KONSISTENT KARAKTERGIVNING MED KUNSTIG INTELLIGENS

Et eksperimentelt studie om støjreduktion og tillid ved brug af en LL-model som medbedømmer

Anne Balsby Roersen
Line Højris Revsbech
Maria Laursen

Titelblad

Titel:	Konsistent karaktergivning med kunstig intelligens - Et eksperimentelt studie om støjreduktion og tillid ved brug af en LL-model som medbedømmer
Tema:	LLM som medbedømmer i karaktergivningen
Udgivet:	Maj 2024
Vejleder:	Christoffer Koch Florczak
Anslag:	294402
Normalsider:	123
Bilag:	15

AALBORG UNIVERSITET
STUDENTERRAPPORT

Gruppemedlemmer	Specialisering	Studienummer
Anne Balsby Roersen	Social Data Science	20184588
Line Højris Revsbech	Ledelse & Forvaltning	20194884
Maria Laursen	Ledelse & Forvaltning	20195491

Forord

I dette projekt skal der lyde en tak til surveyeksperimentets pilottestere, som har taget sig tid til at afprøve og evaluere eksperimentets indhold. Ydermere skal der lyde en tak til studiets respondenter, som har taget sig tid til at deltage i eksperimentet og bidrage med store indsigter i vurderingspraksissen. En særlig tak skal der lyde til vores vejleder Christoffer Koch Florczak, som gennem kyndig vejledning, kritiske spørgsmål og sine tværfaglige indsigter har dannet grundlag for faglige diskussioner og en øget kvalitet af dette speciale.

1. Abstract

The grade of students' final exams in the Danish elementary school system has increasing importance for their future possibilities. However, the grading of the papers is associated with a great amount of noise in the grading process. Earlier research on the subject has shown that this is especially the case with the final exam in written Danish, where only one assessor grades the final papers, resulting in varying outcomes ranging from the bottom to the top grade on the Danish grading scale. This tendency is alarming; it can potentially lead to distrust in the grading and educational system and have a substantial and lasting impact on the students' futures. This study seeks to examine whether recent advances in large language models have paved the way for using these models in grading. Moreover, the study aims to add to the existing literature regarding human-AI collaboration and the role trust plays in this collaboration. Therefore, this study seeks to determine whether an LLM can reduce the noise in grading the final exam in written Danish in the Danish elementary school system. The method of the study is based on a survey experiment.

This study also found noise - and therefore variation in the grading of the final papers, but when using the LL-model as a co-assessor the study showed how the noise could be decreased. It also showed a 35 % decrease in noise of the grading, regardless of whether the co-assessor is believed to be an AI or a human. Furthermore, the LL-model was found to be less noisy than the average Danish teacher. Additionally, this study observed a negative effect from trust on the under-use of the co-assessor in grading the papers and under-use of LLMs. This study found evidence of higher distrust in AI models compared to a human co-assessor.

The study concludes that even though information about an AI co-assessor leads to more under-use, it does not necessarily impact the overall noise reduction in the final grade. Finally, the study discusses the consequences of implementing AI models in the grading system of the Danish elementary school, highlighting points of attention. These include the significance of knowledge and trust in the LL-models, which is of great importance in the use or disuse of the models. This leads to the study’s final recommendation: sharing knowledge concerning the competencies of the models and the continued professional responsibility of the teachers in grading is crucial.

2. Indholdsfortegnelse

3. Indledning

"Det er rart at have en medbedømmer (...) med en medbedømmer bliver bedømmelsen mere fair overfor eleven (...) fejl kan altid ske og er hyppigere ved en bedømmer frem for to"

Citatet stammer fra en af studiets surveyrespondenter. Det er med til at italesætte udfordringen ved den nuværende bedømmelsespraksis ved folkeskolens skriftlige afgangseksamener, som blev indført som følge af folkeskolereformen i 2014. Hvor det tidligere var to bedømmere, som havde ansvaret for at vurdere en elevs præstation og fastsætte en karakter, så hviler denne byrde i dag på skuldrene af én bedømmer. At karakterfastsættelsen udelukkende baseres på én censors vurdering er problematisk, når studier viser, hvordan karaktergivningen særligt i dansk skriftlig fremstilling ved folkeskolens afgangseksamen er præget af en varierende karaktergivning. Med andre ord betyder det hermed, at lignende præstationer vurderes til forskellige karakterer, hvilket er med til at betvivle karaktergivningen, som burde være baseret på ensartethed og pålidelighed.

I Danmark er karakterer en uundgåelig del af det danske uddannelsessystem. Det starter allerede i folkeskolen, hvor eleverne for første gang i udskolingen oplever at blive bedømt med karaktererne. Disse har endvidere afgørende betydning videre i uddannelsessystemet, hvor karaktererne besidder en adgangsgivende indflydelse på de unges uddannelsesmuligheder. Det foreligger hermed yderst problematisk, at karakterne har en afgørende betydning for de unges fremtid, når karaktererne i princippet forekommer tilfældige og upålidelige.

Gennem tiden har karakterer været til megen debat. Afsættet for nærværende studie er ikke, at karakterer i folkeskolen bør afskaffes. Det handler derimod om at undersøge, hvorvidt der findes mulige løsninger, hvor bedømmelsen af elevernes præstationer ved folkeskolens afgangseksamen kunne blive mere ensartede og pålidelige - og hvor censorerne vil opleve en sparring og understøttelse af deres faglighed.

God læselyst!

Anne Balsby Roersen, Line Højris Revsbech og Maria Laursen

4. Problemanalyse og eksisterende forskning

I det følgende kapitel vil studiets problemanalyse blive fremlagt. Her vil det indledningsvist blive tydeliggjort, hvordan karakterer siden årtusindskiftet har fået en stigende betydning. Med en forståelse for denne betydning vil det i forlængelse heraf blive klarlagt, hvordan ensartetheden og pålideligheden i karaktergivningen er udfordret særligt ved dansk skriftlig fremstilling i folkeskolen. Med henblik på at styrke karaktergivningen vil det efterfølgende blive belyst, hvordan de nyere computationelle metoder vil kunne bidrage hertil, og hvordan en anvendelse af disse nyere metoder kan betragtes som en tillidsafhængig relation. Af den grund vil tilliden til disse teknologier og interaktionen blive belyst i problemanalysens sidste afsnit.

4.1 Karakterernes betydning siden årtusindeskiftet

Karaktererne i det danske uddannelsessystem har siden årtusindeskiftet oplevet en øget opmærksomhed og betydning for børn og unge i uddannelsessystemet. Den stigende opmærksomhed kan siges at være forbundet med en række begivenheder. Først og fremmest kan det tydeliggøres, hvordan år 2000 var året, hvor den første PISA-måling blev gennemført i Danmark. Dette betød dermed, at den danske folkeskole fra år 2000 kunne måles og sammenlignes både nationalt og internationalt. Som følge af '*Loven om gennemsigtighed og åbenhed i uddannelserne*' blev der i 2005 for første gang udført en systematisk og offentlig statistik med henblik på at vurdere kvaliteten af undervisningen på de enkelte skoler og institutioner. Dette resulterede i, at alle borgere i dag kan tilgå offentlig tilgængelige statistikker om forskelle i karaktergennemsnit på tværs af skoler, køn, etniske grupper m.m. (Andreasen, 2020: 48-49; *Bekendtgørelse af lov om gennemsigtighed og åbenhed i uddannelserne m.v.,* 2005). Lovgivningen medførte dermed et nyt, målbart og fuldt tilgængeligt konkurrenceparameter i folkeskolen, hvilket var med til at øge opmærksomheden omkring karaktererne.

4.1.1 Karakterer som sorteringsnøgle i uddannelsessystemet

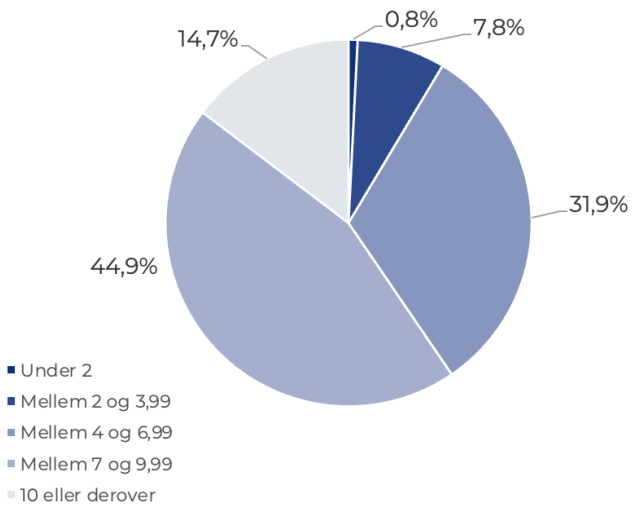
Foruden ovenstående kan det også argumenteres, at den øgede opmærksomhed omkring karakterer kan relateres til deres betydning for unges udvalg af uddannelsesmuligheder. Dette skyldes, at karakterne anvendes som sorteringsnøgle gennem hele det danske uddannelsessystem, hvor karaktererne bliver brugt til at differentiere mellem eleverne. Når danske folkeskoleelever i 9.

klasse modtager de afsluttende karakterer ved folkeskolens afgangseksamen, så er det for mange unge den første gang, at de modtager en kvantitativ bedømmelse, der får direkte betydning for deres muligheder videre i uddannelsessystemet. Dette bunder i, at mulighederne for en ungdomsuddannelse afhænger af de unges karaktergennemsnit fra folkeskolen. Ud fra formålet om at styrke optaget på de danske erhvervsuddannelser, indførte regeringen i 2019 et skærpet adgangskrav til de gymnasiale uddannelser samtidig med, at kravet om uddannelsesparathed vedblev. For at blive optaget på STX, HHX eller HTX skal eleven have et karaktergennemsnit på 5,0, mens direkte optagelse på HF kræver et snit på 4,0, hvor det på erhvervsuddannelserne gør sig gældende, at der er et karakterkrav på 02 i dansk og matematik (Hjorth-Trolle & Jonassen, 2023; Regeringen, 2014). Hermed kan det tydeliggøres, hvordan karaktererne har en betydning for optaget på alle uddannelsesniveauer i det danske uddannelsessystem, hvormed karaktererne også besidder en rolle i at skabe et effektivt uddannelsessystem. Dette sker ud fra en forståelse af, at karaktererne er med til at sikre, at eleverne fordeles og ender på den uddannelsesinstitution, som matcher deres kvalifikationer og kompetencer, hvormed karakterens funktion som sorteringsnøgle sikrer den bedste benyttelse af samfundets ressourcer.

4.1.2 Adgangskravets betydning for eleverne

I det forrige afsnit blev det belyst, hvordan opmærksomheden omkring og betydningen af karaktererne i folkeskolen er steget, hvor karakterer i dag har fået en direkte betydning for elevernes videre muligheder i uddannelsessystemet. Men hvor stor en andel af eleverne får begrænset deres uddannelsesmuligheder på baggrund af deres karakterer fra folkeskolen? Dette tydeliggøres i følgende figur, hvor fordelingen af karaktergennemsnit ved folkeskolens afgangseksamen i skoleåret 2022/23 er visualiseret.

Figur 4.1: Fordelingen over karaktergennemsnittet til folkeskoleelevernes afgangseksamen 2022/2023



Kilde: Uddannelsesstatistik (2023: 2)

På figuren kan det ses, hvordan 8,6 % af de danske folkeskoleelever, der var til afgangseksamen i sommeren 2023, ikke havde mulighed for frit at vælge ungdomsuddannelse som følge af de tidligere nævnte adgangskrav vedrørende karaktergennemsnit (Børne- og Undervisningsministeriet, 2023a). Ydermere gør det sig også gældende, at 10 % af eleverne ikke bestod enten dansk eller matematik, hvilket forhindrer deres direkte adgang til erhvervsuddannelserne. Eksempelvis gælder dette for hele 26 % af eleverne ved folkeskolens afgangseksamen i Lollands Kommune (DI, 2023). Hermed kan det klarlægges, hvordan karaktergennemsnittet i folkeskolen udgjorde en aktiv sorteringsnøgle for knap 10 % af de elever, som afsluttede folkeskolen i skoleåret 2022/23.

Foruden at de nuværende karakterkrav fratager omtrent 10 % af eleverne i den danske folkeskole muligheden for selv at vælge ungdomsuddannelse, bør det videre bemærkes, at regeringen i foråret 2024 fremlagde planer om at hæve det gymnasiale adgangskrav yderligere til et karaktergennemsnit på 6,0 fremfor det nuværende på 5,0. Sådan en skærpelse i adgangskravene vil medføre, at 20 % af de danske folkeskoleelever vil opleve, at karaktergennemsnit fra folkeskolen begrænser de videre muligheder i uddannelsessystemet (Mainz, 2024).

4.2 Støj i karaktergivningen

Med en forståelse for karakterernes stigende betydning qua deres rolle som sorteringsnøgle i unges uddannelsesmuligheder, foreligger det problematisk, når det gennem forskning bliver tydeligt, hvordan der kan stilles tvivl ved, hvorvidt karaktergivningen særligt i den danske folkeskole forekommer ensartet og pålidelig (Dolin et al., 2018; Troelsen, 2020; Roersen et al., 2022).

4.2.1 Varians i karaktergivning

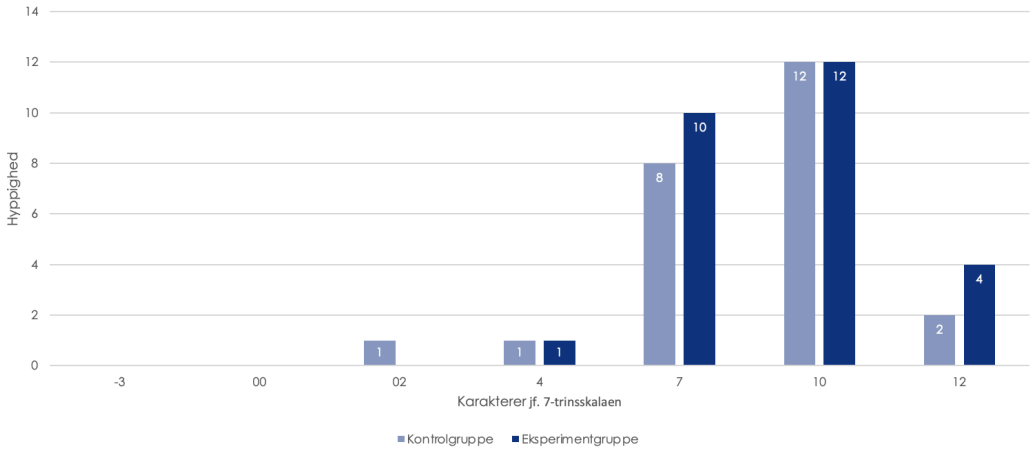
Troelsen (2020) klarlægger i sin ph.d.-afhandling, hvordan karaktergivningen ved folkeskolens afgangseksamen i dansk skriftlig fremstilling kan variere fra bedømmer til bedømmer. I Troelsens (2020: 154) afhandling blev seks beskikkede censorer bedt om at læse og vurdere syv danskopgaver fra folkeskolens afgangsprøve i dansk skriftlig fremstilling. Her viste det sig, at selvom det var de samme danskopgaver, der blev vurderet, så udmærkede bedømmelserne sig med vidt forskellige karakterer. Den opgave, som én censor fandt fremragende og bedømte til karakteren 12, vurderede en anden censor som værende en jævn præstation og gav karakteren 4. En anden danskopgave ville én censor bedømme til karakteren 4 og dermed en jævn præstation, mens to andre censorer ville dumpe opgaven som værende en utilstrækkelig præstation og give karakteren 00 (jf. Børne- og Undervisningsministeriet, n.d.a).

Med afsæt i ovenstående tydeliggør Troelsen (2020: 156-157) på bekymrende vis i sin forskning, hvordan der foreligger en varians i karaktergivningen i dansk skriftlig fremstilling i folkeskolen, hvor danskopgaver bedømmes med varierende karakterer afhængig af den pågældende bedømmer. At der foreligger en varians i karaktergivningen kan argumenteres at være problematisk, når det kommer til eleverne og deres fremtidsmuligheder. Foruden at variansen kan være til ulempe for individet og dets fremtidsmuligheder, så kan det argumenteres, hvordan denne manglende ensartethed kan foreligge problematisk i et demokratisk perspektiv. Dette skyldes, at eleverne i den danske folkeskole ud fra en meritbaseret tilgang skal bedømmes retfærdigt og forekomme ligestillede i bedømmelsen, hvilket kræver en ensartet karaktergivningsproces af høj kvalitet (jf. Troelsen, 2020: 153). Denne manglende ensartethed og dermed retfærdighed i karaktergivningen kan i værste tilfælde underminere de demokratiske principper

omkring lige muligheder for alle borgere, hvilket videre kan have en negativ indflydelse på tilliden og opbakningen til de demokratiske institutioner (jf. Nedergaard, 2022).

I tråd med Troelsen (2020) klarlægger Roersen et al. (2022) også på bekymrende vis gennem et surveyeksperiment, hvordan karaktergivningen i dansk skriftlig fremstilling i folkeskolen er præget af varians. I surveyeksperimentet blev 51 dansklærere tilfældigt inddelt i en kontrol- og eksperimentgruppe, hvor begge grupper blev bedt om at læse og bedømme en dansk skriftlig afgangsprøve fra 9. klasse. Her skulle kontrolgruppen bedømme danskopgaven på normalvis i form af en holistisk bedømmelse, hvor danskopgaven vurderes som en helhed. Eksperimentgruppen blev derimod bedt om at opdele deres bedømmelse og give seks delkarakterer. Denne opdelte karaktergivning kaldes en analytisk vurdering. Bedømmelserne er vist nedenfor:

Figur 4.2: Fordelingen af kontrol- og eksperimentgruppens bedømmelser



Figur 4.2 viser fordelingen af karakterer i Roersen et al.'s (2022: 111) surveyeksperiment.

Af figuren fremgår det, hvordan den selvsamme danskopgave i surveyeksperimentet blev bedømt til alle karakterer fra 02 til 12. Hermed påviser Roersen et al. (2022: 110-112) i tråd med Troelsen (2020), hvordan karaktergivningen i folkeskolen er udfordret, når det kommer til ensartethed og pålidelighed som følge af ovenstående varians.

Troelsen (2020) og Roersen et al.'s (2022) undersøgelsesresultater er herved konsistente med Kahneman et al.s (2021) teoretiske værk om 'støj'. Teorien om støj

tydeliggør, hvordan beslutninger i høj grad er baseret på tilfældighed frem for præcision, når simple såvel som komplekse beslutninger varierer fra individ til individ selvom, at disse i princippet burde være overensstemmende. Ifølge Kahneman et al. (2021) skal støj herved forstås som den usystematiske varians, der er mellem vurderinger foretaget af mennesker. Til forskel fra støj, består bias i højere grad af systematiske fejl, hvormed bias ikke på samme måde er baseret på tilfældigheder. Dermed er forskellen på bias og støj, at bias er systematisk, mens støj er vilkårlig. Af den grund vil støj ikke kunne forudsiges på samme måde som bias til trods for, at de begge er menneskelige fejl (Kahneman et al., 2021: 10). Kahneman et al. (2021) påpeger dog, hvordan bias kan føre til støj. I tilfælde hvor forskellige bedømmere begår forskellige systematiske fejl, så vil disse systematiske fejl i en aggregeret form fremstå som støj. På den måde vil bias kunne resultere i støj, mens dette ikke vil kunne gøre sig gældende omvendt. Ifølge Kahneman et al. (2021: 12-16, 516) vil der altid være støj i menneskelige vurderinger. Dette gælder i alt lige fra lægevidenskaben til retsvæsenet til karaktergivningen i folkeskolen. Uanset hvor støjen optræder, så påpeger Kahneman et al. (2021), hvordan støj er en udfordring, der kan have store omkostninger. Det er her afgørende for beslutningens pålidelighed, at støjen reduceres mest muligt. Dog er et fuldkomment fravær af støj i menneskelige vurderinger hverken opnåeligt eller ønskværdigt (Kahneman et al., 2021: 14-16). Men vil det være muligt at reducere støjen i karaktergivningen i folkeskolen?

4.2.2 Praksis ved karaktergivningen i folkeskolen

Både Troelsen (2022) og Roersen et al. (2022) fremlægger på bekymrende vis, hvordan karaktergivningen i dansk skriftlig fremstilling i folkeskolen er påvirket af støj. For begge undersøgelser gør det sig gældende, at der sættes fokus på karaktergivningen i folkeskolen. At folkeskolen er genstandsfeltet skyldes, at bedømmelsespraksissen ved folkeskolens afgangsprøve forholder sig anderledes sammenlignet med de øvrige institutioner i uddannelsessystemet. Dette skyldes, at der som følge af folkeskolereformen i 2014 blev indført en ny ordning ved bedømmelsen af skriftlige prøver i 9. og 10. klasse (jf. folkeskoleloven, 2014). Før folkeskolereformen gjorde det sig gældende, at skriftlige prøver blev bedømt af både elevens lærer og en beskikket censor. Denne praksis blev dog ændret som følge af reformen i 2014, hvilket betyder, at de skriftlige prøver i 9. og 10. klasse

alene bedømmes af én beskikket censor i dag (jf. §53 i prøvebekendtgørelsen, 2019). Med indførelsen af enebedømmerordningen adskiller folkeskolen sig fra resten af det danske uddannelsessystem, idet folkeskolen er den eneste uddannelsesinstitution, som udelukkende baserer karaktergivningen ved skriftlige prøver på én bedømmelse. Men hvad betyder det for karakterernes pålidelighed mere konkret?

Når det kommer til karakterernes pålidelighed, viser Dolin et al.'s (2018) rapport, hvordan disse er udfordret med enebedømmerordningen. I rapporten klarlægges det i tråd med Troelsen (2020) og Roersen et al. (2022), at karakterer kan variere fra en censor til en anden ved de skriftlige prøver. Undersøgelsen var baseret på 150 elevopgaver fra sommeren 2016 og viste, at sandsynligheden for, at to censorer vil give den samme karakter for en skriftlig danskopgave, var 40 %, mens det for en skriftlig matematikopgave var 72 % (Dolin et al., 2018: 17). At matematik udmærker sig med en højere sandsynlighed for samme karakter skyldes, at matematik er et fag, der i højere grad baserer sig på entydige krav og retningslinjer sammenlignet med dansk. I matematik bedømmes opgaverne hovedsageligt ud fra et facit, hvilket ikke på samme måde gør sig gældende i dansk. I dansk skal censorerne i højere grad bruge deres fagprofessionelle skøn til at bedømme opgaven, som grundet subjektivitet i højere grad kan støje og dermed påvirke karaktergivningens pålidelighed negativt. Dette sker ud fra en forståelse, at et fagprofessionelt skøn kan anses som en vurdering, der befinder sig et sted mellem regler og subjektiv viden (jf. Møller et al., 2021: 83).

Rapporten fra Dolin et al. (2018) er med til at tydeliggøre, hvordan karaktergivningens pålidelighed udfordres, når bedømmelsen udelukkende baseres på en subjektiv bedømmelse. Dette kan anskues som en udfordring, idet enebedømmerordningen blev indført som den gældende praksis ved karaktergivningen som følge af folkeskolereformen. At to censorer 60 % af gangene ikke vil vurdere den samme elevtekst i dansk skriftlig fremstilling til samme karakter kan herved siges at være med til at indikere støj i karaktergivningen. Dette kan betragtes som problematisk, idet eleverne i princippet burde være ligestillede (jf. Holm-Larsen & Plischewski, 2017: 12-13; Kahneman et al., 2021: 416). Men hvordan kan denne støj i karaktergivningen reduceres?

Af litteraturen kan det udledes, hvordan det ikke forekommer overraskende, at karaktergivningen er påvirket af støj, når denne udelukkende baseres på én bedømmelse (jf. Galton, 1886; Pearson, 1924: 400, Pearson, 1930: 581; Kahneman & Tversky, 1974; Campbell & Kenny, 1999; Kahneman et al., 2021: 101-104). Dette sker ud fra en forståelse af, at en karakter baseret på én bedømmelse altid vil forekomme mere støjende end en karakter baseret på to uafhængige bedømmelser. Årsagen hertil er, at hvis karaktergivningen foretages ud fra gennemsnittet af to uafhængige bedømmelser, så vil den samlede bedømmelse altid bevæge sig mod gennemsnittet. Denne implikation stammer oprindelig fra statistikken og omtales som *'regression mod gennemsnittet'*. Dette fænomen handler om, at ekstreme målinger vil have en naturlig tendens til at blive efterfulgt af en knap så ekstrem og mere middelmådig måling, hvormed disse vil bevæge sig tættere på det generelle gennemsnit (jf. Galton, 1886; Pearson, 1924: 400, Pearson, 1930: 581; Kahneman et al., 2021: 101-104). Siden Galton (1886) introducerede dette statiske princip første gang i det 19. århundrede, er det blevet overført til en række andre områder såsom beslutningsprocesser, hvormed dette også kan påpeges at have relevans, når det kommer til karaktergivning. Med en forståelse for fænomenet omkring 'regression mod gennemsnittet' kan det derfor argumenteres, hvordan det vil være ønskværdigt, at der indgik to bedømmere i karaktergivningsprocessen ved dansk skriftlig fremstilling i folkeskolen. Men er det en mulighed? Og hvad vil det kræve?

4.2.3 Hvad med den politiske opbakning?

Fra år 2024 og frem mod år 2029 har regeringen afsat 2,690 millioner kroner årligt til et kvalitetsløft af folkeskolen (Børne- og Undervisningsministeriet, 2023b: 1). Dette fremgår af udspillet *'Forberedt på fremtiden II'*, som er regeringens samlede plan for at skabe en styrket folkeskole med fokus på frihed og fordybelse (Regeringen, 2023: 4-5). Udspillet er med til at tydeliggøre, hvordan der fra politisk side forekommer en villighed til at investere i folkeskolen. Dog er der i den forbindelse ikke afsat midler til et kvalitetsløft af karaktergivningen. Set i lyset af at karaktergivningen er et område, hvor dele af den siddende regering med folkeskolereformen tidligere har foretaget besparelser ved at netop indføre enebedømmerordningen, må det antages, at mulighederne for en ændring af

dette på nuværende tidspunkt efter al sandsynlighed vil være lille. Men kunne det tænkes, at der fandtes et mere overskueligt, omkostnings- og ressourcelet initiativ med to bedømmere, som vil kunne reducere støjen i karaktergivningen? Måske kunne den seneste tids fremadstormende udvikling inden for kunstig intelligens bidrage til, at enebedømmerordningen bliver mindre støjende?

4.3 ML's betydning og anvendelse

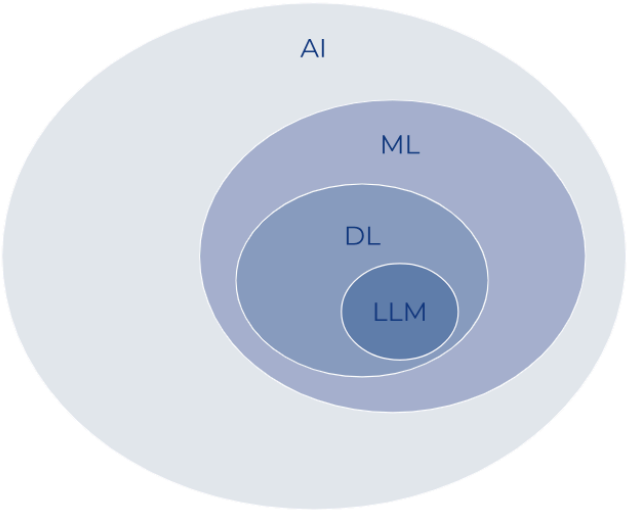
I de forrige afsnit er det blevet klarlagt, hvordan karakterer har fået en øget opmærksomhed og betydning siden årtusindskiftet. At karaktererne besidder en nøglerolle i de unges uddannelsesmuligheder kan anskues som problematisk, når flere undersøgelser påviser, hvordan karaktergivningen i folkeskolen er præget af støj (jf. Dolin et al., 2018; Troelsen, 2020; Roersen et al., 2022). Med henblik på at mindske støjen i vurderinger fremlægger Kahneman et al. (2021: 101, 363-365), at to uafhængige vurderinger sammen kan bidrage til støjreduktion i beslutningsprocesser eksempelvis ved karaktergivning. Kahneman et al. (2021: 164-174) påpeger dog, at disse uafhængige vurderinger ikke nødvendigvis skal foretages af mennesker, idet modeller baseret på algoritmer også vil kunne benyttes både alene og i samspil med en menneskelig vurdering. Med afsæt heri opstår der dermed en nysgerrighed i at undersøge, hvorvidt det er muligt at reducere støjen i karaktergivningen i folkeskolen ved brug af nyere AI-teknologi samt, hvilke fordele og ulemper dette vil have.

4.3.1 Udviklingen af større sprogmodeller

For at afdække hvorvidt kunstig intelligens (AI) i dag kan bruges til at foretage vurderinger i karaktergivningen, er det nødvendigt at forstå, hvordan AI-teknologien herunder machine learning (ML) har udviklet sig de seneste år. Machine learning (ML) skal ses som et begreb for de metoder under AI-teknologien, som tillader, at et program eller en model kan 'lære' ved at genkende sammenhænge i datasæt. Fordelen ved denne 'læring' er, at modellen ikke kræver programmering af hvert eneste element og sammenhæng for at kunne komme med et output, hvilket ellers tidligere har været tilfældet (Géron, 2022: 4). I 2006 blev den første komplekse ML-model kaldet 'dybe neurale netværker' trænet af Geoffrey Hinton (jf. Hinton et al., 2006). De dybe netværker skabte derved et subfelt af ML-modellerne. Disse netværker er grundet deres

kompleksitet karakteriseret ved at være bygget af ikke-transparente funktioner, bias og vægte, hvorfor det ikke er muligt at fortolke direkte på sammenhængen mellem modellens input og output (Geron, 2022: 131). Hinton et al.'s (2006) 'dybe neurale netværk' blev skelsættende for, at den dybe læring i dag kan anvendes til mere komplekse opgaveløsninger. Til forskel fra tidligere er styrkerne ved maskinlæring gennem de dybe netværker, at modellerne evner at genkende komplekse sammenhænge i større datamængder, hvor menneskelige kognitive ressourcer kan være begrænset (Géron, 2022: 7; Faraj, Pachidi, & Sayegh, 2018: 64). Den seneste banebrydende udvikling inden for DL-feltet har været udviklingen af 'transformers'-modellerne. Disse består af en modelarkitektur, hvori flere dybe netværksmodeller kombineres i én samlet model. Udviklingen heraf har gjort det muligt, at AI-modeller nu kan kapere og generere mere komplekse data såsom længere tekster, billeder og lyde (Rokani & Kaminaris, 2020). Disse transformermodeller har blandt andet udvist særlige kompetencer i at kunne håndtere og generere sprog i mere komplekse og bredere sammenhænge. Dette har ledt frem til, at der de seneste år har været en stor fremvækst af særligt de større og komplekse sprogmodeller kaldet 'large language models' (LLM), hvor særligt offentliggørelsen af ChatGPT kan siges at have igangsat diskussioner om mulighederne ved brug af LLM i problemløsning på tværs af forskellige sektorer (Duan et al., 2019; Bach et al, 2019: 168; Johansson, 2023; Chandler et al., 2011; Obermeyer & Emanuel, 2016; Kleinberg et al., 2015). På figur 4.3 er det visualiseret, hvordan LL-modellerne udgør en undergren inden for AI:

Figur 4.3: AI-modellernes hierarki



Dét at de nyere LLM'er er blevet så komplekse, at de kan nedbryde og analysere komplekse inputdata, giver anledning til at overveje, hvorvidt denne evne til at håndtere kompleksitet kan anvendes til at bedømme danskopgaver ved folkeskolens afgangseksamen mere ensartet og pålideligt. Når det kommer til karaktergivning gør det sig gældende, at de amerikanske stater Ohio og Utah allerede gør brug af LL-modeller til karaktergivningen. Her anvendes LLM'erne både til at foretage formative og summative vurderinger samt give feedback til flere millioner skriftlige opgaver årligt (Ramesh & Sanampudi, 2022). Dette giver hermed også anledning til at overveje hvilke muligheder og begrænsninger, der kan være ved brugen af LLM i karaktergivningen ved folkeskolens afgangseksamen i dansk skriftlig fremstilling, og hvorvidt dette vil kunne reducere støj-problemet.

4.3.2 Eksisterende litteratur om styrker og svagheder ved LL-modeller

Med henblik på at undersøge, hvorvidt en LL-model kan anvendes som værktøj til støjreduktion i karaktergivningen i folkeskolen, vil det følgende afsnit præsentere eksisterende forskning på området med formål om at belyse, hvilke styrker og svagheder disse modeller har, samt hvordan dette kan have betydning for modellens evne til at foretage en vurdering af en dansk skriftlig fremstilling i folkeskolen.

Inden for forskningslitteraturen er der siden 1973 blevet undersøgt muligheden for at anvende computationelle metoder til bedømmelsen af skriftlige opgaver. Dette forskningsfelt omtales som 'automatiseret essay scoring' (AES). En af de første færdigheder, som blev udviklet inden for AES var grammatikkontrollen, der blev benyttet til at bedømme skriveegenskaber i form af grammatik, diktion og konstruktion (jf. Ajay et al., 1973). Allerede i 1990'erne stod det tydeligt, at AI-modellerne var stærkest i at kontrollere systematiske fejl som grammatik i elevers skriftlige arbejde (Brock, 1990). For at undersøge, hvorvidt en automatiseret grammatikkontrol kan erstatte den menneskelige bedømmelse, har Crossley et al. (2019) videre fundet en høj grad af sammenhæng mellem den fagprofessionelles rettelser og AI-modellens rettelser. Selvom AI-grammatikkontrollen har været under en stor udvikling og forbedring, viser Nova (2018) videre, at

AI-genererede kontroller af grammatikken ligesom menneskelige bedømmelser ikke forekommer uden fejl.

Trods AI-modellens styrke i grammatikkontrol bliver det gennem litteraturen også tydeligt, hvordan AI-modellerne har begrænsninger. Disse begrænsninger kommer blandt andet til udtryk ved, at AI-modellens kompetencer er varierende alt efter sprog. Shehab et al. (2018) belyser, hvordan mange af de automatiserede karaktergivningssystemer, som er blevet testet i løbet af de seneste årtier, er blevet udviklet på engelsksproget tekstdata, hvilket kan anskues som en udfordring for modellernes evne til at foretage vurderinger af mindre sproggrene (Shehab et al., 2018). På baggrund heraf kunne det tyde på, at en AI-model vil have bedre forudsætninger for at foretage en vurdering af en engelskopgave sammenlignet med en danskopgave.

Foruden at AI-modeller har vist sig at have udfordringer i bedømmelserne af mindre sproggrene, viser Ramesh & Sanampudi (2022), hvordan det i løbet af det seneste årti har været muligt at udvikle pålidelige computationelle værktøjer til karaktergivning i 'multiple choice'-opgaver. De påpeger dog, at udviklingen af computationelle metoder til at bedømme komplekse opgaveløsninger ikke er tilstrækkeligt avancerede endnu. Dette skyldes, at AI-modellerne er udfordret, når det kommer til at evaluere på sammenhæng og relevans i det skriftlige produkt. Omvendt italesætter Ramesh & Sanampudi (2022) dog fordelene ved et domænespecifikt træningsdata, hvorved modellen i højere grad kan blive i stand til at bedømme forskellige sproglige stilistiske genrer.

Med henblik på at undersøge, hvorvidt pålideligheden af karaktergivningen kan styrkes ved at bruge AI-modeller, har Hannah et al. (2023) undersøgt korrelationen mellem to menneskelige bedømmere samt en menneskelig bedømmer og en AI-model. Studiet viste, at der på tværs af alle tre vurderingsområder forekom højere korrelation mellem bedømmelserne, hvis en AI-model var medbedømmer, end hvis det var et menneske. Ud fra dette studie tyder det hermed på, at der forekommer større enighed mellem en AI-model og den menneskelige bedømmelse end, hvad der gør sig gældende for to menneskelige bedømmelser. Selvom det tyder på, at AI-modellen kan være en mere pålidelig medbedømmer, så bliver det gennem Miller (2018) og Wilson &

Daugherty (2018) tydeliggjort, hvordan det vil være nødvendigt med et samarbejde mellem en AI-model og en censor. Dette skyldes, at AI-modellen ikke vil være i stand til at se andre mønstre og kontekster end dem, som indgår i træningen af modellen. Denne begrænsning vil en censor derimod kunne kompensere for, idet en menneskelig bedømmer i højere grad kunne tage dette i betragtning i vurderingen.

I tråd med ovenstående klarlægger Cremer og Kasparov (2021) også, at modeller og mennesker ikke besidder de samme evner, hvorfor modeller ikke bør erstatte mennesker fuldkommen. Ydermere påpeger Mosier & Skitka (1996), hvordan opgaveløsningen gennem et samarbejde vil kunne udføres med en højere kvalitet end, hvad modellen eller mennesket på egen hånd kunne have præsteret. Et samarbejde skal derfor anses som fordelagtigt, fordi modellerne kan være med til at udbygge menneskets evner. Ifølge Puranam (2021) kræver dette dog, at der foreligger en klar arbejdsdeling mellem modellen og den fagprofessionelle med henblik på, at de hver især udfører de opgaver, som de er bedst til. Hermed er litteraturen med til at belyse, hvordan et samarbejde kunne være til gavn i karaktergivningsprocessen, hvis samarbejdet indebærer en komplementering af de to bedømmeres styrker og svagheder (jf. Parasuraman & Riley, 1997; Lee & See, 2004: 55; Madhavan & Wiegmann, 2007; De-Arteaga et al., 2020; Lebovitz et al., 2022; Børne- og Undervisningsministeriet, 2022: 21).

4.3.3 Manglende gennemsigtighed i vurderingen

Gennem ovenstående afsnit er det fremlagt, hvordan det vil være mest fordelagtigt for karaktergivningens pålidelighed, hvis en LL-model indgik i et samarbejde med en fagprofessionel bedømmer. Et andet argument for, at en karaktergivning bør ske i kraft af både en fagprofessionel og AI-baseret vurdering kan ses i forlængelse af LL-modellernes manglende gennemsigtighed. Dette omtales gennem litteraturen som 'black box'-fænomenet. Udtrykket 'black box' anvendes, idet LL-modellernes beslutningsproces forekommer uigennemsigtige, hvorved modellens parametre og beregninger bliver for komplekse til at den menneskelige fortolkning kan forstå sammenhængen mellem input og output (Lundberg og Lee, 2017). Denne manglende transparens kan fremhæves som et problem i en forvaltningsmæssig kontekst. Dette sker ud fra en forståelse af, at

den offentlige sektor baserer sig på et princip om, at alle processer i den offentlige sektor skal forekomme saglige, lige og gennemsigtige for alle borgere (KL, 2017). Hermed kan det udledes, hvordan den manglende transparens ved en LL-model kan udfordre brugen heraf ved karaktergivning i folkeskolen, da dette vil foregå inden for rammerne af forvaltningen i den offentlige sektor. At den manglende transparens ved LL-modeller særligt er en udfordring i den offentlige sektor, skyldes problematikken i henhold til at opretholde borgernes retssikkerhed. Her gør det sig gældende, at modeller ikke på samme måde kan retsforfølges ligesom mennesker, da beslutningsprocessen vil være mere besværlig at kortlægge (jf. Akhtar, 2021: 118; Motzfeld, 2021: 164-166). Herigennem kan det argumenteres, hvordan lærernes brug af en LLM kan være begrænset af denne manglende gennemsigtighed.

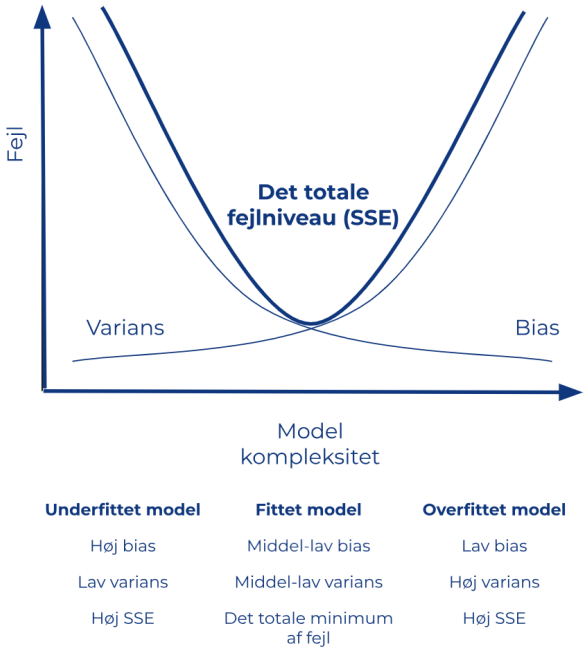
Kritikere af denne black-box-skepsis pointerer dog, at den menneskelige beslutningsproces kan forekomme lige så ugennemsigtig som ved dybere netværker. Med andre ord vil det ligeledes heller ikke være muligt at kortlægge alle de parametre, som har betydning for en menneskelig beslutning (Bonezzi et al., 2022). Det er dog væsentligt at bemærke, hvordan nyere forskningsmetoder arbejder med at åbne denne 'black box', således det bliver mere gennemskueligt, hvad der har betydning for de dybere netværkers output (Lundberg og Lee, 2017). Med de nye muligheder for at 'åbne' modellen er det herved blevet lettere at kortlægge rationalet bag en algoritmisk bedømmelse, hvilket kan være med til at styrke modellens berettigelse i karaktergivningen.

4.3.4 Modellernes trade-off i fejlniveauet

Den sidste begrænsning, der kan italesættes ved brugen af LL-modeller i karaktergivningen, omhandler modellernes fejlniveau. Tidligere blev det beskrevet, hvordan der i beslutningsprocesser kan forekomme fejl i form af bias og varians (støj). Inden for ML-litteraturen opereres der med de samme de begreber for fejl, men her er det afgørende, at ML-litteraturens brug af begreberne 'bias' og 'varians' er forskellig fra den, der findes i litteraturen om beslutningsprocesserne. Hvor Kahneman et al. (2021) som en del af litteraturen omkring beslutningsprocesser fokuserer på, hvordan både bias og varians kan mindskes, så fokuserer litteraturen indenfor ML i højere grad på, at der sker en

afvejning mellem bias og varians. Dette italesættes gennem det såkaldte 'bias-varians-trade-off', som er karakteriseret ved, at ML-modeller ikke både kan udmærke sig med lave niveauer af varians og bias på samme tid, hvorfor der derimod må foretages en afvejning, når det kommer til niveauet af disse (Bach et al. 2019: 177-178). Dette trade-off mellem bias og varians indenfor ML, kan visualiseres som i nedenstående figur:

Figur 4.4: Den klassiske visualisering af bias-varians tradeoff'et



Kilde: Med inspiration fra Guan & Burton (2020).

Med udgangspunkt i ovenstående trade-off betyder det dermed, at en model med en høj grad af varians vil kunne genkende alle unikke mønstre, men ikke nødvendigvis vil kunne 'forstå' data, som afviger fra træningsdataet, hvormed overførbarheden er nedsat. Modsat vil en underfittet model have en høj grad af bias og lav grad af varians, hvorved den i højere grad vil kunne overføre sammenhængene fra træningsdataet til andre datasæt. Udfordringen ved disse underfittede modeller er dog, at de ofte forekommer biased (Bach et al., 2019: 175-176). Af figuren bliver det tydeligt, at det aldrig vil være muligt at eliminere fejl i en ML-model. Af den grund anvendes modellens samlede fejlniveau derfor til at bestemme den optimale model, der opnås ved at acceptere et vist niveau af bias og varians.

Med en viden om, at det ikke er muligt at eliminere fejl i modeller baseret på algoritmer, kunne en overvejelse være, hvorvidt det overhovedet er relevant at anvende dem. Ifølge Jensen et al. (2018) og Magerman (2022) er fordelen ved algoritmiske modeller dog, at fejl foreligger nemmere at identificere og dermed håndtere sammenlignet med bias og støj i menneskelige vurderinger. Dette sker ud fra en forståelse af, at det vil være sværere at ændre på et fejlniveau, som er menneskeskabt, idet uvidenheden om både de systematiske (bias) og usystematiske fejl (støj) er større ved mennesker end modeller (jf. Kahneman et al., 2021: kapitel 5).

Med viden om de fordele og udfordringer, der er ved LL-modeller, kunne en overvejelse i den forbindelse være, hvorvidt en kombination af en menneskelig og computational bedømmelse kunne udgøre en effektiv og støjreducerende karaktergivningsproces i den danske folkeskole. Men hvordan vil censorerne i dansk reagere på en implementering af en LL-model i karaktergivningen? Og vil de have tillid til, at modellen vil kunne fastsætte en retvisende karakter på niveau med deres egen vurdering?

4.4 Tillid i samarbejdet med modellen

I forrige afsnit er det blevet belyst, hvordan LL-modeller kan betragtes som et bud på en mulig medbedømmer ved karaktergivning i skriftlig dansk i folkeskolen. Her gør det sig gældende, at en LL-model som medbedømmer vil have en række fordele, idet denne har potentiale til at kompensere for nogle af de begrænsninger, som optræder ved en karaktergivning udelukkende baseret på én menneskelig bedømmer. Foruden fordelene blev det også tydeliggjort, hvordan LL-modellerne vil kunne besidde en begrænsning, når det kommer til transparens og bias-varians-trade-off'et. Med forståelse herfor blev det dermed fremlagt, hvordan et samarbejde mellem et menneske og model kunne være fordelagtigt. Men hvordan bliver dette samarbejde vellykket?

Gennem litteraturen fremgår det, hvordan tillid skal betragtes som et vigtigt element i et samarbejde mellem mennesker. Dette skyldes, at tillid i hverdagens interaktioner er afgørende for, at samarbejder opleves som velfungerende, og at

opgaven løses på den mest hensigtsmæssige måde (Olesen & Hassle, 2011: 197-198). Dette gælder både mellem mennesker, men det kan også påpeges at spille en rolle, når det drejer sig om samarbejder, der strækker sig ud over menneskelige relationer (jf. Fan et al., 2008; Hoff & Bashir, 2014: 410; Ramchurn et al., 2016; Gao et al., 2021; Puranam, 2021). Dette kommer blandt andet til udtryk gennem følgende definition af Mayer et al. (1995), som er en af de mest anvendte definitioner af tillid (jf. Lee og See, 2004: 53):

"En parts villighed til at være sårbar over for en anden parts handlinger baseret på forventningen om, at den anden vil udføre en bestemt handling, der er vigtig for tillidsgiveren uanset evnen til at overvåge eller kontrollere den anden part"
(Mayer et al., 1995: 712; Glikson & Woolley, 2020: 9, oversat)

Af definitionen fremgår det, hvordan tillid i et samarbejde indebærer en villighed til at gøre sig sårbar mod en positiv forventning til modpartens bidrag af værdi. Her er det afgørende, at tilliden opstår uanset parternes evne til at kunne overvåge eller kontrollere hinanden. I praksis vil det betyde, at denne tillid opstår ved, at den menneskelige bedømmer i form af censoren stoler på, at LL-modellen formår at give en karakter på ensartet og pålidelig vis selvom, at censoreren ikke nødvendigvis vil have mulighed for at eftertjekke, hvordan LL-modellen kommer frem til fastsættelsen af den pågældende karakter.

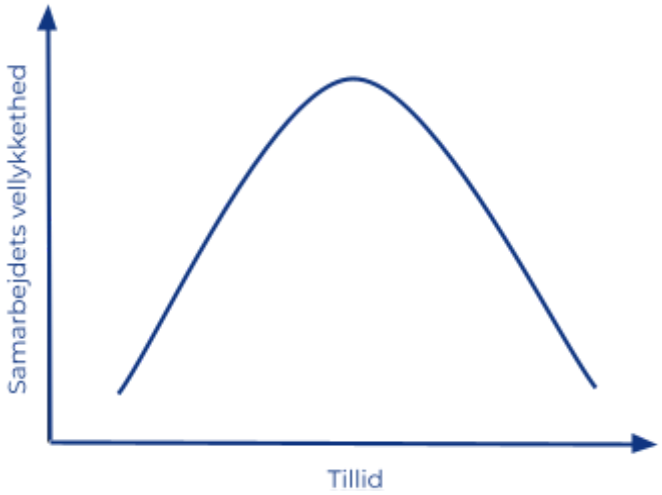
4.4.1 Tillidsniveauet har betydning samarbejde

At tillid også er afgørende i et samarbejde mellem mennesker og modeller understreges yderligere af Glikson & Woolley (2020: 8), som klarlægger, hvordan tillid påvirker anvendelsen af AI-modeller. Her gør det sig gældende, at nogle anvendelsesformer anses som mere hensigtsmæssige end andre, hvor dette i sidste ende har betydning for samarbejdets vellykkethed.

