KØBENHAVNS UNIVERSITET
INSTITUT FOR NATURFAGENES DIDAKTIK

ChatGPT som formativ feedbackgiver på gymnasieelevers design af biologieksperimenter

Anne Sofie Berendt
Speciale i biologi og didaktik

Fra Københavns Universitet & Aarhus Universitet

Vejledere:
Bob Evans

15. April 2024

IND's studenterserie nr. 122, 2024

INSTITUT FOR NATURFAGENES DIDAKTIK, www.ind.ku.dk
Alle publikationer fra IND er tilgængelige via hjemmesiden.

IND's studenterserie

90. Maria Anagnostou: Trigonometry in upper secondary school context: identities and functions (2020)
91. Henry James Evans: How Do Different Framings Of Climate Change Affect Pro-environmental Behaviour? (2020)
92. Mette Jensen: Study and Research Paths in Discrete Mathematics (2020)
93. Jesper Hansen: Effekten og brugen af narrative læringsspil og simuleringer i gymnasiet (2020)
94. Mie Haumann Petersen: Bilingual student performance in the context of probability and statistics teaching in Danish High schools (2020)
95. Caroline Woergaard Gram: "Super Yeast" - The motivational potential of an inquiry-based experimental exercise (2021)
96. Lone Brun Jakobsen: Kan man hjælpe elevers forståelse af naturvidenskab ved at lade dem formulere sig om et naturvidenskabeligt emne i et andet fag? (2021)
97. Maibritt Oksen og Morten Kjøller Hegelund: Styrkelse af motivation gennem Webinar og Green Screen (2021)
98. Søren Bystrup Jacobsen: Peer feedback: Fra modstand til mestring? (2021)
99. Bente Guldbrandsen: Er der nogen, som har spurgt en fysiklærer? (2021)
100. Iben Vernegren Christensen: Bingoplader i kemiundervisningen – en metode til styrkelse af den faglige samtale? (2021)
101. Claus Axel Frimann Kristinson Bang: Probability, Combinatorics, and Lesson Study in Danish High School (2021)
102. Derya Diana Cosan: A Diagnostic Test for Danish Middle School Arithmetics (2021)
103. Kasper Rytter Falster Dethlefsen: Formativt potentiale og udbytte i Structured Assessment Dialogue (2021)
104. Nicole Jonasen: A diagnostic study on functions (2021)
105. Trine Nørgaard Christensen: Organisatorisk læring på teknisk eux (2021)
106. Simon Funch: Åben Skole som indgang til tværfagligt samarbejde (2022)
107. Hans-Christian Borggreen Keller: Stem som interdisciplinær undervisningsform (2022)
108. Marie-Louise Krarup, Jakob Holm Jakobsen, Michelle Kyk & Malene Hermann Jensen: Implementering af STEM i grundskolen (2022)
109. Anja Rousing Lauridsen & Jonas Traczyk Jensen: Grundskoleelevers oplevelse af SSI-undervisning i en STEM-kontekst. (2022)
110. Aurora Olden Aglen: Danish upper secondary students' apprehensions of the equal sign (2023)
111. Metine Rahbek Tarp & Nicolaj Pape Frantzen: Machine Learning i gymnasiet (2023)
112. Jonas Uglebjerg: Independence in Secondary Probability and Statistics: Content Analysis and Task Design (2023)
113. Hans Lindebjerg Legard: Stopmotion som redskab for konceptuel læring. (2023)
114. Caroline Woergaard Gram & Dan Johan Kristensen: The ice algae Ancylonema as icebreakers: A case study on how the international Deep Purple Research Project can create meaningful outreach in Greenland. (2023)
115. Julie Sloth Bjerrum: 'KLIMA HISTORIER' The Art Of Imagining A Green Future. (2023)
116. Emilie Skaarup Bruhn: Muligheder og udfordringer ved STEM-undervisning (2024)
117. Milla Mandrup Fogt: Undersøgelsesbaseret undervisning i Pascals trekant (2024)
118. Mille Bødstrup: P-hacking (2024)
119. Nynne Milthers & Amanda Wedderkopp: Inquiry of the Past and Reflection on the Present: Teaching Rigour and Reasoning in Area Determination through Authentic Historical Sources (2024)
120. Pelle Bøgild: Med kroppen ind i fysikken (2024)
121. Anne Jensen & Charlotte Puge: Modellering som bro mellem teoretisk viden og praktisk laboratoriearbejde (2024)
122. **Anne Sofie Berendt: ChatGPT som formativ feedbackgiver på gymnasieelevers design af biologieksperimenter (2024)**

IND's studenterserie *omfatter kandidatspecialer, bachelorprojekter og masterafhandlinger skrevet ved eller i tilknytning til Institut for Naturfagenes Didaktik. Disse drejer sig ofte om uddannelsesfaglige problemstillinger, der har interesse også uden for universitetets mure. De publiceres derfor i elektronisk form, naturligvis under forudsætning af samtykke fra forfatterne. Det er tale om studenterarbejder, og ikke endelige forskningspublikationer.*
Se hele serien på: www.ind.ku.dk/publikationer/studenterserien/

Abstract

This study examines how formative feedback from ChatGPT influences students' design of biology experiments in two upper secondary education classes in the stx programme, a 1.g class with a basic course in natural science called naturvidenskabeligt grundforløb and a 3.g class with biology at B-level. Sixty students participated in the study, which is based on an inquiry-based approach, and the results are based on 19 anonymized student reports, 56 anonymous questionnaire responses, and 6 group interviews involving a total of 17 students. The students' assessment and implementation of feedback were examined by comparing ChatGPT's feedback with their experimental designs. Additionally, the experimental designs were quality-scored before and after feedback from ChatGPT based on a rubric. Furthermore, interview data and open-ended questions in the questionnaire survey were analyzed using thematic analysis. The results indicate that, on average, the groups implement 4.44 distinct alterations related to ChatGPT in their experimental designs, and the quality of the experimental designs increases significantly ($P<0.001$) for all groups collectively after feedback from ChatGPT. More than half of the students perceive ChatGPT's feedback as useful and trustworthy, but several believe that a critical approach should be maintained towards the feedback, which they feel does not adequately consider their practical reality. Thus, the study concludes that formative feedback from ChatGPT can assist upper secondary programme students in designing biology experiments to improve the quality of their experimental designs, although a more nuanced approach to ChatGPT's feedback is needed.

Tak

Først og fremmest vil jeg gerne takke Novo Nordisk Fonden for at tildele mig et stipendie, der har gjort det muligt at tage en master i scienceundervisning. De medvirkende elever fortjener desuden en særlig tak for at engagere sig i projektet og stille deres rapporter og erfaringer med ChatGPT til rådighed. Også en stor tak til Bob Evans for kyndig vejledning og for at udvise stor interesse for emnet.

Indhold

1. Indledning 6
   1.1 Problemformulering 7
2. Teoretisk grundlag 8
   2.1 IBSE 8
   2.2 Formativ evaluering 9
   2.3 Formativ feedback 10
   2.4 Socialkonstruktivisme 11
3. Litteraturreview 12
   3.1 Hvordan påvirker formativ feedback elever, der designer eksperimenter? 13
   3.2 Vil AI kunne fungere som feedbackgiver på elevers forsøgsdesign? 14
   3.3 Hvordan opfatter elever ChatGPT som redskab i undervisningen? 15
4. Undersøgelsesmetoder og design 16
   4.1 Undervisningens organisering og kontekst 16
   4.2 Udformning af elevvejledninger 19
   4.3 Empirisk design og data 22
      4.3.1 Metode til analyse af elevrapporter 24
      4.3.2 Metode til spørgeskema 26
      4.3.3 Metode til interview 27
      4.3.4 Begrænsninger og muligheder i det empiriske design 28
5. Resultater 30
   5.1 Kvantitativ analyse af elevrapporter 30
      5.1.1 Elevers bedømmelse af formativ feedback fra ChatGPT 30
      5.1.2 Implementering af feedback fra ChatGPT i forsøgsdesign 32
      5.1.3 Type af feedback fra ChatGPT, som implementeres i forsøgsdesign 33
      5.1.4 Kvalitet af forsøgsdesign før og efter feedback 35
   5.2 Kvantitativ analyse af spørgeskemaer 37
   5.3 Kvalitativ analyse af spørgeskemaer 40
      5.3.1 Tema 1: ChatGPT´s feedback kan forbedre forsøgsbeskrivelse og design 40
      5.3.2 Tema 2: ChatGPT´s feedback tager ikke nok højde for elevernes praktiske virkelighed 41
      5.3.3 Tema 3: ChatGPT´s feedback er brugbar for de fleste 42
      5.3.4 Tema 4: ChatGPT´s feedback kræver, at elever er udførlige og forholder sig kritisk 42
   5.4 Kvalitativ analyse af interviews 43

5.4.1 Tema 1: ChatGPT´s feedback kan skabe overblik og give nye ideer til forsøgsdesign 43
5.4.2 Tema 2: ChatGPT´s feedback kan forbedre forsøgsbeskrivelse og design 44
5.4.3 Tema 3: ChatGPT´s feedback har svært ved at ramme elevernes niveau 45
5.4.4 Tema 4: ChatGPT´s feedback tager ikke nok højde for elevernes praktiske virkelighed 45
5.4.5 Tema 5: ChatGPT´s feedback er målrettet det enkelte forsøg og troværdig 46
6. Diskussion 47
6.1 Elevers brug af ChatGPT´s feedback på forsøgsdesign 47
6.2 Feedback fra ChatGPT og kvaliteten af elevers forsøgsdesign 49
6.3 ChatGPT´s styrker og svagheder som formativ feedbackgiver på forsøgsdesign 51
7. Konklusion og perspektivering 56
8. Bibliografi 59
Bilag 63
Bilag 1: Temaopgave 2. Empiriske metoder/Temakursus, MiSU 63
Bilag 2: Elevvejledning 1.g 63
Bilag 3: Elevvejledning 3.g 63
Bilag 4: Eksempel på elevrapport fra 1.g og 3.g 63
Bilag 5: Interviewguide 63
Bilag 6: Rubrik til scoring af forsøgsdesign 63
Bilag 7: Spørgeskema uden svar 1.g 63
Bilag 8: Spørgeskema uden svar 3.g 63
Bilag 9: Pilotspørgeskema med svar 63
Bilag 10: Spørgeskema med svar 1.g 63
Bilag 11: Spørgeskema med svar 3.g 63
Bilag 12: Transskriberede interviews 63
Bilag 13: GAI-deklaration 63

1. Indledning

Den seneste udvikling inden for generativ AI[1], der er en gren inden for kunstig intelligens, har udløst diskussion om, hvordan fremtidens uddannelsessystem skal formes. Et af de store fokusområder i debatten har været brug af generativ AI til snyd med skriftlige opgaver (Rønberg, 2023), som lader til at være udbredt på landets gymnasier (Skovhus, Dupont, & Szocska, 2023). Samtidig har generativ AI potentialet til at fungere som et værktøj til at støtte læring, men der ligger en stor opgave i at håndtere kunstig intelligens i uddannelsessystemet på en pædagogisk forsvarlig, etisk og effektiv måde både i Danmark og internationalt (Vedersø, et al., 2023; Unesco, Fengchun, & Holmes, 2023). Det gør sig også gældende på gymnasieområdet, hvor der forestår et "massivt didaktisk udviklingsarbejde" (GL, 2023). En af konsekvenserne er, at undervisningen frem over bl.a. bør lægge vægt på mere undersøgende opgaver (Friisberg, 2023). Dette er interessant i forhold til undervisningen i de naturvidenskabelige fag, hvor det eksperimentelle arbejde står centralt, fordi læringsudbyttet ved en undersøgelsesbaseret tilgang er større end ved en kogebogspræget øvelse (Berg, Bergendahl, & Lundberg, 2003). Undersøgelsesbaseret naturvidenskabsundervisning indebærer ofte, at elever designer deres egne forsøg og har brug for løbende feedback (Madsen, Evans, & Bruun, 2020) f.eks. i form af formativ feedback på deres forsøgsdesign.

Selvom der er et stort læringspotentiale forbundet formativ feedback (Harlen, 2013a), viser forskning, at mangel på tid er en af de væsentligste årsager til, at undervisere underlader at bruge formativ evaluering i undervisningen (Dolin, Harlen, Black, & Tiberghien, 2018). Som tidligere gymnasielærer i biologi på stx er det min erfaring, at det kunne være svært at nå at give alle elever feedback på deres forsøgsdesign, når jeg underviste undersøgelsesbaseret. Da den gennemsnitlige klassekvotient er på 26-28 elever på de gymnasiale uddannelser (Epinion, 2018), kan man antage, at det samme gør sig gældende for andre gymnasielærere. Samtidig ser kunstig intelligens ud til at

[1] Generativ AI er en gren af kunstig intelligens (AI), der automatisk genererer nyt indhold som svar på opgaver, der er blevet skrevet med almindeligt sprog. Svarene er baseret på eksisterende data fra websteder, sociale medier med videre og bliver skabt ud fra analyse af statistiske fordelinger af ord, pixels eller andre data. AI genereret indhold kan optræde i forskellige formater f.eks. almindeligt sprog, billeder eller video. Generativ AI forstår ikke virkelige objekter og sociale relationer og kan heller ikke generere nye ideer (Unesco, Fengchun, & Holmes, 2023).

have potentialet til at effektivisere og forbedre feedback på uddannelsesområdet (Friisberg, 2023; Debuse & Lawley, 2016), men da feltet er relativt nyt, mangler der generelt forskning om gymnasieelevers brug af generativ AI. Det gælder også formativ feedback fra generativ AI i relation til elevers design af eksperimenter i naturvidenskab, der ikke lader til at være beskrevet i forskningslitteraturen. Som underviser i biologi på stx var det mit indtryk, at eleverne var særligt glade for at bruge ChatGPT som generativ AI, og derfor valgte jeg at arbejde med følgende problemformulering i mit masterprojekt:

1.1 Problemformulering

Hvordan påvirker formativ feedback fra ChatGPT gymnasieelevers design af biologieksperimenter?

Hvad gør elever med formativ feedback fra ChatGPT, når de skal designe biologieksperimenter?
Hvordan påvirker formativ feedback fra ChatGPT kvaliteten af elevers forsøgsdesign i biologi?
Hvordan opfatter elever ChatGPT som feedbackgiver på forsøgsdesign i biologi?

I denne opgave skelnes ikke mellem betegnelserne eksperiment og forsøg, samtidig er det elevernes anvendelse af ChatGPT 3.5, der ligger til grund for undersøgelsen. Da elevers design af forsøg ofte indgår som en del af undersøgelsesbaseret naturvidenskabsundervisning indledes kapitel 2 med en teoretisk introduktion til IBSE (Inquiry Based Science Education), som følges op af en gennemgang af formativ evaluering og kriterier for god formativ feedback fra lærer til elev, da disse felter ligeledes er centrale i forhold til problemformuleringen. Dernæst præsenteres en litteraturgennemgang i kapitel 3, som fokuserer på formativ feedback på elevers arbejde med at designe forsøg og undersøger, om AI kan fungere som feedbackgiver i denne sammenhæng. I kapitel 4 gennemgås empirisk design og metoder for den undersøgelse, der skal besvare problemformuleringen. Herefter følger en gennemgang af undersøgelsens resultater i kapitel 5. I kapitel 6 diskuteres resultaterne med fokus på elevernes brug af feedback fra ChatGPT og dens påvirkning af kvaliteten af deres forsøgsdesign. Hertil kommer en diskussion af ChatGPT´s styrker og svagheder som formativ feedbackgiver på forsøgsdesign, som leder frem til kapitel 7, hvor opgavens konklusioner og perspektiver samles.

Tidligere i masteruddannelsen har vi afleveret et udkast til en indledning og et review, som danner udgangspunkt for indledning og review i denne opgave. For en god ordens skyld er den oprindelige opgave vedlagt i bilag 1, selvom delelementer, der er genanvendt, er omformuleret.

2. Teoretisk grundlag

I dette kapitel bliver det teoretiske grundlag gennemgået med særligt fokus på undersøgelsesbaseret undervisning, formativ evaluering, formativ feedback og socialkonstruktivisme.

2.1 IBSE

Der eksisterer forskellige tilgange til undersøgelsesbaseret læring, men fælles for dem alle er et fokus på, at læring stimuleres ved undersøgelse af f.eks. et problem eller et spørgsmål. Derudover tager læringen udgangspunkt i en proces, hvor elever konstruerer viden, der er ny for dem, igennem en aktiv og elevcentreret tilgang, som indebærer "learning by doing". Lærerens rolle er faciliterende eller vejledende, og der lægges op til, at eleverne i stigende grad tager ansvar for deres egen læring (Spronken-Smith & Walker, 2010).

Undersøgelsesbaseret læring i de naturvidenskabelige fag er ligeledes beskrevet med forskellige betegnelser f.eks. IBSE (Inquiry Based Science Education), UBNU (Undersøgelsesbaseret Naturfags Undervisning) og IBSME (Inquiry Based Science and Math Education). I dette projekt benyttes betegnelsen IBSE, der kan anvendes som en paraplybetegnelse for forskellige tilgange (Frisdahl, 2014). Ud over de ovenfornævnte kendetegn lægges der i IBSE-tilgangen vægt på autentiske spørgsmål med relation til elevernes egne erfaringer. Samtidig er der fokus på, at eleverne udvikler viden om og forståelse for naturvidenskabelige begreber, processer og arbejdsmetoder (Anderson, 2002). En af de modeller, der anvendes i undersøgelsesbaseret undervisning, er 6F-modellen, som er struktureret omkring faserne: **F**orudsætning, **F**ang, **F**orsk, **F**orklar, **F**orlæng og **F**eedback. Elevernes arbejde med at designe biologieksperimenter i dette masterprojekt tager udgangspunkt i 6F-modellen, bl.a. fordi læreren kan skabe "en situation, hvor eleverne udforsker fænomener og opnår læring af et forudbestemt fagligt indhold gennem diskussioner og lærerinstruktion" (Madsen, Evans, & Bruun, 2020, s. 39). Hertil kommer, at 6F-modellen tilbyder "en undervisningsform der ligger mellem traditionelle kogebogsøvelser der er svære at begrunde

læringsmæssigt, og helt åbne undersøgelsesbaserede øvelser hvor elevernes spørgsmål alene er styrende, der både tidsmæssigt og i relation til at opnå bestemte læringsmål passer dårligt i de danske ungdomsuddannelser" (Madsen, Evans, & Bruun, 2020, s. 40). Derudover spiller formativ feedback en central rolle i 6F-modellen (Madsen, Evans, & Bruun, 2020), hvilket er relevant i forhold fokus i dette masterprojekt.

2.2 Formativ evaluering

Evaluering kan have forskellige formål. Hvis formålet er at understøtte læring, er der tale om formativ evaluering, mens evaluering betegnes som summativ, hvis formålet er evaluering af læring, som det f.eks. er tilfældet med karaktergivning. Der findes flere forskellige definitioner på formativ evaluering, men de mest udbredte har fokus på formativ evaluering som en proces (Dolin, Harlen, Black, & Tiberghien, 2018). Èn af dem er den følgende definition, som bliver brugt som udgangspunkt i dette projekt: "Formative assessment is a process used by teachers and students during instruction that provides feedback to adjust ongoing teaching and learning to improve students' achievement of intended instructional outcomes." (McManus, 2008, s. 3).

Udover at være én af de mere udbredte definitioner på formativ evaluering, er definitionen valgt, fordi den er forenelig med, at der leveres beviser, der kan bruges af lærere og elever til at tilpasse undervisningen (McManus, 2008). I dette projekt har ChatGPT overtaget lærerens rolle og fungerer som virtuel vejleder, mens elevernes beskrivelse af deres forsøgsdesign fungerer som beviser, der skal bruges som udgangspunkt for, at ChatGPT kan guide den videre læreproces. McManus (2008) betegner det samtidig som vigtigt, at læreren deler læringsmål og succeskriterier, ligesom eleverne bør deltage aktivt i processen med at forbedre deres læring. I dette masterprojekt bliver eleverne gjort bekendt med læringsmål og evalueringskriterier for deres forsøgsdesign samtidig med, at de aktivt skal forholde sig til ChatGPT´s forbedringsforslag og beslutte, hvad der skal implementeres i deres reviderede forsøgsdesign. Endelig har den ovenstående definition af formativ evaluering den fordel, at den ikke abonnerer på én bestemt evalueringsstrategi, men åbner op for, at der kan anvendes flere forskellige tilgange i den formative evalueringsproces (McManus, 2008).

2.3 Formativ feedback

Et af de helt centrale elementer i formativ evaluering er feedback (Harlen, 2013a). Dolin, Harlen, Black og Tiberghien (2018) definerer feedback som det at give respons på et produkt eller en proces for at forbedre performance. Den mest åbenlyse lærerfeedback gives mundtligt eller skriftligt til elever, men feedbacken kan have mange former og kan også gives mere ubevidst gennem gestikulation, intonation eller handlinger, når læreren tildeler elever opgaver (Harlen, 2013b).

Ifølge forskningen er god formativ feedback kendetegnet ved at være givet som kommentarer uden karakterer eller andre scores og undgår sammenligninger med andre elever (Harlen, 2013a). Dette bakkes op af andre forskere, der påpeger, at sammenligning med bedre elever eller opgaver ikke er motiverende for elever (ARG, 2002). Derudover skal den formative feedback identificere, hvad der er blevet gjort godt, hvad der kan forbedres, og hvordan eleven kan gå i gang med forbedringen (Harlen, 2013a; McManus, 2008). Det stemmer fint overens med, at feedback bl.a. skal adressere spørgsmålene, ”Hvordan klarer jeg mig?”, og ”Hvad er næste skridt?” (Hattie & Timperley, 2007). Fokus i den givne feedback er vigtig, fordi den påvirker, hvad eleverne lægger vægt på. Derfor bør feedbacken være relateret til læringsmål og succeskriterier for opgaven (McManus, 2008), hvilket er i overensstemmelse med, at feedback skal forholde sig til spørgsmålet, ”Hvor skal jeg hen?” (Hattie & Timperley, 2007). Ifølge forskningen skal formativ feedback desuden være relevant og effektiv i forhold til elevens arbejde med forbedringer (Dolin, Harlen, Black, & Tiberghien, 2018). Derfor bør formativ evaluering tage udgangspunkt i den enkelte elev, da manglende differentiering mellem elever for eksempel kan være demotiverende for svagere elever (Harlen, 2013a), hvilket også må antages at gælde for feedbacken. Ifølge forskningen bør feedback desuden hjælpe eleverne med at blive opmærksomme på, hvad de har lært, ligesom læreren bør sikre sig, at eleverne forstår deres kommentarer (Harlen, 2013b). McManus (2008) tilføjer desuden, at feedbacken bør være konkret og rettidig. Endelig bør formativ evaluering være empatisk, da enhver evaluering har en følelsesmæssig indvirkning (ARG, 2002), hvilket også må antages at gælde formativ lærerfeedback. Kravene til god formativ feedback er opsummeret i figur 1.

God formativ feedback fra lærer til elev

- Bliver givet som kommentarer uden karakterer eller scores
- Undgår sammenligning med bedre elever eller opgaver
- Identificerer, hvad der er blevet gjort godt, hvad der kan forbedres, og hvordan man kan gå i gang med forbedringen
- Er effektiv i forhold til elevens arbejde med forbedringer
- Er rettidig
- Tager udgangspunkt i læringsmål og succeskriterier for opgaven
- Er tilpasset den enkelte elevs niveau og formåen
- Gør eleverne opmærksomme på, hvad de har lært
- Er forståelig for eleverne
- Er empatisk og tager hensyn til elevernes følelser

Figur 1. Oversigt over krav til god formativ feedback fra lærer til elev.

