Analyse av flervalgstest som eksamensform på bachelorutdanning i biologi ved UiB

Masteroppgave i biologididaktikk

av

Sigrid Booman Folkvord

Institutt for biologi

Universitetet i Bergen

Juni 2016

Forord

Denne masteroppgaven markerer slutten på noen fine og lærerike år ved Lektorutdanningen ved Universitetet i Bergen.

En stor takk til min veileder, førsteamanuensis Tom Olav Klepaker for veiledning og støtte gjennom hele prosessen.

Videre må jeg få takke:

Professor Torbjørn Torsheim som ga meg gode råd i forbindelse med testteori og analyse av data.

bioCEED for masterstipend og interessante møter med bioCEEDs forskningsgruppe som jeg lærte mye av.

Arild Folkvord, for at du til tider har fungert som en ekstra veileder og orakeltjeneste, men først og fremst takk for at du er *Pappa*.

Lars, for all støtte og omsorg.

Bergen, 01.06.2016
Sigrid Booman Folkvord

Sammendrag

Eksamen i høyere utdanning har tradisjonelt fokusert på kontrollaspektet ved vurdering. I dag er det et økende fokus på læringsaspektet ved vurdering noe som vises i interessen for autentiske og varierte vurderingsformer. Flervalgstester er en vurderingsform som har blitt mer utbredt i Norge og blir brukt i flere emner ved Institutt for biologi ved Universitetet i Bergen. Digitale vurderingsplattformer har bidratt til å gjøre flervalgstester til en tidsbesparende vurderingsform, men gode flervalgstester er krevende å konstruere. Målet med denne masteroppgaven er å gjennomføre en forskningsbasert vurdering av en flervalgseksamen, basert på analyse av oppgaver og hvordan studentene svarer.

Utvalget bestod av resultatene fra en flervalgseksamen i biologi bestående av 60 oppgaver. Data fra 88 respondenter ble analysert kvantitativt og kvalitativt med utgangspunkt i klassisk testteori.

Resultatene viste at det er en overvekt av oppgaver som tester på lavt kognitivt nivå jamfør Blooms taksonomi. Gjennomsnittlig vanskelighetsgrad for testen ble beregnet til å være lettere enn det som er anbefalt og noen oppgaver diskriminerer dårlig mellom dyktige og mindre dyktige studenter. De fleste oppgavene hadde svaralternativer som svært sjeldent ble valgt. Det anbefales derfor å redusere antall svaralternativer fra fem til fire. Statistiske metoder indikerte at testresultatene er pålitelige. Det skyldes i stor av at testen har et høyt antall oppgaver.

Oppgaven belyser flere utfordringer knyttet til flervalgstest som eksamensform og illustrerer hvor viktig det er å analysere oppgaver og testresultater for å utbedre oppgavene før de eventuelt brukes på nytt i en ny eksamen.

Innholdsfortegnelse

Kapittel 1 – Innledning

1.1 Bakgrunn for oppgaven

I en nylig utgitt bok om eksamen og alternative vurderingsformer etterlyser Raaheim (2016) en eksamensrevolusjon. Det foregår allerede en bred satsing i andre land der intensjonen er å utvikle mer *autentiske* prøveformer som er tilpasset de kompetanser som er viktige for dagens samfunn og fremtidig læring (Kunnskapsdepartementet, 2000, Kapittel 13). Den økende interessen for eksamensspørsmål kan ha sammenheng med endringer i høyere utdanning som har ført til en sterk økning i antall eksamener (Kunnskapsdepartementet, 2000, Kapittel 13). Tidligere Kunnskapsminister Kristin Halvorsen påpekte behovet for mer norsk forskning på vurdering i høyere utdanning (Kunnskapsdepartementet, 2006).

Flervalgstester, eller multiple-choice tester, har lenge vært utbredt i land som USA og England. En av vurderingsformens styrker er at det tar kort tid å rette flervalgsoppgaver og det kan spekuleres i om det er en viktig årsak til dens popularitet. Vurderingsformen har imidlertid flere styrker, og det er ikke uten grunn at den benyttes både i PISA- og TIMSS-undersøkelser. Den økte bruken av flervalgstester i Norge i nyere tid kan delvis være et resultat av den raske utviklingen av digitale verktøy og vurderingsplattformer som itslearning, Inspera Assessment, Kahoot med flere. Slike vurderingsplattformer kan skåre flervalgsoppgaver automatisk, og dermed blir vurderingsformen ytterligere tidsbesparende. Det er imidlertid mindre kjente sider ved flervalgstester som gjør formatet mer tidkrevende og komplisert enn mange er klar over. En flervalgstest som skal teste faglig kompetanse er mer krevende å konstruere enn en triviell quiz. En overordnet begrunnelse for å bruke flervalgstester til vurdering i høyere utdanning er tanken om at studentene skal møte ulike vurderingsformer i løpet av studiet.

1.3 Problemstilling

Hensikten med denne masteroppgaven er å gjennomføre en forskningsbasert vurdering av en flervalgseksamen, basert på analyse av oppgaver og hvordan studentene svarer. En overordnet begrunnelse er viktigheten av å forbedre kvaliteten i vurderingen, for det er alltid forbedringspotensiale.

Problemstillinger for oppgaven er:

- Hvilke styrker og svakheter har eksamenssettet?
    - Hvordan diskriminerer oppgavene mellom dyktige og mindre dyktige studenter?
    - Hvordan er vanskelighetsgraden til oppgavene?
- Hvordan kan eksamenssettet utvikles for å styrke testens reliabilitet og validitet?

Begrepene reliabilitet og validitet beskrives i delkapittel 2.4.

For å besvare problemstillingene vil jeg bruke testresultatene fra en eksamen i emnet BIO102 - Organismebiologi 2 som ble gitt digitalt (Institutt for biologi, 2016b). Det aktuelle eksamenssettet bestod av 60 flervalgsoppgaver og ble besvart av 88 respondenter.

Kapittel 2 – Teori

2.1 Vurdering *av* læring i høyere utdanning

I litteraturen er det vanlig å skille mellom *summativ* og *formativ* vurdering. Summativ vurdering omtales ofte som vurdering *av* læring og finner sted når det er forventet at læringsmål skal være nådd (Woolfolk, 2004, side 398. Eksamen og prøver som danner grunnlag for en karakter er typisk summativ vurdering. Formativ vurdering omtales ofte som vurdering *for* læring eller underveisvurdering. Vurdering for læring innebærer at vurderingen skal bidra til å forme undervisningen i etterkant av vurderingen. Denne formen for vurdering har et diagnostisk aspekt ettersom undervisere kan få informasjon om misoppfatninger blant elever eller studenter.

I skolen har vurdering for læring vært en nasjonal satsning siden 2010 på bakgrunn av betydningen vurdering har for elevenes læring (Utdanningsdirektoratet, 2014). I høyere utdanning har det tradisjonelt vært større fokus på vurdering av læring og kontrollaspektet, men mye tyder på at læringsaspektet ved vurdering stadig får mer oppmerksomhet (Kunnskapsdepartementet, 2000, Kapittel 13).

2.1.1 Eksamen

Eksamen i høyere utdanning er en form for summativ vurdering der hensikten er å kontrollere studentenes læringsresultater (/kompetanse). Studentene blir gitt en form for sertifisering eller karakter som reflekterer i hvilken grad læringsmålene er oppnådd (Eggen, 2008). Eksamen kan likevel ha et formativt aspekt. I følge Dysthe (2008) er det ikke *form* som skiller mellom summativ og formativ vurdering, men hensikten med vurderingen og hvordan den brukes. Med andre ord kan en eksamen ha et formativ aspekt om det bidrar til å justere fremtidig undervisning. Dessuten vil studentenes tilnærming til stoffet, altså læringsstrategier, påvirkes av det faktum at de skal ta en eksamen (Raaheim, 2016, side 27).

I Norges offentlige utredninger, nr. 14, om høyere utdanning og forskning i Norge viser Kunnskapsdepartementet (2000) til at vi i Norge fortsatt befinner oss i en utviklingsfase på

eksamensområdet der tradisjonelle eksamener er mest utbredt. I andre land er bruken av psykometriske tester mer utbredt. Dette er en form for flervalgstester som objektivt skal måle respondentens egenskaper eller personlighet. Ambisjonen er at eksamen består av "autentisk" prøving, som vil si prøving der det legges mer vekt å *anvendelse* av kunnskap og et større fokus på (formativ vurdering og) læringsaspektet. Denne formen for vurdering er det økende interesse for. Raaheim (2016) beskriver en rekke alternative vurderingsformer i boken *Eksamensrevolusjonen – råd og tips om eksamen og alternative vurderingsformer.* En utfordring er at eksamen også skal ha et kontrollaspekt som setter krav til objektivitet og reliabilitet (pålitelighet). En løsning på denne utfordringen kan være å kombinere en portefølje/mappevurdering (for eksempel lab.-rapporter og skriftlige oppgaver) med en avsluttende prøve. Nevnte NOU (Kunnskapsdepartementet, 2000, Kapittel 13) trekker frem dette som en vurderingsform med flere fordeler. En fordel er at det legger opp til at studentene må jobbe jevnt med faget. Den avsluttende prøven bør dekke store deler av pensum og legge vekt på forståelse. En flervalgstest kan derfor være en passende vurderingsform. Raaheim (2016) er imidlertid mindre positiv til objektive vurderingsformer ettersom det kan påvirke studentenes læringsstrategier i en retning der det fokuseres mer på overflatisk læring enn dyp læring.

Valg av eksamensform vil naturligvis ha et kostnadsaspekt. Å gjennomføre, for eksempel, muntlig eksamen av 100 studenter er ikke minst tidkrevende, men også dyrt for universitetet. Bruk av eksterne sensorer er et tiltak som har til hensikt å kvalitetssikre eksamen, men det medfører en stor utgiftspost (Kunnskapsdepartementet, 2000, Kapittel 13). En flervalgstest som kan rettes automatisk vil kunne gi mye informasjon på en økonomisk måte og er sannsynligvis en av grunnene til at denne eksamensformen blir stadig mer brukt. Det må likevel nevnes at digitale vurderingsplattformer som Inspera Assessment ikke er gratis.

2.1.2 Vurdering ved Institutt for biologi

Ved bachelorutdanningen i biologi ved UiB ble det i 2011 innført en reform der et viktig mål var å øke studentaktive lærings- og vurderingsformer. I dag har de fleste biologiemnene kombinert ulike lærings- og vurderingsformer for å fremdyrke og vurdere ulike ferdigheter

og typer kunnskap (bioCEED, 2014). I bioCEED sin årsrapport for 2015 vises det til at det har vært gjennomført en kritisk evaluering av vurderingsformene som blir brukt ved Institutt for biologi. Det har blant annet vært fokus på å sikre at vurderingen samsvarer med oppgitt læringsutbytte (bioCEED, 2015). Noen emner har én slutteksamen som er 100 % av grunnlaget for endelig karakter i emnet, men det har blitt mer vanlig at emner består av flere vurderingssituasjoner som tilsammen danner grunnlaget for endelig karakter. Et eksempel er emnet BIO100 – Innføring i evolusjon og økologi, der karaktergrunnlaget består av tre deleksamener og én slutteksamen (Institutt for biologi ved Universitetet i Bergen, 2016a).

2.1.3 Hva skal vurderes?

I en beskrivelse av bachelorprogrammet ved UiB og i emnebeskrivelser av enkeltemner oppgis det forventet læringsutbytte etter fullført studium. Ofte er læringsmålene delt inn i kategoriene *kunnskaper*, *ferdigheter* og *generell kompetanse* (Universitetet i Bergen, 2016). Dette er begreper det kan være vanskelig å skille. I Nasjonalt kvalifikasjonsrammeverk for livslang læring (NKR) beskrives kategoriene kunnskaper, ferdigheter og generell kompetanse på følgende vis:

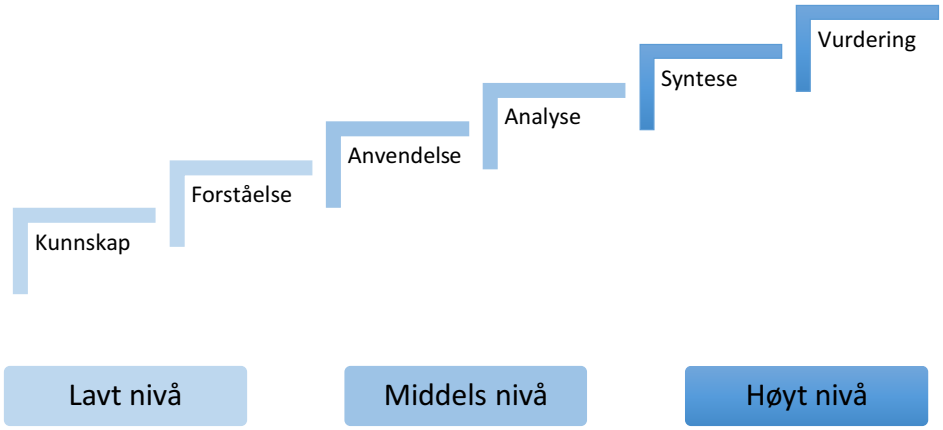
> - **Kunnskaper**: Kunnskaper er forståelse av teorier, fakta, begreper, prinsipper, prosedyrer innenfor fag, fagområder og/eller yrker.
> - **Ferdigheter**: Evne til å anvende kunnskap til å løse problemer og oppgaver. Det er ulike typer ferdigheter – kognitive, praktiske, kreative og kommunikative ferdigheter.
> - **Generell kompetanse**: Generell kompetanse er å kunne anvende kunnskap og ferdigheter på selvstendig vis i ulike situasjoner gjennom å vise samarbeidsevne, ansvarlighet, evne til refleksjon og kritisk tenkning i utdannings- og yrkessammenheng. (Kunnskapsdepartementet, 2011, side 16.)