Gennem litteraturen fremstår det klart, hvordan en for høj tillid til modeller kan medføre, at individet, som benytter modellen, ikke forholder sig kritisk til modellens output og formentlig vil bruge modellens vurdering mere, end hvad der er hensigtsmæssigt. Dette kan italesættes som 'overbrug' af modellen (Parasuraman & Riley, 1997). Sådan et 'overbrug' påviste Dijkstra (1999) også i sin undersøgelse, hvor næsten 80 % af deltagerne rådgav forkert på baggrund af en

models vurdering, fordi de tilsidesatte deres egne overbevisninger og lod sig styre af et uforholdsmæssigt højt niveau af tillid til modellen. Modsat kortlægger litteraturen også, hvordan et lavt niveau af tillid til modellen kan være problematisk, hvis individet vælger at tilsidesætte modellen fuldkomment til trods for, at denne kunne have forbedret opgaveløsningen (Glikson & Woolley, 2020: 8-9; Ghazizadeh, Lee & Boyle, 2012: 40; Parasuraman & Riley, 1997: 230). Denne sammenhæng mellem tillid og samarbejdets vellykkethed, som tydeliggøres i litteraturen, kan visualiseres som en ikke-lineær sammenhæng.

Figur 4.5: Den u-formede samarbejdsfunktion



Kilde: inspireret af Parasuraman & Riley (1997)

Set i lyset af karaktergivningen i dansk skriftlig fremstilling i folkeskolen, kan der med udgangspunkt i ovenstående opstå en udfordring ved brugen af en LL-model som medbedømmer i karaktergivningen. Dette ses ud fra en forståelse af, at den fagprofessionelle bedømmers tillid til modellen er afgørende for, at modellen udnyttes hensigtsmæssigt og samarbejdet derved bliver succesfuldt. I den forbindelse må det antages, at hvis LL-modellen skal kunne bidrage til karaktergivningsprocessen, så vil det stadig være afgørende, at den fagprofessionelle bedømmer er i stand til at bibeholde sin faglighed og faglige skøn i bedømmelsen. Dette vil jævnfør ovenstående visualisering betyde, at den menneskelige bedømmer skal besidde det rette optimum af tillid til LL-modellen og kritisk tænkning, hvis der skal dannes grundlag for en støjreduktion i karaktergivningen.

Når drejer sig om menneskers tillid til AI-modeller, så fremlægger litteraturen flere bud på faktorer, der kan påvirke denne tillid. Hoff & Bashir (2015) finder i overensstemmelse med litteraturen om social tillid (jf. Gronemann & Jespersen, 2015: 68), hvordan tilliden til AI-modeller kan være afhængig af forskellige baggrundsfaktorer såsom alder, køn, kultur, viden om AI, dets muligheder og begrænsninger samt individets generelle tillidsniveau. Ydermere fremlægger Wærn og Ramberg (1996), hvordan en lavere selvtillid til egne kompetencer hos den fagprofessionelle kan medføre højere tillid til det beslutningsunderstøttende computationelle system. Endvidere har Dzindolet et al. (2003: 697) også vist, at en viden om, hvorfor automationen begår fejl, har en positiv effekt på tilliden hertil. Af litteraturen bliver det tydeligt, at tilliden til en AI-model i karaktergivningen kan være påvirket af flere forskellige faktorer. Men er denne tillid til mennesker og modeller helt ens?

4.4.2 Eksisterende forskning: tillid til mennesker sammenlignet med AI-teknologi

Madhaven & Wiegmann (2007) fremlægger gennem et litteraturreview, hvordan tillid opstår gennem den samme proces, uanset om det gælder tilliden til et andet menneske eller til en AI-model. Der er dog forskel på de faktorer, der bruges til at vurdere beslutningens validitet alt efter, om det er et menneske eller en AI. I den forbindelse vil der ved en AI-model være fokus på invarians, mens der ved et menneske vil være fokus på dets tilpasningsevne (Madhaven & Wiegmann, 2007: 289). Heraf kan det antages, at der i karaktergivningen vil være en sammenlignelig proces med at opbygge tillid uanset, om medbedømmeren er en fagprofessionel censor eller en AI-model. Men har mennesker mere tillid til mennesker end til modeller?

Gennem litteraturen foreligger der dog ikke konsensus om, hvorvidt mennesket har mest tillid til andre mennesker eller modeller. Dette kommer til udtryk ved, at Madhavan & Wiegmann (2007: 283-286) i deres undersøgelse finder en tendens til, at mennesker udviser mere tillid overfor automatiserede modeller, mens Lerch et al. (1997) i deres undersøgelse finder modsatrettede resultater i form af en større tillid til menneskelige eksperter sammenlignet med computationelle systemer.

Gennem et eksperimentelt studie undersøgte Jakesch et al. (2019), hvorvidt mennesker havde mere tillid til de informationer, som de troede var skrevet af en AI-model frem for mennesker. Resultaterne viste først og fremmest, at hvis modtageren havde mistanke om, at informationerne var skrevet af en AI-model, så havde de mindre tillid til informationen (Jakesch et al., 2019: 6). Foruden dette undersøgte Jakesch et al. (2019) videre, hvorvidt modtagerne var i stand til at differentiere de 'AI-skrevne' og 'menneske-skrevne' informationer. Resultaterne heraf var, at tilliden forekom større til de AI-genererede informationer sammenlignet med informationer skrevet af mennesker, når modtageren ikke vidste, hvilke informationer der var produceret af henholdsvis mennesket eller AI-modellen (Jakesch et al., 2019: 7). Ud fra Jakesch et al. (2019) tyder det dermed på, at der er en effekt på tilliden, som afhænger af, 1) hvorvidt informationen er produceret af et menneske eller en AI-model, og 2) hvorledes mennesket har en viden om, hvem der har produceret informationen. På baggrund heraf kan det udledes, hvordan det tyder på, at mennesker har størst tillid til mennesker, når de ved, at informationen kommer fra et menneske.

Ovenstående resultater belyser hermed, hvordan der er nogle tilfælde, hvor mennesket har mere tillid til AI-genererede løsninger, mens der er andre tilfælde, hvor tilliden er større til ekspertens bedømmelse. Dette kan herved være med til at indikere, hvordan tilliden til AI-teknologien kan være kontekstafhængig. Gennem afsnit 4.4.1 blev det tydeliggjort, hvordan tilliden har betydning for samarbejdet med en AI-model, hvormed dette videnshul om tilliden til en AI-model kan siges at være afgørende for, hvorvidt LL-modellen som medbedømmer kan bidrage med en forbedring af karaktergivningen.

4.5 Forskningsfeltets videnshul og forsknings-spørgsmål

Gennem ovenstående problemanalyse er det blevet tydeliggjort, hvordan karakterer i de seneste årtier har fået øget opmærksomhed og stigende betydning, idet karaktererne er blevet indført som sorteringsnøgle for optagelsen på ungdomsuddannelser. Dette foreligger at være problematisk, når

folkeskoleelevers afgangsopgaver ikke bedømmes ens, hvilket særligt gør sig gældende i dansk skriftlig fremstilling. Denne variabilitet beskriver Kahneman et al. (2021) som 'støj', hvilket anses som en udfordring, da dette er et udtryk for manglende ensartethed og pålidelighed i karaktergivningen. For individet kan dette forekomme problematisk, hvis elevernes muligheder videre i uddannelsessystemet begrænses uberettiget. Samtidig kan den manglende ensartethed og pålidelighed også være med til udfordre karaktergivningssystemets troværdighed. Slutteligt kan støjen også betragtes som problematisk ud fra et demokratisk princip om, at alle borgere bør have lige rettigheder, hvormed dette i værste tilfælde kan resultere i en manglende opbakning og tillid til det uddannelsessystem, som karaktergivningen er en del af.

Selvom det ikke er muligt at eliminere støjen fuldkommen i en beslutningsproces som karaktergivningen, hvor det faglige skøn udgør en væsentlig del, så bør der stadig være et ønske om at mindske denne støj mest muligt således, at de fagprofessionelles bedømmelser er mere overensstemmende. Med henblik på at reducere støjen i karaktergivningen i folkeskolen er det jævnfør Kahneman et al. (2021) en udfordring, at folkeskolen som den eneste uddannelsesinstitution i Danmark kun har én bedømmer ved folkeskolens afgangsprøve. I beslutningsprocesser, hvor der er to uafhængige bedømmere, vil støjen altid være mindre end, hvad der er tilfældet ved udelukkende én bedømmer. Denne praksis i folkeskolen med enebedømmerordningen giver derfor de bedste forudsætninger til at undersøge, hvorvidt en LL-model som medbedømmer vil kunne bidrage til at reducere støjen i karaktergivningen. Dette sker ud fra forståelsen af, at de nyere computationelle metoder har udviklet sig til at kunne håndtere og respondere på relativt komplekse problemer. Samtidig påpeger Kahneman et al. (2021), hvordan det kan være fordelagtigt for reduceringen af støjen, når algoritmiske modeller benyttes i beslutningsprocesser.

Trods potentialerne ved brugen af en LL-model, er det blevet tydeliggjort, at den mest retvisende og pålidelige karaktergivning forventes at opstå ved et samarbejde, hvor LL-modellen kan være med til at forbedre menneskets evne til at give mindre støjende karakterer. Dette skyldes, at mennesket og modellerne besidder forskellige styrker og svagheder, hvormed begge parters bedømmelse bør spille en aktiv rolle i karaktergivningen i form af et samarbejde.

For at relationen mellem LL-modellen og den fagprofessionelle kan betragtes som et samarbejde, er det gennem litteraturen tydeliggjort, at tilliden er en afgørende forudsætning. Hvorvidt den fagprofessionelle forventes at have mest tillid til mennesket eller AI-teknologien er der dog i litteraturen ikke enighed om. Grundet denne faglige diskrepans har nærværende projekt en interesse i at belyse, hvorvidt der er mere tillid til AI eller menneskelige vurderinger ved karaktergivningen i dansk skriftlig fremstilling, samt hvorvidt denne tillid har betydning for brugen af en LL-model som medbedømmer i karaktergivningen.

Med afsæt i antagelsen om, at to bedømmere kan reducere støjen i karaktergivningen i folkeskolen samt en viden om LL-modellernes styrker og svagheder opstår der herved en interesse i at afdække, hvorvidt en LL-model som medbedømmer kan bidrage til en støjreduktion ved karaktergivningen i dansk skriftlig fremstilling. Ydermere har diskrepansen om tilliden til AI-modeller i den eksisterende litteratur ledt videre til også at undersøge, hvorvidt en AI-model som medbedømmer både har betydning for tilliden og evnen til at bruge denne medbedømmer hensigtsmæssigt. Dette leder frem til følgende forskningsspørgsmål:

I hvilken grad kan en LL-model som medbedømmer være med til at reducere støjen i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen, og har tillid til modellen betydning for, om støjen kan reduceres?

5. Teori

I følgende kapitel vil studiets teoretiske grundlag blive fremlagt. Gennem studiets problemanalyse blev det tydeliggjort, hvordan karaktergivningen i folkeskolen er præget af en varians, som er med til at udfordre karakterernes ensartethed og pålidelighed. Denne varians blev refereret, som dét Daniel Kahneman, Olivier Sibony og Cass R. Sunstein (2021) omtaler som støj i deres seneste værk *"Støj - sådan træffer du bedre beslutninger"*. Heri fremlægger de, hvordan alle beslutninger er påvirket af støj i form af tilfældigheder, hvilket forekommer at være en udfordring for beslutningers nøjagtighed.

Med henblik på at kunne afdække, hvorvidt brugen af en LL-model som medbedømmer vil kunne reducere støjen i karaktergivningen, vil Kahneman et al.'s (2021) teori om støj blive fremlagt i den første del af studiets teoriafsnit. I den anden del af studiets teoriafsnit vil Parasuraman & Rileys (1997) teori om tillidens betydning for brugen af modeller blive fremlagt, da en forståelse herfor vil være afgørende for at kunne afdække, hvorvidt en LL-model som medbedømmer vil kunne bidrage til støjreduktion i karaktergivningen.

5.1 Støj

I det følgende afsnit vil Kahneman et al.'s (2021) teori vedrørende støj som fejlkilde i beslutningstagning blive udfoldet. Med teorien om støj kan det argumenteres, hvordan Kahneman et al. (2021) skriver sig ind i paradigmet omkring 'bounded rationality'. Dette paradigme bygger videre på antagelserne fra 'rational choice''-paradigmet, som er baseret på klassisk økonomisk teori. Her antages det, at alle individer i en beslutningsproces handler rationalitet samtidig med, at deres præferencer foreligger eksogene, stabile og konsistente. Disse antagelser videreudvikles efterfølgende i 'bounded rationality'-paradigmet, hvor det erkendes, at beslutninger aldrig foretages på baggrund af fuld information (Simon, 1947; Nannestad, 2009: 838-852). I stedet påpeges det inden for 'bounded rationality', hvordan menneskelig beslutningstagning er begrænset af en kognitiv kapacitet samt tilgængelige informationer. Med en viden herom vil de kommende teoretiske antagelser basere sig på en forståelse af, at mennesker sjældent har adgang til al den nødvendige information og/eller ressourcer til at foretage den rette beslutning, hvormed det kan argumenteres, at støj i karaktergivningen kan skyldes naturlige begrænsninger blandt censorerne.

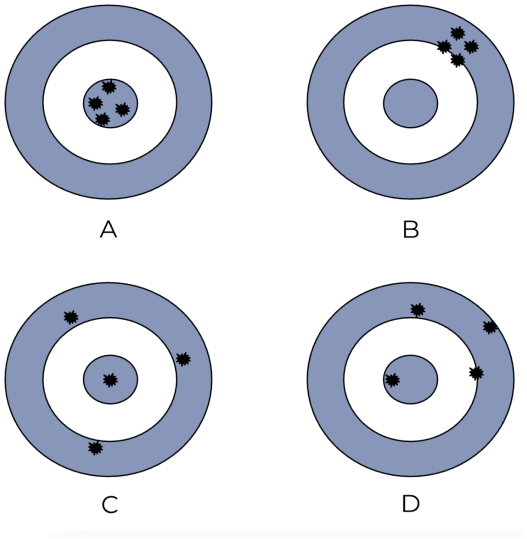
Når det kommer til, hvordan denne begrænsede rationalitet kan udfordre den menneskelige beslutningstagning, kan Daniel Kahneman og Amos Tversky (1974) påpeges at være nogle af foregangsmændene til at udvide forståelsen heraf. Kahneman & Tversky (1974) belyser, hvordan bias i form af systematiske vurderingsfejl kan udgøre en udfordring i menneskelig beslutningstagning. I det seneste værk af Kahneman et al. (2021) har forfatterne videre sat fokus på fejl i beslutningstagningen gennem det hidtil underbelyste fænomen omkring 'støj'. Dette sker ud fra en forståelse af, at støj såvel som bias kan resultere i vurderingsfejl og dermed udfordre kvaliteten af den menneskelige beslutningstagning (Kahneman et al., 2021: 12). Behovet for at fokusere på støjen frem for bias skyldes, at undersøgelser af støj og muligheden for støjreduktion ikke har været genstand for megen forskning sammenlignet med bias (jf. Kahneman et al., 2021: 12). Med en forståelse herfor vil dette studie undersøge mulighederne for støjreduktion i karaktergivningen i folkeskolen.

Det følgende afsnit vil først have til formål at fremlægge differentieringen mellem bias og støj baseret på Kahneman et al. (2021). I forlængelse heraf vil det blive tydeliggjort, hvordan der findes forskellige typer af støj, og hvordan denne støj kan måles. Afslutningvist vil der med afsæt i Kahneman et al. (201: 146-162) blive argumenteret for, at brugen af en LLM med fordel kunne anvendes som medbedømmer i karaktergivningen i folkeskolen som understøttende værktøj til enebedømmerordningen for at reducere støjen heri.

5.1.1 Støj og bias som delkomponenter af fejlniveauet i vurderinger

Gennem Kahneman et al. (2021) bliver det tydeliggjort, hvordan det at lave en vurdering kan sammenlignes med det at skyde mod et mål. I skydning er formålet at ramme pletskud på en skydeskive, som vist på skydeskive A nedenfor.

Figur 5.1: Skydeskive-metaforen



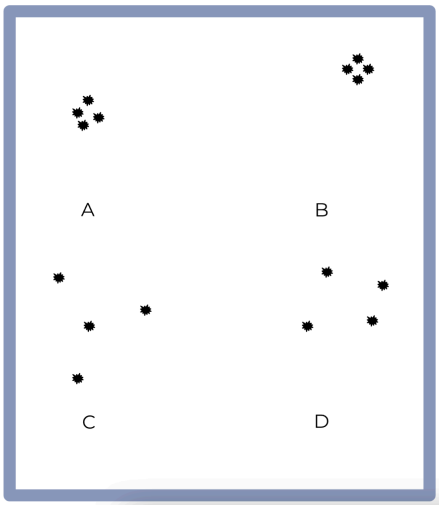
Kilde: Inspireret af Kahneman et al. (2021: 10).

På en skydebane vil skytterne forsøge at ramme midten ved hvert skud, hvor skuddenes præcision afhænger af forskellige faktorer. Her gør det sig gældende, at nogle af skytterne vil være bedre end andre til at ramme midten af skydeskiven - men selv den bedste skytte vil dog opleve variationer i sine skud (Kahneman et al., 2021: 97-113).

Ifølge Kahneman et al. (2021) kan denne skydeskive-metafor anvendes til at give et indblik i forskellene mellem bias og støj, som begge kan anses som fejl i vurderinger. Bias kan ifølge Kahneman et al. (2021: 10) illustreres ved en situation, hvor træfningerne forskydes systematisk til én side af målskiven, mens træfningerne stadig ligger relativt tæt på hinanden (jf. skydeskive B).

I vurderinger, hvor målet ikke er forhåndsbestemt, kan det være svært at identificere bias, fordi vurderingerne forekommer konsistente og afviger systematisk. Dette er visualiseret nedenfor. Her fremgår det, hvordan det kan være svært at afgøre, hvorvidt det er skydeskive A eller B, som er biased, når det korrekte mål ikke kendes på forhånd (se figur 5.2).

Figur 5.2: Skydeskive-metaforen uden skiverne



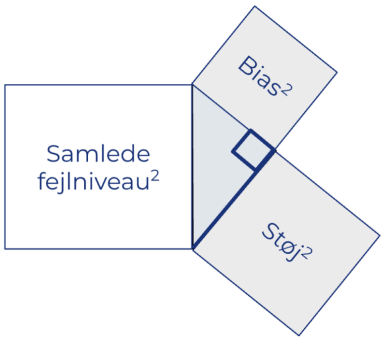
Kilde: Inspireret af Kahneman et al. (2021: 11).

Til forskel fra bias kan støj i vurderinger registreres som værende tilfældige og uregelmæssige fejl. Dette kan gennem skydeskive-metaforen illustreres ved, at træfningerne ikke er centreret omkring målet, men spredt tilfældigt over hele målskiven (jf. skydeskive C).

I praksis vil forskellen mellem bias og støj ikke foreligge så tydelig, idet vurderinger ofte vil være påvirket af begge typer fejlkilder (jf. skydeskive D). I sådanne tilfælde er fordelen ved støj ifølge Kahneman et al. (2021: 11), at den kan identificeres og måles uden kendskab til det korrekte mål eller sande værdier på forhånd. Derudover er fordelen ved støj og en reduktion heraf også, at omfanget af bias i vurderingerne fremstår tydeligere, når støjen bliver reduceret (ibid: 80-81).

Ovenstående figur med skydeskiverne er med til at indikere, hvordan både støj og bias udgør centrale andele i det samlede fejlniveau, som optræder, når der foretages vurderinger (Kahneman et al., 2021: 9-12). Til forståelsen af det samlede fejlniveau, kan geometrien i retvinklede trekanter være med til at forklare sammenhængen mellem bias, støj og det samlede fejlniveau.

Figur 5.3: Fejlniveauets to dele



Note: Den lyseblå trekant kan anvendes til at udlede sammenhængen mellem bias, støj og fejlniveau. Kilde: Inspireret af Kahneman et al. (2021: 78).

Af ovenstående visualisering fremgår det, hvordan det samlede fejlniveau i vurderinger både består af systematiske (bias) og tilfældige fejl (støj). Her gør det sig gældende, at den lyseblå trekant er med til at visualisere, hvordan den kvadrerede bias og kvadrerede støj svarer til det samlede kvadrerede fejlniveau i vurderingen.

5.1.2 Støj som uønsket variabilitet

I ovenstående afsnit er det blevet klarlagt, hvordan både bias og støj bidrager til det samlede fejlniveau i vurderinger, men at disse repræsenterer to forskellige typer af fejlkilder. Ud fra studiets interesse i at undersøge, hvordan det er muligt at reducere støjen ved karaktergivningen i folkeskolen, vil de følgende afsnit have til hensigt at skabe en dybere forståelse af støj som fejlkilde.

Når det drejer sig om støj i vurderinger, definerer Kahneman et al. (2021) dette således:

"(...) den uønskede variabilitet i vurderinger af et tilbagevendende problem (...)"
(Kahneman et al., 2021*: 47)*

Af citatet kan det udledes, hvordan støj forekommer, når mennesker gentagne gange skal foretage en vurdering, hvor disse vurderinger udmærker sig med en uønsket variabilitet. Denne uønskede variabilitet kan derfor være med til at skabe en usikkerhed omkring beslutninger, da disse i tilfælde af støj kan opfattes som upålidelige. I karaktergivning kan støj i form af en uønsket variabilitet komme til udtryk ved, at en lignende præstation bedømmes med forskellige karakterer, hvorved denne manglende ensartethed er med til at udfordre karaktergivningens kvalitet.

Kahneman et al. (2021: 51) påpeger, hvordan menneskelige vurderinger altid vil være påvirket af variabilitet, hvorfor en vis grad af støj vil være acceptabelt. Udfordringen er dog, at variabilitet og dermed støj i menneskelige vurderinger typisk er større end, hvad der er ønskværdigt (ibid: 43). Udover at der er mere støj end, hvad der er ønskværdigt, så er der typisk også mere støj til stede end, hvad mennesker selv tror og er bevidste om.

Ifølge Kahneman et al. (2021:83) skal systemstøj forstås som en fællesbetegnelse for al den støj, der finder sted i et system såsom karaktergivningen i folkeskolen. Med henblik på at bekæmpe den uønskede variabilitet i beslutninger kan det tydeliggøres, at Kahneman et al. (2021: 83) differentierer mellem forskellige typer af støj herunder niveau- og mønsterstøj. Disse typer af støj vil blive uddybet i følgende afsnit.

5.1.3 Niveaustøj og mønsterstøj

Gennem Kahneman et al. (2021) kan det klarlægges, hvordan der foreligger to overordnede måder, hvorpå støj kan opstå ved fagprofessionelle vurderinger. Den ene måde er, når en bedømmer har tendens til at være hårdere i sine vurderinger end andre. Resultatet af dette omtaler Kahneman et al. (2021) som 'niveaustøj', idet der foreligger en generel uoverensstemmelse omkring niveauet i bedømmelserne. Inden for karaktergivning kan niveaustøjen ses, når nogle censorer har en tendens til at give højere karakterer end andre.

Hermed kan det udledes, hvordan niveaustøj kan defineres som variabiliteten i bedømmernes gennemsnitlige vurderinger (Kahneman et al., 2021: 90-91). Denne variabilitet i bedømmernes gennemsnitlige vurderinger er visualiseret i nedenstående figur med udgangspunkt i karaktergivningen (jf. figur 5.4). Her fremgår det, hvordan censor A's gennemsnitlige vurdering er lavere end censor B's, hvormed der kan siges at forekomme niveaustøj.

Figur 5.4: Niveaustøj

Kilde: Inspireret af Kahneman et al. (2021: 87).

Foruden niveaustøj består systemstøjen som tidligere nævnt også af mønsterstøj (Kahneman et al., 2021: 238). Ifølge Kahneman et al. (2021) er mønsterstøj karakteriseret ved, at variabiliteten mellem de forskellige bedømmere er opstået som følge af komplekse mønstre. Disse komplekse mønstre kan opstå på grund af utallige faktorer, hvor nogle af de mest nævneværdige eksempelvis er bedømmernes forskellige faglige præferencer, filosofier og erfaringer (Kahneman et al., 2021: 91-93). Denne form for støj betegner Kahneman et al. (2021) som mønsterstøj. Et eksempel på mønsterstøj i karaktergivningen kunne være, at én censor vægter grammatik højt i sin vurdering, mens en anden i højere grad vægter argumentationskæden. Disse forskellige mønstre vil resultere i, at det vil være forskelligt, hvornår de to censorer giver høje og lave karakterer.

I afsnit 5.1.4 vil det blive tydeliggjort, hvordan dette faktisk er et eksempel på dét, som kaldes stabil mønsterstøj. For mønsterstøjen gør det sig gældende, at den ikke på samme vis kan måles på baggrund af gennemsnittet for en bedømmers vurderinger. Dette er visualiseret på nedenstående figur (jf. figur 5.5), hvor det bliver tydeligt, hvordan mønsterstøjen kun kan uddeles på tværs af forskellige vurderinger ved forskellige censorer.

Figur 5.5: Mønsterstøj

	Eksamensopgave 1	Eksamensopgave 2	Mean
Censor A	7	10	8,5
Censor B	10	12	11
Censor C	12	10	11

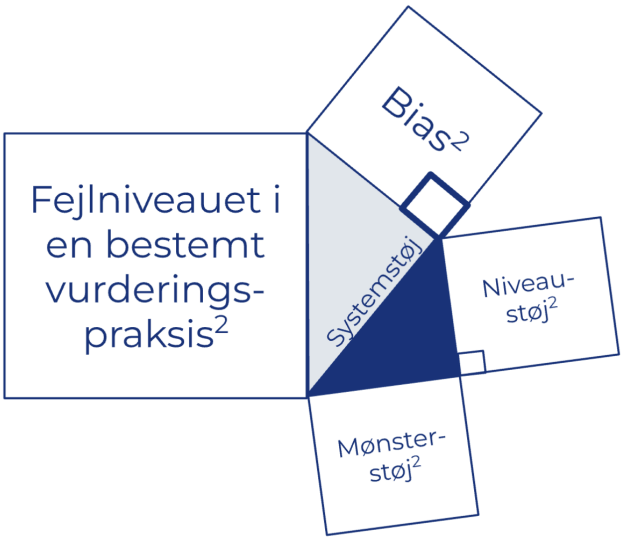
Mønsterstøj

Kilde: Inspireret af Kahneman et al. (2021: 87).

Af figuren kan det ses, hvordan censor B har givet karakteren 10 til eksamensopgave 1 og 12 til eksamensopgave 2, mens censor C har bedømt eksamensopgave 1 til karakteren 12 og eksamensopgave 2 til karakteren 10. Det tyder hermed på, at censor B og C ikke har vægtet opgavens kriterier ens, hvormed der kan argumenteres at forekomme mønsterstøj.

Med afsæt i ovenstående i blev det fremlagt, hvordan systemstøjen ifølge Kahneman et al. (2021) består af to delkomponenter herunder niveaustøj og mønsterstøj, hvilket giver anledning til at udvide modellen for det samlede fejlniveau:

Figur 5.6: Støj opdeles i niveaustøj og mønsterstøj



Kilde: Inspireret af Kahneman et al. (2021: 94).

Af figuren ses det, hvordan arealet, som tidligere udgjorde en firkant med den samlede kvadrerede systemstøj (jf. figur 5.5) nu er blevet erstattet af to firkanter.

Hermed fremgår det af figuren, hvordan systemstøj består af henholdsvis den kvadrerede niveaustøj og mønsterstøj.

Når det kommer til mønsterstøjen påpeger Kahneman et al. (2021: 238-245), hvordan denne kan nuanceres yderligere i to delkomponenter: 'stabil mønsterstøj' og 'situationsstøj'. Disse vil blive udfoldet i det følgende afsnit.

5.1.4 Stabil mønsterstøj og situationsstøj

Ifølge Kahneman et al. (2021: 238-245) opstår den stabile mønsterstøj i kraft af, at forskellige fagprofessionelle beslutningstagere typisk vil have hver sit faglige fokus, som ligger til grund for beslutningen. Den stabile mønsterstøj kan i høj grad sættes lig med den mønsterstøj, som blev eksemplificeret tidligere (jf. afsnit 5.1.3), hvor to forskellige censorer kunne have divergerende faglige præferencer, som forårsagede, at de vægtede forskellige elementer i deres vurdering.

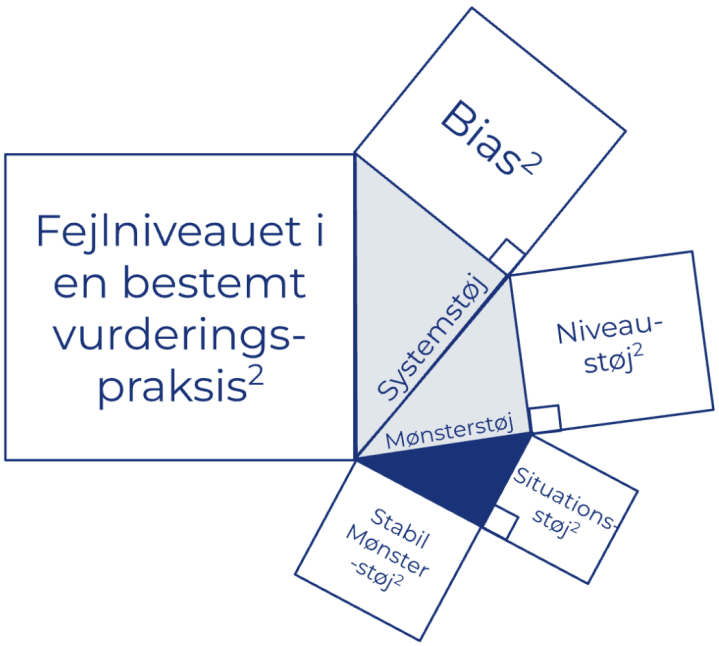
Dermed kan det tydeliggøres, hvordan fagprofessionelles vurderinger er påvirket af de samme mønstre over tid. Dette kan skyldes forskelle i bedømmernes viden og erfaringer (ibid:241). I den forbindelse påpeger Kahneman et al. (2021: 388, 404-407, 409-411), hvordan det i henhold til den stabile mønsterstøj er nødvendigt at diskutere i hvor høj grad, der skal være plads til variabilitet, når det drejer sig om fagprofessionelle og deres skøn. Dette sker ud fra forståelsen af, at det netop er præferencerne i dette faglige skøn, som kan resultere i stabil mønsterstøj.

Til forskel fra den stabile mønsterstøj, kan det tydeliggøres, hvordan graden af situationsstøj i vurderinger varierer over tid. Dette skyldes, at situationsstøjen er karakteriseret ved at opstå som følge af midlertidige eller uforudsigelige faktorer, der påvirker vurderingen i en given situation. Dette kunne omfatte eksterne faktorer såsom distraktioner, forstyrrelser eller uventede begivenheder, som ses at påvirke vurderingen og dermed blive udslagsgivende for den endelige beslutning. Foruden eksterne faktorer påpeger Kahneman et al. (2021: 104-109) også, hvordan situationsstøj kan opstå som følge af interne faktorer hos bedømmerne. Her kan faktorer såsom selektiv opmærksomhed, humør eller træthed påvirke bedømmerens evne til at foretage en vurdering (Kahneman et al., 2021: 104-109). Med udgangspunkt i karaktergivningen i folkeskolen tyder det

dermed på, hvordan både interne og eksterne faktorer vil kunne påvirke den endelige karakter. Eksempelvis kunne elevens endelige karakter være påvirket af karakteren for eksamensopgaven, som blev bedømt forinden? Gennem Kahneman et al. (2021) kan det udledes, hvordan en høj grad af situationsstøj foreligger problematisk for karaktergivningens kvalitet, da dette er med til at udfordre den saglige begrundelse herfor.

Med udgangspunkt i ovenstående kan det igen argumenteres, hvordan forståelsen for det samlede fejlniveau er blevet udvidet. Dette giver dermed anledning til at udvide modellen en sidste gang, hvilket fremgår nedenfor.

Figur 5.7: Mønsterstøj opdeles i stabil mønsterstøj og situationsstøj



Kilde: Inspireret af Kahneman et al. (2021: 239 og 247).

Af figuren for det samlede fejlniveau fremgår det, hvordan forståelsen for de forskellige typer af fejlkilder er blevet styrket yderligere. Her er firkanten med mønsterstøj blevet erstattet af to nye firkanter (jf. figur 5.6). Disse nye firkanter repræsenterer henholdsvis den stabile mønsterstøj og situationsstøjen i deres kvadrerede form.

5.1.5 Måling af støj

Efter at have skabt en forståelse for systemstøjen og dens forskellige komponenter, vil følgende afsnit have til formål at klarlægge, hvordan denne støj i systemerne kan måles. I den forbindelse bliver det gennem Kahneman et al. (2021) klarlagt, hvordan støjen bør identificeres og måles ud fra vurderingens numeriske afvigelse fra gennemsnittet. Derudover påpeger Kahneman et al. (2021: 76) også, hvordan det forekommer afgørende, at værdien af den numeriske afvigelse kvadreres. Dette skyldes, at det ifølge Kahneman et al. (2021: 76) ikke bør være hensigten at lave et mål for støj, som statistisk set 'straffer' de små afvigelser fordi menneskelige beslutninger altid vil indebære en smule variabilitet, som vil resultere i små afvigelser (jf. afsnit 5.1.2). Af den grund finder Kahneman et al. (2021: 76) en fordel ved at kvadrere afvigelserne, idet de særligt støjende vurderinger dermed vil blive 'straffet' hårdere og de støjende vurderinger vægtes mere i målingen af netop støjen.

Kahneman et al.'s (2021: 77) er med ovennævnte tilgang i forhold til de kvadrerede afvigelser inspireret af Gauss' lov om de mindste kvadraters metode (mean squared error), hvor *y* er den målte vurdering, og *p* er gennemsnittet for denne vurdering:

$$MSE = \frac{\Sigma(y_i - p_i)^2}{n}$$

Med ovenstående ligning er det hermed muligt at beregne støjniveauet, som i henhold til karaktergivningen kunne se således ud:

$$Støj = \frac{(12-10)^2+(4-7)^2}{2} = 6,5$$

Af ligningen fremgår det, at hvis elev A har fået karakteren 12 (y_1) ved en eksamensopgave, hvor gennemsnittet af censorer ville give denne opgave karakteren 10 (p_1), mens elev B har fået karakteren 4 (y_2), hvor gennemsnittet af censorer ville have givet karakteren 7 (p_2), så er det samlede støjniveau ved disse to vurderinger 6,5.

5.1.6 Støj udlignes ikke

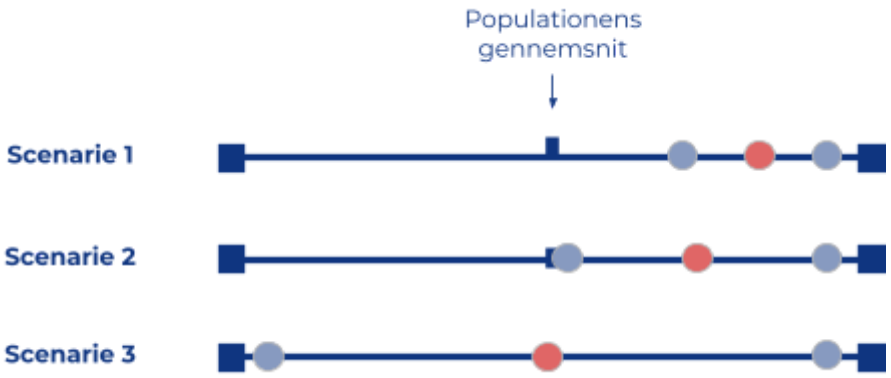
Tidligere blev det fremlagt, hvordan støj hidtil har været en underbelyst fejlkilde i vurderinger sammenlignet med bias, hvorfor støjen ikke har fået meget organisatorisk eller videnskabelig opmærksomhed (jf. afsnit 5.1). En af årsagerne hertil er ifølge Kahneman et al. (37-39, 69) en fejlagtig forestilling om, at vurderingernes afvigelser udligner hinanden. Dette modargumenterer Kahneman et al. (2021: 37-39, 69), da støj aldrig må betragtes som at kunne udlignes men derimod ophobes til ulempe for både individet og samfundet. Dette skyldes, at hver eneste vurdering har en værdi for modtageren af denne. Ved en karaktergivning, hvor en elev på tilfældig vis får en lavere karakter af en censor end, hvad vedkommende ville have fået af andre censorer, så bliver denne fejl ikke mindre af, at en anden elev får en højere karakter. Tværtimod er det blot flere fejl, som ophober sig i systemet.

5.1.7 To bedømmere støjer mindre

Ovenfor er det blevet belyst, hvordan det gennem Gauss' metode er muligt at måle systemstøjen i karaktergivningen. Med afsæt heri vil følgende afsnit have til formål at kortlægge, hvordan det ifølge Kahneman et al. (2021) kan være muligt at reducere støj med henblik på at styrke karaktergivningen i folkeskolen.

I studiets problemanalyse blev det fremlagt, hvordan bedømmelsespraksissen ved folkeskolens afgangsprøver i skriftlig fremstilling i dag baserer sig på en enebedømmerordning (jf. afsnit 4.2.2). At der kun forekommer én bedømmer i karaktergivningen kan gennem Kahneman et al. (2021) siges at være en vigtig faktor, som kan lede til for megen støj. Denne støj vil ifølge Kahneman et al. (2021: kap 7) kunne reduceres, hvis der i stedet indgik to bedømmere i karaktergivningen. Dette er baseret på antagelsen om, at gennemsnittet af to uafhængige bedømmelser altid vil kunne reducere mængden af støj, idet gennemsnittet mellem to bedømmeres vurderinger aldrig kan blive mere støjende end den mest støjende vurdering (ibid: 101-102). Ud fra formålet om at illustrere, hvordan gennemsnittet af to bedømmere altid vil støje mindre end den mest støjende bedømmer, er der i figur 5.8 fremlagt tre forskellige scenarier. Scenarierne repræsenterer tre vurderinger med to uafhængige bedømmere (blå prikker) og deres gennemsnitlige bedømmelse (rød prik).

Figur 5.8: Gennemsnittet for to uafhængige bedømmelser



Af figuren fremgår det, hvordan gennemsnittet af de to bedømmeres vurderinger i alle tre scenarier befinder sig tættere på populationens gennemsnittet end, hvad den mest støjende bedømmer gør. Hermed fremstår det klart, hvordan støjen i en vurdering vil kunne reduceres med to uafhængige bedømmelser. I vurderingspraksisser, hvor der ikke forekommer en sand værdi, kan det ifølge Kahneman et al. (2021) argumenteres, hvordan populationens gennemsnit udgør den bedste målestok til at vurdere de uafhængige bedømmelsers støj. Dette bygger på fænomenet omkring 'the wisdom of crowds', som klarlægger, hvordan de mest korrekte svar ofte findes ved at tage gennemsnittet af en større gruppes forskellige vurderinger (jf. Galton, 1907; Kahneman et al., 2021: 101-104). Med afsæt heri kan det dermed ifølge Kahneman et al. (2021) antages, at des mere vurderingen afviger fra populationens gennemsnit, desto mere støjende er vurderingen.

At ændre bedømmelsespraksissen fra én til to bedømmere ved karaktergivningen vil dog som minimum fordoble ressourceforbruget ved den enkelte vurdering. Som nævnt i problemanalysen blev denne ekstra bedømmelse sparet væk, hvorfor der nu kun er én bedømmer ved karaktergivningen til folkeskolens afgangseksamen (jf. afsnit 4.2.2). Det vil derfor være relativt usandsynligt, at en ordning med to bedømmere vil blive den gældende praksis igen foreløbigt. Ifølge Kahneman et al. (2021) findes der dog alternative metoder til at opnå samme støjreducerende effekt som ved en tobedømmerordning. Her fremhæves det, hvordan algoritmer og machine learning kunne være en anden

måde, hvorpå det kunne være muligt at reducere støjen i vurderinger. De potentielle muligheder herfor vil blive fremlagt i følgende afsnit.

5.1.8 Hvordan kan algoritmer bidrage til støjreduktion i karaktergivningen?

Med henblik på at forstå algoritmer benytter Kahneman et al. (2021) en bred definition i form af: "*en proces eller en række af regler, der opstår i forbindelse med problemløsende beregninger*" (Kahneman et al., 2021: 147). Algoritmer er hermed defineret som værende alle regler eller processer for beregninger, der skal bidrage med at løse et problem. Disse består herved af både vægte og funktioner, hvorved det er forhåndsbestemt, hvordan forskellige delelementer i vurderingen skal vægtes forskelligt (jf. Géron, 2022, s. 312-313, 310; Mbali, 2023; Kalirane, 2022)

Hvad angår brugen af algoritmer i vurderinger, har Kahnemann et. al. (2021) peget på en række fordele ved at anvende algoritmiske værktøjer i beslutningsprocesser. I den forbindelse argumenteres der for, at algoritmer bygger vurderinger på komplekse mønstre, som mennesket ikke på samme vis evner at konstruere. At algoritmer er bygget på fastsatte funktioner og vægte kan i sig selv være et argument for at legitimere anvendelsen af dem i beslutningsprocesser, fordi vægte og funktioner er bestemt ud fra en tilgang om at minimere afvigelsernes kvadrat.

Det andet argument for anvendelsen af algoritmer er ifølge Kahneman et al. (2021), at algoritmer ikke er støjende. Dette skyldes, ifølge Kahneman et al. (2021) at algoritmernes forhåndsbestemte regler styrkes ved, at alle vurderinger foretages på baggrund af akkurat samme funktioner og vægte. Dette betyder dermed, at hvis en algoritme skulle lave samme vurdering i dag og i morgen, så ville der ifølge Kahneman et al. (2021: 148) ikke forekomme nogen variabilitet i vurderingen. Af den grund kan algoritmerne ifølge Kahneman et al. (2021: 149, 159) betragtes som et sæt af regler, der ikke afviges fra, og som er fastsat med formål om at lave den mest pålidelige vurdering. Det bør dog bemærkes, at der gennem forskningslitteraturen eksempelvis findes evidens for, at algoritmer også kan forekomme støjende, idet to forskellige modeller kan give to forskellige

vurderinger (Gunes & Florczak, 2023). Dette betyder med andre ord, at Kahneman et al. (2021) kan have ret i, at modeller i højere grad end menneskelige vurderinger kan være konsistente. Der forekommer dog ikke megen viden om, hvorvidt det i henhold til karaktergivning kan siges, at forskellige AI-modeller støjer mere eller mindre end censorerne i skriftlig dansk.

I afsnit 5.1 blev det fremlagt, hvordan Kahneman et al. (2021) differentierer mellem forskellige typer af støj. Herigennem blev det blandt andet tydeliggjort, hvordan særlig situationsstøj kan betragtes som uhensigtsmæssig i fagprofessionelle vurderinger. Når det kommer til algoritmer påpeger Kahneman et al. (2021: 159) det som fordelagtigt, at algoritmerne vil kunne eliminere situationsstøjen. Foruden at algoritmiske vurderinger ikke vil blive påvirket af situationsstøj såsom vejret, humøret eller træthed, så vil den algoritmiske vurdering heller ikke blive foretaget på baggrund af tilfældige fokuspunkter. Hvor en fagprofessionel vurdering altid vil være præget af en selektiv hukommelse, så vil en algoritmisk vurdering altid anvende dets forudbestemte regler på tværs af al den tilgængelige data. Dette betyder dermed, at algoritmen i sin vurdering i højere grad vil kunne eliminere flere elementer af situationsstøjen, hvilket anses som afgørende for den ensartede bedømmelse (Kahneman et al., 2021: 58, 159).

Med henblik på at foretage mere komplekse vurderinger anbefaler Kahneman et al. (2021: 153-156) at anvende nyere algoritmer såsom ML-modeller, som besidder en større kompleksitet. Til forskel for simple algoritmer kan ML-modellerne håndtere mere komplekse statistisk funderede mønstre (Kahneman et al., 2021: 153). At algoritmer og særligt ML-modeller er i stand til at foretage vurderinger på baggrund af komplekse mønstre samtidig med, at de formår at være konsistente i brugen af disse, er ifølge Kahneman et al. (2021: 152, 156, 161) årsagen til, at algoritmer kan være i stand til at foretage mere pålidelige vurderinger end menneskelige bedømmere. Disse ovennævnte styrker ved brug af ML-modeller gør ifølge Kahneman et al. (2021: 156-159), at modellerne ved støjreduktionen bliver de menneskelige bedømmere overlegne. Men hvorfor anvende den menneskelige bedømmer i karaktergivningen, hvis den algoritmiske bedømmelse kan siges at være mere præcis og mindre støjende?

Til trods for algoritmernes evne til at være støjreducerende, fremhæver Kahneman et al. (2021), hvordan der bør forekomme en opmærksomhed omkring, at anvendelsen af algoritmer i beslutningsprocesser ikke er uproblematisk. Selvom der i ovenstående afsnit er argumenteret for, at algoritmer overgår de menneskelige evner til at foretage vurderinger, så erkender Kahneman et al. (2021) samtidig, at algoritmerne heller ikke er fejlfrie, når det kommer til at foretage vurderinger. Af den grund er det ifølge Kahneman et al. (2021: 161) afgørende, at en fagprofessionel indgår i vurderingen og står med ansvaret for disse fejl. Dette sker ifølge Kahneman et al. (2021: 158-161) ud fra en forståelse af, at individer har større tendens til at have mistro til institutionelle systemer, når en algoritme begår fejl sammenlignet med et menneske. I relation til karaktergivning vil det dermed kunne betyde, at eleven og det resterende samfund vil opleve en større mistillid til karaktersystemet, hvis fejlene i karaktergivningen var forårsaget af en algoritme frem for en censor. Hermed kan det tydeliggøres, hvordan det ved brugen af algoritmer i en beslutningsproces kan være nødvendigt, at der forekommer en menneskelig bedømmer, som har mulighed for at træffe en anden beslutning, hvis modellen foretager en fejlagtig vurdering.

5.1.9 Studiets hypoteser om støj

Vurderinger foretaget af algoritmer kan ifølge Kahneman et al. (2021) være præget af uhensigtsmæssige fejl, der gør det nødvendigt, at fagprofessionelle bedømmere afgiver den endelige vurdering. Ifølge Kahneman et al. (2021) bør det dog anerkendes, at de menneskelige fejl i vurderingsprocesser kan være problematiske, hvorfor det også betragtes som nødvendigt, at de fagprofessionelle gør brug af algoritmens vurderinger. Med disse antagelser åbner Kahneman et al. (2021) op for brugen af algoritmer som redskaber til at forbedre beslutningsprocesser. Med afsæt heri bliver det dermed muligt at betragte algoritmer som en mulig medbedømmer i karaktergivningen. Selvom Kahneman et al. (2021) argumenterer for, at algoritmer skal anvendes i samspil med de fagprofessionelle vurderinger, så er der i studiets problemanalyse ikke fremlagt nogen evidens for, at algoritmerne vil bidrage til mindre støj i vurderinger, hvis den fagprofessionelle selv bestemmer, hvorvidt vedkommende vil anvende algoritmen som værktøj i karaktergivningen.

Der er gennem Kahneman et al. (2021) opsat en implikation om, at to bedømmere altid vil bidrage til mindre støj. Men hvilken betydning har det for støjreduktionen, når censorerne selv har indflydelse på, hvor meget de anvender modellen? Denne undren leder frem til studiets første hypotese:

H1: En LL-model som medbedømmer kan reducere støjniveauet i karaktergivningen i dansk skriftlig fremstilling

Ifølge Kahneman et al. (2021) er algoritmer altid mindre støjende end den menneskelige bedømmer. Men betyder dette, at LL-modellen afviger mindre fra den 'sande karakter' end censorerne? Gennem Galton (1907) og Kahneman et al. (2021) er det klargjort, hvordan et mål for den 'sande karakter' kan findes gennem en gruppes gennemsnit (jf. afsnit 5.1.7). For selvom AI-modeller formentlig er mere konsistente i vurderinger samtidig med, at de gennem H1 også kan vise sig at bidrage med mindre systemstøj i karaktergivningen, er de ikke nødvendigvis selv mindre støjende. Hvorvidt selve LL-modellen i dette studie er mindre støjende end den gennemsnitlige dansklærer vil blive undersøgt gennem hypotese 2:

H2: LL-modellen kan være mindre støjende end den gennemsnitlige dansk censor

5.2 Tillid

Efter at have skabt en forståelse for den første del af studiets teoretiske apparat omhandlede støj, vil følgende afsnit have til formål at klarlægge den anden del af det teoretiske apparat, som vil omhandle tillid og dets betydning for menneskets brug af en AI-model. De teoretiske antagelser om tillid i dette afsnit, vil primært basere sig på Parasuraman & Rileys (1997) artikel *"Humans and Automation: Use, Misuse, Disuse, Abuse"*. Denne sætter fokus på, hvordan menneskets tillid til en automation har betydning for brugen heraf, hvor Parasuraman & Riley (1997: 230) differentierer mellem forskellige former brug.