2.4 Socialkonstruktivisme

Kravene til god formativ feedback fra lærer til elev, som er skitseret i det foregående afsnit, skulle gerne bidrage til, at eleven deltager aktivt i at konstruere sin egen læring. Ud fra et socialkonstruktivistisk perspektiv sker dette gennem social interaktion (Dolin, Harlen, Black, & Tiberghien, 2018). Da fokus i denne master er på elevernes design af eksperimenter, er det særligt Forsk-fasen i 6F-modellen, der er omdrejningspunktet. Denne fase ”indeholder elementer af, hvad Vygotsky (1978) ville kalde leg: Man prøver, fejler måske og opbygger viden – hele tiden i et socialt fællesskab” (Madsen, Evans, & Bruun, 2020, s. 39). Et andet interessant aspekt er *Zonen for nærmeste udvikling* (ZNU), som Vygotsky definerer som: ”the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance or in collaboration with more capable peers” (Vygotsky, 1978, s. 86). ZNU er bl.a. interessant, fordi den kan sættes i relation til vigtigheden af det tidligere nævnte krav om differentieret feedback, der er tilpasset den enkelte elevs formåen og niveau. Den beskrevne kobling mellem socialkonstruktivisme og henholdsvis Forsk-fasen og formativ feedback gør det oplagt at bruge socialkonstruktivismen som læringsteoretisk udgangspunkt i dette masterprojekt.

Dermed udgør IBSE, formativ evaluering og feedback og socialkonstruktivisme det teoretiske grundlag for besvarelsen af problemformuleringen. I det følgende kapitel præsenteres de forskningsresultater, som masterprojektet også tager udgangspunkt i.

3. Litteraturreview

Formativ feedback fra ChatGPT i relation til elevers forsøgsdesign ser ikke ud til at være velbeskrevet i forskningslitteraturen i hverken dansk eller international gymnasiekontekst. Samtidig konkluderer et metodisk grundigt reviewstudie af Ouyang, Dinh og Xu (2023), at brugen af AI til formativ evaluering inden for STEM-uddannelserne ligeledes er sparsomt beskrevet. Studiet viser, at der er udgivet 11 engelsksprogede peer reviewed artikler med empiriske studier, som beskriver AI i relation til formativ evaluering inden for STEM-uddannelserne i tidsperioden januar 2011 til april 2023. I de fleste af studierne er fokus på AI-genereret feedback, der ikke bare er simple scores, begrænset. Derudover er Kina og USA angivet som de mest publicerende lande (Ouyang, Dinh, & Xu, 2023).

Da forskningslitteraturen er sparsom, og dette projekt undersøger, hvordan formativ feedback fra ChatGPT påvirker gymnasieelevers design af biologieksperimenter, er det i stedet oplagt at undersøge forskningslitteraturens beskrivelse af formativ feedback i relation til elevers design af eksperimenter. Selvom ChatGPT´s rolle som feedbackgiver på forsøgsdesign ikke er beskrevet i forskningen, kan litteraturen bruges til at belyse AI´s potentiale til at kunne fungere som feedbackgiver i denne sammenhæng. Endelig må man forvente, at elevers holdninger til brug af ChatGPT i undervisningssammenhæng kan påvirke deres opfattelse af ChatGPT som feedbackgiver på deres forsøgsdesign, og derfor vil reviewet tage udgangspunkt i følgende spørgsmål:

1) Hvordan påvirker formativ feedback elever, der designer eksperimenter?
2) Vil AI kunne fungere som feedbackgiver på elevers forsøgsdesign?
3) Hvordan opfatter elever ChatGPT som redskab i undervisningen?

Følgende prioriteringer er blevet anvendt i besvarelsen af de ovenstående spørgsmål:
Det mest hensigtsmæssige er at anvende peer reviewed empiriske studier på gymnasieniveau inden for biologi fra lande med uddannelsessystemer, der minder om det danske, til at belyse de

ovenstående spørgsmål. I de tilfælde, hvor det ikke er muligt, er der suppleret med empiriske studier fra andre uddannelsesniveauer og fra lande med anderledes uddannelsessystemer. Hvis der er inddraget empiriske studier med andre fag end biologi, er studier med naturvidenskabelige fag blevet prioriteret over studier med andre fag. Endelig er studier med ChatGPT blevet prioriteret over studier med andre AI-værktøjer.

3.1 Hvordan påvirker formativ feedback elever, der designer eksperimenter?

Når elever skal arbejde med forsøgsdesign, kan formativ feedback have en positiv effekt på deres tilegnelse af viden og indre motivation. Det viser et studie af Eckes & Wilde (2019), der undersøger effekten af informative tutoring feedback givet af lærere i biologiundervisningen blandt 165 6. og 7. klasses elever i Tyskland. Informative tutoring feedback er karakteriseret ved, at eleven bliver vejledt til at opdage fejl, overvinde forhindringer og anvende effektive opgaveløsningsstrategier. Andre formative feedbackformater ser også ud til at have en effekt. For eksempel viser et studie blandt fire 8. og 9. klasser i Tyskland, at 80 % af eleverne reviderer deres eksperimentelle forsøgsdesign i biologiundervisningen efter at have deltaget i en proces med peer feedback (Anker-Hansen & Andrée, 2019). Derudover præsenterer Gajanová et al. (2021) resultater fra et studie med 121 high school-elever i Slovakiet, der viser, at formativ evaluering forbedrer elevernes færdigheder inden for forsøgsdesign. Studiet er udført indenfor forskellige naturvidenskabelige fag, herunder biologi og skelner ikke mellem forskellige formative feedbackformater. Forskellige typer af feedback har imidlertid ikke samme effekt, f.eks. ser informative tutoring feedback ud til at være korreleret med en højere videnstilegnelse hos elever sammenlignet med feedback, hvor læreren har holdt sig mere i baggrunden og i stedet ladet eleverne diskutere tvivlsspørgsmål (Eckes & Wilde, 2019). I den sammenhæng er det interessant, at elever, der modtager undervisning helt uden formativ evaluering, klarer sig signifikant dårligere, når det gælder deres forbedring af færdigheder inden for forsøgsdesign end elever, der har modtaget undervisning med formativ evaluering (Ganajová, et al., 2021). Da formativ evaluering ser ud til at have positiv indflydelse på elevers tilgang til forsøgsdesign, er det relevant at undersøge, om en AI-model som ChatGPT kan give brugbar formativ feedback til elever, der skal designe eksperimenter.

3.2 Vil AI kunne fungere som feedbackgiver på elevers forsøgsdesign?

En af forudsætningerne for, at ChatGPT vil kunne fungere som feedbackgiver på elevers design af biologieksperimenter er, at AI kan levere faglig korrekt vejledning. Et studie af Zhang (2023) viser, at studerende ville klare sig bedre i en multiple choice-test, hvis de stolede mere på ChatGPT. Svagheden ved studiet er, at det ikke er udgivet i et peer reviewed tidsskrift, men det har den styrke, at det er baseret på 2828 multiple choice-spørgsmål, herunder biologispørgsmål på high school-niveau. Samtidig understøttes den fundne tendens til, at ChatGPT kan levere faglig vejledning, af to andre studier, der konkluderer, at ChatGPT giver god faglig vejledning inden for geometri (Wardat, Tashtoush, AlAli, & Adeeb, 2023) og i fysik på bachelorniveau (Ding, Li, Jiang, & Gapud, 2023). Her er det værd at bemærke, at ChatGPT´s faglige vejledning ikke er fejlfri (Ding, Li, Jiang, & Gapud, 2023; Wardat, Tashtoush, AlAli, & Adeeb, 2023), f.eks. viser resultater fra Ding, Li, Jiang og Gapud (2023), at kun 85 procent af ChatGPT´s svar er korrekte.

En anden forudsætning for at bruge ChatGPT som feedbackgiver på elevers forsøgsdesign er, at AI kan bruges til at foretage en korrekt bedømmelse af opgaver med naturvidenskabeligt indhold, der ikke udelukkende er baseret på udenadslære. Et studie, der har brugt machine learning, som er en tilgang inden for kunstig intelligens, til at bedømme opgaver inden for fysik og kemi, der krævede dybere forståelse og brug af ræsonnement, viser, at machine learning kan score opgaver på high school-niveau med samme nøjagtighed som menneskelige eksperter (Maestrales, et al., 2021). Samme tendens findes i et andet studie, der har brugt machine learning til at vurdere svar på åbne spørgsmål i biologiopgaver fra 669 elever fra 25 high schools (Ariely, Nazaretsky, & Alexandron, 2023). Desuden viser et tredje studie, at et Inquiry Intelligent Tutoring System, som er et system inden for kunstig intelligens, der er designet til at facilitere og understøtte læring, kan bruges til at bedømme elevers færdigheder inden for design af eksperimenter på mellemtrinnet (Gobert, Pedro, Raziuddin, & Baker, 2013). Det bakkes op af anden forskning, som viser, at der lader til at være et stort potentiale for at bruge AI-teknologier til at give feedback til studerende (Debuse & Lawley, 2016). I den sammenhæng er det værd at bemærke, at en tredje vigtig forudsætning for at bruge ChatGPT som feedbackgiver på elevers forsøgsdesign i dette projekt er, at AI kan levere feedback, der kan bruges formativt.

I flere studier inden for STEM leveres AI-genereret feedback i form af en score (Ariely, Nazaretsky, & Alexandron, 2023; Gobert, Pedro, Raziuddin, & Baker, 2013), men i et studie, der er udført på kandidatniveau på et datalogikursus i Italien, modtog de studerende AI-feedback i form af scores og korte sætninger, der var skrevet i almindeligt sprog. Resultaterne viser, at studerende, der brugte feedbacken formativt, klarede sig bedre end studerende, der ikke brugte den formativt (Vittorini, Menini, & Sara Tonelli, 2021). Endelig konkluderer et australsk studie, at algoritmer inden for machine learning kan bruges til at levere feedback med kommentarer af høj kvalitet til studerende på universitetsniveau. I definitionen af god feedback lægger studiet vægt på feedbackens konsistens, mængde, detaljeniveau, læsbarhed, specificitet og personalisering (Debuse & Lawley, 2016). Der er således overlap med nogle, men ikke alle de teoretiske kriterier for god formativ feedback fra lærer til elev, som er beskrevet i afsnit 2.3.

Ovenstående viser, at AI ser ud til at kunne bruges til at bedømme elevopgaver og give feedback med faglig vejledning inden for STEM-uddannelserne, herunder biologi på gymnasieniveau, men ud fra de beskrevne studier er det usikkert, om alle kriterier for god formativ lærerfeedback vil være opfyldte. I forlængelse heraf er det interessant at undersøge elevers opfattelse af ChatGPT i undervisningssammenhæng.

3.3 Hvordan opfatter elever ChatGPT som redskab i undervisningen?

Ding, Li, Jiang og Gapud (2023) har undersøgt, i hvilken grad 40 studerende, der var i begyndelsen af deres videregående uddannelse, stolede på, at de fik korrekte svar af ChatGPT efter at have brugt den som virtuel vejleder op til en multiple choice-test i deres fysikundervisning. Selvom de studerende var informeret om, at ChatGPT kan give forkerte svar, og 15 procent af de svar, som de modtog, var forkerte, stolede næsten halvdelen af eleverne fuldt ud på ChatGPT´s svar. Dette resultat er i modsætning til det tidligere nævnte studie af Zhang (2023), der konkluderer, at elever ikke altid stoler nok på råd fra ChatGPT, og at de kunne opnå bedre testresultater ved at følge vejledningen fra ChatGPT. Ding, Li, Jiang og Gapud (2023) forklarer den høje tillid i deres studie med, at ChatGPT´s funktion som virtuel vejleder i fysikundervisningen igennem et stykke tid kan have øget de studerendes tillid. Et andet relevant resultat fra studiet er, at de studerende, der stolede fuldt ud på ChatGPT, også lod til at være mere tilbøjelige til at ville anvende ChatGPT i fremtiden end andre studerende.

I forlængelse heraf er det interessant, at et studie fra Ungarn viser, at brug af ChatGPT er udbredt blandt high school-elever. 74,6 % af eleverne angiver, at de oplever ChatGPT som et nyttigt værktøj, når de arbejder med skoleopgaver og -projekter, ligesom 69% er enige eller meget enige i, at ChatGPT har forbedret deres forståelse af komplekse koncepter. Derudover er 54,9 % enige eller meget enige i, at de er afhængige af ChatGPT for at opnå akademisk succes, mens 45,1 % er uenige. Studiet, der også kommer frem til, at der er brug for, at elever bliver bedre uddannet i brug af ChatGPT til skolearbejde, er baseret på en spørgeskemaundersøgelse med 71 besvarelser og en svarprocent på 59,2 % (Forman, Udvaros, & Avornicului, 2023), hvilket gør konklusionerne mindre robuste.

Det ovenstående review viser, at formative feedbackformater ser ud til at have en positiv effekt på elevers tilgang til forsøgsdesign på grundskole og high school-niveau i udlandet, hvilket formentlig også vil gøre sig gældende på gymnasieniveau i Danmark. Derudover lader det til, at AI kan levere faglig vejledning i naturvidenskab, der er overvejende korrekt, ligesom kunstig intelligens kan bedømme biologiopgaver på high school-niveau og generere brugbar skreven feedback til universitetsstuderende. Endelig viser forskningen, at mange studerende er villige til at tage mod vejledning fra ChatGPT, og derfor tyder det på, at AI har potentialet til at kunne give feedback på elevers forsøgsdesign i biologi på gymnasieniveau.

Resultaterne, der er blevet præsenteret i dette kapitel, danner udgangspunkt for den videre undersøgelse af, hvordan formativ feedback fra ChatGPT påvirker gymnasieelevers design af biologieksperimenter, som er beskrevet i det følgende kapitel.

4. Undersøgelsesmetoder og design

4.1 Undervisningens organisering og kontekst

Data til projektet blev indsamlet i en 3.g biologi B-klasse og en 1.g klasse, der havde biologi som en del af naturvidenskabeligt grundforløb (NV). Udvælgelsen var styret af, hvilke hold jeg havde adgang til, men sikrede samtidig en god spredning på de elever, der indgik i projektet, hvad angik deres erfaringer med naturvidenskab i gymnasiet. Desuden harmonerede masterprojektets fokus på forsøgsdesign med læreplanerne i begge fag. Det fremgår af NV-lærerplanen, at ”Det

eksperimentelle og undersøgende arbejde såvel i laboratoriet som i felten skal stå centralt i undervisningen" (BUVM, 2017a, s. 1). Dette er yderligere specificeret i fagets vejledning, hvor der står, at eleverne kan "foreslå forsøgsdesign for enkelte af forsøgene" (BUVM, 2023a, s. 14). Det samme gør sig gældende for læreplanen for biologi B, hvor der står, at eleverne skal kunne "tilrettelægge og udføre eksperimenter og undersøgelser i laboratoriet" (BUVM, 2017b, s. 1), ligesom "inquirybaseret læring" er nævnt i fagets vejledning (BUVM, 2023b, s. 15).

1.g klassen arbejdede med biobrændsel som tema og designede et forsøg om gæring, mens 3.g-klassen arbejdede med et forløb om nervesystemet og designede et forsøg om tærskelværdi i relation til smagssansen. Man kan argumentere for, at det havde været en fordel, hvis klasserne havde arbejdet med det samme forsøg, men det passede desværre ikke sammen med de forløb, som klasserne skulle arbejde med. De to klasser havde følgende læringsmål om forsøgsdesign til fælles:

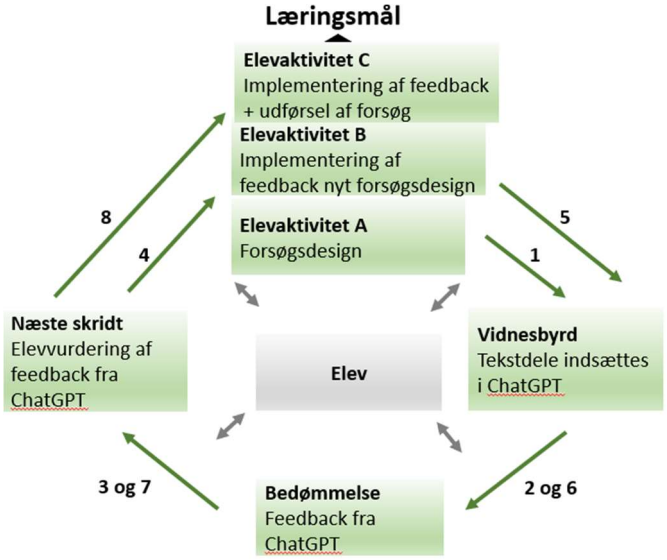
Eleverne skulle kunne:

- Designe et forsøg, der testede et undersøgelsesspørgsmål
- Udarbejde en forsøgsvejledning, der var så præcist beskrevet, at andre ville kunne udføre forsøget.
- Inddrage variabelkontrol og et kontrolforsøg i deres forsøgsdesign.

Læringsmålene om kontrolforsøg og variabelkontrol blev valgt, fordi de er centrale for arbejdet med forsøgsdesign. Samtidig viser forskningen, at universitetsstuderende på første år på et biologikursus havde svært ved at bruge kontrolforsøg rigtigt (Coleman, et al., 2023), ligesom Gobert, Pedro Raziuddin og Baker (2013) nævner, at flere studier konkluderer, at studerende har en tendens til at ændre for mange variabler i samme eksperiment. Dermed var det oplagt at antage, at elever på gymnasieniveau kunne have gavn af, at disse områder blev adresseret i undervisningen. Da kvaliteten af de svar, som ChatGPT kommer med, afhænger af de input, som brugeren leverer (Wardat, Tashtoush, AlAli, & Adeeb, 2023), havde læringsmålet om en præcis forsøgsbeskrivelse desuden til formål at hjælpe eleverne til at få bedre input fra ChatGPT.

Rammen for elevernes arbejde med læringsmålene var, at begge klasser arbejdede undersøgelsesbaseret med forsøgsdesign ud fra en elevvejledning. Den var struktureret ud fra en fælles grundskabelon, som er gennemgået i afsnit 4.2. Arbejdet foregik i grupper af tre, hvilket stemmer fint overens med et socialkonstruktivistisk læringssyn, hvor social interaktion danner basis for individets videnskonstruktion (Dolin, Harlen, Black, & Tiberghien, 2018). Årsagen, til at grupperne ikke var større, var, at der kun var mulighed for at inddrage to klasser i projektet. Som det bliver beskrevet senere i afsnit 4.3, er en del af dataindsamlingen i projektet baseret på elevernes rapporter om de forsøg, de havde designet. Hvis grupperne havde været større, ville det betyde, at der ville indgå færre elevrapporter i projektet, og det ville give et dårligere datagrundlag. Argumentet, for at grupperne skulle bestå af mere end to personer, var, at arbejdet med forsøget foregik over to særskilte moduler. Det betød, at der var risiko for frafald af elever mellem første og andet modul og dermed større risiko for, at eleverne endte med at skulle arbejde individuelt, hvis de kun havde arbejdet i par fra starten, hvilket ikke var hensigtsmæssigt i forhold til projektets socialkonstruktivistiske tilgang. Da projektet undersøger, hvordan formativ feedback fra ChatGPT påvirker elevers forsøgsdesign, holdt jeg mig som lærer i baggrunden og lod i stedet ChatGPT fungere som en virtuel vejleder.

Som tidligere nævnt har formativ feedback et stort potentiale i forhold til at øge læring (Harlen, 2013a), og derfor var begge klassers eksperimentelle arbejde struktureret ud fra den grundmodel for evaluering, som er vist på figur 2. Inden eleverne gik i gang med at udarbejde deres forsøgsdesign, var de blevet introduceret til læringsmål og evalueringskriterier for forsøgsdesignet, som det er anbefalet i litteraturen (Dolin, Harlen, Black, & Tiberghien, 2018). De formative processer i projektet er beskrevet i figurteksten til figur 2.



Figur 2. Figuren viser en gennemgang af de formative processer i projektet. Eleverne blev indledningsvis bedt om at opstille undersøgelsesspørgsmål og komme med bud på forsøgsdesign. Arbejdet skulle formidles skriftligt, så eleverne efterfølgende kunne sætte deres tekstarbejde ind i ChatGPT (1) og få forbedringsforslag (2). Forslagene blev efterfølgende diskuteret i gruppen (3) og implementeret i en revideret tekst (4). Herefter satte eleverne det reviderede forsøgsdesign ind i ChatGPT (5) og bad om feedback (6), hvorefter de diskuterede feedbacken (7) og reviderede forsøgsdesignet en sidste gang, hvis der var behov for det (8). Til sidst udførte de forsøget. Figuren er modificeret efter Harlen (2013b).

4.2 Udformning af elevvejledninger

De elevvejledninger, der blev brugt i projektet, er inspireret af elementer fra Forsk- og Forklar-faserne i 6F-modellen, da de øvrige faser blev dækket på anden vis i undervisningen (Vejledningerne kan ses i bilag 2 og 3). Forsk-fasen kan bl.a. være karakteriseret ved, at eleverne udvikler ideer, afprøver dem og indsamler data, mens Forklar-fasen bl.a. er kendetegnet ved, at erfaringer fra Forsk-fasen bliver koblet sammen med faglig viden (Madsen, Evans, & Bruun, 2020). Det lagde elevvejledningerne op til ved at lade eleverne udvikle undersøgelsesspørgsmål, som de ikke selv kendte svaret på til at begynde med, og et tilhørende forsøgsdesign, der skulle hjælpe dem med at finde svar. Derudover indsamlede de data, da de afprøvede deres forsøgsdesign og koblede resultaterne sammen med biologifaglig viden.

Læringsmålene fra afsnit 4.1 var blevet omformuleret i elevvejledningerne, så de indgik som krav til elevernes forsøgsdesign. Derfor dækkede elevvejledningens krav også kun nogle af de elementer, der skal indgå i et godt forsøgsdesign. Denne tilgang blev valgt for at stilladsere elevernes arbejde med forsøgsdesign.

I elevvejledningerne var der desuden en materialeliste, som både indeholdt materialer, der var relevante for elevernes forsøg, og materialer, som ikke var relevante. Hensigten med denne tilgang var at rammesætte elevernes arbejde ved at afgrænse deres materialevalg, da helt åbne forsøg, hvor det kun er elevernes spørgsmål, der styrer forsøgsdesignet, kan være svære at styre tidsmæssigt, ligesom det kan være udfordrende at få dem til at passe sammen med bestemte læringsmål (Madsen, Evans, & Bruun, 2020). Hensigten med også at tilbyde irrelevante materialer var at få eleverne til at forholde sig kritisk. Dette blev yderligere understøttet i elevvejledningerne, hvor eleverne blev bedt om at markere relevante og irrelevante/forkerte forbedringsforslag, som et led i deres arbejde med ChatGPT´s formative feedback. Denne tilgang stemmer fint overens med anbefalingerne om, at kritisk tænkning[2] indgår som en del af et AI-curriculum (Unesco, 2022; Long & Magerko, 2020). Den bliver desuden understøttet af en ekspertgruppe, der blev nedsat af Børne- og Undervisningsministeriet, som skriver følgende: ”En kritisk og konstruktiv brug af kunstig intelligens er endnu en vej mod moderne dannelse” (Vedersø, et al., 2023). Endelig er den kritiske tilgang nødvendig, fordi forskningen viser, at ChatGPT ikke leverer 100 procent korrekt vejledning jf. afsnit 3.2. Dette kan være problematisk, fordi forkert brug af AI-genereret feedback kan føre til fremtidige læringsmæssige misforståelser (Ding, Li, Jiang, & Gapud, 2023).