Begrepet *kompetanse* spesielt kan ha noe ulik betydning fra land til land. Det er et vidt begrep som omfatter både kunnskap og ferdighet, og evnen til å anvende disse. Begrepet

kompetanse er heller ikke begrenset til kognitive dimensjoner. Det kan inkludere tekniske ferdigheter, relasjonelle egenskaper, holdninger og etiske verdier.

Blooms kognitive taksonomi

For å klassifisere læringsmål benyttes ofte anerkjente Blooms taksonomi over kognitive ferdigheter (Sirnes, 2005). Blooms taksonomi er delt inn i seks hierarkiske kunnskapsnivåer: kunnskap, forståelse, anvendelse, analyse, syntese og vurdering (Bloom, 1956). Kunnskap og forståelse kan plasseres i kategorien *lavt* kunnskapsnivå, anvendelse og analyse kan plasseres i kategorien *middels* kunnskapsnivå og syntese og vurdering kan plasseres i kategorien *høyt* kunnskapsnivå (Figur 2.1).



Figur 2.1. Adaptert fremstilling av Blooms taksonomi over kognitive nivåer.

Basert på Blooms taksonomi kan en generelt plassere læringsmål knyttet til *ferdigheter* på middels eller høyt nivå ettersom ferdigheter er evnen til å anvende kunnskap og løse problemer og oppgaver. Blooms taksonomi kan være et nyttig verktøy når man skal formulere læringsmål og planlegge vurdering som inkluderer oppgaver som tester på lavere, middels og høyt nivå. Haladyna (1994, side 7) påpeker at testing av høyere kunnskapsnivå sjeldent er adekvat. For å lettere kunne kategorisere oppgaver har Sirnes (2005, side 23) presentert nøkkelverb for de ulike nivåene (Tabell 2.1).

Tabell 2.1: Nøkkelverb for kunnskapsnivåene i Blooms taksonomi

Kunnskapsnivå	Nøkkelverb
Kunnskap	beskrive, definere, gjengi, presentere, regne opp
Forståelse	bevise, forklare, oversette, skjelne, tolke
Anvendelse	avlese, bruke, demonstrere, måle, registrere
Analyse	dele opp, identifisere, klassifisere, skille ut, sammenligne
Syntese	forstå, generalisere, organisere, produsere, trekke slutninger
Vurdering	avgjøre, bedømme, kritisere, skille mellom, velge

Ved noen vurderingsformer er det muligens ikke utelukkende kompetanse i biologi som vurderes. For eksempel kan evne til å uttrykke seg skriftlig påvirke vurderingen ved åpne drøftingsoppgaver. Enhver vurderingsform vil ha sine styrker og svakheter, og det viktigste er ikke *hvilken* vurderingsform en bruker, men *hvordan.*

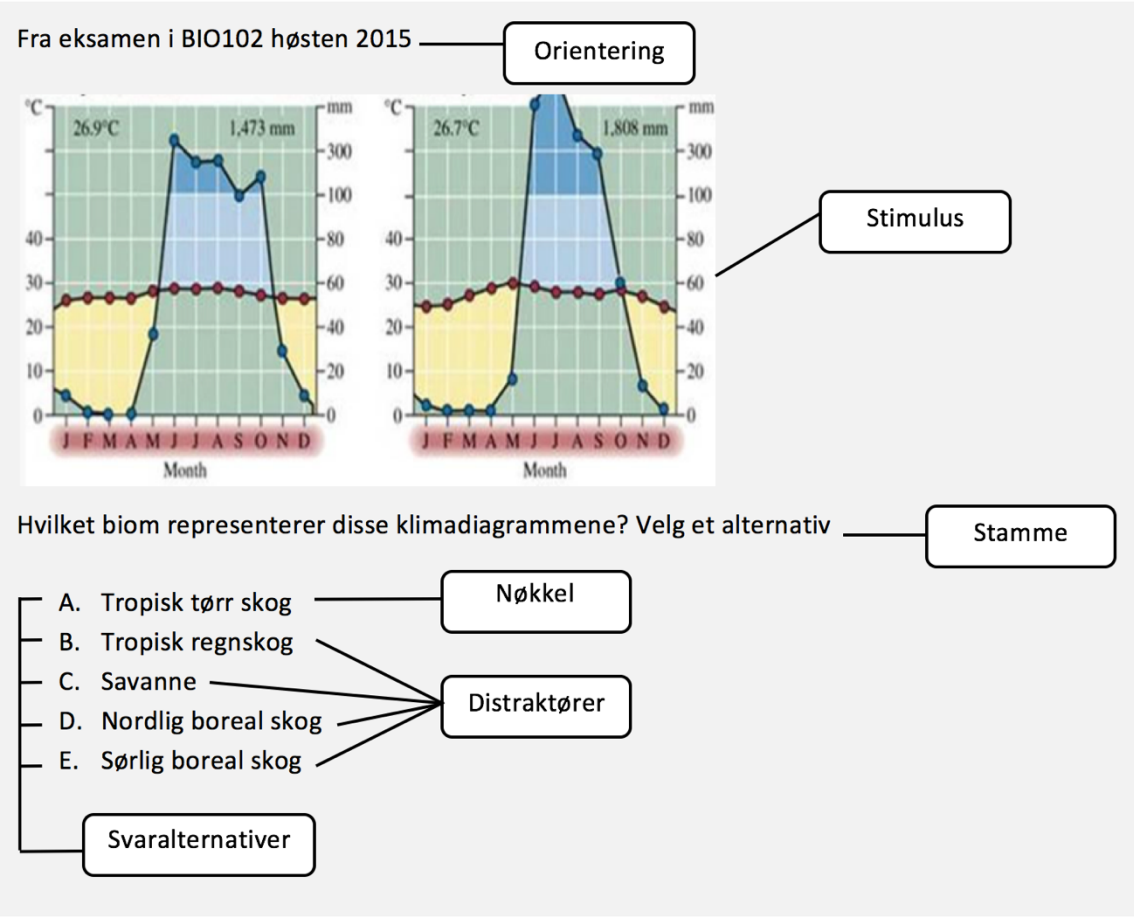
2.2 Flervalgsoppgaver

En flervalgsoppgave er en lukket oppgave der respondenten velger mellom to eller flere formulerte svaralternativer. Til sammenligning krever en åpen oppgave at respondenten besvarer oppgaven med egne ord og med slike oppgaver vil det være rom for flere tolkninger. I motsetning til åpne oppgaver er flervalgsoppgaver en objektiv vurderingsform ettersom det riktige svaret er gitt og vurderingen ikke avhenger av hvem som er sensor. En flervalgsoppgave kalles også et *testledd*, eller *item* på engelsk.

2.2.1 Oppbygging av en flervalgsoppgave

En konvensjonell flervalgsoppgave består av en *stamme,* som er oppgavens problemstilling. Deretter følger det minst to svaralternativer. Det riktige svaret er oppgavens *nøkkel*, mens gale svaralternativer kalles *distraktører*. Vanligvis er det kun én nøkkel, men det er mulig å lage oppgaver med flere nøkler. Noen flervalgsoppgaver har i tillegg en *stimulus* som kan være en illustrasjon (Boks 2.1), en tekst eller et lydelement dersom flervalgstesten er digital.

Boks 2.1: Oppbygging av en flervalgsoppgave med stimulus og fem svaralternativer der én av dem er nøkkel



Varianter av flervalgsoppgaver

Det finnes andre varianter av flervalgsoppgaver enn eksempelet i Boks 2.1. Sant-usant-oppgaver er en form for flervalgsoppgaver som er enkle å lage, men er ikke godt egnet til summativ vurdering da gjetting er en betydelig faktor. I tillegg gir det ingen verdifull diagnostisk informasjon (Sirnes, 2005). Kombinasjonsoppgaver består av to kolonner der ord eller setninger fra den ene kolonnen skal kombineres med ord eller setninger fra den andre kolonnen. Slike oppgaver er ofte enkle å lage, men er et relativt nytt format som man har begrenset kunnskap om (Amin *et al.,* 2016; Haladyna *et al.,* 2002).

Flervalgsoppgaver med flere nøkler (Complex multiple choice) blir sett på som vanskeligere enn oppgaver med én nøkkel. Formatet er imidlertid mindre effektivt med tanke på tid og er generelt ikke anbefalt (Haladyna *et al.,* 2002). I en studie av Kubinger et al. (2010) ble to ulike format av flervalgsoppgaver sammenlignet. Resultatet viste at flervalgsoppgavene med

formatet to nøkler og tre distraktører var vanskeligere enn formatet med én nøkkel og fem distraktører. Oppgavene ble skåret dikotomt, noe som innebærer at alle nøkler og ingen av distraktørene må være valgt for å få riktig. Formatet kan dermed redusere effekten av gjetting. I følge Thayn (2011) kan flervalgsoppgaver med flere nøkler være et godt alternativ, men de tar lengre tid å besvare, noe som kan ha konsekvenser for antall oppgaver i en test.

Oppgaver med stimulus (Context-dependent items) er et interessant format som åpner for flere muligheter. En stimulus kan være et virkemiddel for å lage oppgaver som testet høyere kunnskapsnivå (Haladyna *et al.,* 2002). Et oppsett som kan teste evnen til problemløsing er en stimulus i form av et tekstutdrag som introduserer et problem etterfulgt av flere oppgaver knyttet til tekstutdraget (Haladyna, 1994, side 47). Oppgaver med stimulus er et vanlig format i PISA-undersøkelsene (Universitetet i Oslo, 2016) og TIMSS-undersøkelsene (Universitetet i Oslo, 2006).

2.2.2 Fordeler med flervalgsoppgaver

Noen fordeler med flervalgsoppgaver har allerede blitt nevnt. Objektiv vurdering har sine svakheter, men en fordel er at vurderingen ikke påvirkes av respondentens evne til å uttrykke seg. Flervalgstester er tidsbesparende av flere grunner og dette er trolig en av de største grunnene til at vurderingsformens popularitet. Skåringen av en flervalgstest kan gjøres raskt. Dersom testen er gitt digitalt og skåres automatisk, så kan respondenten få tilbakemelding umiddelbart. Det tar mindre tid å besvare en flervalgsoppgave enn en åpen oppgave. En flervalgstest kan med andre ord bestå av et vesentlig større antall oppgaver enn en skriftlig eksamen med essay-oppgaver. Et resultat av dette er at testen kan dekke flere deler av pensum noe som potensielt kan øke reliabilitet og validitet. Reliabilitet og validitet er beskrevet senere i oppgaven. Gode flervalgsoppgaver kan brukes om igjen til testing av en ny gruppe. Fra et didaktisk ståsted er mulighetene for å analysere testresultatene og dermed få verdifull diagnostisk informasjon kanskje den mest interessante fordelen med flervalgstester. Oppgaveanalyse vil bli beskrevet nærmere i delkapittel 2.2.5.

2.2.3 Ulemper med flervalgsoppgaver

Flervalgsoppgaver blir ofte kritisert for å kun måle faktakunnskaper og evne til å gjenkjenne pensum. Det er imidlertid mulig å lage flervalgsoppgaver som tester høyere kognitive ferdigheter som for eksempel analyse og vurdering, men det kan være utfordrende (Haladyna, 1994; Sirnes, 2005; Woolfolk, 2004). Fakta- og detaljorienterte oppgaver er ikke unikt for flervalgsoppgaver. Kortsvarsoppgaver som spør om *hva*, *hvor* og *når* kan være vel så detaljorienterte. I realiteten er det flere eksempler på analyser av tidligere eksamensoppgaver og flervalgstester som viser at det er en stor overvekt av flervalgsoppgaver med lett vanskelighetsgrad og lav diskrimineringsevne (Madhav, 2015) og at de sjeldent tester høyere kognitive ferdigheter (Domyancich, 2014). Dette understreker hvor krevende det kan være å lage gode flervalgsoppgaver. En kan si at etterarbeidet med en flervalgstest tar mindre tid enn andre vurderingsformer, men forarbeidet må ikke undervurderes. En grunn til at gode flervalgsoppgaver er tidkrevende å lage er prosessen med å finne egnede distraktører. Er distraktørene lite plausible så vil det påvirke vanskelighetsgraden til oppgaven.

Raaheim (2016) uttrykker skepsis til flervalgstest som summativ vurdering. Det antas at studentenes læringsstrategi påvirkes av vurderingsform. En undersøkelse utført av Scouller (1998) viste at studenter som forberedte seg på en flervalgstest hadde en tendens til å benytte seg av en overfladisk tilnærming til stoffet, det vil si fokus på hukommelse og gjengivelse. Til sammenligning brukte studenter som forberedte seg på skrive et essay en tilnærming som involverte dypere læring, det vil si fokus på forståelse. Videre oppfattet studentene som tok flervalgstesten at den testet lavere nivåer av kognitiv tenking.
En annen studie der tidligere eksamensoppgaver ble analysert viste at de sjeldent testet høyere kognitive ferdigheter (Domyancich, 2014). Dette illustrerer at analyse og revidering av oppgaver er viktig for å sikre at flervalgstesten også tester høyere kognitive ferdigheter.

