0,134	S175	0,806	0,131
S64	0,807	0,149	S180	0,762	0,137
S71	0,941	0,127	S183	0,161	0,084
S73	0,812	0,114	S185	0,294	0,139
S74	0,909	0,123	S186	0,322	0,079
S75	0,577	0,148	S190	0,607	0,080
S78	0,779	0,143	S194	0,849	0,095
S83	0,476	0,137	S198	0,132	0,040
S84	0,909	0,027	S204	0,495	0,119
S85	0,345	0,106	S206	0,255	0,119
S90	0,191	0,147	S247	0,779	0,116
S92	0,188	0,134	S257	0,109	0,070
S124	0,100	0,130	S260	0,151	0,134
S125	0,971	0,054	S279	0,522	0,130
S132	0,112	0,102			

4.4.3 Dimensjonalitet

En modifisert parallellanalyse ble brukt til å teste dimensjonaliteten til modellen 1PLM. Den andre egenverdien til de observerte dataene var 4,797. Den andre egenverdien for de syntetiserte dataene var 5,005, med p-verdi på 0,921. Den andre egenverdien fra de observerte dataene er ikke vesentlig høyere enn den andre egenverdien for de syntetiserte dataene og dette indikerer dermed at dataene ikke er flerdimensjonale.

For å teste dimensjonaliteten til modellen 2PLM ble det utført en likelihood-ratio-test mellom to 2PLM-modeller, den ene med én latent variabel, navngitt 2PLM I, og den andre med to latente variabler, navngitt 2PLM II. Tabell 4.3 viser AIC- og BIC-verdiene for 2PLM-testen med én latent variabel og to latente variabler. 2PLM II har både lavere AIC og BIC verdier, selv om det her ikke er en stor differanse. Modellen med én latent variabel har lavest verdier for AIC og BIC og dette støttet dermed bruk av en endimensjonal modell.

Tabell 4.3: *Likelihood-ratio-test mellom 2PLM-modellen med én latent variabel (2PLM I) og 2PLM-modellen med to latente variabler (2PLM II)*

	AIC	BIC	p.value
2PLM I	105729.2	107684.2	
2PLM II	105804.1	108736.6	<0.001

4.4.4 Valg av modell

For å teste mellom modellene 1PLM og 2PLM ble det utført en likelihood ratio test mellom modellene 1PLM og 2PLM. Resultatene er gitt i tabell 4.4. Modellen 1PLM hadde lavere verdier for både AIC og BIC enn modellen 2PLM. Modellen 2PLM viste ikke signifikant forbedring av tilpasning til dataene.

Tabell 4.4: *Likelihood-ratio-test mellom modellene 1PLM og 2PLM*

	AIC	BIC	p,value
1PLM	105463,5	106444,9	
2PLM	105729,2	107684,2	0,723

Resultatene fra oppgavetilpasningsstatistikkene (item-fit) viste at ti oppgaver hadde dårlig tilpasning til dataene i modellen 1PLM, mens 16 oppgaver hadde dårlig tilpasning til dataene i modellen 2PLM, med p-verdi <0,05. Oppgaver med dårlig tilpasning er gitt i tabell 4.5.

Tabell 4.5: *Oppgavetilpasningsstatistikk og p-verdier for oppgaver med p-verdi < 0,05*

1PLM			2PLM		
Oppgave	**X^2**	**Pr(>X^2)**	**Oppgave**	**X^2**	**Pr(>X^2)**
8	21,6054	0,0099	13	19,4191	0,0128
13	22,3638	0,0099	32	19,5047	0,0124
28	18,7385	0,0396	43	16,1797	0,0399
105	15,6685	0,0396	51	19,3829	0,0129
174	21,7486	0,0198	66	16,7128	0,0332

178	20,1406	0,0099	128	17,1729	0,0284
179	16,1481	0,0495	143	18,786	0,016
207	15,6754	0,0297	154	16,3201	0,038
248	21,8548	0,0099	174	19,4204	0,0128
259	17,5732	0,0396	178	19,9822	0,0104
			191	20,2523	0,0094
			230	15,7297	0,0464
			258	21,482	0,006
			259	22,973	0,0034
			276	17,035	0,0297
			289	15,7472	0,0461

Både likelihood-ratio-testen og oppgavetilpasningsstatistikken indikerer at det beste valget av modell er 1PLM. Modellen 1PLM ble valgt og ettersom oppgavene 8, 13, 28, 105, 174, 178, 179, 207, 248 og 259 hadde dårlig tilpasning til denne modellen (Tabell 4.5) ble disse fjernet før videre analyse ble utført.

Egenskapene til oppgavene som ble fjernet på grunn av dårlig tilpasning er gitt i tabell 4.6.

Tabell 4.6: *Vanskelighetsgrad og diskrimineringsparameter for oppgaver med dårlig tilpasning til modellene 1PLM og 2PLM*

	1PLM		**2PLM**	
Oppgave	**Vanskelighetsgrad**	**Diskriminering**	**Vanskelighetsgrad**	**Diskriminering**
S8	-0,319	0,520	-0,267	0,987
S13	-1,771	0,520	-2,581	0,357
S28	-1,580	0,520	-1,724	0,497
S105	-0,626	0,520	-0,53	0,768
S174	-1,421	0,520	-1,473	0,531
S178	-4,447	0,520	-4,366	0,541

S179	-0,626	0,520	-0,973	0,354
S207	2,744	0,520	2,882	0,477
S248	1,259	0,520	0,867	0,73
S259	2,348	0,520	2,421	0,484

4.4.5 Lokal uavhengighet

Lokal uavhengighet ble testet ved Yen´s Q_3 statistikk. Par av oppgaver som hadde en korrelasjon lavere enn -0,20 eller høyere enn 0,20 ble regnet som lokalt avhengige. Resultatene i Tabell 4.7 viser at parene S36 og S150, S267 og S289, S143 og S290 var lokalt avhengige.

Tabell 4.7: *Par av oppgaver som er lokalt avhengige gitt ved Q_3-verdier*

Oppgavepar	Q_3
S36 og S150	-0,213
S267 og S289	-0,241
S143 og S290	-0,208

Tabell 4.8: *Point biserialkorrelasjon og p-verdi for oppgaver som er lokalt avhengige*

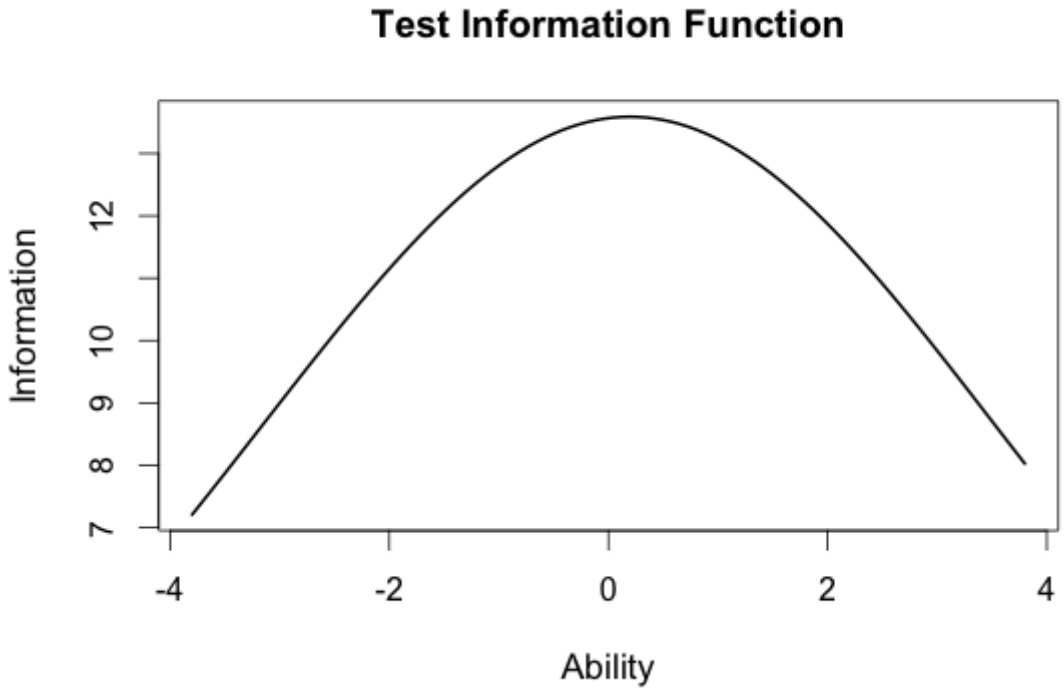
Oppgave	Point biserialkorrelasjon	P(x=1|z=0)
S36	0,227	0,516
S150	0,2312	0,243
S267	0,2655	0,470
S289	0,2839	0,500
S143	0,2158	0,567
S290	0,251	0,449

Oppgave S150 hadde en p-verdi på 0,243, mens oppgave S36 hadde en p-verdi på 0,516. Oppgave S150 ble derfor fjernet fra testen. Blant paret S267 og S289 hadde S267 lavest

point-biserialkorrelasjon, og S267 ble derfor fjernet fra testen. Blant paret S143 og S290 hadde S143 lavest point-biserialkorrelasjon og S143 ble derfor fjernet fra testen.

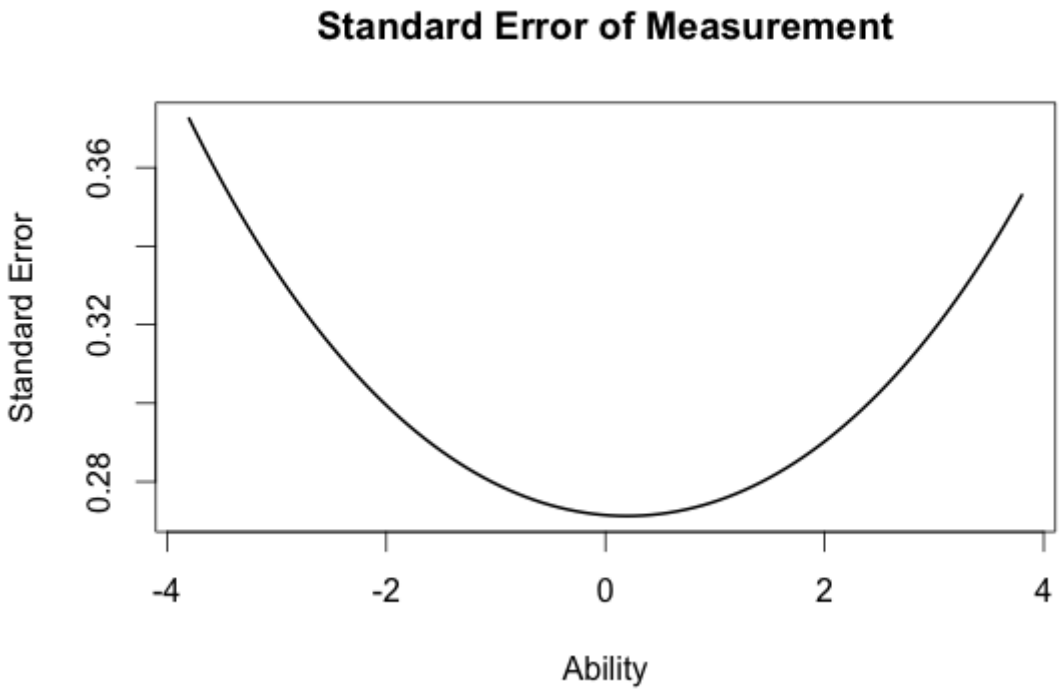
4.4.5 Reliabilitet

Testens informasjonsfunksjon ble brukt til å se på intervallet av dyktighet der testen var reliabel. Figur 4.1 gir kurven til funksjonen og viser at testen var reliabel i dyktighetsintervallet -2,5 til 3, ettersom testinformasjonen da ligger på 10.



Figur 4.1: *Testinformasjonsfunksjonen der x-aksen viser skalaen for dyktighet og y-aksen viser graden av informasjon.*

Siden standardfeilen er invers relatert til testinformasjonen i et gitt punkt på dyktighetsskalen viser figur 4.2 også at testen er mest reliabel for respondenter med et dytighetsnivå mellom -2,5 og 3. Testen er mest reliabel på et dyktighetsnivå på like over 0, altså rett over gjennomsnittet.



Figur 4.2: *Graf for standardfeil, der x-aksen viser skalaen for dyktighet og y-aksen viser graden av standardfeil.*

Cronbachs alphakoeffisient ble kalkulert for de ulike settene hver for seg før analyser og fjerning av oppgaver ble utført, i tillegg til å bli kalkulert for den totale testen etter at analyser og utvelgelse av oppgaver ble utført. Tabell 4.9 viser disse resultatene og for den totale testen var en alpha-verdi på 0,932 som indikerer at testen hadde svært høy reliabilitet.

Tabell 4.9: *Cronbachs koeffisient alpha gitt for sett 1-7 og for endelig datasett*

Datasett	Cronbachs koeffisient alpha
Sett 1	0,84
Sett 2	0,66
Sett 3	0,735
Sett 4	0,765
Sett 5	0,804
Sett 6	0,668
Sett 7	0,816

Endelig datasett	0,932

4.5 Beskrivelse av endelig produkt

Ved en siste gjennomgang av oppgavene ble det oppdaget at oppgave S195 hadde en feil i nøkkelen, og denne oppgaven ble derfor fjernet. Som et endelig produkt endte vi opp med 239 oppgaver, fra et utgangspunkt på 290 oppgaver. Fordelingen av antall oppgaver per tema er gitt i tabell 4.10. Innenfor hvert tema (ut i fra temastrukturen i Campbell Biology) er det også gitt antallet oppgaver som er kategorisert i hver av kategoriene *kunnskap*, *anvendelse* og *vurdering*.