I arbejdet med at udforme de endelige elevvejledninger blev der udført et pilotforsøg i en 1.g-klasse med 26 elever, der blev undervist i NV. Elevernes forsøgsrapporter blev afleveret og dannede sammen med mine observationer fra undervisningen og et pilotspørgeskema, som de havde besvaret, udgangspunkt for følgende justeringer i elevvejledningerne og i undervisningens organisering.

Pilotforsøget viste, at eleverne ikke gav sig ret meget tid til at arbejde med forsøgsdesign og implementering af feedback fra ChatGPT. Det skyldtes formentlig, at de hellere ville i gang med at udføre forsøget med det samme. Problemet var, at det påvirkede validiteten negativt af den undersøgelse, der ligger til grund for dette masterprojekt. Validitet, der handler om undersøgelsens gyldighed, refererer i dette projekt til, i hvilket omfang en given undersøgelse

[2] Kritisk tænkning kan betragtes som en intellektuel proces, der har at gøre med evnen til at tænke kritisk. Det kan f.eks. indebære, at man aktivt udfordrer, undersøger og evaluerer information, hvilket kræver mere end blot at huske fakta. Ovenstående er inspireret af materiale fra University of Louisville, som præsenterer en mere fyldestgørende definition (Louisville, u.d.).

måler et koncept præcist. Definitionen er inspireret af Heale og Twycross (2015). For at øge validiteten blev arbejdet med forsøget fordelt på to moduler i stedet for et, så eleverne fik god tid til at designe forsøg i det ene modul og udførte deres forsøg i det næste modul.

Et andet interessant resultat fra pilotundersøgelsen var, at eleverne havde svært ved at navigere i den vejledning, som de skulle bruge. På side 2 i pilotvejledningen blev de f.eks. bedt om at skrive en prompt i ChatGPT, der indeholdt forskellige oplysninger bl.a. evalueringskriterier for et godt forsøgsdesign, som stod på side 1. Årsagen var som tidligere nævnt, at god formativ feedback fra lærer til elev bl.a. tager udgangspunkt i læringsmål (McManus, 2008). Da der var alignment mellem evalueringskriterier og læringsmål, var det kun evalueringskriterierne, der blev nævnt i prompten. Pilotundersøgelsen viste, at eleverne havde svært ved at hente oplysninger på en anden side i elevvejledningen end den side, som de var kommet til. Samtidig var det svært for dem at kombinere forskellige oplysninger om undersøgelsesspørgsmål, krav til materialer og forsøgsdesign, samt deres bud på et forsøgsdesign i én prompt. Da formålet med dette masterprojekt ikke er at lære eleverne at prompte, men at undersøge, hvordan feedback fra ChatGPT påvirker deres arbejde, blev løsningen, at de reviderede elevvejledninger indeholdt en næsten færdig prompt, hvor eleverne kun skulle indsætte deres undersøgelsesspørgsmål og forsøgsdesign. Som tidligere nævnt afhænger kvaliteten af de svar, som ChatGPT kommer med af brugerens input, og derfor var en af fordelene ved at designe en stor del af prompten for eleverne, at kvaliteten af den feedback, som de fik fra ChatGPT, formentlig blev mere ensartet på tværs af grupper. Det førte til udviklingen af en promptskabelon, der indgår i elevvejledningerne, som bl.a. tager højde for følgende kriterier for gode prompts: præcision, logisk opbygning og eksplicitte specifikationer til det ønskede output (Lo, 2023). En anden fordel ved at benytte en promptskabelon var, at det gav mulighed for at designe en prompt, hvor ChatGPT gav eleverne forslag til forbedringer i små afgrænsede enheder. Dermed blev kravet om, at god formativ feedback gives som kommentarer fra lærer til elev (Harlen, 2013a) og undgår sammenligning med andre elever opgaver eller elever (ARG, 2002), desuden opfyldt. Nedenfor ses et eksempel på en feedbackenhed med et forbedringsforslag fra ChatGPT.

”Kontrolgruppe: Ud over de 5 glas med forskellige mængder citronsaft skal I inkludere en kontrolgruppe med 0 dråber citronsaft. Dette vil hjælpe med at afgøre, om deltageren i forsøget kan smage citronen i forhold til en ren vandsmag.” (Bilag 4)

Indførelsen af promptskabelonen var desuden relevant, fordi flere elevrapporter fra pilotundersøgelsen viste, at eleverne havde en tendens til bare at kopiere ChatGPT´s svar, hvis de blev præsenteret for en færdig øvelsesvejledning. Det fremgår også af spørgeskemaundersøgelsen fra pilotstudiet, hvor en elev, der var blevet spurgt, hvordan de havde brugt ChatGPT´s feedback, svarede ”vi kopierede bare det meste af det” (Bilag 9).

ChatGPT´s svar på elevernes anden prompt indeholdt desuden en kort indledende tekst med angivelse af, hvad der fungerede godt i deres forsøgsdesign, som er vist i eksemplet nedenfor:

”Dit foreslåede forsøgsdesign er generelt godt, da det tager højde for mange af de vigtige aspekter ved at undersøge tærskelværdien for smagssansen for citronsaft. Det er også godt, at du har inkluderet flere forsøgspersoner for at tage højde for individuelle variationer i smagssansen.” (Bilag 4)

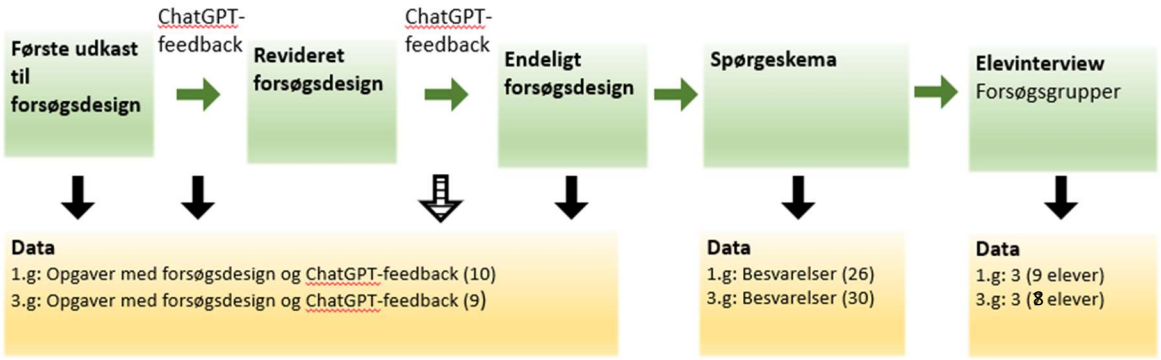
De reviderede elevvejledninger indgik i undersøgelsens empiriske design, som er beskrevet i næste afsnit.

4.3 Empirisk design og data

Projektet blev gennemført som et mixed method-studie, idet der blev anvendt kvantitative og kvalitative metoder i det empiriske design, som er vist på figur 3 på næste side. Alle data blev indsamlet i september 2023.

Den kvantitative metode i projektet var bl.a. baseret på data fra 19 anonymiserede elevrapporter, hvor elevernes bud på forsøgsdesign fremgik før og efter feedback fra ChatGPT (bilag 4). Derudover havde eleverne som tidligere nævnt markeret relevante og irrelevante/forkerte forbedringsforslag fra ChatGPT. Fordelen ved denne tilgang var, at den for det første gav indblik i, hvilken effekt ChatGPT´s forbedringsforslag havde på det faglige niveau i elevernes opgaver. For det andet gav metoden indsigt i elevernes evne til at identificere relevante og irrelevante/forkerte forbedringsforslag fra ChatGPT. Metoden kunne derimod ikke afdække, hvad der lå til grund for, at

eleverne brugte ChatGPT, som de gjorde, og hvordan de oplevede ChatGPT´s feedback. Derfor blev dataindsamlingen suppleret af en spørgeskemaundersøgelse.



Figur 3. Empirisk design. De grønne kasser illustrerer de aktiviteter, som eleverne har indgået i, og de grønne pile viser rækkefølgen af aktiviteterne. Elevernes rapporter med forsøgsdesign og feedback fra ChatGPT indgik i elevinterviewene sammen med deres spørgeskemabesvarelser. De gule kasser viser hvilke data, der er blevet indsamlet, mens de sorte pile illustrerer, hvilket trin i processen data stammer fra. Den stribede sorte pil henviser til, at eleverne i anden runde med feedback fra ChatGPT kun markerede feedback, som de ville bruge til revision af deres forsøgsdesign. Derfor er der nogle grupper, som slet ikke har markeret feedback fra denne runde.

Fordelen ved at anvende en spørgeskemaundersøgelse var, at den kunne bruges til at undersøge de to ovenstående spørgsmål anonymt blandt alle eleverne. Det kan gøre elever mere trygge i undersøgelser som denne, fordi der er et asymmetrisk magtforhold mellem lærer og elev, som blev forstærket af, at jeg skulle give begge klasser karakterer. En anden fordel ved anonymitet er, at man undgår, at eleverne svarer strategisk for at fremstå mere positivt i lærernes øjne, hvilket sænker validiteten. Ulempen var, at det var svært at fange nuancer i respondenternes meninger og holdninger, fordi det ikke var muligt at følge op på besvarelserne med den enkelte elev. Derfor blev spørgeskemaundersøgelsen suppleret med et kvalitativt elevinterview med udvalgte elevgrupper.

Formålet med at interviewe eleverne var, at formatet er fleksibelt og gav mulighed for at gå i dybden med interessante emner, der dukkede op undervejs. Det kunne bl.a. bidrage med en dybere indsigt i de overvejelser, som eleverne gjorde sig, da de skulle designe forsøg og arbejdede med feedbacken fra ChatGPT, ligesom interviewmetoden gav bedre mulighed for at få eleverne til at sætte flere ord på, hvordan de opfattede ChatGPT som feedbackgiver på forsøgsdesign. Ulempen ved metoden var, at eleverne ikke længere kunne være anonyme. For at undgå det eventuelle ubehag, som en elev kan opleve, ved at skulle sidde alene og blive interviewet

dybdegående af sin lærer, blev interviewdelen organiseret som semistrukturerede gruppeinterviews, hvor eleverne blev interviewet sammen med de grupper, som de havde lavet forsøg med.

Forud for interviewene blev der udarbejdet en interviewguide (bilag 5), som skulle sikre en mere ensartet tilgang på tværs af grupper, hvilket kan bidrage til at øge reliabiliteten. Reliabilitet, der handler om pålidelighed, refererer i dette projekt til, i hvilket omfang en given undersøgelse er konsistent og giver de samme resultater ved gentagne målinger af samme situation. Definitionen er inspireret af Heale og Twycross (2015). I denne del af projektet blev der benyttet metodeintegration, idet elevernes rapporter og spørgeskemaundersøgelsen blev brugt som udgangspunkt for at udarbejde en interviewguide (Frederiksen, 2020). I forbindelse med elevrapporterne og spørgeskemaundersøgelserne var eleverne desuden blevet bedt om at vælge et kodenavn. De elever, der indvilligede i at lade sig interviewe, blev bedt om at oplyse deres kodenavn, og derfor var det muligt at lave dataintegration, hvor de enkelte elevers rapporter og spørgeskemabesvarelser blev inddraget i selve interviewet.

Samlet set giver kombinationen af de beskrevne metoder i dette masterprojekt mulighed for triangulering, hvilket kan bidrage til at øge undersøgelsens validitet, fordi fund, der bekræftes af mere end en metode, typisk antages at være bedre underbygget (Frederiksen, 2020). I dette projekt ønsker jeg imidlertid ikke kun at fokusere på fund, der bekræftes af mere end én metode, da de forskellige metoder som nævnt ovenfor også har potentiale til at belyse forskellige aspekter af problemformuleringen. Derfor bygger det empiriske design også på en antagelse om komplementaritet, hvor flere metoder giver mere viden (Frederiksen, 2020), hvilket kan bidrage til at give en mere dækkende beskrivelse af, hvordan formativ feedback fra ChatGPT påvirker gymnasieelevers design af biologieksperimenter. I det følgende afsnit præsenteres en detaljeret beskrivelse af de valgte metoder.

4.3.1 Metode til analyse af elevrapporter

På baggrund af gennemlæsning af seks tilfældige anonymiserede elevrapporter, hvoraf tre var fra 1.g og tre fra 3.g, blev der udarbejdet en rubrik, der blev brugt til at kvalitetsscore elevernes første og sidste forsøgsdesign i alle elevrapporter (Bilag 6). Formålet med rubrikken var både at øge

reliabiliteten og validiteten i bedømmelsesprocessen. For at sikre validiteten var kvalitetsscoren i rubrikken baseret på de tidligere nævnte læringsmål, som handlede om, at eleverne skulle designe et forsøg, der testede deres undersøgelsesspørgsmål og udarbejde en forsøgsvejledning, der var så præcist beskrevet, at andre vil kunne udføre forsøget. Derudover skulle de inddrage variabelkontrol og et kontrolforsøg i deres forsøgsdesign. Ved gennemlæsningen af elevrapporterne blev det tydeligt, at feedbacken fra ChatGPT også havde ført til andre forbedringer af elevernes forsøgsdesign, som handlede om at gentage eller replikere forsøget, antal stikprøver og inkorporering af dataindsamling som en del af forsøgsdesignet. Derfor blev disse elementer medtaget i kvalitetsscoren. I arbejdet med at kvalitetsscore elevernes forsøgsdesign blev det første forsøgsdesign først læst og kvalitetsscoret på baggrund af rubrikken, hvorefter det sidste forsøgsdesign blev læst og scoret efter samme rubrik. Processen er vist på figur 4. Da formatet var formativt, og det desuden først var muligt at kvalitetsscore rapporterne efter, at jeg var holdt op med at undervise eleverne, fik de ikke udleveret kvalitetsscoren for deres forsøgsdesign.

Undersøgelsesspørgsmål
Vi vil gerne undersøge hvad der sker med mængden af co_2 som gær producerer, hvis vi ændrer temperaturen.

Første design

1) Hæld 20g sukker i 4 koniske kolber.
2) Hæld 100 ml koldt vand i 1 kolbe.
3) Hæld 100 ml lunkent vand i 2 kolbe.
4) Hæld 100 ml varmt vand (over 30 grader) i 3 kolbe.
5) Hæld 100 ml vand i 4 kolbe, som er vores kontrolkolbe.
5) Smuldr 5 gram gær i de første tre kolber.
6) Sæt en ballon på hver af de fire kolber.
7) Vent 30 minutter og aflæs hvad der er sket med ballonerne.

Kvalitetsscore: 19
Delkvalitetsscorer:
Undersøgelsesspørgsmål: 4
Præcis forsøgsbeskrivelse: 3
Variabelkontrol: 3
Kontrolforsøg: 5
Gentagelse/stikprøver: 3
Dokumentation: 1

Sidste design

1) Hæld 20g sukker i 4 koniske kolber.
2) Hæld 100 ml koldt vand i 1 kolbe.
3) Hæld 100 ml lunkent vand i 2 kolbe.
4) Hæld 100 ml varmt vand (over 30 grader) i 3 kolbe.
5) Hæld 100 ml vand i 4 kolbe, som er vores kontrolkolbe.
6) Sørg for at måle temperaturen på det vand man putter i og noter dem.
7) Smuldr 5 gram gær i de første tre kolber. Kom ikke gær i kontrolkolben.
8) Sæt en ballon på hver af de fire kolber, det er vigtigt at de bliver sat på, på samme tid, da man sikrer sig at gæren har lige lang tid til at producere co_2. Sørg for at alle balloner er lige store.
9) Vent 30 minutter og aflæs hvad der er sket med ballonerne. Ballonerne skal måles med et målebånd for at se om der er sket en markant forskel.
10) Forsøget burde laves over flere omgange på forskellige dage for at sikre at man kan indsamle nok og præcist viden.

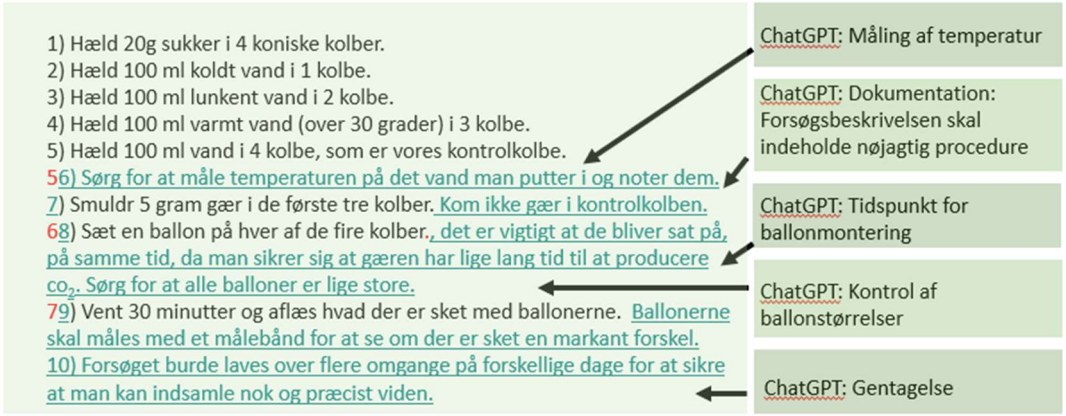
Kvalitetsscore: 23
Delkvalitetsscorer med forbedring:
Præcis forsøgsbeskrivelse 5, variabelkontrol 4, dokumentation 2

Figur 4. Eksempel på kvalitetsscore af elevrapporter. Ændringer mellem første og sidste forsøgsdesign er markeret med rødt for at gøre det lettere at aflæse figuren.

For at undersøge elevernes evne til at afkode ChatGPT´s feedback blev de som tidligere nævnt bedt om at indsætte al feedback fra ChatGPT i den første feedbackrunde og farvekode relevant og

irrelevant eller forkert feedback. Data blev kategoriseret i forhold til, om eleverne havde markeret relevante og irrelevante/forkerte fra input ChatGPT rigtigt eller forkert i første feedbackrunde i forhold til deres første forsøgsdesign. Derudover blev der tilføjet en kategori for input fra ChatGPT, som eleverne havde svært ved at bedømme, enten fordi feedbacken både indeholdt relevante og irrelevante oplysninger, eller fordi den var svær at forstå. Af hensyn til elevernes tidsforbrug blev de ikke bedt om at indsætte og farvekode al feedback fra ChatGPT i anden runde. Her skulle de blot indsætte ChatGPT´s feedback, hvis de ville bruge den til at ændre deres forsøgsdesign. Derfor var det kun data fra første feedbackrunde, der indgik i analysen af elevernes afkodning af feedback fra ChatGPT.

Endelig blev elevernes implementering af ChatGPT´s forbedringsforslag undersøgt ved at sammenligne deres første udkast til forsøgsdesign med deres sidste forsøgsdesign. Sammenligningen af teksterne blev udført ved at bruge "juridisk sammenligning"-funktionen i Word. Irrelevante ændringer som for eksempel ændring i formatering blev sorteret fra, så det kun var ændringer af indholdsmæssig karakter, der indgik i den videre analyse. Herefter blev hver enkelt ændring i forsøgsdesignet analyseret og sammenholdt med ChatGPT´s feedback fra både runde 1 og 2, som det er vist på figur 5.



Figur 5. Eksempel på sammenligning af forsøgsdesign før og efter feedback fra ChatGPT. Ændringerne, der er markeret med grøn tekst, er blevet sammenholdt med forbedringsforslag fra ChatGPT.

4.3.2 Metode til spørgeskema

Som tidligere nævnt var formålet med spørgeskemaet at undersøge, hvad elever gør med formativ feedback fra ChatGPT, og hvordan de opfatter ChatGPT som feedbackgiver, når de skal designe eksperimenter. Det var ikke muligt at finde et tidligere anvendt valideret spørgeskema om emnet, og derfor blev der udarbejdet et nyt anonymt spørgeskema, som er vist i bilag 7 og 8.

Spørgeskemaet blev udarbejdet med udgangspunkt i flere anbefalinger fra EVA (Danmarks Evalueringsinstitut), som bl.a. dækker neutrale, genkaldelige spørgsmål med udtømmende afbalancerede svarskalaer (EVA, 2017). Svarkategorierne var inspireret af Likert-skalaer, som typisk er nemme at forstå for respondenterne, hvilket øger undersøgelsens validitet. Derudover var de anvendte svarkategorier i overensstemmelse med anbefalingerne om tre til syv svarkategorier til de enkelte spørgsmål (Madsen B. S., 2017). Der indgik fire lukkede spørgsmål i spørgeskemaet, som havde til formål at give et let kvantificerbart overblik over elevernes holdninger. De lukkede spørgsmål blev suppleret med syv åbne spørgsmål, der skulle give et mere nuanceret indblik i elevernes holdninger. I forbindelse med udviklingen af det endelige spørgeskema blev en tidlig udgave af spørgeskemaet som tidligere nævnt testet i en klasse, der havde udført en pilottest af elevvejledningerne (bilag 9). Desværre var det ikke muligt at interviewe eleverne om spørgeskemaet bagefter. For at tjekke forståeligheden af de stillede spørgsmål blev en anden elevgruppe, der godt nok ikke havde udført biologiøvelsen, bedt om at læse spørgsmålene igennem og forklare, hvad de mente, de dækkede over. Gennemlæsningen førte til to revisioner. De lukkede spørgsmål i spørgeskemaet blev behandlet kvantitativt. De åbne spørgsmål blev behandlet ved at lave en kvalitativ indholdsanalyse, hvor jeg gjorde brug af en datadreven tilgang med tematisk analyse, som var inspireret af Braun og Clarke (2006). Først blev alle spørgeskemabesvarelser læst, hvorefter jeg foretog en tematisk kodning, som resulterede i fire temaer, der er formidlet i resultatafsnit 5.3. Elevernes besvarelse af spørgeskemaet og deres rapport blev som tidligere nævnt brugt som udgangspunkt for et kvalitativt interview.

4.3.3 Metode til interview

Formålet med interviewet var som tidligere nævnt dels at få et mere dybdegående indblik i, hvad eleverne havde gjort med feedbacken fra ChatGPT og hvorfor, dels at undersøge, hvordan eleverne opfattede ChatGPT som feedbackgiver. Interviewet foregik en til tre dage efter, at eleverne havde udført deres forsøg, så processen var i frisk erindring. Der var stor interesse blandt eleverne for at deltage i interviewdelen. Interviewgrupperne blev udvalgt, så der var tre grupper fra 3.g-klassen og tre grupper fra 1.g-klassen. I 3.g-klassen blev grupperne udvalgt, så der både var informanter med naturvidenskabelig og samfundsvidenskabelig studieretning repræsenteret. Dette kriterie var ikke relevant i 1.g-klassen, hvor eleverne endnu ikke var fordelt på studieretninger. Et andet kriterie for

udvælgelsen var, at der var flest mulige gruppemedlemmer til stede på interviewdagen, og at gruppemedlemmerne havde deltaget i det første modul med forsøgsdesign, hvor de havde brugt ChatGPT. Interviewet blev udført som et semistruktureret gruppeinterview ud fra en interviewguide (Brinkmann & Tanggaard, 2015) (bilag 5). I det analytiske arbejde med interviewene gjorde jeg brug af en datadreven tilgang med en tematisk analyse, som er inspireret af Braun og Clarke (2006) og vist på figur 6. Analysen resulterede i fem essentielle temaer med relevans for problemformuleringen.