Muligheten til å gjette riktig svar er en svakhet ved flervalgsoppgaver. Et raskt Google-søk gir en rekke tips og strategier for å øke sannsynligheten for å gjette riktig. Tipsene basers ofte på statistikk som for eksempel viser at noen svaralternativer oftere eller sjeldnere er riktige, eller at det er større sannsynlighet for at det lengste svaralternativet er riktig. For de som lager flervalgstester kan det være lurt å være oppmerksom på dette og kontrollere at

oppgavene ikke følger bestemte mønstre. Når flervalgstester gis digitalt er det enkelt å sørge for at rekkefølgen på svaralternativene randomiseres. For å minimere effekten av gjetting kan en øke antall oppgaver i testen (Sirnes, 2005). Videre finnes det flere skåringsmetoder som korrigerer eller justerer skårer med hensyn til gjetting. Skåringsmetoder og gjetting blir beskrevet nærmere i neste delkapittel.

2.2.4 Konstruksjon av flervalgstester

Når en skal lage en test må en blant annet overveie hva hensikten med testen er, hva den skal teste og hva den skal inneholde (Cohen *et al.,* 2011). Konstruksjon av flervalgs*tester* består i stor grad av konstruksjon av en rekke flervalgs*oppgaver,* men det er flere ting å ta hensyn til. Med tanke på innholdet til testen bør man kontrollere at det har sammenheng med viktige læringsmål for emnet (Haladyna *et al.,* 2002). En av styrkene til flervalgstester er at de kan dekke store deler av pensum. Derfor er det i utgangspunktet ingen grunn til at noen læringsmål ikke testes, spesielt hvis flervalgstesten er eneste summative vurdering i emnet. Videre må det sørges for at nøkkelens posisjon varierer eller er randomisert. Ideelt sett bør oppgaver pre-testes for å avdekke eventuelle svakheter som bør utbedres (Sirnes, 2005, side 42).

Språket bør holdes enkelt og konsist for å hindre misforståelser og unødig ulempe for respondenter med lesevansker (Haladyna *et al.,* 2002). Om det er faglig kompetanse en ønsker å teste bør en unngå «lurespørsmål» og humor som hører bedre hjemme i en uformell quiz.

Stammen

Stammen skal inneholde selve oppgaven som kan være formet som et spørsmål eller et ikke-avsluttet utsagn (Sirnes, 2005). Det anbefales at stammen ikke er for lang eller inneholder overflødig informasjon (Haladyna *et al.,* 2002). Som hovedregel bør spørsmålet i stammen være såpass fokusert og tydelig at det er mulig å svare på spørsmålet uten å se svaralternativene. Negasjoner i stammen som *ikke, unntatt* og *aldri* kan fort bli oversett av

respondenten og bør unngås (Sirnes, 2005). I tilfeller der det er ønskelig å teste om respondenten vet hva som *ikke* er riktig, så bør negasjonen være i kursiv eller fremhevet på en annen måte slik det er gjort i denne setningen (Haladyna *et al.,* 2002). En alternativ formulering kan være: ”hvilket svaralternativ er galt?”. Da skal det mer til for at respondenten misforstår oppgaven.

Svaralternativene

Fire eller fem svaralternativer der én av dem er nøkkelen er et vanlig format. Oppgavens vanskelighetsgrad øker med antall svaralternativer, men det er utfordrende og tidkrevende å lage gode distraktører (Sirnes, 2005). Videre viser det seg at flervalgstester sjeldent har mer enn tre effektive svaralternativer (Haladyna & Downing, 1993). Distraktører som sjeldent blir valgt er ineffektive som svaralternativer. Den åpenbare ulempen med færre svaralternativer er at det øker sjansen for å gjette riktig. For å lage egnede distraktører bør en ta utgangspunkt i vanlige misoppfatninger hos studentene.

Studier har vist at det ofte er en tendens til at det lengste svaralternativet er det riktige (Mentzer, 1982). Når en lager en flervalgstest kan det være lurt å av og til la kortere svaralternativer være riktige. Der det er mulig bør en likevel forsøke å la svaralternativene være omtrent like lange og detaljerte.

Bruk av svaralternativet ”ingen av svaralternativene er riktige” kan øke vanskelighetsgraden til en oppgave (DiBattista *et al.,* 2014). Likevel blir det frarådet å bruke dette som et svaralternativ med mindre det tester et relevant læringsmål (DiBattista *et al.,* 2014; Haladyna *et al.,* 2002). Spesielt bør en unngå at det dannes et mønster der svaralternativet ”ingen av svaralternativene er riktige” vanligvis er enten distraktør eller nøkkel. Dersom dette svaralternativet er nøkkelen i oppgaven, er det mulig å skåre riktig på oppgaven selv om respondenten ikke kan det riktige svaret. Det er stor enighet om at svaralternativet ”alle svaralternativene er riktige” bør unngås (Haladyna *et al.,* 2002).

Skåringsmetoder og gjetting

Den enkleste måten å skåre en flervalgsoppgave er dikotom (todelt) skåring der respondenten blir belønnet med riktig svar, men ikke straffet for galt svar. I mange tilfeller vil det bli gitt 1 ”poeng” for riktig, og 0 for galt. ”Negative marking” er en metode som innebærer at respondenten får trekk for å svare feil. Hensikten er å forhindre gjetting (Lesage *et al*., 2013).

Sirnes (2005, side 45) oppgir den vanligste justeringsformelen for gjetting:

Skåre = Rett – (Galt/n-1)

$$skåre = rett - \frac{galt}{n-1}$$

I formelen er *n* antall svaralternativer i hver enkeltoppgave. Dersom en bruker justeringsformler, så må respondentene informeres om det. For en respondent som ikke aner hva som er riktig svar, vil det lønne seg å ikke svare på oppgaven fremfor å svare. Dersom en respondent klarer å eliminere noen av svaralternativene vil det lønne seg å gjette. Man kan diskutere om ”negative marking” fører til en fordel for strategiske respondenter og en ulempe for forsiktige respondenter. Det utvikles stadig nye skåringsmetoder. Lesage *et al*. (2013) beskriver skåringsmetoder som anerkjenner delvis mestring.

I klassisk testteori kan effekten av gjetting ignoreres dersom testen har mange nok oppgaver (Haladyna, 1994, side 152). Sannsynligheten for at en respondent får en ufortjent høy skår ved å gjette minker jo flere oppgaver testen inneholder. I tillegg kan terskelverdiene for de ulike bokstavkarakterene heves sammenlignet med typiske terskelverdier for åpne oppgaver. Innenfor IRT blir påvirkningen av parameteren *gjetting* beskrevet som en parameter med mindre påvirkningskraft en parameteren *diskriminering* (Hambleton *et al.,* 1991).

2.2.5 Oppgaveanalyse

Vanskelighetsgrad

Vanskelighetsgraden til en enkeltoppgave kan beregnes ved å fastsette andel av respondentene som har svart riktig på oppgaven. Denne verdien kalles gjerne p-verdi.

Gronlund referert i Sirnes (2005, side 66) beskriver formelen for beregning av p-verdi:

$$P = \frac{R}{T} \times 100$$

der

P = prosentandelen som svarte rett

R = antallet testtakere som svarte rett

T = det totale antall testtakere som svarte på oppgaven

P-verdien oppgis fra 0 – 1. Det vil si at en oppgave som 50 % av respondentene har svart riktig på tilsvarer en p-verdi på 0,50. P-verdien 1,00 vil si at alle kandidatene har svart riktig. Tilsvarende vil p-verdien 0,00 si at alle kandidatene har svart galt Sirnes (2005, s. 66). For en test bør målet være en gjennomsnittlig vanskelighetsgrad på rundt 0,50 (Sirnes, 2005, side 68). P-verdi er et enklere uttrykk å bruke enn ”andel respondenter som svarte riktig”.

Oppgavenes diskrimineringsevne

Diskriminering i denne sammenhengen betyr en oppgaves evne til å bli besvart riktig av respondenter som innehar kompetansen oppgaven er ment å teste og til å bli besvart feil av respondenter som ikke innehar den kompetansen (Cohen *et al.,* 2011, side 484). Det er med andre ord ønskelig at en test består av oppgaver som diskriminerer godt mellom dyktige og mindre dyktige respondenter.

Distraktøranalyse

En fungerende distraktør kan defineres som en distraktør valgt av ≥ 5 % av respondentene og oftere av lavt-skårende respondenter enn høyt-skårende respondenter (Ali & Ruit, 2015;

Hingorjo & Jaleel, 2012). Distraktører som svært få eller ingen har valgt bør vurderes nøye for å avklare hva som er årsaken til dette.

2.3 Testteori

2.3.1 Klassisk testteori

Klassisk testteori (KTT) har lenge vært benyttet til analyse av tester innenfor psykologi og utdanning (Hambleton, *et al.,* 1991). KTT antar at det er en *sann skår* for hver respondent som respondenten vil oppnå for hver gang testen tas, så lenge målingen er uten feil (Cohen *et al.*, 2011). I virkeligheten vil det ofte være feil i målingen som gjør at testen ikke gir en *sann skår*, men en *observert skår.*

Dette uttrykkes i følgende formel:

$$X = T + E$$

der

X = observert skår

T = sann skår

E = error (feil)

Resultater basert på analyser i KTT avhenger av utvalget som besvarer testen Cohen *et al.*, 2011). Vanskelighetsgraden som beregnes for en oppgave er basert på andel respondenter som svarte riktig. En oppgave som blir kategorisert som enkel basert på analyser av en testadministrasjon vil kunne bli kategorisert som vanskelig i en annen testadministrasjon. Resultatenes testavhengighet gjør det utfordrende å sammenligne resultater av respondenter fra ulike testadministrasjoner Cohen *et al.*, 2011).

2.3.2 Item Response Theory

Item Response Theory (IRT) ble utviklet som et svar på utfordringene knyttet til KTT (Hambleton *et al*., 1991). Teorien antar at det er et forhold mellom en respondents ferdighet eller egenskap og hvordan han/hun svarer på et testledd (Cohen *et al.,* 2011). IRT består av

flere modeller som har vist seg å være nyttige når det gjelder konstruksjon og evaluering av tester (Hambleton, *et al.,* 1991). Den største ulempen med IRT er at modellene stiller krav til større utvalg enn det som ofte er tilgjengelig (Hula *et al.,* 2012). For den enkle logistiske Rach-modellen med én parameter (respondentens dyktighet) anbefales det en minimum utvalgsstørrelse på mellom 50 og 200 (Hula *et al.,* 2012; Linacre, 1994). For en modell med to parametere (respondentens dyktighet og oppgavens diskriminering) er det anbefalt med utvalg på minst 350 (Embretson & Reise, 2000). For mer komplekse modeller stilles det enda høyere krav til størrelse på testutvalg. IRT har dermed begrenset nytte for analyse av testresultater med mindre utvalgsstørrelse.

2.4 Validitet og reliabilitet av flervalgstester

2.4.1 Validitet

Validitet, eller *gyldighet* som det også kalles, omhandler i hvilken grad en test måler det den er ment å måle (Cohen *et al.,* 2011, side 483). I følge Haladyna (1994, side 27) er det essensielt å evaluere om vurderingsformen

Innholdsvaliditet omhandler i hvilken grad oppgavene som testen inneholder, er representative for det faget eller emnet som elevene skal testes i (Sirnes, 2005, side 81). For å sikre innholdsvaliditet er det med andre ord viktig at det er en sammenheng mellom oppgitte læringsmål og vurderingen. Flervalgstester kan dekke store deler av pensum, noe som potensielt kan øke dens innholdsvaliditet (Sirnes, 2005, side 10). Det er et poeng at en vurderingsform ikke er mer gyldig/valid enn en annen (Schuwirth og van der Vleuten, 2004). For å sikre validiteten til en test er det viktig at den inneholder oppgaver som tester høyere kognitivt nivå (Haladyna, 1994).

2.4.2 Reliabilitet

Reliabilitet, eller *pålitelighet* som det også kalles, omhandler i hvilken grad testresultatene er *pålitelige* (Cohen *et al.,* 2011, side 483). Det er flere forhold som påvirker reliabiliteten til

testen. Antall observasjoner (oppgaver) er en nøkkelfaktor og en test med 60 flervalgsoppgaver er mer reliabel enn en test med 20 oppgaver. Dersom en test med flervalgsoppgaver gjør at en får testet større deler av pensum, så kan dette øke testens reliabilitet og innholdsvaliditet (Sirnes, 2005, side 10) Testens reliabilitet øker i utgangspunktet med antall distraktører per testledd, men det avhenger av at distraktørene fungerer godt (Haladyna & Downing, 1993).

Ulike vurderingsformer har ulike styrker og svakheter. Schuwirth og van der Vleuten (2004) påpeker at ingen vurderingsform er automatisk *upålitelige* og alle vurderingsformer kan potensielt være tilstrekkelig *pålitelige* så lenge de brukes på en passende måte. I følge Raaheim (2016) er dette en god grunn til å variere bruken av vurderingsformer. Haladyna (1994, side 27) argumenterer for at flervalgstester generelt har høyere reliabilitet enn essayoppgaver.