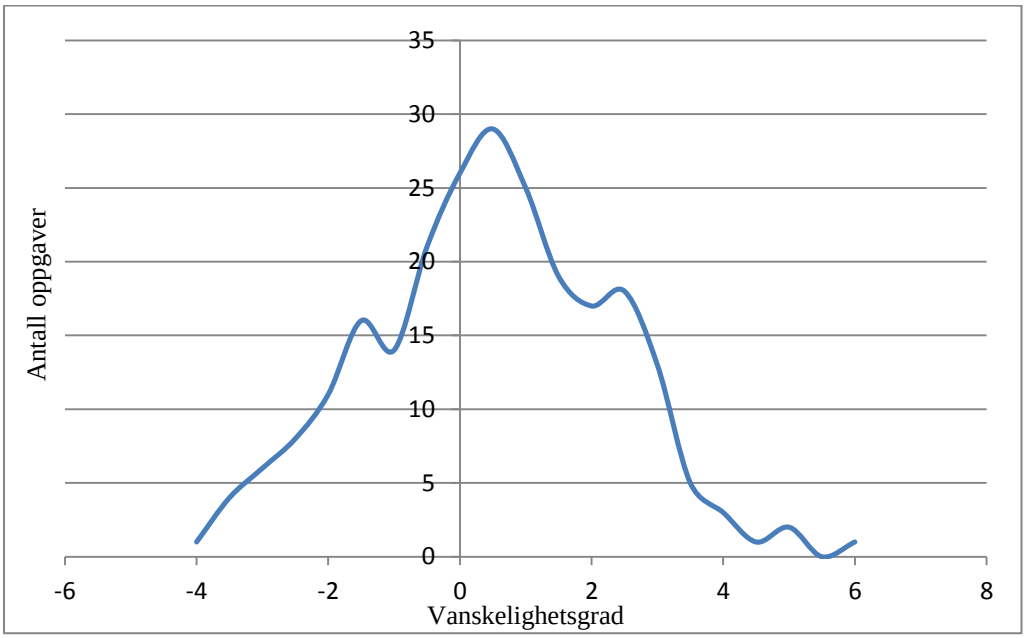
Tabell 4.10: *Oversikt over antall kunnskaps-, anvendelses- og vurderingsoppgaver i hvert av temaene i Campbell Biology i det endelige produktet*

Tema	**Antall oppgaver *kunnskap***	**Antall oppgaver *anvendelse***	**Antall oppgaver *vurdering***	**Antall oppgaver totalt**
Livets kjemi	2	0	1	3
Cellen	24	3	3	30
Genetikk	22	10	9	41
Evolusjonære mekanismer	9	6	8	23
Den biologiske diversitetens evolusjonære historie	35	4	3	42
Planters oppbygning og funksjon	16	8	7	31
Dyrs oppbygning og funksjon	30	15	10	55
Økologi	9	2	3	14
Total	147	48	44	239

Tabell 4.10 viser at det er svært få oppgaver innunder temaet livets kjemi. Tidlig i prosessen valgte vi, for å begrense antall sett og antall oppgaver i hvert sett, vekk en del av oppgavene

som var samlet innenfor dette temaet til fordel for andre oppgaver som var mer direkte tilknyttet biologi. Totalt sett er det 147 oppgaver som går direkte på kunnskap, mens 48 oppgaver går på anvendelse av kunnskap og 44 oppgaver på vurdering av kunnskap.

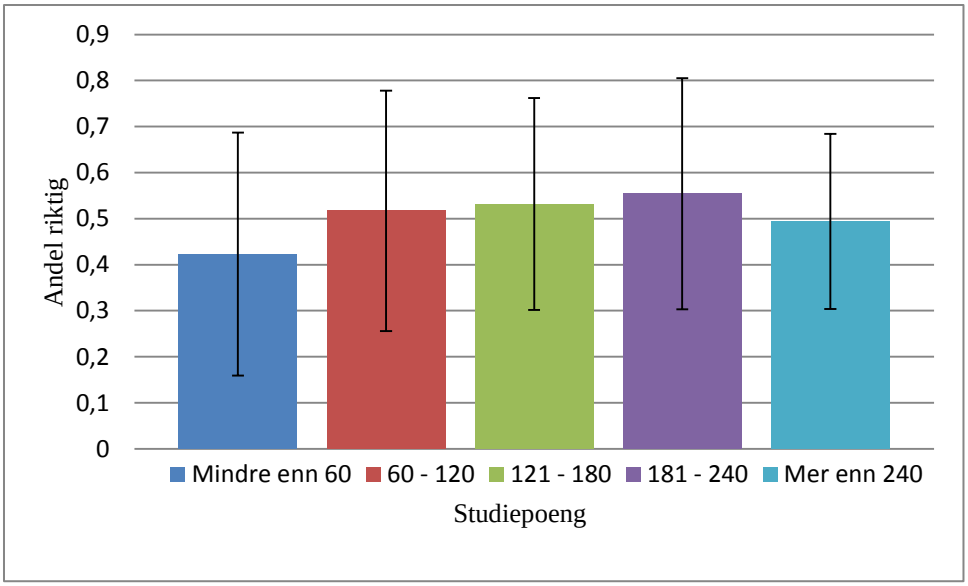
Fordelingen av antall oppgaver i forhold til vanskelighetsgrad på oppgavene er vist i figur 4.3 Her ser vi at det er en overvekt av oppgaver på vanskelighetsgrad litt over 0.



Figur 4.3: *Fordeling av antall oppgaver etter vanskelighetsgrad*

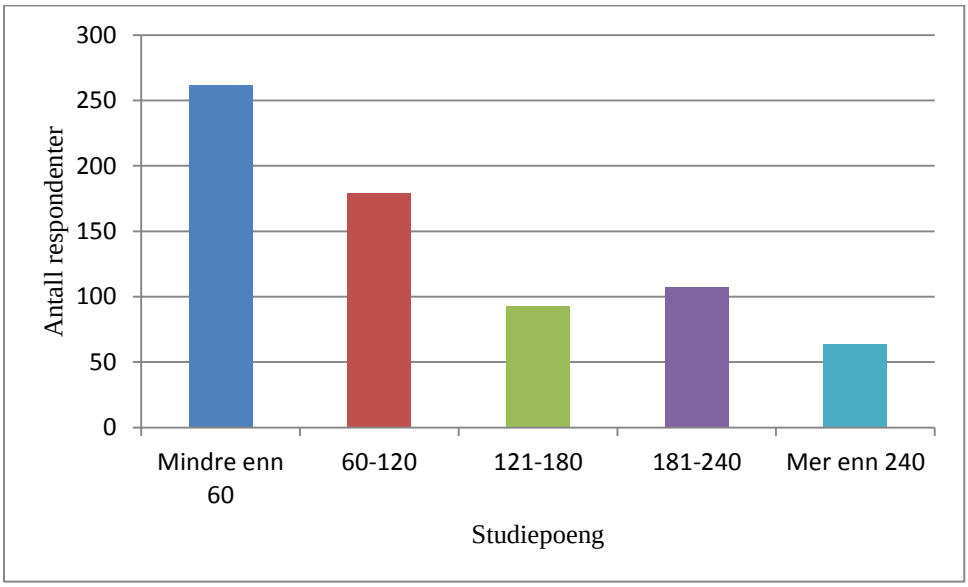
4.6 Studentenes dyktighet i forhold til avlagte studiepoeng

Fordelingen av dyktighet (andel riktige svar) mellom grupper av studenter med ulike antall studiepoeng er vist i figur 4.3. Figuren viser at den gjennomsnittlige andelen riktige svar øker med antall studiepoeng respondentene har, med unntak av den siste gruppen der respondentene har mer enn 240 studiepoeng, altså at respondentene er på siste året på masterstudiet dersom de følger et normalt studieløp. Konfidensintervallet er vist i figuren for hver gruppe med respondenter.



Figur 4.4: *Oversikt over gjennomsnittslig andel riktige svar i de ulike gruppene med studiepoeng*

Figur 4.5 viser hvor mange respondenter det er innenfor hver gruppe fordelt etter antall studiepoeng. Det er klart flest respondenter som har mindre enn 60 studiepoeng, og færrest respondenter med mer enn 240 studiepoeng.



Figur 4.5: *Oversikt over antall respondenter fordelt på de ulike gruppene studiepoeng*

For å teste om det er en signifikant forskjell mellom gruppene med respondenter, ble det utført en one-way ANOVA test. Analysen viste en signifikant forskjell mellom minst to av gruppene av respondenter (p-verdi = 9,392e-08).

En Tukey HSD test ble utført for å se hvilke grupper av respondenter som varierte. Tabell 4.12 viser at gruppen som har mindre enn 60 studiepoeng skiller seg ut med lavere dyktighet enn de andre.

Tabell 4.11: *Variasjon mellom ulike grupper av studiepoeng undersøkt opp mot hverandre ved Tukey-HSD-test*

Grupper av studiepoeng undersøkt mot hverandre	Differanse	p-verdi
181-240 mot 121-180	0,027	0,885
60-120 mot 121-180	-0,016	0,973
Mer enn 240 mot 121-180	-0,022	0,973
Mindre enn 60 mot 121-180	**-0,115**	**< 0,05**
60-120 mot 181-240	-0,043	0,558
Mer enn 240 mot 181-240	-0,049	0,692
Mindre enn 60 mot 181-240	**-0,142**	**< 0,05**
Mer enn 240 mot 60-120	-0,006	0,999
Mindre enn 60 mot 60-120	**-0,099**	**< 0,05**
Mindre enn 60 mot Mer enn 240	**-0,093**	**< 0,05**

Kapittel 5 – Diskusjon

Målet med denne masteroppgaven er å utvikle et kvalitetssikret verktøy som kan teste studenters kompetanse innenfor biologi, og at verktøyet skal kunne brukes til forskning på læring, utbytte av undervisning og studenters progresjon gjennom studiet. Det er derfor viktig at metodevalg og resultater blir diskutert opp mot teori for å kunne gi en vurdering av hvor godt verktøyet egner seg til dette. Dette kapittelet starter med en diskusjon av metoden som ble brukt for å samle inn og distribuere flervalgstestene til studentene. Etter dette følger en gjennomgang av funnene gjort i analysene i lys av teori om IRT og teori om flervalgsoppgaver. Validiteten og reliabiliteten til verktøyet blir diskutert og til slutt kommer vi med anbefalinger for videre bruk av verktøyet.

5.1 Diskusjon av metode:

5.1.1 Innsamling av flervalgsoppgaver

Det å samle inn flervalgsoppgaver fra nettet har både fordeler og ulemper fremfor å lage oppgaver selv. En fordel er at man sparer mye tid, og man får dermed mulighet til å lage en større bank med oppgaver. En annen fordel er at man kan finne eksempler på oppgaver som ikke bare tester reproduksjon. Slike oppgaver er vanskelig å lage og det å få idéer til hvordan man kan utforme slike oppgaver kan være svært nyttig. I tillegg kan det være nyttig for å se hvordan man kan skape variasjon blant oppgavene. En ulempe med å samle oppgaver fra internett er at disse ofte må redigeres. Det er to grunner til det: For det første kan kvaliteten på mange av oppgavene som ligger på internett være ganske lav i forhold til hvordan flervalgsoppgaver bør skrives (Boks 2.13); For det andre må oppgavene passe innholdsmessig og nivåmessig til det man ønsker å måle med testen. Dersom man benytter seg av internettsider tilknyttet diverse læreverk vil det alltid være en sjanse for at noen av respondentene har vært inne på disse sidene og besvart noen av oppgavene tidligere. I vårt tilfelle var nok sjansen stor for at noen av respondentene hadde vært inne på siden tilknyttet Pearson Education Ltd©, siden dette læreverket benyttes ved UiB. Men ettersom oppgavene vi hentet fra denne nettsiden var blandet med andre oppgaver, og i tillegg redigert, vurderte vi det som lite sannsynlig at dette ville påvirke resultatet i vår utprøving ved at respondentene til vår test kjente igjen oppgavene og dermed husket hvilket svar som var riktig.

5.1.2 Matrisesamling

Oppgavebanken ble delt inn i mindre sett med utgangspunkt i resultatene fra pre-testene. Pre-testene inneholdt 50 oppgaver, og respondentene brukte gjennomsnittlig 30 minutter på å gjennomføre dem. Dette var en indikasjon på at 50 flervalgsoppgaver per sett var tilstrekkelig siden ønsket var at testene ikke skulle ta mer enn 30 minutter å gjennomføre. Det er mulig at færre oppgaver per sett kunne ha økt svarprosentene testene. For d å få testet ut samme antall oppgaver, måtte vi ha delt oppgavebanken inn i flere sett. Da ville hver enkeltoppgave til blitt besvart av færre respondenter, og dette ville ført til en større statistisk usikkerhet knyttet til analysen av oppgavene. Et annet alternativ ville selvfølgelig vært å redusere det totale antall oppgaver men da kunne man risikert og ikke få testet ut potensielt verdifulle oppgaver.