Fase	**Beskrivelse af proces**
1. At blive bekendt med data.	Gennemlytning og transkribering af alle interviews (Bilag 12). Ideer og forslag til koder blev noteret.
2. Udarbejdelse af første forslag til koder.	Data blev organiseret i 17 forskellige koder, som blev brugt til at farvekode så mange udsagn som muligt i alle interviews.
3. Udarbejdelse af første forslag til temaer.	Koderne blev sorteret i 11 forskellige temaer, som alle havde relation til underspørgsmål fra problemformuleringen.
4. Gennemgang af temaer	Alle udsagn under de forskellige temaer blev læst igennem og enkelte blev omstruktureret. Dette førte til fem temaer. Validiteten af temaerne blev tjekket i forhold til hele datasættet.
5. Definition og navngivning af temaer	Alle udsagn til hvert tema blev læst igennem, og hvert tema blev tildelt en dækkende overskrift.
6. Rapportskrivning	De endelige temaer blev skrevet sammen til en rapport.

Figur 6. Beskrivelse af faser i analyse af interviews.

4.3.4 Begrænsninger og muligheder i det empiriske design

I de forrige afsnit har jeg skitseret udvalgte fordele og ulemper ved de valgte metoder. I dette afsnit vil jeg komme ind på udvalgte muligheder og begrænsninger i det empiriske design.

I analysen af elevrapporter blev der som tidligere nævnt anvendt en rubrik. Kvaliteten af rubrikken kunne være øget, hvis flere end jeg selv havde været inde over processen med at udarbejde den. Eleverne burde desuden være blevet introduceret til rubrikken og have diskuteret den, inden de gik i en gang med opgaverne, da ville have gjort det tydeligere for dem, hvad de blev evalueret på (Andrade, 2005). Derfor kan man argumentere for, at validiteten i undersøgelsen kunne øges, hvis rubrikken ikke kun var blevet brugt til den afsluttende kvalitetsscore. Derudover havde det været hensigtsmæssigt at lade flere end jeg selv kvalitetsscore elevrapporterne og derefter undersøge inter-rater reliability. I tilfælde af uoverensstemmelser mellem bedømmere havde det været oplagt at undersøge, om reliabiliteten kunne forbedres ved at foretage justeringer i rubrikken.

Hvad angår spørgeskemaundersøgelsen, så havde man kunnet øge validiteten ved at interviewe de elever, der havde deltaget i pilotundersøgelsen og udfyldt spørgeskemaet, i stedet for elever, der ikke havde afprøvet elevvejledningen. I interviewdelen kunne validiteten være øget ved at lade en kollega, der ikke skulle give eleverne karakterer, stå for at interviewe eleverne for at kompensere for, at det asymmetriske magtforhold mellem lærer og elev kan få elever til at give svar, som de forestiller sig, at læreren gerne vil høre.

Derudover er det en begrænsning i det empiriske design, at der kun indgik to klasser, ligesom man kan argumentere for, at det er uhensigtsmæssigt, at 1.g-klassen har arbejdet med gærforsøg, mens 3.g-klassen har arbejdet med smagsforsøg, fordi det gør det svært at sammenligne resultaterne på tværs af klassetrin. På den anden side så giver anvendelsen af forskellige forsøg i de to klasser mulighed for at sige noget om fælles tendenser, som ikke er afhængige af det enkelte forsøg eller klassetrin.

Endelig var ChatGPT´s rolle som feedbackgiver begrænset til to prompts fra elevvejledningen, men lagde ikke op til, at eleverne ændrede mange gange i deres prompts og stillede opfølgende spørgsmål, hvis der f.eks. var noget, som de ikke forstod. Derfor kan man argumentere for, at det valgte empiriske design ikke kan belyse ChatGPT´s fulde potentiale som feedbackgiver på elevers forsøgsdesign. Til gengæld kan undersøgelsen bruges til at sige noget om, hvordan formativ feedback fra ChatGPT påvirkede gymnasieelevers design af biologieksperimenter i et gruppearbejde i den undersøgte undervisningskontekst, herunder elevernes afkodning og implementering af feedback på forsøgsdesign fra ChatGPT, udvikling i kvalitet af forsøgsdesign og elevernes opfattelse af ChatGPT som feedbackgiver.

5. Resultater

5.1 Kvantitativ analyse af elevrapporter

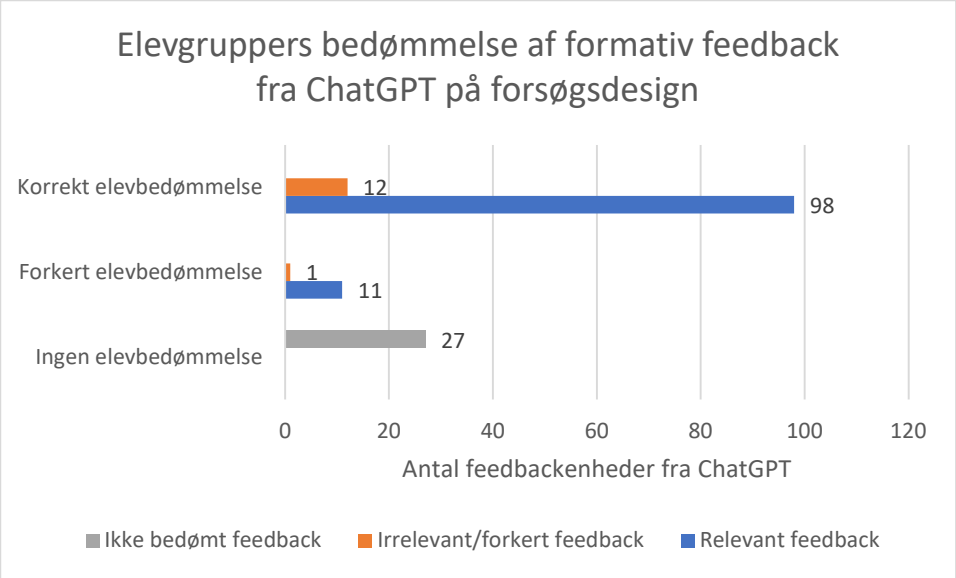
I det følgende afsnit bliver spørgsmålet om, hvordan elever opfatter ChatGPT som feedbackgiver belyst ved at præsentere resultater, der viser, 1) hvordan de bedømmer feedbacken. Derudover beskæftiger afsnittet sig med spørgsmålet om, hvad elever gør med formativ feedback fra ChatGPT med udgangspunkt i resultater, der viser, 2) i hvilken grad ændringerne i deres endelige forsøgsdesign kan sættes i relation til ChatGPT, og 3) hvilke dele af feedbacken fra ChatGPT de vælger at implementere. Endelig sætter resultater, der viser, 4) hvordan kvalitetsscorer af elevers forsøgsdesign ændrer sig efter feedback fra ChatGPT, fokus på spørgsmålet om, hvordan formativ feedback fra ChatGPT påvirker kvaliteten af elevers forsøgsdesign i biologi. De resultater, der bliver præsenteret, er baseret på kvantitativ analyse af 19 elevrapporter, hvor ti af rapporterne er fra 1.g og ni fra 3.g. I to af 1.g-rapporterne og en 3.g-rapport havde eleverne ikke foretaget nogle ændringer i deres forsøgsdesign efter at have modtaget formativ feedback fra ChatGPT. Derfor indgår disse rapporter ikke i resultaterne for punkt 2 og 3, der handler om implementering af feedback.

5.1.1 Elevers bedømmelse af formativ feedback fra ChatGPT

Resultaterne viser data fra 19 grupper, der har bedømt feedback fra ChatGPT, som de modtog i den første feedbackrunde, hvor de præsenterede ChatGPT for deres første forsøgsdesign. Den prompt, som grupperne brugte, var som tidligere nævnt designet, så de fik feedback fra ChatGPT i små afgrænsede enheder, som de blev bedt om at bedømme som henholdsvis relevante eller forkerte/irrelevante i den første runde af feedback. Relevant feedback er defineret som feedback, hvor eleverne fik hjælp til at forbedre forsøgsdesignet f.eks. ved at indtænke et kontrolforsøg, eller feedback, hvor eleverne fik bekræftet den fremgangsmåde, som de havde valgt. Irrelevant eller forkert feedback er f.eks. feedback, hvor ChatGPT foreslog udstyr, som eleverne ikke havde adgang til eller uhensigtsmæssige sikkerhedsforanstaltninger. En korrekt bedømmelse er kendetegnet ved, at relevant feedback fra ChatGPT blev vurderet som relevant, mens irrelevant/forkert feedback blev vurderet som irrelevant/forkert. En forkert bedømmelse er defineret som en bedømmelse, hvor relevant feedback fra ChatGPT blev bedømt som irrelevant/forkert eller en bedømmelse, hvor irrelevant/forkert feedback blev bedømt som værende relevant.

I datasættet indgår der 166 enheder med feedback fra ChatGPT, som grupperne er blevet bedt om at bedømme. Heraf er 17 feedbackenheder sorteret fra, da de var svære at kategorisere som enten relevante eller forkerte/irrelevante, hvilket f.eks. skyldtes, at nogle enheder indeholdt både relevante og forkerte eller irrelevante oplysninger. Derfor er resultaterne i dette afsnit baseret på 19 gruppers bedømmelse af i alt 149 feedbackenheder. Heraf er der 27 enheder, som grupperne enten undlod at bedømme eller noterede, at de havde svært ved at forstå. Grupperne modtog tilsammen 109 enheder med feedback fra ChatGPT, som kan kategoriseres som relevante, og 13 enheder, der kan kategoriseres som forkerte eller irrelevante. Figur 7A viser, at grupperne kunne identificere al irrelevant/forkert feedback på nær én enhed, og at de lavede en korrekt bedømmelse af 98 enheder med relevant feedback fra ChatGPT, men fejlbedømte 12 relevante enheder med feedback som værende irrelevante.

A)



B)

Feedbackkategori ChatGPT	Elevbedømmelse			Elever
	Korrekt	Forkert	Ikke bedømt	
Relevant	5,16 (1,71)	0,58 (0,77)	1,50 (1,62)	
Forkert/irrelevant	0,63 (0,83)	0,06 (0,24)		**Alle grupper**
Relevant	5,20 (1,87)	0,6 (0,52)	1,60 (1,35)	
Forkert/irrelevant	0,8 (0,92)	0 (0)		**1.g**
Relevant	5,11 (1,62)	0,56 (0,53)	1,38 (2,0)	
Forkert/irrelevant	0,44 (0,73)	0,11 (0,33)		**3.g**

Figur 7. A) 19 elevgruppers bedømmelse af feedback fra ChatGPT på deres første forsøgsdesign. N=149. B) Beskrivende statistik for elevers bedømmelse af feedback fra ChatGPT angivet som middelværdier (standardafvigelse) for hver enkelt gruppe.

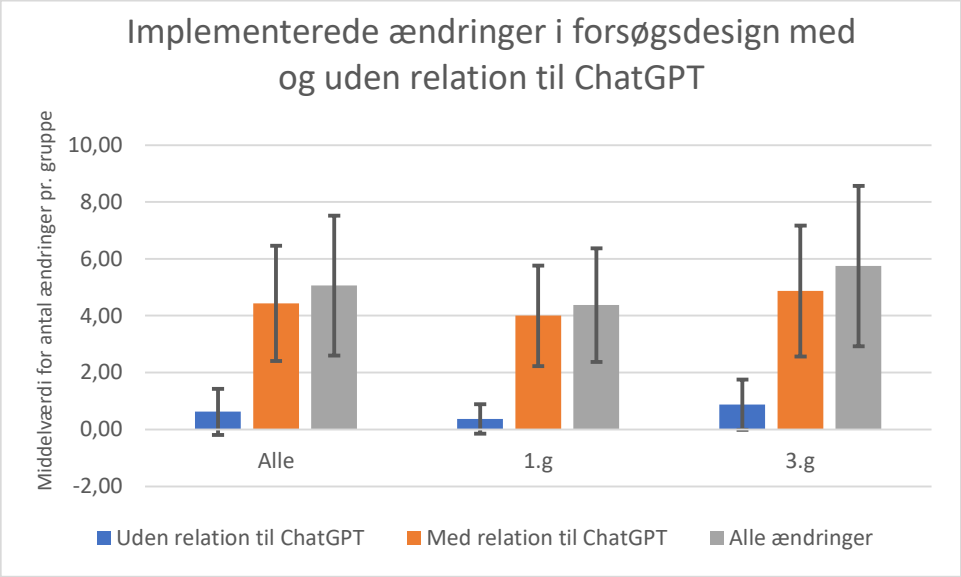
Figur 7B viser beskrivende statistik for elevernes bedømmelse af feedback fordelt pr. gruppe. Middelværdien for relevant feedback er 5,16 korrekte vurderinger pr. gruppe for alle grupper samlet. Som det fremgår af figur 7B, undlod 1.g-klasserne at bedømme mest feedback med en middelværdi på 1,6 undladte bedømmelser pr. gruppe sammenlignet med 1,38 undladte bedømmelser pr. gruppe for 3.g-grupperne. I 1.g undlod 80 procent af grupperne at bedømme en eller flere feedbackenheder, mens tallet er 44 procent for 3.g., hvor to grupper desuden stod for 81 procent af de undladte bedømmelser. Der ser ikke ud til at være et mønster i typen af feedbackenheder, som ikke blev bedømt.

Samlet set viser resultaterne, at størstedelen af den feedback, som eleverne modtog fra ChatGPT, var relevant for deres forsøgsdesign, og at grupperne i langt de fleste tilfælde kunne bedømme feedbacken korrekt. Samtidig er der en tendens til, at flere grupper i 1.g undlod at bedømme feedback sammenlignet med 3.g. I det følgende afsnit præsenteres resultater, der viser, om eleverne valgte at implementere den feedback, som de havde bedømt.

5.1.2 Implementering af feedback fra ChatGPT i forsøgsdesign

Som tidligere nævnt blev ændringer i forsøgsdesign analyseret ved at sammenholde gruppernes første forsøgsdesign med det sidste forsøgsdesign, hvorefter ændringerne blev sammenholdt med feedbacken fra ChatGPT. 16 grupper havde foretaget ændringer og indgår dermed i datagrundlaget, heraf er otte grupper fra 1.g og otte grupper fra 3.g. I alt er der registreret 81 ændringer, heraf kan ti ikke sættes i relation til ChatGPT. Det indikerer, at eleverne fortrinsvis har brugt ChatGPT´s formative feedback til at ændre i deres forsøgsdesign, hvilket understøttes af fordelingen på figur 8A, der ses på næste side, og viser det gennemsnitlige antal ændringer pr. gruppe og ændringernes relation til ChatGPT. Figur 8B viser, at 3.g-eleverne gennemsnitligt implementerede 5,59 ændringer pr. gruppe, mens 1.g-eleverne implementerede 4,38 ændringer pr. gruppe. Samme tendens gør sig gældende for ændringer med relation til ChatGPT, hvor 3.g-eleverne gennemsnitligt implementerede 4,71 ændringer pr. gruppe, mens tallet for 1.g er 4,00 ændringer pr. gruppe. Forskellen i implementering mellem 1.g og 3.g er ikke signifikant (Student t-test – se figurtekst til figur 8).

A)



B)

Ændringer	Alle	1.g	3.g
Uden relation til ChatGPT	0,63 (0,81)	0,38 (0,52)	0,88 (0,99)
Med relation til ChatGPT	4,44 (2,03)	4,00 (1,77)	4,71 (2,30)
Alle ændringer	5,06 (2,46)	4,38 (2,00)	5,59 (2,82)

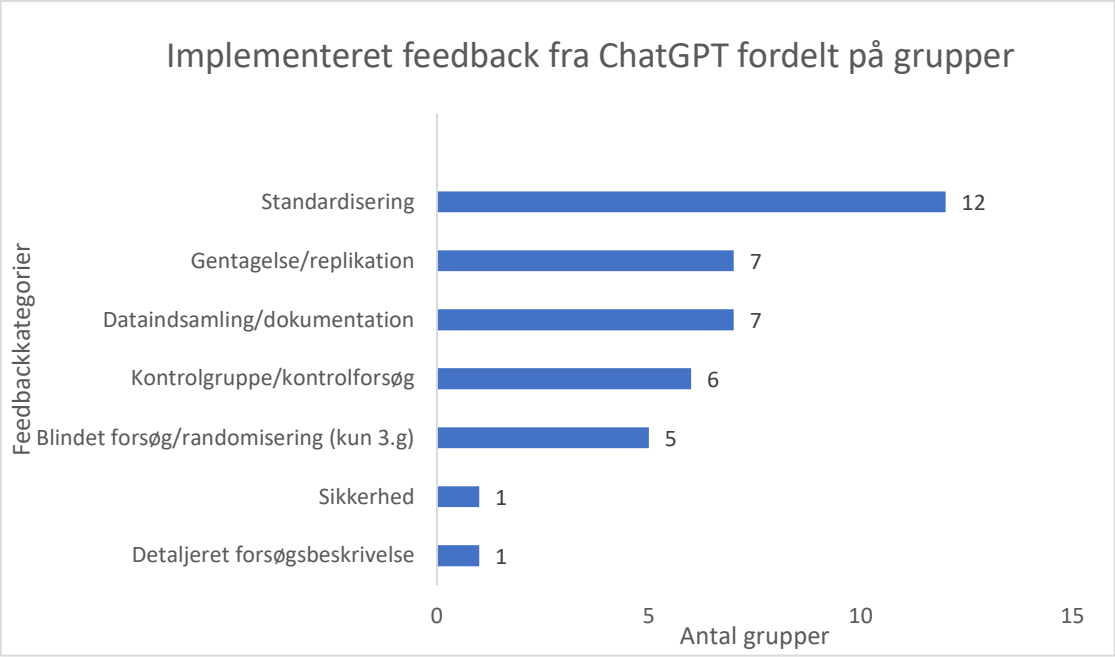
Figur 8. A) Gennemsnitligt antal implementerede ændringer pr. gruppe og ændringernes relation til ChatGPT for alle grupper og fordelt på klassetrin. B) Middelværdi og standardafvigelse for antal implementerede ændringer pr. gruppe og ændringernes relation til ChatGPT for alle grupper og fordelt på klassetrin.

Figurerne viser, at 3.g´erne har lidt flere ændringer relateret til ChatGPT end 1.g´erne. Da der ikke på forhånd er en forventning om, at den ene klasse ville bruge flere ændringer fra ChatGPT end den anden, er der anvendt i en to-sidet t-test til at undersøge, om der er forskel på implementering af ændringer fra ChatGPT. T-værdien er 0,85 numerisk, hvilket svarer til sandsynligheden 0,4. Dermed er der ikke signifikant forskel på de to klassetrin.

Samlet set viser resultaterne altså, at eleverne implementerede formativ feedback fra ChatGPT i deres endelige forsøgsdesign, og derfor er det interessant at undersøge, hvilke dele af feedbacken, de valgte at bruge.

5.1.3 Type af feedback fra ChatGPT, som implementeres i forsøgsdesign

16 grupper havde implementeret feedback fra ChatGPT i deres endelige forsøgsdesign. Den implementerede feedback fra ChatGPT blev inddelt i syv kategorier afhængig af, hvilke dele af forsøgsdesignet feedbacken dækkede. Figur 9, som ses på næste side, viser hvilke feedbackkategorier grupperne valgte at implementere. Selvom en gruppe kan have foretaget flere ændringer inden for samme kategori, er hver gruppe maksimalt registreret én gang i hver kategori.



Figur 9. Kategorisering af 16 gruppers implementering af feedback fra ChatGPT i deres endelige forsøgsdesign.

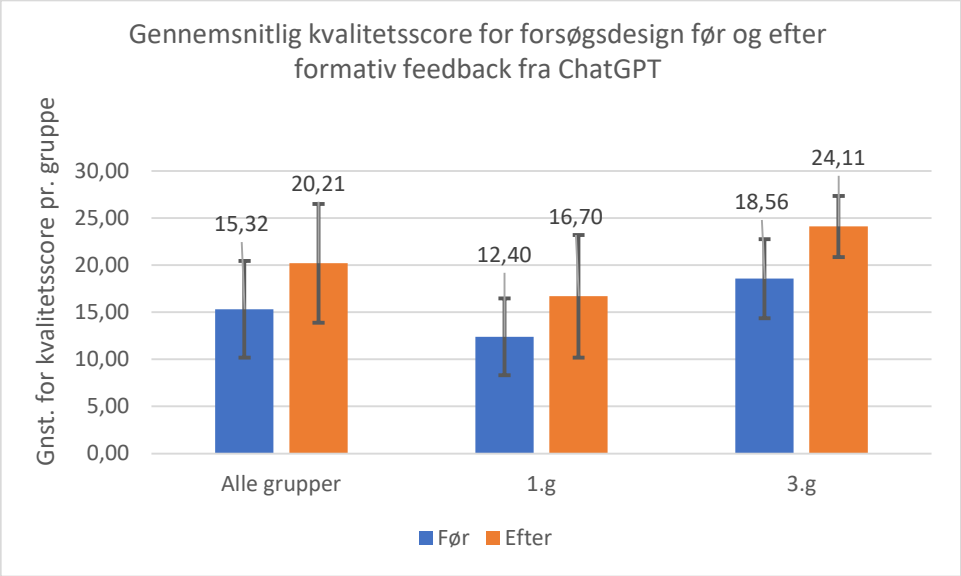
Resultaterne viser, at 12 grupper implementerede én eller flere ændringer, der har at gøre med standardisering af deres forsøgsdesign. Standardisering dækker over ændringer, der typisk kan sættes i relation til variabelkontrol. Et eksempel kan være temperaturkontrol i et forsøg om gærvækst eller mundskyldning mellem prøver i et smagsforsøg. Derudover implementerede syv grupper gentagelse eller replikation som en del af deres forsøgsdesign, hvilket f.eks. kan dække over, at grupperne valgte at opstille flere gærflasker ved hver enkelt temperatur i et forsøg, der undersøgte temperaturens indflydelse på gærvækst. Tilsvarende havde syv grupper implementeret ændringer, der har at gøre med dokumentation, så deres reviderede forsøgsdesign f.eks. indeholdt oplysninger om, hvornår og eventuelt hvordan data skulle noteres. Seks grupper implementerede ændringer med relation til kontrolforsøg, hvilket f.eks. kunne være en prøve kun med vand i et smagsforsøg. Kategorien ”Blindet forsøg/randomisering” skilte sig ud ved kun at være implementeret af 3.g-klasser, da det kun var relevant for deres forsøg. Man kan desuden argumentere for, at denne kategori kan sættes i relation til variabelkontrol og dermed indgå i tallene for standardisering. Kategorierne ”Sikkerhed” og ”Detaljeret forsøgsbeskrivelse” var begge kun blevet implementeret af en gruppe.

I det følgende afsnit præsenteres resultater for, hvordan den implementerede feedback påvirkede kvaliteten af gruppernes forsøgsdesign.

5.1.4 Kvalitet af forsøgsdesign før og efter feedback

Alle 19 grupper havde afleveret et første udkast til forsøgsdesign og et endeligt forsøgsdesign efter feedback fra ChatGPT, som begge blev tildelt en kvalitetsscore. Et godt forsøgsdesign kunne maksimalt opnå 30 point, og den samlede kvalitetsscore var baseret på seks delkvalitetsscorer, der kunne give op til fem point hver jf. rubrikken i bilag 6. Figur 10 viser data for kvalitetsscore før og efter feedback fra ChatGPT.