Det finnes flere metoder for å estimere reliabiliteten til en test. Metoder som krever at respondentene testes flere ganger er vanskeligere å gjennomføre av praktiske årsaker) og vil ikke bli beskrevet her. Indre konsistens-metoder krever at respondentene tar kun én test og er dermed enklere å bruke (Sirnes, 2005).

Indre konsistens-metoder

Split-half-metoden går ut på å skåre oddetalls- og partalssoppgaver hver for seg (Sirnes, 2005, side 83). Korrelasjonskoeffisienten *r* angir i hvilken grad de to delene av testen gir samme resultat. Med utgangspunkt i korrelasjonskoeffisienten *r* for de to delene av testen kan korrelasjonskoeffisienten for hele testen bestemmes ved å bruke Spearman-Brown-formelen.

Spearman-Brown-formelen:

$$\text{Reliabilitet til hele testen} = \frac{2r}{1+r}$$

I formelen er *r* korrelasjonskoeffiesienten for de to halvdelene av testen. Korrelasjonskoeffisienten vil være høyere for hele testen enn for de to halvdelene (Sirnes, 2005, side 83). Det viser hvor viktig antall oppgaver er for reliabiliteten av testresultatene.

En annen enkel måte å estimere reliabiliteten til testskårer er Kuder-Richardson formel 21. Den beregnes ut fra antall oppgaver i testen, gjennomsnittet og standardavviket.

Versjonen gitt i Sirnes (2005, side 84) ser slik ut:

$$KR21 = 1 - \frac{M(K - M)}{K(s^2)}$$

der

K = antall enkeltoppgaver i testen

M = gjennomsnittet på testskårene

s = standardavviket til testskårene

Er reliabilitetskoeffisienten (KR21) 0,00, så er det ingen reliabilitet, og om den er 1,00 så er reliabiliteten total. Verdier mellom 0,60 og 0,80 er vanlige for testet som tas av en enkelt klasse eller grupper studenter, mens ferdighetstester gjerne har verdier over 0,90 (Sirnes, 2005, side 84).

Cronbach's koeffisient alpha er et annet mål på intern konsistens (Sirnes, 2005, side 84).

Formelen for Cronbach's alpha ser slik ut:

$$\alpha = \frac{n(1 - \sum \sigma_i^2/\sigma_t^2}{n - 1}$$

der

n = antall oppgaver i testen

σ_i^2 = leddvarians

σ_t^2 = varians i sumskåren

Høy Cronbach's koeffisient alpha indikerer høy reliabilitet (Tabell 2.2) (Sirnes, 2005, side 86).

Tabell 2.2: Reliabilitetskoeffisienten Cronbach's alpha

Cronbach's alpha	**Reliabilitet**
> 0,90	svært høy
0,80 – 0,90	høy
0,70 – 0,80	middels
0,60 – 0,70	minimal
< 0,60	uakseptabel

Kapittel 3 – Metode

For å besvare problemstillingen ble det valgt å bruke både kvantitative og kvalitative metoder. Denne masteroppgaven er på mange måter et case-studie ettersom case-studier bruker spesifikke og aktuelle hendelser til å forklare noe mer generelt (Cohen *et al.,* 2011). Det bør likevel utvises forsiktighet med å generalisere om flervalgstest som vurderingsform basert på resultatene fra én flervalgstest.

3.1 Utvalg og datainnsamling

Utvalget består av 88 biologistudenter ved UiB som gikk opp til ordinær eksamen i Organismebiologi 2 (BIO102). Emnet er et av de obligatoriske grunnemnene i biologi ved UiB og tas normalt i tredje semester av bachelorprogrammet i biologi.

Ved valg av emne og eksamenssett til analysen ble følgende punkter tatt hensyn til:

- Antall studenter som tok eksamen gitt semester, jo flere jo bedre
- Antall flervalgsoppgaver i eksamenssettet, jo flere jo bedre
- Hvor tilgjengelig resultatene fra eksamen er

Organismebiologi 2 er det emnet som har flest flervalgsoppgaver (60). Høsten 2015 ble en 3-timers eksamen gitt digitalt via den digitale vurderingsplattformen Inspera Assessment (Inspera, 2016) og det var et relativt høyt antall studenter (88) som tok eksamen. Emneansvarlig for faget var positiv til en evaluering av vurderingen i emnet. Eksamenssettet gitt høsten 2015 i Organismebiologi 2 ble av disse grunner valgt som datasett for analysen av flervalgstest som eksamensform.

Det aktuelle eksamenssettet består av 60 flervalgsoppgaver som alle har fem svaralternativer og én nøkkel. For hver oppgave ble det gitt ett ”poeng” for riktig, og 0 for galt. Flere oppgaver har vært benyttet i tidligere eksamenssett.

Eksamensbesvarelsene ble manuelt oversatt (konvertert) fra PDF-format til datasett i Excel. Det viste seg at eksamensresultatene for alle respondentene som lå lagret på Inspera Assessment kun var tilgjengelig som et PDF-dokument på 3015 sider. De virkelige

kandidatnumrene til respondentene var ikke tilstede i dette dokumentet. I stedet ble de kalt kandidat 1, 2, 3 og så videre. Resultatene var med andre ord anonymisert.

Til analysene ble det behov for to versjoner av datasettet. Begge formatene inneholdt informasjon om de 88 respondentenes svar på de 60 oppgavene. Det mest detaljerte formatet inneholdt informasjon om hvilket svar respondenten har valgt, der svaralternativ A, B, C, D eller E ble kodet henholdsvis 1, 2, 3, 4 eller 5. Det andre formatet var et dikotomt datasett der riktig svar ble kodet 1 og galt svar ble kodet 0. Dette ga totalskåren til hver enkelt respondent.

3.2 Kvantitativ analyse

Kvantitative analyser baserer seg på hvordan studentene har besvart oppgavesettet. Disse av datasettet ble gjennomført i (Microsoft Excel og) programmet R versjon 3.2.2 for Mac (R Core Team, 2015). R-pakken ”ltm” ble brukt til deskriptive analyser av datasettet. (R Development Core Team, 2006). Pakken er tilgjengelig fra CRAN: (https://cran.r-project.org/web/packages/ltm/ltm.pdf).

I utgangspunktet var det planlagt å gjøre analyser direkte basert på Item Response Theory, men fordi denne analysemåten krever større utvalg enn hva som var tilgjengelig for å gi pålitelige resultat, ble det meste av analysene utført med utgangspunkt i klassisk testteori.

3.2.1 Oppgaveanalyse

Vanskelighetsgrad

Oppgavenes vanskelighetsgrad ble beregnet ut fra p-verdien, som er andelen av respondentene som har svart riktig på den oppgaven. Dette er beskrevet i delkapittel 2.2.5 i teoridelen av oppgaven.

Oppgavenes diskrimineringsevne

Point-biserialkorrelasjoner brukes innenfor klassisk testteori som et mål på en oppgaves diskrimineringsevne. Det er en Pearson-korrelasjon mellom skåren på hver oppgave, som kan være 0 eller 1, og totalskåren på testen. Verdiene vil være mellom – 1 (negativ korrelasjon) og 1 (positiv korrelasjon). El-Uri & Malas (2013) beskriver hvilke verdier regnes som svært god, god, middels og minimal diskrimineringsevne (Tabell 3.1). Hva som regnes som akseptable verdier varierer, men oppgaver med verdier nær eller mindre enn null bør fjernes. I følge Kibble &Johnson (2011) bør gjennomsnittlig point-biserialkorrelasjon for en test ligge i nærheten av 0,5.

Tabell 3.1: Point-biserialkorrelasjon som et mål på diskrimineringsevnen til en oppgave.

Point-biserialkorrelasjon	**Diskrimineringsevne**
≥ 0.40	svært god
0,30 – 0,40	god
0,10 – 0,30	middels
0,001 – 0,0099	minimal

Attali *et al.* (2000) retter kritikk til bruk av point-biserialkorrelasjoner som diskrimineringsindeks i flervalgsoppgaver. Det finnes en rekke metoder for å beregne diskrimineringsevne til oppgaver, men til tross for at de gir ulike numeriske verdier, så vil konklusjonen for om en oppgave bør forkastes ofte være den samme (Attali *et al.,* 2000).

Distraktøranalyse

Distraktørene ble analysert ved å se på svarfordelingen på de ulike svaralternativene (beskrevet i delkapittel 2.2.5 i teoridelen av oppgaven).

3.2.2 Reliabilitet

For å estimere reliabiliteten til testskårene ble følgende indre konsistens-metoder benyttet: Spearman-Brown-formelen, Kuder-Richardson formel 21 og Cronbach's koeffisient alpha (beskrevet i delkapittel 2.4.2).

3.3 Kvalitativ analyse

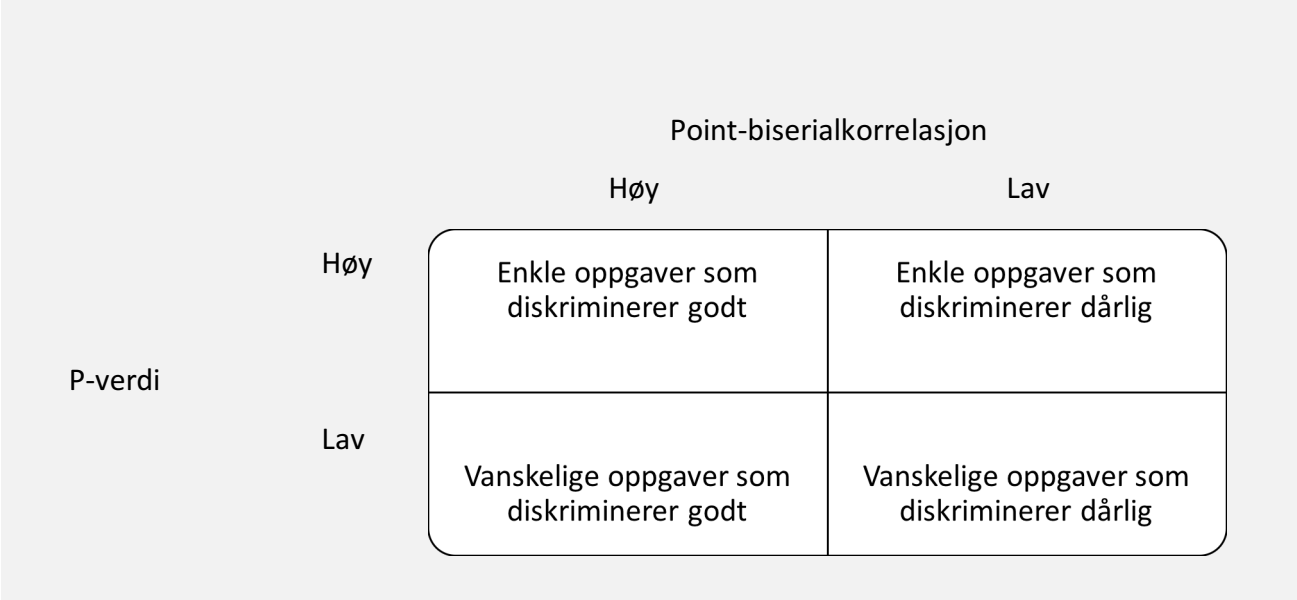
3.3.1 Kategorisering av oppgaver

Cohen *et al.*, (2011, side 482) foreslår en matrise som indikerer vektlegging av ulike tema og læringsmål. Med utgangspunkt i en slik matrise ble oppgave kategorisert utfra oppgitt læringsutbytte (læringsmål) eller oppgavens tema der det var vanskelig å koble oppgaven til et bestemt læringsmål, i tillegg til å kategorisere basert på om de tester kompetanse på *lavt* eller *høyt* kognitivt nivå jamfør Blooms taksonomi. Med lavere nivå menes *kunnskap* og *forståelse*. Høyt kognitivt nivå inkluderer vanligvis *syntese* og *vurdering.* I kategoriseringen ble oppgaver som testet på middels kognitivt nivå (*anvendelse* og *analyse*) plassert under *høyt* nivå for å gjøre kategoriseringen mer treffsikker. Oppgaver som tester evnen til å anvende og analysere kunnskap har elementer av høyere kognitivt nivå ettersom de krever mer kompleks kognitiv tenkning enn oppgaver som tester ren kunnskap. Hensikten med denne kategoriseringen av oppgavene er å vurdere testens innholdsvaliditet.

3.3.2 Kvalitativ vurdering av oppgaver basert på vanskelighetsgrad og point-biserialkorrelasjon

Den kvalitative vurderingen av oppgavene tok utgangspunkt i resultater fra kvantitativ oppgaveanalyse. Spesielt viktig ble det å se nærmere på oppgaver med enten veldig høy eller veldig lav vanskelighetsgrad, i tillegg til oppgaver med lav point-biserialkorrelasjon. Med utgangspunkt i kvantitative resultater for vanskelighetsgrad og diskrimineringsevne kan

oppgaver plasseres i en 2x2-matrise (Figur 3.1).