Inndelingen av flervalgsoppgaver i mindre sett har både fordeler og ulemper. Fellesoppgaver er fordelaktig fordi det øker sammenlignbarheten mellom de forskjellige settene (Dings, Childs & Kingston, 2002). I tillegg vil man kunne bruke resultatene fra hvert sett til å beregne universale statistikker, som om man hadde distribuert alle testleddene til alle respondentene (Shoemaker & Shoemaker, 1981). Item-sampling-modellen (Tabell 2.1) ble valgt bort på grunn av den dårlige sammenlignbarheten mellom settene (Aningbo, 2011). Genuine-matrix-modellen (Tabell 2.3) krever at antall respondenter per oppgave er høyt (Childs & Jaciw, 2003), og derfor vurdert som lite aktuell for vår uttesting. Partial-matrix-modellen (Tabell 2.2) ble valgt fordi flere oppgaver kunne bli testet, samtidig som testene ikke var avhengige av like mange respondenter sammenlignet med genuine-matrix-modellen. En fordel med genuine-matrix-modellen er likevel at hver oppgave besvares av flere respondenter, og at flere oppgaver fungerer som koblingspunkter mellom de ulike settene (Childs & Jaciw, 2003). Dette øker i sin tur sammenlignbarheten mellom oppgavene. På grunn av dette ble vi anbefalt å bruke nettopp med denne modellen, men denne anbefalingen kom først etter at vi hadde distribuert flervalgstestene. Dersom vi skulle utviklet flervalgstesten på nytt, ville vi vurdert å bruke genuine-matrix-modellen, slik at sammenlignbarheten mellom oppgavene og settene hadde blitt større.

5.2.3 Distribusjon via SurveyXact (SX)

En ulempe med å bruke SX til vårt formål var at systemet ikke var laget for flervalgstester som skulle ha *ett* korrekt svaralternativ (dikotome data). Det var derfor ikke mulig å merke av hvilket av svaralternativene som var korrekt.. Dette var noe vi ønsket for at respondentene

etter hvert spørsmål kunne få vite om de hadde svart riktig eller galt, for å øke motivasjonen for å gjennomføre testene. Slik SX var laget var dette ikke mulig. Dette ble løst ved å legge inn en videresending etter fullført test, hvor respondentene fikk opp en egen side med fasiten for oppgavene. Vi tror at enda flere ville fullført testen dersom respondentene hadde fått opp fasit direkte etter hver oppgave, og ville nok vurdert et annet program dersom vi skulle valgt på nytt. Problemet er at mange slike systemer/programmer er kommersielle, og vi måtte da velge mellom å bruke SX, som vi fikk tilgang til via UiB, eller bruke et annet gratis system, hvor forhold som reklame og kontroll på dataene var usikre faktorer.

Forholdet at det ikke var mulig å merke av hvilket svaralternativ som var riktig hadde også implikasjoner for datamatrisen som skulle danne grunnlaget for analysen. Datamatrisen som ble generert i SX kom bare til å inneholde de verdiene som de ulike svaralternativene ble tildelt (1 til 4) samt hva de ulike respondentene hadde svart, men ikke hvilket svaralternativ som var det riktige på hvert spørsmål. Dette var en ulempe siden vi i analysen ønsket en enkel måte å se om hver respondent hadde svart riktig eller galt. Dette ble løst ved å legge inn det korrekte svaralternativet som første svaralternativ i datamatrisen. Her måtte vi selvsagt sørge for at de ulike svaralternativene i testen ble randomisert for hver respondent, slik at det riktige svaralternativet ikke kom først i hver eneste oppgave.

En fordel med systemet var at det var flere personer på UiB med kunnskap om systemet, slik at vi hadde mulighet til å få direkte hjelp til tekniske problemer. SX hadde også den fordelen at flervalgstestene var enkle å distribuere blant studentene, og at vi lett kunne eksportere data etter at testen var gjennomført.

I etterkant har vi sett på dataene fra testen at flere respondenter valgte å avslutte før de hadde fullført alle oppgavene. En mulig grunn til dette er at de synes testen var for lang. En annen mulig årsak er at respondentene ikke var forberedt på at linken de fikk tilsendt var en kunnskapstest. Dette kunne vi nok ha informert bedre om. Enkelte av oppgavene på de første sidene kunne nok i tillegg oppleves som ganske vanskelige, selv om vi flyttet noen av de lettere oppgavene frem til den første siden, og dette kan ha svekket motivasjonen for å gjennomføre resten av oppgavene. For videre bruk av oppgavebanken bør antall oppgaver i hvert sett og rekkefølgen på oppgavene vurderes nøye. Dette må selvfølgelig også vurderes opp mot hvilke områder som skal testes.

Vi valgte å sende ut testene til flere universiteter og høyskoler, utenom UiB, for å øke antall respondenter. Respondentene har derfor noe ulik bakgrunn, men siden testen bare ble sendt ut til de som studerer biologi bør ikke forskjellen være så stor. Det er og mulig de bruker andre lærebøker, men siden testen skal teste grunnleggende biologikompetanser så vi ikke på dette som et stort problem.

Et annet initiativ som ble gjort for å øke svarprosenten på testene våre var å få støtte til kjøp av en iPad som gevinst til en tilfeldig uttrukket respondent. Slike gevinster har vist seg å ha en effekt på respondenters motivasjon for å besvare og gjennomføre undersøkelser, selv om denne effekten er begrenset (Singer & Ye, 2013). I ettertid så ser vi allikevel at vi kunne gjort en større innsats for å fått sponset en mer attraktiv og mindre ordinær gevinst enn en iPad. Siden testene i var biologikunnskapstester så så vi på det som svært aktuelt å få støtte til kjøp av et hobbymikroskop. Dessverre ble ikke dette aktuelt. Et hobbymikroskop noe vi tror ville slått an hos respondentene (som jo er biologistudenter). Dette er noe å vurdere for fremtidig bruk av denne testen dersom målet er å få en høy svarprosent.

5.2 Diskusjon av analyser

5.2.1 Estimering av manglende verdier gjennom imputering

Det kvadratiske gjennomsnittsavviket for differansen $alpha_{etter}$-$alpha_{før}$ var 0,034 (Tabell 4.1). Dette indikerte at endringen i oppgavenes indre konsistens som følge av imputeringsteknikken var 3,400 %. Dette er et relativt lavt avvik, men allikevel en kilde til usikkerhet.

Det kvadratiske gjennomsnittsavviket for differansen $andel_{etter}$-$andel_{før}$ var 0,1138 (Tabell 4.1). Dette indikerte at den gjennomsnittlige endringen i respondentenes dyktighet som følge av imputeringen var 11,38 %. Dette er et noe høyere avvik, og en kilde til usikkerhet. En interessant observasjon er at imputeringsteknikken tilsynelatende reduserer det gjennomsnittlige antallet riktige svar respondenten har (at imputeringsteknikken gjør respondentene "dummere").

Avvikene mellom verdiene før og etter imputeringen indikerer at imputeringsmetoden kan ha vært problematisk. Det er ikke så ulogisk siden det å eksempelvis estimere 285 verdier ut av kun 5 besvarte oppgaver vil skape en viss usikkerhet. En annen metode vi kunne ha brukt for å vurdere imputeringen var å gjennomføre noen endimensjonale rasch-analyser på det

originale datasettet, samt på det imputerte datasettet, og vurdere om dyktighetsestimatene (θ) var tilstrekkelig like. Dersom dyktighetsestimatene hadde vært dårlige så burde imputeringsmetoden ha vært gjentatt flere ganger for å oppnå et sett av dyktighetsverdier Chalmers (2012). Dette ble ikke gjort ettersom nødvendigheten av å bruke imputering ble oppdaget for sent. Dersom vi skulle gjort nye imputeringer så ville vi fulgt Chalmers råd og beregnet ett sett med dyktighetsestimater.

5.2.2 Flervalgsoppgaver som ble tatt bort i analysen

Det første steget i analysen hvor flervalgsoppgaver ble fjernet fra databasen var etter at point-biserialkorrelasjonene var blitt undersøkt. Point-biserialkorrelasjoner gir en indikasjon på hvor godt oppgaver diskriminerer mellom sterke og svake respondenter, men point-biserialkorrelasjonen må ikke forveksles med diskrimineringsfaktoren i 2PLM (Kavitha *et al.*, 2012). Det kan være mange faktorer som kan føre til at oppgaver diskrimerer dårlig. Noen ganger kan årsaken til den lave point-biserialverdien være én slik faktor, andre ganger kan det være en kombinasjon av flere. Det er verdt å merke seg at selv om det kan hypotetiseres om hva som var årsakene til de lave point-biserialverdiene, så kan man aldri vite noe sikkert uten å se nærmere på hvilke respondenter som svarte hva. I tillegg så vil oppgaver ofte ligge i gråsoner hvor det vil kunne være vanskelig å hypotetisere noe i det hele tatt.. Til slutt er det verdt å nevne at point-biserialskalaen i seg selv kan inneholde gråsoner, spesielt for oppgaver med point-biserialverdier rundt 0,15. Jo lavere en point-biserialverdi er, jo mer alvorlig er diskrimineringsfeilen.

En mulig årsak til at flere av oppgavene diskriminerte dårlig kan være at oppgavene enten hadde svært høye p-verdier eller svært lave p-verdier, altså at oppgavene ble korrekt besvart av for mange eller for få til at de gav noen særlig statistisk informasjon (Tabell 4.2). Slike oppgaver vil ofte diskriminere dårlig. I eksemplenene i dette avsnittet er svarprosenten på de ulike svaralternativene gitt i parentes og nøkkelen symbolisert med en *:

Boks 5.1: *Oppgave S53 som ble fjernet på grunn av point-biserialkorrelasjon <0,15*

S53: Hva er funksjonen til rothår på planterøtter?

A. De øker absorbsjonen av vann og mineraler. (*96*) *
B. De gir økt opptak av CO_2 og økt utslipp av O_2. (*0*)
C. De fungerer i vegetativ reproduksjon. (*0*)
D. De lagrer mat og vann. (*4*)

96,4 % av respondentene som besvarte oppgave 53 besvarte oppgaven korrekt. Bare 3,6 % av respondentene svarte feil på denne oppgaven. I dette tilfellet, så kan den lave point-biserialverdien (0,052) indikere at oppgaven rett og slett var for lett for nivået den ble testet på. På et lavere nivå ville oppgaven kanskje ha diskriminert bra. Det trenger altså ikke ha noe med mangler ved oppgaven å gjøre (Osterlind, 2002).

En annen årsak til at flere av oppgavene diskriminerte dårlig kan være dårlig bruk av språk. Dårlig språkbruk kan gå på tekstmengden eller innholdet, og vil enten kunne øke sannsynligheten for at respondenter gjetter ved at oppgaver fremstår som uoverkommelige eller mindre forståelig, eller redusere vanskelighetsgraden ved at oppgaver inneholder elementer som kan gi hint om hvilket svaralternativ som er nøkkelen (Brame, 2015; Sirnes, 2005).

Den lave point-biserialverdien på oppgave 198 (0,040) indikerer at store mengder tekst kan være problematisk:

Boks 5.2: *Oppgave S198 som ble fjernet på grunn av point-biserialkorrelasjon <0,15*

S198: På hvilken måte fører Casparianske bånd i planterøtter til at vann og mineraler må passere gjennom plasmamembranen i endodermale celler før det kan gå inn i den vaskulære sylinderen?

A. Casparianske bånd sørger for at vann og mineraler som går inn i roten før eller siden må ta den apoplastiske ruten, og dermed passere en plasmamembran, før det går inn i endodermalcellene. (*29*)
B. Prosent Casparianske bånd er lokalisert i celleveggene til- og mellom alle endodermalceller og blokkerer passeringen av vann og løste mineraler. (*16*) *
C. Casparianske bånd er selektivt permeable membraner som ligger utenfor endodermis, og slipper inn det planten trenger av vann og mineraler direkte inn til plasmamembranen i endodermalcellene. (*32*)
D. Casparianske bånd ligger på innsiden av endodermale celler og inneholder store mengder løste ioner som danner en gradient som tvinger vann inn i de endodermale cellene ved diffusjon (mineraler følger etter ved kotransport). (*23*)

I oppgave 198 må man huske ordlyden i spørsmålet samtidig som man skal evaluere hvilket av de lange svaralternativene som er riktig. Det at oppgaven kan være naturlig vanskelig og at svaralternativene er relativt like gjør det ikke akkurat lettere. Dette kan virke demotiverende, og kan føre til en større grad av gjetting (Lykknes & Smidt, 2009). En annen indikasjon på at

respondenter kan ha gjettet på oppgave 198 er at alle de ulike svaralternativene er relativt godt representert.

Mens tekstmengden var det tilsynelatende største problemet med oppgave 198, så kan den lave point-biserialverdien i oppgave 124 (0,130) knyttes til innholdet, og da spesielt ordet "galt" i spørsmålsteksten og ordet "bare" i det første svaralternativet:

Boks 5.3: *Oppgave S124 som ble fjernet på grunn av point-biserialkorrelasjon <0,15*

S124: Hva er GALT om telomerer?

A. Telomerer inneholder flere repetisjoner av korte nukleotidesekvenser. (*14*)
B. Normal forkortning av telomerer kan beskytte organismer mot kreft. (*50*)
C. I eukaryote kimceller vil telomerer gjenopprette sine originale lengder, katalysert av enzymet telomerase. (*25*)
D. Telomerer inneholder bare få gener. (*11*) *

Svaralternativer som inneholder absolutte utsagn har en tendens til å bli bortvalgt av respondenter, siden "det finnes alltid et unntak" kan være en naturlig tankerespons (Haladyna *et al.*, 2002). Dette er uheldig siden respondenter ledes bort fra nøkkelen i oppgave 124 på grunn av en dårlig formulert setning og ikke på grunn av manglende kunnskap. I tillegg kan negasjonen i spørsmålsteksten lett overses, og respondenter kan bli "lurt" til å hake av på noen av distraktorene da disse setningene oppleves som riktige, siden påstandene i og for seg er korrekte, men ukorrekte i forhold til det oppgaven spør om.