A)



B)

	Alle	1.g	3.g
Procentvis forbedring	33,94 (28,60)	32,15 (24,34)	35,93 (29,39)

Figur 10. A) *Den gennemsnitlige kvalitetsscore* pr. gruppe *før og efter feedback fra ChatGPT for alle grupper, 1.g og 3.g.*

Der indgik 19 observationer for kvalitetsscore før feedback og efter feedback for de enkelte elevrapporter, og derfor blev der anvendt en parret t-test. Da man vil forvente en højere score efter brug af ChatGPT, blev der brugt en ensidet test. For alle grupper samlet var forskellen signifikant med p-værdien < 0,001, for 1.g var forskellen signifikant med p-værdien < 0,002, og for 3.g var forskellen signifikant med p-værdien < 0,005. Sammenlignes ændringerne for de to klassetrin med en to-sidet t-test, opnås t-værdien 0,66 numerisk svarende til p-værdien 0,52, og derfor er der ikke forskel på de to klassetrin.

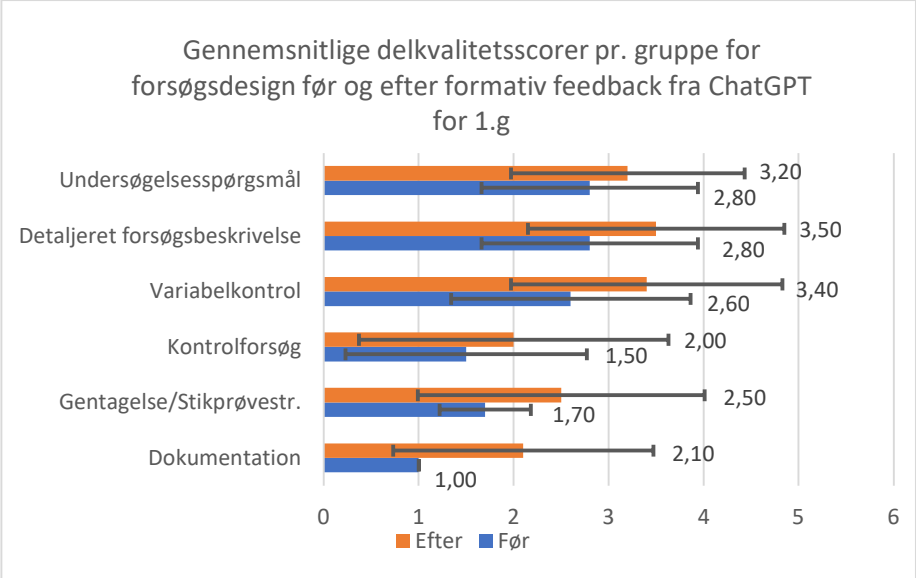
B) Middelværdi og standardafvigelse for procentvis forbedring i kvalitetsscore i forsøgsdesign før og efter feedback fra ChatGPT.

Figur 10A viser, at den gennemsnitlige kvalitetsscore stiger efter feedback fra ChatGPT for alle grupper samlet og fordelt på klassetrin. I alle tilfælde var stigningen i kvalitetsscore efter feedback fra ChatGPT signifikant med en p-værdi under 0,001 for alle grupper samlet, en p-værdi under 0,002 for 1.g og en p-værdi under 0,005 for 3.g (Parret t-test – se figurtekst 10). Resultaterne viser desuden, at den gennemsnitlige kvalitetsscore er på 15,32 for 1.g-grupperne før feedback fra ChatGPT og højere for 3.g-grupperne, der har en gennemsnitlig kvalitetsscore på 18,56. Samme tendens gør sig gældende efter feedback fra ChatGPT, hvor den gennemsnitlige kvalitetsscore pr. gruppe er 16,70 for 1.g mod 24,11 for 3.g.

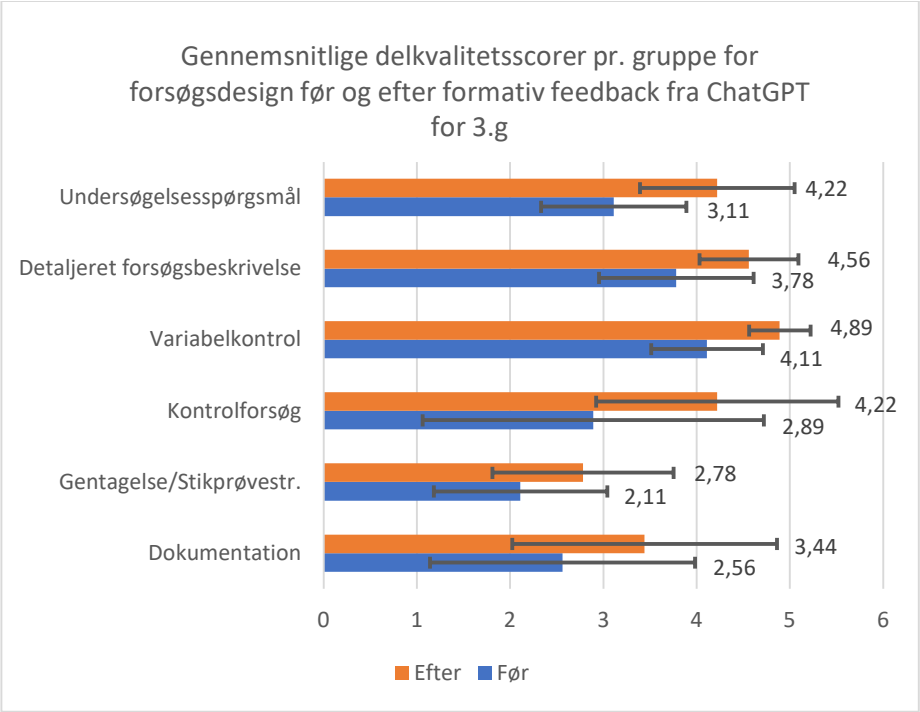
Som man kan se på figur 10B forbedrer alle grupperne gennemsnitligt deres kvalitetsscore med 33,94 procent efter feedback fra ChatGPT. 3.g-grupperne har en gennemsnitlig forbedring på 35,93 procent, mens 1.g-grupperne ligger lidt lavere med 33,58 procent, men forskellen mellem 1.g og 3.g er ikke signifikant (Student t-test – se figurtekst 10).

Endelig viser resultaterne, at den gennemsnitlige stigning i kvalitetsscore pr. gruppe efter feedback fra ChatGPT er fordelt på alle delkvalitetsscorer i både 1.g (figur 11A) og 3.g (figur 11b).

11A)



11B)



Figur 11. Gennemsnitlige kvalitetsscore pr. gruppe *fordelt på delkvalitetsscorer før og efter feedback fra ChatGPT vist for 1.g i figur A og 3.g i figur B). Tallene angiver gennemsnitscoren pr. gruppe for de enkelte kategorier.*

Samlet set viser resultaterne altså en gennemsnitlig og signifikant stigning i kvalitetsscore efter feedback fra ChatGPT for alle grupper samlet. Selvom 3.gernes forsøgsdesign har en højere kvalitetsscore og procentvis forbedring sammenlignet med 1.gerne, kan forskel mellem klassetrin ikke vises statistisk. På begge klassetrin ser den gennemsnitlige stigning i kvalitetsscore efter feedback fra ChatGPT desuden ud til at være baseret på stigninger i alle delkvalitetsscorer.

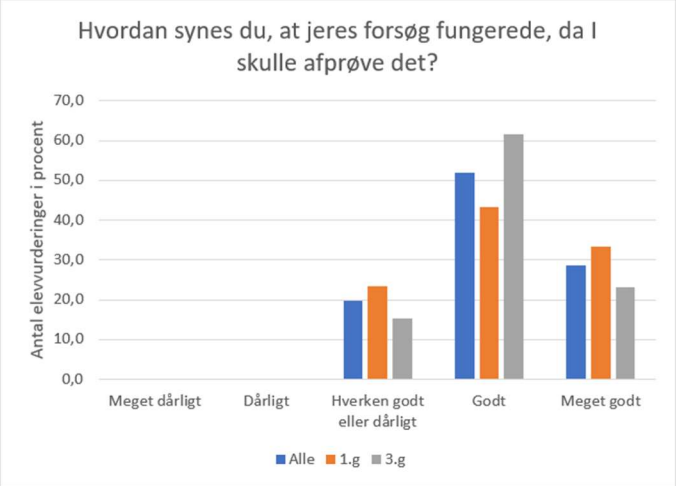
5.2 Kvantitativ analyse af spørgeskemaer

De resultater, der bliver præsenteret i de følgende afsnit, er baseret på en undersøgelse, hvor 60 elever er blevet bedt om at besvare et spørgeskema, 30 elever i 1.g og 30 elever i 3.g. Fire elever i 3.g undlod at besvare spørgeskemaet, hvilket giver en samlet svarprocent på 93 % for alle elever samlet, 100 % for 1.g og 87% for 3.g. De lukkede spørgsmål fra spørgeskemaet danner grundlag for en kvantitativ analyse, mens de åbne spørgsmål er blevet analyseret kvalitativt. Resultaterne i dette afsnit er baseret på en kvantitativ analyse og viser, 1) hvordan eleverne vurderer kvaliteten af det forsøg, som de har udført. Disse data er relevante, fordi en sådan vurdering kan betragtes som en indikator for elevernes oplevelse af kvaliteten af deres forsøgsdesign. Derudover bliver

elevernes opfattelse af ChatGPT som feedbackgiver belyst ved at præsenterer resultater, der viser, i hvilken grad eleverne oplever feedback fra ChatGPT som 2) brugbar, 3) til at stole på og 4) værd at bruge næste gang, de skal designe forsøg i biologi. De kvantitative data er præsenteret i figur 12.

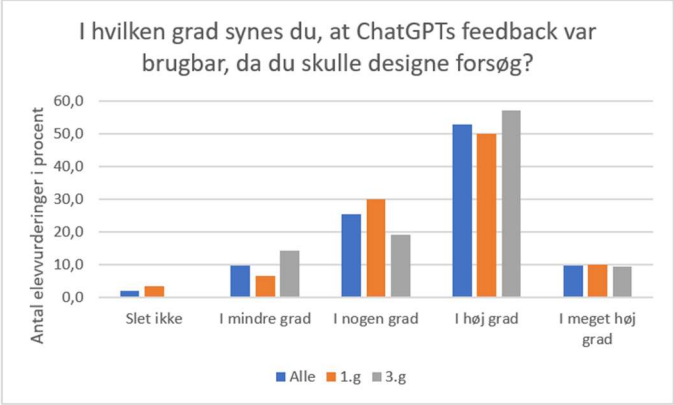
12A

Svarkategori	Pct.	Kumulativ pct.
Meget godt	29	29
Godt	52	81
Hverken godt eller dårligt	20	101
Dårligt	0	101
Meget dårligt	0	101
I alt	101	



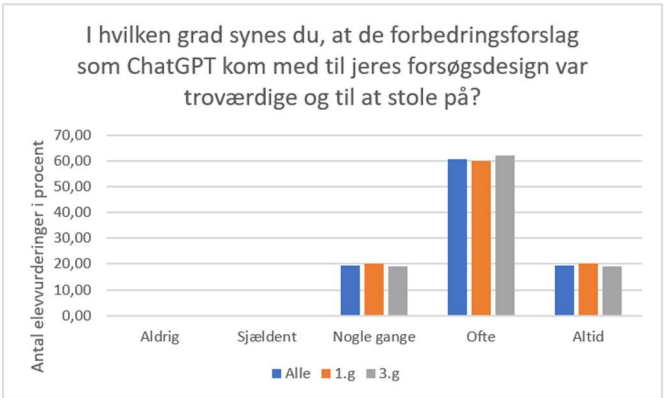
12B

Svarkategori	Pct.	Kumulativ pct.
I meget høj grad	10	10
I høj grad	53	63
I nogen grad	26	89
I mindre grad	10	99
Slet ikke	2	101
I alt	101	



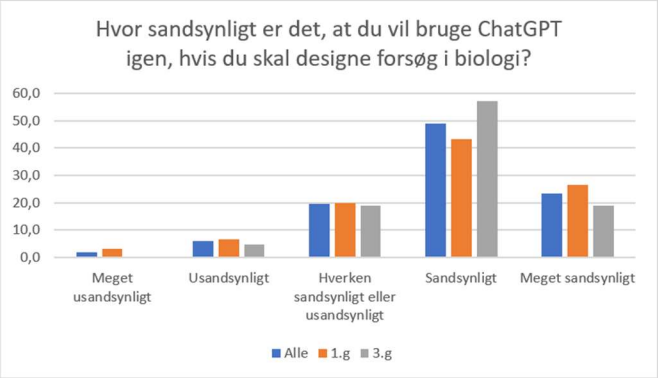
12C

Svarkategori	Pct.	Kumulativ pct.
Altid	20	20
Ofte	61	81
Nogle gange	20	101
Sjældent	0	101
Aldrig	0	101
	101	



12D

Svarkategori	Pct.	Kumulativ pct.
Meget sandsynligt	24	24
Sandsynligt	49	73
Hverken sandsynligt eller usandsynligt	20	93
Usandsynligt	6	99
Meget usandsynligt	2	101
I alt	101	



Figur 12. Elevvurderinger angivet i procent af A) forsøgsdesign efter udførelse af forsøget B) brugbarheden af ChatGPT´s feedback på forsøgsdesign C) troværdigheden af ChatGPT´s feedback og D) sandsynligheden for, at de vil bruge feedback fra ChatGPT, hvis de skal designe forsøg igen. Vurderingerne er angivet grafisk for alle grupper samlet og fordelt på klassetrin, mens de tal, der er vist i tabellerne til venstre, kun er for alle grupper samlet. Den samlede procentsats overstiger 100 procent, hvilket skyldes afrunding.

Data for elevernes vurdering af deres forsøgsdesign er vist i figur 12A. Den kumulative fordeling for alle elever samlet viser, at 81 procent af eleverne svarede ”Godt” eller ”Meget godt”, når de blev spurgt, hvordan deres forsøg fungerede, da de skulle afprøve det. Medianen er ”Godt”, som opnås ved 81 pct. af fordelingen. Samme tendens fremgår af den grafiske fremstilling, der viser en fordeling, som er tilnærmelsesvis symmetrisk omkring ”Godt” med en svag højreskævhed, hvilket også ser ud til at gøre sig gældende, hvis data opdeles i 1.g og 3.g.

Figur 12B viser resultaterne for elevernes vurdering af brugbarheden af ChatGPT´s feedback på deres forsøgsdesign. Resultaterne er baseret på data fra 51 respondenter, 30 fra 1.g og 21 fra 3.g, da besvarelserne fra fem respondenter i 3.g er sorteret fra, fordi de ikke var til stede i det 1. modul, hvor eleverne arbejdede med feedback fra ChatGPT. Det fremgår af den kumulative fordeling for alle elever samlet, at 63 procent af eleverne svarede ”I høj grad” eller ”I meget høj grad”, når de blev spurgt, i hvilken grad de syntes, at ChatGPT´s feedback var brugbar, da de skulle designe forsøg. Medianen er ”I høj grad”, som opnås ved 63 pct. af fordelingen. En grafisk fremstilling af data viser en venstreskræv fordeling, som også træder frem, hvis data opdeles i 1.g og 3.g

I spørgeskemaet blev eleverne desuden spurgt, ”I hvilken grad oplevede du, at de forbedringsforslag, som ChatGPT kom med til jeres forsøgsdesign, var troværdige og til at stole på?”. Her viser resultaterne på figur 12C, at 81 procent af eleverne svarede ”Ofte” eller ”Altid”. Medianen, som er ”Ofte”, opnås ved 81 pct. af fordelingen. Den grafiske fremstilling af data viser en fordeling, der er tilnærmelsesvis symmetrisk omkring ”Ofte” for alle data samlet og ved

opdeling i 1.g og 3.g. Ligesom med de foregående data indgår der kun svar fra respondenter, som var til stede i det 1. modul, hvilket også gør sig gældende for de følgende data.

Resultaterne fra besvarelsen af spørgsmålet, ”Hvor sandsynligt er det, at du vil bruge ChatGPT igen, hvis du skal designe forsøg?”, er vist i figur 12D. Den kumulative fordeling for alle elever samlet viser, at 73 procent af eleverne svarede ”Sandsynligt” eller ”Meget sandsynligt” med medianen ”Sandsynligt”. En grafisk fremstilling af data viser en venstreskæv fordeling for alle data samlet, og hvis data opdeles i 1.g og 3.g

De kvantitative analyse af spørgeskemaer viser, at de fleste eleverne oplevede feedbacken fra ChatGPT som brugbar og troværdig. Derudover syntes de fleste, at deres forsøg fungerede godt, ligesom de fandt det sandsynligt, at de ville bruge ChatGPT igen til at designe forsøg. I det følgende afsnit bliver resultaterne fra den kvalitative analyse af spørgeskemabesvarelserne præsenteret.

5.3 Kvalitativ analyse af spørgeskemaer

Resultaterne i dette afsnit forholder sig til spørgsmålet om, hvordan formativ feedback påvirker kvaliteten af elevers forsøgsdesign ved at præsentere resultater fra spørgeskemaundersøgelsen, der beskæftiger sig med 1) forbedringer i forsøgsdesign i relation til ChatGPT. Spørgsmålet om, hvordan elever opfatter ChatGPT som feedbackgiver på forsøgsdesign i biologi, bliver desuden belyst ud fra tre centrale temaer, som handler om, 2) at ChatGPT´s rådgivning ikke altid tager højde for elevernes praktiske virkelighed, men 3) er brugbar for de fleste. Samtidig reflekterer eleverne over, hvilke krav feedback fra ChatGPT stiller til dem med fokus på 4) vigtigheden af at være udførlig og forholde sig kritisk. Elevernes svar er præsenteret i bilag 10 for 1.g og bilag 11 for 3.g. I det følgende betyder en henvisning til R.1.2, at respondenten er fra 1.g og har nr. 2 i oversigten over respondentsvar i bilag 10. Tilsvarende vil R. 3.2 henvise til en respondent fra 3.g, som har nummer 2 i svaroversigten i bilag 11.

5.3.1 Tema 1: ChatGPT´s feedback kan forbedre forsøgsbeskrivelse og design

Eleverne nævnte, at feedback fra ChatGPT hjalp dem med at gøre deres forsøg mere ”præcist” (18 respondenter), og at deres ”forsøgsdesign blev mere detaljeret” (R. 1.2), fordi forsøgsvejledningen blev ”meget mere beskrivende” (R 1.11). Tre respondenter svarede desuden, at feedbacken

påvirkede strukturen. Den øgede præcision skyldtes bl.a., at eleverne ”undgik flere fejlkilder” (R 3.4), fordi forslag fra ChatGPT hjalp med at ”udpege fejlkilder” (R 1.29) og gav ”gode råd til at mindske fejlkilder” (R 3.26). I den forbindelse er det interessant, at syv respondenter gav udtryk for, at feedbacken fik dem til at implementere et kontrolforsøg. Derudover svarede syv respondenter, at ChatGPT´s forslag om gentagelser var relevante for deres forsøgsdesign. Endelig gav feedbacken anledning til, at eleverne skulle forholde sig til forskellige aspekter af variabelkontrol. Blandt 1.g-eleverne nævnte fem respondenter variablen ”temperatur”, ligesom en respondent svarede, at de udførte eksperimentet ”simultant så produktet ikke ville have usikkerheder” (R. 1.7). I 3.g var det særligt ”randomisering” (5 respondenter) og forskellige former for blindtest (4 respondenter), der blev nævnt. Selvom ovenstående tyder på, at eleverne brugte feedback fra ChatGPT til at forbedre deres forsøgsdesign, var der også flere elever, der forholdt sig kritisk til ChatGPT som feedbackgiver.

5.3.2 Tema 2: ChatGPT´s feedback tager ikke nok højde for elevernes praktiske virkelighed

50 respondenter svarede, at der var feedback fra ChatGPT, som de valgte ikke at bruge, bl.a. fordi de ikke havde tid til det (8 respondenter). Blandt 1.g eleverne nævnte 12 respondenter, at de ikke brugte ChatGPT´s feedback, fordi den ”foreslog at bruge instrumenter som” de ”ikke havde til rådighed” (R. 1.4), ligesom tre respondenter gav udtryk for, at ChatGPT´s feedback om sikkerhed var overdreven i forhold til situationen. En skrev ”Den sagde et tidspunkt at vi skulle huske briller siden vi arbejde med kemikalier (vi skulle bare lave en dej af mel, vand og gær)” (R. 1.8). Blandt 3.g-respondenterne svarede fire, at de ikke brugte feedbacken, fordi det krævede for mange forsøgspersoner. Samtidig mente fire 3.g-respondenter, at ChatGPT´s feedback om etik ”ikke rigtig gav mening” (R. 3.1.) i situationen. En skrev ”Den forslog blandt andet at søge etisk godkendelse hvilket vi ikke følte var relevant eftersom vi i forsøget kun har tænkt os at bruge vand og sukker.” (R. 3.3). 15 respondenter fra 3.g nævnte desuden, at de var nødt til at justere på koncentrationerne af sukker og citron, da de skulle udføre deres smagsforsøg, bl.a. fordi deres ”koncentrationer var alt for små” (R. 3.7) eller ”alt for høje” (R. 3.9). Endelige svarede 12 respondenter fordelt på 1.g og 3.g, at de ikke valgte at gå videre med dele af ChatGPT´s feedback, fordi de allerede havde taget højde for det i deres design ”Mange af tingene gjorde vi i forvejen, men det havde den ikke opfaffet.” (opfattet) (R. 3.26). Samtidig nævnte otte respondenter, at dele af feedbacken ikke blev brugt, fordi den var irrelevant. Samlet set viser resultaterne, at eleverne syntes, at ChatGTP ikke tog nok højde

for deres praktiske virkelighed, men resultaterne i næste afsnit tyder på, at feedbacken alligevel var brugbar.

5.3.3 Tema 3: ChatGPT´s feedback er brugbar for de fleste

Adspurgt om der var noget af ChatGPT´s feedback, som respondenterne havde med i deres forsøgsdesign, som de ikke ville tage med, hvis de skulle gentage forsøget, svarede 27 nej, mens en ville bruge mere gær (R.1.19), og en anden ville undlade at måle kuldioxid (R. 1.26). Resten af respondenterne forholdt sig ikke til spørgsmålet. Andre kaldte ChatGPT ”et fedt virkemidler” (virkemiddel) (R. 1.9), ”et godt værktøj som giver flere ideer” (R. 3.8) og en ”mega god hjælp” (R. 1.3), ligesom en syntes ”at chat gpts svar var optimalt” (R. 1.5). En respondent opsummerede brugen af feedback fra ChatGPT med følgende kommentar: ”Den gav os flere forbedringer som vi kunne bruge til vores forsøg, der gjorde vi lidt nemmere kunne komme i mål med havde vi havde i tankerne.” (R 1.12), mens en anden svarede: ”Det føltes lidt som en ekstra lærer og det gav nogle gode pointer, som vi ellers ikke havde inddraget, samtidig med at man ikke som sådan følte, at man "snød"” (R. 3.12). Selvom der var fem respondenter fra 1.g og en fra 3.g, som gav udtryk for, at de ændringer, de lavede, ikke havde den store effekt på deres forsøgsdesign, viser resultaterne, at ChatGPT´s feedback blev oplevet som brugbar af de fleste. I det følgende afsnit præsenteres resultater, der viser, hvordan eleverne mente, at man bør forholde sig til ChatGPT som feedbackgiver.