5.2.1 Bounded rationality nuanceret med paradigmet om social kapital

I det tidligere afsnit blev det fremlagt, hvordan Kahneman et al. (2021) med teorien omkring støj i beslutningstagning skrev sig ind i et teoretisk paradigme omkring begrænset rationalitet. Med en teori om tillid til automation kan det argumenteres, hvordan Parasuraman & Riley (1997) til gengæld skriver sig ind i et større paradigme, der kan siges at være i tråd med paradigmet om social kapital (jf. Bourdieu, 1986; Coleman, 1990; Putnam, 2000). Gennem paradigmet belyses det, hvordan tillid kan betragtes som social kapital. Her udgør tilliden en ressource, som opstår på baggrund af sociale relationer og netværk (jf. Putnam, 2000). På baggrund heraf argumenteres det inden for dette paradigme, at individer kan trække på deres sociale kapital herunder tillid, når de skal foretage en beslutning.

At studiets to teorier befinder sig indenfor to forskellige teoretiske paradigmer, gør det nødvendigt at afsøge, hvorvidt disse paradigmer er forenelige. I stedet for at betragte disse to paradigmer som modstridende kan det argumenteres, hvordan social kapital i form af tillid kan være med til at udbygge forståelsen af, hvordan begrænset rationalitet kan opstå eller kompenseres. Dette sker ud fra en forståelse af, at det i studiets problemanalyse blev belyst, hvordan både en uforholdsmæssig lav og høj tillid til en automatiseret model kan medføre uhensigtsmæssige anvendelse af AI-modeller (jf. figur 4.4, afsnit 4.4.1). Herved kan det argumenteres, at et passende niveau af tillid kan fremme rationaliteten i karaktergivningen, idet censoreren i samarbejde med en AI-model teoretisk set kan give en mindre støjende karakter, hvis der udvises et passende niveau af tillid til en AI-model.

Med en forståelse for, at studiets teoretiske paradigmer er forenelige, vil Parasuraman & Rileys (1997) teori om tillidens betydning for brugen af automatiserede modeller blive udfoldet i det følgende. At studiet vælger at basere sine teoretiske antagelser omkring tillid i menneskers samarbejde med AI-modeller på Parasuraman & Riley (1997) skyldes, at de formår at præcisere deres teoretiske termer omkring 'tillid' og 'brugen af automation'. Foruden deres

præcise konceptualisering, kan det også fremhæves, hvordan deres teoretiske termer om 'tillid' og 'brugen af automation' er differentierbare. Ydermere kan det også fremhæves som fordelagtigt, at deres teoretiske antagelser både foreligger genkendelige og resonante, da dette er med til at styrke forståelsen og anvendeligheden af begreberne både inden for en akademisk- og offentlig kontekst (jf. Gerring, 1999). Med afsæt heri kan det tydeliggøres, hvordan Parasuraman & Riley (1997) vil kunne bidrage til en nuancering af censorernes villighed til at bruge en LL-model ved karaktergivningen i folkeskolen, og hvorvidt LL-modeller kan anvendes til støjreduktion ved karaktergivningen i skriftlig dansk fremstilling.

5.2.2 Teoriens relevans i en nutidig kontekst

Inden Parasuraman & Rileys (1997) teori om tillid til automation fremlægges, vil der indledningsvist forekomme en klarlægning af, hvordan Parasuraman & Riley (1997) med deres reference til automation kan anvendes i dette studie. Af studiets 'empiriske fund' bliver det tydeligt (jf. afsnit 4.4) , hvordan Parasuraman & Rileys (1997) studie qua deres fokus på tilliden til og brugen af automation er en del af en større litterær bølge, der har undersøgt forholdet mellem mennesker og automation. I den sammenhæng kan Parasuraman & Rileys (1997) definition af automation være med til at afgøre, hvorvidt det er muligt at overføre de teoretiske antagelser omkring automation og tillid til indeværende studie, som tager afsæt i LL-modeller. Definitionen lyder således:

> *"Vi definerer automatisering som udførelsen af en maskinagent (normalt computere) af en funktion, der tidligere blev udført af et menneske"*
> (Parasuraman & Riley, 1997: 231, oversat)

Af definitionen fremgår det, hvordan Parasuraman & Riley (1997: 231) definerer automation som en maskine, der udfører en funktion, som tidligere er blevet udført af et menneske. Ydermere belyser Parasuraman & Riley (1997: 231) også, hvordan betragtningen af automation vil ændre sig med tiden i takt med, at funktionen af disse automatiserede modeller forventes at udvikle sig.

På udgivelsestidspunktet for Parasuraman & Rileys (1997) artikel foreligger der ganske få eksempler på, at maskiner kunne overtage opgaver med store kognitive krav såsom komplekse beslutningstagningsprocesser. Parasuraman & Riley påpeger dog allerede dengang, hvordan neurale netværker (NN) kunne være en måde, som kunne muliggøre en overførsel af opgaver, der kræver store kognitive forudsætninger fra menneske til maskine i fremtiden (Parasuraman & Riley, 1997: 232). Med afsæt heri kan det udledes, at ML, DL og LL kan betragtes som en undergren af automationen, idet disse udgør en videreudvikling af NN. Af den grund er dette med til at indikere, hvordan Parasuraman & Rileys (1997) teoretiske definitioner og antagelser stadig foreligger anvendelige og overførbare til indeværende studie, der fokuserer på brugen af LL-modeller i karaktergivningen.

5.2.3 Synergi i samarbejdet mellem automationen og mennesket

Gennem Parasuraman & Riley (1997) bliver det tydeliggjort, hvordan automatisering tidligere omfattede en udvikling, hvorved automationen overtog processer, som ikke længere nødvendigvis krævede menneskelig indgriben (Parasuraman & Riley, 1997: 231). I kraft af at det i dag i højere grad handler om, at automatiseringen inddrages i processer, hvor den menneskelige deltagelse stadig er nødvendig, har tilgangen til automationen ændret sig. Dette skyldes ifølge Parasuraman & Riley (1997: 232), at automatisering og udførelsen heraf ikke altid foreligger fejlfri, hvorfor de påpeger, hvordan en menneskelig vurdering af den automatiserede models output er nødvendig, inden modellen benyttes. Dette behov for den menneskelige interaktion med automationen, sker ifølge Parasuraman & Riley ud fra en forståelse af, at mennesker og automatiserede modeller hver især har deres styrker og svagheder. Dette eksemplificerer Parasuraman & Riley (1997) gennem følgende citat, hvor de tydeliggør, hvordan mennesket besidder en force, når det kommer til at være fleksibel, tilpasningsdygtig og kreativ:

> *"(...) mennesker er mere fleksible, tilpasningsdygtige og kreative end automatisering og er derfor bedre i stand til at reagere på ændringer i uforudsete forhold"* (Parasuraman & Riley, 1997: 232, oversat).

Af citatet fremgår det, hvordan mennesker i højere grad er i stand til at håndtere og reagere på ændringer og uforudsete forhold sammenlignet med automatiserede systemer. Det er yderligere værd at bemærke, hvordan denne påstand er overensstemmende med Kahneman et al.'s (2021) forbehold overfor, at vurderinger ikke alene skal foretages af algoritmer, da mennesker i nogle tilfælde vil besidde en viden om sammenhænge og kontekst, som vil kunne mindske fejlniveauet.

Denne forståelse for menneskers og modellers divergerende forcer er med til indikere, hvorfor et samarbejde mellem mennesker og automatiserede modeller kunne være oplagt, når det drejer sig om komplekse beslutningsprocesser såsom karaktergivning af dansk skriftlige opgaver. Ved at kombinere de menneskelige forcer vedrørende fleksibilitet, tilpasningsdygtighed og kreativitet med modellernes styrker inden for præcision og effektivitet, kan det herved argumenteres, at der skabes grundlag for en synergieffekt. En synergi hvor mennesket og modellen sammen må forventes at være bedre rustet til at håndtere komplekse beslutningsprocesser end, hvad mennesket og modellen hver især vil kunne. For at denne synergi vil kunne opstå, kan tillid til den automatiserede model dog fremhæves som værende afgørende (jf. Parasuraman & Riley, 1997: 236-237). Denne tillid vil blive beskrevet i nedenstående afsnit.

5.2.4 Definition af tillid

Når det kommer til dette studies teoretiske forståelse af tillid, vil der i dette studie blive anvendt en definition af Mayer et al. (1995). Herved vil det også blive tydeliggjort, hvordan denne definition går i tråd med Parasuraman & Rileys (1997) definition af tillid. Ifølge Mayer et al. (1995) skal tillid forstås som:

"En parts villighed til at være sårbar over for en anden parts handlinger baseret på forventningen om, at den anden vil udføre en bestemt handling, der er vigtig for tillidsgiveren uanset evnen til at overvåge eller kontrollere den anden part"
(Mayer et al., 1995: 712; Glikson & Woolley, 2020: 9, oversat)

Ovenstående citat viser hermed, hvordan tillid kan være en kompleks størrelse, når det drejer sig om at udvikle og opretholde denne i relationer. Samtidig kan

det også udledes, at tilliden består af en række forventninger til modpartens evne og vilje til at handle hensigtsmæssigt uagtet af, om det er muligt at kontrollere dette. I studiet problemanalyse blev det fremlagt, hvordan LL-modellernes manglende gennemsigtighed kan forekomme som en begrænsning for anvendelsen heraf (jf. afsnit 4.3.3). Med ovenstående definition af tillid, kan det siges, at samarbejdet mellem automationen og mennesket ikke nødvendigvis bliver begrænset af den manglende gennemsigtighed så længe, at forventningen om, at modellen vil foretage pålidelige beslutninger, ikke modbevises. Hermed kommer tilliden til udtryk ved, at mennesket forekommer villig til at gøre sig sårbar i relationen uden mulighed for at kunne overvåge eller kontrollere den automatiserede models fremgangsmåde.

Ved at betragte Mayer et al.'s (1995) definition af tillid i henhold til Parasuraman & Riley (1997), kan det fremhæves, hvordan der forekommer en mulighed for overførbarhed af Mayer et al.'s tillidsbegreb. I studiets problemanalyse fremlægges det, hvordan tillid mellem et menneske og en automatiseret model kan betragtes som værende påvirket af de samme faktorer, som tilliden mellem to mennesker (jf. Madhaven & Wiegmann, 2007). Ifølge de teoretiske antagelser baseret på Parasuraman & Riley (1997) har tilliden til automatiseringen betydning for, hvorvidt tillidsgiveren er villig til at bruge automatiseringen og dermed sætte sig selv på spil ved at overgive en del af opgaveløsningen til automationen. Denne overgivelse kan argumenteres at være overensstemmende med Mayer et al's definition af tillid, som indebærer en villighed til at være sårbar. Af den grund er der herved et solidt argument for, at Mayer et al.'s (1995) definition af tillid kan anvendes i tråd med Parasuraman & Rileys (1997) teoretisering af tillidens betydning for samarbejdet mellem mennesker og automation.

Med henblik på at undersøge, hvorvidt LL-modeller kan anvendes i karaktergivningen, kan det gennem denne definition af tillid siges, at censorens tillid til AI-teknologien har betydning for censorens villighed til at gøre sig sårbar og herved bruge LL-modellen som medbedømmer i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen (jf. Mayer et al., 1995; Parasuraman & Riley, 1997).

5.2.5 Brug af modellerne

Med afsæt i ovenstående forståelse af tillid påpeger Parasuraman & Riley (1997: 136-137), hvordan niveauet af tillid til automationen er afgørende for menneskets brug af automation. Når det kommer til brugen af en automatiseret model, tydeliggør Parasuraman & Riley (1997: 230), hvordan der kan forekomme forskellige niveauer af tillid, som ikke alle er hensigtsmæssige. I den forbindelse fremlægger de henholdsvis 'overbrug' og 'underbrug', som værende begreber for uhensigtsmæssige brugsformer af automationen.

Ifølge Parasuraman & Riley (1997: 238) kan 'overbrug' karakteriseres som det niveau af brug, der opstår, når automationen bruges, hvor den ikke bør bruges eller, hvor den bruges uden kritisk stillingtagen hertil. Dette 'overbrug' er typisk resultatet af et uforholdsmæssigt højt niveau af tillid til automationen. En 'overbrug' af automatiserede processer på grund af for høj tillid er problematisk, idet mennesket tilsidesætter sin kritiske tænkning, hvilket kan resultere i fejl. Disse fejl kan ifølge Kahneman et al. (2021) være karakteriseret som støj.

Som nævnt tidligere kan forskellige bias måles som systemstøj, idet afvigelserne fra gennemsnittet forekommer forskellige (jf. afsnit 5.1.5). Med henblik på at forstå, hvorfor overbrug forekommer, tydeliggør Parasuraman & Riley (1997), hvordan overbrug kan opstå som følge af brugen af heuristikker i beslutningstagning (jf. Tversky & Kahneman, 1974; Parasuraman & Riley, 1997: 239). Heuristikker er karakteriseret som kognitive genveje, der kan anvendes i beslutningstagning, hvor de kan benyttes til at mindske ressourceforbruget ved en kognitiv vurdering. Idet brugen af disse heuristikker skyldes en 'besparelse' i de kognitive ressourcer anvendt i vurderingen, opstår der herved også en bagside ved vurderingerne, fordi de er baseret på en intuitiv og ikke-anstrengende tænkning og dermed i højere grad kan føre til fejl. Disse kognitive genveje kan opfattes som mønstre, der kommer til udtryk gennem bias hos individet. Som tidligere nævnt kan forskellige bias i en aggregeret form komme til udtryk som mønsterstøj (jf. system-1-tænkning i Tversky & Kahneman, 1974; Kahneman et al., 2021: 191). Med afsæt i Parasuraman & Riley (1997) kan det dermed udledes, hvordan det kan svække kvaliteten af karaktergivningen, hvis censorerne overbruger LL-modellen som en heuristik til at foretage en hurtigere og mere effektiv karaktergivning.

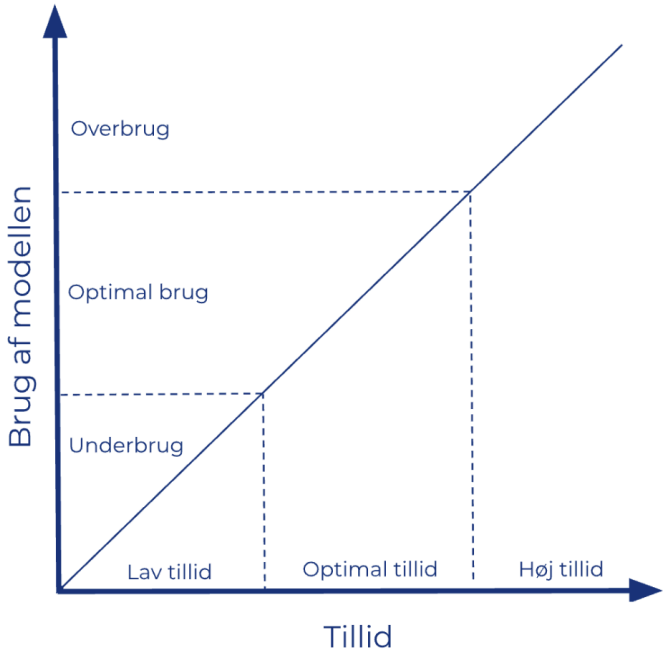
Dette skyldes, at der ikke vil opstå en synergieffekt, som censoren og LL-modellen med deres forskellige forcer hver især ellers ville kunne have bidraget til.

Foruden overbrug fremlægger Parasuraman & Riley (1997) også, hvordan 'underbrug' også kan forekomme uhensigtsmæssig. Ifølge Parasuraman & Riley (1997: 244) er en underbrug af modellen ofte et resultat af, at mennesket besidder et lavt niveau af tillid overfor modellen. At et uhensigtsmæssigt lavt niveau af tillid kan føre til underbrug foreligger særligt problematisk i situationer, hvor opgaveudførelsen eller beslutningsprocessen kunne have været mere pålidelig med modellens input (Parasuraman & Riley, 1997: 233). Gennem Kahneman et al. (2021: 113) blev det tydeliggjort, hvordan alle menneskelige vurderinger indebærer fejl i form af støj (jf. afsnit 5.1.2). Når bedømmeren underbruger modellen, vil de menneskelige fejl, som finder sted ved karaktergivningen, ikke blive reduceret. Denne underbrug kan herved, på samme måde som overbrug have negative konsekvenser for fejlniveauet ved karaktergivningen. I henhold til karaktergivningen vil et underbrug herved også kunne antages at medføre en fortsat upålidelig karaktergivning, hvor elevens udvalg af ungdomsuddannelse indskrænkes unødigt eller, at befolkningen finder karaktersystemet utroværdigt og af den grund mister tilliden til uddannelsessystemet.

5.2.6 Visualisering af sammenhængen

Gennem de ovenstående afsnit er det blevet klarlagt, hvordan der ifølge Parasuraman & Riley (1997) forekommer en sammenhæng mellem tillid og brugen af en automatiseret model. Niveauerne af 'tillid' og 'brug' operationaliseres i dette studie som to kontinuummer, der spænder fra henholdsvis et lavt niveau af tillid og underbrug til højt niveau af tillid og overbrug. Sammenhængen mellem disse kontinuummer er visualiseret nedenfor:

Figur 5.9: Sammenhæng mellem tillid og brug af modellen



Figuren er et illustreret eksempel på sammenhængen mellem 'brug' og 'tillid' udformet på baggrund af Parasuraman & Riley (1997)

Af visualiseringen fremgår det, hvordan sammenhængen mellem tillid og brug af en automatiseret model kan anskues som værende kompleks, idet grænserne for den optimale tillid og dermed den hensigtsmæssige brug af automationen ikke nødvendigvis er givet på forhånd. I den forbindelse gør det sig dog gældende, at der forekommer to yderpunkterne: underbrug og overbrug. Underbrug, der opstår som følge af en lav grad af tillid samt overbrug, der opstår som følge af en høj grad af tillid. Begge disse yderpunkter indikerer to former for brug, hvor mennesket gennem brugen ikke formår at opnå en synergieffekt med modellen. Det synergiske samarbejde opstår i stedet, når tillidsniveauet indfinder sig på midten af x-aksen. Det er herved afgørende, at den menneskelige bedømmer på den ene side er åben over for LL-modellens vurderinger, hvorved den menneskelig bedømmer til en vis grænse gør sig sårbar, samtidig med at den menneskelige bedømmer fastholder den kritiske tænkning i anvendelsen af modellen output, som er væsentligt for, at det fagproffesionelle skøn opretholdes.

5.2.7 Studiets hypoteser om tillid

Gennem det ovenstående er Parasuraman & Rileys (1997) teoretiske antagelser om tillids betydning for brugen af en automatiseret model blevet klarlagt. Med

udgangspunkt heri forventes det, at censorernes tillid til algoritmer og maskinlæring vil have betydning for villigheden til at bruge en LL-model ved karaktergivningen i folkeskolen. Her forventes det, at en lav grad af tillid til AI-modeller vil resultere i underbrug af LL-modellen, mens en høj grad af tillid omvendt vil kunne resultere i, at censoren tilsidesætter sit eget faglige skøn og overbruger LL-modellen i karaktergivningen.

Gennem Parasuraman & Riley (1997) kan det hermed udledes, hvordan den mest optimale brug af LL-modellen i karaktergivningen vil være karakteriseret ved et passende niveau af tillid fra censoren til LL-modellen, hvis samarbejdet i karaktergivningen skal kunne bidrage til en reduktion af det samlede fejlniveau. Vi betragter hermed tilliden til automation som værende afgørende for, hvorvidt brugen af LL-modellen kan reducere støjniveauet i karaktergivningen. I studiets problemanalyse blev det ydermere tydeliggjort, hvordan dele af litteraturen fandt indikationer på, at tilliden til en vurdering blev påvirket af informationen om, at vurderingen var foretaget af en AI-model frem for et menneske (jf. afsnit 4.4; Jakesch et al., 2019). Med en viden herom ønskes dette undersøgt yderligere i indeværende studie.

Med udgangspunkt i denne teoretiske antagelse om, at tillid har betydning for brugen af en AI-model og dermed støjen, og at informationen om, hvorvidt medbedømmeren er et menneske eller en model, har betydning for tilliden og dermed brugen, har studiet opsat følgende hypotese:

H3: Informationen om en AI-model som medbedømmer påvirker tilliden, der gennem brugen har betydning for støjniveauet

Af studiets tredje hypotese kan det udledes, hvordan denne har til hensigt at teste to underliggende sammenhænge. Af den grund vil der yderligere blive opsat to delhypoteser med formålet at afprøve hovedhypotesen. Den første af disse delhypoteser er udformet på baggrund af en uvished i litteraturen om, hvorvidt tilliden påvirkes positivt eller negativt af informationen om, at en vurdering er genereret af en AI-model eller et menneske (jf. Jakesh et al, 2019; Lerch et al 1997; Madhaven & Wiegmann 2007). Derfor opstilles følgende delhypotese:

H3.A: Informationen om en AI-model som medbedømmer har betydning for tilliden til medbedømmeren

Foruden ovenstående delhypotese vil studiet også teste en anden delhypotese, som tager afsæt i Parasuraman & Rileys (1997) teoretiske antagelse om, at tilliden til en automatiseret model har betydning for, hvordan mennesket vil gøre brug af modellen. Brugen kan gennem Parasuraman & Riley (1997) både forekomme som værende hensigtsmæssig eller være karakteriseret som underbrug og overbrug. Dette har det ledt til følgende delhypotese:

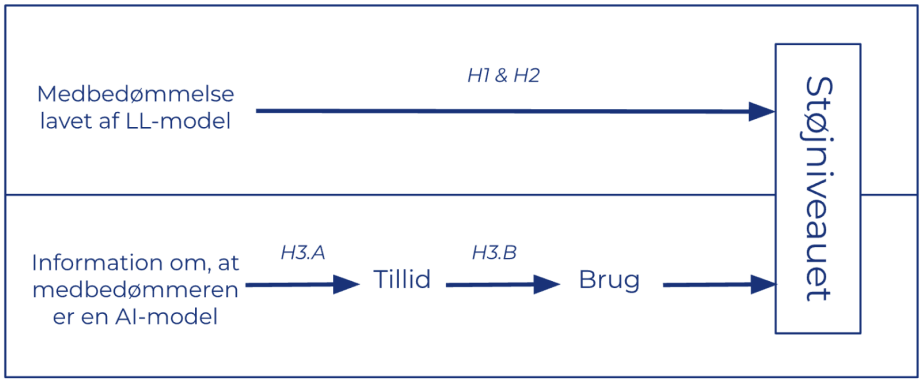
H3.B: Tilliden til medbedømmeren har betydning for brugen af medbedømmeren

Hermed kan det opsummeres, hvordan de to delhypoteser (jf. H3.A og H3.B) vil have til hensigt at teste to underliggende sammenhænge med formålet at afprøve hovedhypotesen (jf. H3).

5.3 Opsamling på studiets teoretiske antagelser

I det indeværende kapitel er studiets teoretiske apparat vedrørende støj og tillidens betydning for brugen af automatiserede modeller samt studiets tre hypoteser blevet fremlagt. Med henblik på at danne en samlet forståelse for sammenhængen mellem støj, tillid og brugen af en LL-model, så er følgende kausalmodel blevet udviklet:

Figur 5.10: Studiets kausalmodel



Gennem ovenstående kausalmodel antages det, at der er flere faktorer, som forventes at have indflydelse på støjniveauet. Først kan det gennem Kahneman et al. (2021) siges, at der er en forventning om, at en LL-model som medbedømmer vil kunne reducere støjniveauet. Dette er dog forudsat at LL-modellen vil være mindre støjende end den gennemsnitlige censor.

For det andet er der en forventning om, at tilliden, brugen og dermed støjniveauet påvirkes af informationen om, hvorvidt medbedømmeren er en LL-model eller en anden censor. Dette sker ud fra en forståelse af, at overbrug og underbrug i dette studie vil blive defineret ud fra fejlniveauet, hvor der ved karaktergivningen kunne være forekommet en støjreduktion ved henholdsvis mere eller mindre brug af modellen.

6. Forskningsdesign og metode

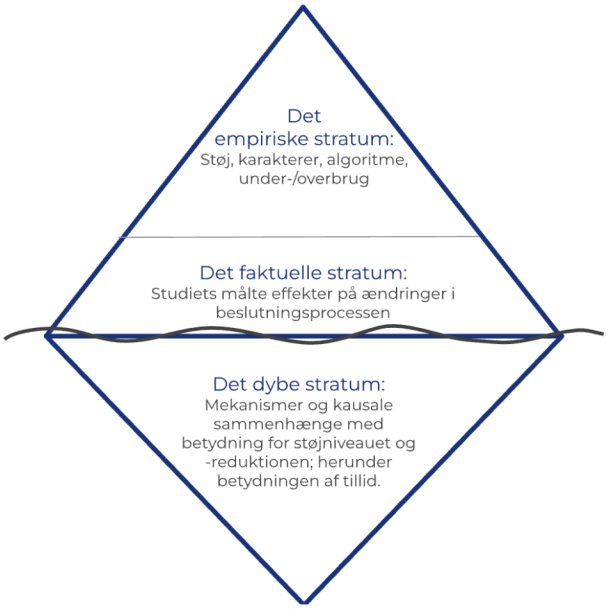
I følgende kapitel vil studiets forskningsdesign og metodiske overvejelser blive fremlagt. Dette sker ved, at studiets videnskabsteoretiske orientering først vil blive tydeliggjort i afsnit 6.1. Herefter vil argumentationen for studiets forskningsdesign i form af et surveyeksperiment blive fremlagt i afsnit 6.2. I forlængelse heraf vil afsnit 6.3 have til formål at klarlægge indholdet i studiets surveyeksperiment, mens selve udførelsen af eksperimentet samt en sammenligning af surveyeksperimentets deltagere og populationen vil blive afdækket i afsnit 6.4. Herefter vil studiets analysestrategi blive udfoldet i afsnit 6.5. Afslutningsvis vil der i afsnit 6.6 være en opsamlende vurdering af studiets kvalitet.

6.1 Videnskabsteoretiske forhold

I det følgende afsnit vil studiets videnskabsteoretiske ståsted herunder ontologi og epistemologi blive tydeliggjort. Dette sker ud fra en forståelse af, at studiets videnskabsteoretiske position besidder en afgørende rolle i udformningen af studiets metodologi (jf. Ingemann, 2017: 24-25).

Indeværende studie har en interesse i at undersøge, hvorvidt en LL-model kan bidrage til at reducere støjen i karaktergivningen i folkeskolen. Med henblik på at skabe forståelse for ontologien bag dette fænomen kan Kahneman et al. (2021) benyttes. Gennem Kahneman et al. (2021) kan det klarlægges, hvordan menneskelige beslutningsprocesser er komplekse og uigennemskuelige, mens støjeffekten ved disse beslutningsprocesser foreligger direkte målbar. Af den grund kan det ifølge Kahneman et al. (2021) argumenteres, at karaktergivningen i den danske folkeskole bør betragtes som værende lagdelt. Det synlige målbare lag består dermed af den givne karakter, den målbare støj og de faktuelle forhold, som bedømmelsen er givet under (jf. det empiriske stratum på figur 6.1). Foruden disse målbare faktorer, argumenterer Kahneman et al. (2021), at der i karaktergivningen forekommer sammenhænge, som ikke er 'direkte målbare'. Dette er visualiseret gennem det dybe stratum, hvor de kausale sammenhænge, der ligger til grund for støjen samt reduktionen af støjen, kunne være eksempler på sammenhænge, der ikke er direkte målbare. At der foreligger en række ikke direkte målbare faktorer i det dybe stratum medvirker til, at det faktuelle stratum må benyttes til at skabe en idé om sammenhængene i det dybe stratum.

Figur 6.1: Studiets ontologiske isbjerg



Kilde: Med inspiration fra Bhaskar et al.'s (2013:41) tre domæner

Med afsæt i ovenstående ontologiske tilgang til karaktergivningen leder det videre til, at studiet epistemologisk erkender, at støj kan måles. Effekten af LL-modellen som medbedømmer i karaktergivningen skal derimod testes, hvilket vil ske på baggrund af en måling af de synlige faktorers indvirkning på støjen. Med henblik på at undersøge, hvorvidt karaktergivningen i folkeskolen kan blive mindre støjende ved brugen af LL-modeller, vil det være essentielt at måle den afledte støj. Hermed kan det påpeges, hvordan dette studies ontologiske og epistemologiske erkendelser forekommer at være overensstemmende med den kritiske realisme (jf. Bhaskar et al., 2013).

6.2 Surveyeksperiment og randomisering

Ud fra ovenstående klarlægning af de ontologiske og epistemologiske antagelser, som vil ligge til grund for indeværende studie, vil følgende afsnit have til hensigt at redegøre for studiets metodologi herunder den valgte metode samt den tilhørende strategi for undersøgelsen. Denne undersøgelsesstrategi vil være formet af studiets forskningsspørgsmål:

I hvilken grad kan en LL-model som medbedømmer være med til at reducere støjen i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen, og har tillid til modellen betydning for, om støjen kan reduceres?

Af forskningsspørgsmålet fremgår det, hvordan der i høj grad er tale om en problemstilling, som vil indbefatte en undersøgelse af kausalsammenhængen mellem brugen af en LL-model og støjreduktion i karaktergivningen samt tillidens betydning herfor.

I henhold til at undersøge årsag-virkningssammenhænge kan det tydeliggøres, hvordan der på tværs af den samfundsvidenskabelige forskning og andre forskningsområder optræder et såkaldt 'kausalitetsproblem'. Dette sker ud fra en forståelse af, at det ikke foreligger muligt at replikere to identiske verdener, hvori to situationer vil kunne sammenlignes med henblik på at isolere effekten af en ændret variabel. At det ikke forekommer muligt at opstille to identiske verdener, hvor kun én variabel varierer, mens alle andre faktorer forbliver konstante, gør det vanskeligt at fastslå den nøjagtige kausalitet, hvoraf 'det fundamentale kausalitetsproblem' opstår (jf. Hariri, 2012: 196).

Med henblik på at adressere dette kausalitetsproblem spiller eksperimenter en afgørende rolle, idet det eksperimentelle design er den metode, som kommer dét at opsætte faktiske og kontrafaktiske situationer nærmest (Hariri, 2021: 196). Hermed kan det klarlægges, hvordan eksperimentet som metode formår at håndtere kausalitetsproblemet. Dette skyldes, at fordelen ved eksperimenter er, at disse kan *"(...) bruges til at etablere relationen mellem årsag og virkning"* samt *"påvise kausalrelationer"* (Petersen et al., 2007: 5-7). Ligeledes argumenterer McDermott (2002) for, at eksperimenter besidder potentiale i at forfine en viden om kausaliteter i statskundskaben.

Med henblik på at afdække, hvorvidt LL-modeller kan reducere støjniveauet ved karaktergivningen i folkeskolen, vil dette studie ud fra ovenstående argumentation bruge det eksperimentelle design gennem en deduktiv tilgang. I den forbindelse vil studiet teste de teoretisk baserede hypoteser om, hvordan censorernes brug af LL-modellen og tilliden hertil forventes at påvirke støjreduktionen (Andersen et al., 2020: 70, 77, 79).

For at kunne konkludere, hvorvidt der foreligger kausal inferens og dermed en klar årsag-virkningssammenhæng i studiets eksperiment, klarlægger Mutz (2011: 9), hvordan tre kriterier skal være opfyldt. Disse tre kriterier indebærer, at der 1) skal forekomme kovarians over tid, 2) årsagen skal tidsmæssigt komme før effekten og 3) relationen må ikke kunne skyldes ukendte tredje variabler og dermed alternative forklaringer. Ifølge Mutz kan eksperimentet gennem opfyldelse af disse kriterier hermed fastslå, at der er tale om kausalsammenhænge og ikke blot korrelationer eller spuriøse effekter. Hermed formår det eksperimentelle design at løse endogenitetsproblemet og kontrolproblemet, hvormed eksperimenterne besidder en styrke i at kunne afdække kausal inferens (jf. Blom-Hansen & Serritzlew, 2014).

På baggrund af eksperimenternes force i håndtere 'kausalitetsproblemet' betragtes disse eksperimentelle kontrollerede forsøg ofte som den gyldne standard inden for den videnskabelige metode (Bækgaard et al., 2020: 124). Når det kommer til eksperimenter kan det tydeliggøres, hvordan der findes forskellige typer såsom laboratorieeksperimenter, felteksperimenter, kvasieksperimenter, naturlige eksperimenter og surveyeksperimenter (Blom-Hansen & Serritzlew, 2014: 12-20; Bækgaard et al., 2020: 126). Her gør det sig gældende, at laboratorieeksperimentet anses for at være det mest rendyrkede eksperimentelle design. Dette er kendetegnet ved, at det i sammenligningen mellem kontrolgruppen og eksperimentgruppen er muligt at kontrollere alle variable, så det kan argumenteres, at eksperimentets grupper ikke varierer på andre måder end den intenderede variabel. Herved isolerer laboratorieeksperimenter kausaliteten og afdækker kun effekten af den manipulerede intervention, hvorved dens interne validitet forekommer særligt stærk (Blom-Hansen & Serritzlew, 2014: 12-13; McDermott, 2022: 32; Serritslew, 2007).

I tråd med laboratorieeksperimentet kan det også påpeges, hvordan surveyeksperimentet tilnærmelsesvist har samme stærke interne validitet, idet surveyeksperimentet på samme vis som laboratorieeksperimentet indeholder en randomiseret inddeling af kontrol- og eksperimentgruppe samt en eksogoen og forskermanipuleret intervention. Forskellen mellem de to typer af eksperimenter er blot, at den eksogene manipulerede intervention forekommer gennem mindre

ændringer af spørgsmålsformuleringer eller informationer i et survey, hvorefter effekten af denne intervention ved eksperimentgruppen vil blive målt gennem surveyen (Blom-Hansen & Serritzlew, 2011: 16-17).

Surveyeksperimentet adskiller sig fra øvrige eksperimenter i social- og samfundsvidenskaben ved at være relativt nemmere at gennemføre, idet eksperimentet ikke afhænger af en fysisk tilstedeværelse. Ydermere er antallet af respondenter i surveyeksperimentet ubegrænset, idet der under surveyen bliver randomiseret løbende. På den måde behøver det eksakte antal respondenter ikke være bestemt på forhånd i et surveyeksperiment. I tilfælde af at n-respondenter nærmer sig undersøgelsens population (N) skaber surveyeksperimentet blot mulighed for en højere ekstern validitet (Hendrickson et al., 2019: 85-86). Selvom surveyeksperimentet besidder en række fordele, kan det gennem denne type af eksperiment være sværere at påvise en effekt, da interventionerne typisk vil være mere lavintensive sammenlignet med andre eksperimenter, fordi det typisk er baseret på mindre ændringer. Endvidere kan det også i surveyeksperimentet sammenlignet med laboratorieeksperimenter være sværere at dokumentere, at andre faktorer ikke kan have påvirket effekten af interventionen (Blom-Hansen & Serritzlew, 2011: 16-17).

Med udgangspunkt i ovenstående er det herved belyst, hvordan validiteten kan variere afhængigt af eksperimenttypen. Disse forskelle i eksperimenterne er opsamlet i nedenstående tabel:

Tabel 6.2: Validitet i eksperimenter

	Intern validitet	Ekstern validitet	Økologisk validitet
Eksperimenter generelt	Høj	Lav	Varierende
Laboratorieeksperiment	Høj	Lav	Lav
Surveyeksperiment	Høj	Varierende	Varierende

Kilde: Frit fra Blom-Hansen & Serritslev, 2014

Af ovenstående fremgår det, hvordan eksperimenter generelt besidder en høj intern validitet, hvilket bunder i forcen vedrørende kausal interferens. Dog kan det

også tydeliggøres, at den eksterne validitet i eksperimenter ofte er svækket, hvilket typisk sker i kraft af, at eksperimenternes respondenter ikke er repræsentative for populationen. I den forbindelse besidder surveyeksperimentet dog en fordel i, at denne tillader et stort n, hvilket i nogle tilfælde kan være med til at styrke den eksterne validitet. Afslutningsvist kan surveyeksperimentet også have en styrket økologisk validitet, da denne type af eksperiment afhængigt af den undersøgte kausalitet kan have bedre forudsætninger for at kunne afspejle elementer fra respondenternes hverdag (Blom-Hansen & Serritzlew, 2014: 6; Bryman, 2016: 42; Bækgaard et al., 2020: 125-127).

Med afsæt i ovenstående vil der i dette studie blive foretaget et surveyeksperiment med henblik på at afdække sammenhængen mellem brugen af en LL-model og støjreduktion ved karaktergivningen i folkeskolen samt tillidens betydning for denne støjreduktion.

6.3 Surveyeksperimentets udformning

Med en viden om, at dette studie vil basere sin undersøgelse på et surveyeksperiment, vil det følgende afsnit have til hensigt at skitsere tilrettelæggelsen af selve surveyeksperimentet. Indledningsvist vil der være en redegørelse for studiets randomisering samt en gennemgang af surveyens baggrundsvariabler. Efterfølgende vil overvejelserne vedrørende udvælgelsen af den dansk skriftlige opgave samt deltagernes første karaktergivning blive fremlagt. Dernæst vil LL-modellens bedømmelse af elevteksten samt målingen af deltagernes tillid hertil og sidste karaktergivning blive udfoldet. Afslutningsvis vil surveyens debriefing blive fremlagt.

6.3.1 Randomisering af respondenterne

I studiets surveyeksperiment vil respondenterne ved starten af surveyen blive tilfældigt inddelt i en kontrol- og eksperimentgruppe på baggrund af en simpel randomisering (Noor et al., 2022). Dette sker ud fra forståelsen af, at denne randomisering foreligger afgørende for studiets evne til at afdække en kausal inferens og dermed, hvorvidt der er tale om en årsags-virkningssammenhæng mellem brugen af LL-modellen og støjreduktion i karaktergivningen (jf. McDermott, 2022: 33).

Med henblik på at tilnærme sig principperne for en komplet randomisering vil studiet foretage en randomisering baseret på starttidspunkt. Herved vil sekundet for respondentens påbegyndelse af spørgeskemaet være udslagsgivende for, hvorvidt vedkommende bliver en del af kontrol- eller eksperimentgruppen (jf. Sniderman, 2011: 103-104; Bækgaard et al., 2020: 131-132).

Figur 6.3 : Den simple randomisering

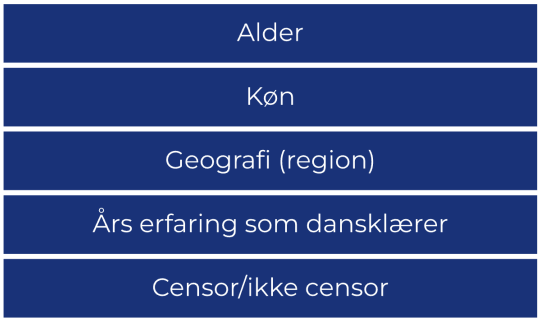
Kontrolgruppe	Eksperimentgruppe
Respondenter som påbegynder spørgeskemaet i et **lige sekund**	Respondenter som påbegynder spørgeskemaet i et **ulige sekund**

At respondenterne inddeles tilfældigt baseret på sekundet sker ud fra antagelsen om, at dette ikke vil skabe nogen systematisk forskel på grupperne i henhold til den undersøgte kausalitet. Eventuelle forskelle, som studiet vil finde mellem grupperne, vil dermed argumenteres at kunne tilskrives manipulationen af den uafhængige variabel. På den måde vil studiet på baggrund af den tilfældige inddeling af respondenterne kunne fremsætte mere pålidelige konklusioner omkring kausal inferens i henhold til LL-modellens betydning for støjen i karaktergivningen. Senere i studiets metode vil det på baggrund af uafhængighedstests blive tydeliggjort, hvordan denne randomisering er lykkedes efter hensigten, idet der på tværs af grupperne forekommer tilnærmelsesvise lige fordelinger ved eksperiment- og kontrolgruppens baggrundsvariable (jf. afsnit 6.4.5).

6.3.2 Baggrundsvariable

Efter randomiseringen vil respondenterne blive bedt om at svare på en række baggrundsspørgsmål. Med en bevidsthed om, at deltagelsen i surveyen kan være tidskrævende, er antallet af baggrundsspørgsmål forsøgt begrænset med henblik på at minimere frafald undervejs. På baggrund heraf har studiet udvalgt fem baggrundsspørgsmål, som kommer til at vedrøre følgende:

Figur 6.4: Baggrundsvariable



Indsamlingen af disse informationer har to formål for studiets undersøgelse. Først og fremmest vil baggrundsvariablene gøre det muligt at eftertjekke randomiseringen af studiets kontrol- og eksperimentgruppe. Gennem en sammenligning af kontrol- og eksperimentgruppens baggrundsvariable kan det gennem statistiske uafhængighedstest (Chi^2-test og Cramer's V-test) sikres, at der ikke forekommer forskelle mellem grupperne, hvilket er afgørende, når eksperimentet har som forudsætning, at grupperne forekommer ens (jf. Bækgaard et al., 2020: 136).

For det andet vil baggrundsvariablerne kunne give et indblik i, hvorvidt studiets stikprøve forekommer tilnærmelsesvis repræsentativ for populationen. Selvom det som udgangspunkt ikke er afgørende at have en repræsentativ stikprøve i et eksperimentelt design, så kan en viden om stikprøvens repræsentativitet alligevel give en indikation om studiets eksterne validitet og dermed i hvor høj grad, at resultaterne ville kunne forventes at besidde en generaliserbarhed (jf. Hartman, 2021). Denne sammenligning fremlægges i afsnit 6.4.4.

6.3.3 Den anvendte elevtekst

Efter at respondenterne har besvaret de ovennævnte fem baggrundsspørgsmål, vil de blive præsenteret for den udvalgte elevtekst, som de efterfølgende skal bedømme. I den første del af eksperimentet skal alle respondenterne vurdere den udvalgte elevtekst på samme måde, som de vil gøre under normale omstændigheder i karaktergivningen ved folkeskolens afgangseksamen (jf. UVM, 2023). Denne 'normale' bedømmelsespraksis kommer hermed til at fungere som studiets baseline, der senere vil blive anvendt som kontrol i eksperimentet.

Udvælgelsen af danskopgaven til surveyeksperimentet er sket på baggrund af flere kriterier. Med henblik på at styrke den økologiske validitet har der først og fremmest været ønske om, at elevteksten skulle være en reel opgavebesvarelse udarbejdet af en elev i 9. klasse til folkeskolens afgangseksamen i dansk skriftlig fremstilling. Dette blev en realitet, da studiet fik adgang til en elevtekst fra folkeskolens afgangseksamen i sommeren 2021.

Foruden overvejelserne vedrørende den økologiske validitet kan det også fremhæves, hvordan studiet har udvalgt elevteksten ud fra teoretiske overvejelser. Med ønsket om at måle støjniveauet og afdække muligheden for støjreduktion, blev opgaven udvalgt i henhold til Kahneman et al.'s (2021) argumentation om, at der forekommer størst variabilitet i vurderingen af 'det gode'. Dette skyldes, at der ifølge Kahneman et al. (2021: 56) typisk forekommer uenighed om, hvilke kriterier der berettiger en god bedømmelse og dermed én høj karakter, mens der er større enighed om, hvad der karakteriseres som en dårlig præstation og dermed en lav karakter, når ingen kriterier er opfyldt. Af den grund søgte studiet bevidst efter en danskopgave, hvor karakteren enten var 7 eller 10, mens opgaver med karakteren 12 bevidst blev undgået, da dette kunne medvirke til en begrænset mulighed for at måle varians.

Med henblik på at styrke den økologiske validitet yderligere har studiet også haft et ønske om at benytte en elevtekst, som forekommer relativt ny. Da det ikke har været muligt at finde en elev med adgang til sin eksamensopgave, som enten havde fået karaktererne 7 eller 10 til folkeskolens afgangsprøve i 2023, vil der i stedet blive anvendt en eksamensopgave, som er blevet skrevet til folkeskolens afgangseksamen i sommeren 2021 og bedømt til karakteren 10 (se elevteksten i bilag 1).

Efter at have udvalgt en elevtekst søgte studiet aktindsigt i opgavesættet, rettevejledningen samt vurderingsskemaet fra folkeskolens afgangseksamen i sommeren 2021 (jf. bilag 4 og bilag 13). Dette skete ud fra en forståelse af, at alle tre dele er elementer, som censorerne skal anvende og have adgang til i surveyeksperimentet. Studiet blev tildelt aktindsigt i alle tre dele. Dog forekommer aktindsigten i opgavesættet begrænset i den forstand, at opgavesættet er omfattet af ophavsretten (jf. §1 i ophavsretsloven, 2014). Dette

betyder dermed, at det ikke er tilladt at dele opgaveformuleringen. Af den grund har studiet selv været nødt til retrospektivt at lave en opgaveformulering, som passer til den anvendte elevtekst, som efterfølgende blev valideret af en dansklærer. Opgaveformuleringen lød som følger:

"Skriv et essay om tro. I essayet skal du reflektere over din personlige relation til tro, og hvordan andres tro kan være forskellig fra din. Essayet skal skrives som et avisindlæg"

At studiet har været nødsaget til selv at udforme en opgaveformulering er naturligvis med til at svække den økologiske validitet en smule, da ovenstående opgaveformulering og dens krav ikke er direkte svarende til den formulering, som elevteksten til afgangseksamen er skrevet på baggrund af. Selvom studiets økologiske validitet kan være en smule svækket gennem den nye opgaveformulering, kan det dog påpeges, hvordan studiets økologiske validitet generelt er styrket ved, at surveyeksperimentet vil indeholde konkrete hjælpeværktøjer fra karaktergivningen i folkeskolen (jf. vurderingsskemaet og rettevejledningen). Dette skulle gerne sikre en vis genkendelighed for respondenterne. Afslutningsvis kan det argumenteres, at den økologiske validitet også er styrket ved, at karaktergivningen i surveyeksperimentet foregår digitalt, hvilket også er gældende praksis ved folkeskolens afgangseksamen. Det skal dog bemærkes, at filadgangen i den 'virkelige karaktergivning' kan forekomme anderledes end i dette surveyeksperiment, hvor elevteksten er tildelt via et link i surveyen, mens opgaveformuleringen, rettevejledningen og vurderingsskemaet er fremlagt inde i surveyen (se bilag 2).

6.3.4 Deltagernes første vurdering

Efter gennemlæsningen af elevteksten skal respondenterne foretage deres første vurdering, hvor de i overensstemmelse med bedømmelsespraksissen ved dansk skriftlig fremstilling vil gøre brug af vurderingsskemaet, som er udviklet af Børne- og Undervisningsministeriet (jf. UVM, 2023: 24). Ved hjælp af vurderingsskemaet skal respondenterne afgive en karakter til hvert vurderingsområde, hvorved der foretages en udsat holistisk bedømmelse. Nedenfor ses en oversigt over de

vurderingsområder, som respondenterne skal foretage deres vurdering ud fra, mens det fulde vurderingsskema med tilhørende forklaringer kan ses i bilag 3.

Tabel 6.5: Vurderingsområder ved dansk skriftlig fremstilling til FP9

Vurderingsdimension	Vurderingsområder	Vurdering
Funktion	Skrivesituation	
	Opgavekrav	
Indhold	Mening	
	Ressourcer	
Form	Tekstsammenhæng	
	Skrift og andre modaliteter	
Samlet karakter		
Begrundelse		

Efter at respondenten har afgivet de seks delkarakterer, skal de afgive en samlet karakter og en begrundelse herfor. For denne første vurdering gør det sig gældende, at denne senere vil blive anvendt som baseline med henblik på longitudinalt at kunne sammenligne effekten ved, at alle respondenterne ved deres anden vurdering har fået en medbedømmer (jf. afsnit 6.3.8).

At respondenterne skal afgive en begrundelse kan både siges at være overensstemmende med vurderingspraksissen i dag samtidig med, at begrundelsen også anvendes som en validering af respondentens karaktergivning herunder særligt outliers. Endvidere udgør den også et 'attention check' med henblik på at sikre, at respondenten ikke besvarer surveyen uopmærksomt eller med andet motiv for øje (jf. Muszynski, 2023: 6-8).