Et annet eksempel på at dårlig språkbruk kan ha negative konsekvenser er oppgave 132:

Boks 5.4: *Oppgave S132 som ble fjernet på grunn av point-biserialkorrelasjon <0,15*

S132: Eukaryote cellevegger, til forskjell fra prokaryote, består av

A. en karbohydrat-matriks, kryssbundet av korte polypeptider. (*15*)
B. glykolipider og proteinfibre. (*59*)
C. kitin. (*13*)
D. cellulosefibre som ligger i en matriks. (*13*) *

Det kan være flere årsaker til den lave point-biserialverdien i oppgave 132 (0,102). For det første så kan flere respondenter ha forvekslet "cellevegg" med "cellemembran", noe som isåfall styrker svaralternativ 3. For det andre kan "til forskjell fra" i stammen lett overses, noe som isåfall også styrker svaralternativ 3. Som om ikke det var nok; Den tenkte nøkkelen er

ikke det eneste korrekte svaret på oppgaven. Både planter og sopp er eukaryoter, men celleveggene i planteceller består av cellulose (blant annet) og celleveggene i soppceller består av kitin (blant annet). Dermed er både svaralternativ 1 og svaralternativ 4 korrekte svar på oppgaven. Dette kan være både forvirrende og misledende.

Et annet innholdsproblematisk aspekt ved dårlig språkbruk er heterogene svaralternativer. I oppgave 73 kan den lave point-biserialverdien (0,114) kobles til en høy p-verdi, men også svaralternativer koblet på en heterogen måte:

Boks 5.5: *Oppgave S73 som ble fjernet på grunn av point-biserialkorrelasjon <0,15*

S73: Hvilket av de følgende alternativene forklarer best hvorfor det å ha både klorofyll a og b er bedre for planter, enn å bare ha ett av de?

A. Klorofyll a er primært involvert i elektronoverføring; klorofyll b er i hovedsak involvert i absoberingen av lysenergi. (*13*)
B. Klorofyll a og b absorbererer i noe ulike deler av det elektromagnetiske spekteret; dette øker bredden av bølgelengder som kan bli absorbert. (*81*) *
C. Klorofyll a fanger all solenergien; Klorofyll b beskytter klorofyll a fra overeksitasjon. (*0*)
D. Klorofyll a er i tylakoidmembranen og fanger lysenergi der; klorofyll b er løst i stroma i kloroplasten og fanger lysenergi der. (*6*)

Det riktige svaralternativet i oppgave 73 er klart overrepresentert. Det kan selvsagt ha noe med oppgavens vanskelighetsgrad å gjøre. Men det kan også ha noe med koblingen mellom de tre andre svaralternativene å gjøre. De tre siste svaralternativene inneholder alle formuleringer som ligner på "Klorofyll a gjør X; Klorofyll b gjør Y". Svaralternativ 1 er ikke formulert på denne måten. En slik 3:1-fordeling er uheldig. Ideelt sett så bør det ikke være noen kobling mellom svaralternativene, men dersom det finnes en kobling, så bør alle svaralternativene kunne kobles med minst ett annet svaralternativ (Brame, 2015; Kehoe, 1995; Haladyna *et al.*, 2002; Sirnes, 2005).

Lurespørsmål kan påvirke besvarelsen av flervalgsoppgaver i en negativ retning ved at respondenter blir lurt til å velge en av distraktorene:

Boks 5.6: *Oppgave S53 som ble fjernet på grunn av point-biserialkorrelasjon <0,15*

S257: 1,000,000 J solenergi holder vanligvis til rundt 10,000 J primærprodusenter, 1,000 J primærkonsumenter, 100 J sekundærkonsumenter og rundt 10 J tertiærkonsumenter. Dette viser at

A. organismer er ineffektive til å konvertere energien de konsumerer til biomasse. (*12*) *
B. de trofiske effektivitetene er rundt 10 % for alle trofiske nivåer. (*80*)
C. produsenter (for eksempel planter) ofte er tyngre enn konsumenter (for eksempel fugler). (*0*)
D. toppnivåpredatorer bruker veldig mye energi på å fange byttedyr. (*8*)

Det er kanskje ikke så åpenbart hvorfor svaralternativ 1 er (den tenkte) nøkkelen her og ikke svaralternativ 2 som hele 80 % av respondentene har trodd. Tanken bak oppgaven var at den første overgangen fra 1.000.000 J til 10.000 J innebærerer en trofisk effektivitet på 1% og ikke 10 % (10.000/1.000.000=0,01). Det viste seg at svært få respondenter fikk tak i denne ideen. Dette er negativt siden man da blir avhengig av å spotte en liten detalj for å besvare spørsmålet riktig. Besvarelsene skjer dermed ikke på grunnlag av dyktighet, men på grunnlag av hint i feil retning, eventuelt manglende hint i riktig retning.

Oppgave 257 illustrerer også et annet dilemma, nemlig at ukritisk definisjonsbruk kan skape uklarheter rundt hvordan oppgaver leses, forstås og besvares. For hva er egentlig et trofisk nivå? Og hva er egentlig trofisk effektivitet? For hver 1.000.000 J solenergi primærprodusentene tar opp gjennom fotosyntese så overføres 10.000 J til primærkonsumentene. Dette tilsvarer selvsagt en energieffektivitet på 1%, men solenergien kan ikke akkurat karakteriseres som et trofisk nivå, så kan man eller kan man ikke si at primærprodusentene har en *trofisk* effektivitet på 1%? Eller begynner denne regnemåten først ved overgangen fra primærprodusenter til primærkonsumenter? Dette viser at ulike måter å tenke på kan medføre at ulike respondenter velger ulike svaralternativer. At ulike svaralternativer kan regnes som riktig svar avhengig av hvordan man leser oppgaven bør man selvsagt unngå.

5.2.3 Dimensjonalitet

En viktig antagelse for bruk av 1PLM-modellen er at dataene er tilstrekkelig endimensjonale, ettersom studier har vist at dataene ikke trenger å være fullstendig endimensjonale, men tilstrekkelig endimensjonale (Budescu *et al.*, 1997). Vi ønsket at testen skulle måle biologikompetanse, men andre mulige dimensjoner kan for eksempel være språkforståelse. Vi

brukte en modifisert parallellanalyse for å sjekke om dataene var tilstrekkelig endimensjonale. Metoden har vist seg å være nyttig og bedre egnet enn tradisjonelle metoder som eksempelvis faktoranalyser (Drasgow and Lissak, 1983). Nye metoder er hele tiden under utvikling og Budescu *et al.* (1997) påpeker noen ulemper ved bruk av modifiserte parallellanalyser. Blant annet mangler modellen en mekanisme for å luke ut spørsmål som fører til at testen er mer flerdimensjonal, og da skape et undersett med spørsmål, der undersettet er endimensjonalt. Resultatene våre indikerer at bruk av en endimensjonal modell passer til dataene våre (Tabell 4.3), men vi bruker da altså ingen metode til å identifisere oppgaver som kunne vært fjernet for å skape et enda mer endimensjonalt datasett. Dersom analysene hadde vist at dataene var flerdimensjonale, kunne vi undersøkt om for eksempel leseferdigheter også ble testet. Noen av studentene tok testen på et annet språk enn sitt eget morsmål, og da kan evnen til å forstå innholdet i testen ha spilt inn. Hvis dette skal undersøkes senere vil det være nødvendig med et større antall respondenter som tar testen på et annet språk enn sitt morsmål, eventuelt la norske studenter ta testen på begge språk og deretter sammenligne resultatene. En slik undersøkelse kan være aktuell i forhold til fag som blir undervist i ett språk, men har eksamen i et annet, eller andre varianter.

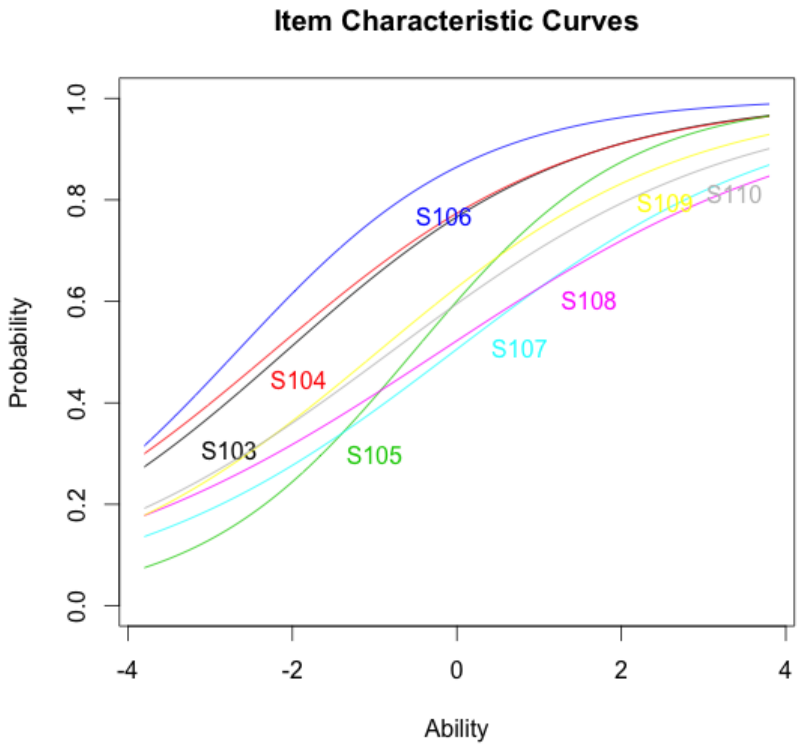
5.2.4 Valg av modell

Ved data som er tilstrekkelig endimensjonale kan det innenfor IRT brukes ulike modeller. 1PLM som er den enkleste modellen tar bare med én parameter, vanskelighetsgrad. 2PLM tar med en ekstra parameter, diskrimineringsparameteren, mens 3PLM som er den mest komplekse modellen av disse tre tar med gjetting som en tredje parameter. Data fra virkeligheten vil sjelden møte antagelser perfekt og innebærer ofte at oppgaver diskriminerer ulikt og at respondenter gjetter på noen oppgaver (Sick, 2010). Utfra denne beskrivelsen ville vi ved første øyekast valg den mest komplekse modellen ettersom det alltid vil være en sjanse for at noen av respondentene gjetter på noen av spørsmålene. Men ettersom kravet til utvalg endrer seg mellom de ulike modellene, var det ikke mulig å utføre analyser innenfor 3PLM siden denne modellen krever minst 1000 respondenter for gode estimater (Lord, 1968).

For å velge mellom 1PLM og 2PLM ble det brukt to ulike metoder, en likelihood ratio test og en sjekk av oppgavetilpasningsstatistikk (item fit). Likelihood ratio testen viste at 2PLM ikke var bedre egnet enn 1PLM (Tabell 4.4). Ettersom modeller med flere parametere generelt er bedre tilpasset dataene, var dette et overraskende resultat (De Ayala, 2009). Resultatene fra

oppgavetilpasningsstatistikken viste at modellen 2PLM hadde flere oppgaver som var dårlig tilpasset modellen enn 1PLM (Tabell 4.5), og indikerte dermed at det beste valget var 1PLM. Embretson & Reise (2000) anbefaler et utvalg på minst 350 for bruk av 2PLM. Vi har i analysene et utvalg på 352, og antallet respondenter ligger derfor helt på grensen til hva som er akseptabelt for at parameterne i modellen skal estimeres nøyaktig. Basert på både resultatene fra analysene og med tanke på utvalget valgte vi å gå videre med 1PLM.

Oppgavene kan ha hatt dårlig tilpasning til modellen 1PLM av ulike grunner. En av grunnene kan ha vært at oppgaver har hatt avvikende diskrimineringsparametre. 1PLM setter diskrimineringsparameteren lik for alle oppgavene (Sick, 2010). Dette kan vi se på ved å undersøke diskrimineringsparameteren en gitt oppgave har fått i 2PLM. Som et eksempel har oppgave S105 en ganske høy diskrimineringsparameter (Tabell 4.6), og dette kan også sees ut fra kurvene i figur 5.1, der man ser at oppgavens karakteristiske kurve skiller seg fra de syv andre kurvene den blir sammenlignet med.



Figur 5.1: *Oppgavekarakteristisk kurve for S105 sammenlignet med syv andre oppgaver*

Fra tabell 4.6 finner vi at det samme sannsynligvis gjelder for oppgave S8, S13, S105, S179 og S248, der diskrimineringsparameteren endret seg betydelig fra 1PLM til 2PLM. Ved

oppgave S13, S179 og S248 gav dette også et utslag på vanskelighetsgraden til oppgavene, der oppgave S13 eksempelvis gikk fra vanskelighetsgrad -1,771 i modellen 1PLM til -2,582 i modellen 2PLM. Disse oppgavene egnet seg ikke ved bruk av modellen 1PLM.

En annen grunn til at oppgavene ikke passer til modellen kan være at gjetting spiller inn. Modellen 1PLM tar ikke hensyn til gjetting som en parameter og dersom gjetting er noe som spiller inn på en gitt oppgave kan dette muligens gjøre at oppgaven ikke passer inn i modellen. Som eksempel kan man se på oppgave S259 som ble identifisert som en oppgave som ikke passet til modellen 1PLM, og som heller ikke varierte i stor grad i diskrimineringsparameteren mellom modellene 1PLM og 2PLM. Svarprosenten på alle de fire svaralternativene i oppgave S259 lå mellom 16 og 31 prosent. Dette er en forholdsvis jevn fordeling av svar på de fire alternativene, og dette kan indikere at en del respondenter gjettet på svaret:

Boks 5.7: *Oppgave S259 som ble fjernet på grunn av dårlig tilpasning til 1PLM*

S259: Hvilket av følgende utsagn om pollinering er GALT?