5.3.4 Tema 4: ChatGPT´s feedback kræver, at elever er udførlige og forholder sig kritisk

18 respondenter gav udtryk for, at brug af ChatGPT som feedbackgiver krævede, at de forholdt sig kritisk til den feedback, som de brugte. Otte respondenter brugte formuleringen ”kritisk”, en ”skepisk” (skeptisk) (R. 1.23), og en tredje mente, at man skulle være opmærksom på ”ikke at følge den 100%” (R. 3.10). Samtidig mente 10 respondenter, at det var nødvendigt at være udførlig, når man bad ChatGPT om feedback, hvilket blev beskrevet med formuleringer som, at man skulle være ”konkret” (3 respondenter), huske at være ”specifik” (R.1.12 og 1.24) og have mange ”detaljer” (R.1.8) med.

Samlet set viser den kvalitative analyse af spørgeskemaundersøgelsen, at de fleste elever oplevede ChatGPT´s feedback som værende brugbar og syntes, at den kunne bruges til at forbedre

forsøgsbeskrivelse og design, selvom de ikke mente, at ChatGPT tog nok højde for deres praktiske virkelighed. Samtidig mente flere elever, at det var vigtigt at være udførlig og forholde sig kritisk til ChatGTP´s feedback. I det følgende afsnit bliver resultaterne fra den kvalitative analyse af elevinterviews præsenteret.

5.4 Kvalitativ analyse af interviews

Denne del af resultatafsnittet beskæftiger sig med spørgsmålet om, hvad elever gør med formativ feedback fra ChatGPT, når de skal designe forsøg med udgangspunkt i to centrale temaer fra de seks elevinterviews. Temaerne dækker elevernes brug af ChatGPT 1) til at skabe overblik og give nye ideer til deres forsøgsdesign, og 2) til at forbedre deres forsøgsdesign og beskrivelse. I det følgende afsnit bliver spørgsmålet om, hvordan elever opfatter ChatGPT som feedbackgiver på forsøgsdesign i biologi, desuden belyst ud fra temaer, som handler om ChatGPT´s feedback i relation til elevernes 3) niveau, 4) praktiske virkelighed og 5) opfattelse af målrettethed og troværdighed i feedbacken. Transkriberinger af elevernes interviews er præsenteret i bilag 12. I det følgende betyder en henvisning til T. 1.2, at citatet er fra et interview i 1.g med gruppe nummer 2 og en elev med forbogstav T i det elevvalgte kodenavn. Tilsvarende vil V. 3.2 henvise til et interview i 3.g med gruppe nummer 2 og en elev med forbogstav V i kodenavnet. I tilfælde, hvor der refereres til fund i gruppen som helhed, er forbogstavet udeladt, og dermed betyder (1.3), at der henvises til et interview i 1.g med gruppe 3.

5.4.1 Tema 1: ChatGPT´s feedback kan skabe overblik og give nye ideer til forsøgsdesign

Eleverne gav udtryk for, at feedback fra ChatGPT bidrog til at skabe ”overblik” (1.3, 3.2) og hjalp med at gøre arbejdsprocessen med forsøgsdesign mere ”overskuelig” (3.1). I den sammenhæng er det interessant, at to grupper nævnte, at det var en god tjekliste (1.2, 3.3). En elev sagde ”Jeg tror, at det er godt til lige at tjekke bagefter, om man har alt med, fordi man kan jo godt lige lave nogle fejl, så glemmer man en ting eller noget” (T. 1.2). Derudover nævnte en anden elev, at ChatGPT kom med feedback, som de ”ikke havde tænkt på” (A. 3.3), og den holdning delte alle andre grupper. En elev mente desuden, at feedbacken bidrog med nye ideer og vinkler, som man ikke altid ville få fra en lærer, og sagde, at en lærer typisk ikke ”giver dig nye forslag til, altså sådan hey du kunne også gøre det her eller prøve måske at lave en anden del af forsøget, hvor du bliver til at tage en nye variabel eller noget. Og det tror jeg måske, at den var meget god til.” (V. 3.2). Samlet

set viser analysen, at eleverne bl.a. brugte feedback fra ChatGPT til at skabe overblik, som tjekliste og til at overveje nye elementer i deres forsøgsdesign. Spørgsmålet er, hvilke dele af feedbacken de oplevede som brugbar.

5.4.2 Tema 2: ChatGPT´s feedback kan forbedre forsøgsbeskrivelse og design

I fem interviews gav eleverne udtryk for, at ChatGPT kunne bruges til at forbedre det grundlæggende forsøgsdesign. En elev forklarede det ved at sige, at ChatGPT ”kan bidrage med, hvordan et sådan overordnet forsøg ligesom skal sættes op” (A. 3.3). Tre af grupperne fremhævede ChatGPT´s forslag om kontrolgrupper som noget positivt og beskrev det med udtryk som ”kontrolforsøg” (1.1, 3.3) og ”kontrolgruppe” (3.1). De tre andre grupper mente, at ChatGPT´s forslag, om kontrolforsøg var irrelevante, fordi de allerede havde indtænkt det, men en gennemgang af elevrapporterne viste, at to af grupperne (1.2; 1.3) ikke havde indtænkt et korrekt kontrolforsøg, mens den tredje gruppe ikke havde skrevet det ind i deres endelige forsøgsdesign, da de først havde ”implementeret det på dagen” (H. 3.2), hvor de udførte forsøget.

I tre grupper gav eleverne desuden udtryk for, at de godt kunne se relevansen af ChatGPT´s råd om gentagelser (1.2) eller replikation (1.2, 1.3, 3.1). En elev sagde ”Vi brugte ligesom den der gentagelse. Den synes jeg i hvert fald var ret god, fordi […] hvis der var sådan, vi havde kun tre hver, og der var en fejl med en af dem, så ville vi måske ligesom ikke få det rigtige svar, som når vi nu havde lavet ekstra.” (P. 1.1). Citatet viser også, at ChatGPT´s feedback kunne bruges til at få mere præcise resultater, hvilket to andre grupper var enige i, da de udtalte, at feedbacken mindskede ”fejl” (1.2) og ”fejlkilder” (3.3).

Endelig gav tre grupper (1.3, 3.2, 3.3) udtryk for, at ChatGPT´s feedback påvirkede kvaliteten af deres forsøgsbeskrivelse positivt, hvilket illustreres af følgende udsagn, hvor en elev sagde, at feedbacken ”hjalp også lidt med at få struktur” (T. 1.3), og at de ”satte flere detaljer på” (T. 1.3). En anden elev sagde ”Jeg synes nemlig, altså meget af det her havde vi så ikke kunne skrive uden chatbotten, eller vi havde i hvert fald glemt det, men når vi ligesom selv går ind og bearbejder det, den har sagt, så synes jeg det måske er blevet bedre end vores udgangspunkt uden den (E. 3.3.), hvilket vidner om en mere generel positiv påvirkning på deres forsøgsbeskrivelse og -design. Ovenstående viser, at feedbacken fra ChatGPT havde potentiale til at forbedre elevernes forsøgsdesign, men resultaterne i næste afsnit viser, at det var svært at ramme det rigtige niveau i feedbacken.

5.4.3 Tema 3: ChatGPT´s feedback har svært ved at ramme elevernes niveau

En elev påpegede, at ChatGPT´s feedback blandede ”noget fra både høje niveauer og lave niveauer” (H. 3.2). I den sammenhæng er det interessant, at interviewundersøgelsen viste, at fire grupper (1.1, 1.3, 3.1, 3.3) havde svært ved at forstå dele af feedbacken, f.eks. udtalte en elev, at ”hvis du ved, hvad anova og t-test er, så tror jeg godt, man forstår, hvad den siger, men når vi ikke er helt sikre på, hvad det er, så er det lidt sort.” (E. 3.3). Samtidig gav en 3.g-elev med naturvidenskabelig baggrund udtryk for, at dele af feedbacken var på et lavt niveau og udtalte, at ”nogle af de ting, den nævnte var meget basic, altså, at man sku lave en dataindsamling, det tænker jeg er meget intuitivt.” (H. 3.2). Samme elev tilføjede, at de ikke fik ”nævnt præcis, hvad det er for et niveau det ligger på” (H. 3.2), hvilket er relevant i forhold til udformningen af promptskabelonen. ChatGPT´s feedback blev ikke kun kritiseret for ikke altid at ramme elevernes niveau, flere elever mente også, at den manglede en kobling til virkeligheden.

5.4.4 Tema 4: ChatGPT´s feedback tager ikke nok højde for elevernes praktiske virkelighed

I interviewet gav eleverne udtryk for, at de på nogle punkter mente, at ChatGPT havde ”en mangel på situationsfornemmelse” (A. 1.3), og at ”man kunne godt mærke, at […] det praktiske ikke lige var på plads (T. 1.3). Det kom til udtryk på flere forskellige niveauer. I den prompt fra elevvejledningen, som eleverne brugte, da de designede deres forsøg, stod der f.eks., hvilke ressourcer de havde til rådighed (bilag 2 og 3). Alligevel sagde en elev ”Vi havde en ressourceliste, som vi satte ind også ikk, og den går lidt udenom.” (T. 1.3). Problematikken gjaldt alle interviewede 1.g-grupper, som fik feedback, der krævede en gasmåler (1.1, 1.2) og et ”pH-meter” (1.3), som de ikke havde adgang til. For alle 3.g klasserne var problemet derimod, at de manglede rådgivning om, hvilke mængder sukker og citronsaft de skulle bruge for at finde tærskelværdien for de to smage, hvilket illustreres af følgende citat: ”Men altså det eneste jeg tror, man måske ikke skulle bruge ChatGPT til, var nok mængder” (A. 3.3). Derudover gav en gruppe udtryk for at få sikkerhedsinstrukser af ChatGPT, som de fandt irrelevante i forhold til situationen. De skulle arbejde med ”gær og vand og sukker” (A. 1.1) og blev bedt om at bruge beskyttelsesbriller og handsker. Tilsvarende mente en anden gruppe ikke, at ChatGPT´s forslag om ”etiske overvejelser” (3.2) var relevant for deres forsøg. Tre grupper (1.2, 3.1, 3.2) gav desuden udtryk for, at ChatGPT kom med forbedringsforslag, som de allerede havde indtænkt i deres forsøgsdesign. En elev sagde f.eks. ”Vi havde allerede tænkt, at det var sådan, vi skulle gøre” (L. 3.1). Selvom resultaterne i dette

afsnit viser, at ChatGPT´s feedback ikke altid var relevant, viser resultaterne i næste afsnit, at den indeholdt positive elementer.

5.4.5 Tema 5: ChatGPT´s feedback er målrettet det enkelte forsøg og troværdig

Tre grupper sagde, at de var glade for, at den feedback, som de modtog fra ChatGPT, ikke bare var en generel liste med ting, de skulle huske, når de designede forsøg (1.1, 3.1, 3.3). Det lader til at være en fordel, at ChatGPT´s feedback passer til netop deres forsøg, hvilket en elev forklarede med ordene, ”det var også målrettet til vores forsøg […] altså det var nemmere sådan at skrive om, fordi den ligesom havde lavet det til det, vi skulle undersøge” (B. 1.1). Desuden tilføjede to grupper (3.1, 3.3), at det var positivt, at den kom ”med nogle begrundelser for, hvorfor man gør det” (E. 3.3). Derudover var der i alle grupperne mindst en elev, der gav udtryk for, at man kunne stole på ChatGPT, når det gjaldt forsøgsdesign i biologi. En elev sagde f.eks., at ”tingene den sagde […] egentlig virkede troværdige nok” (C. 3.2), mens en anden brugte formuleringen ”Alt sammen er meget troværdigt” (T. 1.2). Selvom eleverne mente, at feedbacken var troværdig, så gav nogle også udtryk for, at man skulle forholde sig kritisk, eller som en sagde ”Selvfølgelig skal man ikke bare tage det blindt” (E. 3.3), hvilket understreger elevernes behov for selv at være inde over processen med at designe forsøg.

Analysen af de seks gruppeinterview viser, at flere elever oplevede, at ChatGPT´s feedback hjalp med at skabe overblik og give nye ideer til forsøgsdesign, ligesom den kunne bruges til at forbedre forsøgsbeskrivelse og design. Selvom feedback fra ChatGPT i flere tilfælde havde svært ved at ramme elevernes niveau og ikke tog nok højde for deres praktiske virkelighed, så havde feedbacken den fordel, at den var målrettet deres forsøg. Samtidig blev feedbacken oplevet som troværdig, men eleverne vurderede stadig, at det var vigtigt, at man ikke udelukkende forlod sig på ChatGPT. I det følgende afsnit bliver resultaterne fra interviewanalysen diskuteret sammen med undersøgelsens øvrige analyser.

6. Diskussion

I dette afsnit samles undersøgelsens resultater i en diskussion af, hvordan formativ feedback fra ChatGPT bliver brugt af elever og påvirker kvaliteten af deres design af biologieksperimenter. De resultater, der beskæftiger sig med elevernes opfattelse af feedback fra ChatGPT, vil desuden danne grundlag for en diskussion af ChatGPT´s styrker og svagheder som formativ feedbackgiver på forsøgsdesign.

6.1 Elevers brug af ChatGPT´s feedback på forsøgsdesign

Resultaterne fra den kvantitative analyse af elevrapporterne viser, at eleverne gennemsnitligt implementerer 4,44 ændringer pr. gruppe med relation til ChatGPT mod et gennemsnit på 0,63 ændringer pr. gruppe uden relation til ChatGPT (Figur 8). Dermed ser det ud til, at eleverne bruger feedbacken fra ChatGPT til at ændre i deres forsøgsdesign. Det fund er i overensstemmelse med et tidligere studie, der har vist, at formativ feedback bliver brugt til at foretage ændringer i forsøgsdesign (Anker-Hansen & Andrée, 2019).

Da 3.g-grupperne gennemsnitligt implementerer 4,71 ændringer med relation til ChatGPT pr. gruppe, mens tallet for 1.g-grupperne kun er 4,00 ændringer pr. gruppe, kan man argumentere for, at 3.g grupperne er bedre til at bruge feedbacken fra ChatGPT. Her er det værd at bemærke, at forskellen mellem de to klassetrin ikke er statistisk signifikant. Det kan skyldes, at der kun indgår 16 elevrapporter i denne del af undersøgelsen, men resultaterne viser samtidig, at den procentvise forbedring i kvalitetsscore for forsøgsdesign pr. gruppe ligger tæt for 1.g (32,15%) og 3.g (35,93 %), og at der heller ikke er signifikant forskel mellem de to klasser, når det gælder forbedring i kvalitetsscore (Figur 10). Hertil kommer, at 1.g og 3.g-eleverne har arbejdet med forskellige øvelser, hvilket betyder, at forskelle mellem de to grupper ikke nødvendigvis kun kan sættes i relation til forskellige klassetrin. Derfor kan denne undersøgelses resultater ikke bruges til at vise, at 3.g er bedre til at bruge feedback fra ChatGPT end 1.g.

Til gengæld viser analysen af elevrapporter, spørgeskemasvar og interview, at eleverne i både 1.g og 3.g implementerer rettelser, der har at gøre med variabelkontrol, kontrolforsøg og gentagelse eller replikation af forsøg. I den sammenhæng er det interessant, at resultaterne fra elevrapporterne viser, at feedback med relation til variabelkontrol implementeres af flest grupper

(figur 9). Relevansen bakkes op af spørgeskemaundersøgelsen, hvor kontrol af temperatur og blindtest bliver nævnt som eksempler på variabelkontrol, ligesom reduktion af fejlkilder nævnes både i spørgeskemaundersøgelse (afsnit 5.3.1) og interview (afsnit 5.4.2). At eleverne har brug for feedback om variabelkontrol stemmer fint overens med, at studerende ofte ændrer for mange variabler i samme eksperiment (Gobert, Pedro, Raziuddin, & Baker, 2013).

Derudover viser resultaterne, at syv grupper implementerer ændringer i kategorien kontrolforsøg (figur 9). Tendensen understøttes af, at syv respondenter nævner, at feedbacken fik dem til at implementere kontrolforsøg i deres spørgeskemabesvarelse (afsnit 5.3.1), ligesom elever i halvdelen af informantgrupperne forholder sig positivt til ChatGPT´s forslag om kontrolforsøg (afsnit 5.4.2). Behovet for feedback om kontrolforsøg er i overensstemmelse med et studie af Coleman et al. (2023), der finder, at 1-års studerende på et biologikursus har svært ved at bruge kontrolforsøg rigtigt. Endelig har syv af grupperne implementeret feedback inden for kategorien gentagelse/replikation (figur 9), ligesom syv respondenter i spørgeskemaundersøgelsen og halvdelen af de interviewede grupper nævner det som relevant (afsnit 5.3.1 og 5.4.2). Disse resultater stemmer fint overens med et studie, der viser, at en del førsteårsstuderende i biologi ikke har en god forståelse af vigtigheden af at gentage eksperimenter (Brownell, et al., 2014). Derfor er det oplagt at antage, at eleverne har haft brug for denne type feedback i undersøgelsen.

Man kan undre sig over, at der kun er syv respondenter i spørgeskemaundersøgelsen, der nævner kontrolforsøg og gentagelse/replikation, når feedbacken i begge tilfælde implementeres af syv ud af 19 grupper. Forklaringen kan være, at eleverne kun er blevet bedt om at give et eller flere konkrete eksempler på anvendt feedback i spørgeskemaet, og at de derfor har nævnt færre eksempler, end de har implementeret. I denne sammenhæng er det desuden værd at bemærke, at både elevrapporter, spørgeskemaundersøgelse og interview viser, at eleverne bruger ChatGPT's feedback om variabelkontrol, kontrolforsøg og gentagelse/replikation til revision af deres forsøgsdesign, hvilket styrker validiteten af de beskrevne resultater med det forbehold, at datagrundlaget i undersøgelsen er begrænset til to klasser.

Et andet relevant aspekt i elevernes brug af feedback fra ChatGPT er, at de indsamlede interviewdata viser, at ChatGPT´s svar bl.a. bliver brugt som tjekliste til at huske at få det hele med. Hertil kommer, at analysen af interview og spørgeskemaer viser, at mange elever oplever at få

feedback, som de ”ikke havde tænkt på” (A. 3.3) og giver udtryk for, at ChatGPT er ”et godt værktøj som giver flere ideer” (R. 3.8). Udsagnene er interessante, fordi de viser, at ChatGPT bliver brugt som en kognitiv partner, der hjælper eleverne med at få nye ideer og tilgange til forsøgsdesign, som de ikke selv har overvejet. Når teknologier bliver brugt på denne måde, lærer eleven med teknologien i stedet for fra den, og elevrollen skifter fra at reproducere information til at konstruere viden (Jonassen, 1995). Det passer godt sammen med en undersøgelsesbaseret tilgang, hvor elever netop konstruerer viden, der er ny for dem, igennem en aktiv og elevcentreret tilgang til læring (Spronken-Smith & Walker, 2010). I den undervisningskontekst, der danner grundlag for dette masterprojekt, er ChatGPT´s rolle som tidligere nævnt stramt styret, fordi eleverne udelukkende brugte ChatGPT´s svar på promptskabeloner og ikke indgik yderligere i dialog med ChatGPT. Dermed kan man argumentere for, at teknologiens underliggende agens ikke er fuldt udfoldet, men resultaterne tyder på, at ChatGPT har potentiale til at fungere som kognitiv partner, når elever skal designe biologieksperimenter.

Samlet set viser ovenstående, at eleverne bruger formativ feedback fra ChatGPT til at revidere deres forsøgsdesign i den undersøgte undervisningskontekst. I det følgende afsnit bliver det diskuteret, om det påvirker kvaliteten af deres forsøgsdesign.

6.2 Feedback fra ChatGPT og kvaliteten af elevers forsøgsdesign

Da den gennemsnitlige kvalitetsscore for elevers forsøgsdesign går fra 15,32 point til 20,21 point pr. gruppe efter feedback fra ChatGPT for alle grupper samlet, og stigningen er signifikant med $P<0{,}001$ (figur 10), tyder resultaterne på, at feedbacken hjælper eleverne med at forbedre deres forsøgsdesign. I den sammenhæng er det værd at bemærke, at en alternativ forklaring på stigningen kan være, at eleverne har forbedret deres forsøgsdesign, fordi de samlet set har brugt mere tid på det sidste end det første forsøgsdesign. Det er imidlertid i modstrid med følgende udsagn fra et elevinterview: ”Jeg synes nemlig, altså meget af det her havde vi så ikke kunne skrive uden chatbotten, eller vi havde i hvert fald glemt det, men når vi ligesom selv går ind og bearbejder det, den har sagt, så synes jeg, det måske er blevet bedre end vores udgangspunkt uden den.” (E. 3.3). Udsagnet tyder på, at ChatGPT´s feedback har passet til elevernes ZNU og hjulpet dem med at forbedre kvaliteten af deres design, hvilket også kommer til udtryk i spørgeskemaundersøgelsen, hvor en elev siger ”Den gav os flere forbedringer som vi kunne bruge

til vores forsøg, der gjorde vi lidt nemmere kunne komme i mål med havde vi havde i tankerne." (R 1.12). De nævnte fund i interview- og spørgeskemaundersøgelse sammenholdt med den tidligere nævnte kobling mellem implementerede ændringer og feedback fra ChatGPT i elevrapporterne viser derfor, at stigningen i kvalitetsscore i elevernes forsøgsdesign sandsynligvis er relateret til formativ feedback fra ChatGPT.

Påstanden om, at det er formativ feedback fra ChatGPT, der bidrager til stigningen i kvalitetsscore, er desuden understøttet af, at resultaterne fra elevrapporterne samtidig viser, at der ses en stigning i alle delkvalitetsscorer i både 1.g og 3.g efter feedback fra ChatGPT (figur 11). Variabelkontrol, kontrolforsøg og gentagelse/replikation af forsøg indgår alle i delkvalitetsscorerne, og dermed viser resultaterne, at den stigende kvalitet i elevernes forsøgsdesign efter feedback bl.a. skyldes forbedringer i disse kategorier. Det er i fin overensstemmelse med undersøgelsens øvrige fund, der som tidligere nævnt viser, at det er i netop de nævnte kategorier, at eleverne implementerer ændringer. Her er det værd at bemærke, at kun en gruppe implementerer ændringer i kategorien "Detaljeret forsøgsbeskrivelse", når ChatGPT´s feedback sammenholdes med implementerede ændringer (figur 9). Alligevel stiger denne kvalitetsscorer gennemsnitligt fra 2,80 til 3,50 point pr. gruppe i 1.g og fra 3,78 til 4,56 point i 3.g (figur 11). Forklaringen er formentlig, at de øvrige ændringer, som eleverne foretager, bidrager til en mere detaljeret forsøgsbeskrivelse. Det harmonerer med fundene fra interview- og spørgeskemaundersøgelse, hvor en elev giver udtryk for, at deres "forsøgsdesign blev mere detaljeret" (R. 1.2), mens en anden siger, at det, som de skriver "lyder meget mere præcist, end hvis vi ikke havde brugt" ChatGPT (A 3.3). Resultaterne indikerer derfor, at eleverne også bliver bedre til at beskrive deres forsøg.