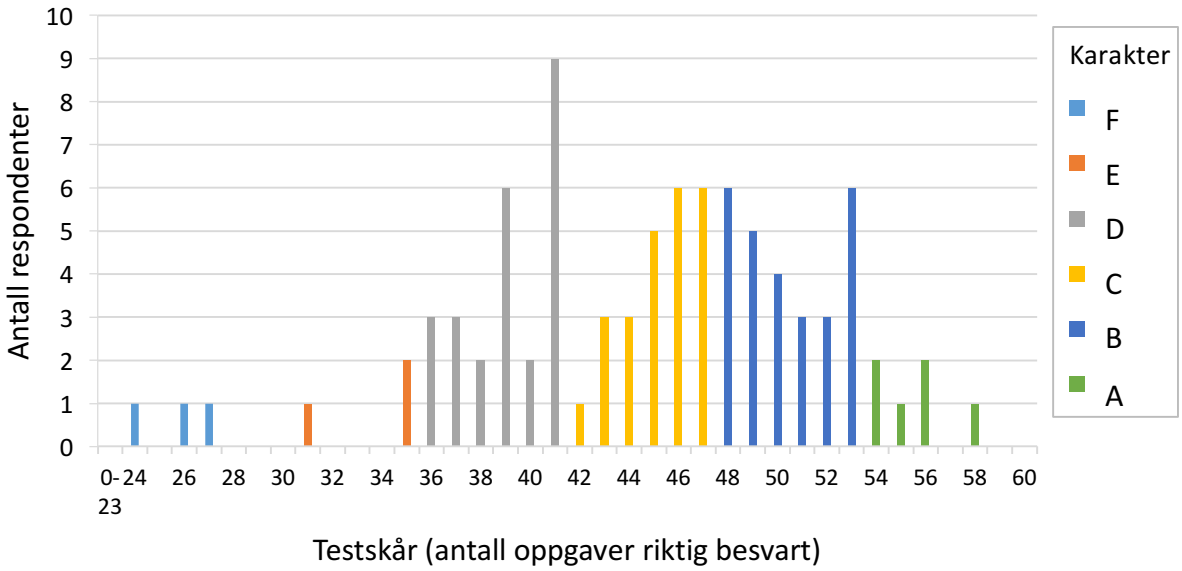
Figur 3.1: Matrise for kategorisering av oppgaver med hensyn til vanskelighetsgrad (andel riktig besvart, fra 0 til 1) og diskrimineringsevne (point-biserialkorrelasjon, fra -1 til +1).

Distraktøranalysen er utgangspunktet for å vurdere om noen distraktører bør fjernes eller revideres.

Kapittel 4 – Resultater

4.1 Testresultater

Testskår er beskrevet som antall oppgaver riktig besvart, der høyeste mulige skår er 60. Gjennomsnittlig skår er 44,7, medianen er 46, laveste skår er 24 og høyeste skår er 58 (Figur 4.1). Fordelingen er venstreskjev med noen få lave skårer. Gjennomsnittskarakteren er C. Tre studenter fikk karakteren F, som tilsvarer stryk.



Figur 4.1: Fordeling av testskår basert på antall totalt antall riktige svar der høyeste mulige skår er 60.

Terskelverdier for de ulike bokstavkarakterene ble bestemt av emneansvarlig (Tabell 4.1).

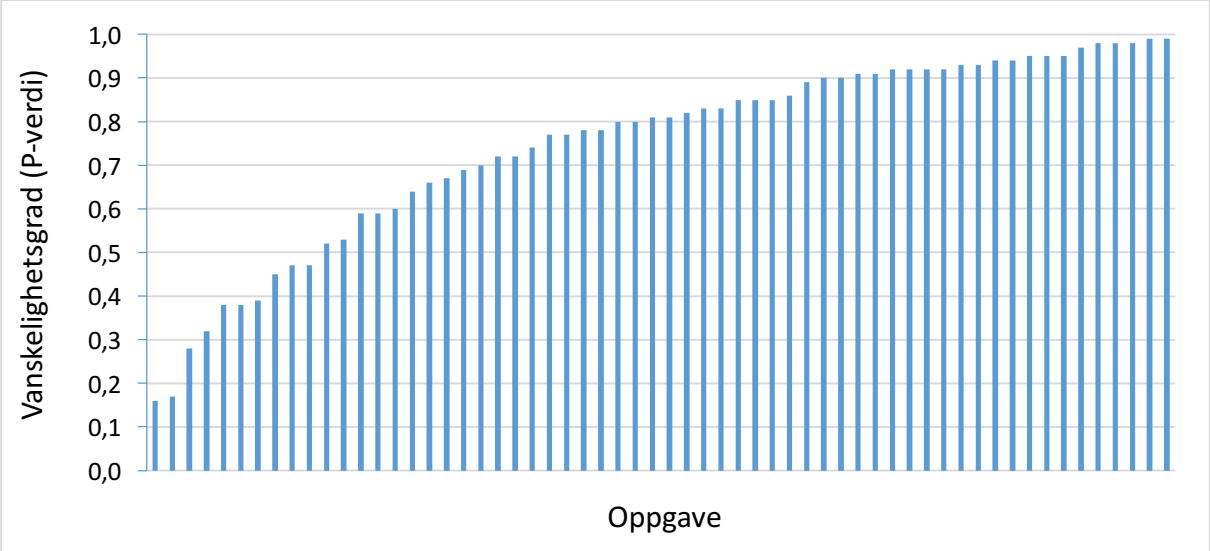
Tabell 4.1: Terskelverdier for bokstavkarakterer.

Bokstavkarakter	Terskelverdier
A	54-60
B	48-53
C	42-47
D	36-41
E	30-35
F	0-29

4.2 Oppgaveanalyse

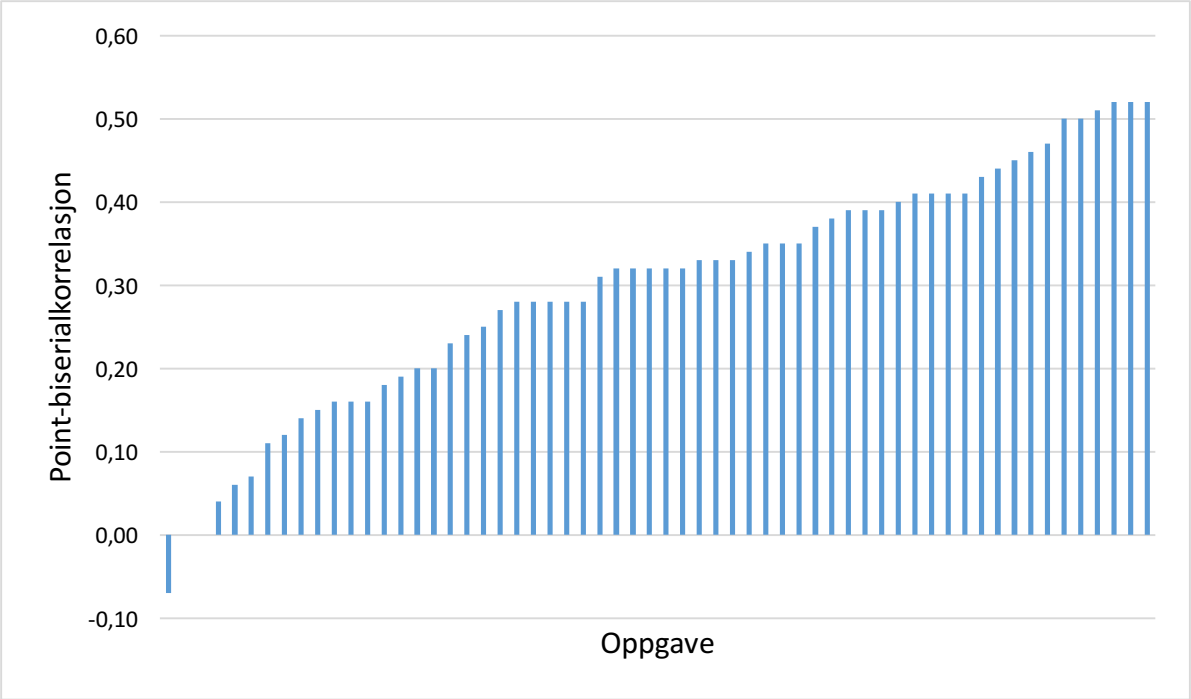
4.2.1 Oppgavenes vanskelighetsgrad og diskrimineringsevne

Oppgavenes vanskelighetsgrad ble beregnet ut fra p-verdien. P-verdier over 0,90 indikerer veldig lette oppgaver, mens verdier under 0,20 indikerer veldig vanskelige oppgaver. For dette eksamenssettet er 21 veldig lette oppgaver og 2 veldig vanskelige oppgaver (Figur 4.2). Gjennomsnittlig vanskelighetsgrad for oppgavene i testen er 0,75 med et standardavvik på 0,219. P-verdiene ligger i intervallet 0,16 – 0,99.



Figur 4.2: Vanskelighetsgrad beregnet utfra p-verdi for de 60 oppgavene i eksamenssettet rangert fra lavest p-verdi til høyest p-verdi. Gjennomsnittlig vanskelighetsgrad = 0,75.

Point-biserialkorrelasjonen for oppgavene ligger i intervallet -0,07 – 0,52 (Figur 4.3) Gjennomsnittlig point-biserialkorrelasjon er 0,30 med et standardavvik på 0,144. Kun én oppgave har negativ korrelasjon.



Figur 4.3: Point-biserialkorrelasjon for de 60 oppgavene i eksamenssettet ranger fra lavest til høyest.

For detaljerte p-verdier og point-biserialkorrelasjon for oppgavene i eksamenssettet se Tabell 7.1 i vedlegg.

4.2.2 Distraktøranalyse

De fleste oppgavene hadde én eller flere distraktører som ble valgt av svært få respondenter (Tabell 4.2). 46 av 60 oppgaver hadde én eller flere distraktører som ble valgt av 2 % av respondentene, eller færre. Det utgjør 77 % av oppgavene i eksamenssettet. 47 % av oppgavene hadde én eller flere distraktører som ikke ble valgt av noen av respondentene.

Tabell 4.2: Svarfordelingen til svaralternativene for de 60 oppgavene i testen. Nøkkelen for hver oppgave er markert med fet skrift. Svaralternativ A er oftest nøkkel (er randomisert i Inspera Assessment).

Oppgave	Svaralternativ					Oppgave	Svaralternativ				
	A	B	C	D	E		A	B	C	D	E
S1	0,05	0	0,01	0	**0,94**	S31	**0,95**	0	0,02	0,01	0,01
S2	0,02	0	**0,98**	0	0	S32	**0,98**	0	0	0	0,02
S3	0,01	0,06	0,01	**0,92**	0	S33	**0,66**	0,16	0,02	0,09	0,07
S4	0,26	0,01	0,11	**0,59**	0,02	S34	**0,91**	0,03	0	0	0,06
S5	**0,90**	0,03	0	0,07	0	S35	**0,69**	0,11	0,07	0,03	0,09
S6	0,01	0	0	**0,78**	0,20	S36	**0,80**	0	0,07	0,13	0,01
S7	0,05	0,02	**0,83**	0,08	0,02	S37	**0,99**	0,01	0	0	0
S8	0,38	**0,32**	0,09	0,08	0,14	S38	**0,98**	0,02	0	0	0
S9	**0,72**	0	0	0,11	0,16	S39	**0,81**	0,05	0,08	0,07	0
S10	0,49	0,10	0,14	**0,17**	0,10	S40	**0,85**	0,02	0	0,07	0,06
S11	0,01	0,05	0,05	**0,28**	0,61	S41	**0,16**	0,70	0,05	0,06	0,03
S12	**0,53**	0,08	0,01	0,14	0,24	S42	**0,83**	0,08	0,05	0,03	0,01
S13	0,10	0,07	**0,74**	0	0,09	S43	**0,90**	0	0,07	0,02	0,01
S14	0,18	**0,59**	0,03	0,05	0,15	S44	**0,93**	0,01	0,01	0	0,05
S15	0,14	0,01	**0,47**	0,35	0,03	S45	**0,92**	0,02	0,05	0	0,01
S16	**0,47**	0,15	0,18	0,17	0,03	S46	**0,91**	0,02	0,03	0,02	0,01
S17	**0,81**	0,09	0,06	0,01	0,03	S47	**0,92**	0	0,07	0,01	0
S18	**0,38**	0,43	0,14	0,02	0,03	S48	**0,95**	0	0,01	0	0,03
S19	**0,64**	0,07	0,17	0,11	0,01	S49	**0,93**	0,03	0	0,01	0,02
S20	**0,72**	0,20	0,01	0,02	0,05	S50	**0,70**	0,13	0,15	0	0,02
S21	**0,80**	0,02	0,07	0,01	0,10	S51	**0,77**	0	0,19	0,03	0
S22	**0,89**	0,01	0,02	0,06	0,02	S52	**0,99**	0	0,01	0	0
S23	**0,77**	0,02	0,05	0,10	0,06	S53	**0,85**	0,01	0,09	0,01	0,03
S24	**0,85**	0,03	0	0,09	0,02	S54	**0,45**	0,11	0,10	0,07	0,26
S25	**0,94**	0	0,02	0,01	0,02	S55	**0,38**	0,25	0,09	0,10	0,18
S26	**0,97**	0	0	0	0,03	S56	**0,86**	0,01	0,09	0,01	0,02
S27	**0,78**	0	0,03	0,14	0,05	S57	**0,39**	0,14	0,15	0,07	0,26
S28	**0,82**	0,01	0,10	0,02	0,05	S58	**0,67**	0,15	0,09	0,08	0,01
S29	**0,92**	0	0,01	0,01	0,06	S59	**0,52**	0,01	0,15	0,30	0,02
S30	**0,95**	0	0,05	0	0	S60	**0,60**	0,05	0,06	0,15	0,15

4.3 Kvalitativ vurdering av oppgavene

4.3.1 Kategorisering av oppgaver

Oppgavene ble kategorisert utfra tema/læringsmål, og om den tester kompetanse på lavere, eller høyt kognitivt nivå (Tabell 4.3). I tillegg er det oppgitt hvilke oppgaver som har stimulus i form av figur eller bilde. Antall oppgaver med stimulus er 10, hvorav 8 er bilder av arter der respondenten skal gjenkjenne arten. Seksten oppgaver ble kategorisert som høyt kognitivt nivå og 44 oppgaver ble kategorisert som lavt kognitivt nivå. Gjennomsnittlig p-verdi for oppgaver i kategorien *høyt* kognitivt nivå er 0,61. Gjennomsnittlig p-verdi for oppgaver i kategorien *lavt* kognitivt nivå er 0,80. Det vil si at oppgaver i kategorien *høyt* kognitivt nivå har høyere vanskelighetsgrad.