A. Mange blomstrende arter har koevolvert med spesifikke pollinatorer. (*16*)
B. Vind-pollinerte blomster er ofte små og grønne, og produserer vanligvis hverken nektar eller lukt. (3*1*)
C. I pollineringsprosessen overføres pollenkorn fra pollenbærer til fruktknute. (*24*) *
D. Insekter pollinerer omtrent 65 % av alle blomstrende planter. (*29*)

Det er ikke gitt at man tar vekk oppgaver som er dårlig tilpasset modellen, uten videre ettertanke. Dersom oppgavene er nødvendig for å dekke testens bredde i innhold og nivå, vil det være nødvendig å vurdere om oppgavene skal beholdes i testen og eventuelt redigeres dersom mangelen ved oppgaven kan påpekes. Dersom oppgaver som blir tatt vekk er nødvendige for testens helhet vil validiteten av testen svekkes (Bohlig *et al.*, 1998). Oppgavene som hadde dårlig tilpasning var fordelt mellom temaene *cellen*, *genetikk*, *evolusjonære mekanismer*, *planters oppbygning og funksjon*, *dyrs oppbygning og funksjon* og *økologi*. Temaene *cellen*, *genetikk, planters oppbygning og funksjon* og *dyrs oppbygning og funksjon* har alle 30 eller flere oppgaver etter at oppgaver ble fjernet i analysen, og vi mener derfor at disse temaene er innholdsmessig godt dekket. Tre oppgaver (S13, S174, S207) ble tatt vekk fra temaet evolusjonære mekanismer, og dette temaet står dermed igjen med 23 oppgaver. Oppgave S13 er allerede identifisert med en endring i diskrimineringsparameteren

dersom man beveger seg fra modellen 1PLM til 2PLM, og det er derfor tydelig at det er riktig å fjerne denne oppgaven fra testen.

Når det gjelder oppgave S207 så har 42 % svart et av de gale svaralternativene og dette kan tyde på at mange svarer konsekvent feil svar på denne oppgaven:

Boks 5.8: *Oppgave S259 som ble fjernet på grunn av dårlig tilpasning til 1PLM*

S207: En populasjon med villblomster har ett gen med to alleler, A1 og A2. Tester viser at 70% av pollen som blir produsert i populasjonen inneholder A1-allelet. Dersom populasjonen er i Hardy-Weinberg-likevekt, hvor stor andel blomster er heterozygoter?

A. 0,21 (*10*)
B. 0,42 (*26*) *
C. 0,49 (*42*)
D. 0,70 (*22*)

Det er grunn til å tro at oppgave S207 er formulert på en måte som kan være misvisende selv for dyktige respondenter. Kanskje det kan ha noe med å gjøre at oppgaven krever kjennskap til én spesifikk formel og evne til regning. Vi vil derfor argumentere for at det er riktig å fjerne oppgaven.

Oppgave S174 fra evolusjonære mekanismer og S178 fra økologi viste ikke noen tydelig feil utfra svarprosentene, men begge oppgavene hadde ganske høy svarprosent på korrekt svar, og kunne dermed betegnes for lette. Det kan likevel begrunnes for at disse oppgavene burde vært beholdt i oppgavesettet ettersom temaene evolusjonære mekanismer og økologi ikke har like stor innholdsdekning som de andre temaene. Oppgave S177 går inn på begreper innenfor evolusjon og ved nærmere ettersyn ser vi at dette dekkes godt av andre spørsmål. Det samme gjelder for oppgave S178 som går inn på egenskaper ved ulike biomer, og dette er også et tema og en type oppgave som dekkes godt av andre oppgaver i testen. Vi konkluderte derfor med at innholdsvaliditeten i testen ikke blir svekket ved å fjerne disse ti oppgavene.

5.2.5 Lokal uavhengighet

Lokal uavhengighet mellom oppgaver betyr at oppgavene ikke er relatert til hverandre (Baghaei, 2008). Oppgavene er lokalt uavhengige når den statistiske avhengigheten mellom oppgaver bare er en funksjon av parmeterene i modellen, som i Rasch-modellen er dyktighet (Wilson, 1998). Når vi undersøkte om noen av oppgavene var lokalt avhengige, fant vi tre par som var lokalt avhengige. Oppgavene hadde en negativ verdi på under -0,2 ved bruk av Yen´s

Q_3-statistikk (Tabell 4.7). Negativ assosiasjon mellom oppgaver kan være et tegn på at flerdimensjonale data er lagt inn i en endimensjonal IRT modell (Habing & Roussos, 2003). Ettersom det å sjekke dataene opp mot en flerdimensjonal modell ligger utenfor oppgavens omfang, valgte vi her å fjerne en av oppgavene i hvert par, etter en vurdering av hvilken av oppgavene i parene som var minst egnet. Negativ assosiasjon mellom to oppgaver kan også komme av at de to oppgavene diskriminerer mest i motsatte deler av den latente skalaen (Zhang, 2007). Innenfor hvert par som ble identifisert var vanskelighetsgraden på en av oppgavene under gjennomsnittet og den andre over gjennomsnittet. Vi ser dette mest tydelig på oppgavene S143 med vanskelighetsgrad -0,519 og oppgave S290 med vanskelighetsgrad 0,394. Det er vanskelig å konkludere noe videre angående årsaken til de negative Q_3 verdiene uten å gjøre ytterlige analyser av parene som er identifisert som lokalt avhengige. Dersom oppgavene hadde vært nødvendige å ha med videre, hvis de for eksempel dekket et spesielt tema som ikke ble dekket av andre oppgaver, kunne vi valgt å beholde oppgavene, men sørge for at disse ikke ble brukt i det samme delsettet med oppgaver senere. Ettersom det ikke er vi selv som skal bruke oppgavene videre, vurderte vi det som mest oversiktlig å fjerne oppgavene slik at vi kunne ende opp med en bank med oppgaver som kan deles opp uten begrensninger, og uten lokalt avhengige oppgaver.

5.3 Reliabilitet

En testinformasjonsfunksjon brukes til å se på intervallet av dyktighet hvor testen måler best, og det er viktig å merke seg at der hvor standardfeilen er høy vil ikke dyktighetsnivået bli estimert presist, men den sanne scoren på testen vil bli estimert presist (Doran, 2005). Vi valgte å følge kriteriet til Hambleton & Lam (2009) der testinformasjonen på 10 er valgt som en kritisk verdi. Testen er derfor mest reliabel for respondenter med dyktighet på mellom -2,5 og 3,0 (Figur 4.1). Dette må taes hensyn til ved senere bruk av testen hvis man forventer at repondenter vil ha svært lav eller høy dyktighet. Man bør da bruke delsett av testen som er tilpasset nivået av dyktighet til respondentene. Doran (2005) presiserer også at man ikke måler reliabilitet som stabilitet, ettersom man ikke måler replikerbarhet i testen, og at man gjennom bruk av standardfeil ikke måler variasjonen rundt en sann score, men hvor presis testen er rundt en viss dyktighet. Ved videre bruk av testen kan reliabilitet som stabilitet måles ved å se på andre grupper av respondenter med likt antall studiepoeng eller gjennomføre testen for den samme gruppen respondenter etter et visst tidsrom.

Indre konsistens ble målt ved bruk av Cronbach`s alpha. Etter fjerning av oppgaver fra testen endte vi opp med en koeffisient på 0,932 som indikerer svært høy reliabilitet. Det er dermed god indre konsistens mellom ulike testledd. Før oppgaver ble fjernet fra testen (Tabell n undersøkte vi også Cronbach`s alpha for de ulike settene med oppgaver. Hvis vi for eksempel ser på sett 2, hadde dette settet en koeffisient på 0,66, som ifølge Cohen *et al.* (2011) indikerer minimal reliabilitet. I sett 2 ble 12 av oppgavene identifisert som uegnet grunnet lav point-biserial korrelasjon. Ettersom point-biserial korrelasjonen kan brukes som et mål på hvor godt en oppgave diskriminerer mellom respondenter med lav og høy dyktighet (Kavitha *et al.*, 2012), kan dette forklare den lave reliabilitetskoeffisienten på dette settet. Dette viser tydelig hvorfor det er viktig å undersøke oppgavene i en test for å sikre at oppgavene har en indre konsistens.

5.4 Validitet

Innholdsvalditet

En flervalgstests innholdsvalididitet kan knyttes til grad av innholdsdekning; Hvor godt innholdet i testen representerer fagområdet som testen er ment å dekke (Cohen *et al.*, 2011). Når vi sammenlignet de tre lærebøkene *Biology - How Life Works*, *Campbell Biology* og *Life - The Science of Biology* ble konklusjonen at den faglige overlappingen mellom bøkene var ganske stor. Dette indikerer at Campbell Biology var en egnet bok å ta utgangspunkt i ut fra et generelt biologisk kunnskapssyn. I tillegg så er de ulike temaene i Campbell Biology (Tabell 3.1) representert nokså likt prosentvis i den endelige oppgavesamlingen (Tabell 4.10) sammenlignet med boken. Unntaket er tema 1 - livets kjemi. Dette temaet handler om biokjemiske prosesser, og en god del kjemisk bakgrunnsstoff er også presentert her. Fordi vår flervalgstest skulle teste biologisk kunnskap og ikke kjemisk kunnskap var det naturlig å kutte ut enkelte av flervalgsoppgavene som kun omhandlet det kjemiske bakgrunnsstoffet. I ettertid innså vi at vi kunne dekket molekylærbiologiske temaer bedre, spesielt det som dreier seg om biologiske makromolekyler som karbohydrater, fett og proteiner og deres biologiske funksjoner. Utenom dette vurderte vi overensstemmelsen mellom temaene i flervalgstesten og de obligatoriske emnene i biologi ved bachelorutdanningen i biologi ved UiB som god. Mye av dette indikerer at flervalgstesten har en god innholdsdekning. Det at Campbell Biology benyttes som lærebok ved UiB støtter også dette.

En negativ trend med flervalgsoppgavene er oppgavenes fordeling i kategoriene *kunnskap*, *anvendelse* og *vurdering* (Tabell 4.10). Vi oppdaget fort at det uten tvil var lettest å lage kunnskapsoppgaver, og at anvendelses- og vurderingsoppgaver var noe vanskeligere å lage. Ettersom biologi er et detaljorientert fag så kan det argumenteres for at det naturlig vil være et stort behov for oppgaver som tester faktakunnskaper. Dette stemmer nok til en viss grad, men vi vil også argumentere for at det er fullt mulig -og ønskelig å lage oppgaver som tester anvendelses- eller vurderingsevner. En skjev fordeling mellom kategoriene kan være fordelaktig for enkelte respondenter. En jevn fordeling er derfor mer rettferdig. Ideelt sett bør flervalgstester derfor ha omtrent like mange oppgaver innenfor hver kategori (Haladyna, 2004). Vår flervalgstest dekker ikke alle tre kategoriene like godt, og dette svekker innholdsvalditeten.

Et annet aspekt ved innholdsvaliditet er at enhver oppgave bør reflektere den underliggende kompetansen som oppgavene er ment å måle (Cohen *et al.*, 2011). Dette er det samme som å si at oppgaver bør være endimensjonale, og dersom analyser viser at oppgaver måler noe annet enn det de er ment å teste vil dette svekke innholdsvaliditeten (Messick, 1989). Siden modellen 1PLM ble brukt til å analysere dataene og siden en av antagelsene for å bruke denne modellen er at dataene er tilstrekkelige endimensjonale indikerer dette at oppgavene målte det de var ment å måle. Unntaket er selvfølgelig oppgavene som ikke var godt tilpasset modellen, men disse oppgavene ble fjernet, så flervalgstesten regnes som tilstrekkelig endimensjonal. Dette styrker flervalgstestens innholdsvaliditet.

Et annet negativt aspekt ved innholdsvaliditeten til flervalgsoppgavene er at bare oppgavene i tema 4 - evolusjonære mekanismer gjennomgikk ekspertvurdering. Oppgavene i tema 4 ble gjennomgått av en professor ved UiB som vi har hatt kontakt med tidligere i studiet, og fordi han har kjennskap til flervalgsoppgaver og flervalgstester. Ideelt sett burde vi ha sendt alle oppgavene til vurdering for innhold til ulike eksperter (Haladyna, 2004), men dette ble ikke gjort. Vi innså ikke viktigheten av ekspertvurdering tidsnok, og tidspress førte til at vi måtte nøye oss uten. Dette er en svakhet ved flervalgstesten. Dersom vi skulle laget flervalgstesten på nytt hadde vi så langt det hadde latt seg gjøre sørget for at alle oppgavene gjennomgikk ekspertvurdering.

Kriterievalditet

En flervalgstests kriterievaliditet kan knyttes til sammenlignbarhet med andre, fremtidige eller tidligere tester (Cohen *et al.*, 2011; DeVillis, 1991; Messick, 1989). Siden formålet med denne masteroppgaven var å utvikle et produkt vil det ikke være hensiktsmessig om å snakke om sammenligning av produktet med tidligere resultater. I denne sammenhengen kan kriterievaliditet derfor bare brukes om potensiell sammenlignbarehet; at fremtidige tester kan sammenligne sine resultater med denne studien.

Konstruktvaliditet

En flervalgstests konstruktvaliditet innebærer at testen er forankret i tilstrekkelig relevant litteratur. I tillegg innebærer konstruktvaliditet at masteroppgavens design korrelerer positivt med tilsvarende designer og at potensielle moteksempler som kan falsifisere designet er presentert (DeVellis, 1991; Grimm & Widaman, 2012). Man er først i stand til å vurdere konstruktvaliditet når bekreftende og avkreftende argumenter for et design er vurdert og balansert (Cohen *et al.*, 2011). Dette innebærer å vurdere positive og negative sider ved metodikken brukt i denne masteroppgaven. Ettersom de fleste aspekter ved designet allerede har blitt diskutert vil negative og positive sider ved designet bli gjennomgått kort.