Forbedringerne i elevernes forsøgsbeskrivelser er bemærkelsesværdige, fordi de antyder, at brug af formativ feedback fra ChatGPT i arbejdet med at designe eksperimenter understøtter udviklingen af elevernes kommunikationsevner i relation til forsøgsdesign. Samtidig tyder den beskrevne stigning i delkvalitetsscoren for variabelkontrol på, at det valgte undervisningsdesign også fremmer elevernes variabelforståelse. Det er interessant, fordi kommunikationsevne og variabelforståelse indgår som vigtige elementer i den eksperimentelle problemløsningskompetence, der handler om at kunne "opnå forståelse for og træning i at løse problemer ved at angribe problemerne empirisk-

eksperimentelt" (Jacobsen, 2008, s. 3), hvilket også kan være et vigtigt formål med eksperimentelt arbejde i undervisningen (Jacobsen, 2008).

Som det fremgår af det foregående, så kan den signifikante stigning i kvalitetsscore, som er observeret i elevernes forsøgsdesign, med stor sandsynlighed sættes i relation til formativ feedback fra ChatGPT. Samtidig lader det til, at arbejdet med forsøgsdesign også har andre gavnlige effekter. I den sammenhæng er det værd at bemærke, at et tidligere studie har vist, at formativ feedback forbedrer elevernes færdigheder inden for forsøgsdesign (Ganajová, et al., 2021). Samtidig påpeger Anker-Hansen og Andrée (2019), at ændringer i forsøgsdesign i sig selv ikke er nok til at vurdere potentialet for formativ peer feedback, bl.a. fordi der også kan være værdifulde refleksioner i gruppediskussioner, som ikke nødvendigvis resulterer i ændringer i forsøgsdesignet. Det samme kan antages gøre sig gældende i denne undersøgelse, hvor eleverne diskuterer ChatGPT´s feedback i grupper. Dermed er oplagt at antage, at kvaliteten af elevernes forsøgsdesign ikke alene kan bruges til at bedømme udbyttet af at bruge ChatGPT som formativ feedbackgiver i den undersøgte undervisningskontekst, selvom det ikke er blevet undersøgt nærmere. I det følgende afsnit bliver ChatGPT´s potentiale som feedbackgiver på forsøgsdesign derfor diskuteret med særligt fokus på elevernes perspektiv og de tidligere beskrevne krav til formativ feedback fra lærer til elev i forskningslitteraturen.

6.3 ChatGPT´s styrker og svagheder som formativ feedbackgiver på forsøgsdesign

Da resultaterne fra spørgeskemaundersøgelsen viser, at 63 procent af eleverne vurderer, at ChatGPT´s feedback i høj grad eller i meget høj grad var brugbar til forsøgsdesign, lader det til, at mange elever kan se potentialet i at bruge ChatGPT som feedbackgiver på forsøgsdesign (figur 12B). Det bliver yderligere understreget af, at 73 procent af eleverne tilkendegiver, at det er sandsynligt eller meget sandsynligt, at de vil bruge ChatGPT igen, hvis de skal designe forsøg (figur 12D). I den sammenhæng er det værd at bemærke, at spørgeskemaundersøgelsen har en svarprocent på 93, og at data derfor kan antages at være repræsentative for de to klasser. Samtidig harmonerer resultaterne med en spørgeskemaundersøgelse blandt high school-elever, hvor 74,6 % angiver, at de oplever ChatGPT som et nyttigt værktøj til skolearbejde.

Årsagen til, at så mange eleverne vurderer ChatGPT som brugbar og værd at bruge igen, er formentlig, at feedbacken opfylder flere af de krav, der er til god formativ feedback fra lærer til elev. Det bliver bl.a. understreget af, at en elev i interviewundersøgelsen fremhæver det som positivt, at ChatGPT´s feedback passede til netop deres forsøg ved at sige "det var også målrettet til vores forsøg […] altså det var nemmere sådan at skrive om, fordi den ligesom havde lavet det til det, vi skulle undersøge" (B. 1.1). God formativ feedback fra lærer til elev er kendetegnet ved at være differentieret (Harlen, 2013a). At ChatGPT differentierer feedbacken, så den passer til den enkelte gruppe, fremgår desuden af det følgende forbedringsforslag fra en elevrapport.

"**Kontrolgruppe:** Ud over de 5 glas med forskellige mængder citronsaft skal I inkludere en kontrolgruppe med 0 dråber citronsaft. Dette vil hjælpe med at afgøre, om deltageren i forsøget kan smage citronen i forhold til en ren vandsmag." (Bilag 4).

Eksemplet viser samtidig, at ChatGPT kan identificere, hvad der kan forbedres, og hvordan man kan gå i gang med forbedringen, hvilket også karakteriserer god formativ feedback fra lærer til elev. Feedbacken bør samtidig identificere, hvad der er blevet gjort godt (McManus, 2008; Harlen, 2013a), og det viser følgende eksempel fra en elevrapport, at ChatGPT også er i stand til:

"Dit foreslåede forsøgsdesign er generelt godt, da det tager højde for mange af de vigtige aspekter ved at undersøge tærskelværdien for smagssansen for citronsaft. Det er også godt, at du har inkluderet flere forsøgspersoner for at tage højde for individuelle variationer i smagssansen." (Bilag 4).

De to ovenstående eksempler viser desuden, at ChatGPT begrunder sine svar, hvilket to grupper nævner som positivt i interviewundersøgelsen. Samlet set fremhæver elevernes positive udsagn potentialet i at bruge ChatGPT som feedbackgiver på forsøgsdesign. Det bliver også understreget af et udsagn fra spørgeskemaundersøgelsen, hvor en elev opsummerer ChatGPT´s rolle som feedbackgiver med ordene "Det føltes lidt som en ekstra lærer" (R. 3.12).

Selvom ChatGPT´s feedback på forsøgsdesign opfylder mange af kravene til god formativ feedback fra lærer til elev, viser spørgeskemaundersøgelsen også, at 12 procent af eleverne vurderer, at ChatGPT´s feedback i mindre grad eller slet ikke var brugbar til forsøgsdesign (figur 12B), ligesom otte procent tilkendegiver, at det er usandsynligt eller meget usandsynligt, at de vil bruge ChatGPT til at designe forsøg igen (figur 12D).

Den kritiske holdning skal formentlig tilskrives mangler i ChatGPT's feedback og understøttes af fund i spørgeskema- og interviewundersøgelserne, hvor eleverne påpeger, at ChatGPT ikke tager nok højde for deres praktiske virkelighed. En elev betegner det som "en mangel på situationsfornemmelse" (A 1.3), hvilket ikke er overraskende, da generativ AI ikke forstår virkelige objekter (Unesco, Fengchun, & Holmes, 2023). Det kan være en af forklaringerne på, at eleverne oplever, at ChatGPT har svært ved at rådgive om mængder i det testede undervisningsdesign. Derudover forholder eleverne sig kritisk til dele af feedbacken, hvor ChatGPT foreslår forbedringer, som allerede er indtænkt i deres forsøgsdesign, eller udstyr, som de ikke har til rådighed i både spørgeskema- og interviewundersøgelse (afsnit 5.3.2 og 5.4.4). I den sammenhæng er det værd at bemærke, at elevernes liste over tilgængelige materialer indgik i den anvendte promptskabelon, og at manglerne formentlig ikke kan tilskrives deres prompt, men skyldes ChatGPT´s måde at generere feedback på. Forskningen viser, at formativ feedback skal være effektiv i forhold til elevernes arbejde med forbedringer (Dolin, Harlen, Black, & Tiberghien, 2018). Derfor sænker forslag om udstyr, som ikke er tilgængeligt for eleverne, kvaliteten af feedbacken. Det er formentlig særligt udtalt i den undersøgte undervisningskontekst, hvor eleverne indsatte deres prompt, men ikke indgik yderligere i dialog med ChatGPT. Man kan derfor argumentere for, at denne begrænsning kan imødegås ved brug af feedbackformater, hvor eleverne har mulighed for at gå mere i dialog med ChatGPT og specificere, at de vil have et nyt forslag uden det pågældende udstyr, som de ikke har til rådighed. Endelig kan nogle elevers kritiske holdning formentlig også forklares med, at de ikke har været udførlige i beskrivelserne af deres forsøgsdesign, hvilket kan have betydet, at kvaliteten af den feedback, som de har fået, har været ringere. Det bakkes op af fund i spørgeskemaundersøgelsen, hvor eleverne har erfaret, at det er vigtigt at være specifik og have mange detaljer med (afsnit 5.3.4), ligesom nogle af forsøgsbeskrivelserne i elevrapporterne lader til ikke at være særlig udførlige.

Fund i både spørgeskema- og interviewundersøgelse viser desuden, at eleverne modtager feedback med, hvad de opfatter som overdrevne sikkerhedsinstrukser og etiske hensyn (afsnit 5.3.2 og 5.4.4). Her er forklaringen formentlig, at den anvendte promptskabelon ikke direkte præciserede elevernes niveau, hvilket må betegnes som en svaghed i undersøgelsens empiriske design. Den manglende præcisering af elevniveau i promptskabelonen er desuden interessant, fordi den sandsynligvis kan forklare fund i interviewundersøgelsen, hvor fire ud af seks informantgrupper nævner, at de havde svært ved at forstå dele af feedbacken (afsnit 5.4.3). Da god formativ feedback bør være forståelig for eleverne (Harlen, 2013b), og følelsen af at være kompetent eller dygtig fremmer indre motivation (Ryan & Deci, 2000), påvirker det kvaliteten af feedbacken negativt, at den er svær at forstå, fordi det kan få eleverne til at føle sig mindre kompetente. I forlængelse heraf er det relevant at nævne, at en elev har lagt mærke til, at ChatGPT´s feedback blander "noget fra både høje niveauer og lave niveauer" (H. 3.2). Det vidner om, at ChatGPT i nogle tilfælde har haft svært ved at ramme elevernes ZNU, hvilket også sænker kvaliteten af feedbacken, fordi god formativ evaluering er kendetegnet ved at være tilpasset niveau og formåen for den enkelte elev (Harlen, 2013a) eller gruppe.

Her er det værd at bemærke, at data fra elevrapporterne samtidig viser, at eleverne har bedømt 110 enheder med feedback fra ChatGPT korrekt og fejlbedømt 12 enheder, da de skulle tage stilling til, om feedbacken var relevant eller irrelevant/forkert. Derudover har eleverne undladt at bedømme 27 enheder og i nogle tilfælde angivet, at det var fordi feedbacken var svær at forstå (figur 7). Dermed tyder resultaterne på, at en stor del af feedbacken alligevel har været forståelig for eleverne. Samtidig er det interessant, at eleverne overvejende foretager korrekte bedømmelser, fordi andre studier har fundet en tendens til, at studerende stoler for meget (Ding, Li, Jiang, & Gapud, 2023) eller for lidt (Zhang, 2023) på ChatGPT. Forklaringen er sandsynligvis, at dette projekts socialkonstruktivistiske tilgang har betydet, at eleverne har haft mulighed for at drøfte feedbacken i grupper, og at det har ført til en bedre bedømmelse end i de andre studier, hvor eleverne tilsyneladende har arbejdet mere individuelt.

Resultaterne fra elevrapporterne viser desuden, at dele af feedbacken fra ChatGPT er ukorrekt (figur 7), hvilket er i overensstemmelse med andre studier (Wardat, Tashtoush, AlAli, & Adeeb, 2023; Ding, Li, Jiang, & Gapud, 2023) og forventeligt, da generativ AI ikke kan regnes for at være præcis (Unesco, Fengchun, & Holmes, 2023). Det er sandsynligvis forklaringen på, at fund i både spørgeskema- og interviewundersøgelse viser, at flere elever mener, at man bør være kritisk over for ChatGPT´s feedback og opmærksom på "ikke at følge den 100%" (R. 3.10), hvilket stemmer fint overens med Unescos anbefalinger om, at elever altid skal foretage en kritisk vurdering af informationer fra generativ AI (Unesco, Fengchun, & Holmes, 2023). Ukorrekt feedback kan påvirke ChatGPT´s potentiale som formativ feedbackgiver på forsøgsdesign negativt, men kan måske reduceres, hvis fremtidig udvikling forbedrer sprogmodellen. Tilstedeværelsen af ukorrekt feedback giver desuden mulighed for at designe undervisning, der opfordrer eleverne til at stille spørgsmålstegn ved AI-teknologiers pålidelighed og derved træner deres kritiske tænkning, som det er anbefalet i litteraturen (Long & Magerko, 2020). I den sammenhæng er en undersøgelsesbaseret tilgang, som den, der er anvendt i dette masterprojekt, oplagt, fordi eleverne også får mulighed for at teste ChatGPT´s feedback i praksis, når de udfører deres forsøg.

På trods af at eleverne identificerer ukorrekt feedback, svarer 81 procent af respondenterne i spørgeskemaundersøgelsen, at man altid eller ofte kan stole på ChatGPT´s forbedringsforslag til deres forsøgsdesign. Tendensen bakkes op af fund i interviewundersøgelsen, hvor en elev siger, at "tingene den sagde […] egentlig virkede troværdige nok" (C. 3.2). Et fund fra spørgeskemaundersøgelsen, hvor 81 procent af eleverne vurderer, at deres forsøg fungerede godt eller meget godt, da de skulle afprøve det, indikerer, at en forklaring på, at eleverne i overvejende grad stoler på ChatGPT, kan være, at de har oplevet, at ChatGPT´s feedback har hjulpet dem til at udvikle velfungerende forsøg.

Samlet set viser ovenstående, at eleverne oplever ChatGPT´s feedback på forsøgsdesign i den undersøgte undervisningskontekst som brugbar og værd at bruge igen. De beskriver det som positivt, at feedbacken er begrundet og differentieret ved at være målrettet deres forsøg, og den feedback, som de har modtaget, viser, at feedbacken bl.a. formår at identificere, hvad der er blevet gjort godt, hvad der kan forbedres, og hvordan man kan gå i gang med forbedringen. Selvom

eleverne ikke giver direkte udtryk for det i undersøgelsen, så er det sandsynligt, at deres positive holdning også kan forklares med, at kravene til god formativ feedback fra lærer til elev også er tænkt ind i det undervisningsdesign, der danner grundlag for undersøgelsen. Det gælder f.eks. kravene om, at feedbacken er rettidig (McManus, 2008), bliver givet som kommentarer (Harlen, 2013a) og undgår sammenligning med andre elever eller opgaver (ARG, 2002). Derudover er kravet om, at feedbacken tager udgangspunkt i læringsmål (McManus, 2008) indlejret i udformningen af den anvendte promptskabelon.

Det foregående viser desuden, at eleverne også forholder sig negativt til dele af feedbacken, som de oplever mangler en kobling til virkeligheden og til tider er svær at forstå. Det betyder, at kravene om forståelighed (Harlen, 2013b) og passende differentiering (Dolin, Harlen, Black, & Tiberghien, 2018) i god formativ feedback fra lærer til elev ikke er opfyldt alle steder i det testede undervisningsformat, hvilket kan imødegås ved enkelte justeringer i fremtidig undervisning. Kravet om, at god formativ feedback fra lærer til elev skal være empatisk (ARG, 2002) og gøre eleverne opmærksomme på, hvad de har lært ((Harlen, 2013b), er ikke blevet undersøgt. Derudover kan man argumentere for, at kriterierne for god formativ feedback fra lærer til elev kan udvides og derved give et mere nuanceret billede af ChatGPT´s potentiale som feedbackgiver. Hertil kommer, at de følelsesmæssige interaktioner, der er forbundet med feedbackprocessen, må forventes at være anderledes og i visse tilfælde gå tabt, når elever interagerer med kunstig intelligens i stedet for en lærer.

7. Konklusion og perspektivering

I dette projekt er det blevet undersøgt, hvordan formativ feedback fra ChatGPT påvirker gymnasieelevers design af biologieksperimenter i en 1.g klasse på 30 elever med naturvidenskabeligt grundforløb og en 3.g klasse på 30 elever med biologi på B-niveau på stx. Undersøgelsens resultater, der er baseret på en undersøgelsesbaseret tilgang med 19 elevrapporter, 56 elevers besvarelse af spørgeskemaer og 6 gruppeinterviews med i alt 17 elever, når frem til følgende hovedkonklusioner: 1) Eleverne bruger formativ feedback fra ChatGPT til at implementere ændringer i deres forsøgsdesign. Ændringerne kan bl.a. sættes i relation til kontrolforsøg, variabelkontrol og forsøgsgentagelse. 2) Kvaliteten af elevernes forsøgsdesign stiger signifikant ($P<0,001$) for alle grupper samlet efter feedback fra ChatGPT og kan sandsynligvis

tilskrives ChatGPT´s feedback. 3) De fleste elever opfatter ChatGPT´s feedback som brugbar og værd at anvende igen, bl.a. fordi den er målrettet deres forsøg og begrundet, men flere er samtidig kritiske over for dele af feedbacken, som de ikke oplever, tager nok højde for deres praktiske virkelighed. 4) Selvom dele af ChatGPT´s feedback er svær at forstå for eleverne, er de gode til at bedømme feedbackens korrekthed i grupper, og på trods af at flere elever mener, at man bør forholde sig kritisk til feedbacken, mener de fleste, at den overordnet set er til at stole på. Dermed konkluderer undersøgelsen, at formativ feedback fra ChatGPT kan hjælpe gymnasieelever, der skal designe biologieksperimenter med at forbedre kvaliteten af deres forsøgsdesign, selvom der er brug for en mere nuanceret tilgang til ChatGPT's feedback.

Resultaterne blev indsamlet i september 2023, hvor eleverne i undersøgelsen ikke havde ret meget erfaring med lærerorganiseret brug af ChatGPT, og resultaterne skal derfor ses i denne sammenhæng. Fremadrettet kunne det være interessant at undersøge, hvordan elevernes interaktion og oplevelse af ChatGPT som feedbackgiver udvikler sig over længere tidsperioder i takt med, at de gør sig flere erfaringer. Hertil kommer, at undersøgelsen belyser ChatGPT´s potentiale som formativ feedbackgiver på forsøgsdesign til grupper, og derfor kunne det være oplagt at udforske, hvordan en individuel tilgang påvirker elevens interaktion med ChatGPT i den sammenhæng. Endelig er det en begrænsning ved den anvendte tilgang, at ChatGPT´s rolle som feedbackgiver var stramt styret, fordi eleverne udelukkende brugte ChatGPT´s svar på promptskabeloner og ikke indgik yderligere i dialog med ChatGPT. En anden relevant retning for fremtidige undersøgelser er derfor at undersøge, hvordan formativ feedback fra ChatGPT påvirker elevers design af eksperimenter, hvis der anvendes en mere åben og dialogbaseret tilgang til elevernes interaktion med ChatGPT.

Da formativ feedback fra ChatGPT i relation til elevers design af eksperimenter i gymnasieskolen endnu ikke lader til at være beskrevet i forskningslitteraturen, og denne undersøgelse beskæftiger sig med generelle principper inden for forsøgsdesign, kan resultaterne også bruges som inspiration til planlægning af undersøgelsesbaseret undervisning med inddragelse af ChatGPT i andre naturvidenskabelige fag. Derfor er erfaringerne fra udformningen af den undersøgte undervisning og opgavens hovedkonklusioner blevet brugt som udgangspunkt for følgende praktiske råd til

tilrettelæggelse af undervisning, hvor elever skal bruge ChatGPT som formativ feedbackgiver til at lære at designe deres egne eksperimenter i naturvidenskab.

- Sørg for, at eleverne selv skaber deres eget forsøgsdesign. Det kan f.eks. gøres ved at lade eleverne udarbejde et forsøgsdesign, som ChatGPT giver feedback på i punkter, i stedet for at feedbacken bliver leveret som en færdig øvelsesvejledning.
- Gør eleverne opmærksomme på, at det er vigtigt, at de laver en udførlig beskrivelse af deres forsøgsdesign, inden de beder om feedback fra ChatGPT.
- Sørg for, at eleverne bruger prompts til at generere feedback, hvor deres niveau, samt evalueringskriterier og rammer for opgaveløsningen er beskrevet.
- Gør det tydeligt for eleverne, at de ikke nødvendigvis kan stole på al feedback. Giv dem mulighed for at diskutere feedbacken og undersøge feedback, som de ikke forstår.
- Lad eleverne implementere feedbacken i et nyt forsøgsdesign, som de afprøver.
- Lav en tydelig tidsmæssig opdeling af forsøgsdesign og afprøvning for at sikre, at eleverne bruger tilstrækkeligt med tid på at arbejde med feedback på deres forsøgsdesign.

8. Bibliografi

Anderson, R. D. (2002). Reforming Science teaching: What Research says about Inquiry. *Journal of Science teacher Education, 13*(1), pp. 1-12.

Andrade, H. G. (2005). Teaching With Rubrics: The Good, the Bad, and the Ugly. *College Teaching, 53*(1), pp. 27-31. doi:10.3200/CTCH.53.1.27-31

Anker-Hansen, J., & Andrée, M. (2019). Using and rejecting peer feedback in the science classroom: a study of students´negotiations on how to use peer feedback when designing experiments. *Research in Science & Technological Education, 37*(3), pp. 346-365 https://doi.org/10.1080/02635143.2018.1557628.

ARG. (2002). *(Assessment Reform Group).* Retrieved from Assessment for learning: Ten principles: https://www.researchgate.net/publication/271849158_Assessment_for_Learning_10_Principles_Research-based_principles_to_guide_classroom_practice_Assessment_for_Learning

Ariely, M., Nazaretsky, T., & Alexandron, G. (2023). Machine learning and hebrew NLP for automated. *International Journal of Artifcial Intelligence in, 33*(1), pp. 1-34. doi:https://doi.org/10.1007/s40593-021-00283-x

Berg, C. A., Bergendahl, V., & Lundberg, B. (2003). Benefiting from an open-ended experiment? A comparison of attitudes to, and outcomes of, an expository versus an open-inquiry version of the same experiment. *International Journal of Science Education, 25*(3), pp. 351-372.

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research In Psychology, 3*(2), pp. 77-101. doi:10.1191/1478088706qp063oa

Brinkmann, S., & Tanggaard, L. (2015). 1. Interviewet: Samtalen som forskningsmetode. In S. Brinkmann, & L. Tanggaard, *Kvalitative metoder. En grundbog. 2. udgave.* (pp. 38-42). Hans Reitzels Forlag.