Med henblik på at styrke validiteten har studiet valgt at benytte det gældende og opdaterede vurderingsskema fra Børne- og Undervisningsministeriet fremfor det, som blev anvendt til folkeskolens afgangsprøve 2021. At studiet har taget et valg om at anvende det nyeste vurderingsskema fra 2023 sker på baggrund af en

pilottest og feedback fra en beskikket censor i dansk (FP9). Feedbacken gik på, at en dansklærer som udgangspunkt altid vil tage afsæt i det gældende vurderingsskema. Hermed kan det argumenteres, hvordan gyldigheden af studiets konklusioner bevares ved, at undersøgelsen tager afsæt i den nuværende vurderingspraksis og dermed afspejler bedømmernes virkelighed, hvilket også er med til at styrke den økologiske validitet.

6.3.5 Den anvendte LL-model

Efter at have afgivet den første vurdering af elevteksten, vil alle respondenterne blive forelagt LL-modellens vurdering. I studiets teorikapitel er det blevet beskrevet, hvordan anvendelsen af en medbedømmer i form af en LL-model forventes at have en støjreducerende effekt. Dette gør sig gældende uanset om respondenterne ved, at det er en LL-model eller ej.

Indenfor LL-modeller kan der sondres mellem de generalistiske og de domænespecifikke modeller. De generalistiske modeller er trænet på et bredt sæt af data, og kan benyttes til mange forskellige formål. Modsat er de domænespecifikke modeller trænet på mere afgrænset og relevant data, hvorved de er udviklet til et mere specifikt formål. Til dette studie er der valgt en mere generalistisk sprogmodel, som ikke er udviklet specifikt til formålet omkring vurdering af skriftlige danskopgaver. På baggrund heraf forventes det, at en generalistisk model vil have dårligere forudsætninger for at bedømme elevteksten korrekt end en domænespecifik model (jf. Pattam, 2023; Pal et al., 2024; Tariq et al., 2024; Ramesh & Sanampudi, 2022). Det kan derfor argumenteres, hvordan udvælgelsen af den anvendte LL-model kan betragtes som en hård test af, hvorvidt sprogmodeller kan bidrage til støjreduktion i karaktergivningen. Ud fra principperne om en least-likely case kan det hermed argumenteres, at hvis en generalistisk LL-model kan reducere støjen i karaktergivningen, kan det med stor sandsynlighed også antages, at domænespecifikke LL-modeller vil kunne bidrage til en yderligere støjreduktion (jf. Andersen et al., 2020: 90).

Mere specifikt er sprogmodellen, der benyttes i dette studie til karaktergivningen af elevteksten, 'PDF Reader', som er udviklet af OpenAI. Denne er baseret på en

deep learning-algoritme, som evner at læse PDF-filer. At studiet har valgt at benytte sig af PDF Reader skyldes, at denne havde forudsætninger for at vurdere de funktionsmæssige krav i danskopgaven. Dette skyldes, at der i opgaveformuleringen foreligger direkte krav vedrørende funktion i form af layout, idet essayet skal skrives som et avisindlæg (jf. rettevejledningen i bilag 13). Hermed er det afgørende, at den anvendte LL-model i studiet evner at læse PDF-formater og dermed forventes at have bedre forudsætninger for at vurdere layoutet (jf. Børne- & Undervisningsministeriets vurderingsskema i bilag 4).

Når det kommer til LL-modellen er det også væsentligt at overveje, hvorvidt denne vil være i stand til at gentage sin vurdering af den samme elevtekst, da dette vil kunne have en indflydelse på studiets gentagelighed. I den forbindelse gør det sig gældende, at der kan være flere parametre, som kan påvirke, i hvor høj grad en AI-model vil forekomme konsistent i sine svar (jf. temperatur, top_p og top_k). Dette skyldes, at parametrene kan justeres, hvormed det er muligt at foretage en afvejning mellem ensartetheden og tilfældigheden i LL-modellens output (Blete, 2023; Gunes & Florczak, 2023:10; OpenAI, n.d; Singh, 2023). I den anvendte model gør det sig gældende, at disse parametre er indstillet til at have en middelværdi, hvorfor den opererer med en vis grad af tilfældighed, hvormed der som udgangspunkt ikke vil forekomme en fuldkommen konsistens i modellens output (OpenAI, n.d.).

I henhold til en karaktergivning vil det dermed forventes, at LL-modellen ikke nødvendigvis vil kunne gentage sin vurdering og komme frem til den samme karakter, hvis LL-modellen blev bedt om at foretage flere vurderinger af den samme elevtekst uden at ændre på standardindstillingen. Af den grund kan det argumenteres, at studiets eksterne reliabilitet forekommer svækket, idet det ikke som udgangspunkt vil være muligt at sikre, at LL-modellen udviklet af OpenAI vil kunne gentage en konsistent vurdering af elevteksten på et andet tidspunkt. I den forbindelse er det dog værd at bemærke, at der kan justeres i LL-modellernes parametre, hvormed det vil være muligt at øge LL-modellens konsistens og dermed styrke den eksterne reliabilitet. Det forventes dog ikke at være muligt at gøre modellen fuldkommen deterministisk, hvilket heller ikke nødvendigvis vil være gavnligt for karaktergivningen og reliabiliteten af denne (jf. Blete, 2023).

Dette sker ud fra en forståelse for, at der forekommer et trade-off, hvor modellen med højere frihedsgrader kan give mindre mekaniske vurderinger.

6.3.6 LL-modellens vurdering

Børne- og Undervisningsministeriets vurderingsskema blev benyttet i prompten til LL-modellen inden, at denne skulle foretage en vurdering af elevteksten. Efter træningen blev LL-modellen bedt om at give en karakter for hvert vurderingsområde (jf. analytisk vurdering, afsnit 4.2.1), foretage en samlet vurdering af elevteksten (jf. udsat holistisk vurdering, afsnit 4.2.1) samt give en begrundelse for karaktergivningen. Modellens vurdering er fremlagt gennem følgende tabel:

Tabel 6.6: LL-modellens vurdering af opgaven.

Vurderingsdimension	Vurderingsområde	Vurdering
Funktion	Skrivesituation	10
	Opgavekrav	10
Indhold	Mening	12
	Ressourcer	10
Form	Tekstsammenhæng	10
	Skrift og andre modaliteter	7
Samlet karakter		10
Begrundelse	Samlet set er elevens arbejde af høj kvalitet og viser en moden tilgang til essay-skrivning. Elevens evne til at reflektere over og diskutere et komplekst emne som tro gør vedkommendes essay både engagerende og tankevækkende. Mens der er plads til forbedring, specielt i forhold til sproglig finpudsning, er den overordnede præstation stærk og imødekommer opgavens krav effektivt.	

Note: Se promptene i bilag 4

Den ovenstående vurdering foretaget af LL-modellen anvendes i surveyeksperimentet. Efter at respondenterne har foretaget deres første vurdering af danskopgaven, vil de blive informeret om, at der er en medbedømmer, hvor de bliver vist dennes vurdering. Denne medbedømmer vil for alle respondenterne være LL-modellen. Forskellen og dermed interventionen er blot, at kontrolgruppen bliver fortalt, at vurderingen er udformet af en anden censor, mens eksperimentgruppen bliver fortalt, at den er udformet af en AI-algoritme:

Figur 6.7: Interventionen

Af visualisering fremgår det, hvordan surveyseksperimentets intervention vil bestå af en ændring i spørgsmålsformuleringen. Effekten af interventionen vil studiet undersøge gennem en hypotesetest, hvilket vil blive uddybet i analysestrategien (jf. afsnit 6.6)

6.3.7 Tillid til medbedømmeren

Foruden studiet ønsker at afdække, hvorvidt en LL-model kan være med til at reducere støjen i karaktergivningen, fremgår det gennem problemanalysen og teoriafsnittet også, hvordan studiet ønsker at undersøge, hvorvidt tilliden til modellen har betydning for støjreduktionen (jf. afsnit 4.4 og 5.2).

Idet tilliden fremlagt af Parasuraman & Riley (1997) kan betragtes som et abstrakt fænomen, er tilliden som teoretisk begreb ikke simpelt at måle. Derfor må målingen heraf udformes som et refleksivt indeks (jf. Andersen et al., 2010: 413-414). Dette indeks vil i studiet blive dannet på baggrund af spørgsmålsformuleringer fra European Social Survey (ESS) samt Den Danske Værdiundersøgelse, som er en del af The European Value Survey (European Social Survey, 2021; Rigsarkivet, 2017; World Value Survey, n.d.). Dette kan siges at højne

den interne validitet, da det kan argumenteres for at reducere risikoen for målefejl.

I forhold til ESS gør det sig gældende, at disse tillidsspørgsmål er angivet som dikotomier, hvor respondenten skal indplacere sig på en skala. Da alle dikotomier ikke forekom overførbare til studiets surveyeksperiment, er de omsat til udsagn, hvor respondenterne skal angive deres grad af enighed (se operationaliseringen i bilag 6). Dette resulterer samlet set i, at respondenterne skal besvare følgende fem tillidsspørgsmål:

Tabel 6.8: Måling af tillid

European Social Survey (ESS): Hvor enig er du i følgende udsagn?						
	Meget uenig	Uenig	Hverken enig eller uenig	Enig	Meget enig	Ved ikke
1. Jeg kan stole på medbedømmerens evne til at give karakterer						
2. Jeg bør være forsigtig, når det kommer til at anvende ovenstående karaktergivning						
3. Den ovenstående begrundelse og karakterer er fair						
4. Ovenstående karaktergivning kan være hjælpsom i min egen karaktergivning						
Den Danske Værdiundersøgelse						
	Ingen tillid	Næsten ingen tillid	Nogen tillid	Stor tillid	Ved ikke	
5. Hvor stor tillid har du til medbedømmerens karaktergivning?						

Note: En uddybende operationalisering kan ses i bilag 6.

Kilde: European Social Survey (2021); Rigsarkivet (2017); World Value Survey (n.d.)

6.3.8 Den endelige karaktergivning

Efter at respondenterne har foretaget deres første vurdering og efterfølgende er blevet præsenteret for medbedømmerens vurdering og tillidsspørgsmålene, vil respondenterne få mulighed for at revidere deres bedømmelse og fastsætte den endelige karakter. Dermed vil respondenterne som tidligere nævnt longitudinalt blive deres egen kontrol, når det kommer til at undersøge, hvorvidt en medbedømmer i form af LL-modellen kan bidrage til en støjreduktion (jf. hypotese 1; Taris, 2010).

Med afsæt i respondentens egen vurdering samt medbedømmerens vurdering i form af LL-modellen skal respondenten udfylde Børne- og Undervisningsministeriets vurderingsskema på ny inden, at der skal fastsættes en endelig og samlet karakter.

Tabel 6.9: Spørgsmål om anledning til revidering af karaktergivningen

" Nedenfor er din egen og din medbedømmers vurdering: *[Præsentation af respondenten og LL-modellens vurderinger]*   *Hvad ville du vurdere, at elevtekstens endelige karakterer skulle være ud fra vurderingerne ovenfor?*   *[De seks vurderingsområder; skrivesituation, opgavekrav, mening, ressourcer, tekstsammenhæng samt skrift og andre modaliteter] "*

At respondenterne skal udfylde vurderingsskemaet igen, har til formål at give en indsigt i, hvorvidt der er nogle vurderingsdimensioner eller -områder, hvor respondenterne er mere villige til at ændre deres karakterer efter at have set LL-modellens vurdering.

6.3.9 Debriefing

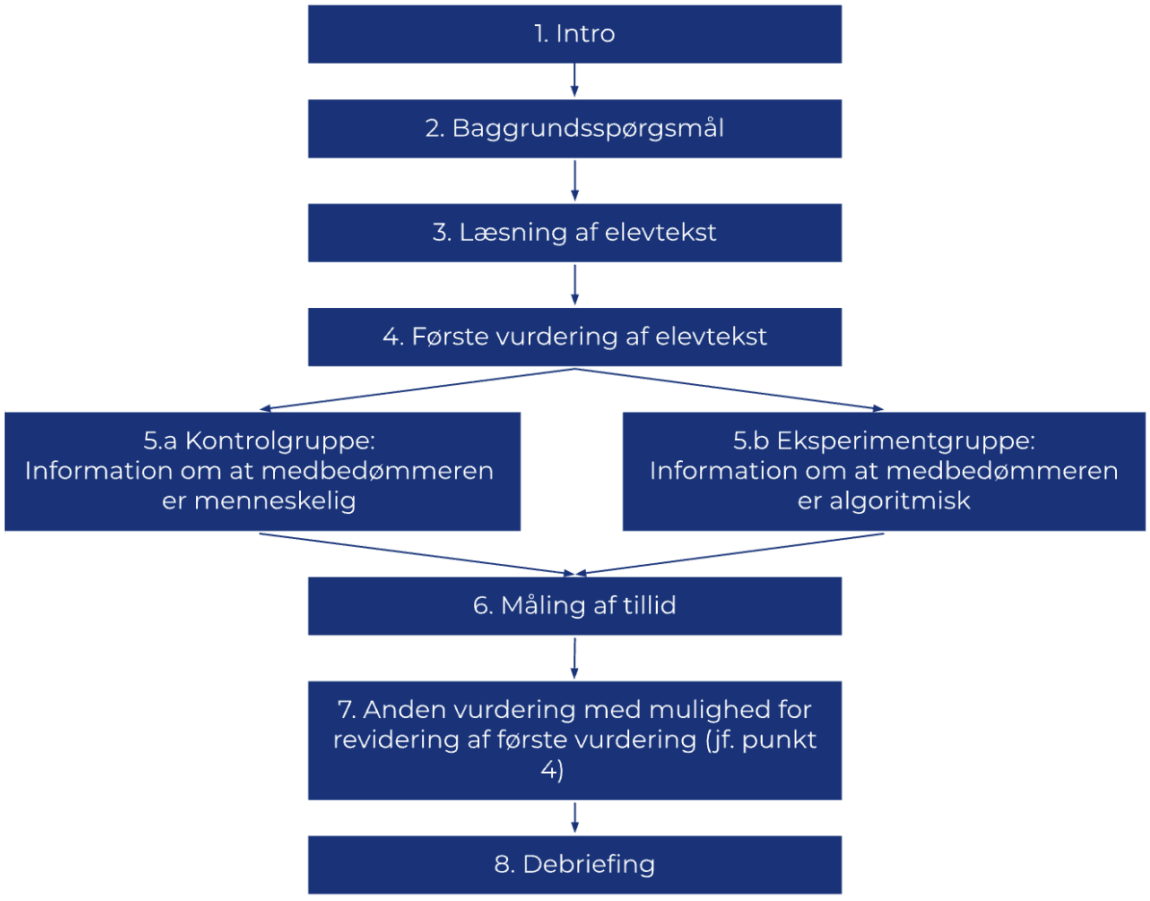
Respondenterne vil afslutningsvist modtage en debriefing. Dette sker ud fra en forståelse af, at der i eksperimenter altid vil være en vis mængde af tilbageholdte informationer, idet effekten af interventionen ikke kan måles, hvis

respondenterne kender til denne (Petersen et al., 2007: 11). En måde at imødekomme dette er at foretage en debriefing, hvor respondenterne på bagkant informeres om, at de har været med i et eksperiment og formålet med dette. I debriefingen bliver respondenterne bekendte med, at de har været en del af et eksperiment omkring brugen af AI i karaktergivningen i folkeskolen. Med henblik på at sikre effekten ved interventionen bliver respondenterne ydermere opfordret til ikke at drøfte indholdet i undersøgelsen med deres kollegaer, som endnu ikke har deltaget (jf. Bækgaard et al., 2020: 147). Se debriefingen i bilag 2.

6.3.10 Hele surveyen

I de forrige afsnit er studiets surveyeksperiment blevet skitseret. Dette leder videre til, at det nu er muligt at visualisere strukturen herfor, hvilket fremgår af nedenstående figur:

Figur 6.10: Surveyeksperimentets struktur



Note: En uddybning af udformningen af surveyeksperimentet findes i bilag 2.

6.4 Surveyeksperimentets kvalitet

Gennem det forrige afsnit er studiets undersøgelsesstrategi blevet klarlagt, hvormed der er skabt en forståelse for surveyeksperimentets design samt det tilhørende forskningsparadigme. I forlængelse heraf vil det følgende afsnit have til formål at diskutere forskningsdesignet og metodens kvalitet.

6.4.1 Kriterier for eksperimentets respondenter

Med henblik på at sikre validiteten har det været afgørende, at der udelukkende søges respondenter, som forekommer kvalificeret til at give karakterer til en dansk skriftlig fremstilling. Til folkeskolens afgangseksamen er det et beskikket censorkorps, som foretager vurderingen af elevernes opgaver. Disse censorer *"(...) skal have gennemført et undervisningsforløb, der i faget fører frem til folkeskolens prøve"* samtidig med, at der foreligger en forventning om, at de deltager i censorkurser for at blive opdateret på regler og procedurer under deres beskikkelse (Børne- og undervisningsministeriet, n.d.b).

Med en viden herom kan det argumenteres, hvordan det ville være fordelagtigt, hvis surveyeksperimentets respondenter bestod af medlemmer fra det danske censorkorps. Grundet GDPR er det dog ikke muligt at få tildelt kontaktinformation på de beskikkede censorer. Studiet gør i stedet brug af dansklærere fra udskolingen, idet disse må antages at udgøre det bedste alternativ af de foreliggende muligheder i henhold til den intenderede population. Dette kan argumenteres at svække surveyeksperimentets validitet, da alle dansklærere nødvendigvis ikke vil besidde de samme kvalifikationer vedrørende karaktergivning. Dog må det forventes, at dansklærerne har en vis erfaring med at give karakterer som led i undervisningen og ved prøveforberedende terminsprøver, som gør dem i stand til at give karakterer.

6.4.2 Pilotundersøgelsen

Inden distribueringen af surveyeksperimentet blev der foretaget en pilotundersøgelse af to dansklærere, hvor den ene var beskikket censor i mundtlig og skriftlig dansk (FP9). Pilottesten blev foretaget med formålet om at

identificere potentielle problemer eller uklarheder i eksperiment inden, eksperimentet blev distribueret. Pilottesten har været med til at sikre, at surveyeksperimentet forstås, virker og besvares efter hensigten (Bækgaard et al., 2020: 130). I forlængelse af disse tests er enkelte rettelser blevet foretaget. Dette har primært omhandlet mindre udfordringer vedrørende formuleringer i spørgeskemaet, hvor den efterfølgende justering har bidraget til en forbedring i klarheden af disse, hvilket er med til at styrke grundlaget for valide konklusioner (Andersen et al., 2020: 112). Pilottesten har afslutningsvist været med til at sikre, at surveyeksperimentet ifølge de to dansklærere afspejler en praksisnær kontekst, hvorfor studiet forventes at besidde en høj økologisk validitet.

6.4.3 Distribuering

Efter pilotundersøgelsen blev surveyeksperimentet distribueret. Her gør det sig gældende, at studiet har benyttet sig af flere forskellige distributionsmetoder med formålet om at maksimere antallet af respondenter. Først og fremmest er der i distributionen gjort brug af personlige relationer og netværk samt forskellige facebookgrupper for dansklærere i udskolingen og censorer i dansk. Derudover er surveyeksperimentet også blevet distribueret via kontakttagen til folkeskoler og efterskoler i Danmark. Her gør det sig gældende, at studiet har kontaktet cirka 500 folkeskoler og 400 efterskolelærere fordelt på alle kommuner i Danmark. Det bør her bemærkes, at disse distributionsmetoder ikke kan siges at påvirke den interne validitet negativt grundet randomisering. Se distributionsteksten i bilag 5.

Gennem distributionen og den tilhørende kommunikation har der med henblik på en sikring af den interne validitet været en opmærksomhed omkring, at de potentielle respondenter ikke bliver bekendte med undersøgelsens eksperimentelle præmis (jf. Hawthorne-effekten). Ved at holde respondenterne uvidende om eksperimentet kan det dermed argumenteres, at studiet vil opnå mere umiddelbare og naturlige reaktioner, hvilket styrker den interne validitet (Aarhus Universitet, n.d.).

Med henblik på at maksimere antallet af respondenter i surveyen, har studiet holdt surveyen åben indtil den daglige frekvens af respondenter faldt til et niveau

på under én besvarelse om dagen. Dette resulterede i, at surveyen var mulig at tilgå i en periode på tre uger.

6.4.4 Surveyeksperimentets deltagere

Efter dataindsamlingen er der foretaget en sammenligning af respondenterne i eksperimentets stikprøve og populationen. Hertil har det ikke været muligt at få adgang til specifikke data på dansklærere i udskolingen, hvorfor dataudtræk fra Danmarks Lærerforening primært vil blive anvendt som det bedste mål og dermed indblik i studiets population. I den forbindelse bør det derfor nævnes, at Danmarks Lærerforenings medlemmer i 2022 udgjorde 94 % af lærerne i Danmark, som er repræsenteret på tværs af alle fag (jf. Roersen et al., 2022: 95). Nedenfor ses en sammenligningen af surveyeksperimentets stikprøve sammenholdt med medlemmerne af Danmarks Lærerforening:

Tabel 6.11: Sammenligning af Danmarks Lærerforening og studiets stikprøve

Kønsfordeling		
	Danmarks Lærerforening	**Stikprøven**
Kvinde	73 %	74 %
Mand	27 %	26 %
Aldersfordeling		
	Danmarks Lærerforening	**Stikprøven**
18-30	11 %	7 %
31-40	23 %	23 %
41-50	31 %	25 %
51-60	25 %	32 %
61+	10 %	13 %
Geografisk fordeling		
	Danmarks Lærerforening	**Stikprøven**
Region Hovedstaden	26 %	13 %
Region Sjælland	14 %	12 %
Region Syddanmark	21 %	17 %
Region Midtjylland	24 %	34 %
Region Nordjylland	12 %	23 %
Medlem af censorkorps		
	Censorkorps	**Stikprøven**
Medlem	Censorer: 295	22 %
Ikke medlem	Dansklærer i udskolingen: 6935	78 %
Medlemmer og respondenter i alt		
	Antal medlemmer i DLF	**Antal respondenter**
I alt	50.427	93

Kilde: DLF, 2023; Kilde DLF, 2024, Styrelsen for undervisning og kvalitet, 2024; Bilag 7; Bilag 8.

Gennem ovenstående sammenligning kan det ses, hvordan studiets stikprøve besidder nogenlunde samme fordeling som medlemmerne i Danmarks Lærerforening, når det kommer til både variablene 'køn' og 'alder'. Det bliver dog også tydeligt, at fordelingen af dansklærere i dette studie geografisk set ikke forekommer at være overensstemmende med fordelingen i Danmarks Lærerforening. Selvom det ved et eksperimentelt design ikke har betydning, hvorvidt fordelingen af deltagere er repræsentativ for den samlede population, havde det alligevel været en styrkelse af studiets eksterne validitet, hvis den geografiske fordeling af deltagere i studiet havde indikeret en repræsentativ fordeling.

Med afsæt i, at den ovenstående tabel kun indikerer en begrænset repræsentativitet af populationen, og at data om denne population er baseret på Danmarks Lærerforenings medlemmer, som dækker en begrænset del af populationen på tværs af alle faggrupper, kan resultaterne fra surveyeksperimentet derfor ikke nødvendigvis generaliseres til andre kontekster (jf. Blom-Hansen & Serritzlew, 2014). Denne begrænsede eksterne validitet vil blive taget i betragtning i studiets analyse og konklusion.

6.4.5 Randomiseringens vellykkethed

Det følgende afsnit vil have til formål at foretage en kvalitetsvurdering af studiets randomisering ved at vurdere og reflektere over eventuelle begrænsninger og fejlkilder, da dette er med til at muliggøre en diskussion af studiets validitet. I den forbindelse er det først og fremmest muligt at kigge på den indledende fordeling af respondenterne i de to grupper. Her gør det sig gældende, at 339 respondenter i alt har åbnet og påbegyndt surveyeksperimentet, hvor henholdsvis 51 % er endt i kontrolgruppen, hvorimod 49 % er endt i eksperimentgruppen (jf. bilag 9: 4). Med en forskel på 2 %-point kan det argumenteres, at respondenterne er blevet fordelt stort set ligeligt. Hermed tyder det på, at randomiseringen ved respondentens påbegyndelse af surveyeksperimentet har virket efter hensigten.

Af de 339 påbegyndte besvarelser er der samlet set 93 dansklærere, der i løbet af distributionsperioden har gennemført surveyeksperimentet (jf. afsnit 6.4.4). Her befinder 52 af respondenterne sig i kontrolgruppen, mens de resterende 41

respondenter udgør eksperimentgruppen. Dette er svarende til en fordeling på henholdsvis 56 % i kontrolgruppen og 44 % i eksperimentgruppen. Den større uligevægt i fordelingen af respondenterne skyldes hovedsageligt et større frafald blandt respondenterne i eksperimentgruppen. Her var 82 % (9 ud af 11) af dem, der faldt fra lige efter interventionen, i eksperimentgruppen. Respondenterne i eksperimentgruppen tyder dermed på i højere grad at reagere negativt og modvilligt, når de bliver præsenteret for interventionen omkring en LL-model som medbedømmer. En umiddelbar årsag til denne negative reaktion kunne være en AI-skepsis.

At der er nogle respondenter i eksperimentgruppen, som falder fra - muligvis på grund af en AI-skepsis - kan medføre, at der opstår en sampling bias. Her vil eksperimentgruppens gennemsnitlige tillid formentligt havde været lavere, end hvad der gør sig gældende for nærværende studie, hvis de mere AI-skeptiske respondenter ikke var frafaldet eksperimentet undervejs (jf. afsnit 7.4.1). Herved betragtes frafaldet i eksperimentgruppen dermed ikke som en svækkelse af studiets interne validitet, idet studiet blot udsættes for en hårdere test af interventionens betydning for tilliden til medbedømmeren.

Ud fra formålet om at vurdere studiets randomisering og dens indvirkning på validiteten, fremlægger følgende tabel lighederne og forskellene i kontrol- og eksperimentgruppen i henhold til studiets fem baggrundsvariable.

Tabel 6.12: Sammenligning af kontrol- og eksperimentgruppen

Kønsfordeling		
	Kontrolgruppe	**Eksperimentgruppe**
Kvinde	75 %	73 %
Mand	25 %	27 %
	Chi2 Statistik: 0.0, P-værdi: 1.0 (Bilag 10, s. 6)	
Aldersfordeling		
18-30	8 %	7 %
31-40	25 %	20 %
41-50	21 %	29 %
51-60	35 %	29 %
61+	12 %	15 %
	Cramérs V: 0.118 (bilag 10, s. 6-7)	
Geografisk fordeling		
Region Hovedstaden	8 %	20 %
Region Sjælland	8 %	17 %
Region Syddanmark	17 %	17 %
Region Midtjylland	38 %	29 %
Region Nordjylland	29 %	17 %
	Cramérs V: 0.256 (Bilag 10, s. 7)	
Erfaring		
0-3 år	12 %	5 %
4-10 år	23 %	29 %
11-20 år	17 %	17 %
20+ år	48 %	49 %
	Cramérs V: 0.128 (Bilag 10, s. 8)	
Medlem af censorkorps		
Censor	17 %	27 %
	Chi2 Statistik: 0.732, P-værdi: 0.3923 (Bilag 10, s. 8)	

Af tabellen kan det ses, hvordan der forekommer en lille forskel på 2 %-point mellem de to gruppers kønsfordeling, og at denne forskel gennem en Chi^2-test ikke er signifikant. Ved aldersfordelingen bliver det tydeligt, at der igen er tale om mindre forskelle mellem kontrol- og eksperimentgruppen. Den mindste forskel på 1 %-point findes ved aldersgruppen '18-30 år', mens den største forskel på 8 %-point omvendt ses ved aldersgruppen '41-50 år'. Med en Cramer's V-værdi på 0,12 kan dette betragtes som en meget lille forskel mellem grupperne (jf. Barts, 1999: 184). Hermed kan det også argumenteres, at studiets interne validitet ikke svækkes, idet den tilnærmelsesvist lige fordeling er med til at indikere, hvordan de fremlagte effekter i analysen, kan tilskrives eksperimentets uafhængige variabel og dermed ikke køn eller alder som en spuriøs faktor.

I henhold til de respondenternes geografiske bopæl fremgår det af tabellen, hvordan kontrol- og eksperimentgruppens geografiske forskelle er lidt større end, hvad der var tilfældet ved køn og alder. Her er 17 % af respondenterne i eksperimentgruppen bosiddende i Nordjylland, mens det samme gør sig gældende for 29 % af respondenterne i kontrolgruppen. Med en Cramer's V-værdi på 0,26 kan dette betragtes som en lille forskel (jf. Barts, 1999: 184). Dette anses dog ikke som værende problematisk, idet studiet ikke er bekendt med et teoretisk grundlag omkring, at geografiske forskelle har betydning for karaktergivning, hvorfor denne forskel ikke vurderes at ville svække studiets interne validitet.

Med hensyn til respondenternes erfaring som dansklærere fremlægges det, hvordan kontrol- og eksperimentgruppen fordeler sig næsten ligeligt ved erfaringskategorierne '11-20 år' og '+20 år'. Ved de to resterende erfaringskategorier er der dog en lille forskel, hvilket ses ved, at der er 7 %-point flere deltagere med 0-3 års erfaring i kontrolgruppen end eksperimentgruppen. 'Cramer's V'-testen viste med en værdi på 0,13, at der ikke forekom en bemærkelsesværdig forskel på grupperne (jf. Barts, 1999: 184), hvorfor denne fordeling ikke vil svække studiets interne validitet eller have en indvirkning på de effekter, der vil blive fremlagt i den kommende analyse.

Sidst er andelen af censorer, som er en del af censorkorpset ved skriftlig dansk FP9, blevet undersøgt i de to grupper (jf. afsnit 6.4.5). Her er der 10 %-point flere respondenter i eksperimentgruppen end kontrolgruppen, som er en del af det danske censorkorps. Dette er svarende til, at der i eksperimentgruppen er 11 beskikkede censorer, mens dette gør sig gældende for 9 respondenter i kontrolgruppen. Roersen et al. (2022: 135-136) fandt indikationer på, at censorerne er mere støjende end de resterende dansklærere. Denne fordeling kunne derfor potentielt have betydning for studiets uafhængige variabel i form af gruppernes støjniveau. Chi^2-testen med en p-værdi på 0,4 er dog med til at påpege, hvordan der ikke forekommer nogen signifikant forskel mellem grupperne, hvorved studiets interne validitet er sikret.

Af ovenstående kan det opsummeres, at studiets randomisering har været vellykket. Dette er også til trods for, at der blev fremlagt et større frafald blandt respondenterne i eksperimentgruppen, da denne sampling bias blot bidrager til, at eksperimentets effekter testes under strengere forudsætninger (jf. afsnit 6.4.5). Ydermere er forskellene mellem kontrol- og eksperimentgruppen ikke signifikante (jf. Chi^2-testene) og samtidig karakteriseret som værende ubetydelige (jf. Cramer's V'-test). Med afsæt heri kan det fremhæves, hvordan randomisering ikke har svækket gyldigheden af studiets resultater, men derimod skabt grobund for en høj intern validitet.

6.5. Analysestrategi

Med afsæt i ovenstående kortlægning af studiets undersøgelsesstrategi og undersøgelsesmetoder, vil studiets analysestrategi blive introduceret i det følgende afsnit. Gennem analysestrategien vil det blive kortlagt, hvordan studiets tre hypoteser vil blive undersøgt.

6.5.1 H1: Støjreduktion med LL-modellen som medbedømmer

Studiets første hypotese vil have til formål at undersøge effekten af at have en medbedømmer i karaktergivningen, hvor denne medbedømmer er en LL-model, hvilket leder frem til følgende hypotese:

H1: En LL-model som medbedømmer kan reducere støjniveauet i karaktergivningen i dansk skriftlig fremstilling

Studiets første hypotese vil hermed undersøge, hvorvidt brugen af en LL-model kan reducere støjen i karaktergivningen. I studiets teori blev det beskrevet, hvordan graden af støj kan måles gennem følgende formel:

$$Støj = \frac{\Sigma(y_i - p_i)^2}{n}$$

For at teste, hvorvidt der forekommer en signifikant støjreduktion ved brugen af LLM-vurderingen, bliver der anvendt en F-test og Levene's test, som benyttes til at teste forskelle i varians mellem grupperne (Zhou, Zhu & Wong, 2023; Iversen, 2014). Et opmærksomhedspunkt ved F-testen er, at den er sensitiv for outliers (Agresti, 2018: 375; Zhou, Zhu & Wong, 2023). Omvendt er Levene's test i lavere grad overfølsom overfor outliers i data, men i stedet udfordret, når det drejer sig om mindre datasæt (Zhou, Zhu & Wong, 2023). Af den grund finder studiet det fordelagtigt at benytte både F-testen og Levene's test, idet disse vil kunne komplementere hinanden.

I studiets teori blev det beskrevet, hvordan det forventes, at to bedømmere vil bidrage med støjreduktion. Herved vil der først blive undersøgt, hvorvidt der forekommer en effekt ved studiets longitudinale eksperiment. Dette sker ved at teste forskellen mellem lærernes første vurdering, og deres anden vurdering efter interventionen. Efter at have testet, hvorvidt LL-modellen samlet set bidrager med en støjreduktion, vil dette blive nuanceret ved at udføre den samme test for henholdsvis kontrol- og eksperimentgruppen. Selvom begge grupper i princippet har den samme medbedømmer i form af LL-modellen, er det i henhold til hypotesen relevant at afdække, hvorledes støjreduktionen afhænger af om respondenten tror, at medbedømmerens vurdering er udformet af en anden censor eller en AI-algoritme.

6.5.2 H2: LL-modellens støjniveau

Gennem hypotese 1, blev det undersøgt, hvorvidt LL-modellen bidrager til en støjreduktion, hvorigennem støjreduktionen på gruppens samlede variabilitet måles. Selvom den samlede variabilitet forventes mindsket (jf. hypotese 1), kan LL-modellen i sig selv være støjende. Som nævnt i studiets teori sker dette, hvis LL-modellens vurdering afviger mere fra gennemsnittet end den gennemsnitlige bedømmer gør (jf. afsnit 5.1.7). For at undersøge dette er studiets hypotese 2 opstillet:

H2: LL-modellen kan være mindre støjende end den gennemsnitlige dansk censor

For at teste denne hypotese vil LL-modellens vurdering blive sammenlignet med respondenternes første vurdering, hvor deres afvigelser fra gennemsnittet vil være omdrejningspunktet for denne sammenligning. I den forbindelse vil gennemsnittet for respondenternes første vurdering blive anvendt. Dette skyldes, at deltagernes første vurdering kan betragtes som værende den karakterfastsættelse, der afspejler karaktergivningen uden brugen af LL-modellens vurdering. Med henblik på at nuancere resultaterne vil en sammenligning af LL-modellen og deltagernes afvigelser også blive foretaget ved alle vurderingsområderne. LL-modellens støjniveau på de enkelte vurderingsområder kan derved give indsigt i, hvorvidt der er nogle områder, som LL-modellen har sværere ved at bedømme end andre. I tilfælde af at modellen afviger mere fra gennemsnittet end, hvad den gennemsnitlige deltager gør, må hypotesen forkastes.

6.5.3 H3: AI som medbedømmer og tilliden hertil påvirker brugen

Med henblik på at undersøge betydningen af tillid blev det i studiets problemanalyse tydeliggjort (jf. afsnit 4.4), hvordan det forventes, at en information om, at medbedømmeren er en AI-model, vil have betydning for den fagprofessionelles tillid til denne medbedømmer. I litteraturen var der dog ikke enighed om, hvorvidt denne information kunne forventes at have en positiv eller negativ effekt på tilliden. Foruden en forventet ændring i tilliden til medbedømmeren blev det samtidig også fremlagt, at denne tillid til

medbedømmeren kunne forventes at påvirke brugen af medbedømmeren. Her vil en højere tillid forventes at øge sandsynligheden for overbrug af modellen, mens en lavere tillid omvendt forventeligt vil øge sandsynligheden for underbrug (jf. afsnit 5.2). Ud fra disse antagelser blev følgende hypotese blive opstillet:

H3: Information om en AI-model som medbedømmer påvirker tilliden, der gennem brugen har betydning for støjniveauet

Med henblik på at teste ovenstående hypotese, blev det argumenteret, at denne skal undersøges gennem to delhypoteser. Disse blev fremlagt i studiets teoriafsnit (jf. afsnit 5.2.7), hvor den første delhypotese lød som følger:

H3.A: Informationen om en AI-model som medbedømmer har betydning for tilliden til medbedømmeren

Med henblik på at måle effekten af informationen om medbedømmeren på tilliden til denne medbedømmer, vil der indledningsvist blive undersøgt, hvorvidt de anvendte spørgsmål om tillid kan udgøre et samlet mål (jf. afsnit 6.5.3). For at ensliggøre tillidsspørgsmålenes skalaer, er der foretaget en standardisering af disse, således at spørgsmålene kan forenes i et samlet mål for tillid.

Efter at have sikret, at der ikke er multikollineariteten mellem tillidsspørgsmålene, er der foretaget en PCA, hvorved loadingværdierne giver indsigt i, hvorvidt de enkelte spørgsmål bidrager til det samlede mål (jf. Guadagnoli & Velicer, 1988: 266). Disse loadingværdier viser, at alle spørgsmålene bidrager til det samlede mål (jf. bilag 11). En KMO-test er herefter udført for at vurdere, hvorvidt andelen af varians blandt alle de observerede spørgsmål om tillid besidder en fælles varians. Herved forekommer en høj KMO-værdi på 0,8 (over 0,6 og tættere på 1), som en indikation på, at der er en betydelig andel af variansen, der kan forklares af underliggende faktorer (datacamp.com, 2019; bilag 11). Dette leder videre til at afdække, hvorvidt der er tale om et samlet mål for det latente begreb 'tillid' ved en faktoranalyse. Til forskel fra PCA'en bliver det gennem faktoranalysen undersøgt, hvorvidt der forekommer underliggende faktorer i det samlede mål (jf. Tavakol & Dennick, 2011: 54; Tavakol & Wetzel, 2020). Faktoranalysen viser her, at tillidsspørgsmålet om, hvorvidt begrundelsen og karaktererne var fair, ikke i

samme grad som de andre tillidsspørgsmål, passede ind i det samlede mål (jf. spørgsmål 3 på tabel 6.8). Afslutningsvis er 'Cronbachs Alpha'-testen (CBA) blevet brugt til at teste, hvorvidt spørgsmålene hænger sammen internt (Tavakol & Dennick, 2011: 53). Her blev det fundet, at der i det samlede mål forekom en CBA-værdi på 0,8. I forlængelse af disse reliabilitetestest er det hermed konkluderet, at alle spørgsmålene om tillid kan indgå i det samlede mål. Dog er der en bevidsthed om, at spørgsmålet om, hvorvidt begrundelsen og karaktererne var fair, udgør den svageste del i dette samlede mål (se reliabilitetstestene i bilag 11). Med afsæt i disse tests kan det dog generelt argumenteres, at studiets interne reliabilitet ved målet for tillid er høj.

Efter at have foretaget en reliabilitetstest for det samlede mål for tillid, vil H3.A omhandlende forskellen på kontrol- og eksperimentgruppens gennemsnitlige tillid blive testet. Denne test mellem gruppernes tillid vil både blive vist ved studiets samlede mål for tillid samt hvert enkelt spørgsmål om tillid. Forskellen mellem denne tillid vil blive testet gennem en t-test. I tilfælde af at der med 95 % sikkerhed forekommer en ændret tillid til medbedømmeren, kan H3.A ikke forkastes.

Efter at have undersøgt den første delhypotese, vil tillidens betydning for graden af underbrug og overbrug blive undersøgt gennem den anden delhypotese:

H3.B: Tilliden til medbedømmeren har betydning for brugen af medbedømmeren

I undersøgelsen af tillidens betydning for 'brugen' af medbedømmeren er målingen af 'brug' opdelt i målinger af henholdsvis 'underbrug' og 'overbrug'. Som nævnt i studiets teoriafsnit vil brugen blive klassificeret som underbrug i tilfælde af, at bedømmeren kunne have reduceret støjniveauet ved at have brugt LL-modellens karaktergivning i den endelige vurdering af elevteksten. Videre bliver variablen for overbrug oprettet ud fra en klassificering af de deltagere, som er blevet mere støjende ved brug af LL-modellen. For at danne et overblik over sammenhængen mellem tillid og brug, er der ved både under- og overbrugen lavet en deskriptiv kvartilinddeling fordelt på tilliden. Disse deskriptive resultater viser andelen af under- og overbrug ved henholdsvis 1) nedre kvartil af tillid, 2)

interkvartilet (IQR) og 3) øvre kvartil af tillid. Herved vil det med udgangspunkt i disse kvartiler blive fremlagt indikationer på, hvordan tillidsniveauet har betydning for brugen.

Med et videre ønske om at teste sammenhængen mellem tillid og brug, vil der blive foretaget logistiske regressioner for både under- og overbrug. Gennem de logistiske regressioner vil det ud fra forklaringsgraden blive klarlagt, hvorvidt tillid kan forklare variansen af brugen.

Analysen vil være opdelt således, at det først bliver undersøgt, hvorvidt tilliden har betydning for brugen ved karaktergivningens vurderingsområder. Herefter vil samme analyse blive foretaget ved den samlede karakter. Begge analysedele vil både anvende de ovennævnte deskriptive fordelinger samt logistiske tests for sammenhænge. For at validere sammenhængen mellem tilliden og brugen er der afslutningsvist foretaget en multipel logistisk regression, hvor det undersøges, hvorvidt dele af brugen kan forklares gennem interventionen i studiet.

6.6 Opsamling på studiets kvalitet

Gennem ovenstående kapitel er de forskellige metodiske valg løbende blevet reflekteret i henhold til deres indvirkning på studiets kvalitet, hvilket har betydning for studiets resultater og konklusionernes gyldighed, pålidelighed og gentagelighed. Med afsæt i disse overvejelser kan det opsummerende argumenteres, at nærværende studie kan karakteriseres ved følgende validitet og reliabilitet:

Figur 6.13: Studiets kvalitet

Validitet	
Intern validitet	Høj
Ekstern validitet	Lav
Økologisk validitet	Høj
Reliabilitet	
Intern reliabilitet	Høj
Ekstern reliabilitet	Middel-høj

Efter ovenstående indføring i studiets metodologi og den valgte undersøgelses- og analysestrategis indvirkning på studiets validitet og reliabilitet, leder det videre til studiets næste kapitel, som er analysen. Her vil studiets indsamlede data blive fremlagt, analyseret og fortolket.

7. Analyse

I det følgende kapitel vil analysen for dette studie blive præsenteret. I den forbindelse gør det sig gældende, at analysen vil være opdelt i henhold til studiets tre hovedhypoteser, som blev fremsat på baggrund af studiets teoretiske grundlag (jf. afsnit 5; Parasuraman & Riley, 1997; Kahneman et al., 2021). En præmis for at undersøge studiets hypoteser er, at der forekommer støj i karaktergivningen ved studiets surveyeksperiment. Af den grund vil der i det første analyseafsnit indledningsvist blive gjort rede for de karakterer, som surveyeksperimentets deltagere har givet ved deres første vurdering af danskopgaven. Dernæst vil studiets tre hypoteser blive undersøgt i de efterfølgende analysedele. Dette vil blive gjort gennem følgende kronologi:

Tabel 7.1: Kronologi for studiets analyse

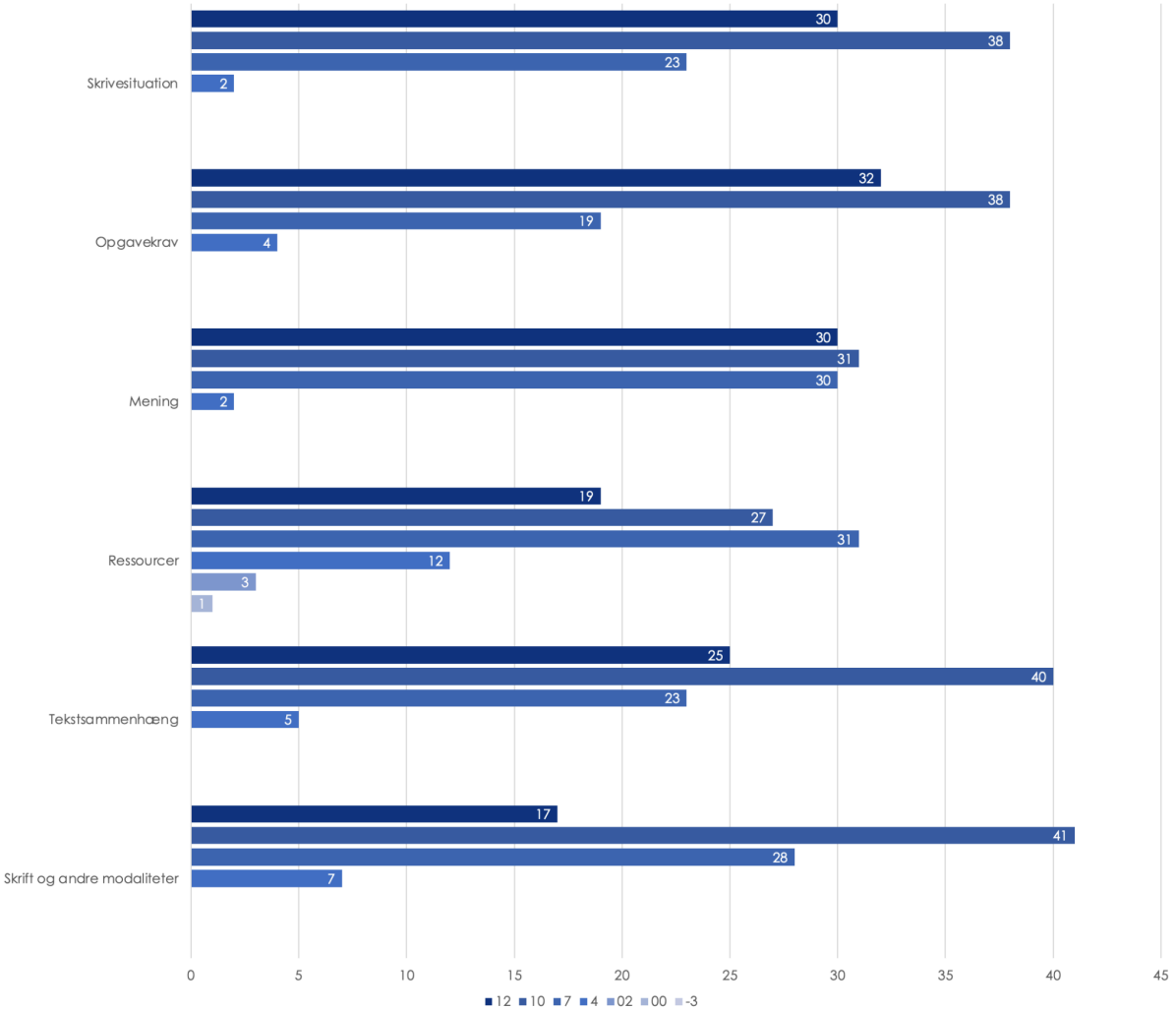
Analysens fire analysedele	
Afsnit 7.1	Præsentation af karaktererne givet i respondenternes første vurdering og dermed en indsigt i surveyeksperimentets støjresultater
Afsnit 7.2	Hypotese 1: En LL-model som medbedømmer kan reducere støjniveauet i karaktergivningen i dansk skriftlig fremstilling
Afsnit 7.3	Hypotese 2: LL-modellen kan være mindre støjende end den gennemsnitlige dansk censor
Afsnit 7.4	Hypotese 3 - herunder H3.A og H3.B: Informationen om en AI-model som medbedømmer påvirker tilliden, der gennem brugen har betydning for støjniveauet

7.1 Støjen i den første vurdering

Gennem problemanalysen blev det tydeliggjort, hvordan karaktergivningen ved folkeskolens afgangseksamen særligt i dansk skriftlig fremstilling er præget af støj. Dette fremgik både af Dolin et al. (2018), Troelsen (2020) og Roersen et al. (2022) undersøgelser, som på bekymrende vis belyste, hvordan karaktergivningen i folkeskolen er præget af en manglende ensartethed og pålidelighed (jf. afsnit 4.2.2). Af Børne- & Undervisningsministeriets prøvevejledning til folkeskolernes afgangseksamen i dansk skriftlig fremstilling fremgår det, hvordan der foreligger et krav om, at bedømmeren i forbindelse med karaktergivningen skal udfylde et vurderingsskema (Børne- & Undervisningsministeriet, 2022: 21-22) Vurderingsskemaet er udviklet af Børne- & Undervisningsministeriet og

indeholder tre vurderingsdimensioner herunder funktion, indhold og form, som hver består af to vurderingsområder. Disse seks vurderingsområder skal indgå som en samlet helhed i bedømmelsen af elevens besvarelse og skal som udgangspunkt vægtes ligeligt i den endelige karakterfastsættelse. I tilfælde af tvivl skal funktionsdimensionen dog vægtes lidt højere (Børne- & Undervisningsministeriet, 2022: 21-22). På baggrund heraf blev respondenterne i surveyeksperimentet bedt om at foretage en vurdering med udgangspunkt i Børne- & Undervisningsministeriets vurderingsskema. Fordelingen af karaktererne for denne analytiske vurdering ses nedenfor:

Figur 7.2: Hyppigheden af karakterer fordelt på vurderingsområder



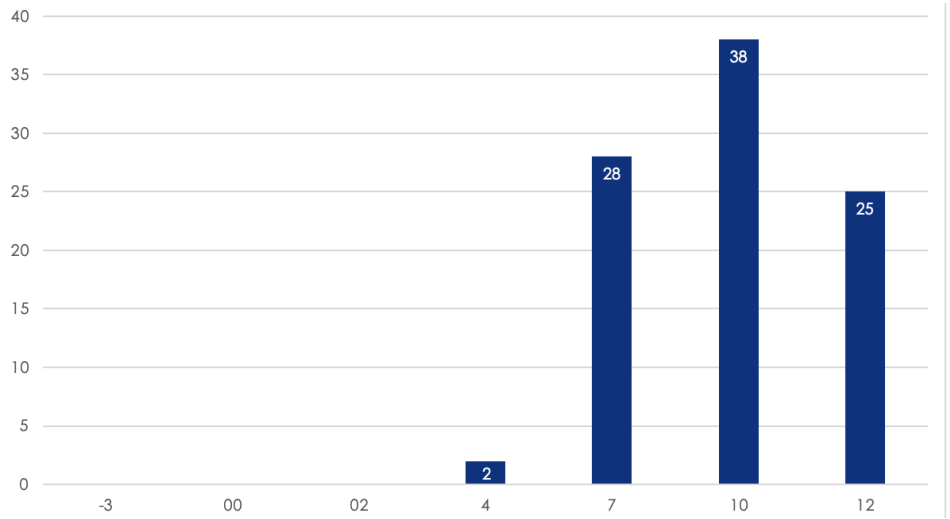
Kilde: bilag 10, s. 4-5

På ovenstående figur fremgår det, hvordan karaktergivningen ved de seks forskellige vurderingsområder udmærker sig med varierende karakterer. Her gør

det sig gældende, at der ved vurderingsområderne 'skrivesituation', 'opgavekrav', 'mening', 'tekstsammenhæng' samt 'skrift og andre modaliteter' er givet karakterer i spændet fra 4 til 12. Det sidste vurderingsområde angående 'ressourcer' bemærker sig ved, at der forekommer en større variation i karaktergivningen sammenlignet med de andre områder. Dette skyldes, at når det kommer til vurderingen af, hvorvidt elevteksten anvender relevante ressourcer såsom opgaveforlægget, faglig viden og internettet, så bliver elevteksten vurderet til alle karakterer fra 00 til 12. Hermed kan det tydeliggøres, hvordan der inden for alle seks vurderingsområder foreligger en variation i karaktergivningen, hvor denne spredning er størst ved vurderingsområdet angående 'ressourcer'. På tværs af de seks vurderingsområder kan det endvidere fremhæves, hvordan hyppigheden forekommer størst ved karakteren 10. Dette kommer til udtryk ved, at karakteren 10 er den karakter, som er givet flest gange ved alle vurderingsområderne pånær 'ressourcer', hvor elevteksten er blevet vurderet til karakteren 7 flest gange.