Tabell 4.3: Kategorisering av oppgaver utfra tema/læringsmål og hvorvidt de tester kunnskap, anvendelse eller vurdering. Tabellen viser hvilke oppgaver som har stimulus i form av figur eller bilde.

Oppgave	Tema / Læringsmål	Kognitivt nivå	Stimulus
S1	Metoder i økologien	Høyt	
S2	Interaksjoner mellom arter	Lavt	
S3	Artskunnskap / Livshistorietrekk	Lavt	
S4	Vegetasjonssoner	Lavt	
S5	Naturtype	Lavt	
S6	Biodiversitet	Lavt	
S7	Marine soner	Høyt	
S8	Økosystem	Høyt	
S9	Økosystem	Høyt	
S10	Artskunnskap / Begrepskunnskap	Lavt	
S11	Artenes utbredelse / Klima	Høyt	
S12	Senglacial flora	Høyt	
S13	Biogeokjemisk syklus	Høyt	
S14	Flathogst	Høyt	
S15	Økosystem	Høyt	
S16	Jern i havet	Høyt	
S17	Artenes utbredelse	Lavt	
S18	Biom / Klima	Høyt	Stimulus
S19	Artenes utbredelse	Lavt	
S20	Artskunnskap	Lavt	
S21	Artskunnskap	Lavt	
S22	Biodiversitet	Lavt	
S23	Artskunnskap / Artenes utbredelse	Lavt	

S24	Fundamental nisje	Lavt	
S25	Interaksjoner mellom arter	Lavt	
S26	Vegetasjonssoner / naturtyper	Lavt	
S27	Artskunnskap/ Vegetasjonssoner / naturtyper	Lavt	
S28	Intermediate disturbance	Lavt	
S29	Benthos	Lavt	
S30	Mutualisme / Interaksjoner	Lavt	
S31	Uniform fordeling / Interaksjoner	Lavt	
S32	Biogeografi	Lavt	
S33	Biogeografi	Høyt	Stimulus
S34	Suksesjon	Lavt	
S35	Artskunnskap / Dyregrupper	Lavt	
S36	Biodiversitet	Lavt	
S37	Samfunnsøkologi	Lavt	
S38	Miljø /økosystem	Lavt	
S39	Artskunnskap / Artenes utbredelse	Høyt	
S40	Suksesjon	Lavt	
S41	Artenes utbredelse	Høyt	
S42	Nøkkelart	Lavt	
S43	Suksesjon	Lavt	
S44	Biom / Klima	Lavt	
S45	Populasjonsøkologi	Lavt	
S46	Populasjonsøkologi	Høyt	
S47	Biodiversitet	Høyt	
S48	Populasjonsøkologi	Lavt	
S49	Artskunnskap	Lavt	Stimulus
S50	Artskunnskap	Lavt	Stimulus
S51	Artskunnskap	Lavt	Stimulus
S52	Artskunnskap	Lavt	Stimulus
S53	Artskunnskap	Lavt	Stimulus
S54	Artskunnskap	Lavt	Stimulus
S55	Artskunnskap	Lavt	Stimulus
S56	Artskunnskap	Lavt	Stimulus
S57	Artskunnskap	Lavt	Stimulus
S58	Artskunnskap	Lavt	Stimulus
S59	Artskunnskap	Lavt	Stimulus
S60	Artskunnskap	Lavt	Stimulus

4.3.2 Kvalitativ vurdering av oppgaver basert på vanskelighetsgrad og point-biserialkorrelasjon

I avsnittene nedenfor vurderes noen utvalgte oppgaver som hadde enten høye eller lave verdier for vanskelighetsgrad og point-biserialkorrelasjon med utgangspunkt i 2x2-matrisen i Figur 3.1. Nøkkelen er markert med fet skrift og svaralternativenes svarprosent er gitt i parentes.

Vanskelige oppgaver som diskriminerer dårlig

Oppgave 41 er oppgaven som færrest av respondentene fikk til, og har derfor den laveste p-verdien (0,16) (Boks 4.1). Riktig svaralternativ er A, men veldig mange (70,5 %) respondenter svarer svaralternativ B. En mulig forklaring kan være negasjonen *ikke* i stammen av oppgaven. Negasjoner bør helst unngås da respondenter kan overse disse. Er det nødvendig å bruke negasjoner bør disse fremheves, for eksempel ved bruk av kursiv skrift (Haladyna *et al.,* 2002; Sirnes, 2005). En kan også spekulere i om de som eventuelt har gjettet på oppgaven har valgt svaralternativ B siden det er det lengste svaralternativet.

Boks 4.1: Oppgave 41. P-verdi = 0,16. Point-biserialkorrelasjon = 0,20.

S41: Kva for eit alternativ nedanfor kan ikkje forklara samanhengen mellom artstal og arealstorleik?

A. **Større område er gunstigare («favourable») (15,9 %)**
B. Større område gjev fleire artar på grunn av større innsamlingsinnsats (70,5 %)
C. Større område gjev meir heterogenitet (4,5 %)
D. Større område husar fleire individ (5,7 %)
E. Øybiogeografiteorien (3,4 %)

Totalt to oppgaver hadde p-verdi som var lavere enn 0,20 og dermed kategorisert som veldig vanskelige oppgaver. Lav p-verdi kan bety at studentene ikke har forstått konseptet, men det kan også indikere at det er et problem med selve oppgaven.

Oppgave 10 er er den andre oppgaven som er kategorisert som veldig vanskelig (Boks 4.2). Point-biserialkorrelasjonen er blant de laveste (0,16), og oppgaven diskriminerer dermed dårlig mellom dyktige og mindre dyktige respondenter.

Boks 4.2: Oppgave 10. P-verdi = 0,17. Point-biserialkorrelasjon = 0,16.

S10: Kva for ein av artane nedanfor er eit mesokratisk tre?

A. Bjørk (48,9 %)
B. Osp (10,2 %)
C. Rogn (13,6 %)
D. Alm (17,1 %)
E. Selje (10,2 %)

Vanskelige oppgaver som diskriminerer godt

Det er få eksempler på oppgaver i eksamenssettet som kan plasseres i denne kategorien. Oppgave 54 er en av oppgavene med best diskrimineringsevne (point-biserialkorrelasjon = 0,52) (Boks 4.3). Samtidig er p-verdien forholdsvis lav (p-verdi = 0,45) sammenlignet med andre oppgaver som diskriminerer bra. Ved å endre på en distraktør kan muligens oppgaven gjøres vanskeligere. Blæretang er en art som ligner på spiraltang, men for en som kjenner til artene er det enkelt å skille mellom de to. Det kan derfor være en idé å erstatte distraktøren som færrest respondenter har valgt, D. Sagtang, med Blæretang.

Boks 4.3: Oppgave 54. P-verdi = 0,45. Point-biserialkorrelasjon = 0,52.

S54: Hvilken art er avbildet ovenfor?

A. Spiraltang (Fucus spiralis) (45,4 %)

B. Skolmetang (Halidrys siliquosa) (11,4 %)

C. Japansk drivtang (Sargassum muticum) (10,2 %)

D. Sagtang (Fucus serratus) (6,8 %)

E. Grisetang (Ascophyllum nodosum) (26,1 %)

Enkle oppgaver som diskriminerer dårlig

21 av oppgavene hadde p-verdi som var over 0,90 og dermed kategorisert som veldig enkle oppgaver. Disse oppgavene ville i mange tilfeller blitt anbefalt å fjerne. Alternativet er å vurdere om oppgavene kan gjøres vanskeligere. En måte å gjøre det på er å finne bedre distraktører. Emneansvarlig har fortalt at mange av flervalgsoppgavene blir brukt om igjen fra år til år, noen ganger i revidert form. Dette kan være en mulig årsak til at en tredjedel av oppgavene er besvart riktig av så stor andel studenter. Det er ikke utenkelig at noen studenter har sett tidligere eksamenssett.

Oppgave 37 var en av oppgavene flest fikk til, og har en svært høy p-verdi (0,99) (Boks 4.4). Point-biserialkorrelasjonen er 0, noe som vil si at oppgaven ikke diskriminerer mellom dyktige og mindre dyktige respondenter. Dette illustrerer at point-biserialkorrelasjonen er følsom for svært høye eller lave p-verdier. Med denne oppgaven er det altså vanskelighetsgraden som er det største problemet.

Boks 4.4: Oppgave 37. P-verdi = 0,99. Point biserialkorrelasjon = 0.

S37: Kva for ei utsegn om samfunn er riktig?

A. **Eit samfunn er ei samling av populasjonar av ulike artar som lever nært nok kvarandre for potensiell interaksjon (98,9 %)**
B. Eit samfunn fungerer som ein super-organisme der artane eksisterer til det beste for samfunnet (1,1 %)
C. Eit samfunn er ei samling av populasjonar av same art (0 %)
D. Samfunn består av nokre få arter med høg tettleik av individ, medan dei andre artene har låg tettleik (0 %)
E. Dei fleste artene i eit samfunn er dominante (0 %)

Enkle oppgaver som diskriminerer godt

Det er flere oppgaver som kan plasseres i denne kategorien. Oppgave 40 er blant oppgavene som veldig mange fikk til (p-verdi = 0,85) som samtidig diskriminerer noenlunde godt (point-biserialkorrelasjon = 0,51) (Boks 4.5). Oppgaven tester kunnskap om grunnleggende prinsipper i økologi. Det er flere eksempler på oppgaver som tester evnen til tenkning på lavere kognitive nivåer, har høy p-verdi og samtidig diskriminerer bra. Disse oppgavene er imidlertid ikke oppgaver som ber om detaljkunnskap, men kunnskap om sentrale begreper innenfor emnet.

Boks 4.5: Oppgave 40. P-verdi = 0,85. Point-biserialkorrelasjon = 0,51.

S40: Kva er skilnaden på primær og sekundær suksesjon?

A. **Ved sekundærsuksesjon vil noko av jordsmonnet vere intakt, men ikkje i ein primærsuksesjon**
B. Det er fleire artar involvert i ein sekundærsuksesjon
C. Sekundærsuksesjon gjeld berre dyra, medan primærsuksesjon gjeld for plantene
D. Sekundærsuksesjon startar med ei forstyrring. Det gjer ikkje primærsuksesjon
E. Primærsuksesjon er det same som evolusjon, medan sekundærsuksesjon er utvikling i økosystema

4.3.3 Revidering av oppgaver

Tabell 4.3 indikerer at det stor overvekt av oppgaver i kategorien kunnskap. Slik vil det ofte være. Desto høyere kunnskapsnivå, desto vanskeligere blir det å lage gode flervalgsoppgaver. Jeg vil her komme med noen forslag til revidering av oppgaver som skiller seg ut i eksamenssettet.

I oppgave 56 skal respondenten gjenkjenne arten *Lyr* basert på et bilde. *Sei* er en annen torskefisk som ligner på lyr, men er ikke et av svaralternativene. Det anbefales derfor å erstatte en av distraktørene som få respondenter har valgt med distraktøren *Sei*.

For noen oppgaver kan det være spesielt krevende å finne egnede distraktører. Oppgave 1 er et eksempel på dette (Boks 4.6). Ikke alle distraktørene er plausible. En student som ikke husker formelen for fangst-gjenfangst, vil likevel kunne utelukke noen svaralternativer, som for eksempel 167. Det er vanskelig å se for seg at det er det riktige svaret når verdiene som er opplyst er 100, 50 og 20.

100/20 * 50 = 250. (riktig formel)
100/50 * 20 = 40 (ikke en god distraktør siden de fleste vil skjønne at populasjonen ikke kan bestå av færre individer enn det antallet individer som ble merket ved første fangst.)

100*20 – 50 = 1950 (bedre distraktør)

På en side så er kanskje ikke hensikten å teste om studenten husker formelen, men om de kan finne frem til et riktig estimat. Denne oppgaven kan eventuelt gis som en ”fyll inn riktig svar”-type oppgave. Slike oppgaver kan også rettes automatisk i digitale vurderingsplattformer som for eksempel Inspera Assessment.

Boks 4.6: Oppgave 1. P-verdi = 0,94. Point-biserialkorrelasjon = 0,31.

S1: For å finna ut kor mange individ det er i ein populasjon er det vanleg å bruka fangst-gjenfangst- metoden. Kor mange individ vil vi estimera at det er i ein populasjon med følgjande resultat av fangst-gjenfangst? Første fangst: 100 individ. Andre fangst: 50 individ, og av desse var 20 individ òg fanga første gongen.