Følgende aspekter ved designet av denne flervalgstesten ble vurdert som en trussel for oppgavens konstruktvaliditet: Bare et fåtall oppgaver gjennomgikk ekspertvurdering; Bruk av SurveyXact gjorde det umulig å gi umiddelbar tilbakemelding om rett svaralternativ; Oppgavene som innledet flervalgstestene kan ha vært for vanskelige; Partial-matrix-modellen (Tabell 2.2) førte til en mindre sammenlignbarhet mellom de ulike settene i forhold til hva genuine-matrix-modellen (Tabell 2.3) ville ha gjort; Imputeringsteknikken ble ikke vurdert på riktig punkt i analyseprosessen; Designet er ikke sammenlignbart med andre designer siden formålet er å utvikle et produkt.

Følgende aspekter ved designet av denne masteroppgaven ble vurdert som en støtte for oppgavens konstruktvaliditet: Oppgaven er godt empirisk forankret; Alle oppgavene ble kvalitetssikret på flere måter (foruten gjennom ekspertvurdering); Partial-matrix-modellen (Tabell 2.2) førte til en økt sammenlignbarhet mellom settene enn item-sampling-modellen ville gjort (Tabell 1.1); Pretestene ga verdifull informasjon om flervalgstestenes og flervalgsoppgavenes gjennomførbarhet; Bruk av gevinst som motivasjonsmiddel kan ha økt

motivasjonen for å gjennomføre flervalgstestene; Distribueringen og markedsføringen av flervalgstestene gjennom Mi side, Webmail, Facebook, TV-skjermer, forelesningsbesøk og utsendelse til andre institusjoner med påminnelser økte respondentantallet; Testing av hvilken modell gjorde at 1PLM-modellen ble valgt fordi det var den beste løsningen, ikke fordi det var det eneste alternativet; Fjerning av utilpassede oppgaver gjorde flervalgstesten mer reliabel.

5.5 Avsluttende vurdering av verktøyet

Som en avsluttende vurdering av selve verktøyet som har blitt utviklet gjennom denne masteroppgaven, vil vi trekke frem svake og sterke sider ved banken med flervalgsoppgaver. En svakhet med banken med flervalgsoppgavene er at oppgavene som ble "godkjent" gjennom analysene ikke ble redigert i etterkant av svarene kom inn fra respondentene. Noen svaralternativer kunne vært forbedret, da det er noen svaralternativ som ingen respondenter har valgt. Det er også en ujevn fordeling av oppgaver klassifisert som kunnskap, anvendelse og vurdering. Her ligger en hovedvekt på kategorien kunnskap, men dette var noe forventet ettersom ved bruk av flervalgsoppgaver vil det være krevende å konstruere gode oppgaver som tester anvendelse og vurdering. En usikkerhet ved de estimerte vanskelighetsgradene til oppgavene er at disse er estimert utfra datasettet med imputerte data. Vi har likevel valgt å ta disse med i det endelige verkøyet slik at de som skal bruke verktøyet kan sette sammen sett med oppgaver der det er en god spredning av vanskelighetsgrad. Da vil ikke usikkerheten i den nøyaktige vanskelighetsgraden ha like stor betydning.

Utenom området livets kjemi, har banken med flervalgsoppgaver god innholdsdekning innenfor grunnleggende kompetanse i biologi, og antallet oppgaver er stort nok til å kunne lage flere delsett med oppgaver som måler biologikompetanse. En god spredning av oppgaver langs dyktighetsskalaen gir mulighet for å tilpasse tester til ulike temaer og dyktighetsnivåer. Gjennom kvalitetssikring både i forkant av distribueringen og som resultat av analysene vil oppgavene som ligger i banken med flervalgsoppgaver være godt egnet til å teste biologikompetanse innenfor områdene cellen, genetikk, evolusjonære mekanismer, den biologiske diversitetens evolusjonære historie, planters oppbygning og funksjon, dyrs oppbygning og funksjon og økologi. Det vil også være en fordel for de som skal bruke verktøyet videre at det nå finnes statistikk som beskriver svarprosenter på de ulike

svaralternativene, fordeling av andel riktige svar for ulike grupper av respondenter (gruppert etter antall studiepoeng) og vanskelighetsgrad til oppgavene.

Gjennom denne oppgaven har vi forsøkt å argumentere for at vi mener verktøyet er egnet til å måle biologikompetanse på bachelornivå. Vi mener å ha argumentert for at verktøyet er valid og reliabel, selv om testen har sine mangler.

5.6 Studentenes dyktighet i forhold til avlagte studiepoeng

Når vi ser på fordelingen av gjennomsnittet av andel riktige svar i forhold til antall studiepoeng ser vi at andel riktige svar stiger i takt med antall studiepoeng, unntatt i den siste gruppen der respondentene har over 240 studiepoeng. Stigningen er i tråd med det man kan forvente siden kunnskapen bør økes etterhvert som man kommer lenger i studiet. Det er også her et spørsmål om hvor reell forskjellen mellom gruppene er, ettersom konfidensintervallene er ganske store. Lengre konfidensintervaller signaliserer større usikkerhet enn kortere intervaller (Store Norske Leksikon). Det korteste konfidensintervallet er i gruppen med mer enn 240 studiepoeng, på tross av at det er denne gruppen som har færrest respondenter. Ved få respondenter vil man ofte få et lengre konfidensintervall. Ut fra dette kan resultatene indikere at gruppen med respondenter med mer enn 240 studiepoeng er mer homogen enn de andre gruppene.

Selv om gjennomsnittet av andel riktige svar viser en forskjell mellom gruppene av respondentene, viser Tukey HSD testen at det bare er gruppen mindre enn 60 studiepoeng som er signifikant forskjellig fra de andre gruppene. Derfor kan vi bare si sikkert at det er gruppen med mindre enn 60 studiepoeng som har lavere andel av korrekte svar. Denne gruppen av respondenter går altså på første året av bachelor studiet i biologi, og det er derfor forventet at disse har lavere grad av kompetanse i biologi, enn respondenter som har kommet lengre i studiet. Gruppen med respondenter som har mindre enn 60 studiepoeng har det største utvalget respondenter, og det ville vært interessant å se med et større utvalg på de andre gruppene om det da ville vært en signifikant stigning i andel riktige svar, og om nedgangen i gruppen med respondenter med mer enn 240 studiepoeng faktisk gjør det dårligere enn de andre gruppene. Dersom nedgangen i andel korrekte svar hadde vist seg å være signifikant ved et større utvalg ville dette vært noe overraskende ettersom disse respondentene har

kommet lengst i studiet. En mulig forklaring kan være at studentene på masterdelen av studiet fordyper seg mer innenfor ett emne, og ettersom det er gått en stund siden disse studentene har hatt om de spesifikke emnene det spørres om, kan det være en mulighet for at kunnskapen glemmes. Dette er selvsagt bare spekuleringer fra vår side, siden det ligger utenfor denne oppgaven å undersøke årsaker til nedgangen i andel riktige svar. Men dette er et spennende område som er aktuelt for videre forskning.

Kapittel 6 – Anbefalinger for bruk av verktøyet?

Vi har utviklet et verktøy som skal teste biologi kompetanse blant studenter. Verktøyet består etter analysene av en bank med 239 flervalgsoppgaver innenfor temaene livets kjemi, cellen, genetikk, evolusjonære mekanismer, den biologiske diversitetens evolusjonære historie, planters oppbygning og funksjon, dyrs oppbygning og funksjon og økologi. Temaet livets kjemi er ikke dekket noe godt. Dersom man skal undersøke fagområdet økologi for seg selv vil vi anbefale å legge til spørsmål ettersom det er litt få spørsmål til å dekke området. Området evolusjonære mekanismer bør sees i sammenheng med den biologiske diversitetens evolusjonære historie dersom området evolusjon skal undersøkes. Ellers er resten av temaene dekket godt med oppgaver.

Vi har tatt utgangspunkt i analysene når vi har fjernet oppgaver, men det vil være lurt vurdere de gjenværende oppgavene før videre bruk. Her tenker vi spesielt på at noen av distraktorene som ikke har vært benyttet. Ved slike tilfeller bør man vurdere om disse distraktorene bør endres. Det er også en mulighet for å slette disse distraktorene dersom man ikke er velger å ha like mange svaralternativ på hver oppgave. Men det er også en mulighet at distraktorer ikke er valgt på grunn av få respondenter, og at hvis utvalget var større, ville noen respondenter valgt disse svaralternativene.

Antall oppgaver hver student bør få vil være avhengig av hvor stort område testen skal dekke. Dersom det er hele området med alle fagområdene, vil vi anbefale 40-50 spørsmål for å få dekket alle områdene tilstrekkelig. Er det ulike fagområdet som skal testes, kan en minske antall spørsmål i settet. Antall oppgaver vil også ha sammenheng med om hvordan testen skal gjennomføres. Dersom flervalgsoppgavene senere skal brukes som en test av kunnskap underveis i et emne og at dette skal være frivillig, vil vi anbefale at testen legges inn som en del av en forelesning der det settes av tid til gjennomføring. Dette vil sannsynligvis øke svarprosenten. Hvis testen er frivillig og skal gjennomføres av studentene hjemme på fritiden anbefaler vi å korte ned på antall oppgaver.

Vanskelighetsgraden til oppgavene kan også være et hjelpemiddel når man skal sette sammen sett med oppgaver, og det er anbefalt å ha en spredning langs intervallet av vanskelighetsgrad med hovedvekt av oppgaver rundt vanskelighetsgrad 0 (Varma, 2006). Dersom man har en bakgrunnsinformasjon om dyktigheten til studentene, kan man tilpasse settet med oppgaver

slik at spørsmålene er mer tilpasset nivået. Testen er i utgangspunktet mest reliabel for bruk hos respondenter med dyktighet litt over 0. Ved homogene grupper der man forventer at dyktigheten er under 0, kan man velge ut oppgaver som har en lavere vanskelighetsgrad. Ved bruk av hele oppgavebanken må man være klar over at det er større usikkerhet knyttet til vanskelighetsgraden til oppgaver under - 2,5 og over 3.

Kapittel 7 – Litteraturliste:

Aningbo, L. C. (2011) 'Demonstrating of Multiple Matrices Sampling Technique In Establishing The Psychometric Characteristics Of Large Samples', *Journal of Education and Practice*, vol. 2, nr. 3, s. 19-25.

Baker, F. B. (2001) *The Basics of Item Response Theory* (2. utgave). ERIC Clearinghouse on Assessment and Evaluation, College Park, MD.

Baghaei, P. (2008) 'Local Dependency and Rasch Measures', *Rasch Measurement Transactions*, vol. 21, nr. 3, s. 1105-1106.

bioCEED: <http://bioceed.no/> [04.05.2015].

Bloom, B. (1956) *Taxonomy of Educational Objectives: the classification of educational goals* (6. utgave). New York: David McKay Company, inc.

Bohlig M., Fisher W.P. Jr., Masters, G. N. & Bond, T. (1998) 'Content Validity and Misfitting Items', *Rasch Measurement Transactions*, vol. 12, nr. 1, s. 607.

Bortolotti, S. L. V.; Tezza, R.; F. de Andrade, D.; Bornia, A. C. & F. de Sousa Júnior (2012) 'Relevance and advantages of using item response theory', *Quality and quantity*, vol. 47, nr. 4, s. 2341-2360.

Brame, C. J. (2015) *Writing Good Multiple Choice Test Questions*. The Center for Teaching, Vanderbilt University, Nashville. Tilgjengelig fra: <http://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/> [16.01.2015].

Brennan, R. (1998) 'Misconceptions at the Intersection of Measurement Theory and Practice', *Educational measurement*, vol. 17, nr. 1, s. 5-9.

Budescu, D. V., Cohen, Y. & Bensimon, A. (1997 'A revised modified parallel analysis for the construction of unidimensional item pools', *Applied psychological measurement*, vol. 21, nr. 3, s. 233-252.

Campbell, N.; Cain, M. L; Jackson, R. B.; Minorsky, P.V.; Reece, J. B.; Urry, L. A. & Wasserman, S. A. (2011) *Campbell Biology*. Pearson Education, U.S.A.

Cavanaugh, J. E. (2012) *Model Selection Lecture V: The Bayesian Information Criterion*, s. 171-290. Department of Biostatistics, Department of Statistics and Actuarial Science, The University of Iowa.

Chalmers, R. P. (2012) 'mirt: A multidimensional Item Response Theory Package For the R Environment', *Journal of Statistical Software*, vol. 48, nr. 6, s. 1-29.

Chen, W. H. & Thissen, D. (1997) 'Local Dependence Indexes for Item Pairs Using Item Response Theory', *Journal of Educational and Behavioral Statistics*, vol. 22, nr. 3, s. 265-289.

Childs, R. A. & Jaciw, A. P (2003) ‘Matrix sampling of items in large-scale assessments’, *Practical Assessment, Research & Evaluation*, vol. 8, nr. 16, s. 1-11.

Cizek, G. J., Robinson, K. L. & O'Day, D. M (1988) ‘Nonfunctioning options: a closer look’, *Educational and psychological measurement*, vol. 58, nr. 4, s. 605-611.

Cohen, L., Manion, L. & Morrison, K. (2011) *Research Methods in Education* (7. utgave). London & New York: Routledge.

De Ayala, R. J. (2009) *The Theory and Practice of Item Response Theory*. The Guilford Press. A Division of Guilford Publications, Inc. New York.

DeVellis, R. F. (1991) *Scale Development* (utgave). Applied Social Research Methods Series, Vol. 26. California, U.S.A., SAGE Publications.