Efter at respondenterne havde vurderet elevteksten med udgangspunkt i de seks vurderingsområder, blev respondenterne bedt om at foretage en samlet vurdering på baggrund heraf. Ved vurderingsområderne skal der gives delkarakterer og dermed en analytisk vurdering, som danner grundlag for den endelige holistiske vurdering. Fordelingen af de samlede karakterer fremgår nedenfor, hvor fordelingen er fremlagt som hyppighed:

Figur 7.3: Hyppighed af karakterer ved respondenternes første vurdering



Kilde: bilag 10: 4

Af ovenstående figur ses fordelingen af karakterer ved respondenternes samlede vurdering. Her fremgår det, hvordan der i karaktergivningen forekommer variation, idet de samlede vurderinger af elevteksten befinder sig i karakterspændet fra karakteren 4 til karakteren 12. Det foreligger dermed bemærkelsesværdigt, hvordan respondenterne har forskellige opfattelser af elevteksten og dens præstationsniveau, hvorfor den er blevet vurderet til forskellige karakterer. Her gør det sig gældende, at 2 respondenter giver eleven karakteren 4, hvormed de vurderer elevteksten til at være en jævn præstation, der demonstrerer en mindre grad af opfyldelse af danskfagets mål med adskillige væsentlige mangler. Derimod er der 25 respondenter, som giver karakteren 12 og dermed vurderer elevteksten til at være en fremragende præstation, som demonstrerer en udtømmende opfyldelse af danskfagets mål med ingen eller få uvæsentlige mangler. For 38 og herved størstedelen af respondenterne gør det sig dog gældende, at de vurderer elevteksten som værende en fortrinlig præstation, der demonstrerer en omfattende opfyldelse af danskfagets mål med nogle mindre væsentlige mangler, hvorfor de bedømmer opgaven til karakteren 10.

Hermed kan det opsummeres, hvordan respondenterne har forskellige opfattelser af elevtekstens niveau. Disse forskellige opfattelser udmønter sig i en karaktergivning præget af varierende karakterer, der ifølge Kahneman et al. (2021: 47) kan karakteriseres som støj. At det kan karakteriseres som værende støj sker ud fra en forståelse af, at karaktergivningen udmærker sig med en uønsket variabilitet. Det tyder herved på, at det danske uddannelsessystem er præget af systemstøj, når det kommer til karaktergivningen i dansk skriftlig fremstilling ved folkeskolens afgangseksamen (jf. Kahneman et al., 2021: 83). Denne uønskede variabilitet kan endvidere argumenteres at bidrage til en usikkerhed omkring karaktergivningen, som kan have store konsekvenser både på individuelt og samfundsmæssigt plan. Det skyldes, at der kan drages tvivl ved karaktergivningens pålidelighed, hvis denne i højere grad er præget af tilfældighed frem for ensartethed (jf. afsnit 4.2).

Foruden variabiliteten i delkaraktererne ved vurderingsområderne og de samlede karakterer er med til at belyse, hvordan respondenterne har divergerende opfattelser af elevtekstens præstationsniveau, så kommer det også til udtryk

gennem de begrundelser, som respondenterne har angivet i forlængelse af deres samlede vurdering. I nedenstående figur er et udsnit af respondenternes begrundelser præsenteret ud fra formålet om at illustrere den forskel, der forekommer i disse. De fire udvalgte begrundelser lyder således:

Figur 7.4: Begrundelse for første vurdering

Kilde: bilag 15

I samspil med de endelige karakterer er de fire ovenstående begrundelser yderligere med til at tydeliggøre, hvordan respondenterne har forskellige opfattelser af den elevtekst, som de blev præsenteret for i eksperimentet. Dette eksemplificeres ved, at respondent A oplever opgavebesvarelsen som værende velskrevet, velbegrundet og på højt niveau, hvorimod respondent D mener, at eleven har misforstået dele af opgaven, har mangelfuld tegnsætning og et rodet sprog. Disse forskellige opfattelser resulterer som fremlagt i to divergerende karakterer, hvor respondent A samlet set vurderer elevteksten til karakteren 12 og dermed en fremragende præstation, mens respondent D giver karakteren 4, hvormed respondent D samlet set vurderer elevteksten til at være en jævn præstation.

På tværs af de fire begrundelser kan det udledes, hvordan respondenterne særligt opfatter det sproglige niveau i elevbesvarelsen forskelligt, hvilket ifølge bedømmerne har været udslagsgivende i fastsættelsen af den samlede karakter.

At der blandt respondenterne foreligger en uoverensstemmelse omkring elevtekstens sproglige niveau, kan være et udtryk for, at der forekommer niveaustøj, som i sidste ende udmønter sig i en støjende karaktergivning. Dette skyldes, at niveaustøj ifølge Kahneman et al. (2021: 87) opstår, når en bedømmer har tendens til at være hårdere i sine vurderinger end andre bedømmere. I henhold til de fire begrundelser ovenfor kan det argumenteres, hvordan en elev potentielt vil have nemmere ved at opnå en høj karakter ved respondent A, idet respondent A tyder på at være mindre hård i sine bedømmelser sammenlignet med de tre andre respondenter.

Ovenstående resultater klarlægger på overbevisende vis, hvordan der i studiets surveyeksperiment er tale om en uønsket variabilitet på tværs af vurderingerne, hvormed der jævnfør Kahneman et al. (2020) forekommer støj i karaktergivningen. Karaktererne og de tilhørende begrundelser er med til at synliggøre, hvordan der blandt dansklærerne findes divergerende forestillinger om, hvad der karakteriseres som en god besvarelse. Endvidere kan det også med afsæt i ovenstående resultater udledes, hvordan dette studie i tråd med Dolin et al. (2018), Troelsen (2020) og Roersen et al. (2022) finder, at der forekommer støj i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen.

Gennem problemanalysen (jf. afsnit 4) blev det fremlagt, hvordan en støjende karaktergivning kan anses som en udfordring først og fremmest på individuelt plan. Dette skyldes, at karakterer i dag har en afgørende betydning for elevernes videre muligheder i uddannelsessystemet, da det kræver et karaktergennemsnit på 4 eller 5 for at blive optaget på henholdsvis HF eller en gymnasial uddannelse. Samtidig kræver det, at eleven minimum har fået karakteren 02 i fagene dansk og matematik for at blive optaget på en erhvervsuddannelse (jf. afsnit 4.1.1). Udover en støjende karaktergivning kan ses som en udfordring for individet, kan det også argumenteres, hvordan det kan udgøre en udfordring på samfundsmæssigt niveau. Dette sker ud fra en forståelse af, at en støjende karaktergivning kan være med til at underminere troværdigheden af det danske uddannelsessystem som helhed, hvor både elever, forældre og andre interessenter kan miste tilliden til karaktergivningen og tvivle på værdien heraf, hvis dette ikke kan anvendes som et pålideligt mål for elevens præstation (jf. afsnit 4.2). Af den grund bliver det hermed gennem studiets eksperiment relevant at undersøge, hvorvidt det er

muligt at reducere støjen med en LL-model som medbedømmer, og dermed skabe en mere ensartet og pålidelig karaktergivning.

7.2 H1: Støjreduktionen med en medbedømmer

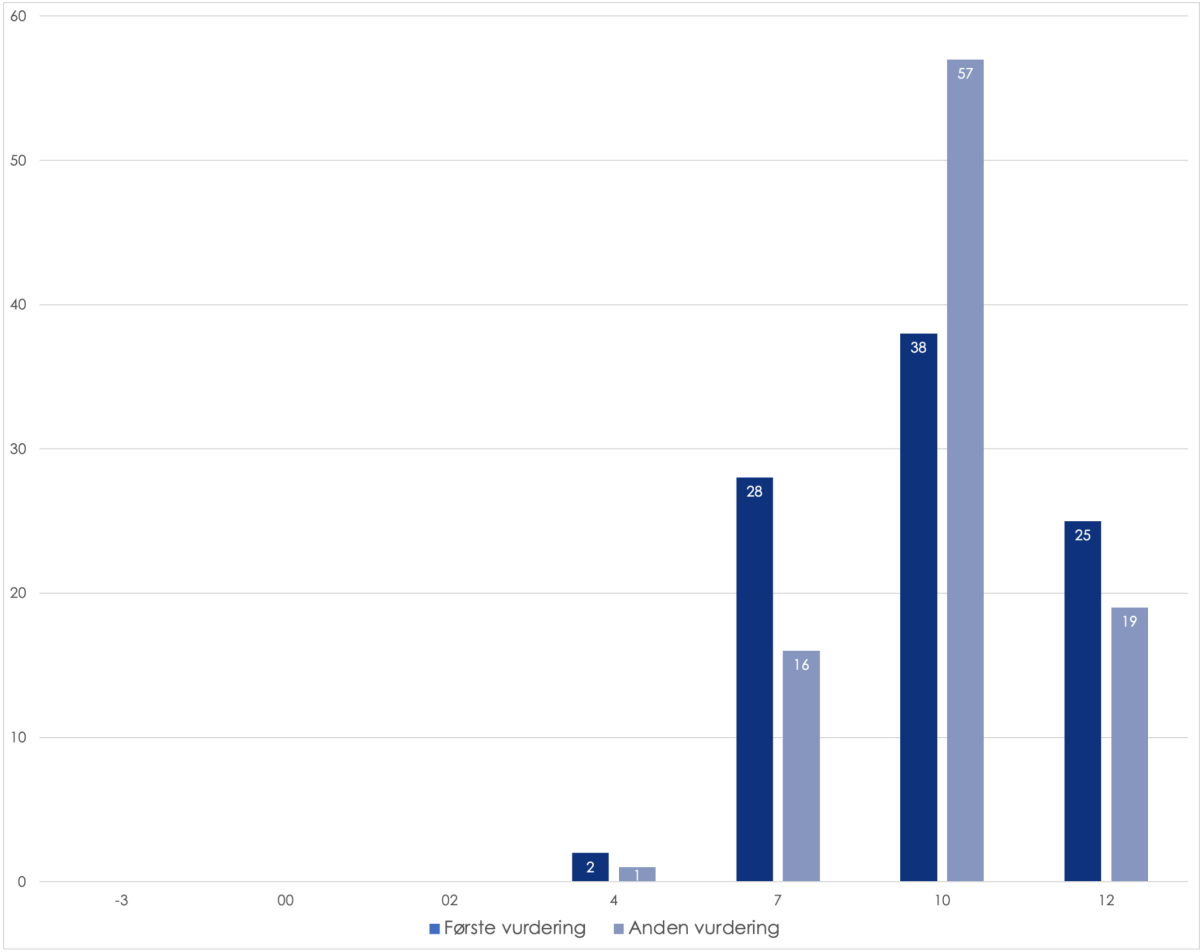
I forrige afsnit blev det kortlagt, hvordan dette studie i overensstemmelse med tidligere undersøgelser finder en variabilitet og dermed støj i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen. Dette studie vil gennem den første hypotese have til formål at undersøge, i hvilken grad en LL-model som medbedømmer kan være med til at reducere støjen i karaktergivningen:

H1: En LL-model som medbedømmer kan reducere støjniveauet i karaktergivningen i dansk skriftlig fremstilling

Af studiets første hypotese kan det udledes, hvordan denne har til formål at undersøge effekten af at have en medbedømmer ved karaktergivningen i skriftlig dansk i folkeskolen. Fælles for alle respondenterne i eksperimentet er, at deres medbedømmer er en LL-model, hvorved studiets første hypotese vil blive anvendt til at undersøge, hvorvidt brugen af en LL-model kan reducere støjen i karaktergivningen. Jævnfør Kahneman et al. (2021) kan støj inden for statistikken ligestilles med varians, hvorfor studiet har foretaget en række statistiske analyser herunder F-test og Levene's test med formålet om at afdække studiets første hypotese.

I analysens første afsnit blev det klarlagt, hvordan studiet i tråd med Dolin et al. (2018), Troelsen (2020) og Roersen et al. (2022) finder, at der er støj i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen. Med et ønske om at reducere støjen og dermed styrke karaktergivningen blev respondenterne efter, at de havde foretaget deres første vurdering præsenteret for en medbedømmer, hvorefter de skulle foretage deres anden og endelige vurdering. Resultaterne for de to vurderinger er visualiseret nedenfor:

Tabel 7.5: Sammenligning af karaktergivningen ved første og anden vurdering



Kilde: bilag 10, s. 4

Ovenfor ses de karakterer, som respondenterne har vurderet elevteksten til ved henholdsvis deres første og anden vurdering. Forskellen mellem de to vurderinger er, at respondenterne mellem de to vurderinger er blevet præsenteret for en medbedømmer og dens vurdering af teksten. Denne medbedømmer er som tidligere nævnt en LL-model, som på baggrund af vurderingsskemaet fra Børne- og Undervisningsministeriet samlet set havde vurderet elevteksten til karakteren 10. Respondenternes anden vurdering er hermed foretaget efter interventionen, hvormed denne udgør den endelige vurdering af elevteksten.

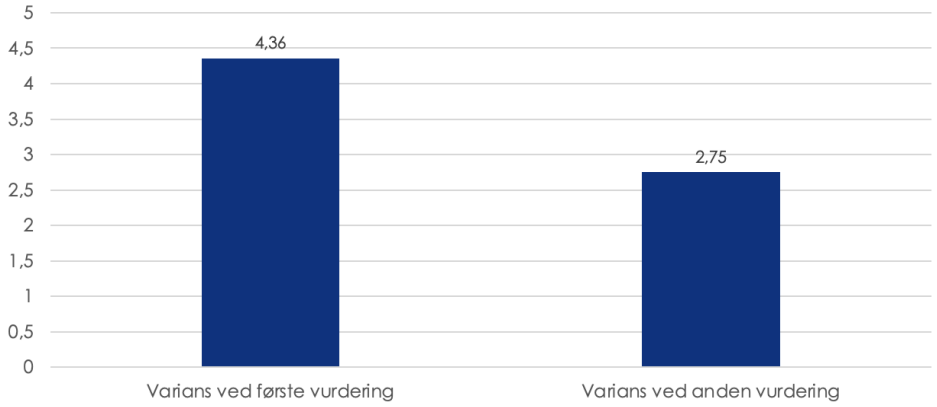
Af ovenstående sammenligning kan det fremhæves, hvordan der blandt respondenterne sker en ændring i karakterfastsættelsen fra første til anden vurdering. Dette kommer særligt til udtryk ved, at 57 respondenter ender med at give karakteren 10 ved den anden vurdering, hvilket er højere end den første

vurdering, hvor 38 respondenter havde givet karakteren 10. At en stor del af respondenterne ændrer deres endelige karakter i kølvandet på interventionen fremgår ligeledes ved, at antallet af respondenter, som tildeler elevteksten karakteren 7, reduceres med 12, mens der ved karakteren 12 er en reduktion på 6 respondenter.

På tværs af kontrol- og eksperimentgruppen gør det sig gældende, at 24 % af respondenterne bliver mindre støjende fra første til anden vurdering. Med andre ord reducerer knap en fjerdedel af alle dansklærerne deres støjniveau ved at have en LL-model som medbedømmer, hvilket kan anses som bemærkelsesværdigt. I den forbindelse er det også blevet testet, hvorvidt der er nogle af respondenterne, som er blevet mere støjende fra første til anden vurdering, hvilket ikke har været tilfældet. Med afsæt i Parasuraman & Riley (1997) kan det dermed siges, at der ikke har været et overbrug af modellen ved den samlede karakter, som har ført til yderligere støj (jf. bilag 10: 11-24). Det tyder i stedet på, at der er sket det, som Kahneman et al. (2021: 101-104) og Galton (1886) omtaler som 'regression mod gennemsnittet', idet flere af respondenterne samler sig om karakteren 10. Hermed kan det argumenteres, at eksperimentet er med til at synliggøre, hvordan en karaktergivning med to uafhængige bedømmelser, herunder en menneskelig og algoritmisk, kan være fordelagtig, når det kommer til at støjreducere og dermed styrke kvaliteten af karaktergivningen i folkeskolen.

At en bedømmelsespraksis med to bedømmere er fordelagtig for reduktionen af støj ved dansk skriftlig fremstilling i folkeskolen kommer også til udtryk, når der kigges nærmere på variansen ved respondenternes første og anden vurdering:

Tabel 7.6: Støjniveauet ved første og anden vurdering for hele stikprøven



F-test: 1,584409, P- Værdi: 0,0142; Levene's test: 8,017693, P-værdi: 0,005

Kilde: bilag 10, s. 12-14

Af tabellen fremgår det, hvordan variansen ved respondenternes første vurdering er 4,36, mens variansen er 2,75 ved respondenternes anden vurdering. Hermed kan det tydeliggøres, hvordan variansen reduceres med 1,61, når respondenterne udsættes for interventionen og introduceres til en medbedømmers vurdering. Denne forskel på variansen mellem de to vurderinger er svarende til en reduktion på 37 %, hvilket er med til underbygge Kahneman et al.'s (2021) antagelse om, at en bedømmelsespraksis med to bedømmere er fordelagtig, når det kommer til støjreduktion. Årsagen hertil er, at det gennem Kahneman et al. (2021) kan argumenteres, at roden til megen af den påviste støj ved karaktergivningen i folkeskolen kan skyldes det faktum, at karakterfastsættelsen udelukkende er baseret på én enkelt bedømmers vurdering. Dette sker ud fra en forståelse, at en bedømmelsespraksis med to bedømmere jævnfør Kahneman et al. (2021) har bedre forudsætninger for at reducere den uønskede variabilitet, idet to bedømmere vil kunne sammenligne, diskutere og afveje hinandens vurderinger (jf. afsnit 5.1.7).

Med afsæt i en p-værdi på 0,0142 ved F-testen og 0,005 ved Levene's test kan det fremhæves, hvordan forskellen i støjniveauet ved respondenternes første vurdering og anden vurdering foreligger højsignifikant på et 99 % konfidensniveau. Signifikanserne er med til at understrege, hvordan det medfører en bemærkelsesværdig forskel i støjen ved karaktergivningen, når en dansklærer tildeles en LL-model som medbedømmer. Endvidere er signifikansen for både F-testen og Levene's test også med at bekræfte studiets første hypotese om, at en

LL-model kan reducere støjniveauet i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen.

Med en viden om, at en LLM på højsignifikant vis kan reducere støjen i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen, vil det yderligere blive undersøgt, hvorvidt det samme gør sig gældende, når det kommer til delkarakterne ved de seks vurderingsområder fra Børne- og Undervisningsministeriets vurderingsskema. I nedenstående tabel vil støjreduktionen ved vurderingsområderne blive fremlagt:

Tabel 7.7: Støjniveauet ved første og anden vurdering fordelt på delkarakterer

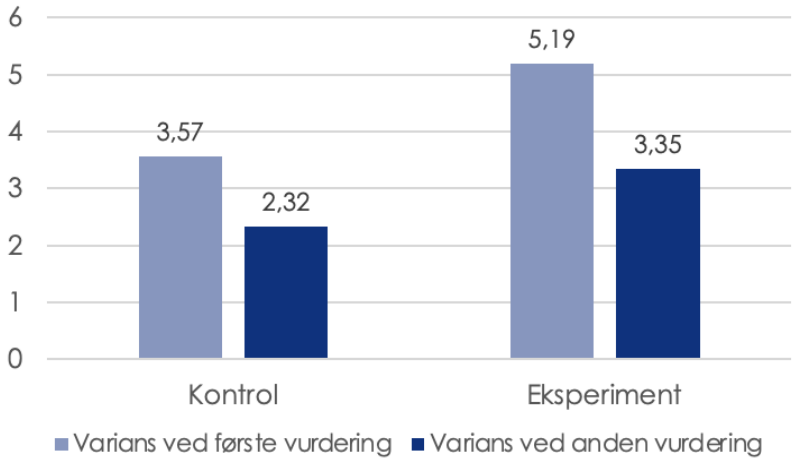
Vurderings-dimensioner	Vurderings-områder	Variansen ved første og anden vurdering	P-værdier for test
Funktion	Skrivesituation	Første: 4,29 Anden: 2,47	F-test: 0,004 Levene's: 0,006
	Opgavekrav	Første: 4,78 Anden: 3,37	F-test: 0,0472 Levene's: 0,1418
Indhold	Mening	Første: 4,82 Anden: 3,44	F-test: 0,0539 Levene's: 0,0566
	Ressourcer	Første: 8,70 Anden: 5,52	F-test: 0,0151 Levene's: 0,004
Form	Tekstsammenhæng	Første: 5,01 Anden: 3,27	F-test: 0,0211 Levene's: 0,0683
	Skrift og andre modaliteter	Første: 5,23 Anden: 4,15	F-test: 0,1351 Levene's: 0,800

Kilde: bilag 10: 16-20

På baggrund af resultaterne i ovenstående tabel bliver det igen tydeligt, at der forekommer dét, som Kahneman et al. (2020) italesætter som reduktion af støj. Dette fremgår ved, at der på tværs af alle seks vurderingsområder er sket en reduktion i variansen fra første til anden vurdering. I den forbindelse gør det sig gældende, at der ved fem ud af de seks vurderingsområder er tale om en signifikant forskel på støjniveauet i karaktergivningen efter, at respondenterne bliver præsenteret for deres medbedømmer i form af LL-modellens vurdering. For de fem vurderingområder er der fundet en procentvis støjreduktion på 29 % til 42 %, hvor 'skrivesituation' er det område, som udmærker sig med den største reduktion. Støjreduktionen ved vurderingsområdet 'skrift og andre modaliteter' forekommer at være den eneste reduktion, som gennem Levene's test og F-testen ikke er reduceret signifikant. På trods af, at denne ikke er signifikant, er der stadig tale om en reduktion i variansen på 22 %.

Nu hvor det er blevet afdækket, at der forekommer en signifikant støjreduktion ved både den samlede karaktergivning samt ved fem ud af seks vurderingsområder, leder det videre til, at det afslutningsvist vil blive undersøgt, hvorvidt der forekommer en forskel i støjniveauet ved første og anden vurdering på tværs af kontrol- og eksperimentgruppen. Forskellen mellem de to grupper er som tidligere nævnt, at kontrolgruppen bliver informeret om, at medbedømmeren er en anden censor, mens eksperimentgruppen informeres om, at medbedømmeren er en AI-algoritme. Sandheden er dog, at begge grupper bliver præsenteret for den samme vurdering, der er genereret af en LL-model, hvormed interventionen ligger i informationen om, hvorvidt det er en menneskelig eller algoritmisk medbedømmer (jf. afsnit 6.3.6). Forskellen i kontrol- og eksperimentgruppens varianser ved første og anden vurdering tydeliggøres i følgende tabel:

Figur 7.8: Støjniveauet ved første og anden vurdering fordelt på grupper



Kontrolgruppen:
F-test: 1,5400, P-værdi: 0,0632
Levene's test: 5,5699, P-værdi: 0,0202

Eksperimentgruppen:
F-test: 1,5505, P-værdi: 0,0849
Levene's test: 2,6586, P-værdi: 0,1069

Kilde: bilag 10, s. 14-16

Resultaterne fra ovenstående figur er med til at belyse, hvordan både kontrol- og eksperimentgruppen oplever en støjreduktion ved respondenternes anden vurdering. Dette ses ved, at kontrol- og eksperimentgruppens varians reduceres med henholdsvis 1,25 og 1,84, hvilket for begge grupper er svarende til en støjreduktion på 35 %. Denne støjreduktion ved både kontrol- og eksperimentgruppen er med til at understøtte Kahneman et al.'s (2021) antagelse om, at to bedømmere støjer mindre end én bedømmer i vurderingssituationer såsom karaktergivningen i folkeskolen (jf. 5.1.7). At både kontrol- og eksperimentgruppen ses at have en støjreduktion på 35 % er yderligere med til at indikere, hvordan informationen om, at medbedømmeren er en AI-model, ikke har den store betydning i forhold til støjreduktion.

Af tabellen fremgår det, hvordan der er en relativt stor forskel i variansen ved kontrol- og eksperimentgruppens første vurdering (jf. tabel 7.8). Dette skyldes blandt andet, at de to deltagere, som har givet karakteren 4, befinder sig i eksperimentgruppen. At to karakterer ud af 41 i eksperimentgruppen kan påvirke støjniveauet så meget, skyldes, at støjen beregnes efter den kvadrerede værdi af afvigelserne, hvorved de mest støjende deltagere bliver 'straffet' særligt hårdt (jf. afsnit 5.15). Deltagere, der har givet karakteren 4, får herved et støjniveau på 36 (6*6), mens deltagere, der har givet 12, kun har et støjniveau på 4 (2*2). Forskellen

mellem de to gruppers varianser ved første vurdering forventes derfor at kunne tilskrives stikprøvefordelingens tilfældigheder.

Når det kommer til p-værdierne fra F-testen og Levene's testen, gør det sig gældende, at disse foreligger signifikante på et 90 % konfidensniveau. I henhold til Levene's test kan det dog tydeliggøres, hvordan denne foreligger signifikant på et 95 % konfidensniveau ved kontrolgruppens støjreduktion. At der forekommer denne forskel i konfidensniveauet og dermed signifikansen mellem grupperne kunne skyldes, at der i eksperimentgruppen er 11 færre respondenter end i kontrolgruppen. Dette er særligt en ulempe for Levene's testen, da denne kan være sensitiv over for mindre grupper (jf. afsnit 6.5.1). Det kan dermed tydeliggøres, hvordan der ikke forekommer den samme sikkerhed som ved den forrige analyse, hvor støjreduktionen på tværs af hele stikprøven blev undersøgt. Inden for samfundsvidenskaberne er det dog ikke atypisk at arbejde med konfidensniveauer på 90 %. Med afsæt heri vil studiet derfor fortsat argumentere for, at en LL-model som medbedømmer kan bidrage til støjreduktion i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen, hvormed den første hypotese kan bekræftes.

Opsummerende kan det fremhæves, hvordan en LL-model som medbedømmer har bidraget til en signifikant reducerende effekt på støjniveauet ved karaktergivningen i dansk skriftlig fremstilling i folkeskolen, idet respondenternes anden vurdering er mindre støjende sammenlignet med den første vurdering (jf. tabel 7.8). At respondenternes anden vurdering forekommer mindre støjende, er med til at underbygge Kahneman et al's (2021: 101-102) teoretiske antagelse om, at en karaktergivning baseret på to uafhængige bedømmelser vil reducere støjen (jf. afsnit 5.1.7). Disse teoretiske antagelser stammer fra Galtons (1886) statistiske princip omkring 'regression mod gennemsnittet', hvor gennemsnittet af to bedømmere alt andet lige vil støje mindre. Denne bevægelse mod gennemsnittet kan også siges at gøre sig gældende i nærværende eksperiment, selvom dansklærerne ikke er tvunget til at bruge medbedømmeren, og det herved er en frivillig bevægelse mod gennemsnittet. En reduktion i variansen og dermed støjen har den betydning, at vurderingernes afvigelse fra gennemsnittet er blevet mindre, hvilket må antages at være med til at styrke ensartetheden og pålideligheden ved karaktergivningen i folkeskolen.

7.3 H2: Den mindst støjende bedømmer

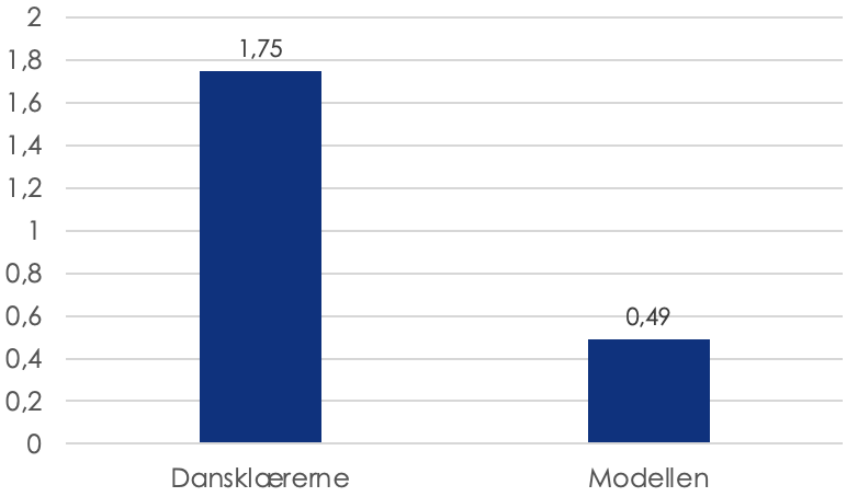
I det forrige afsnit blev det på signifikant vis fremlagt, hvordan en LL-model kan bidrage til støjreduktion ved karaktergivningen i dansk skriftlig fremstilling i folkeskolen, hvormed studiets første hypotese blev bekræftet. I forlængelse heraf vil det følgende afsnit teste studiets anden hypotese:

H2: LL-modellen kan være mindre støjende end den gennemsnitlige dansk censor

Her gør det sig gældende, at hvis LL-modellen forekommer at være mere støjende end den gennemsnitlige dansklæreres, så vil studiets anden hypotese blive forkastet. Dette vil blive undersøgt gennem en sammenligning af modellens afvigelse fra gennemsnittet samt den gennemsnitlige dansklærers reelle afvigelse fra gennemsnittet. Dette sker ud fra en forståelse af, at hvis modellen viser sig at være mere afvigende end den gennemsnitlige dansklæreres absolutte afvigelse, vil modellen ikke kunne bidrage til en mere ensartet og pålidelig karaktergivning, men derimod blot medvirke til yderligere støj i karaktergivningen og dermed forringe kvaliteten heraf.

Med henblik på at teste studiets anden hypotese, er gennemsnittet for dansklærernes første vurdering samt dansklærernes og modellens afvigelse herfra blevet udregnet. I den forbindelse gør det sig gældende, at dansklærernes gennemsnit ved første vurdering vil blive anvendt til at undersøge, hvorvidt det er modellen eller dansklærerne, der støjer mindst. At det er dansklærernes gennemsnit, som anvendes, sker med afsæt i fænomenet om *'the wisdom of the crowds',* som fremsætter, hvordan kollektive vurderinger kan bruges som et udtryk for den sande værdi i vurderinger, hvor denne ikke er kendt på forhånd. På baggrund heraf anses dansklærernes gennemsnit ved første vurdering inden interventionen for at være det bedste udgangspunkt for en pålidelig karakter, der kan anvendes i sammenligningen af respondenternes og modellen afvigelse (jf. Kahneman et al., 2021: 101-104; Galton, 1907). Dansklærernes gennemsnitlige afvigelse ved den første vurdering samt modellens afvigelse fra gennemsnittet ses i følgende tabel:

Tabel 7.9: Dansklærernes og modellens afvigelse fra gennemsnittet



Note: Den absolutte afvigelse fra gennemsnittet ved første vurdering (9,56)

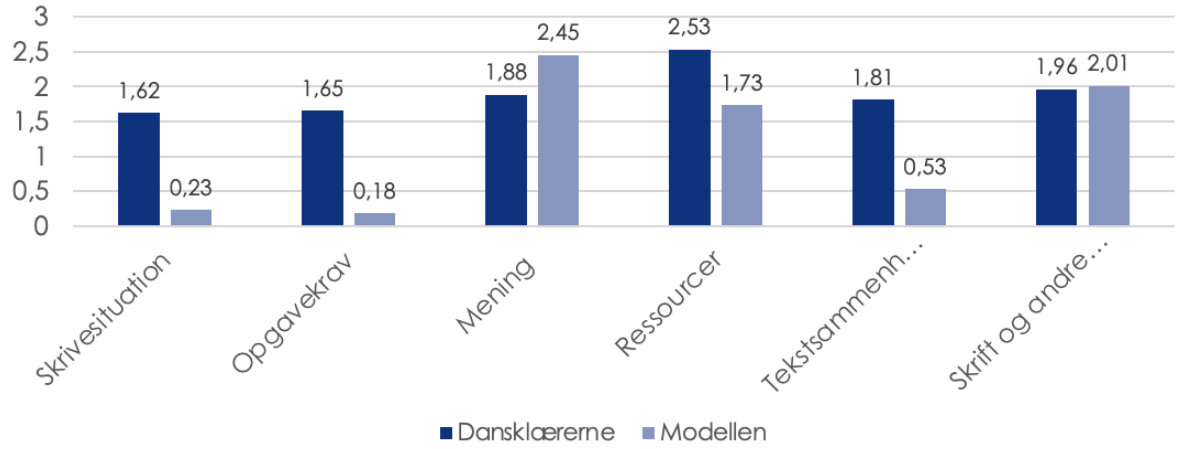
Kilde: bilag 10, s. 20

Af figuren bliver det tydeligt, at dansklærernes afvigelse fra gennemsnittet er 1,75, mens modellens afvigelse er 0,49. Med en afvigelse på 1,75 fra gennemsnittet kan det påpeges, hvordan dansklærerne afviger mere fra gennemsnittet end, hvad der er tilfældet for modellen. Hermed kan det udledes, hvordan modellen foreligger mindre støjende end den gennemsnitlige dansklærer, hvorfor det for nærværende eksperiment vil kunne antages, at LL-modellen vil kunne bidrage positivt til en mere ensartet og pålidelig karaktergivning.

Til trods for at modellen overordnet set er mindre støjende end den gennemsnitlige dansklærer, så er det ikke utænkeligt, at der kan være forskel, hvis der ses nærmere på afvigelserne ved de forskellige vurderingsområder. I den forbindelse var antagelsen, at LL-modellen vil have de bedste forudsætninger for at styrke karaktergivningen, når det kommer til vurderingsområderne 'tekstsammenhæng' samt 'skrift og andre modaliteter'. Angående vurderingsområderne vedrørende 'skrivesituation' og 'mening' var antagelsen dog omvendt, at modellen vil have svære ved at foretage en retvisende vurdering ved disse områder. Dette bunder i, at 'skrivesituation' og 'mening' er to vurderingsområder, der i højere grad kræver en kontekstforståelse i det fagprofessesionelle skøn, som modellen ikke forventes at evne på samme vis som dansklærerne (jf. afsnit 4.3.2; Puranam, 2021).

Med henblik på at skabe en mere nuanceret forståelse for, hvorvidt modellen er mere eller mindre støjende på de enkelte vurderingsområder sammenlignet med dansklærerne, er følgende figur opstillet:

Figur 7.10: Afvigelser ved de seks vurderingsområder



Kilde: bilag 10, s. 21-22

I tabellen ovenfor ses henholdsvis dansklærernes og modellens afvigelse fra gennemsnittet ved de seks vurderingsområder. Her kan det i overensstemmelse med ovenstående udledes, hvordan den gennemsnitlige dansklærer afviger mere fra gennemsnittet end modellen. Dette bliver tydeligt ved, at modellen på fire af de seks vurderingsområder er mindre afvigende end dansklærerne, hvor modellen særligt udmærker sig ved at være mindst afvigende på dimensionen vedrørende 'funktion'. For både dansklærerne og modellen gør det sig gældende, at funktionsdimensionen indeholder de to vurderingsområder, hvor både dansklærerne og modellen afviger mindst fra gennemsnittet. Forskellen er dog, at modellen afviger væsentlig mindre fra gennemsnittet sammenlignet med dansklærerne, idet modellen ved vurderingsområdet 'skrivesituation' afviger 0,23 fra gennemsnittet, mens afvigelsen fra gennemsnittet på vurderingsområdet 'opgavekrav' er 0,18. At 'skrivesituation' er et af de vurderingsområder, hvor modellen afviger mindst fra gennemsnittet og dermed ifølge Kahneman et al. (2021) er mindst støjende, forekommer bemærkelsesværdigt. Dette sker ud fra en forståelse af, at en vurdering af 'skrivesituationen' besidder en vis kompleksitet, da det typisk vil kræve en vis grad af kontekstforståelse i det fagprofessionelle skøn at vurdere, hvorvidt elevens besvarelse fungerer i den situation, som eleven skal

sætte sig ind i, hvilket LL-modellen ifølge litteraturen som udgangspunkt vil have svært ved at vurdere (jf. afsnit 4.3.2; Puranam, 2021).

Af tabellen kan det yderligere bemærkes, hvordan 'ressourcer' er det vurderingsområde, hvor dansklærerne har den største afvigelse, idet de på dette område har en gennemsnitlig afvigelse på 2,53 fra gennemsnittet. Dette betyder dermed, at 'ressourcer' er det vurderingsområde, hvor forekomsten af støj ses at være størst hos dansklærerne sammenlignet med modellen, hvilket også stemmer overens med hyppighederne, der blev fremlagt i tabel 7.2.

Med afsæt i afvigelserne kan det argumenteres, hvordan modellen ifølge Kahneman et al. (2021) vil kunne bidrage til at styrke karaktergivningen, når det kommer til vurderingsområderne 'skrivesituation', 'opgavekrav', 'ressourcer' og 'tekstsammenhæng'. Dette skyldes, at modellen på disse fire vurderingsområder ved den undersøgte elevtekst forekommer mindre støjende end den gennemsnitlige dansklærer. Omvendt kan det påpeges, hvordan modellen ved vurderingsområderne 'mening' samt 'skrift og andre modaliteter' er mere støjende end dansklærerne, idet modellens afvigelse fra gennemsnittet er større end den gennemsnitlige dansklærers absolutte afvigelse fra gennemsnittet ved disse to vurderingsområder. At modellen afviger mere fra gennemsnittet end dansklærerne er ifølge Kahneman et al. (2021) med til at indikere, hvordan modellen på disse områder som udgangspunkt ikke vil bidrage til at styrke karaktergivningen. Det er imidlertid ikke uventet, at modellen oplever en vis grad af udfordring og dermed udviser en større afvigelse, når det kommer til vurderingsområdet omkring 'mening'. Dette sker ud fra en forståelse af, at modellen jævnfør Puranam (2021) kan have svært ved at håndtere den kompleksitet, som det kræver at vurdere, i hvilken grad elevens besvarelse udtrykker et meningsfuldt indhold. Dette formår dansklærerne som forventet i højere grad på grund af deres evne til at forstå kompleksitet, helhedsforståelse og samt deres fagprofessionalisme. Hermed kan det ifølge Kahneman et al. (2021) fremhæves, at dansklærerne på vurderingsområdet 'mening' vil have de bedste forudsætninger for at vurdere en dansk skriftlig fremstilling frem for modellen.

De fremlagte resultater ovenfor er med til at bekræfte studiets anden hypotese om, at LL-modellen er mindre støjende end den gennemsnitlige dansklærer.

Foruden dette er resultaterne også med til at indikere, hvordan det vil være fordelagtigt, hvis bedømmelsespraksissen ved karaktergivningen i dansk skriftlig fremstilling i folkeskolen bestod af et samarbejde mellem en menneskelig bedømmer og en LL-model. Dette skyldes, at ovenstående viden om dansklærernes og modellens afvigelser fordelt på de forskellige vurderingsområder kan være med til at indikere, at der bør være en opmærksomhed omkring, hvilke områder modellen vil kunne supplere og dermed styrke censoren i sin karaktergivning. Ydermere vil det også kunne indikere hvilke områder, hvor censoren bør stole på og følge sin egen fagprofessionelle skønsmæssige vurdering. Netop denne arbejdsdeling påpeger Puranam (2021) som værende vigtig, hvis modellen skal kunne understøtte censoren i karaktergivningen og dermed skabe grobund for et vellykket samarbejde. Et vellykket samarbejde hvor opgaveløsningen og dermed karaktergivningen bliver udført med en større ensartethed, højere pålidelighed og stærkere kvalitet end, hvad censoren eller modellen på egen hånd ville kunne have præsteret (jf. afsnit 4.3.2 og 4.4.1 ; Parasuraman & Riley, 1997: 236-237; Mosier & Skitka, 1996).

7.4 H3: Tilliden til en AI-model som medbedømmer

I de forrige afsnit blev studiets første og anden hypotese undersøgt. Her blev det bekræftet, at en LL-model kan bidrage til at reducere støjniveauet ved karaktergivningen i folkeskolen samtidig med, at det også blev klarlagt, at LL-modellen er mindre støjende end den gennemsnitlige dansklærer. Efter at have bekræftet de to første hypoteser, så vil følgende afsnit have til formål at undersøge studiets tredje hypotese, der lyder som følger:

H3: Informationen om en AI-model som medbedømmer påvirker tilliden, der gennem brugen har betydning for støjniveauet

Som nævnt i det tidligere metodekapitel består den tredje hypotese af to delhypoteser (jf. H3.A og H3.B). Disse to delhypoteser vil blive undersøgt hver for sig, hvorfor det først og fremmest vil blive undersøgt, om tilliden til medbedømmeren er forskellig afhængig af, hvorvidt respondenten er blevet informeret om, at det er en anden censor eller en AI-model, som er

medbedømmer. Efterfølgende vil det blive undersøgt, hvorvidt denne tillid påvirker sandsyndsynligheden for, at respondenten enten over- eller underbruger LL-modellens vurdering i sin karaktergivning.

7.4.1 Hypotese 3.A: Informationen om en AI-model som medbedømmer og tilliden hertil

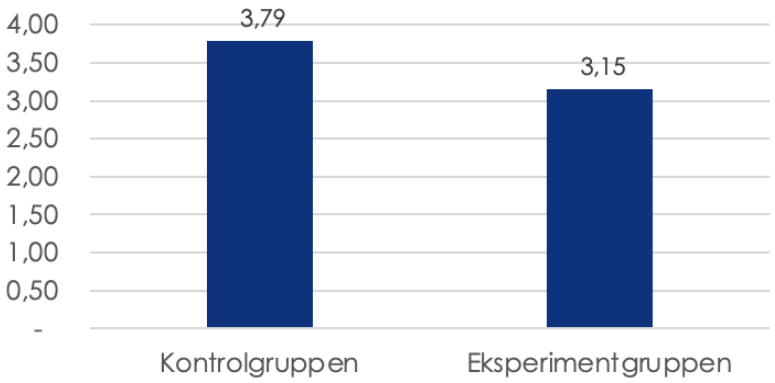
Som tidligere nævnt adskiller kontrol- og eksperimentgruppen sig ved, at kontrolgruppen bliver informeret om, at medbedømmeren består af en anden censor, mens eksperimentgruppen får information om, at det er en AI-model, som har foretaget en vurdering af elevteksten (jf. afsnit 6.3.6). Efter at have modtaget denne information samt set medbedømmerens vurdering af elevteksten, skal respondenterne besvare fem spørgsmål vedrørende deres tillid til den pågældende medbedømmer (jf. afsnit 6.3.7). Disse besvarelser vil i dette afsnit blive brugt til at undersøge den første delhypotese ved studiets tredje hypotese:

H3.A: Informationen om AI-modellen som medbedømmer har betydning for tilliden til medbedømmeren

I studiets metodekapitel blev det fremlagt, hvordan det foreligger nødvendigt at afdække, hvorvidt de anvendte spørgsmål om tillid i surveyeksperimentet kan udgøre et samlet mål for tillid (jf. afsnit 6.3.7). Af den grund foretog studiet en række reliabilitetstests. Disse viste, hvordan der overordnet set befinder sig et solidt grundlag for at anvende alle fem spørgsmål omkring tillid i et samlet mål.

For at undersøge den første delhypotese er der foretaget en t-test. Hensigten med denne t-test er at afgøre, hvorvidt kontrol- og eksperimentgruppens samlede tillid er signifikant forskellig. Resultaterne fra denne test ses på figuren nedenfor, hvor skalaen for tillid går fra 1 til 5:

Figur 7.11: T-test for gruppernes samlede tillid til medbedømmeren (skala: 1-5)



T-test: 4,5060, P-værdi: 0,00001

Kilde: bilag 10, s. 29

Af tabellen fremgår det, hvordan kontrolgruppen har en gennemsnitlig tillid på 3,79, mens eksperimentgruppens gennemsnitlige tillid ligger på 3,15. Dette betyder dermed, at kontrolgruppen udmærker sig med en samlet tillid, der er 0,64 højere end eksperimentgruppen. Denne forskel på 0,64 kan argumenteres at udgøre en væsentlig forskel, idet dette er en gennemsnitlig reduktion af tilliden på 23 % taget i betragtning, at skalaen kun har fire trin. Hermed kan det udledes, hvordan tilliden til en medbedømmer i karaktergivningen foreligger større, når respondenterne tror, at medbedømmeren er en anden censor fremfor en AI-model. Dette betyder med afsæt i Mayer et al. (1995) på, at respondenterne i kontrolgruppen må forventes at besidde en større villighed til at være sårbar over for medbedømmeren, og dermed tage medbedømmerens vurdering i betragtning ved karaktergivningen (jf. afsnit 5.2.4).