Brownell, S. E., Wenderoth, M. P., Theobald, R., Okoroafor, N., Koval, M., Freeman, S., . . . Crowe, A. J. (2014). How Students Think about Experimental Design: Novel Conceptions Revealed by in-Class Activities. *BioScience, 64*(2), pp. 126-137 https://doi.org/10.1093/biosci/bit016. doi:10.1093/biosci/bit016

BUVM. (2017a). *Naturvidenskabeligt grundforløb - stx august 2017.* Retrieved 03 27, 2024, from www.uvm.dk Stx-læreplaner 2017: https://www.uvm.dk/-/media/filer/uvm/gym-laereplaner-2017/stx/naturvidenskabeligt-grundforloeb-stx-august-2017-ua.pdf

BUVM. (2017b). *www.uvm.dk.* doi:https://www.uvm.dk/-/media/filer/uvm/gym-laereplaner-2017/stx/biologi-b-stx-august-2017.pdf

BUVM. (2023a, Juni). *www.uvm.dk.* Retrieved 03 27, 2024, from Stx - læreplaner 2017: https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf23/aug/vejledninger/230814-naturvidenskabeligt-grundforloeb-stx.pdf

BUVM. (2023b, Juni). *www.uvm.dk.* Retrieved Marts 27, 2024, from Stx - læreplaner 2017: https://www.uvm.dk/-/media/filer/uvm/udd/gym/pdf23/jun/230608-vejledning-til-biologi-a--b-og-c--stx.pdf

Coleman, A. B., Lorenzo, K., McLamb, F., Sanku, A., Khan, S., & Bozinovic, G. (2023). Imagining, designing, and interpreting experiments: Using quantitative assessment to improve instruction in scientific reasoning. *Biochemistry and Molecular Biology Education, 51*, pp. 286-301. doi:10.1002/bmb.21727

Debuse, J. C., & Lawley, M. (2016). Benefits and drawbacks of computer-based assessment and feedback systems: Students and educator perspectives. *British Journal of Educational Technology, 47*(2), pp. 294-301. doi:10.1111/bjet.12232

Ding, L., Li, T., Jiang, S., & Gapud, A. (2023). Students' perceptions of using ChatGPT. *International Journal of Educational Technology in Higher Education*(20:63), pp. 1-18. doi:http://dx.doi.org/10.1186/s41239-023-00434-1

Dolin, J., Harlen, W., Black, P. J., & Tiberghien, A. (2018). Exploring Relations Between Formative and Summative Assessment. In J. Dolin, & R. Evans, *Transforming Assessment* (pp. 53-80). Springer.

Eckes, A., & Wilde, M. (2019). Structuring experiments in biology lessons through teacher feedback. *International Journal of Science Education, 41*(16), pp. 2233-2253 . doi:10.1080/09500693.2019.1668578

Epinion. (2018). *Evaluering af det fleksible klasseloft på 28 elever i de gymnasiale uddannelser.* Undervisningsminiseriet. Retrieved 08 22, 2023, from https://www.uvm.dk/publikationer/gymnasiale-uddannelser/2018-evaluering-af-det-fleksible-klasseloft-paa-28-elever-i-de-gymnasiale-uddannelser

EVA. (2017). *Danmarks Evalueringsinstitut.* Retrieved 08 22, 2023, from https://www.eva.dk/eva-tilbyder/gode-spoergeskema

Forman, N., Udvaros, J., & Avornicului, M. S. (2023). ChatGPT: A new tool shaping the future for high school students. *International Journal of Advances Natural Sciences and Engineering Researches, 7*, pp. 95-102. doi:10.59287/ijanser.2023.7.4.562

Frederiksen, M. (2020). Mixed methods-forskning. In S. Brinkmann, & L. Tanggaard, *Kvalitative metoder* (pp. 257-277). Hans Reitzels Forlag.

Friisberg, K. (2023, 05 25). *7 ting du skal vide om AI i skolen*. Retrieved 01 23, 2024, from www.sdu.dk: https://www.sdu.dk/da/nyheder/7-ting-du-skal-vide-om-ai-i-skolen

Frisdahl, K. (2014). *Kompendium: Inquiry Based Science Education - IBSE.* Retrieved 02 21, 2024, from Institut for Naturfagenes Didaktik: https://www.ind.ku.dk/publikationer/inds_skriftserie/2014-36/Kompendie-IBSE_ny_web2.pdf

Ganajová, M., Sotáková, I., Lukáč, S., Ješková, Z., Jurková, V., & Orosová, R. (2021). Formative assessment as a tool to enhance the development of inquiry skills in science education. *Journal of Baltic Science Education, 20*(2), pp. 204-222.

GL. (2023, 10 13). *Kunstig intelligens (AI) og ChatGPT*. Retrieved 04 12, 2024, from www.gl.org: https://www.gl.org/detmenergl/Sider/Kunstigintelligens.aspx

Gobert, J. D., Pedro, M. S., Raziuddin, J., & Baker, R. S. (2013). From Log Files to Assessment Metrics: Measuring Students´ Science Inquiry Skills Using Educational Data Mining. *The Journal of the Learning Sciences, 22*(4), pp. 521-563. doi:0. 1 080/ 1 0508406.20 13.837391

Harlen, W. (2013a). 3: Assessment purposes and uses. In W. Harlen, *Assessment and Inquiry-based science education* (pp. 16-25). Trieste: Global Network of Science Academies (IAP) Science Education Programme (SEP).

Harlen, W. (2013b). 5: Implementing formative assessment og IBSE. In W. Harlen, *Assessment and Inquiry-based science education* (pp. 35-47). Trieste: Global Network of Science Academies (IAP) Science Education Programme.

Hattie, J., & Timperley, H. (2007). The Power of Feedback. *Review of Educational Research, 77*(1), pp. 81-112.

Heale, R., & Twycross, A. (2015). Validity and reliability in quantitative studies. *Evid Based Nurs, 18*(3), pp. 66-67. doi:10.1136/eb-2015-102129

Jacobsen, L. B. (2008). Formål med eksperimentelt arbejde i fysikundervisningen. *MONA*(4), pp. 22-41.

Jonassen, D. H. (1995). Computers as Cogntive Tools: Learning with Technology, Not from Technology. *Journal of Computing in Higher Education, 6*(2), pp. 40-73.

Lo, L. S. (2023, April 18). The CLEAR path: A framework for enhancing information literacy through prompt engineering. *The Journal of Academic Librarianship, 49*. Retrieved from https://doi.org/10.1016/j.acalib.2023.102720

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. Atlanta, USA. doi: https://doi.org/10.1145/3313831.3376727

Louisville, U. o. (n.d.). *What is Critical Thinking?* Retrieved April 10, 2024, from https://louisville.edu/ideastoaction/about/criticalthinking/what

Madsen, B. S. (2017). Kapitel 4 Stikprøver og spørgeskemaundersøgelser. In B. S. Madsen, *Statistik for ikke-statistikere* (pp. 65-91). Samfundslitteratur.

Madsen, L. M., Evans, R., & Bruun, J. (2020). Undersøgelsesbaseret undervisning: 6F-modellen - dens tilblivelse og udvikling i Danmark. *MONA, 1*, pp. 26-44.

Maestrales, S., Zhai, X., Touitou, I., Baker, Q., Schneider, B., & Krajcik, J. (2021, Marts 26). Using Machine Learning to Score Multi-Dimensional Assessments. *Journal of Science Education and Technology*(30), pp. 239-254. doi:https://doi.org/10.1007/s10956-020-09895-9

McManus, S. (2008). *Attributes of effective Formative Assessment.* Washington, DC: CCSSO (The Council for Chief State School Officers).

Ouyang, F., Dinh, T. A., & Xu, W. (2023, November 9). A Systematic Review of AI-Driven Educational Assessment. *Journal for STEM Education Research, 6*, pp. 408-426. doi:https://doi.org/10.1007/s41979-023-00112-x

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and Extrinsic Motivations: Classic Definitions and New Directions. *Contemporary Educational Psychology, 25*, pp. 54-67. doi:10.1006/ceps.1999.1020

Rønberg, A. (2023). ChatGPT presser danske uddannelser: "Det kan ende i et våbenkapløb". *Ingeniøren* . doi:https://www.version2.dk/artikel/chatgpt-presser-danske-uddannelser-det-kan-ende-i-et-vaabenkaploeb

Skovhus, S., Dupont, F., & Szocska, H. (2023, 1 12). *Hver sjette gymnasieelev benytter chatbot til snyd*. Retrieved Januar 15, 2024, from www.gymnasieskolen.dk: https://gymnasieskolen.dk/articles/hver-sjette-gymnasieelev-benytter-chatbot-til-snyd/

Spronken-Smith, R., & Walker, R. (2010). Can inquiry-based learning strengthen the links. *Studies in Higher Education, 35*(6), pp. 723-740. doi:10.1080/03075070903315502

Unesco. (2022). *K-12 AI curricula. A mapping of government-endorsed AI curricula.* Paris: Unesco. Retrieved 12 12, 2023, from https://unesdoc.unesco.org/ark:/48223/pf0000380602

Unesco, Fengchun, M., & Holmes, W. (2023). *Guidance for generative AI in education and research.* Paris: Unesco. Retrieved December 12, 2023, from https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Vedersø, B., Aslak, M., Andreasen, M., Lauridsen, P. S., Augustinus, I. B., Andersen, H. L., . . . Jensen, T. W. (2023, December 17). Ekspertgruppe: ChatGPT hører til i skolen – bare ikke til alle eksamener. *Altinget*. Retrieved Marts 26, 2024, from

https://www.altinget.dk/uddannelse/artikel/ekspertgruppe-chatgpt-hoerer-til-i-skolen-bare-ikke-til-alle-eksamener

Vittorini, P., Menini, S., & Sara Tonelli. (2021). An AI-Based System for Formative and Summative. *International Journal of Artificial Intelligence in Education, 31*, pp. 159-185. Retrieved from https://doi.org/10.1007/s40593-020-00230-2

Vygotsky, L. S. (1978). Interaction between Learning and Development. In L. Vygotsky, *Mind in Society. The Development of Higher Psychological Processes* (pp. 79-91). USA: Harvard University Press.

Wardat, Y., Tashtoush, M. A., AlAli, R., & Adeeb, J. M. (2023). ChatGPT: A revolutionary tool for teaching and learning mathematics. *EURASIA Journal of Mathematics, Science and Technology Education, 19*(7), pp. 1-18. Retrieved from https://doi.org/10.29333/ejmste/13272

Zhang, P. (2023). *Taking advice from ChatGPT.* Retrieved Juni 23, 2023, from Cornell University: https://arxiv.org/abs/2305.11888

Bilag

Bilag 1: Temaopgave 2. Empiriske metoder/Temakursus, MiSU
Bilag 2: Elevvejledning 1.g
Bilag 3: Elevvejledning 3.g
Bilag 4: Eksempel på elevrapport fra 1.g og 3.g
Bilag 5: Interviewguide
Bilag 6: Rubrik til scoring af forsøgsdesign
Bilag 7: Spørgeskema uden svar 1.g
Bilag 8: Spørgeskema uden svar 3.g
Bilag 9: Pilotspørgeskema med svar
Bilag 10: Spørgeskema med svar 1.g
Bilag 11: Spørgeskema med svar 3.g
Bilag 12: Transskriberede interviews
Bilag 13: GAI-deklaration

Bilag 2

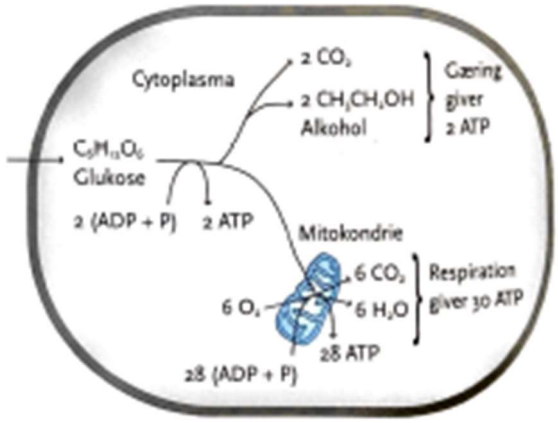
"Gær og CO_2"

Formål

I skal designe et forsøg, hvor I undersøger, hvordan man kan få en gær til at producere mest muligt CO_2 på 30 min.

Baggrundsviden

Figuren nedenfor viser respiration og gæring.



Materialer

- Balloner i forskellige størrelser
- Gær i pakker
- Sukker
- Flasker, der passer til balloner
- Vægt
- Isterninger og varmt vand til at lave vandbade med forskellige temperaturer
- Termometre
- Målebånd
- Mel
- Bægerglas
- Vitawrap
- Elastikker
- Ur
- Alkohol
- Rørepinde

Forsøgsmodul 1

Undersøgelsesspørgsmål

1) **Hvad vil I gerne undersøge om, hvordan gær kan producere mest muligt CO_2?**
    - I skal undersøge noget, som I ikke kender svaret på i forvejen
    - Undersøgelsesspørgsmålet skal kunne undersøges eksperimentelt ud fra de materialer, I har til rådighed
    - Undersøgelsesspørgsmålet skal være formuleret som et spørgsmål

Skriv jeres undersøgelsesspørgsmål her:

Bilag 2

Fremgangsmåde

Krav til jeres forsøgsdesign

Forsøgsdesignet skal:

- kunne teste jeres undersøgelsesspørgsmål
- være så præcist beskrevet, at andre vil kunne udføre forsøget ud fra jeres forsøgsbeskrivelse.
- være udformet, så der kun er én faktor ad gangen, der varierer
- indeholde et kontrolforsøg.

2) Hvordan vil I undersøge jeres undersøgelsesspørgsmål? Lav en detaljeret beskrivelse af, hvordan I vil udføre jeres forsøg. Jo flere detaljer, jo bedre ☺

3) Sæt jeres undersøgelsesspørgsmål og forsøgsdesign ind i teksten nedenfor, og indsæt hele teksten som en prompt i ChatGPT.

Vi skal undersøge følgende undersøgelsesspørgsmål "Sæt jeres undersøgelsesspørgsmål fra punkt 1 ind her". Vi har følgende materialer til rådighed: Balloner i forskellige størrelser, gær i pakker, sukker, flasker der passer til balloner, vægt, isterninger og varmt vand til at lave vandbade med forskellige temperaturer, termometre, målebånd, mel, bægerglas, vitawrap, elastikker, ur, alkohol, rørepinde. Vores forsøgsdesign bliver evalueret ud fra følgende kriterier: Forsøgsdesignet skal kunne teste undersøgelsesspørgsmålet og være så præcist beskrevet, at andre vil kunne udføre forsøget ud fra forsøgsbeskrivelsen. Forsøget skal være designet, så der kun er én faktor ad gangen, der varierer, og der skal indgå et kontrolforsøg. Vores foreløbige forsøgsdesign ser sådan her ud "Sæt jeres foreløbige forsøgsdesign fra punkt 2 ind her". Vær opmærksom på, at der kun er 30 min til hele forsøget. Kan du komme med forbedringsforslag til vores forsøgsdesign?

4) Sæt ChatGPTs svar ind nedenfor, og marker eventuelle relevante forbedringsforslag med grøn farve og eventuelle forkerte eller irrelevante forbedringsforslag med rød farve.

Bilag 2

5) Brug feedbacken fra ChatGPT til at lave en ny beskrivelse af jeres forsøgsdesign, og sæt den ind her:

6) Nu er I klar til at få ChatGPT til at give en sidste vurdering af, om jeres forsøgsdesign er godt nok til, at I kan udføre forsøget. Sæt jeres undersøgelsesspørgsmål og forsøgsdesign fra punkt 5 ind i teksten nedenfor, og indsæt hele teksten som en prompt i ChatGPT.

Vi skal undersøge følgende undersøgelsesspørgsmål "Sæt jeres undersøgelsesspørgsmål fra punkt 1 ind her". Vi har følgende materialer til rådighed: Balloner i forskellige størrelser, gær i pakker, sukker, flasker der passer til balloner, vægt, isterninger og varmt vand til at lave vandbade med forskellige temperaturer, termometre, målebånd, mel, bægerglas, vitawrap, elastikker, ur, alkohol, rørepinde. Vores forsøgsdesign bliver evalueret ud fra følgende kriterier: Forsøgsdesignet skal kunne teste undersøgelsesspørgsmålet og være så præcist beskrevet, at andre vil kunne udføre forsøget ud fra forsøgsbeskrivelsen. Forsøget skal være designet, så der kun er én faktor ad gangen, der varierer, og der skal indgå et kontrolforsøg. Vores foreløbige forsøgsdesign ser sådan her ud "Sæt jeres seneste version af forsøgsdesignet fra punkt 5 ind her". Vær opmærksom på, at der kun er 30 min til hele forsøget. Kan du komme med en vurdering af, om vores forsøgsdesign er godt nok til at besvare undersøgelsesspørgsmålet, eller om vi bør foretage ændringer, inden vi udfører forsøget?

7) Sæt ChatGPTs svar ind nedenfor. Hvis der stadig er behov for forbedringer, så marker de dele af ChatGPTs forbedringsforslag, som I vil bruge til at forbedre jeres forsøgsdesign.

8) Hvis I har fundet relevante forbedringsforslag fra ChatGPT i punkt 7, så brug dem til at lave en ny og sidste beskrivelse af jeres forsøgsdesign. Sæt jeres endelige forsøgsdesign ind her:

Forsøgmodul 2

9) Sæt jeres forsøg op efter beskrivelsen i punkt 8, og tag et par billeder af forsøgsopstillingen. Sæt billederne ind nedenfor.

Resultater

10) Tag billeder af jeres resultater, og sæt dem ind her.

11) Lav et diagram eller en tabel, der viser resultaterne.

12) Skriv en tekst, hvor I beskriver jeres resultater.

Bilag 2

Diskussion

13) Hvordan kan I forklare jeres resultater om gær og CO_2? Inddrag resultaterne og biologifaglig viden i en diskussion af resultaterne.

14) Overvej om fejlkilder eller usikkerheder har påvirket jeres resultater, og om de skyldes jeres forsøgsdesign eller fejl i udførelsen af jeres forsøg. Skriv jeres svar her:

15) Giv et eksempel på minimum én ting, som I ville ændre i jeres forsøgsdesign, hvis I havde mulighed for at udføre forsøget én til gang, og forklar, hvorfor.

Konklusion

16) Her skal I med få sætninger opsummere, hvad I er kommet frem til ud fra jeres undersøgelse af, hvordan man kan få gær til at producere mest mulig CO_2 på 30 min.

Bilag 3

"Smagssansen og tærskelværdi"

Formål

I skal designe et forsøg, hvor I undersøger tærskelværdien for jeres smagssans.

Baggrundsviden

Når man for eksempel spiser noget sødt, påvirker det nogle receptorer, som er i forbindelse med nervesystemet. Hvis koncentrationen af det kemiske stof (fx sukker), som bindes til receptorerne, er høj nok, bliver der skabt en elektrisk impuls, som når over den såkaldte tærskelværdi i de nerveceller, der registrerer smagspåvirkningen. Det betyder, at der bliver udløst et aktionspotentiale, som giver hjernen besked om smagsoplevelsen.

Materialer

- Sukker
- Citroner
- Mel
- Engangspipetter
- Vand
- Vægt
- Glas
- Servietter
- Beholdere til smagsopløsninger
- Balloner
- Skeer

Forsøgsmodul 1

Undersøgelsesspørgsmål

1) Hvad vil I gerne undersøge om tærskelværdien for jeres smagssans?

- I skal undersøge noget, som I ikke kender svaret på i forvejen.
- Undersøgelsesspørgsmålet skal kunne undersøges eksperimentelt ud fra de materialer, som I har til rådighed.
- Undersøgelsesspørgsmålet skal være formuleret som et spørgsmål

Skriv jeres undersøgelsesspørgsmål her:

Bilag 3

Fremgangsmåde

Krav til jeres forsøgsdesign

Forsøgsdesignet skal:

- kunne teste jeres undersøgelsesspørgsmål
- være så præcist beskrevet, at andre vil kunne udføre forsøget ud fra jeres forsøgsbeskrivelse.
- være udformet, så der kun er én faktor ad gangen, der varierer
- indeholde et kontrolforsøg.

2) Hvordan vil I undersøge jeres undersøgelsesspørgsmål? Lav en detaljeret beskrivelse af, hvordan I vil udføre jeres forsøg. Jo flere detaljer, jo bedre ☺

3) Sæt jeres undersøgelsesspørgsmål og forsøgsdesign ind i teksten nedenfor, og indsæt hele teksten som en prompt i ChatGPT.

Vi skal undersøge følgende undersøgelsesspørgsmål "Sæt jeres undersøgelsesspørgsmål fra punkt 1 ind her". Vi har følgende materialer til rådighed: Sukker, citroner, mel, engangspipetter, vand, vægt, glas, servietter, beholdere til smagsopløsninger, balloner, skeer. Vores forsøgsdesign bliver evalueret ud fra følgende kriterier: Forsøgsdesignet skal kunne teste undersøgelsesspørgsmålet og være så præcist beskrevet, at andre vil kunne udføre forsøget ud fra forsøgsbeskrivelsen. Forsøget skal være designet, så der kun er én faktor ad gangen, der varierer, og der skal indgå et kontrolforsøg. Vores foreløbige forsøgsdesign ser sådan her ud "Sæt jeres foreløbige forsøgsdesign fra punkt 2 ind her". Kan du komme med forbedringsforslag til vores forsøgsdesign?

4) Sæt ChatGPTs svar ind nedenfor, og marker eventuelle relevante forbedringsforslag med grøn farve og eventuelle forkerte eller irrelevante forbedringsforslag med rød farve.

5) Brug feedbacken fra ChatGPT til at lave en ny beskrivelse af jeres forsøgsdesign, og sæt den ind her:

Bilag 3

6) Nu er I klar til at få ChatGPT til at give en sidste vurdering af, om jeres forsøgsdesign er godt nok til, at I kan udføre forsøget. Sæt jeres undersøgelsesspørgsmål og forsøgsdesign fra punkt 5 ind i teksten nedenfor, og indsæt hele teksten som en prompt i ChatGPT.

Vi skal undersøge følgende undersøgelsesspørgsmål "Sæt jeres undersøgelsesspørgsmål fra punkt 1 ind her". Vi har følgende materialer til rådighed: Sukker, citroner, mel, engangspipetter, vand, vægt, glas, servietter, beholdere til smagsopløsninger, balloner, skeer. Vores forsøgsdesign bliver evalueret ud fra følgende kriterier: Forsøgsdesignet skal kunne teste undersøgelsesspørgsmålet og være så præcist beskrevet, at andre vil kunne udføre forsøget ud fra forsøgsbeskrivelsen. Forsøget skal være designet, så der kun er én faktor ad gangen, der varierer, og der skal indgå et kontrolforsøg. Vores foreløbige forsøgsdesign ser sådan her ud "Sæt jeres seneste version af forsøgsdesignet fra punkt 5 ind her". Kan du komme med en vurdering af, om vores forsøgsdesign er godt nok til at besvare undersøgelsesspørgsmålet, eller om vi bør foretage ændringer, inden vi udfører forsøget?

7) Sæt ChatGPTs svar ind nedenfor. Hvis der stadig er behov for forbedringer, så marker de dele af ChatGPTs forbedringsforslag, som I vil bruge til at forbedre jeres forsøgsdesign.

8) Hvis I har fundet relevante forbedringsforslag fra ChatGPT i punkt 7, så brug dem til at lave en ny og sidste beskrivelse af jeres forsøgsdesign. Sæt jeres endelige forsøgsdesign ind her:

Forsøgmodul 2

9) Sæt jeres forsøg op efter beskrivelsen i punkt 8, og tag et par billeder af forsøgsopstillingen. Sæt billederne ind nedenfor.

Resultater

10) Tag billeder af jeres resultater, og sæt dem ind her.

11) Lav et diagram eller en tabel, der viser resultaterne.

12) Skriv en tekst, hvor I beskriver jeres resultater.

Diskussion

13) Hvordan kan I forklare jeres resultater om tærskelværdi og smagssans? Inddrag resultaterne og biologifaglig viden i en diskussion af resultaterne.

Bilag 3

14) Overvej om fejlkilder eller usikkerheder har påvirket jeres resultater, og om de skyldes jeres forsøgsdesign eller fejl i udførelsen af jeres forsøg. Skriv jeres svar her:

15) Giv et eksempel på minimum én ting, som I ville ændre i jeres forsøgsdesign, hvis I havde mulighed for at udføre forsøget én til gang, og forklar, hvorfor.

Konklusion

16) Her skal I med få sætninger opsummere, hvad I er kommet frem til ud fra jeres undersøgelse af tærskelværdien for jeres smagssans.