Dings, J., Childs, R., & Kingston, N. (2002) ‘The effects of matrix sampling on student score comparability in constructed-response and multiple-choice assessments’, *Washington, DC: Council of Chief State School Officers*.

Doran, H. C. (2005) ‘The information function for the one-parameter logistic model: is it reliability?’, *Educational and Psychological Measurement*, vol. 65, nr. 5, s. 665-675.

Drasgow, F. & Lissak, R. (1983) ‘Modified parallel analysis: a procedure for examining the latent dimensionality of dichotomously scored item responses’, *Journal of Applied Psychology*, vol. 68, nr. 3, s. 363-373.

Ebel, R. L. (1982) ‘Proposed solutions to two problems of test construction’, *Journal of Educational Measurement*, vol. 19, nr. 4, s. 267-278.

Edelen, M. O. & Reeve, B. B. (2007) ‘Applying item response theory (IRT) modeling to questionnaire development, evaluation and refinement’, *Quality of Life Research*, vol. 16, nr. 1, s. 5-18.

Embretson, S. E. & Reise S. P. (2000) *Item response theory for psychologists* (utgave). Lawrence Erlbaum Associates, Mahwah, New Jersey.

Educational Testing Service: <https://www.ets.org/s/gre/pdf/practice_book_biology.pdf> [20.01.2015].

Falk, C. F. & Savalei, V. (2011) ‘The Relationship Between Unstandardized and Standardized Alpha, True Reliability, and the Underlying Measurement Model’, *Journal of personality assessment*, vol. 93, nr. 5, s. 445-453.

Frary, R. B. (1995) ‘More multiple-choice item writing do's and don'ts’. *Practical Assessment, Research & Evaluation*, vol. 4, nr. 11.

Garland Science, Taylor & Francis Group: <http://garlandscience.com/garlandscience_resources/resource_detail.jsf?landing=student&resource_id=9780815341291_CH05_QZ001> [21.01.2015].

Geisinger, K. F., Bracken, B. A., Carlson, J. F., Hansen, J. C., Kuncel, N. R., Reise, S. P. & Rodriguez, M. C. (2013) 'Reliability', *APA handbook of testing and assessment in psychology*, Vol. 1, s. 21-42.

Grimm, K. J. & Widaman, K. F. (2012) 'Construct Validity'. *APA handbook of research methods in psychology*, vol. 1, s. 621-642.

Haladyna, T. M. & Downing, S. M. (1993) 'How many options is enough for a multiple choice test item?', *Educational and Psychological Measurement*, vol. 53, nr. 4, s. 999-1010.

Haladyna, T. M. (2004) *Developing and Validating Multiple-Choice Test Items* (3. utgave). Lawrence Erlbaum Associates. U.S.A.

Haladyna, T. M., Downing, S. M. & Rodriguez, M. C. (2002) 'A review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment'. *Applied Measurements in Education*, vol. 15, nr. 3, s. 309-334.

Hambleton, R. K., & Lam, W. (2009) *Redesign of MCAS Tests Based on a Consideration of Information Functions* (revidert utgave). MCAS Validity Report No. 18; CEA-689. Amherst, M.A., University of Massachusetts, Center for Educational Assessment.

Hambleton, R. K., Swaminathan, H. & Rogers H.J. (1991) *Fundamentals of item response theory*. SAGE Publications Inc., New Bury, California.

Harvey, R. J. & Hammer, A. L. (1999) 'Item response theory', *Counseling Psychologist*, vol. 27, nr. 3, s. 353-383.

HCC Southeast Commons, College Educational Technology Services: http://m.se.hccs.edu/Users/tanvir.khatlani/MyDocuments/Study_Guide_for_Lecture_Exam_2_Biol_1407_03.12.2010.pdf [16.01.2015].

Hopfenbeck, T. N. & Lillejord S. (2013) 'Vurdering etter Kunnskapsløftet.' I Krumsvik, R. J. & Säljö R. (red.) *Praktisk-pedagogisk utdanning. En antologi* (utgave). Fagbokforlaget, Bergen.

Horgen, S. A. (2007) 'Pedagogical use of multiple choice tests, Students create their own tests', *Proceeding of the Informations Education Europe II Conferance, South East European Research Center (SEERC)*, 386-392.

Huisman, M. (2000) 'Imputation of Missing Item Responses: Some Simple Techniques', *Quality & Quantity*, vol. 34, nr. 4, s. 331-351.

Hula, W. D., Fergadiotis, D. & Martin, N. (2012) 'Model Choice and Sample Size in Item Response Theory Analysis of Aphasia Tests', *American Journal of Speech-Language Pathology,* vol. 21, nr. 2, s. 38-50.

Imsen, G. (2009) *Lærerens verden. Innføring i generell didaktikk* (4. utgave). Universitetsforlaget, Oslo.

IndiaBIX: <http://www.indiabix.com/general-knowledge/biology/018002> [26.01.2015].

Institutt for biologi: <http://www.uib.no/studieprogram/BAMN-BIO#uib-tabs-laringsutbyte> [22.04.15].

Kavitha, R., Vijaya, A. & Saraswathi, D. (2012) 'Intelligent item assigning for classified learners in ITS using Item Response Theory and Point Biserial Correlation', *Computer Communication and Informatics (ICCCI),* 2012 International Conference on. IEEE, 2012.

Kehoe, J. (1995) 'Writing multiple-choice test items. Practical Assessment', *Research & Evaluation*, vol. 4, no. 9.

Kline, T. J. B. (2005) *Psychological testing: A practical approach to design and evaluation*, SAGE Publications Inc., Thousand Oaks, CA.

Kromey, J. D. & Hines, C. V. (1994) 'Nonrandomly missing data in multiple regression: An empirical comparison of common missing-data treatments', *Educational and Psychological Measurement*, vol. 54, nr. 3, s. 573-593.

Linacre, J. M. (1994) 'Sample Size and Item Calibration Stability', *Rasch Measurement Transactions,* vol. 7, nr. 1, s. 238.

Little, R. J. A. & Rubin, D. B. (2006) *Incomplete Data*. Encyclopedia of Statistical Sciences. 5.

Loevinger, J. (1957) 'Objective Tests as Instruments of Psychological Theory: Monograph Supplement 9', *Psycholocigal Reports*, vol. 3, nr. 7, s. 635-694.

Lord, F. M. (1968) 'An Analysis of the Verbal Scholastic Aptitude Test Using Birnbaum's Three-parameter logistic model', *Educational And Psychological Measurement*, vol. 28, nr. 4, s. 989-1020.

Lord, F. M. (1977) 'Optimal number of choices per item - A comparison of four approaches', *Journal of Educational Measurement,* vol. 14, nr. 1, s. 33-38.

Louangrath, P. I. (2013) 'Validity Test'. *Center for Family Enterprise Research Center, Bangkok University*.

Lykknes, A. & Smidt, J. (2009) 'Skriving i arbeidsbok i naturfag på ungdomstrinnet: Innhold, form og formål'. I Groven, B., Guldal, T. M., Lillemyr, O. F., Naastad, N. & Rønning, F. (2009) *FoU i Praksis 2008. Rapport fra konferanse om praksisrettet FoU i lærerutdanning*. Trondheim, tapir akademisk forlag.

Marais, S. M. & Andrich, D. H. (2008) 'Formalising dimension and response violations of local independence in the unidimensional Rasch model', *Journal of Applied Measurement*, vol. 9, nr. 3, s. 200-215.

Marion P. v. & Strømme, A. (2008) *Biologididaktikk*. Høyskoleforlaget, Kristiansand.

Mc Graw Hill Education: <http://highered.mheducation.com/sites/0073031208/student_view0/index.html> [16.01.2015].

Morris, J., Hartl, D., Knoll, A. & Lue, R. (2013) *Biology - How Life Works* (1. utgave). U. S. A., W. H. Freeman and Company.

Nasjonalt kunnskapssenter for helsetjenesten: <http://www.psyktest.no/om-m%C3%A5leegenskaper> [16.04.2015].

Osterlind, S. J. (2002) *Constructing Test Items : Multiple-Choice, Constructed-Response, Performance and Other Formats*. Kluwer Academic Publishers. New York, Boston, Dordrecht, London, Moscow.

Ostini, R. & Nering, M. L. (2006) *Polytomous Item Response Theory Models*, SAGE Publications Inc., Thousand Oaks, CA.

Oxford University Press, online resource centres: <http://global.oup.com/uk/orc/pharmacy/ifp_therapeutics/student/mcqs/ch02/> [17.01.2015].

Pearson Higher Education, Biology of Humans: <http://wps.aw.com/bc_goodenough_boh_3/104/26712/6838489.cw/index.html> [26.01.2015].

Pearson Higher Education, Mastering Biology: <http://www.pearsonmylabandmastering.com/northamerica/masteringbiology/> [16.01.2015].

Popham, W. J. (1993) 'Circumventing the high cost of Authentic Assessment', *Phi Delta, Kappa*, vol. 74, nr. 6, s. 470-473.

Reise, S. (1990) 'A comparison of item- and person-fit methods of assessing model-data fit in IRT', *Applied Psychological Measurement*, vol. 14, nr. 2, s. 127-137.

Reise, S. & Revicki D. (2015) *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. Routledge, New York.

Reise, S. P., Ainsworth, A. T. & Haviland, M. G. (2005) 'Item response theory - Fundamentals, applications, and promise in psychological research', *Current Directions In Psychological Science*, vol. 14, nr. 2, s. 95-101.

Resaland, E. (2013) *Rasch-analyse av data fra et spørreskjema om hvordan helsesøstre oppfatter ulike brukergruppers funksjonelle og interaktive nutrition literacy.* Masteroppgave,

Høgskolen i Oslo og Akershus. Tilgjengelig fra: <https://oda.hio.no/jspui/bitstream/10642/1770/2/Resaland_Erik_MAME5910_masteroppgave.pdf> [15.04.2015??].

Rodriguez, M. C. (2005) 'Three Options Are Optimal For Multiple Choice Items: A Meta-Analysis of 80 Years of Research', *Journal of Educational Measurement,* vol. 24, nr. 2, s. 3-13.

Rubin, D. B. (1975) 'Interference and Missing Data', *ETS Research Bulletin Series*, vol. 1975, nr. 1, s. 1-19.

Sande, I. G. (1982) 'Imputation in Surveys: Coping with Reality', *The American Statistician,* vol. 36, nr. 3, s. 145-152.

Savada, D., Hillis, D., Heller, C. & Berenbaum, M. (2014) *Life - The Science of Biology* (10. utgave). U.S.A, Sinauer Associates.

Shoemaker, D. M. & Shoemaker, J. S. (1981) 'Applicability Of Multiple Matrix Sampling To Estimating Effectiveness Of Educational Programs', *Evaluation and Program Planning*, vol. 4, nr. 2, s. 151-161.

Sick, J. (2010) 'Rasch measurement in language education Part 5: Assumptions and requirements of Rasch measurement', *SHIKEN: JALT Testing & Evaluation SIG Newsletter*, vol.14, nr. 2, s. 23-29. International Christian University, Tokyo.

Singer, E. % Ye, C. (2013) 'The Use and Effects of Intencives in Surveys', *The ANNALS of the American Academy of Political and Social Science*, vol. 645, nr. 1, s. 112-141.

Sirnes, S. (2005) *Flervalgsoppgaver - konstruksjon og analyse*. Bergen, Fagbokforlaget.

Store Norske Leksikon: <https://snl.no/konfidensintervall>, [27.05.15]

Tanaka, J. S. (1987) '"How Big Is Enough?": Sample Size and Goodness of Fit in Structural Equation Models with Latent Variables', *Child development*, vol. 58, nr. 1, s. 134-146.

Tutor Vista: <http://www.tutorvista.com/content/science/science-ii/reproduction/multiple-choice.php> [23.01.2015].

Ulysses S. Grant High School: <http://www.google.no/url?sa=t&rct=j&q=&esrc=s&source=web&cd=10&sqi=2&ved=0CGMQFjAJ&url=http%3A%2F%2Fap-bio-a-per-1.granths.org%2Fmodules%2Flocker%2Ffiles%2Fget_group_file.phtml%253Fgid%253D3246698%2526fid%253D17424757&ei=sz3PVOyYI4uwUeichPAJ&usg=AFQjCNH7VVFkRbZULuxd9cDRnOU5xDR0UQ&sig2=HpSjlgEd04I3bElG5xZYjw&bvm=bv.85076809,d.d24> [30.01.2015].

Universitetet i Oslo: <http://www.uv.uio.no/ils/forskning/prosjekt-sider/pisa/frigitte-oppgaver/> [25.05.2015].

Utdanningsdirektoratet (2006) *Læreplan i biologi - programfag i studiespesialiserende utdanningsprogram*. Tilgjengelig fra: <http://www.udir.no/kl06/BIO1-01/Hele/Komplett_visning/> [22.04.15].

Utdanningsdirektoratet, (2014) *Hva er vurdering for læring*.Tilgjengelig fra: <http://www.udir.no/Vurdering-for-laring/Hva-er-Vurdering-for-laring/Hva-er-vurdering-for-laring/> [25.05.2015].

Utdanningsdirektoratet, (2015) *Eksamensveiledning - om vurdering av eksamensbesvarelser*. REA3002 Biologi 2. Sentralt gitt skriftlig eksamen. Tilgjengelig fra: <https://dok.udir.no/DokumenterAndrekataloger.aspx?proveType=Ev> [25.05.2015].

Varma, S. (2006) Preliminary item statistics using point biserial correlation and p value. Educational Data System, Inc. Tilgjengelig fra: <http://www.eddata.com/resources/publications/eds_point_biserial.pdf> [30.04.15].

Wagenmakers E. J. & Farell, S. (2004) 'AIC model selection using Akaike weights', *Psychonomic Bulletin & Review*, vol. 11, nr. 1, s. 192-196.