Den højsignifikante p-værdi ved t-testen for den samlede tillid er også med til at fastslå, hvordan der forekommer en signifikant større tillid til medbedømmeren i kontrolgruppen sammenlignet med eksperimentgruppen. Gennem metoden blev det fremlagt, hvordan særligt eksperimentgruppen havde oplevet et frafald i respondenter muligvis på grund af en skepsis overfor AI-modeller ved karaktergivningen. Det betyder derfor også, at studiets signifikante resultater vedrørende forskellen på kontrol- og eksperimentgruppens tillid til medbedømmeren er en stærk indikation på, at denne forskel både vil gøre sig gældende og formentlig også være større i praksis. Samtidig kan det også

fremhæves, hvordan ovennævnte resultater kan siges at være overensstemmende nogle af Jakesch et al.'s (2019: 6) empiriske fund, som også fandt en mindre tillid blandt dem, der blev fortalt, at information var skrevet af en AI-model, end dem der blev fortalt, at det var skrevet af mennesker (jf. afsnit 4.4.2).

Med henblik på at imødekomme en potentiel centralitet ved gennemsnittet for det samlede tillidsmål samt yderligere at undersøge, hvorvidt denne forskel i tilliden gør sig gældende ved de fem tillidsspørgsmål, har studiet foretaget en t-test ved alle spørgsmålene. Resultaterne herfra tydeliggøres i nedenstående tabel:

Tabel 7.12: T-test for gruppernes tillid til medbedømmeren

	Gnst. for kontrolgruppen og eksperiment-gruppen	T-testens p-værdi
1. Jeg kan stole på medbedømmerens evne til at give karakterer	K: 3,90 E: 3,36	0,0050
2. Jeg bør være forsigtig, når det kommer til at anvende ovenstående karaktergivning	K: 3,10 E: 2,19	0,0000
3. Den ovenstående begrundelse og karakterer er fair	K: 3,67 E: 3,37	0,1617
4. Ovenstående karaktergivning kan være hjælpsom i min egen karaktergivning	K: 4,09 E: 3,41	0,0006
5. Hvor stor tillid har du til medbedømmerens karaktergivning?	K: 4,14 E: 3,34	0,0000

Note: på en skala fra 1-5. Kilde: bilag 10, s. 30-33

Overordnet set kan det ud fra ovenstående t-tests fremhæves, hvordan der på tværs af alle spørgsmålene om tillid til medbedømmeren forekommer en højere tillid til medbedømmeren blandt deltagerne i kontrolgruppen sammenlignet med eksperimentgruppen. Dette understøtter dermed pointen ovenfra, hvor det blev tydeliggjort, hvordan eksperimentet finder en større tillid til en vurdering

foretaget af et menneske fremfor en AI-model. I tabellen kan det ses, at der ved spørgsmål 1, 2, 4 og 5 forekommer signifikante p-værdier ved t-testen, hvormed der er tale om signifikante forskelle mellem kontrol- og eksperimentgruppens tillid ved disse fire tillidsspørgsmål. Ydermere er det bemærkelsesværdigt, at der forekommer den største effekt på tilliden ved spørgsmålet om, hvorvidt respondenten mener, at vedkommende bør være forsigtig ved brugen af medbedømmeren. Med udgangspunkt heri tyder det på, at respondenterne mener, at forsigtighed er vigtigere i brugen af medbedømmeren, hvis det bliver fortalt, at medbedømmeren er en AI-model frem for en menneskelig medbedømmer. I henhold til Mayer et al. (1995) er dette med til at indikere, hvordan eksperimentgruppen i lavere grad vil være villige til at gøre sig sårbar overfor medbedømmeren, når denne er en AI-model. Dette kunne potentielt skyldes, at de finder en begrænset mulighed for at overvåge eller kontrollere medbedømmeren, hvormed de i højere grad mener, at der bør være en forsigtighed omkring at anvende en AI-model i karaktergivningen (jf. afsnit 5.2.4 og 6.4.5).

I modsætning til ovenstående tyder det ud fra tabellen på, at forskellen mellem gruppernes holdning til, hvorvidt karakteren er 'fair', er det eneste spørgsmål, hvor der ikke er signifikant forskel på kontrol- og eksperimentgruppen i forhold til tillid. Trods den manglende signifikans kunne dette spørgsmål dog indikere, at respondenterne forekommer at være mindre påvirket af informationen om, at vurderingen er udarbejdet af en anden censor eller AI-model, når det kommer til deres oplevelse af, hvorledes medbedømmelsen er fair. Dermed tyder det på, at respondenterne besidder en evne til at adskille vurderingen af selve indholdet og afsenderen af dette, når de skal tage stilling til den oplevede fairness.

Med afsæt i ovenstående signifikante resultater kan hypotese 3.A bekræftes, idet det blev påvist, at det havde en signifikant effekt på tilliden, når respondenterne i eksperimentgruppen blev informeret om, at medbedømmeren var en AI-model. Foruden det på signifikant vis kan påvises, at det har en effekt på tilliden, hvorvidt medbedømmeren er en anden censor eller en AI-model, så kan det ydermere klarlægges, at denne effekt forekommer at være negativ. Det betyder dermed, at det har en negativ effekt på respondentens tillid, når vedkommende informeres om, at medbedømmeren er en AI-model.

7.4.2 Hypotese 3.B: Tillidens betydning for brugen af medbedømmeren

I forlængelse af påvisningen om, at tillidsniveauet påvirkes negativt, når respondenterne modtager information om, at medbedømmeren er en AI-model, så leder det videre til at undersøge den anden delhypotese ved studiets tredje hypotese:

H3.B: Tilliden til medbedømmeren har betydning for brugen af medbedømmeren

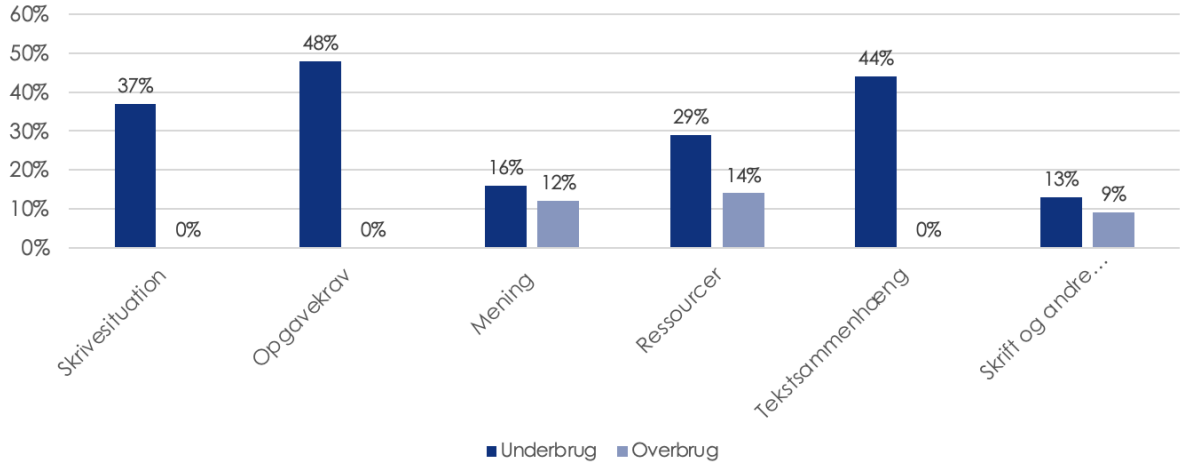
Til forskel fra den første delhypotese i det forrige afsnit, fremgår det af ovenstående, hvordan tillid ved den anden delhypotese vil udgøre den uafhængige variabel. I undersøgelsen af tilliden som den uafhængige variabel vil det samlede mål for tillid blive anvendt, hvilket også gjorde sig gældende i undersøgelsen af hypotese 3.A.

Med henblik på yderligere at operationalisere hypotesens afhængige variabel vil Parasuraman & Rileys (1997) teoretiske antagelser, der blev fremlagt i studiets teoretiske apparat, blive benyttet. Ifølge Parasuraman & Riley (1997) har individers tillid til automatiserede modeller betydning for brugen af dem. Dette sker ud fra en forståelse af, at der både kan opstå under- og overbrug af modellerne, hvor en balancegang herimellem skaber den mest hensigtsmæssige brug (jf. afsnit 5.2.5). Ud fra formålet om at sikre en hensigtsmæssig brug af medbedømmeren i karaktergivningen, så er henholdsvis overbrug og underbrug blevet operationaliseret ud fra en idé om, at brugen af medbedømmeren skal forekomme i de tilfælde, hvor denne brug kan være med til at reducere støjniveauet ved den enkelte karaktergivning. Herved er overbrugen defineret som brug af medbedømmeren i tilfælde, hvor brugen resulterer i, at respondenten bruger medbedømmerens anbefalinger til at ændre sin karakter, selvom dette bidrager til mere støj ved den endelige karakter. Omvendt er underbrug defineret ved en manglende brug af medbedømmeren i tilfælde, hvor denne kunne have reduceret støjniveauet ved, at respondenten nærmede sig medbedømmerens karaktergivning (jf. afsnit 5.1.7).

7.4.2.1 Tillidens betydning for brugen af medbedømmeren ved vurderingsområderne

I dette afsnit vil det blive undersøgt, hvorvidt tilliden har betydning for brugen af medbedømmeren. Forinden tillidens effekt på brugen af medbedømmeren undersøges, vil der indledningsvis foreligge en kortlægning af den faktiske brug, som gør sig gældende i eksperimentet. Dette sker i nedenstående tabel, hvor respondenternes under- og overbrug på tværs af de seks vurderingsområder er fremlagt.

Tabel 7.13: Andele af under- og overbrug



	Skrivesituation	Opgavekrav	Mening	Ressourcer	Tekstsammenhæng	Skrift og andre modaliteter
Lærernes gennemsnit	9,77	9,82	9,55	8,27	9,47	9,01
Medbedømmerens karakter	10	10	12	10	10	7

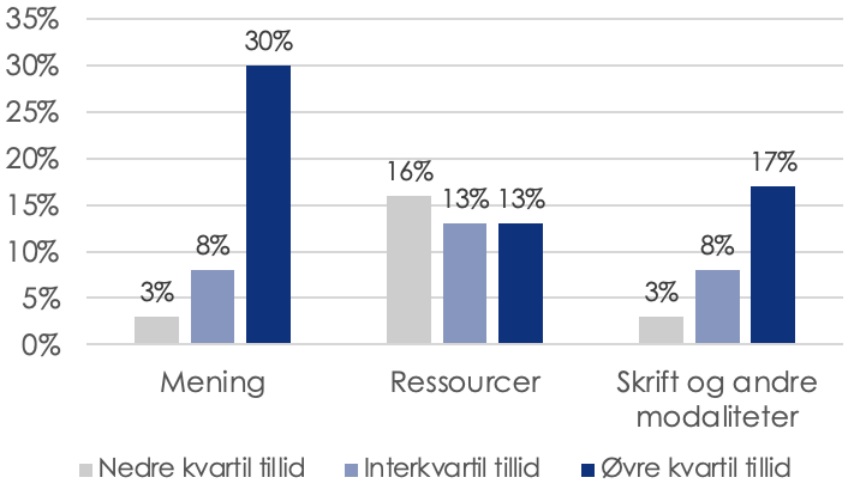
Note: Andelene er angivet i procent, hvormed det skal aflæses som den procentdel af stikprøven, som over- eller underbruger AI-modellen. Kilde: bilag 3; bilag 10, s. 22, 33-37

På tabellen kan det ses, at der på tværs af karaktergivningen ved de forskellige vurderingsområder er en større andel af respondenterne, som underbruger medbedømmeren end, der er respondenter, som overbruger den. Dette ses blandt andet ved, at der kun ved tre af vurderingsområderne sker et overbrug, mens der på tværs af alle vurderingsområderne forekommer underbrug. Ydermere bliver det også belyst, hvordan der er relativt store forskelle i andelen af dansklærere, som underbruger modellen. Her gør det sig gældende, at 'opgavekrav' er det vurderingsområde, hvor der er den største andel, som underbruger medbedømmeren, hvorimod 'skrift og andre modaliteter' er det

vurderingsområde, hvor den mindste andel ses at underbruge medbedømmeren. At der blandt dansklærerne i mindre grad forekommer et overbrug af medbedømmeren kunne skyldes, at det kan være svært at overbruge modellen, når modellens vurdering befinder sig tæt på den gennemsnitlige vurdering, hvilket er tilfældet for nærværende eksperiment. Det er med andre ord ikke muligt at overbruge medbedømmeren på den danske 7-trins-skala, når medbedømmerens karakterfastsættelse kommer så tæt på gennemsnittet som muligt, hvorfor der i dette studie ikke er målt nogen overbrug ved vurderingsområderne 'skrivesituation', 'opgavekrav' og 'tekstsammenhæng'. I henhold til videre forskning kan det derfor argumenteres, at disse resultater skaber et grundlag samt en nødvendighed for at afdække overbrugen yderligere.

Med en indledende viden om dansklærernes faktiske brug af medbedømmeren ved karaktergivningen i eksperimentet, vil den resterende del af afsnittet have til formål at skabe en større indsigt i, hvordan tillid har betydning for denne brug. Her vil sammenhængen mellem tillid og overbrug først og fremmest blive afdækket, hvilket sker i nedenstående tabel, hvor de tre vurderingsområder med overbrug er fremlagt:

Tabel 7.14: Tillidens betydning for overbrug



Note: Lav tillid består af den nedre kvartil på det samlede mål for tillid

Kilde: bilag 10: 56-61

Med afsæt i ovenstående er der ved vurderingsområderne 'mening' samt 'skrift og andre modaliteter' en indikation af, at jo højere tillid dansklærerne har til

medbedømmeren, des højere andel af dansklærerne overbruger medbedømmerens vurdering. Dette ses ved, at der forekommer flere dansklærere, som overbruger medbedømmeren, hvis de har høj tillid sammenlignet med, hvis de besidder en lav tillid. Disse indikationer kan her siges at stemme overens med de teoretiske antagelser jævnfør Parasuraman & Riley (1997), idet det forventes, at et højt niveau af tillid kan føre til overbrug. Et overbrug som ifølge Parasuraman & Riley (1997) kan være problematisk, hvis det er medvirkende til, at dansklæreren tilsidesætter sin egen kritiske tænkning ved brugen af LL-modellen, og på den måde kan være med til at svække karaktergivningen yderligere. Dette kunne eksempelvis ske, hvis dansklærerne anvender LL-modellen som en heuristik og dermed en kognitiv genvej til at foretage en hurtigere og mindre ressourcekrævende karakterfastsættelse. Ud fra en forståelse af, at karaktergivningen kræver en kognitiv vurdering fra dansklærerne kan heuristikkerne være med til at kompensere for dette. Ulempen er herved, at karaktergivning vil blive mere intuitiv som følge af 'system 1'-tænkningen, idet der ikke vil forekomme den samme refleksion og anstrengelse fra dansklærerne. Endvidere vil det også kunne argumenteres, at der ikke vil opstå den ønskede synergieffekt gennem et samarbejde mellem de to bedømmere i karaktergivningen, hvis dansklæreren blot overbruger AI-modellen og derved bidrager til yderligere støj (jf. afsnit 5.2.5).

Foruden ovenstående resultater indikerer, at Parasuraman & Rileys (1997) teoretiske antagelser om tillidens betydning for overbruget kan bekræftes, så bør det dog nævnes, at der ikke forekommer entydig evidens herfor. Dette kommer til udtryk ved vurderingsområdet 'ressourcer', hvor dansklærerne med det laveste tillidsniveau på overraskende vis udgør den gruppe, der har det største niveau af overbrug. Dette fund forekommer overraskende, da resultaterne udarter sig modsat studiets forventninger om, at det største overbrug ville gøre sig gældende ved det højeste tillidsniveau baseret på Parasuraman & Rileys (1997) teori.

Med udgangspunkt i ovenstående blev det tydeligt, hvordan fordelingerne af dansklærernes overbrug ved de tre vurderingsområder ikke foreligger at være i fuld overensstemmelse med studiets teoretiske antagelser, der havde en forventning om, at overbrug primært ville opstå ved et højt tillidsniveau. Denne

manglende overensstemmelse blev særlig tydelig ved vurderingsområdet 'ressourcer', mens vurderingsområderne 'mening' samt 'skrift og andre modaliteter' i højere grad underbygger de teoretiske forventninger. Med henblik på yderligere at undersøge sammenhængen mellem tillid og overbrug af medbedømmeren på disse to vurderingsområder er der foretaget to logistiske regressioner med følgende resultater:

Tabel 7.15: Logistisk regressioner for tillidens betydning for overbrug

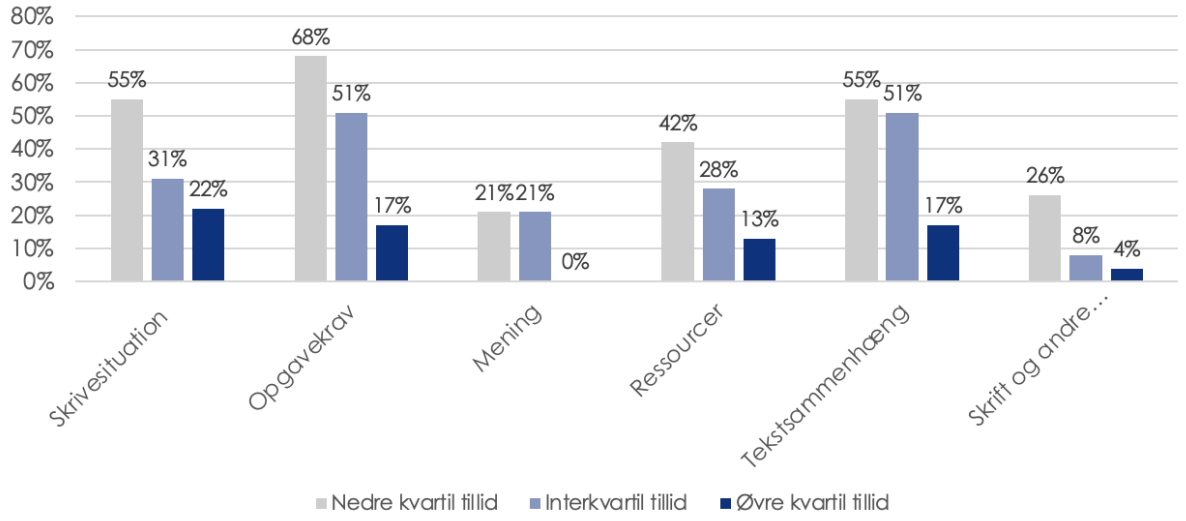
Vurderings-dimension	Vurderingsområder	P-værdi / Pseudo R^2	Intercept / hældnings-koefficient
Indhold	**Mening**	P-værdi: 0,004 R^2: 13 %	I: -7,3646 H: 1,4208
Form	**Skrift og andre modaliteter**	P-værdi: 0,3086 R^2: 2 %	I: -4,2631 H: 0,5245

Kilde: bilag 10, s. 72-74

Med afsæt i de logistiske regressioner kan det ses, at der ved 'skrift og andre modaliteter' ikke forekommer en signifikant sammenhæng mellem tillid og overbrug af medbedømmeren samtidig med, at forklaringsgraden også kun er på 2 %. Omvendt tyder det på, at der ved vurderingsområdet 'mening' forekommer en sammenhæng mellem tillid og overbrug af medbedømmeren. Dette ses ved, at p-værdien forekommer signifikant samtidig med, at forklaringsgraden er på 13 %. Hermed kan det for nærværende studie kun klarlægges, at tilliden har en positiv betydning for overbrugen ved dette ene vurderingsområde (jf. 'mening'). På den måde har det kun været muligt at udlede en sammenhæng mellem tillid og overbrug ved ét ud af de seks vurderingsområder, hvorfor der i henhold til hypotese 3.B ikke kan påvises en tydelig evidens for, at højere tillid generelt set fører til overbrug.

Nu hvor det er blevet klarlagt, at der ikke tyder på at være en klar sammenhæng mellem tilliden til medbedømmeren og overbrugen, vil der i det følgende blive afsøgt, hvorvidt det samme gør sig gældende ved underbrug.

Tabel 7.16: Tillidens betydning for underbrug



Note: Lav tillid består af den nedre kvartil på det samlede mål for tillid

Kilde: bilag 10: 56-61

Af ovenstående tabel tegner sig en indikation af, at jo lavere tillid dansklærerne har til medbedømmeren, des højere andel underbruger medbedømmerens vurdering. Dette ses på tværs af alle seks vurderingsområder, hvor det gør sig gældende, at vurderingsområdet 'opgavekrav' udmærker sig med det største underbrug. Det tyder dermed på, at der foreligger en negativ sammenhæng mellem tillid og underbrug af medbedømmeren, hvilket er med til at understøtte de teoretiske antagelser om, at underbrug ofte sker grundet et lavt niveau af tillid. Med udgangspunkt i Parasuraman & Riley (1997) kan det dermed argumenteres, at dansklærernes underbrug er til ulempe for karaktergivningen, idet medbedømmerens vurdering bliver tilsidesat selvom, at denne kunne have bidraget til støjreduktion og dermed styrket karaktergivningens pålidelighed (jf. afsnit 5.2.5).

Det er dog i henhold til de teoretiske antagelser overraskende, at der foreligger underbrug på tværs af alle tillidsniveauerne. Selvom det gør sig gældende i varierende grad, så tyder det på, at dansklærere med et højt niveau af tillid også underbruger modellen. Dette kunne være med til at indikere, at tillidens betydning for underbrug er kontekstafhængig, hvor det afhænger af, hvilket vurderingsområde dansklærerne skal afgive delkarakterer på. At tilliden tyder på at være kontekstafhængig kan være med til at nuancere Parasuraman & Rileys

(1997) idéer om sammenhængen mellem tillid og underbrug af medbedømmeren.

Med en indikation på, at der forekommer en negativ sammenhæng mellem tillid og underbrugen af medbedømmeren, vil det i følgende blive testet, hvorvidt denne sammenhæng forekommer signifikant ved brug af logistiske regressioner:

Tabel 7.17: Logistisk regressioner for tillidens betydning for underbrug

Vurderings-dimension	Vurderingsområder	P-værdi / Pseudo R^2	Intercept / hældnings-koefficient
Funktion	**Skrivesituation**	P-værdi: 0,0067 R^2: 6 %	I: 2,3063 H: -0,8274
	Opgavekrav	P-værdi: 0,0001 R^2: 12 %	I: 4,3085 H: -1,2477
Indhold	**Mening**	P-værdi: 0,0113 R^2: 8 %	I: 1,5993 H: -0,9741
	Ressourcer	P-værdi: 0,0200 R^2: 5 %	I: 1,6234 H: -0,7343
Form	**Tekstsammenhæng**	P-værdi: 0,0074 R^2: 6 %	I: 2,5399 H: -0,7966
	Skrift og andre modaliteter	P-værdi: 0,0144 R^2: 8 %	I: 1,4593 H: -1,0191

Kilde: bilag 10, s. 66-72

På tabellen ovenfor bliver det tydeligt, at der på tværs af alle vurderingsområderne er en signifikant sammenhæng mellem tilliden til medbedømmeren og sandsynligheden for underbrug. Dette ses ved, at der ved alle seks logistiske regressioner forekommer en signifikant p-værdi for modellerne. Af hældningskoefficienterne bliver det tydeligt, at tilliden har en negativ indvirkning på underbrugen set ud fra log-odds, hvilket også stemmer overens med resultaterne fra figur 7.16. Disse regressioner bekræfter herved de teoretiske antagelser fra Parasuraman & Riley (1997) om, at der ved en højere tillid forekommer en mindre sandsynlighed for underbrug. Ydermere kan det gennem

modellernes R^2-værdi ses, at tilliden forklarer mellem 5-12 % af variansen for sandsynligheden ved underbrug. Dette indikerer hermed, at andre faktorer forklarer 88-95 % af variansen for underbrugen. I den forbindelse kan det argumenteres, at denne viden også kan være med til at nuancere forståelsen af tillids betydning for brug videre end Parasuraman & Rileys (1997) teori, som ikke italesætter, hvordan andre faktorer end tillid også kan have betydning for brugen af automatiserede modeller.

7.4.2.2 Tillidens betydning for brugen af medbedømmeren ved den samlede karakter

Nu hvor det gennem ovenstående analyse er blevet vist, hvorledes tilliden tyder på at have en betydning for både over- og underbrugen på tværs af karaktergivningen ved vurderingsområderne, vil dette afsnit undersøge, hvorvidt det samme gør sig gældende for brugen ved den samlede karakter:

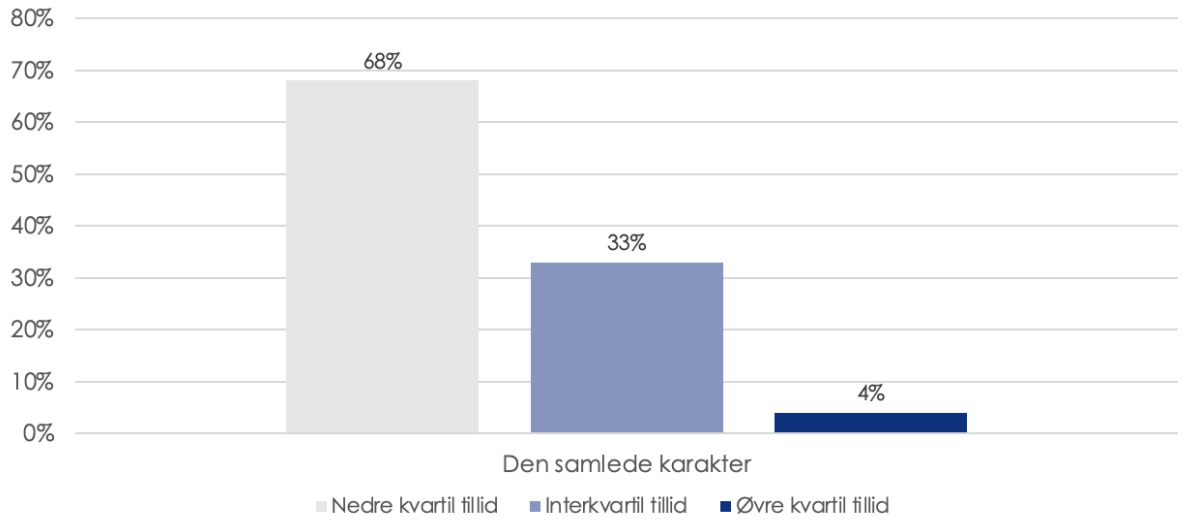
Tabel 7.18: Under- og overbrug for den samlede karakter

	Lærernes gennemsnit	Medbedømmerens karakter	Underbrug	Overbrug
Den samlede karakter	9,51	10	35 %	0 %

Kilde: bilag 3; bilag 10, s. 4, 34

Af tabel 7.18 fremgår det, at der ikke er nogen dansklærere, som overbruger medbedømmerens vurdering i den endelige karakterfastsættelse. Det samme gør sig dog ikke gældende for underbrugen, da 35 % af respondenterne ses at have et underbrug. Hermed tyder det på, at hver tredje dansklærer kunne have reduceret støjniveauet yderligere, hvis de havde gjort brug af medbedømmeren i deres karaktergivning. Da der ved den samlede karaktergivning ikke forekommer nogen overbrug blandt deltagerne, vil omdrejningspunktet for det følgende afsnit være at undersøge, hvorvidt tillid har betydning for underbrugen ved den samlede karakter.

Tabel 7.19: Underbrug i den samlede karakter



Kilde: bilag 10, s. 55

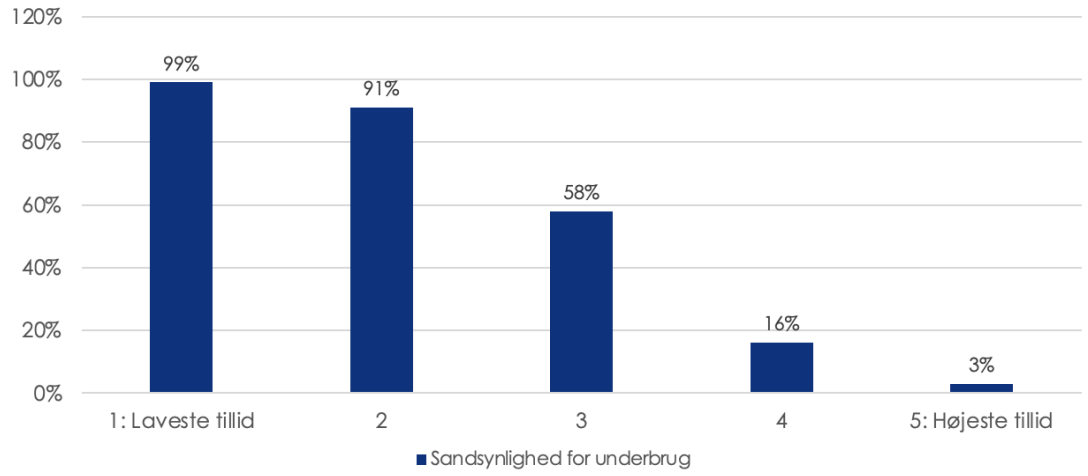
Ovenstående andele er med til at belyse, hvordan der foreligger tydelige tegn på, at jo mindre tillid dansklærerne har til deres medbedømmer, desto større underbrug af medbedømmeren foreligger der. Dette kommer til udtryk ved, at 68 % af dansklærerne i kvartilet med den laveste tillid ses at have et underbrug, mens dette kun gælder for 4 % af dansklærerne i kvartilet med den højeste tillid. Der forekommer hermed en indikation på, at der tale om en negativ sammenhæng mellem tillid og underbrugen af medbedømmeren, hvilket ses at være overensstemmende med Parasuraman & Rileys (1997: 230-238) teoretiske antagelse om, at underbrug ofte sker som følge af en lav grad af tillid. I henhold til Parasuraman & Riley (1997) kan det argumenteres, at dansklærernes underbrug og dermed tilsidesættelse af medbedømmerens input er til ulempe for karaktergivningen. Dette sker ud fra en forståelse af, at LL-modellen som medbedømmer foreligger mindre støjende end dansklærerne, hvormed denne kunne have bidraget til støjreduktion og dermed styrket karaktergivningens ensartethed, pålidelighed og dermed retfærdighed (jf. afsnit 5.2.5).

Når tillidseffekterne sammenlignes (jf. tabel 7.16 og tabel 7.19), kan det fremhæves, hvordan tillidseffekten er større ved den samlede karakter end ved de seks vurderingsområder. Dette ses ved, at differencen mellem den øvre og nedre kvartil for tillid er på hele 64 %-point ved den samlede karakter, mens denne forskel maksimalt er på 41 %-point ved vurderingsområderne. I henhold til

Parasuraman & Riley (1997) kunne det derfor argumenteres, at antagelserne om tillidens betydning for underbrugen kan nuanceres ved, at det ud fra ovenstående resultater tyder på, at tilliden får større betydning for brugen af medbedømmeren, når den fagprofessionelle vurdering bliver mere helhedsorienteret. Herved tyder det på, at når vurderingskriterierne for karaktergivningen bliver mindre konkrete, idet de går fra at være delkarakterer til en samlet karakter, så kan tilliden som en form for social kapital få betydning for, om censorerne bruger modellens vurderinger til at støjreducere (jf. afsnit 7.4).

Med ovennævnte indikation på, at tilliden har en negativ betydning for underbrugen af medbedømmeren, leder det frem til en interesse i at undersøge, hvorvidt denne sammenhæng forekommer at være signifikant. Dette bliver testet gennem en logistisk regression, hvor resultaterne herfra (omregnet til sandsynlighed) er fremlagt:

Tabel 7.20: Sandsynlighed for under-brug ved logistisk regression



Note: Sandsynlighederne ovenfor er udregnet på baggrund af logistiske regressions værdier med Pseudo R^2= 22 %, Interceptværdien (log odds) = 6,16, Hældningskoefficienten (log odds) = -1,94. Begge koefficienter havde p-værdier på 0,000. LLR P-værdi: 0,0000. Kilde: bilag 10, s. 62-63

Den logistiske regression ovenfor er med til at tydeliggøre, hvordan der forekommer en klar sammenhæng mellem niveauet af tillid og sandsynligheden for underbrug. Dette kommer først og fremmest til udtryk gennem forklaringsevnen, som indikerer, at tilliden kan forklare omtrent 22 % af den samlede varians vedrørende sandsynligheden for underbrug. Endvidere er modellens koefficienter for de forskellige log-odds blevet anvendt til at beregne

de ovenstående sandsynligheder. Gennem disse sandsynligheder bliver det tydeliggjort, at hvis dansklæreren har en samlet tillid på værdien 1, forventes det, at der er 99 % sandsynlighed for, at dansklæreren vil underbruge sin medbedømmer. Har dansklæreren derimod en samlet tillid på værdien 5, er sandsynligheden for underbrug af medbedømmeren i stedet 3 %. På tværs af grupperne kan det hermed udledes, hvordan der generelt set er større sandsynlighed for underbrug, desto lavere tillid der er til medbedømmeren. Foruden sammenhængen ses at være signifikant på et 99 % konfidensniveau, kan det også påpeges, hvordan sammenhængen gør sig gældende uanset om dansklærerne er blevet informeret om, at medbedømmeren givetvis er en anden censor eller en LL-model.

Ovenstående er i høj grad med til at underbygge de teoretiske antagelser fra Parasuraman & Riley (1997), som påpegede en sammenhæng mellem en tillidsgiver med et lavt niveau af tillid og underbrugen af den automatiserede model. I nærværende eksperiment foreligger dansklæreren at være tillidsgiver, hvor det gennem ovenstående logistiske regression bliver tydeligt, at sansynligheden for underbrug falder, desto højere tillid dansklærerne har til deres medbedømmer. At der foreligger et underbrug ved de dansklærere, som besidder en lav grad af tillid til medbedømmeren, kan ifølge Parasuraman & Riley (1997) være problematisk. Dette sker ud fra en forståelse af, at dansklærernes tilsidesættelse af medbedømmeren kan argumenteres at være en uhensigtsmæssig brugsform særligt i situationer, hvor opgaveudførelsen i form af karaktergivningen kunne have været styrket ved at inddrage medbedømmeren og derved skabt en støjreduktion.

7.4.2.3 Krydstjek af tillidens betydning for brug

I tråd med studiets teoretiske antagelser baseret på Parasuraman & Riley (1997) er det nu blevet undersøgt, hvorvidt tillid har betydning for brugen af medbedømmeren (jf. hypotese 3.B). I den forbindelse blev det i hypotese 3.A vist, hvordan kontrol- og eksperimentgruppens tillid forekom signifikant forskellige. At gruppernes tillid var signifikant forskellige leder videre til en interesse i at undersøge, hvorvidt studiets intervention herunder informationen om AI-modellen som medbedømmer kan bidrage med en forklarende faktor, når det

drejer sig om tillidens betydning for variansen i underbrugen. Foruden at eftertjekke om interventionen kan siges at have en direkte effekt på underbrugen, vil det ligeledes være relevant at kontrollere interventionens interaktionseffekt på tilliden. På den måde vil det også blive eftertjekket, hvorvidt sammenhængen mellem tilliden og underbrugen afhænger af interventionen.

Fælles for begge undersøgelser er, at disse vil blive foretaget gennem en multipel logistisk regression. Denne multiple logistiske regression har til hensigt at eftertjekke studiets kausalmodel, hvor det ved kontrollen af interventionen vil sikres, at der er en sammenhæng mellem tilliden og underbrugen af medbedømmeren. Dette sker ud fra en forståelse af, at førnævnte interaktioner ikke indgår som en del af studiets teoretiske antagelser (jf. afsnit 5), hvormed dette kan argumenteres at være en kontrol af studiets kausalmodel. Resultaterne fra denne multiple logistiske regression med interaktionsled er fremlagt i nedenstående tabel:

Tabel 7.21: Multipel logistisk regression for underbrug med interaktionsled

	Værdier	P-værdi
Den samlede model		0,0000
Pseudo R^2	23 %	
Intercept	7,5899	0,003
Tillid (koefficient for stigning på tillid)	-2,2833	0,002
Eksperimentgruppe (koefficient for effekten fra kontrolgruppen)	-1,6794	0,619
Eksperimentgruppe x Tillid (koefficient for interaktionsleddet)	0,3232	0,745

Note: Værdierne for koefficienterne i ovenstående er fremlagt i log-odds. Kilde: bilag 10, s. 63-64

Af tabellen fremgår det, at den samlede model besidder en signifikant p-værdi. Dette er med til at antyde, hvordan mindst én af de inkluderede prædiktorer i modellen har en signifikant effekt på den afhængige variabel i form af

underbrugen. På tabellen kan det yderligere ses, at den samlede model har en forklaringsgrad på 23 %, hvilket ikke er en markant stigning (1 %-point) sammenlignet med den simple logistiske regression, hvor der kun blev testet for sammenhængen mellem tillid og underbrug. Dette betyder med andre ord, at den tilføjede variabel i form af interventionen (jf. kontrol-/eksperimentgruppe) ikke bidrager med nogen yderligere bemærkelsesværdig forklaring på variansen af underbrug. Dette kommer ydermere også til udtryk ved koefficienten for eksperimentgruppen, hvor den negative effekt ved at gå fra kontrol- til eksperimentgruppen forekommer insignifikant med en p-værdi på 0,619. Den samme insignifikante effekt ses også ved interaktionsleddet, hvorved interaktionen mellem interventionen og tilliden ikke har en signifikant effekt på underbrugen.

Ovenstående insignifikante resultater er endnu en gang med til at bekræfte den anden delhypotese i studiets tredje hypotese (jf. hypotese 3.B) om, at tilliden til medbedømmeren har betydningen for brugen af denne. Samtidig kan det med afsæt i ovenstående også påpeges, at der ikke findes grundlag for at udvikle studiets kausalmodel yderligere, da kontrollen for interaktioner som nævnt ikke viste sig at være signifikant.

7.5 Opsamling

I løbet af det indeværende kapitel er studiets analyse blevet fremlagt. Gennem det første analyseafsnit blev det på deskriptiv vis klarlagt (jf. afsnit 7.1), hvordan dansklærernes karaktergivning var præget af støj, idet den selvsamme elevtekst blev vurderet til alle karaktererne mellem 4 og 12. Foruden den samlede karakterfastsættelse udmærkede sig som støjende, blev det ydermere også belyst, hvordan der forekom variabilitet i delkarakterne ved de seks vurderingsområder fra Børne- og Undervisningsministeriets vurderingsskema. I den forbindelse var det særligt vurderingsområdet 'ressourcer' som udmærkede sig med en stor spredning, hvor elevteksten havde modtaget alle karakterer fra 00 til 12.

I det andet analyseafsnit blev vist, hvordan dansklærerne i deres første vurdering havde et støjniveau på 4,36, mens dette støjniveau var reduceret til 2,75 ved

dansklærernes anden vurdering efter interventionen. Med andre ord betyder det, at eksperimentet med at tildele dansklærerne en medbedømmer resulterede i en støjreduktion på 37 %, hvilket viste sig at være signifikant på et 99 % konfidensniveau. Foruden støjen forekom reduceret ved den samlede karaktergivning, blev det også påvist, hvordan støjniveauet blev reduceret signifikant ved fem ud af de seks vurderingsområder (jf. tabel 7.6-7.7). Når det kommer til støjreduktionen ved den samlede karakter på tværs af grupperne, blev denne også reduceret med 35 % ved både kontrol- og eksperimentgruppen, hvilket kunne indikere, at interventionen ikke har den store betydning i forhold til støjreduktionen. Med afsæt i den signifikante støjreduktion forekommer der herved grundlag for at bekræfte studiets første hypotese om, at støjniveauet kan reduceres, når en LL-model anvendes som medbedømmer ved karaktergivningen i dansk skriftlig fremstilling i folkeskolen (jf. H1).

Ligesom første hypotese blev studiets anden hypotese også bekræftet, da det blev fremlagt, at LL-modellen overordnet set støjede mindre end, hvad den gennemsnitlige dansklærer gjorde (jf. afsnit 7.3). Dette kom til udtryk ved, at LL-modellens absolutte afvigelse fra gennemsnittet ved den samlede karakter var på 0,49, mens denne afvigelse var på 1,75 hos dansklærerne (jf. tabel 7.9). På tværs af vurderingsområderne viste LL-modellen sig at være mindre støjende på fire ud af seks vurderingsområder, mens vurderingsområderne 'mening' samt 'skrift og andre modaliteter' derimod er de to områder, hvor dansklærerne forekom mindst støjende (jf. tabel 7.10). Dermed kan det siges, at LL-modellen generelt er mindre støjende end dansklærerne, hvorfor studiets anden hypotese kan bekræftes (jf. H2).

I det fjerde analyseafsnit blev det tydeligt, at kontrolgruppen havde signifikant højere tillid til deres medbedømmer sammenlignet med eksperimentgruppen (jf. tabel 7.11). Denne negative effekt på tilliden, når dansklærerne er bekendte med, at medbedømmeren er en AI-model, gjorde sig både gældende ved det samlede tillidsmål og på tværs af de enkelte tillidsspørgsmål. I henhold til de enkelte tillidsspørgsmål fremgik det, hvordan kontrolgruppen udviser en højere tillid til medbedømmeren ved alle spørgsmålene om tillid, men forskellen på kontrol- og eksperimentgruppen kun forekommer signifikant ved fire ud af de fem tillidsspørgsmål (jf. tabel 7.12). På baggrund af disse resultater kan det

argumenteres, at der var grundlag for at retningsbestemme den første delhypotese (jf. H3.A), da det blev påvist, at informationen om AI-modellen som medbedømmer havde en negativ betydning for tilliden til denne medbedømmer.

I forlængelse af ovenstående blev den anden delhypotese også undersøgt (jf. H3.B). I den forbindelse var det ikke fuldt ud muligt at klarlægge, hvordan tilliden havde betydning for overbrugen af medbedømmeren. Dette var både tilfældet ved den deskriptive kortlægning af overbrugen på de enkelte vurderingsområder samt i undersøgelsen af tillidsniveauernes sammenhæng med overbrugen. Disse manglende sammenhænge blev yderligere bekræftet af de logistiske regressioners modsatrettede resultater (jf. tabel 7.13-7.15). Dette var dog ikke tilfældet ved underbrugen, hvor der ved alle seks vurderingsområder forekom signifikante resultater på, at tilliden havde betydning for underbrugen af medbedømmeren (jf. tabel 7.16-7.17). I henhold til Parasuraman & Riley (1997) tyder det dermed på, at tilliden til medbedømmeren særligt har betydning for underbrugen, mens dette nødvendigvis ikke gør sig gældende for overbrugen.

Ved den samlede karaktergivning stod det klart, at der blandt dansklærerne ikke var nogen, som overbrugte medbedømmeren. Til gengæld var der 35 % af dansklærerne som underbrugte medbedømmerens vurdering ved den endelige karakterfastsættelse (jf. tabel 7.18). I den forbindelse blev det tydeligt, hvordan der var mest underbrug ved de dansklærere, som havde det laveste niveau af tillid (jf. tabel 7.19). Dette blev yderligere bekræftet gennem den logistiske regression, som på signifikant vis påviste, hvordan tilliden havde en negativ betydning for sandsynligheden for underbrug (jf. tabel 7.20). Derudover blev det også eftertjekket gennem den multiple regression, hvor det blev klarlagt, at sammenhængen mellem tilliden og brugen af medbedømmeren var direkte korreleret (jf. tabel 7.21). Det tyder dermed på, at studiets anden delhypotese kan bekræftes, når det drejer sig som dansklærernes underbrug af medbedømmeren, da dette viser sig at være signifikant påvirket af tilliden til medbedømmeren (jf. H3.B).

Hermed er det muligt at opsummere, hvordan det i henhold til studiets tredje hypotese betyder, at informationen om, hvorvidt medbedømmeren er en anden censor eller en AI-model, har betydning for dansklærernes tillid. Hertil kan det

dog ikke klarlægges, at tilliden til medbedømmeren har betydning for overbrugen af denne. Derimod foreligger der dog stærke stærke indikatorer på, at tilliden har betydning for underbrugen, hvilket i sidste ende får en betydning for støjniveauet. Dette betyder dermed, at studiets tredje hypotese kun delvist kan bekræftes, idet det ikke foreligger fuldkommen klart, hvordan tilliden til medbedømmeren påvirker brugen og herigennem støjen (jf. H3).

Gennem ovenstående er det blevet undersøgt og analyseret, hvorvidt en LL-model som medbedømmer kan bidrage til at reducere støjen i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen, og om tilliden til denne LL-model har betydning for brugen og dermed støjreduktionen. Hermed er studiets forskningsspørgsmål blevet undersøgt, hvilket leder videre til følgende diskuterende kapitel.

8. Diskussion

I studiets analysekapitel blev det klarlagt, hvordan en bedømmelsespraksis med en LL-model som medbedømmer udmærker sig med en reducerende effekt på støjniveauet ved karaktergivningen i dansk skriftlig fremstilling i folkeskolen. Samtidig blev det også belyst, hvordan tilliden til medbedømmeren har en betydning for i hvor høj grad, at dansklærerne er villige til at bruge modellen og dermed indirekte en betydning for støjniveauet. Disse resultater vil første og fremmest blive diskuteret i henhold til de praktiske implikationer ved en implementering. Dernæst vil det også blive diskuteret, hvordan studiets resultater kan anvendes i henhold til at vurdere og udvikle de eksisterende teoretiske antagelser. Afslutningsvis vil studiets metodiske valg og konsekvenserne heraf blive diskuteret.

8.1 Praktisk diskussion vedrørende brugen af LL-modellen i karaktergivningen

Ud fra studiets ønske om at skabe værdi ved anvendelse af resultaterne i praksis, vil der i det følgende afsnit blive diskuteret, hvilke implikationer implementeringen af en LL-model som medbedømmer kan have på flere niveauer. Her vil følgende diskussion både berøre implikationer vedrørende individet, det organisatoriske niveau og det samfundsmæssige niveau.

8.1.1 Styrker LL-modellen kvaliteten af det fagprofessionelle skøn?

Når en dansklærer skal vurdere en elevtekst i dansk skriftlig fremstilling, udgør dansklærerens fagprofessionelle skøn en afgørende faktor i fastsættelsen af den endelige karakter. Med udgangspunkt i Kahneman et al. (2021: 191-192) kan det dog diskuteres, hvorledes LL-modellen som medbedømmer er med til at styrke kvaliteten af dansklærerens fagprofessionelle skøn.

På den ene side kan det argumenteres, at kvaliteten af det fagprofessionelle skøn styrkes ved, at LL-modellen som medbedømmer skaber grobund for, at censoren kommer til at reflektere over sin egen vurdering, hvilket over tid vil kunne medføre en kontinuerlig forbedring af det fagprofessionelle skøn. I den forbindelse forventes det, at denne refleksion særligt vil gøre sig gældende i forhold til de vurderingsområder, hvor der forekommer uoverensstemmelse

mellem censoren og LL-modellen. Her vil dansklæreren på baggrund af modellens begrundelse kunne opnå indblik i, hvordan en anden vurdering af den givne elevtekst ville kunne se ud, hvilket med stor sandsynlighed vil igangsætte en refleksion og samtidig aktivere dansklærernes 'system 2'-tænkning. At 'system 2'-tænkningen aktiveres kan argumenteres at være en fordel for karaktergivningen, idet den endelige vurdering i sidste ende vil blive mere velovervejet og velbegrundet, fordi dansklæreren indirekte tvinges til at genoverveje sin egen vurdering.

En alternativ måde, hvorpå LL-modellen kunne implementeres, ville være ved, at censorerne tvinges til at bruge LL-modellen som medbedømmer ved karaktergivningen. Jævnfør Kahneman et al. (2021) kan det argumenteres, at en tvungen brug af LL-modellen vil føre til støjreduktion, hvis den endelige karakter blev fastsat med udgangspunkt i gennemsnittet af censorens og LL-modellens vurderinger. Til trods for at en tvungen brug vil kunne reducere støjen, kan det også argumenteres, hvordan den tvungne brug kan have potentielle negative konsekvenser, som vil kunne påvirke kvaliteten af det fagprofessionelle skøn og karaktergivningen negativt.

Først og fremmest kan det argumenteres, at idet et fagprofessionelt skøn er defineret ved at være en vurdering, der ikke udelukkende kan reduceres til regler, så vil skønnet alt andet lige blive svækket, idet karaktergivningen delvist vil blive overladt til en LL-model, der udelukkende er baseret på matematiske funktioner og dermed regler (Møller et al., 2021: 83). Derudover kan det også argumenteres, at en tvungen brug vil svække censorens mulighed for at udøve sit fagprofessionelle skøn, udbygge sin egen dømmekraft og skabe en synergieffekt. Dette sker ud fra antagelsen om, at interaktionen herved ikke på samme måde vil udgøre et samarbejde, hvis brugen forekommer tvungen. Denne tvungne brug vil også kunne resultere i, at censorerne vil kunne opleve, at deres autonomi og fagprofessionelle skøn bliver begrænset, hvilket i værste tilfælde vil kunne resultere i en frustration og en følelse af, at censorernes faglighed ikke værdsættes (jf. Deci og Ryan, 2000).