A. 225 (4,6 %)
B. 120 (0 %)
C. 167 (1,1 %)
D. 200 (0 %)
E. 250 (94,3 %)

4.4 Reliabilitet

Testens reliabilitet ble beregnet utfra metodene beskrevet i delkapittel 2.4.2. Reliabilitetsestimeringen basert på Chronbachs koeffisient alfa er 0,83 og regnes som høy (Tabell 4.4). Reliabiliteten basert på Kuder-Richardson-formelen er 0,75, som også er relativt høy. Korrelasjonskoeffisienten for hele testen basert på Spearman-Brown-formelen er 0,86.

Tabell 4.4: Mål på reliabilitet i prøveutvalget.

Chronbachs koeffisient alfa	0,83
Kuder-Richardson formel 21	0,75
Spearman-Brown formel	0,86

Kapittel 5 – Diskusjon

Hensikten med denne masteroppgaven er å gjøre en vurdering av flervalgstest som eksamensform ved å analysere et eksamenssett i biologi bestående utelukkende av flervalgsoppgaver. Kort oppsummert viser resultatene at eksamenssettet har både styrker og svakheter. På den ene siden er det en overvekt av oppgaver med lav vanskelighetsgrad og oppgaver som diskriminerer dårlig mellom dyktige og mindre dyktige respondenter. I tillegg viser analyse av distraktørene at de fleste oppgavene har én eller flere ineffektive distraktører. På den annen side viser statistiske tester at flervalgstesten har høy reliabilitet. Testen dekker de fleste læringsmålene godt, men noen er underrepresenterte.

Dette kapittelet inneholder diskusjon av metode, analyse, eksamenssettet som en helhet og flervalgstest som eksamensform.

5.1 Diskusjon av metode

Til tross for at eksamen er gitt digitalt via Inspera Assessment, så er det en tidkrevende og omfattende prosess å analysere oppgavene og resultatene fra eksamen. Foreløpig har ikke Inspera Assessment et innebygget analyseverktøy som gjør det enklere for undervisere og andre som konstruerer flervalgstester å vurdere kvaliteten på tester. Forhåpentligvis vil denne masteroppgaven bidra til å belyse hvor nyttig og viktig en slik analyse er, og forhåpentligvis vil slike analyseverktøy være inkludert i fremtidige digitale vurderingsplattformer.

Det er viktig å ikke utelukkende vurdere oppgaver kun basert på kvantitativ oppgaveanalyse, men også gjøre en kvalitativ vurdering av oppgavenes kvalitet og egnethet. Man kan ikke stole blindt på point-biserialkorrelasjonen som mål på diskrimineringsevnen til enkeltoppgaver. Er andel respondenter som besvarte riktig enten veldig høy eller veldig lav, så vil point-biserialkorrelasjonen være lett bli påvirket. Implikasjonen er at for oppgaver med veldig høy eller lav p-verdi, så er point-biserialkorrelasjonen ikke et rimelig mål på diskrimineringsevnen til oppgaven. Hovedproblemet da er at oppgaven enten er for lett, eller for vanskelig. En svakhet ved å bruke klassisk testteori for å beregne oppgavers vanskelighetsgrad er at den avhenger veldig av utvalget (Cohen *et al.,* 2011, side 480).

Kategorisering av oppgaver etter hvilket kognitivt nivå som testes kan være en utfordring. Flervalgsoppgaver inkluderer ikke nødvendigvis nøkkelverb som gjør det enklere å kategorisere oppgavene (Tabell 2.1). Blooms taksonomi består av seks nivåer, så for å gjøre det enklere å kategorisere oppgavene ble nivåene *lavt* og *høyt* valgt. Likevel kan det være vanskelig å plassere noen oppgaver. En oppgave som kan se ut som den tester evnen til å *anvende* kunnskap, kan i realiteten teste på et lavere nivå dersom eksempelet har blitt brukt i undervisningen. Derfor er det mulig at andre ville kategorisert oppgavene annerledes. Det er ikke et stort problem i dette tilfellet ettersom hensikten med kategoriseringen etter kognitivt nivå var å vurdere om det er en overvekt av oppgaver som tester på lavere kognitivt nivå, slik det ofte hevdes at flervalgstester gjør.

5.2 Diskusjon av analyser

Fordelingen av totalskår (Figur 4.1) for eksamenssettet er venstreskjev med kun tre studenter som fikk stryk. En noe venstreskjev fordeling er å forvente siden respondentene kan gjette på oppgavene. I teorien vil en respondent som gjetter det samme svaralternativet på alle oppgavene få riktig på 12/60 oppgaven. Likevel skal det mye til for å gjette seg frem til en kunstig høy skår (Haladyna, 1994). Ni studenter var ett poeng fra å oppnå karakteren C (Figur 4.1), noe som viser at terskelverdier for de ulike bokstavkarakterene har stor betydning. Det stiller krav til kvaliteten på oppgavene og det er ikke rom for oppgaver med store svakheter.

Oppgavene hadde en lav gjennomsnittlig vanskelighetsgrad (p-verdi = 0,75) (Tabell 7.2). I følge Sirnes (2005, side 68) bør gjennomsnittlig vanskelighetsgrad være rundt 0,50. Gjennomsnittlig point-biserialkorrelasjon er 0,30 (Tabell 7.2). Det indikerer at oppgavene diskriminerer middels godt.

Eksamenssettet har en overvekt av oppgaver som tester på lavt kognitivt nivå (Tabell 4.3) og gjennomsnittlig vanskelighetsgrad er lav. Oppgavene som ble kategorisert til å teste høyere kognitivt nivå hadde høyere vanskelighetsgrad enn oppgavene som testet lavere kognitive nivå.

5.2.1 Revidering av oppgaver

Oppgaver med enten veldig høy eller veldig lav vanskelighetsgrad bør vurderes nøye. Oppgave 10 (Boks 4.2) er et eksempel på en oppgave som er kategorisert som veldig vanskelig, samtidig som den diskriminerer dårlig. Årsaken til det kan være at både dyktige og mindre dyktige studenter ikke vet hva begrepet *mesokratisk* betyr eller misforstår begrepet. Ordet *mesokratisk* refererer til en fase i den Nordeuropeiske interglasiale syklusen, men i følge leksikonet Merriam-Websteren betyr ordet *mesokratisk* en stein som har omtrent like mengder mørke og lyse mineraler (Mesocratic, 2016). Det kan være en forklaring på at forholdsvis mange har valgt svaralternativet *Bjørk* (48,9 %). Mitt inntrykk er at dette ikke er et veldig sentralt begrep i emnet og at oppgaven fokuserer på detaljkunnskap. Jeg vil derfor anbefale å fjerne oppgaven. Dersom det virkelig er et viktig begrep, så bør en vurdere hvordan undervisningen kan tilpasses slik at flere studenter forstår begrepet.

Boks 4.1 illustrerer hvorfor en bør unngå negasjoner i stammen av oppgaven. Oppgave 12 har en bedre formulering der det kommer tydeligere frem at en er ute etter svaralternativet som er feil (Boks 5.1).

Boks 5.1: Oppgave S12.

S12: Kva for ei utsegn om den seinglaciale floraen er feil?

A. **Den seinglaciale floraen starta ein sekundær suksesjon (53,41%)**
B. Dei først etablerte elementa var strandplantene, fjellplantene og steppe-plantene. (7,9 %)
C. Yngre Dryas kuldeperiode forårsaka ei sørleg forskyving av grensene til vegetasjonen (1,1 %)
D. Då isen smelta, følgde fjellplantene etter den smeltande breen austover (13,6 %)
E. Nokre seinglaciale strandplanter fekk ei ny nisje som ugrasplanter då mennesket byrja å rydda skog fleire tusen år seinare. (23,9 %)

5.2.2 Reliabilitet og validitet

Testens reliabilitet

Indre konsistens-metodene estimerte reliabiliteten til testen til å være god. Hovedårsaken til dette er et høyt antall oppgaver (60) (Sirnes, 2005, side 83).

Testens innholdsvaliditet

Basert på kategoriseringen av oppgavene utfra tema eller læringsmål er det interessant å se vektleggingen av de ulike læringsmålene. Innholdsvaliditet sier noe om i hvilken grad testens innhold samsvarer med det den er ment å teste, som i dette tilfellet er oppgitt læringsutbytte for emnet (Sirnes, 2005, side 81). Det er viktig å påpeke at emneansvarlig kan tillegge ulik vekt til de ulike læringsmålene. Noen læringsmål favner bredt og flere oppgaver vil dermed kunne kobles til de læringsmålene. Hovedpoenget er å undersøke om noen læringsmål er underrepresenterte.

Det er spesielt to læringsmål som i liten grad blir vurdert gjennom flervalgstesten.
Et av disse læringsmålene er:

> ”Få en forståelse for metodene som brukes i økologien.”

Etter min vurdering er oppgave 1 (Boks 4.6) den eneste oppgaven der dette læringsmålet står sentralt. Metoder som brukes i økologien er sentralt for emnet og studentene lærer trolig mye om dette av å delta på de tre ukene med feltkurs som inngår i emnet. For at det skal være en sammenheng mellom oppgitte læringsmål, undervisning og vurdering, så bør læringsmålet få større plass i den summative vurderingen. Forslag til hvordan dette kan gjøres blir beskrevet i delkapittel 5.3.

Det andre underrepresenterte læringsmålet er:

> ”Kunne identifisere et gitt sett med arter av planter, dyr og sopp, og være i stand til å bruke litteratur for å identifisere andre arter i Vest-Norge.”

Studentene får vist i 8 oppgaver om de kjenner igjen en art basert på et bilde, men de får ikke vist om de kan bruke litteratur for å identifisere arter. Et alternativ kan være å lage oppgaver med bilde av en art det er vanskelig å identifisere basert på bilde i tillegg til en stimulus i form av tekstutdrag fra litteratur. Viser det seg å være vanskelig å teste dette

læringsmålet bør man vurdere om læringsmålet skal omformuleres eller inkluderes i den summative vurderingen på en annen måte.

Kategoriseringen av oppgavene utfra hvilket nivå av kognitiv tenkning de krever viser at det er en betydelig overvekt av oppgaver på lavt kognitivt nivå som hovedsakelig tester evnen til å gjengi kunnskap i den form det er lært (Tabell 4.3). Det skal sies et kategoriseringen er basert på en subjektiv vurdering noe som svekker dens reliabilitet, men tendensen er likevel klar. Resultatet er ikke uventet da det er en utfordring å lage flervalgsoppgaver som tester tenkning på høyere kognitivt nivå (Haladyna, 1994; Sirnes, 2005; Woolfolk, 2004).

5.3 Anbefalinger for utbedring av eksamenssettet

For å oppsummere ser det ut til at testen har både styrker og svakheter. Et høyt antall oppgaver er vesentlig for reliabiliteten og minimerer effekten av gjetting. Sjansen for at en student har fått en kunstig høy skår på grunn av gjetting når testen inneholder 60 flervalgsoppgaver er svært liten (Haladyna, 1994). Det anbefales derfor å fortsette med 60 flervalgsoppgaver til neste eksamen. Antallet oppgaver kan eventuelt økes dersom erfaringen tilsier at de fleste studentene leverer i god tid. Slik testen er nå har studentene 3 minutter per oppgave. Ettersom det kun er én oppgave som krever regning, er dette rikelig med tid. Haladyna (1994) anslår at 1 minutt per oppgave er tilstrekkelig.

En flervalgstest som gis digitalt kan gjerne bestå av ulike varianter av flervalgsoppgaver siden objektive vurderingsformer kan skåres automatisk i digitale vurderingsplattformer. Oppgaver med stimulus har stort potensiale. I følge Haladyna *et al.* (2002) er en stimulus et godt utgangspunkt for å lage oppgaver som tester høyere kognitive nivå. De to oppgavene med stimulus som ikke testet artskunnskap ble kategorisert som høyt kognitivt nivå (Tabell 4.3). Det anbefales derfor å inkludere flere oppgaver med stimulus. Siden formatet er egnet til å lage oppgaver som tester evne til problemløsing, så er det gode muligheter for å lage oppgaver som tester de underrepresenterte læringsmålene. Dersom det viser seg å være utfordrende å teste studentenes kompetanse i noen av læringsmålene basert på en flervalgstest, kan en vurdere mappevurdering. NOU (Kunnskapsdepartementet, 2000, Kapittel 13) nevner mappevurdering som en vurderingsform med flere styrker. En

mappevurdering i emnet Organismebiologi 2 kan for eksempel bestå av feltkursrapportene og en avsluttende flervalgstest. Den avsluttende testen kan alternativt inkludere noen få drøftingsoppgaver som enklere kan formuleres slik at de tester kompetanse på høyere kognitivt nivå (Haladyna, 1994, side 28). Ved å legge til flere oppgaver i eksamenssettet som tester høyere kognitivt nivå kan også vanskelighetsgraden økes.

Flervalgsoppgaver med fire eller fem svaralternativer er det vanligste, men Haladyna *et al.*, (1993) argumenterer for at tre svaralternativer kan være tilstrekkelig. Dette eksamenssettet har mange distraktører som er ineffektive. Jeg vil derfor anbefale å starte med å redusere antall svaralternativer fra 5 til 4. Det gjøres oppmerksom på at terskelverdiene for de ulike bokstavkarakterene bør økes om antall svaralternativer reduseres.