Wikipedia: http://no.wikipedia.org/wiki/Biologi [23.04.15].

Wilson, M. (1998) 'Detecting and Interpreting Local Item Dependence Using a Family of Rasch Models', *Applied Psychological Measurement*, vol. 12, nr. 4, s. 353-364. University of California, Berkeley.

Winter, G. (2000) 'A comparative discussion of the notion of 'validity' in qualitative and quantitative research', *The Qualitative Report*, vol. 4, nr. 3, s. 4.

Wolkowitz, A. A. & Skorupski, W. P. (2013) 'A Method for Imputing Response Options for Missing Data on Multiple-Choice Assessments', *Educational and Psychological Measurement*, vol. 73, nr. 6, s. 1036-1053.

Zhang, J. (2007) *Dichotomous or polytomous model? equating of testlet-based tests in light of conditional item pair correlations*. PhD (Doctor of Philosophy) thesis, University of Iowa.

Zimmaro, D. M. (2004) 'Writing good multiple-choice exams', *Measurement and Evaluation Center: University of Texas, Austin*.

Vedlegg 1: Oversikt over oppgavenes egenskaper I den endelige flervalgstesten

Oppgave	Tema	Vanskelighetsgrad	Kategori
S1	Evolusjon	-1,393	V
S2	Planters oppbygning og funksjon	2,043	A
S3	Evolusjon	-1,140	K
S4	Økologi	-3,620	K
S5	Dyrs oppbygning og funksjon	1,676	K
S6	Den biologiske diversitetens evolusjonære historie	-2,888	K
S7	Cellen	0,239	K
S9	Dyrs oppbygning og funksjon	-0,369	K
S10	Genetikk	-0,205	A
S11	Genetikk	1,102	V
S12	Evolusjon	-0,251	A
S14	Cellen	0,664	K
S15	Cellen	-1,713	K
S16	Cellen	-1,658	K
S17	Evolusjon	1,984	V
S19	Genetikk	1,621	K
S20	Cellen	0,616	V
S21	Genetikk	-0,439	K
S22	Genetikk	1,596	A
S23	Dyrs oppbygning og funksjon	2,316	A
S24	Evolusjon	0,929	V
S25	Genetikk	-2,481	K
S26	Dyrs oppbygning og funksjon	-1,498	A
S27	Dyrs oppbygning og funksjon	2,853	K
S29	Den biologiske diversitetens evolusjonære historie	0,520	K
S30	Økologi	0,379	K
S31	Genetikk	0,121	V
S32	Dyrs oppbygning og funksjon	0,498	V
S33	Den biologiske diversitetens evolusjonære historie	0,099	V
S34	Den biologiske diversitetens evolusjonære historie	1,078	K
S35	Planters oppbygning og funksjon	0,355	A
S36	Evolusjon	-0,111	V
S37	Genetikk	-0,251	K
S38	Den biologiske diversitetens evolusjonære historie	-1,420	K
S39	Den biologiske diversitetens evolusjonære historie	1,702	K
S40	Planters oppbygning og funksjon	-1,795	K
S41	Dyrs oppbygning og funksjon	0,239	A

S42	Planters oppbygning og funksjon	1,331	A
S43	Den biologiske diversitetens evolusjonære historie	-0,205	K
S44	Den biologiske diversitetens evolusjonære historie	1,843	K
S45	Cellen	-0,918	K
S47	Planters oppbygning og funksjon	2,316	K
S48	Den biologiske diversitetens evolusjonære historie	0,687	K
S50	Planters oppbygning og funksjon	2,852	V
S51	Dyrs oppbygning og funksjon	-1,631	A
S52	Evolusjon	-3,958	K
S54	Genetikk	-0,676	A
S55	Planters oppbygning og funksjon	-0,941	K
S56	Genetikk	-3,276	K
S57	Dyrs oppbygning og funksjon	0,075	V
S58	Cellen	-1,164	K
S59	Cellen	0,215	K
S61	Den biologiske diversitetens evolusjonære historie	2,577	K
S62	Genetikk	-2,964	A
S63	Genetikk	3,968	V
S65	Genetikk	1,435	K
S66	Dyrs oppbygning og funksjon	-0,439	K
S67	Dyrs oppbygning og funksjon	-0,439	K
S68	Den biologiske diversitetens evolusjonære historie	3,075	K
S69	Evolusjon	-1,472	A
S70	Genetikk	2,073	V
S72	Den biologiske diversitetens evolusjonære historie	0,028	V
S76	Økologi	-0,868	K
S77	Dyrs oppbygning og funksjon	-1,240	A
S79	Cellen	-0,087	K
S80	Planters oppbygning og funksjon	5,615	K
S81	Dyrs oppbygning og funksjon	2,926	K
S82	Økologi	-4,468	V
S86	Dyrs oppbygning og funksjon	2,889	K
S87	Dyrs oppbygning og funksjon	-1,879	K
S88	Den biologiske diversitetens evolusjonære historie	0,687	A
S89	Cellen	2,014	K
S91	Evolusjon	-3,812	K
S93	Cellen	-2,749	K

S94	Evolusjon	2,641	A
S95	Planters oppbygning og funksjon	1,541	A
S96	Dyrs oppbygning og funksjon	1,409	V
S97	Cellen	0,076	K
S98	Planters oppbygning og funksjon	2,379	K
S99	Cellen	1,229	V
S100	Cellen	0,831	A
S101	Planters oppbygning og funksjon	1,029	V
S102	Cellen	1,383	K
S103	Evolusjon	-2,199	K
S104	Genetikk	-2,291	K
S106	Genetikk	-3,359	V
S107	Genetikk	0,005	K
S108	Genetikk	-0,135	A
S109	Dyrs oppbygning og funksjon	-0,942	K
S110	Dyrs oppbygning og funksjon	-0,700	V
S111	Dyrs oppbygning og funksjon	0,712	K
S112	Evolusjon	-0,298	V
S113	Den biologiske diversitetens evolusjonære historie	-2,080	K
S114	Dyrs oppbygning og funksjon	2,014	K
S115	Den biologiske diversitetens evolusjonære historie	0,474	K
S116	Dyrs oppbygning og funksjon	-0,392	K
S117	Den biologiske diversitetens evolusjonære historie	1,178	K
S118	Evolusjon	-0,748	K
S119	Planters oppbygning og funksjon	1,648	K
S120	Dyrs oppbygning og funksjon	-0,181	A
S121	Den biologiske diversitetens evolusjonære historie	4,630	K
S112	Den biologiske diversitetens evolusjonære historie	0,737	K
S123	Den biologiske diversitetens evolusjonære historie	1,786	A
S126	Planters oppbygning og funksjon	1,759	K
S127	Genetikk	-2,110	A
S128	Økologi	1,541	K
S129	Dyrs oppbygning og funksjon	3,402	K
S130	Planters oppbygning og funksjon	-1,631	A
S131	Den biologiske diversitetens evolusjonære historie	2,043	K
S133	Genetikk	-0,676	K
S134	Genetikk	1,356	A

S125	Cellen	2,852	A
S136	Genetikk	0,520	A
S137	Cellen	2,379	V
S138	Den biologiske diversitetens evolusjonære historie	0,285	K
S139	Dyrs oppbygning og funksjon	3,362	V
S141	Dyrs oppbygning og funksjon	0,403	K
S142	Evolusjon	-3,234	A
S144	Genetikk	-3,808	K
S145	Økologi	-1,040	V
S146	Økologi	-1,525	K
S147	Den biologiske diversitetens evolusjonære historie	-2,890	K
S148	Cellen	-0,820	K
S149	Cellen	-0,510	K
S151	Den biologiske diversitetens evolusjonære historie	0,616	K
S152	Den biologiske diversitetens evolusjonære historie	0,121	K
S153	Genetikk	2,477	K
S154	Genetikk	2,817	V
S155	Genetikk	-0,275	K
S156	Dyrs oppbygning og funksjon	2,543	V
S157	Cellen	2,853	K
S158	Planters oppbygning og funksjon	0,592	V
S159	Økologi	-3,039	A
S160	Genetikk	-1,935	K
S161	Planters oppbygning og funksjon	-1,189	K
S162	Planters oppbygning og funksjon	0,521	K
S163	Dyrs oppbygning og funksjon	1,203	K
S164	Dyrs oppbygning og funksjon	-0,228	K
S165	Dyrs oppbygning og funksjon	-1,714	V
S167	Den biologiske diversitetens evolusjonære historie	0,356	K
S168	Planters oppbygning og funksjon	-0,868	V
S170	Den biologiske diversitetens evolusjonære historie	3,535	K
S171	Planters oppbygning og funksjon	1,203	K
S173	Dyrs oppbygning og funksjon	0,711	A
S176	Den biologiske diversitetens evolusjonære historie	1,871	V
S177	Cellen	1,153	A
S181	Dyrs oppbygning og funksjon	0,640	K
S182	Genetikk	-0,869	K

S184	Dyrs oppbygning og funksjon	-1,630	K
S187	Den biologiske diversitetens evolusjonære historie	-2,855	A
S188	Planters oppbygning og funksjon	0,262	K
S189	Dyrs oppbygning og funksjon	-0,018	K
S191	Dyrs oppbygning og funksjon	1,593	K
S192	Evolusjon	-1,446	V
S193	Genetikk	-0,844	A
S196	Den biologiske diversitetens evolusjonære historie	2,223	K
S197	Genetikk	1,003	V
S199	Cellen	0,808	K
S200	Økologi	2,072	V
S201	Dyrs oppbygning og funksjon	-0,748	A
S202	Den biologiske diversitetens evolusjonære historie	-1,420	K
S203	Den biologiske diversitetens evolusjonære historie	1,435	A
S205	Økologi	-0,252	A
S208	Livets kjemi	-2,261	K
S209	Planters oppbygning og funksjon	1,178	V
S210	Evolusjon	-2,080	K
S211	Livets kjemi	-3,001	K
S212	Genetikk	3,584	K
S213	Økologi	-2,417	K
S214	Planters oppbygning og funksjon	0,735	V
S215	Planters oppbygning og funksjon	-2,680	V
S216	Genetikk	0,759	V
S217	Cellen	-1,936	K
S218	Den biologiske diversitetens evolusjonære historie	-0,392	K
S219	Evolusjon	-0,772	K
S220	Dyrs oppbygning og funksjon	-0,065	K
S221	Cellen	1,102	K
S222	Evolusjon	2,379	A
S223	Dyrs oppbygning og funksjon	1,870	K
S224	Livets kjemi	-1,189	V
S225	Den biologiske diversitetens evolusjonære historie	0,100	K
S226	Dyrs oppbygning og funksjon	-0,844	V
S227	Dyrs oppbygning og funksjon	-0,676	A
S228	Evolusjon	0,616	A
S229	Dyrs oppbygning og funksjon	-1,768	K
S230	Cellen	-1,215	K

S231	Cellen	3,236	K
S232	Den biologiske diversitetens evolusjonære historie	-0,557	K
S233	Økologi	0,784	K
S234	Dyrs oppbygning og funksjon	1,956	K
S235	Planters oppbygning og funksjon	2,542	A
S236	Den biologiske diversitetens evolusjonære historie	-2,169	K
S237	Dyrs oppbygning og funksjon	-0,275	K
S238	Dyrs oppbygning og funksjon	2,710	K
S239	Genetikk	0,520	K
S240	Dyrs oppbygning og funksjon	2,347	V
S241	Dyrs oppbygning og funksjon	-3,194	K
S242	Den biologiske diversitetens evolusjonære historie	-1,265	K
S243	Den biologiske diversitetens evolusjonære historie	-0,557	K
S244	Planters oppbygning og funksjon	1,786	K
S245	Den biologiske diversitetens evolusjonære historie	2,073	A
S246	Genetikk	-1,394	K
S249	Planters oppbygning og funksjon	-0,844	K
S250	Planters oppbygning og funksjon	0,880	A
S251	Dyrs oppbygning og funksjon	0,051	A
S252	Evolusjon	3,001	K
S253	Genetikk	-1,605	K
S254	Cellen	4,991	K
S255	Genetikk	0,215	K
S256	Den biologiske diversitetens evolusjonære historie	-2,514	K
S258	Den biologiske diversitetens evolusjonære historie	-2,322	K
S261	Dyrs oppbygning og funksjon	-0,369	V
S262	Evolusjon	0,640	K
S263	Genetikk	-0,392	K
S264	Den biologiske diversitetens evolusjonære historie	0,309	K
S265	Dyrs oppbygning og funksjon	1,330	A
S266	Evolusjon	-2,714	V
S268	Cellen	0,122	K
S269	Dyrs oppbygning og funksjon	-1,524	K
S270	Cellen	0,053	K
S271	Dyrs oppbygning og funksjon	1,229	A
S272	Dyrs oppbygning og funksjon	2,191	K
S273	Genetikk	2,222	K

S274	Økologi	-2,259	K
S275	Cellen	-0,205	K
S276	Dyrs oppbygning og funksjon	1,927	A
S277	Planters oppbygning og funksjon	-0,990	K
S278	Genetikk	0,784	V
S280	Planters oppbygning og funksjon	0,028	K
S281	Genetikk	-0,392	A
S282	Genetikk	1,178	K
S283	Dyrs oppbygning og funksjon	2,014	A
S284	Dyrs oppbygning og funksjon	3,076	K
S285	Evolusjon	-0,439	V
S286	Den biologiske diversitetens evolusjonære historie	0,098	K
S287	Dyrs oppbygning og funksjon	0,426	K
S288	Økologi	-1,823	K
S289	Planters oppbygning og funksjon	0,006	A
S290	Planters oppbygning og funksjon	0,402	K