Med henblik på at imødekomme ovenstående problemer vil det gennem det følgende blive argumenteret, hvordan en tvungen brug af LL-modellen som

medbedømmer ikke kan betragtes som fordelagtig, når det kommer til at reducere støjen i karaktergivningen. Selvom brugen af LL-modellen ikke er tvungen i studiets eksperiment, viser eksperimentets resultater på en overbevisende måde, hvordan LL-modellen alligevel er med til at reducere støjniveauet i karaktergivningen betydeligt. Dette blev særligt tydeligt ved, at en fjerdedel af alle respondenterne frivilligt ændrede deres endelig karakter efter, at de havde set LL-modellens vurdering. Hermed kan det på baggrund af nærværende studie udledes, hvordan det er muligt at reducere støjen i karaktergivningen i folkeskolen betragteligt uden, at censorerne direkte tvinges til at anvende LL-modellen som medbedømmer.

Modsat kan det dog argumenteres, at kvaliteten af det fagprofessionelle skøn ikke styrkes, hvis der blandt censorerne opstår en form for glidebaneeffekt, som ender med, at de vil overbruge LL-modellen uden at forholde sig kritisk og refleksivt til dens vurdering. Med andre ord er det afgørende, at karaktergivningen ikke må resultere i en situation, hvor censoren kan nøjes med at bruge sin 'system 1'-tænkning, som ifølge Parasuraman & Riley (1997) kan medføre et overbrug. Dette sker ud fra en forståelse af, at censorerne vil sætte hurtighed over grundighed, hvortil LL-modellen og dens vurdering kan ende med at blive brugt som en heuristik til at fastsætte den endelige karakter.

I eksperimentet forekom der eksempler på overbrug ved tre af vurderingsområderne, hvormed dansklærerne i nogle tilfælde blev mere støjende ved at bruge LL-modellen, hvilket er med til at svække kvaliteten af karaktergivningen. Dette sker ud fra en forståelse af, at hvis censorerne som følge af overbrug bliver mere tilbøjelige til at følge LL-modellens vurdering uden at forholde sig kritisk og refleksivt hertil, så vil dette blot svække karaktergivningen og kunne føre til yderligere støj. Foruden at det vil kunne bidrage til en større uønsket variabilitet, så kan det også argumenteres, at det i værste tilfælde vil kunne svække censorernes evne til at vurdere og give karakterer over tid, hvis der opstår en overafhængighed af LL-modellen, og censorerne dermed ender med at stole mindre på deres fagprofessionelle skøn og egen dømmekraft.

8.1.2 Medfører brugen af en LL-model et etisk dilemma?

I nærværende studie er der som nævnt kun påvist overbrug ved tre af vurderingsområderne, hvormed der ikke forekommer overbrug i den samlede karakter. At der ikke opstår overbrug ved den samlede karakter skyldes, at det i nærværende studie ikke har været muligt at overbruge LL-modellen, idet der ikke findes en karakter på 7-trins-skalaen, som er tættere på den 'sande karakter' end karakteren givet af modellen, som var 10 (jf. the wisdom of the crowds). Dog kan det ikke udelukkes, at der vil kunne forekomme et større overbrug ved den samlede karakter, hvis en LLM skulle vurdere andre elevtekster. Dette sker ud fra en forståelse af, at det gennem studiet er blevet tydeliggjort, hvordan den gennemsnitlige dansklærer bliver mindre støjende på bekostning af, at enkelte dansklærere gennem overbrug er blevet mere støjende. Det faktum, at der sker et trade-off, hvor størstedelen af dansklærerne bliver mindre støjende, mens enkelte af dansklærerne bliver mere støjende, kan diskuteres ud fra et etisk perspektiv. Dette sker ud fra en forståelse af, at det omtalte trade-off besidder et iboende dilemma, som handler om, hvorvidt det vil være retfærdigt, at én elev udsættes for en mere støjende karaktergivning, mens andre elever modtager en mindre støjende og dermed mere ensartet og pålidelig karaktergivning.

Svaret hertil foreligger ikke entydigt, da det som nævnt kan anskues som et etisk dilemma. I den forbindelse vil det ud fra et utilitaristisk perspektiv kunne argumenteres, at så længe størstedelen af eleverne modtager en mere ensartet og pålidelig vurdering og støjen i karaktergivningen overordnet set reduceres, så vil det anses som hensigtsmæssigt at benytte LL-modellen som medbedømmer i karaktergivningen i folkeskolen. Dette sker ud fra en forståelse af, at utilitaristerne vil have fokus på maksimering af den samlede nytte, hvormed utilitaristerne vil være tilhænger af den løsning, som vil tilgodese flest mulige elever (jf. Cuneo, 2020). Hermed vil der ud fra et utilitaristisk perspektiv advokeres for, at flere elever vil opleve en mindre støjende karaktergivning og dermed acceptere, at én elev vil opleve det modsatte. Dette skyldes, at alternativet med den eksisterende praksis givetvis ifølge utilitarismen vil være mere uhensigtsmæssigt, fordi flere af eleverne vil kunne opleve, at deres videre muligheder i uddannelsessystemet bliver uberettiget begrænset.

Omvendt vil det ud fra et deontologisk perspektiv kunne argumenteres, at en overordnet reduktion i systemstøjen ved karaktergivningen i folkeskolen ikke legitimerer en øget støj ved nogle af bedømmerne. Dette sker ud fra en forståelse af, at deontologien vil fokusere på moraliteten. Hermed vil det blive anset som værende moralsk fejlagtigt, hvis LL-modellen implementeres som medbedømmer ved karaktergivningen med henblik på at gavne flertallet af eleverne, mens karaktergivningen bliver forringet for enkelte elever. Hermed kan det med afsæt i deontologien udledes, at såfremt brugen af en LL-model som medbedømmer ikke vil kunne sikre det danske retsprincip om ligebehandling for alle eleverne, så vil dette ikke blive betragtet som et legitimt værktøj at benytte til støjreduktion ved karaktergivningen i dansk skriftlig fremstilling i folkeskolen (jf. Cuneo, 2020).

Med afsæt i ovenstående diskussion kan det udledes, hvordan både utilitarismen og deontologien bidrager med to modsatrettede perspektiver, når der kommer til dilemmaet om, hvorvidt en LL-model som medbedømmer potentielt grundet overbrug vil kunne medvirke til, at enkelte censorer bliver mere støjende. I den forbindelse kan det argumenteres, hvordan det vil være nødvendigt at finde frem til en etisk funderet balance, hvis LL-modellen på sigt skulle implementeres i praksis som medbedømmer ved karaktergivningen i folkeskolen.

Gennem studiets problemanalyse blev det fremlagt, hvordan støjen i karaktergivningen er en væsentlig udfordring med konsekvenser for både individet og samfundet. Med udgangspunkt heri vil nærværende studie advokere for, at det både vil være nødvendigt og ønskværdigt, at der implementeres en LL-model som medbedømmer selvom, at dette vil indebære en risiko for, at enkelte elever vil blive udsat for en mere støjende karaktergivning. Hermed kan det udledes, hvordan studiet i overvejende grad vil læne sig op af den utilitaristiske tilgang. Dog vil det også argumenteres at være væsentligt, at der forekommer et fokus på at imødekomme og hindre det uhensigtsmæssige overbrug mest muligt, som kan føre til at enkelte censore bliver mere støjende i deres karaktergivning.

En måde at undgå denne uhensigtsmæssige overbrug kunne være ved at oplære censorerne i korrekt brug af LL-modellen i karaktergivningen. Gennem oplæring må det antages, at censorerne vil blive bedre til at bruge modellen på hensigtsmæssig vis (jf. underbrug og overbrug). I studiets problemanalyse blev det fremlagt, hvordan LL-modeller både har styrker og svagheder, når det kommer til at vurdere tekster. Ved at censorerne er bevidste om dette, vil deres brug forhåbentlig også være præget af en konstruktiv og refleksiv tilgang til modellens vurdering. Det vil med andre ord antages, at der vil kunne forekomme mindre under- og overbrug, hvis censorerne oplæres i at bruge modellen hensigtsmæssigt til fordel for støjreduktionen.

8.1.3 Vil implementeringen af LL-modellen kunne udfordre retsprincipperne?

Såvel som det er relevant at diskutere etikken ved brugen af en LLM i karaktergivningen, er det yderligere væsentligt at diskutere, om brugen af en LL-model udfordrer de danske retsprincipper. Af studiets analyse blev det i tråd med tidligere undersøgelser fremlagt, hvordan karaktergivningen ved dansk skriftlig fremstilling i folkeskolen er præget af støj. Her kan det gennem principperne for god forvaltningsskik argumenteres, at en støjende karaktergivning er med til at udfordre retfærdigheden og lighedsprincippet, når nogle elever grundet manglende ensartethed og pålidelighed modtager en høj karakter, mens andre modtager en lav karakter for lignende præstationer (jf. Folketingets ombudsmand, 2023). Gennem studiet blev det dog på signifikant vis klarlagt, hvordan støjen i karaktergivningen kan reduceres ved brug af en LLM som medbedømmer. Hermed kan det på den ene side argumenteres, at brugen af en LL-model er med til at sikre borgernes grundlæggende demokratiske principper om ligebehandling. Dette skyldes, at en mere ensartet og pålidelig karaktergivning er med til at sikre, at færre unge uberettiget begrænses i deres muligheder, når det kommer til at vælge ungdomsuddannelse.

Omvendt vil det dog kunne argumenteres, hvorvidt de danske retsprincipper også udfordres ved brugen af LL-modeller i karaktergivningen, når det kommer til gennemsigtighed og retssikkerhed. Selvom Kahneman et al. (2021) betragter algoritmiske modeller som værende bedre end de menneskelige vurderinger, bør

det bemærkes, at dybe ML-modeller som nævnt i studiets problemanalyse besidder en vis grad af uigennemsigtighed (jf. afsnit 4.3.2 og 4.3.3). Denne uigennemsigtighed skyldes, at modellerne består af matematiske komponenter, der forekommer så komplekse, at menneskelige censorer ikke kan foretage direkte fortolkninger heraf (jf. Lundberg & Lee, 2017; Merrick & Taly, 2020: 18). Selvom der i de seneste år er udviklet metoder til at fortolke modellernes vægtninger, kan det stadig argumenteres, at der kan forekomme en begrænsning i muligheden for at få en uddybende forklaring på, hvordan modellen er kommet frem til dens vurdering (Mittal, 2023). Denne manglende mulighed for en uddybende forklaring på modellernes vægte, funktioner og bias vil primært være en udfordring for den censor, som skal bruge modellen i sin karaktergivning. Dette skyldes, at det i sidste ende altid vil være censoren, som har ansvaret for karakterfastsættelsen og samtidig en velfunderet og saglig begrundelse herfor. Gennem eksperimentet blev det belyst, hvordan modellen var mere støjende end dansklærerne ved to af vurderingsområderne. Dette er med til at signalere vigtigheden af, at censoren formår at forholde sig kritisk og velreflekteret til modellen og bibeholder sit fagprofessionelle skøn (jf. afsnit 8.1).

På den anden side kan det dog også argumenteres, at modellen ved klagesager kan bidrage med større gennemsigtighed for den enkelte elev. I en klagesag, hvor en elev har krav på at modtage en begrundelse for den givne karakter, kan brugen af LL-modellens vurderingen og censorens afvigelser herfra med andre ord bidrage til en mere velbegrundet og gennemsigtig begrundelse (jf. Børne- og Undervisningsministeriet, 2023c). Dette sker ud fra en forståelse af, at der ved folkeskolens afgangsprøve kun er afsat 16 minutter til vurderingen af den enkelte elevtekst (Børne- og Undervisningsministeriet, 2023d). Af den grund må det antages, at der ved den enkelte karaktergivning forekommer begrænset tid til at formulere og uddybe begrundelsen for karakterfastsættelsen. Dog vil LL-modellen formodentligt kunne bidrage til en mere velbegrundet vurdering, idet censoren vil kunne anvende denne vurdering til at begrunde sin egen bedømmelse yderligere. Hermed vil det i henhold til karaktergivningens gennemsigtighed kunne argumenteres, at retssikkerheden vil forekomme styrket, idet eleverne ved klagesager kan opleve en bedre indsigt i bevæggrundene for karakteren med en velfunderet begrundelse.

Hermed kan det på opsummerende vis udledes, hvordan en LL-model kan bidrage til støjreduktion i karaktergivningen, hvilket er med til at understøtte det demokratiske princip om ligebehandling. Omvendt vil der ved brugen af en LLM ved karaktergivningen kunne opstå en udfordring som følge af modellernes manglende transparens. Denne uigennemsigtighed kan medføre, at censorerne særligt ved uoverensstemmelse vil have svært ved at få en uddybende indsigt i grundlaget for LL-modellens vurdering. Trods en potentiel udfordring vedrørende manglende transparens, vil LL-modellen kunne udgøre et værktøj, hvortil censoren vil kunne foretage en mere velfunderet begrundelse, som både vil være med til at sikre retssikkerheden, og i sidste ende bidrage positivt til at styrke karaktergivning. Med afsæt heri kan det samlet set sige, at implementeringen af en LL-model som medbedømmer både vil kunne udfordre og styrke de danske retsprincipper.

8.1.4 Hvilken betydning har den lovmæssige regulering af AI for brugen af LLM i karaktergivningen?

I henhold til retssikkerheden kan det tydeliggøres, hvordan den danske lovgivning i nyere tid har været udfordret, når det kommer til anvendelsen af kunstig intelligens grundet en manglende regulering (Nisgaard & Bach, 2023). Dette er dog ikke længere tilfældet efter vedtagelsen af EU-forordningen "The AI Act" fra april 2024. Denne har til formål at sikre borgerne og samfundet mod de negative konsekvenser, der kan forekomme ved brugen af kunstig intelligens. Gennem forordningen bliver kunstig intelligens reguleret afhængigt af den risiko, som den udsætter enten EU-borgeren eller -samfundet for. Her vil kunstig intelligens, som indebærer en social scoring baseret på personlige egenskaber, blive klassificeret som værende 'uacceptabel'. Videre vil kunstig intelligens, der begrænser borgernes grundlæggende rettigheder eksempelvis i form af uddannelse, blive klassificeret som værende 'højrisiko' (Europa-Parlamentet, 2024; Europa-Parlamentet, 2023). Hermed kan det argumenteres, hvordan forordningen i høj grad er med til at styrke elevernes retssikkerhed, idet der herigennem forekommer internationale retningslinjer omkring brugen af kunstig intelligens, som den offentlige sektor er forpligtet til at sikre. I henhold til EU's AI Act kan det argumenteres, hvordan det i høj grad vil udfordre elevernes retssikkerhed, hvis karaktergivningsprocessen og herunder karakterfastsættelsen

udelukkende vil bestå af algoritmiske bedømmere. Hermed kan det udledes, hvordan det jævnfør forordningen vil være afgørende, at LL-modellen anvendes på retmæssig vis således, at brugen af LL-modellen ikke krænker elevens rettigheder.

Foruden at EU's AI Act kan være med til at sikre elevernes retssikkerhed, så kan det ligeledes argumenteres, at den også er med til at sikre dansklærernes berettigelse i karaktergivningen. Når det drejer sig om AI-teknologi og implementeringen heraf, tyder det på, at der blandt de fagprofessionelle forekommer bekymringer omkring brugen af denne (Meehl & Grove, 1996). Bekymringer som primært omhandler, hvorvidt implementeringen af AI-teknologien kan betragtes som begyndelsen på en række uønskede inkrementelle forandringer, som i sidste ende vil kunne resultere i en udslusning af det fagprofessionelle skøn og dermed censoren i karaktergivningen (jf. Kahneman et al., 2021: 160). Dette er vedtagelsen af EU's AI Act dog med til at forhindre, da det jævnfør forordningen ikke vil være muligt, at kunstig intelligens på egen hånd skal kunne determinere individets rettigheder og muligheder i samfundet. Dermed kan det diskuteres, hvorvidt grundlaget for ovenstående bekymring om en udslusning af den fagprofessionelle bedømmer i princippet foreligger ubegrundet, idet censorernes tilstedeværelse i karaktergivningen er afgørende for at opretholde elevernes retssikkerhed jævnfør EU's AI Act.

Samlet set kan det herved ud fra ovenstående siges, at EU's AI Act har været med til at sikre elevernes retssikkerhed. På trods af at der kan forekomme en risiko for, at censorerne overbruger LL-modellen i karaktergivningen, vil EU's AI Act dog stadig sikre, at ansvaret for karaktergivningen i sidste ende ligger hos den fagprofessionelle.

8.1.5 Skal dansklærerne uddannes for, at LL-modellen skaber værdi gennem støjreduktion?

Gennem de forrige afsnit er det blevet diskuteret, hvordan brugen af en LL-model i karaktergivningen både kan styrke og svække kvaliteten af det fagprofessionelle skøn samtidig med, at det også blev belyst, hvordan de danske retsprincipper sikres såvel som udfordres. En forudsætning for at kunne sikre både kvaliteten af

det fagprofessionelle skøn, retsprincipperne og støjreduktionen kan dog argumenteres at være, at censorerne formår at bruge LL-modellen hensigtsmæssigt, hvorved tilliden til LL-modellen som medbedømmer, ses at have en betydning.

Gennem studiets surveyeksperiment blev det tydeligt, hvordan der i mødet med LL-modellen var en forskel i måden, hvorpå dansklærerne reagerede. Dette kom til udtryk ved, at der var flere af respondenterne i eksperimentgruppen, som forlod surveyen, da de modtog informationen om, at en AI-model også havde vurderet elevteksten. Helt konkret var det 18 % af respondenterne i eksperimentgruppen, der frafaldt ved informationen om, at medbedømmeren var en AI. Dette kunne indikere, at der ville være en betragtelig andel af censorerne, der forventeligt vil have en modvillighed ved at bruge en LL-model som medbedømmer i karaktergivningen. Denne modvillighed vil være afgørende at håndtere og imødekomme, hvis en LL-model på sigt vil skulle implementeres som medbedømmer i form af et beslutningsunderstøttende værktøj i karaktergivningen.

I forbindelse med en implementering af kunstig intelligens fremlægger Kahneman et al. (2021:160) et studie af Meehl & Grove (1996), som finder en række socialpsykologiske forklaringer på fagprofessionelles modstand mod mekaniske vurderinger. Her bliver det tydeliggjort, hvordan 'frygt for arbejdsløshed' viste sig at være en betydningsfuld faktor for fagprofessionelles modstand til mekaniske vurderinger. Ud fra denne organisatoriske viden kunne der hermed advokeres for, at der med fordel forud for en implementering kunne igangsættes en række initiativer med henblik på at imødekomme en potentiel organisatorisk modstand. Dette kunne eksempelvis være gennem uddannelse. Formålet med initiativer såsom uddannelse vil være at informere og instruere de fagprofessionelle i brugen af LL-modeller samt mulighederne og begrænsningerne ved benyttelse af disse i karaktergivningen. Dette vil formodentlig kunne mindske noget af den modstand, som på baggrund af dette studie vil kunne forventes at opstå i forbindelse med implementeringen. Argumentet vil her være, at hvis de fagprofessionelle opnår en bevidsthed om, hvordan LL-modeller træffer beslutninger, samt hvordan algoritmer ikke kan stilles til ansvar juridisk, og derfor kun kan bruges som et værktøj i karaktergivningsprocessen (jf. EU's AI Act), så må

det antages, at en sådan viden vil kunne reducere modstanden og øge brugen. Således vil uddannelse med andre ord kunne forebygge, at der opstår en unødvendig frygt for, at LL-modellen fratager censorerne deres arbejde, da der som følge af EU's AI Act ikke vil være en reel mulighed herfor.

Selvom det gennem uddannelse af censorerne kan argumenteres at være en fordel, at censorernes modstand til LL-modellen mindskes, kan det omvendt dog argumenteres, hvordan det også kan være en fordel, at censorerne stadig besidder en vis grad af skepsis overfor AI-teknologien og LL-modellen som medbedømmer ved karaktergivningen. Dette sker ud fra en antagelse om, at hvis censorerne besidder en sund grad af skepsis over for den algoritmiske medbedømmer, vil dette også kunne argumenteres at bidrage til en mere ansvarlig, etisk og hensigtsmæssig brug heraf. Taget den ovennævnte modstand til AI-modellen i betragtning er det i den forbindelse dog afgørende, at censorernes skepsis ikke foreligger ubegrundet, men derimod forekommer sagligt begrundet. Dette skyldes, at det med overvejende sandsynlighed vil være nemmere at håndtere og imødekomme en sagligt begrundet skepsis i tilfælde af en implementering. Af den grund vil det på et organisatorisk niveau derfor være relevant at prioritere, at den fagprofessionelle bedømmer tildeles uddannelse i brugen af LL-modellen og dens kompetence. Dette sker ud fra en antagelse om, at det vil kunne have betydning for, hvorvidt censorerne besidder det rette niveau af skepsis, der i sidste ende vil kunne have betydning for støjreduktionen i karaktergivningen.

Foruden en diskussion om nødvendigheden af uddannelse i anvendelsen af LL-modellen kan det ydermere diskuteres, hvordan det vil være fordelagtigt, at dette også blev inkorporeret som en del af læreruddannelsen. Dette sker ud fra en forståelse af, at karaktergivning ifølge '*bekendtgørelsen om uddannelsen til professionsbachelor som lærer i folkeskolen*' ikke på nuværende tidspunkt forekommer at være en del af dansklærernes kompetencemål (jf. Retsinformation, 2015). Med henblik på at skabe en fælles forståelse for fastsættelsen af karakterer, og hvordan en LL-model vil kunne bruges hertil, vil det dermed være oplagt at påbegynde denne indsats under lærernes uddannelse, hvilket i sidste ende formodentligt vil være til fordel for støjreduktionen.

Selvom der kan være en række organisatoriske argumenter for at uddanne de fagprofessionelle i brugen af LL-modellen, så skal det her nævnes, at dette vil kræve en økonomisk prioritering fra politisk side. I studiets problemanalyse blev det italesat, hvordan der i den seneste politiske aftale vedrørende kvalitetsløftet af den danske folkeskole ikke er afsat midler til en forbedring af karaktergivningen (jf. afsnit 4.2.3). Det tyder herved på, at et kvalitetsløft af folkeskolens karaktergivning på nuværende tidspunkt ikke umiddelbart er en politisk prioritering. Dette vil dog være en præmis, hvis karaktergivningen i den danske folkeskole skal styrkes, hvormed det vil være nødvendigt at prioritere ressourcer hertil. I den forbindelse kunne en løsningsmodel være, at det udelukkende var de beskikkede censorer, som blev uddannet i LL-modellens kompetencer for at begrænse omkostningerne og ressourceforbruget. Omvendt vil den almene dansklærer også kunne opleve, at uddannelse i brugen af LL-modellen vil forekomme gavnlig. Dette sker ud fra en forståelse af, at dansklærerne vil kunne bruge denne viden i deres egen karaktergivning, til indsigt i bedømmelsespraksissen samt i kommunikationen med eleverne omkring LL-modellen - særligt hvis LL-modellen ender med at blive implementeret som en del af karaktergivningen på tværs af hele udskolingen. Selvom denne uddannelse af alle dansklærere vil være den løsning, der vil være forbundet med den største ressourcemæssige- og økonomiske byrde, vil det samtidig kunne argumenteres at være en investering i at sikre en hensigtsmæssig brug, den rette tillid og dermed støjreduktion.

8.1.6 Hvordan vil en manglende tillid til modellen hos den brede befolkning have konsekvenser for uddannelsessystemet?

I forrige afsnit blev det tydeliggjort, hvordan det vil være nødvendigt at prioritere ressourcer til at imødekomme den modstand, der potentielt kan opstå blandt de fagprofessionelle ved implementeringen af en LL-model som medbedømmer. Denne prioritering af ressourcer vil også kunne argumenteres at være nødvendig, når det kommer til at opbygge censorernes, elevernes og det resterende samfunds tillid til LL-modellen.

Gennem studiets eksperiment blev det vist, hvordan dansklærerne havde større tillid til medbedømmeren, hvis denne var en anden censor fremfor en LL-model. Ud fra en antagelse om, at dansklærernes tillid forekommer svækket ved en vurdering foretaget af en LLM, kan det herved ikke udelukkes, at det samme potentielt vil gøre sig gældende for den bredere befolkning. På et samfundsmæssigt plan kan en lav grad af tillid netop siges at være problematisk af flere årsager. For det første kan det forventes, at der ved en lav grad af tillid til karaktergivningen vil blive indgivet flere klager, hvilket i sidste ende kan betragtes som en ressourcemæssig byrde for systemet, idet flere klagesager vil medføre en forøgelse af det administrative arbejde. Ydermere kan en lav grad af tillid til karaktergivningen også forventes at bidrage til en forringelse af dennes legitimitet, hvilket vil kunne være til ulempe for befolkningens generelle opbakning og tillid til uddannelsessystemet. I den forbindelse vil det heller ikke kunne udelukkes, hvorvidt denne lave tillid til karaktergivningen i folkeskolen vil kunne forankre sig og påvirke elevernes tillid til karaktergivningen videre i uddannelsessystemet (jf. Nedergaard, 2022).

Med henblik på at imødekomme denne potentielle skepsis og samtidig opbygge befolkningens tillid til LL-modellen i forhold til en fremtidig implementering, vil det kunne argumenteres, hvordan dette kan forebygges gennem information. Herved kan det siges at være fordelagtigt at prioritere informationskampagner om, hvordan LL-modellen kan være med til at forbedre kvaliteten af karaktergivningen. For at kampagnerne vil kunne opnå den tilsigtede effekt, vil det ud fra et forandringsledelses perspektiv være nødvendigt, at disse vil skulle forløbe både forud, under og efter implementeringen af LL-modellen som medbedømmer ved karaktergivning (jf. Kotter, 2012). Her vil formålet være at forklare formatet for samarbejdet, forsikre om retssikkerhedens opretholdelse og bidrage til en mere velfunderet forståelse af LL-modellens begrænsninger og fordele ved benyttelse i praksis. Endvidere vil det også i henhold til kommunikationen foreligge afgørende, at befolkningen bliver vidende om, hvordan LL-modellen udelukkende vil være et beslutningsunderstøttende værktøj, hvormed den enkelte censor altid vil tage den endelige beslutning og dermed har det fulde ansvar for den givne karakter. Selvom dette vil bidrage med en øget omkostning i forbindelse med at udarbejde og lancere kampagnerne, vil fordelene herved formodentligt være, at tilliden fra både eleverne og den brede

befolkning styrkes, hvormed klagesagerne også vil forventes at blive på et minimum. Samtidig kan det også tydeliggøres, hvordan AI-teknologien med stor sandsynlighed også vil blive implementeret i andre dele af den offentlige sektor med formålet om at forbedre eller effektivisere forvaltningen (jf. KL, 2023; Jensen, 2024). Hermed kan det argumenteres, hvordan en informationskampagne, som vil kunne styrke befolkningens generelle tillid til den offentlige sektors brug af AI, kan anskues som en værdifuld investering.

Samlet set kan det ud fra den ovenstående diskussion om muligheder, konsekvenser og dertilhørende opmærksomhedspunkter ved at benytte en LL-model som medbedømmer ved karaktergivningen fremhæves, hvordan det alt andet lige vil kunne medføre et kvalitetsløft af karaktergivningen i folkeskolen. En vellykket implementering vil dog kræve, at der fra politisk side bliver prioriteret at afsætte ressourcer hertil både, når det eksempelvis kommer til at uddanne den fagprofessionelle i at bruge modellen hensigtsmæssigt, men også ressourcer til fortsat at sikre tilliden og opbakningen til uddannelsessystemet blandt både elever, lærere, censorer og det resterende samfund.

8.2 Diskussion af studiets teoretiske begrænsninger og bidrag

I ovenstående afsnit er muligheder og begrænsninger ved at implementere en LL-model som medbedømmer ved karaktergivningen i praksis blevet diskuteret. I forlængelse af diskussionen angående praktiske implikationer, så vil følgende afsnit bevæge sig videre til en mere teoretisk diskussion. Her vil det først og fremmest blive diskuteret, hvorvidt studiets resultater vedrørende informationen om en AI-models betydning for tilliden kan anses som et bidrag og dermed videreudvikling af de eksisterende empiriske fund på området.

8.2.1 Har AI-modellen som medbedømmer en positiv eller negativ effekt på tilliden?

Af dette studies forskningsspørgsmål fremgår det, hvordan der har været en interesse i at undersøge, hvorvidt tillid har betydning for brugen af en AI-model. Gennem en granskning af litteraturen blev det klart, at sammenhængen mellem

tilliden til AI-teknologi inden for konteksten af karaktergivning ikke tidligere var blevet undersøgt, hvormed dette studie ønskede at bidrage til det eksisterende forskningsfelt. Dette sker ud fra en forståelse af, at studiets teoretiske antagelse om tilliden til en vurdering foretaget af en AI-model er funderet på empiri indsamlet fra et bredere forskningsfelt, som har dét til fælles at have undersøgt tilliden til vurderinger foretaget af AI-modeller (jf. afsnit 4.4.2). I studiets problemanalyse blev det vist, hvordan der i litteraturen forekommer en uenighed om, hvorvidt informationen om en AI-model har en positiv eller negativ effekt på tilliden til vurderingen. Denne diskrepans resulterede dermed også i, at studiets første delhypotese ikke forekom retningsbestemt (jf. H3.A).

Resultaterne fra indeværende studie viste, at der forekom mindre tillid til vurderinger foretaget af en AI-medbedømmer frem for en menneskelig censor. Dette betyder dermed, at informationen om en AI-medbedømmer havde en negativ effekt på tilliden til medbedømmeren. Hermed blev det gennem resultaterne muligt at retningsbestemme H3.A, hvilket videre kan lede til en diskussion af, hvorvidt dette fund kan bidrage til en videreudvikling af de teoretiske antagelser om tilliden til vurderinger foretaget af AI-modeller.

At indeværende studie finder en mindre tillid til vurderinger foretaget af en AI-medbedømmer frem for en menneskelig bedømmer, understøtter herved de empiriske fund fra Jakesh et al. (2019) samt Madhaven & Wiegmann (2007), som begge fandt, at tilliden til AI-modeller var lavere end til mennesker. Når det kommer til Madhaven & Wiegmann (2007) samt Jakesh et al.'s (2019) fund, bør det bemærkes, hvordan disse henholdsvis er funderet i et litteraturstudie samt et studie om AI-genererede vurderinger i Airbnb. Herved bliver det tydeligt, at de empiriske fund og dermed teoretiske antagelser er fundet gennem andre kontekster end karaktergivning. Af den grund kan det argumenteres, at dette studies fund om tilliden til AI-genererede vurderinger i karaktergivningen bidrager til et bredere forskningsfelt og dermed andre kontekster.

Selvom der gennem ovenstående argumenteres, at studiets resultater kan bidrage til en bredere teoretisk antagelse om, at der forekommer lavere tillid til vurderinger foretaget af en AI-model, bliver denne mulighed for at anvende disse resultater i andre kontekster dog udfordret af Lerch et al. (1997). Dette ses ved, at

Lerch et al. (1997) i sit studie finder, at mennesker havde større tillid til computationelle ekspertsystemer frem for menneskelige eksperter (jf. Madhaven & Wiegmann, 2007: 285-286). Ud fra en kritisk rationalistisk tilgang, vil det grundet Lerch et al. (1997) derfor kunne argumenteres, at denne brede teoretiske antagelse om tilliden til AI-vurderinger bør forkastes. I den forbindelse bør det dog bemærkes, at det gennem problemanalysen blev tydeliggjort, hvordan der siden 1997, hvor Lerch et al.'s (1997) undersøgelse blev publiceret, har været en stor udvikling af de computationelle metoder. Med en viden om at de komplekse og dybe ML-modeller først blev udviklet efter årtusindeskiftet, må det herved antages, at Lerch et al.'s (1997) undersøgelse har taget afsæt i computersystemer, som ikke var karakteriseret ved den samme kompleksitet, som de nyere computationelle metoder besidder i dag. Af den grund vil det heller ikke være utænkeligt, at interaktionen med og videre tilliden til disse computationelle har ændret sig i takt med denne udvikling.

Med afsæt i ovenstående diskussion bliver det tydeligt, hvordan der forekommer to modsatrettede teoretiske antagelser om tilliden til AI-vurderinger. Mens litteraturen inden årtusindskiftet indikerer en højere tillid til computationelle vurderinger sammenlignet med menneskelige vurderinger (jf. Lerch et al., 1997), så gør det modsatte sig gældende i litteraturen udgivet siden midten af 00'erne. Det kunne herved tyde på, at de teoretiske antagelser om tilliden til AI-modeller forekommer at være kontekstafhængige, hvormed de kan siges at være afhængige af den teknologiske udvikling. Dette betyder derfor, at studiet ikke har fundet belæg for, at mennesker efter udviklingen af de nyere og mere komplekse computationelle systemer har større tillid til computationelle vurderinger sammenlignet med fagprofessionelle vurderinger, hvilket nærværende studie er med til at understøtte.

Ud fra ovenstående kan det dermed argumenteres, hvordan dette studies fund understøtter den del af empirien, som finder, at der forekommer højere tillid til vurderinger foretaget af mennesker sammenlignet med vurderinger foretaget af modeller. I den forbindelse gør det sig dog gældende, at studiets resultater som nævnt i metodekapitlet må antages med en vis grad af forbeholdenhed, når der kommer til generaliserbarhed (jf. afsnit 6.6), hvorfor det i fremtidig forskning kunne være en prioritering at belyse den empiriske diskrepans yderligere.

8.2.2 Påvirker tilliden til medbedømmeren brugen af medbedømmeren?

I forlængelse af ovenstående teoretiske diskussion vil der gennem det følgende afsnit blive diskuteret, hvorvidt studiets anvendte teoretiske antagelser baseret på Parasuraman & Riley (1997) bør nuanceres eller forkastes. Dette vil blive diskuteret med udgangspunkt i kriterierne for en god konceptualisering (jf. Gerring, 1999).

Når det kommer til studiets resultater vedrørende tillidens betydning for brugen af medbedømmeren, blev det først og fremmest klarlagt, hvordan studiet ingen sammenhæng fandt mellem tillid og overbrug. Der blev på trods af de teoretiske antagelser med afsæt i Parasuraman & Riley (1997) ikke fundet belæg for, at en høj tillid til medbedømmeren medførte en uforholdsmæssig høj brug af medbedømmerens vurdering. Disse resultater om den manglende sammenhæng mellem tilliden og overbrugen kan herved danne grundlag for en diskussion om, hvorvidt de teoretiske antagelser fra Parasuraman & Riley (1997) bør forkastes, eller hvorvidt en nuancering af forståelsen for tillidens betydning i henhold til brugen af automatiserede modeller bør finde sted. Hvis sammenhængen mellem tilliden og brugen af medbedømmeren gennem dette studie kun kunne måles ud fra overbrugen, ville det kunne argumenteres, at der forekom belæg for at forkaste Parasuraman & Rileys (1997) teoretiske antagelser. Dette er dog ikke tilfældet, idet der ikke foreligger en tydelig og signifikant evidens herfor, hvormed det kan argumenteres, hvordan den manglende sammenhæng mellem tilliden og overbrugen skaber grundlag for, at de teoretiske antagelser fra Parasuraman & Riley (1997) i stedet bør nuanceres.

Selvom dette studie ikke fandt en direkte sammenhæng mellem tillid og overbrug, så blev det gennem studiets resultater tydeligt, hvordan dette ikke gjorde sig gældende, når det drejede sig om sammenhængen mellem tillid og underbrug. Dette kom til udtryk ved, at der blev fundet en klar sammenhæng mellem tilliden til medbedømmeren og underbrugen af denne ved alle vurderingsområderne fra Børne- og Undervisningsministeriets vurderingsskema (jf. afsnit 7.4.2.1 og 7.4.2.2). Disse resultater om tillidens betydning for underbrugen understøttede dermed de teoretiske antagelser fra Parasuraman & Riley (1997). Til

trods for at studiets resultater kan ses som en bekræftelse af de teoretiske antagelser, så blev det gennem den logiske regression belyst, hvordan tilliden ikke alene udgør en faktor, når det kommer at forklare sandsynligheden for underbrug. Dette danner derved grundlag for en videre undersøgelse af denne sammenhæng med fokus på at skabe indsigt i de andre faktorer, som kan have betydning for underbrugen af medbedømmeren.

Med afsæt i ovenstående kan det opsummeres, hvordan det med udgangspunkt i studiets analytiske resultater kan argumenteres, at Parasuraman & Rileys (1997) teoretiske antagelser på nogle områder er præget af en overvejende grad af sparsommelighed. Gerring (1999) opstiller ni kriterier for en god konceptualisering. Herigennem er sparsommeligheden i Parasuraman & Rileys (1997) teori med til at begrænse muligheden for en nuancering af de analytiske resultater. Dette kan dermed også være med til at svække den teoretiske nytteværdi. Når det kommer til gode konceptualiseringer, påpeger Gerring (1999: 357-361) dog, hvordan der altid vil forekomme et trade-off mellem de forskellige kriterier for en god teoretisk konceptualisering. Dette betyder dermed, at Parasuraman & Rileys (1997) teoretiske sparsommelighed kan siges at blive opvejet, når det kommer til de teoriens styrker inden for genkendelighed, differentiering og sammenhæng. Netop genkendeligheden, differentieringen og sammenhængen har været væsentlige i henhold til muligheden for at undersøge årsags-virkningssammenhængene gennem studiets eksperiment. Dette var endvidere også årsagen til, at netop denne teori blev anvendt som grundlag for studiets teoretiske apparat. Det blev dog efter endt analyse tydeligt, hvordan det forekommer nødvendigt at nuancere og udbygge Parasuraman & Rileys (1997) teori for at styrke den teoretiske nytteværdi.

Med udgangspunkt i at der gennem Parasuraman & Rileys (1997) teori forekommer en manglende sammenhæng mellem tillid og overbrug, vil det ud fra Gerrings (1999) principper om en god konceptualisering afsluttende argumenteres, at indeværende studie motiverer en videre nuancering af den eksisterende antagelse om tillidens betydning for brugen. Teoretisk set udvider studiet herved Parasuraman & Rileys (1997) teori ved at fremhæve, at andre faktorer også må besidde en væsentlig rolle i forklaringen af brugen, hvor det

særligt ved undersøgelsen af overbrugen forekommer nødvendigt at udforske disse komplekse forhold yderligere i videre forskning.

8.3 Metodisk diskussion af studiets kvalitet

I forrige afsnit blev studiets teoretiske begrænsninger og bidrag diskuteret, hvorfor det følgende afsnit videre vil diskutere de potentielle implikationer ved studiets metodiske valg samt diskutere, hvordan dette vil kunne håndteres i fremtidig forskning.

8.3.1 Hvilke potentielle uintenderede effekter har interventionen?

Når det kommer til den eksperimentelle metode, blev det gennem metodekapitlet fremlagt, hvordan interventionen besidder en afgørende rolle i at kunne afdække kausalrelationer (jf. afsnit 6.2). Dette betyder dermed, at interventionen udgør en central del af studiets eksperimentelle design, hvorfor det i det følgende vil blive diskuteret, hvorvidt denne har virket efter hensigten.

I henhold til interventionen kan det fremhæves, hvordan et kritikpunkt i forbindelse med surveyeksperimenter ofte er, at disse indebærer en lavintensiv intervention. Dette skyldes, at interventionen i surveyeksperimenter typisk består af mindre ændringer i spørgsmålsformuleringerne (jf. Blom-Hansen & Serritzlew, 2011: 16-17). Med andre ord betyder dette, at i tilfælde hvor interventionens intensivitet bliver for lav, så vil dette med overvejende sandsynlighed udfordre studiets interne validitet. Dette kunne eksempelvis ske ved, at respondenterne ikke opfangede eller reagerede på interventionen på samme måde, som de ville have gjort i virkeligheden eller ved andre eksperimenttyper. Med udgangspunkt i studiet kunne dette eksemplificeres ved, at der kunne være en risiko for, at dansklærerne i mindre grad reagerer på LL-modellen som medbedømmer, fordi interventionen blot fremgår som en mindre sproglig ændring i surveyen. Hermed vil risikoen være, at nogle respondenter ikke reagerer på samme måde, som de ville have gjort, hvis LL-modellen blev implementeret som en del af karaktergivningsprocessen i praksis, hvor ændringen i højere grad vil være uundgåelig ikke at bemærke.

Selvom surveyeksperimentet som eksperimentel metode besidder denne ulempe, kan det omvendt argumenteres, at det grundet studiets signifikante forskel mellem kontrol- og eksperimentgruppens tillid til medbedømmeren samt eksperimentgruppens frafald forekommer indikationer på, at interventionen har haft en effekt. Det tyder dermed på, at den potentielle udfordring med respondenternes manglende opfangelse af interventionen på grund af en lav intensitet, ikke har haft en problematisk indvirkning på nærværende eksperiment. Derimod kan det argumenteres, at denne lavintensivitet er med til, at studiets kausale model blot testes under strengere forudsætninger.

Foruden at lavintensiviteten kan udgøre et diskussionspunkt i forhold til surveyeksperimenters interne validitet, kunne et andet diskussionspunkt også være udfordringen om, hvorvidt effekten ved interventionen egentlig måler den intenderede kausalrelation - eller om effekten ved interventionen opstår på baggrund af multiple og/eller spuriøse sammenhænge. Gennem litteraturen om kausalsammenhænge bliver denne udfordring omtalt som 'informational equivalence'. Dette defineres ud fra "*(...) antagelsen om, at undersøgelsens manipulation er informationsækvivalent (IÆ) med hensyn til relevante baggrundsfaktorer i scenariet*" (Dafoe et al., 2018: 399-400, oversat). Når det kommer til at måle effekten af en intervention, påpeger Dafoe et al. (2018), hvordan det ikke forekommer tilstrækkeligt blot at udsætte eksperimentgruppen for en intervention, som kausalslutningen ræsonneres ud fra. Hertil mener Dafoe et al. (2018), at det foreligger nødvendigt at sikre, at interventionseffekten ikke skyldes spuriøse sammenhænge, da dette vil forårsage en fejlslutning i tolkningen af denne effekt. Dette kunne en manglende informationsækvivalens eksempelvis lede til, hvilket ifølge Dafoe et al. (2018: 399-400) vil svække den interne validitet. I henhold til dette studie kan dette dermed argumenteres at udgøre en potentiel problematik, idet der kan forekomme afledte ændringer på andre baggrundsvariable med relevans for eksperimentets afhængige variabel. Herved kan det argumenteres, at der på individuelt niveau kan forekomme afledte ændringer på forskellige relevante variable blandt respondenterne i eksperimentgruppen såsom forskellige forståelser for AI's styrker og svagheder, der kan have betydning for informationsækvivalensen.

Gennem Dafoe et al. (2018: 405) kan det dog argumenteres, at denne informationsækvivalens kan imødekommes og forebygges ved at foretage en mere udførlig beskrivelse af scenariet i interventionen. Ydermere argumenterer Dafoe et al. (2018) også for, at der internt i eksperimentet kan indlejres andre eksperimenter, hvorved det kan være muligt at kontrollere for spuriøse effekter (jf. Dafoe et al., 2018: 405-406). På den ene side kan det argumenteres, hvordan studiet indirekte har forsøgt at forebygge denne informationsækvivalens ved at indsætte definitionen af en AI-algoritme, som en del af interventionen med henblik på at sikre en fælles forståelse blandt respondenterne i eksperimentgruppen (se bilag 2). Selvom studiet i forvejen har foretaget initiativer med henblik på at forebygge informationsækvivalensen og de spuriøse sammenhænge, kan det på den anden side med afsæt i Dafoe et al. (2018) også argumenteres, at kausaliteten kunne have været sikret yderligere, hvis der i surveyeksperimentet var blevet testet og herved kontrolleret for andre kausalrelationer.

8.3.2 Kunne studiets kvalitet være forbedret gennem alternative undersøgelsesstrategier?

I det forrige afsnit var omdrejningspunktet studiets interne validitet, hvor det blev diskuteret, hvordan den kunne have været styrket yderligere. Dette vil også være formålet i følgende afsnit, hvor der i højere grad vil være fokus på den økologiske validitet.

Når det kommer til surveyeksperimenter generelt kan det tydeliggøres, hvordan det nødvendigvis ikke foreligger givet, at disse besidder en høj økologisk validitet, da disse ofte kan besidde unaturlige elementer (Andersen et al., 2020). Af studiets metodekapitel fremgik det dog, hvordan nærværende studie trods dette udmærkede sig med en høj økologisk validitet. Dette skyldes, at surveyeksperimentet bestod af elementer, som i høj grad repræsenterede respondenternes virkelige omgivelser, når det kommer til vurderingen af en dansk skriftlig fremstilling ved folkeskolens afgangseksamen. Omvendt er det ikke utænkeligt, at denne økologiske validitet kunne have været styrket yderligere, hvis studiet i stedet havde foretaget et felteksperiment. Dette sker ud fra en forståelse af, at felteksperimentet er karakteriseret ved, at dette udføres i

felten, hvormed dette er et af de eksperimentelle designs, der typisk står stærkest, når det kommer til den økologiske validitet (jf. Eden 2017 i Hansen & Trummer, 2020: 922). Med henblik på at styrke den økologiske validitet vil det jævnfør Eden (2017) dog kræve, at interventionen skal forekomme realistisk. I den forbindelse kan i henhold til nærværende studie diskuteres, hvorvidt scenariet med, at dansklæreren tildeles en LL-model som medbedømmer, vil forekomme realistisk i naturlige omgivelser ved en reel karaktergivning. Samlet set kan det udledes, hvordan det ikke fremstår klart, hvorvidt felteksperimentet som undersøgelsesstrategi potentielt kunne have styrket studiets økologiske validitet. Ydermere er det også værd at bemærke, at felteksperimentet er et ressourcekrævende eksperimentelt design, hvorfor denne undersøgelsesstrategi er fravalgt i indeværende studie (jf. Hansen & Trummer, 2020: 921).

Nu hvor det gennem ovenstående er blevet diskuteret, hvorvidt et feltstudie kunne have bidraget med øget kvalitet af studiets resultater, vil det afslutningsvist blive diskuteret, hvordan opfølgende kvalitative interviews også kan styrke studiets kvalitet. Gennem studiets metodekapitel blev det belyst, hvordan flere respondenter i eksperimentgruppen faldt fra, da de blev præsenteret for, at deres medbedømmer som følge af interventionen var en AI-model (jf. afsnit 6.4.5). I den forbindelse blev det antaget, at dette frafald potentielt kunne være forårsaget af en AI-skepsis blandt respondenterne, som resulterede i, at de reagerede negativt og dermed forlod surveyen. Dette var dog udelukkende baseret på antagelse, idet den reelle årsag til frafaldet ikke var mulig at afdække gennem surveyeksperimentet. At surveyeksperimentet er foregået online har medført, at der gennem denne undersøgelsesstrategi ikke har været en fysisk interaktion med respondenterne, hvilket kan argumenteres at udgøre en ulempe med henblik på at opnå en forståelse for frafaldet i eksperimentgruppen. Denne manglende forståelse kunne dog eksempelvis imødekommes gennem kvalitative interviews, da dette i højere grad vil kunne muliggøre en mere dybdegående undersøgelse af respondenternes oplevelse med en LL-model som medbedømmer karaktergivningen, hvilket også kunne have styrket den interne validitet yderligere (jf. Brinkmann & Tanggaard, 2020: 35)

8.3.3 Ville brugen af en anden LL-model reducere støjen yderligere?

I forlængelse af ovenstående metodiske diskussioner vil det i følgende afsnit blive diskuteret, hvorvidt en mere domænespecifik LL-model som medbedømmer kunne have reduceret støjniveauet yderligere.

Gennem studiets metode blev det fremhævet, hvordan der i dette studie er blevet anvendt en generalistisk LL-model, hvilket betyder, at modellen er trænet på et meget bredt funderet træningsdata, der kvalificerer modellen til at løse forskelligartede opgaver. Med andre ord er den ikke specifikt trænet til formålet om at kunne give karakterer. I studiets metodekapitel blev det italesat, hvordan det forventes, at en mere domænespecifik model formodentligt vil have bedre forudsætninger for at foretage mindre støjende vurderinger i karaktergivningen, fordi denne vil være trænet til formålet. Dette sker ud fra en forståelse af, at det ved træningen af en domænespecifik model i højere grad vil være muligt at kontrollere, at modellen besidder de ønskede kompetencer (jf. Pattam, 2023; Pal et al., 2024; Tariq et al., 2024; Ramesh & Sanampudi, 2022). Ud fra disse antagelser om, at den generalistiske LL-model vil have nogle af de dårligste forudsætninger for at foretage vurderinger af en dansk skriftlig fremstilling, kan det med afsæt heri argumenteres, at støjreduktionen ved dette surveyeksperiment læner sig op af principperne vedrørende en 'least likely'-case (jf. afsnit 6.3.5). På trods af at den eksterne validitet i dette studie forekommer begrænset, kan det dog påpeges, at det forventes, at der ved brugen af en domænespecifik model kan forekomme en endnu større støjreduktion.