5.4 Avsluttende vurdering av flervalgstest som eksamensform

Hensikten med denne masteroppgaven har vært å gjøre en vurdering av en flervalgseksamen ved å analysere oppgaver og hvordan studentene har svart. Resultatene viser at eksamenssettet har flere svakheter. Dette betyr ikke nødvendigvis at en må gå bort fra flervalgstest som eksamensform. Revidering av oppgaver er essensielt for å øke kvaliteten på eksamen.

Mange er kritiske til flervalgstest som eksamensform, blant annet Raaheim (2016, side 120). Det er enklere å se fordelene ved å bruke flervalgstester til formativ vurdering der elever/studenter kan få rask tilbakemelding. Et poeng er at flervalgstester brukt til summativ vurdering også kan ha et formativt aspekt dersom undervisere på bakgrunn av resultater av flervalgstester tilpasser og utvikler undervisningen.

Detaljfokuserte tester kan påvirke læringsstrategien til studentene i en retning der det fokusert på overflatisk læring (Raaheim, 2016). Gjenbruk av oppgaver kan være problematisk av denne grunn, for det er ikke ønskelig at studentene velger en læringsstrategi basert på pugging. Det er derfor en fordel med en stor oppgavebank med oppgaver som har blitt benyttet tidligere. I tillegg kan nye flervalgsoppgaver lages i løpet av semesteret. En idé kan være å skrive ned spørsmål studentene stiller i undervisningen.

Noen læringsmål kan være utfordrende å vurdere med en flervalgstest. Oppgaver med stimulus har stort potensiale og kan gjøre det enklere å lage oppgaver som tester på høyere kognitivt nivå (Haladyna *et al.,* 2002).

Å innføre flervalgstest som eksamensform er mer omfattende enn mange er klar over. En stor fordel med vurderingsformen er at gode oppgaver kan brukes om igjen. Med tiden vil man kunne opparbeide en stor oppgavebank som gjør det enklere å konstruere gode flervalgstester. Vurderingsformen gjør det mulig å dekke store deler av pensum, hvilket øker reliabiliteten til testresultatene (Haladyna, 1994, side 27; Sirnes, 2005, side 10).

Kapittel 6 – Veien videre

Veien videre innebærer en revidering av oppgavene i eksamenssettet basert på anbefalinger gitt i denne masteroppgaven. Det kan bli en tidkrevende prosess for emneansvarlige, men en oppgavebank bestående av gode flervalgsoppgaver har stor verdi. Når en får resultatene fra neste eksamen vil det være interessant å vurdere om dette eksamenssettet har blitt forbedret. Eksamenssettene må sammenlignes med varsomhet ettersom resultater fra metoder brukt i klassisk testteori vil avhenge av oppgaveutvalget .

Jeg ser for meg flere interessante studier av andre eksamensformer som brukes på Institutt for biologi. Noen emner har såkalt mappevurdering som består av ulike vurderingsformer. Det vil være nyttig å sammenligne hvordan studentene presterer i de ulike vurderingsformene med tanke på å få en best mulig læring for studentene.

Referanser

Ali, S., & Ruit, K. (2015). The Impact of item flaws, testing at low cognitive level, and low distractor functioning on multiple-choice question quality. *Perspectives on Medical Education, 4*(5), 244-251.

Amin, Z., Seng, C., & Eng, K. (2006). Extended Matching Items (EMI). In *Practical Guide To Medical Student Assessment* (pp. 43-45). World Scientific Publishing Pte.

bioCEED. (2014). Årsrapport 2014. Hentet fra: http://www.uib.no/sites/w3.uib.no/files/attachments/bioceed_application_text.pdf

bioCEED. (2015). Årsrapport 2015. Hentet fra: https://scholar.uib.no/sites/default/files/bioceed/files/arsrapport2015_bioceed_final_0.pdf

Bloom, B. S. (1956). *Taxonomy of educational objectives; The classification of educational goals, by a committee of college and university examiners.* New York: McKay

Cohen, L., Manion, L. & Morrison, K. (2011). *Research Methods in Education (7th Edition)*. Florence: Taylor and Francis.

DiBattista, D., Sinnige-Egger, J. & Fortuna, G. (2014). The "None of the Above" Option in Multiple-Choice Testing: An Experimental Study. *The Journal of Experimental Education, 82*(2), 168-183.

Domyancich, J. M. (2014). The Development of Multiple-Choice Items Consistent with the AP Chemistry Curriculum Framework to More Accurately Assess Deeper Understanding. *Journal of Chemical Education, 91*(9), 1347-1351.

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

Haladyna, T. M. & Downing, Steven M. (1993). How many options is enough for a multiple-choice test item? (Validity Studies). *Educational and Psychological Measurement, 53*(4), 999. http://epm.sagepub.com.pva.uib.no/content/53/4/999.full.pdf+html

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education, 15*(3), 309-34.

Haladyna, T. (1994). *Developing and validating multiple-choice test items*. Mahwah, N.J: Lawrence Erlbaum Associates.

Hingorjo, M., & Jaleel, F. (2012). Analysis of one-best MCQs: The difficulty index, discrimination index and distractor efficiency. *JPMA. The Journal of the Pakistan Medical Association, 62*(2), 142-7.

Inspera Assessment. (2016). Complete e-Assessment platform. Hentet fra 02. mai fra http://www.inspera.com

Institutt for biologi ved Universitetet i Bergen. (2016a). *Innføring i evolusjon og økologi.* Hentet 02. mai fra: http://www.uib.no/emne/BIO100

Institutt for biologi ved Universitetet i Bergen. (2016b). *Organismebiologi 2.* Hentet 02. mai fra: http://www.uib.no/emne/BIO102

Kibble, Jonathan D., & Johnson, Teresa. (2011). Are Faculty Predictions or Item Taxonomies Useful for Estimating the Outcome of Multiple-Choice

Drasgow, F., Lissak, R., & Guion, Robert. (1983). Modified parallel analysis: A procedure for examining the latent dimensionality of dichotomously scored item responses. *Journal of Applied Psychology, 68*(3), 363-373.

Kubinger, K. D., Holocher-Ertl, S., Reif, M., Hohensinn, C., & Frebort, M. (2010). On minimizing guessing effects on multiple-choice items: Superiority of a two solu- tions and three distractors item format to a one solution and five distractors item format. International Journal of Selection and Assessment, 18(1), 111–115.

Kunnskapsdepartementet. (2006). *Ambisjoner for høyere utdanning.* Hentet fra: https://www.regjeringen.no/no/aktuelt/kontaktkonferanse-med-universiteter-og-h/id712110/)

Kunnskapsdepartementet. (2011). *Nasjonalt kvalifikasjonsrammeverk for livslang læring (NKR).* Hentet fra: http://www.nokut.no/Documents/NOKUT/Artikkelbibliotek/Norsk_utdanning/NKR/250414_Nasjonalt_kvalifikasjonsrammeverk_for_livslang_læring_NKR.pdf

Kunnskapsdepartementet. (2010). *Frihet med ansvar — Om høgre utdanning og forskning i Norge*. (NOU 2010: 14). Hentet fra https://www.regjeringen.no/no/dokumenter/nou-2000-14/id142780/?q=&ch=18

Lesage, E., Valcke, M., & Sabbe, E. (2013). Scoring methods for multiple choice assessment in higher education – Is it still a matter of number right scoring or negative marking?

Studies in Educational Evaluation, 39(3), 188-193.

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

El -Uri, F., & Malas, N. (2013). Analysis of use of a single best answer format in an undergraduate medical examination. *Qatar Medical Journal,* (1), Qatar Medical Journal, 2013(1).

Mentzer, T. (1982). Response Biases in Multiple-Choice Test Item Files. *Educational and Psychological Measurement, 42*(2), 437-448.

Mesocratic. (2016) I *Merriam-Webster.* Hentet 31.05.16 fra: http://www.merriam-webster.com/dictionary/mesocratic

R Development Core Team (2006). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http: //www.R-project.org/

Raaheim, A. (2016) *Eksamensrevolusjonen – Råd og tips om eksamen og alternative vurderingsformer*. Utgivelsessted: Gyldendal akademisk.

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

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. *Higher Education, 35*(4), 453-472.

Thayn, S., Sudweeks, R., Davies, R., Foster, D., Olsen, J., & Wiley, D. (2011). *An Evaluation of Multiple Choice Test Questions Deliberately Designed to Include Multiple Correct Answers,* N/a.

Universitetet i Bergen. (2016). Bachelorprogram i biologi. Hentet 30.05.2016 fra: http://www.uib.no/studieprogram/BAMN-BIO

Universitetet i Oslo. (2006). TIMSS – Frigitte oppgaver. Hentet fra: http://www.timss.no/timss05_frigitte.html

Universitetet i Oslo. (2016, 30.02.2016). Frigitte oppgaver. Hentet fra: http://www.uv.uio.no/ils/forskning/prosjekt-sider/pisa/frigitte-oppgaver/

Utdanningsdirektoratet. (18.08.2014). *Nasjonal satsing på vurdering for læring.* Hentet fra: http://www.udir.no/Vurdering-for-laring/Nasjonal-satsing1/Nasjonal-satsing-pa-Vurdering-for-laring/

Woolfolk, A., Pettersson, T., Nygård, M., & Karlsdóttir, R. (2004). *Pedagogisk psykologi*. Trondheim: Tapir akademisk forlag.

Wright, B. D. (1977). Misunderstanding the Rasch Model. *Journal of Educational Measurement,* 219-25.

Vedlegg

7.1 Læringsutbytte

Boks 7.1: Læringsmål det er forventet at studenten skal kunne etter fullført emne. Hentet fra emnets hjemmeside (Institutt for biologi ved Universitetet i Bergen, 2016b).

Læringsutbytte

Etter fullført emne skal studenten

- Ha en grunnleggende forståelse for hva, populasjonsøkologi, samfunnsøkologi, og økosystemer er.
- Kunne gjøre rede for de forskjellige biomene og biomenes utbredelse i verden
- Ha kunnskap om de viktigste faktorene som påvirker artenes utbredelse globalt og lokalt.
- Forstå hvordan arter interagerer og påvirker hverandre positivt og negativt.
- Kunne beskrive biodiversiteten i et område og diskutere hvilke faktorer som påvirker biodiversiteten.
- Beskrive og forstå dynamiske prosesser både for populasjoner og samfunn både på kortere og lengre tidsskalaer.
- Kunne forklare enkle biogeografiske prinsipper, som for eksempel likevektsmodellen for øybiogeografi.
- Forstå hvordan livshistorietrekk påvirker økologien til artene.
- Gjøre rede for de viktigste truslene mot det biologiske mangfoldet i dag, i Norden spesielt og i verden generelt, og hvilke virkemidler man bruker i bevaringen av det biologiske mangfoldet.
- Kunne identifisere et gitt sett med arter av planter, dyr og sopp, og være i stand til å bruke litteratur for å identifisere andre arter i Vest-Norge.
- Kjenne de viktigste miljøfaktorene for utbredelsen av arter i Norden
- Forstå viktigheten av interaksjoner mellom prokaryote organismer og Eukaryote planter og dyr samt betydning av og funksjon til prokaryote organismer i biokjemiske sykluser.
- Få en forståelse for metodene som brukes i økologien.

7.2 Vanskelighetsgrad og point-biserialkorrelasjon

Tabell 7.1: Vanskelighetsgrad basert på p-verdi og point-biserialkorrelasjon for de 60 oppgavene i eksamenssettet. Gjennomsnittlig vanskelighetsgrad = 0,75. Gjennomsnittlig point-biserialkorrelasjon = 0,30.

Oppgave	Vanskelighets-grad (p-verdi)	Point-biserialkorrelasjon	Oppgave	Vanskelighets-grad (p-verdi)	Point-biserialkorrelasjon
S1	0,94	0,31	S31	0,95	0,24
S2	0,98	0,06	S32	0,98	0,27
S3	0,92	0,11	S33	0,66	0,37
S4	0,59	0,20	S34	0,91	0,40
S5	0,90	0,18	S35	0,69	0,33
S6	0,78	0,35	S36	0,80	0,46
S7	0,83	0,28	S37	0,99	0,00
S8	0,32	0,28	S38	0,98	0,04
S9	0,72	0,45	S39	0,81	0,28
S10	0,17	0,16	S40	0,85	0,51
S11	0,28	0,16	S41	0,16	0,20
S12	0,53	0,41	S42	0,83	0,34
S13	0,74	0,19	S43	0,90	0,52
S14	0,59	0,32	S44	0,93	0,07
S15	0,47	0,33	S45	0,92	0,32
S16	0,47	0,35	S46	0,91	0,28
S17	0,81	0,32	S47	0,92	0,41
S18	0,38	0,33	S48	0,95	0,35
S19	0,64	0,32	S49	0,93	0,38
S20	0,72	0,14	S50	0,70	0,15
S21	0,80	0,32	S51	0,77	0,44
S22	0,89	0,47	S52	0,99	0,00
S23	0,77	0,39	S53	0,85	0,25
S24	0,85	0,52	S54	0,45	0,52
S25	0,94	0,43	S55	0,38	0,41
S26	0,97	0,12	S56	0,86	0,16
S27	0,78	0,28	S57	0,39	0,39
S28	0,82	0,50	S58	0,67	0,39

S29	0,92	0,41	S59	0,52	0,50
S30	0,95	0,23	S60	0,60	-0,07