Selvom der gennem ovenstående er argumenteret for, at en mere domænespecifik model vil kunne have reduceret støjniveauet yderligere, er det dog også værd at bemærke, hvordan der også kan forekomme udfordringer ved brugen af domænespecifikke LL-modeller trænet på færre datapunkter. I henhold til karaktergivningen i dansk skriftlig fremstilling kan det argumenteres, at den anvendte LL-model i karaktergivningen skal kunne vurdere forskellige skrivegenrer og indholdet af forskellige emner, idet opgaveformuleringerne og emnerne heri vil skifte fra år til år ved folkeskolens afgangseksamen. I den forbindelse kan det diskuteres, hvorvidt en domænespecifik model vil være

fordelagtig, da denne vil skulle trænes jævnligt med henblik på at forblive opdateret og derved kunne foretage en vurdering ved vurderingsområder såsom 'mening' og 'skrivesituation'. Af den grund kan det argumenteres, at modellen skal have indsigt i de forskellige emner og kontekster, som eksamensopgaven skal skrives ud fra.

I henhold til dette studie er det afgørende, at der i skrivende stund ikke eksisterer en tilgængelig domænespecifik LL-model, som kan anvendes i karaktergivningen. Dog er det værd at bemærke, at der er ved at blive bygget en dansksproget LL-model (Simonsen, 2024). I henhold til denne danske sprogmodel vil det forventes, at den vil have bedre forudsætninger til at forstå og dermed bedømme danskopgaver. Dette sker ud fra en forståelse af, at den danske sprogmodel udelukkende vil være trænet på dansk tekstdata, og derved kan argumenteres at være mere domænespecifik end den generalistiske model, som blev anvendt i indeværende studie. Gennem studiets problemanalyse blev det belyst, hvordan en af udfordringerne ved de generalistiske modeller kunne være, at de ikke på samme måde som de domænespecifikke modeller vil formå at forstå nuancerne i mindre sproggrene (Shehab et al., 2018). Det vil i henhold til karaktergivningen af en dansk skriftlig fremstilling herved kunne argumenteres, at den danske LL-model forventes at have elementer fra både de generalistiske og domænespecifikke LL-modeller, hvilket ud fra ovenstående diskussion kunne forventes at være fordelagtigt i karaktergivningen i folkeskolen ved dansk skriftlig fremstilling.

Samlet set kan det udledes, hvordan der forekommer en manglende viden om, hvorvidt en domænespecifik LL-model vil være mere fordelagtigt, når det kommer til at vurdere dansk skriftlige fremstillinger. Dette vil dermed være oplagt at undersøge yderligere i fremtiden med henblik på at finde frem til den LL-model, der som medbedømmer ved karaktergivningen vil kunne bidrage med den mest ensartede og pålidelige vurdering til gavn for støjreduktionen.

Med afsæt i diskussionens ovennævnte overvejelser samt resultaterne fra analysen vil studiets konklusion blive fremlagt i det følgende kapitel, hvor der vil blive fremsat en afsluttende policy-anbefaling.

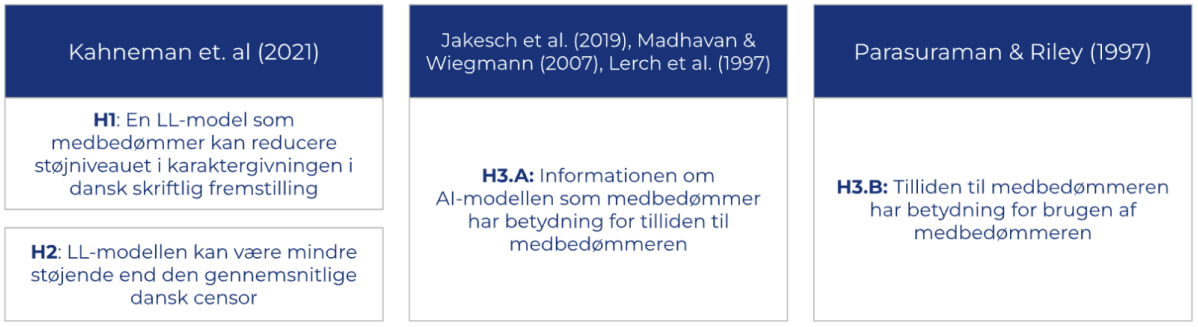
9. Konklusion

I studiets problemanalyse blev det tydeliggjort, hvordan karakterer i folkeskolen siden årtusindeskiftet har fået en stigende betydning, da karaktererne i dag er adgangsgivende for elevernes muligheder videre i uddannelsessystemet (jf. afsnit 4.1). Af den grund kan det anses som problematisk, at flere studier finder variabilitet ved karaktergivningen i folkeskolen, hvoriblandt den samme danskopgave blev vurderet til alle karakterer mellem 02 og 12 (jf. afsnit 4.2). Med formål om at reducere støjniveauet i karaktergivningen blev det fremlagt, hvordan teknologien bag machine learning har udviklet sig i en sådan grad, at den i dag blandt andet bruges til at give karakterer i USA (jf. afsnit 4.3). I kortlægningen af, hvorvidt LL-modeller havde potentiale til at styrke karaktergivningen, blev det afslutningsvist belyst, hvordan flere studier fandt tilliden til automationen som et afgørende element for, at brugen heraf forekom vellykket (jf. afsnit 3.4). Denne undren over, hvorvidt LL-modellerne kunne bruges som værktøj til at mindske støjniveauet i karaktergivningen, og hvordan tilliden har betydning herfor ledte frem til følgende forskningsspørgsmål:

I hvilken grad kan en LL-model som medbedømmer være med til at reducere støjen i karaktergivningen ved dansk skriftlig fremstilling i folkeskolen, og har tillid til modellen betydning for, om støjen kan reduceres?

Med henblik på at teste studiets forskningsspørgsmål blev der gennem Kahneman et al. (2021) opstillet to hypoteser om støjreduktionen ved brugen af en LL-model som medbedømmer. Endvidere blev der med afsæt i studiets problemanalyse samt Parasuraman & Riley (1997) opstillet en tredje hypotese om, hvorvidt informationen om AI-modellen som medbedømmer havde betydning for tilliden til medbedømmeren, og om denne tillid ville have betydning for brugen og dermed støjniveauet. Studiet tre hypoteser er visualiseret nedenfor, hvor det fremgår, at den tredje hypotese består af to delhypoteser:

Figur 9.1: Studiets hypoteser



Studiets hypoteser blev testet gennem et surveyeksperiment, hvor 93 dansklærere gennemførte eksperimentet. Her skulle de indledningsvist foretage deres første vurdering ved at fastsætte en karakter på normal vis, hvorefter de blev præsenteret for LL-modellens vurdering. Her blev halvdelen fortalt, at vurderingen var foretaget af en anden censor, mens den anden halvdel blev informeret om, at vurderingen var udarbejdet af en AI-model. Begge grupper blev efterfølgende bedt om at angive deres tillid til medbedømmeren, hvorefter de skulle foretage deres anden vurdering med mulighed for at revidere på baggrund af medbedømmerens vurdering.

Figur 9.2: Studiets surveydesign

9.1 Resultaterne fra analysen

Gennem studiets analyse blev det indledningsvist fundet, at der til trods for en ændring i vurderingspraksissen i 2023 stadig forekommer støj ved karaktergivningen i skriftlig dansk fremstilling i folkeskolen. Dette fremgik ved, at elevteksten ved den første vurdering samlet set blev bedømt til alle karakterer mellem 4 og 12. Foruden støjen ved den samlede karakter, blev det også fremlagt, hvordan der forekom støj i delkarakterne ved de seks vurderingsområder, hvor

elevteksten blev vurderet til alle karakterer mellem 00 og 12 ved vurderingsområdet 'ressourcer' (jf. afsnit 7.1).

I testen af studiets første hypotesen blev det fastslået, hvordan der forekom en signifikant støjreduktion fra 4,36 til 2,75 ved den samlede bedømmelse, når dansklærerne blev præsenteret for en medbedømmers vurdering. Mere konkret viste det sig, at 24 % af dansklærerne blev mindre støjende fra første til anden vurdering. Resultaterne viste ydermere, at denne støjreduktion var lige stor uanset, om dansklærerne blev fortalt, at medbedømmeren var en anden censor eller en AI-model (jf. afsnit 7.2). På baggrund heraf blev H1 hermed bekræftet.

Gennem studiets anden hypotese blev det vist, hvordan LL-modellen ved den samlede bedømmelse var mindre støjende end den gennemsnitlige dansklærer, hvormed LL-modellen vil kunne bidrage til en mere ensartet og pålidelig karaktergivning. Hertil er det værd at bemærke, at LL-modellen ved to ud af seks vurderingsområder var mere støjende end dansklærerne. Det ændrer dog ikke ved, at LL-modellen samlet set forekom mindst støjende, hvormed H2 overordnet set bekræftes (jf. afsnit 7.3).

Efter at have bekræftet hypoteserne om støjreduktion ved en LL-model som medbedømmer, leder det videre til, at det i studiets første delhypotese blev klarlagt, at informationen om, at medbedømmeren var en AI-model havde en signifikant negativ betydning for tilliden til medbedømmeren. Ved tillidsspørgsmålet om, hvorvidt medbedømmerens karaktergivning forekom at være fair, fandt studiet dog ikke nogen signifikant reduktion i tilliden (jf. afsnit 7.4). På baggrund af disse resultater blev H3.A bekræftet.

Afslutningsvist blev det gennem studiets anden delhypotese påvist, at tilliden har en negativ effekt på underbrugen. At der forekommer underbrug blandt dansklærerne med en lavere grad af tillid betyder med andre ord, at der kunne have været en større støjreduktion, hvis disse dansklærere havde brugt modellen mere i deres karaktergivning. Samtidig blev det kontrolleret, at sammenhængen mellem tillid og underbrug ikke var afhængig af interventionen. Omvendt var det dog ikke muligt at udlede, hvorvidt tilliden har betydning for overbrugen. Af den

grund kunne H3.B kun delvist bekræftes, hvorfor det samme gør sig gældende for H3 samlet set.

Hermed kan det med udgangspunkt i studiets forskningsspørgsmål konkluderes, at støjen i karaktergivningen kan reduceres ved at benytte en LL-model i karaktergivningen, hvor tilliden ses at have betydning for underbrugen af modellen. Det er dog væsentligt at bemærke, hvordan der trods en lavere tillid til LL-modellen som medbedømmer blev fundet en støjreduktion på 35 %, uanset om medbedømmeren var en anden censor eller en LL-model. Det kan hermed konkluderes, at der kan være større sandsynlighed for underbrug, når der er en bevidsthed om, at medbedømmeren er en LL-model, men at dette ikke nødvendigvis har betydning for den samlede støjreduktion.

9.2 Anbefalinger om anvendelsen af LL-modeller som medbedømmer i karaktergivningen

I forlængelse af studiets resultater blev muligheder og konsekvenser ved at benytte en LL-model ved karaktergivningen i folkeskolen diskuteret. Studiet har vist, hvordan en LLM som medbedømmer kan styrke kvaliteten af karaktergivningen. Det kommer til udtryk ved, at LL-modellen bidrager til en signifikant støjreduktion samtidig med, at modellens vurdering ydermere forekommer at være tættere på gennemsnittet end de fleste dansklærere.

På trods af at enkelte dansklærere har vist sig at være i risiko for at øge støjniveauet og dermed svække kvaliteten af deres vurdering, vil et forslag være at implementere en LLM som et understøttende værktøj ved karaktergivningen i den danske folkeskole. Dette skyldes, at karaktergivningens samlede ensartethed og pålidelighed øges, hvilket særligt blev tydeligt ved, at omtrent hver fjerde karaktergivning forbedres som følge af en støjreduktion. Initiativet om at implementere en LL-model som en beslutningsunderstøttende medbedømmer sigter dermed mod at forbedre kvaliteten af karaktergivningen og samtidig bevare såvel som styrke lærernes fagprofessionelle skøn.

Af studiets metode og diskussion kan det udledes, hvordan dilemmaerne ved implementeringen af en LL-model i karaktergivningen rummer tværfaglige dimensioner. Det anbefales derfor, at Børne- & Undervisningsministeriet indledningsvist nedsætter en tværfaglig arbejdsgruppe bestående af statslige beskikkede censorer, folkeskolelærere, datascientister, juridiske fagprofessionelle, relevant administrativt personale samt repræsentanter fra elev- og forældreorganisationer, som inden implementeringen kan diskutere, hvordan uhensigtsmæssigt brug og potentiel manglende tillid til modellen kan forebygges eksempelvis gennem uddannelse og informationskampagner.

Med henblik på at sikre en effektiv og ansvarlig implementering og anvendelse af kunstig intelligens i uddannelsessystemet, anbefales det først og fremmest, at LL-modellen indledningsvist bliver en del af et kontrolleret pilotprojekt. Her vil LL-modellen som medbedømmer kunne afprøves på terminsprøver eller lignende elevbesvarelser, der ikke udgør en direkte og adgangsgivende betydning for elevens videre muligheder i uddannelsessystemet. Gennem et pilotprojekt vil det videre være muligt at vurdere, hvordan dansklærerne tillidsmæssigt reagerer samt, hvorvidt der opstår modstand omkring brugen af LL-modellen. Endvidere vil det også muliggøre en grundig evaluering af AI-teknologiens effektivitet og indflydelse i et kontrolleret miljø før, at denne implementeres bredere. Ydermere anbefales det også, at dansklærerne i pilotprojektet kan give feedback på modellens begrænsninger, hvorved det inden en bredere implementering vil være muligt at træne LL-modellen yderligere på de områder, hvor dansklærerne oplever mangler ved modellen. Herigennem bliver samarbejdet mellem datascientister og dansklærerne afgørende for, hvorvidt det i sidste ende er muligt at skabe en brugbar LL-model, der kan indgå som en beslutningsunderstøttende medbedømmer i karaktergivningen.

Afslutningsvis skal det understreges, at der ved implementeringen af en LL-model i karaktergivningen kræver en række ressourcer. Foruden opbygning af modellen, vedligehold af modellen, uddannelse af fagprofessionelle og information om LL-modellen til den bredere befolkning, så må det også forventes, at brugen af en medbedømmer formentligt ville betyde, at censorerne vil skulle bruge et par ekstra minutter per karaktergivning. Sammenlignet med en to-bedømmerordningen, hvor der som minimum havde været en fordobling af

tids- og ressourceforbruget, kan det argumenteres, at forøgelsen af ressourcer vil være mere begrænset ved en bedømmerordning med en LL-model som medbedømmer. I tilfælde af, at censorerne ikke tildeles ekstra tid til den enkelte vurdering, vil risikoen for modstand mod en bedømmelsespraksis med en LLM forventeligt blive større. Endvidere foreligger der også økonomiske og sociale argumenter for, at et yderligere arbejdspres ved karaktergivningen bør undgås. Dette ændrer dog ikke på faktummet om, at støjen i karaktergivningen skal reduceres.

Dette studie har vist, hvordan 24 % af dansklærerne reducerer deres støj, hvormed en mere ensartet og pålidelig og dermed retfærdig karaktergivning kan opnås med en LL-model som medbedømmer. Derfor anbefales det herved at igangsætte ovennævnte proces, som på sigt skal lede hen mod en implementering af en LL-model, der som medbedømmer kan understøtte den fagprofessionelle bedømmer ved karaktergivningen i folkeskolen. Dette anses som afgørende med henblik på at sikre, at karaktergivningen i folkeskolen karakteriseres ved en større ensartethed og pålidelighed, og det dermed ikke vil være tilfældigheder, som vil være afgørende for elevernes fremtidige muligheder i uddannelsessystemet.

10. Litteraturliste

Aarhus Universitet (n.d.). *Hawthorne-effekten*. Metodeguiden. Available: https://metodeguiden.au.dk/hawthorne-effekten [08.04.2024]

Agresti, Alan. (2018). *Statistical methods for the social sciences*. (5. eds). Pearson.

Ajay HB, Tillett PI, Page EB (1973). *Analysis of essays by computer (AEC-II)* (No. 8-0102). Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education, National Center for Educational Research and Development. Available: https://eric.ed.gov/?id=ED093342 [14.05.2024]

Akhtar, M. (2021). *Transparens i offentliges brug af maskinlæring*. I M. Jarlner & K. Escherich (red.), *Fra velfærdsstat til overvågningsstat: algoritmernes magt i den offentlige forvaltning*. (1. eds). Djøf Forlag.

Alake, Richmond (2022) "*An Introduction To Gradient Descent and Backpropagation In Machine Learning Algorithms*". Available: https://towardsdatascience.com/an-introduction-to-gradient-descent-and-backpropagation-in-machine-learning-algorithms-a14727be70e9 [22.04.2023]

Andersen, J. G. (2004). Et ganske levende demokrati. Aarhus universitetsforlag.

Andersen, L. B., Hansen, K. M., & Klemmensen, R. (Eds.). (2010). *Metoder i statskundskab*. Rosinante&Co.

Andersen, Lotte Bøgh; Binderkrantz, Anne Skorkjær & Hansen, Kasper Møller. (2020). *Forskningsdesign* i Hansen, Kasper Møller; Andersen, Lotte Bøgh & Hansen, Sune Welling (2020). *Metoder i statskundskab*. (3. udg). København: Hans Reitzels forlag.

Andersen, Lotte. B., Laustsen, L. og M. Cecchini. (2020). *Forskningskriterier*. I K. M. Hansen, L. B. Andersen & S. W. Hansen (red). *Metoder i Statskundskab*. (3.udg). København: Hans Reitzels Forlag.

Hansen, K., Andersen, L. og S. W. Hansen. (2020) (red). *Metoder i Statskundskab* (3.udg). København : Hans Reitzels Forlag.

Andreasen, Karen Egedal. (2020). *FAQ - Om karakterer*. (1. udg). København: Hans reitzels forlag,

Bach, A., Svejgaard, J., & Hjort, F. (2019). *Maskinlæring som politologisk værktøj*. Politica 51(2), 168-186. Available: *https://doi.org/10.7146/politica.v51i2.131144* [22.01.2024]

Bartz, A. E. (1999: 184). *Basic statistical concepts*. (4 eds.). Upper Saddle River, NJ: Prentice Hall.

BEK nr 1068. (2015). *Bekendtgørelse om uddannelsen til professionsbachelor som lærer i folkeskolen*. Uddannelses- og Forskningsministeriet. Available: https://www.retsinformation.dk/eli/lta/2015/1068 [14.05.2024]

Bekendtgørelse af lov om gennemsigtighed og åbenhed i uddannelserne m.v.. Børne- og Undervisningsministeriet. LBK nr. 880 af 19/09/2005, Retsinformation. (2005). Available: https://www.retsinformation.dk/eli/lta/2005/880 [14.05.2024]

Bhaskar, R. (1979). *The Possibilities of Naturalism*. *A Philosophical Critique of the Contemporary Human Sciences*. Brighton: Harvester Press.

Bhaskar, R., Archer, M., Collier, A., Lawson, T., & Norrie, A. (Eds.). (2013). *Critical realism: Essential readings*. Routledge

Bjerva, Johannes & Lindsay, Euan. (2024) *AI skal sikre mere feedback til studerende*. VBN. Available: https://vbn.aau.dk/da/clippings/ai-skal-sikre-mere-feedback-til-studerende [22.01.2024]

Blete, Marie-Alice. (2023). *LLMs: Determinism & Randomness*. Medium https://medium.com/@mariealice.blete/llms-determinism-randomness-36d3f3f1f793 [14.05.2024]

Blom-Hansen, Jens & Serritzlew, Søren (2014) *Endogenitet og eksperimenter - forskningsdesignet som løsning*. Politica, Årg. 46 Nr. 1 (2014): Empiriske metoder til studiet af kausalitet i statskundskaben: Available: https://doi.org/10.7146/politica.v46i1.69758 [14.05.2024]

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). *The human black-box: The illusion of understanding human better than algorithmic decision-making*. Journal of Experimental Psychology: General, 151(9), 2250–2258. https://doi.org/10.1037/xge0001181[14.05.2024]

Bourdieu, Pierre (1986). *The forms of capital*. I Richardson, J. (red.). Handbook of Theory and Research for the Sociology of Education. Westport, CT: Greenwood.

Brinkmann, Svend & Tanggaard, Lene (red.) (2020) *Kvalitative metoder: En grundbog*, (3. udg) Hans Reitzels Forlag

Brock, M. N. (1990). *Can the computer tutor? An analysis of a disk-based text analyzer*. System, 18(3), 351-359.

Bryman, Alan. (2016). *Social Research Methods*. (5. udg). Oxford university press.

Bækgaard, Martin; Dahlgaard, Jens Olav & Hansen, Kasper Møller (2020) *Det eksperimentelle design*. i Hansen, Kasper Møller; Andersen, Lotte Bøgh & Hansen, Sune Welling (2020) *Metoder i statskundskab*. Hans reitzels forlag, 3. udg

Børne- og Undervisningsministeriet (n.d., a.). *Karakterer på 7-trins-skalaen*. Available:https://www.uvm.dk/uddannelsessystemet/7-trins-skalaen/karakterer-paa-7-trins-skalaen [23.01.2024]

Børne- og undervisningsministeriet (n.d., b) *Hvordan bliver man statsligt beskikket censor i folkeskolen*. Available: https://www.uvm.dk/folkeskolen/folkeskolens-proever/bedoemmelse-og-censur/information-til-censorer/hvordan-bliver-man-statsligt-beskikket-censor-i-folkeskolen [23.01.2024]

Børne- og Undervisningsministeriet (2022). *Vejledning til folkeskolens prøver i dansk i 9. klasse*. Styrelsen for Undervisning og Kvalitet. Available: https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf22/okt/221028-vejledning-til-folkeskolens-proever-i-dansk-i-9--klasse.pdf [08.02.2024]

Børne- og Undervisningsministeriet (2023, a) *Adgangskrav til uddannelserne til studentereksamen*. Available: https://www.uvm.dk/gymnasiale-uddannelser/adgang-og-optagelse/adgang/adgang-til-uddannelserne-til-studentereksamen [05.02.2024]

Børne- og Undervisningsministeriet (2023, b). *Faktaark - økonomi for forberedt på fremtiden II*. Available: https://www.uvm.dk/-/media/filer/uvm/aktuelt/pdf23/okt/231011-faktaark---oekonomi-i-forberedt-paa-fremtiden-ii.pdf [23.01.2024]

Børne- og Undervisningsministeriet (2023, c). *Klager over prøver. Available:* https://www.uvm.dk/folkeskolen/folkeskolens-proever/bedoemmelse-og-censur/klager-over-proever [23.01.2024]

Børne- og Undervisningsministeriet (2023, d). *Praktiske forhold i forbindelse med skriftlig og mundtlig censur ved folkeskolens prøver*. Available: https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf23/feb/230213-praktiske-forhold-i-forbindelse-med-skriftlig-og-mundtlig-censur-2023.pdf [15.05.2024]

Børne- og Undervisningsministeriet (2024) *Aftale mellem regeringen (Socialdemokratiet, Venstre og Moderaterne) og Liberal Alliance, Det Konservative Folkeparti, Radikale Venstre og Dansk Folkeparti om folkeskolens kvalitetsprogram – frihed og fordybelse*. uvm.dk , 19. marts 2024. Availble: https://www.uvm.dk/-/media/filer/uvm/aktuelt/pdf24/mar/240320-aftale-om-folkeskolens-kvalitetsprogram-–-frihed-og-fordybelse.pdf [05.04.2024]

Campbell, D. T., & Kenny, D. A. (1999). *A Primer on Regression Artifacts*. Guilford Press.

Canger, T., Krøjer, J., & Padovan-Özdemir, M. (2021). Tillid og mistillid i velfærdsprofessionelle relationer. Dansk Paedagogisk Tidsskrift, (1), 5-9.

Chandler, Dana, Steven D. Levitt & John A. List (2011). *Predicting and preventing shootings among at-risk youth*. The American Economic Review 101 (3): 288–292.

Chauhan, Nagesh Singh (2020) *Introduction to RNN and LSTM*. Available: https://www.theaidream.com/post/introduction-to-rnn-and-lstm [23.01.2024]

Coleman, James S. (1990). *Foundations of Social Theory*. Cambridge, MA.: Harvard University Press,

Cremer DD, Kasparov G (2021) AI should augment human intelligence, not replace it. Harvard Bus. Rev. (March 18), https://hbr.org/2021/03/ai-should-augment-human-intelligence-not-replace-it [08.02.2024]

Cuneo, T. (2020). *A Dictionary of Ethics*. Oxford University Press.

Creswell, JW. (2009). *Research Design: Qualitative, Quantitative, and Mixed Methods Approaches* (3. Eds). Thousand Oaks, Cal.: Sage Publications, pp. 3-21, 73-80 & 129-143.

Crossley, S. A., Bradfield, F., & Bustamante, A. (2019). *Using human judgments to examine the validity of automated grammar, syntax, and mechanical errors in writing*. Journal of Writing Research, 11(2), 251-270.

Dafoe, A., Zhang, B., & Caughey, D. (2018). *Information Equivalence in Survey Experiments*. Political Analysis, 26(4), 399–416. Available: https://www.cambridge.org/core/journals/political-analysis/article/information-equivalence-in-survey-experiments/8D134C6387CD7D845249B0712775AB79# [30.05.2024]

Danermark, B. et al. (1997): *Att förklara samhället*. Lund: Studentlitteratur

Datacamp.com (2019) *Introduction to Factor Analysis in Python*. Available: https://www.datacamp.com/tutorial/introduction-factor-analysis [14.05.2024]

Deci, Edward og Ryan, Richard (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. American psychologist. Vol. 55, No. 1, 68-78.

De-Arteaga, Maria; Fogliato, Riccardo & Chouldechova, Alexandra (2020). *Humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores*. CHI 2020, April 25–30, 2020, Honolulu, HI, USA
https://dl.acm.org/doi/pdf/10.1145/3313831.3376638 [14.05.2024]

Dijkstra, JJ. (1999). in Madhavan P. & Wiegmann D. A. (2007) *Similarities and differences between human–human and human–automation trust: an integrative review*, Theoretical Issues in Ergonomics Science, 8:4, 277-301,Available:

https://www-tandfonline-com.zorac.aub.aau.dk/doi/pdf/10.1080/14639220500337708 [14.05.2024]

DI (2023). *To elever i hver klasse består ikke dansk og matematik ved folkeskolens afgangseksamen*. Available:
https://www.danskindustri.dk/arkiv/analyser/2023/9/to-elever-i-hver-klasse-bestar-ikke-dansk-og-matematik-ved-folkeskolens-afgangseksamen/ [14.05.2024]

DLF (2023). *Aktive medlemmer af DLF 2023*. (Danmarks Lærerforening). Available:
https://www.dlf.org/media/13407106/aktive-medlemmer-af-dlf-2023.pdf
[14.05.2024]

Dolin, Jens og K. Nielsen, B. Rangvid. (2018). *Følgegruppen for én bedømmer ved folkeskolen prøver*. Børne- og undervisningsministeriet. Available: https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf18/apr/180425-rapport-fra-foelgegruppen-for-een-bedoemmer-ved-fp.pdf [18.01.2024]

Druckman, J. N., Green D. P., Kuklinski J. H. & Lupia A. et al. (2006). *The Growth and Development of Experimental Research in Political Sciences*. American Political Science Review. 100(4): 627-635.

Jensen, Magnus Stenaa (2024). *Udviklingen af AI har aldrig gået stærkere, men hvor langt er vi egentlig?*. DTU. Available: https://www.dtu.dk/newsarchive/2024/03/udviklingen-har-aldrig-gaaet-staerkere-men-hvor-langt-er-vi-egentlig

Du, Mark (2023) *Machine vs. human, who makes a better judgment on innovation? Take GPT-4 for example*. PubMed Central. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10482032/ [14.05.2024]

Duan, Y., Edwards, J. S., & Dwivedi, Y. K. (2019). *Artificial intelligence for decision making in the era of Big Data–evolution, challenges and research agenda*. International journal of information management, 48, 63-71.

Dzindolet, M.T.; Peterson, S. A.; Pomranky, R. A.; Pierce, L. G.; an Beck H. P. (2003). *The role of trust in automation reliance. International Journal of Human-Computer*. Studies 58:697–719. Available: https://kbdk-aub.primo.exlibrisgroup.com/discovery/fulldisplay?docid=cdi_openaire_primary_doi_833a31180cec06430a619b89d93d117f&context=PC&vid=45KBDK_AUB:AUB&lang=da&search_scope=MyInst_and_CI&adaptor=Primo%20Central&tab=Everything&query=any,contains,Dzindolet,%20%20M.%20%20T.;%20%20Peterson,%20%20S.%20%20A.;%20%20Pomranky,%20%20R.%20%20A.;%20Pierce,%20%20L.%20%20G.;%20%20and%20%20Beck,%20%20H.%20%20P.%20%202003.%20%20The%20%20role%20%20of%20trust%20in%20automation%20reliance.%20International%20Journal%20of%20Human-Computer%20Studies%2058:697–719&offset=0 [14.05.2024]

Erjavec, N. (2011). *Tests for Homogeneity of Variance. In: Lovric, M. (eds) International Encyclopedia of Statistical Science*. Springer, Berlin, Heidelberg. Available: https://doi-org.zorac.aub.aau.dk/10.1007/978-3-642-04898-2_590 [14.05.2024]

Europa-Parlamentet (2024) *Artificial Intelligence Act: MEPs adopt landmark law*. Available: https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law [14.05.2024]

Europaparlamentet (2023) *EU AI Act: first regulation on artificial intelligence*. Available: https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence [14.05.2024]

European value survey (2021) *Source questionnaire round 10 2020/2021*. Available: https://stessrelpubprodwe.blob.core.windows.net/data/round10/fieldwork/source/ESS10_source_questionnaires.pdf [14.05.2024]

Fan, X., Oh, S., McNeese, M., Yen, J., Cuevas, H., Strater, L., & Endsley, M. R. (2008). *The influence of agent reliability on trust in human-agent collaboration. ACM International Conference Proceeding Series*, 369. Available: https://dl-acm-org.zorac.aub.aau.dk/doi/epdf/10.1145/1473018.1473028 [14.05.2024]

Faraj, S., Pachidi, S., & Sayegh, K. (2018). Working and organizing in the age of the learning algorithm. Available: Information and Organization, 28(1), 62-70.

Freed, S. (2019). *AI and human thought and emotion*. CRC Press. (EBOG)

Folkeskoleloven (2014). *Bekendtgørelse af lov om folkeskolen. LBK nr 665 af 20/06/2014*. Børne- og Undervisningsministeriet. Retsinformation. Available: https://www.retsinformation.dk/eli/lta/2014/665 [18.01.2024]

Folketingets ombudsmand (2023). God forvaltningsskik. Available: https://www.ombudsmanden.dk/find-viden/myndighedsguiden/generel-forvaltningsret/1-god-forvaltningsskik [27.06.2024]

Galton, Francis (1886). *Regression towards mediocrity in Hereditary Stature*. Anthropological Miscellanea. Available:

https://lru.praxis.dk/Lru/microsites/hvadermatematik/hem3download/kap8_QR5b_ekstra_galton_regression_artiklen.pdf [26.02.2024]

Galton, Francis (1907). *Vox populi*. Nature 75 450-451. (March 7, 1907.). Available: https://www.nature.com/articles/075450a0 [27.02.2024]

Gao R, Saar-Tsechansky M, De-Arteaga M, Han L, Lee MK, Lease M (2021) *Human-AI collaboration with bandit feedback*. Pre-print, submitted May 22, https://arxiv.org/abs/2105.10614 [14.05.2024]

Géron, A. (2023). *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow* (3. eds.). O'Reilly.

Gerring. J. (1999) *What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences*. Polity , Spring, 1999, Vol. 31, No. 3 (Spring, 1999), pp. 357-393. https://www-jstor-org.zorac.aub.aau.dk/stable/pdf/3235246 [14.05.2024]

Ghazizadeh, M., Lee, J. D., & Boyle, L. N. (2012). *Extending the Technology Acceptance Model to assess automation*. *Cognition Technology and Work*, *14*(39), 49. https://doi.org/10.1007/s10111-011-0194-3

Giselmar A. J.; Hemmert, Laura M.; Schons, Jan Wieseke and Heiko Schimmelpfennig (2018) *Log-likelihood-based Pseudo-R2 in Logistic Regression: Deriving Sample-sensitive Benchmarks*. Available: https://journals-sagepub-com.zorac.aub.aau.dk/doi/pdf/10.1177/0049124116638107 [14.05.2024]

Glikson E, Woolley AW (2020) *Human trust in artificial intelli-gence: Review of empirical research*. Acad. Management Ann. 14(2):627 – 660. https://www.researchgate.net/profile/Anita-Woolley/publication/340605601_Human_trust_in_artificial_intelligence_Review_of_empirical_research_Academy_of_Management_Annals_in_press/links/5e9460ec92851c2f529c406e/Human-trust-in-artificial-intelligence-Review-of-empirical-research-Academy-of-Management-Annals-in-press.pdf [14.05.2024]

Gronemann, C., & Jespersen, M. K. (2015) *TILLID I KRISE*. Available: https://rucforsk.ruc.dk/ws/portalfiles/portal/57645803/Tillid_i_krise_-_finanskrisensp%C3%A5virkning_af_social_og_institutionel_tillid_i_Europa.pdf [14.05.2024]

Guadagnoli, E., & Velicer, W. F. (1988: 266). *Relation of sample size to the stability of component patterns*. Psychological bulletin, 103(2).

Guan, X., & Burton, H. (2022). *Bias-variance tradeoff in machine learning: Theoretical formulation and implications to structural engineering applications*. In *Structures* (Vol. 46, pp. 17-30). Elsevier.

Gunes, E. & C. Florczak (2023). *Multiclass Classification of Policy Documents with Large Language Models*. VBN. Available: https://vbn.aau.dk/da/publications/multiclass-classification-of-policy-documents-with-large-language

Hannah, L., Jang, E. E., Shah, M., & Gupta, V. (2023). *Validity Arguments for Automated Essay Scoring of Young Students' Writing Traits*. Language Assessment Quarterly, 20(4-5), 399-420.

Hansen, Jesper & Trummers, Lars. (2020). A Systematic Review of Field Experiments in Public Administration. Public Administration Review. 80.

Hariri, Jacob Gerner. (2012). *Kausal inferens i statskundskaben.* Politica, 44. årg. nr. 2 2012, 184-201 Available: https://politica.dk/fileadmin/politica/Dokumenter/politica_44_2/hariri.pdf

Hartman, E. (2021). *Generalizing Experimental Results.* In J. N. Druckman & D. P. Green. (Eds.), *Advances in Experimental Political Science* (pp. 385–410). Cambridge: Cambridge University Press.

Hendrickson, A. T., Perfors, A., Navarro, D. J., & Ransom, K. (2019). *Sample size, number of categories and sampling assumptions: Exploring some differences between categorization and generalization*. Cognitive Psychology, 111, 80-102.

Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). *A fast learning algorithm for deep belief nets*. Neural computation, 18(7), 1527-1554.

Hjorth-Trolle, Anders & Jonassen, Anders Bruun Jonassen (2023) *Adgangskrav på de gymnasiale uddannelser og optag på erhvervsuddannelserne*. Available: https://rockwoolfonden.s3.eu-central-1.amazonaws.com/wp-content/uploads/2023/06/RFF-Indlaeg_RFF-Indlaeg_Reform2019.pdf?download=true [14.05.2024]

Hoff, K. A., & Bashir, M. (2015). *Trust in automation: Integrating empirical evidence on factors that influence trust*. Human Factors, 57(3), 407–434. https://journals-sagepub-com.zorac.aub.aau.dk/doi/pdf/10.1177/0018720814547570 [14.05.2024]

Holm-Larsen, Signe & Niels Plischewski. (2017). *Om karaktergivning*. Dafola. 2. udgave.

Infomedia (2024) Søgning ('karakterer') AND ('folkeskole') i perioden 01-01-2000 til 01-01-2005 sammenholdt med 01-01-2019 til 01-01-2024. [20.02.2024]

Ingemann, Jan holm. (2017). *Videnskabsteori - For økonomi, politik og forvaltning*. (1, udg, 3. opl). Samfundslitteratur.

Iversen, Gudmund (2014). *Analysis og variance*. International Encyclopedia of Statistical Science. Available: https://link-springer-com.zorac.aub.aau.dk/referenceworkentry/10.1007/978-3-642-04898-2_117 [14.05.2024]

Jakesch, M., French, M., Ma, X., Hancock, J. T., & Naaman, M. (2019). *AI-mediated communication: How the perception that profile text was written by AI affects trustworthiness*. In Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-13).

Jensen, M. B., Bahnsen, C. H., Nasrollahi, K., & Moeslund, T. B. (2018). *Deep Learning: Et gennembrud inden for kunstig intelligens*. Aktuel Naturvidenskab, (2), 8-13. Available:

https://aktuelnaturvidenskab.dk/fileadmin/Aktuel_Naturvidenskab/nr-2/AN2-2018deep-learning.pdf [14.05.2024]

Johansson, Mikkel Linnemann (2023). *Potentialet for kunstig intelligens er enormt*. Available: https://www.sdu.dk/da/nyheder/potentialet-for-kunstig-intelligens-er-enormt [14.05.2024]

Jordan, M. I. og Mitchell, T. M. (2015) *Machine learning: Trends, perspectives, and prospects*. Available via AUB https://aub.page.link/DLX6 [14.05.2024]

Kahneman, Daniel & Amos Tversky (1974). *Judgment under Uncertainty: Heuristics and Biases*. *Science*. Vol. 185, Issue 4157, pp. 1124-1131. Available: https://www2.psych.ubc.ca/~schaller/Psyc590Readings/TverskyKahneman1974.pdf [26.02.2024]

Kahneman, D.; Sibony, O. & Sunstein, C. (2021). *Støj - Sådan træffer du bedre beslutninger*. 1. eds. Lindhardt og Ringhof

Kalirane, Mbali (2023). *Gradient Descent vs. Backpropagation: What's the Difference?*. Available: https://www.analyticsvidhya.com/blog/2023/01/gradient-descent-vs-backpropagation-whats-the-difference/ [14.05.2023]

Kazemnejad, Amirhossein (2019). *Transformer Architecture: The Positional Encoding Let's use sinusoidal functions to inject the order of words in our model*. Available: https://kazemnejad.com/blog/transformer_architecture_positional_encoding/ [14.05.2024]

KL (2017). *God adfærd i det offentlige*. Available: https://www.modst.dk/media/17831/god-adfaerd-i-det-offentlige_web.pdf [14.05.2024]

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan og Ziad Obermeyer (2015). *Prediction policy problems*. American Economic Review 105 (5): 491-495.

KL (2023). Kunstig intelligens skal løfte fremtidens velfærd. Available: https://www.kl.dk/forsidenyheder/2023/december/kunstig-intelligens-skal-loefte-fremtidens-velfaerd

Kotter, John P. (2012). Leading change. Harvard Buisnes Review Press. Boston. Available: https://web-p-ebscohost-com.zorac.aub.aau.dk/ehost/ebookviewer/ebook/bmxlYmtfXzY3NTE5NV9fQU41?sid=f7809b92-d9eb-4df6-905d-c143c6bdd368@redis&vid=0&format=EB&rid=1

Kulshrestha, Ria (2020). *Transformers, Towards Data Science*. Available: https://towardsdatascience.com/transformers-89034557de14 [14.05.2024]

Kumar, Ranjit. (2011). "Research Methodology - a step-by-step guide for beginners". 3. Udgave. SAGE Publications. Available: http://www.sociology.kpi.ua/wp-content/uploads/2014/06/Ranjit_Kumar-Research_Methodology_A_Step-by-Step_G.pdf [22.12.2020]

Langroudi, H. F., Merkel, C., Syed, H., & Kudithipudi, D. (2019, July). *Exploiting randomness in deep learning algorithms*. In 2019 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. (2022). *To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis*. Organization Science, 33(1), 126-148. https://doi.org/10.1287/orsc.2021.1549 [14.05.2024]

LeCun, Y., Bengio, Y., & Hinton, G. (2015). *Deep learning*. nature, 521(7553), 436-444.

Lee, J. D., & See, K. A. (2004). *Trust in automation: Designing for appropriate reliance. Human Factors*. The Journal of the Human Factors and Ergonomics Society, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392 [14.05.2024]

Lerch, F. J.; Preitula, M. J. & Kulik, C. T. (1997). *The turing effect: the nature of trust in expert systems advice* I *Expertice in context: Human and Machine*. Feltovich, P. J. & Ford K. M. (red..) pp 417-448. Cambridge: The MIT Press.

Lundberg S. & Lee, I. (2017). *A Unified Approach to Interpreting Model Predictions*. Available: https://doi.org/10.48550/arxiv.1705.07874 [27.04.2023]

Noor, S., Tajik, O., & Golzar, J. (2022). Simple random sampling. International Journal of Education & Language Studies, 1(2), 78-82.

Madhavan, P., & Wiegmann, D. A. (2007). *Similarities and differences between human–human and human–automation trust: an integrative review*. Theoretical Issues in Ergonomics Science, 8(4), 277–301. Available: https://www-tandfonline-com.zorac.aub.aau.dk/doi/pdf/10.1080/14639220500337708 [14.05.2024]

Mainz, Pernille. (2024). *Karakterkrav på 6 til gymnasiet vil udelukke hver femte elev*. Politiken. 01/05/2024. Available: https://apps-infomedia-dk.zorac.aub.aau.dk/mediearkiv/link?articles=ea34e427 [14.05.2024]

Magerman, D. (2022). *Systemic Bias in Data Makes AI a Stereotype Machine*. 09/02/2022. Inside BIGDATA. Available: https://insidebigdata.com/2022/02/09/systemic-bias-in-data-makes-ai-a-stereotype-machine/ [14.05.2024]

Mayer, R. C., Davis, J. H., & Schoorman, D. F. (1995). *An integrative model of organizational trust*. The Academy of Management Review, 20(3), 709–734.

McDermott, R. (2002). *Experimental methods in political science*. Annual Review of Political Science, 5(1), 31-61.

Meehl, Paul E. & Grove, William M. (June 1996). *Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction*

procedures: the clinical–statistical controversy. Psychology, Public Policy, and Law. 2 (2): 293–323. Available: https://meehl.umn.edu/sites/meehl.umn.edu/files/files/167grovemeehlclinstix.pdf [14.05.2024]

Merrick, L., & Taly, A. (2020). *The Explanation Game: Explaining Machine Learning Models Using Shapley Values*. I A. Holzinger, P. Kieseberg, A. Tjoa, & E. Weippl (red.) *Machine Learning and Knowledge Extraction*. CD-MAKE 2020. Lecture Notes in Computer Science, 12279. Springer. Available: https://doi.org/10.1007/978-3-030-57321-8_2 [14.05.2024]

Miller, S. (2018) *AI: Augmentation, mere end automatisering*. Asian Management Insights, (2018) , s. 1-20

Mittal, A. (2023) The Black Box Problem in LLMs: Challenges and Emerging Solutions. Unite.AI. Available: https://www.unite.ai/the-black-box-problem-in-llms-challenges-and-emerging-solutions/ [14.05.2024]

Mosier, K.L. and Skitka, L.J. (1996). H*uman decision makers and automated decision aids: made for each other?* In R. Parasuraman and M. Mouloua (Eds). *Automation and Human performance: Theory and Applications*, , pp. 201–220 (Mahwah, NJ: Lawrence Erlbaum Associates). Available: https://www.researchgate.net/publication/230601064_Human_Decision_Makers_and_Automated_Decision_Aids_Made_for_Each_Other [14.05.2024]

Motzfeld, Hanne Marie (2021) *Når algoritmen får det sidste ord*. I Jarlner, Michael & Escherich(red.) (2021) F*ra velfærdsstat til overvågningsstat* (1. eds), Djøf forlag.

Muszynski, Marek. (2023). *Attention checks and how to use them: Review and practical recommendations*. Ask Research and Methods. 32. 3-38. 10.18061/ask.v32i1.0001. Available: https://www.researchgate.net/publication/376340288_Attention_checks_and_how_to_use_them_Review_and_practical_recommendations [14.05.2024]

Mutz, Diana C.. *Population-Based Survey Experiments*, Princeton: Princeton University Press, 2011. Available: https://doi-org.zorac.aub.aau.dk/10.1515/9781400840489 [14.05.2024]

Møller, Marie Østergaard; Hermansen, Anne & Møller, Simon Østergaard (2021). *Faglighed og styring - Fagprofessionel dømmekraft på socialområdet*. Djøfs forlag.

Nannestad, Peter (2009) *Rational choice/public choice* i Kaspersen, Lars bo & Loftager, Jørn (red.). *Klassisk og moderne politisk teori* (2009) (1.eds, 4. opl.), Hans Reitzels forlag.

Nedergaard, Peter (2022). *Tillid til institutioner er nødvendig*. Københavns Universitet.
Det Samfundsvidenskabelige Fakultet. Available: https://samf.ku.dk/presse/kronikker-og-debat/2022/tillid-til-institutioner-er-noedvendig/ [15.05.2024]

Nisgaard, A., & Bach, A. H. (2023). *Over 1000 forskere og techmoguler i opråb; sæt øjeblikkeligt udviklingen af kunstig intelligens på pause*. 29/03/2023. Danmarks Radio. Available at: https://www.dr.dk/nyheder/viden/teknologi/over-1000-forskere-og-techmoguler-i-opraab-saet-oejeblikkeligt-udviklingen [14.05.2024]

Nova, M. (2018). *Utilizing Grammarly in evaluating academic writing: A narrative research on EFL students' experience*. Journal of English Education, 7(1), 80-97.

Obermeyer, Ziad & Emanuel, Ezekiel J. (2016). P*redicting the future: Big data, machine learning, and clinical medicine*. The New England Journal of Medicine 375 (13): 1216-1219.

Olesen, Christian Gylling & Hassle, Peter (2011), *Tillid i virksomheder - samarbejde, kerneopgaver og social kapital* i Hegedahl, Paul & Svendsen, Gunnar (red.), 2011 *Tillid - samfundets fundament*: *Teorier, tolkninger, cases*. kapitel 13. Syddansk Universiteteforlag.

OpenAI (n.d). *API reference*. https://platform.openai.com/docs/api-reference/chat/create#chat-create-temperature [14.05.2024]

Ophavsretsloven (2014). *Bekendtgørelse af lov om ophavsret*. Retsinformation.dk. Available: https://www.retsinformation.dk/eli/lta/2014/1144 [14.05.2024]

Pal, S., Bhattacharya, M., Lee, S. S., & Chakraborty, C. (2024). *A domain-specific next-generation large language model (LLM) or ChatGPT is required for biomedical engineering and research*. Annals of Biomedical Engineering, 52(3), 451-454.

Parasuraman, R., & Riley, V. (1997). *Humans and automation: Use, misuse, disuse, abuse*. Human Factors: The Journal of the Human Factors and Ergonomics Society, 39(2), 230–253. Available: https://www.proquest.com/docview/216444294/fulltextPDF?parentSessionId=eL7n1OouploHwqTv4%2B41rjwktFzfspIkSeCR1gHIP5g%3D&pq-origsite=primo&accountid=8144&sourcetype=Scholarly%20Journals [14.05.2024]

Pattam, Aruna (2023) *Generativ AI: Episode #9: The Rise of Domain-Specific Large Language Models*. Available: https://arunapattam.medium.com/generative-ai-the-rise-of-domain-specific-large-language-models-ec212290c8a2 [14.05.2024]

Pearson, K. (1924). *The Life , Labours and Letters of Francis Galton : Vol. 2, Researches of Middle Life*. Cambridge: Cambridge Univ. Press,

Pearson, K. (1930). *The Life , Labours and Letters of Francis Galton*. Vol. 3B, Characterization , Especially by Letters , & Index. Cambridge Univ. Press, Cambridge.

Petersen, M.B.; Slothuus R.; Stubager R. og Togeby, L. (2007). *Eksperimenter- Et redskab i politologiens værktøjskasse*. Politica, 39(1): 530–550.

Prøvebekendtgørelsen (2019). *Bekendtgørelse om folkeskolens prøver*. BEK nr 1128 af 14/11/2019. Børne- og Undervisningsministeriet. Available: https://www.retsinformation.dk/eli/lta/2019/1128 [18.01.2024]

Puranam P (2021). *Human – AI collaborative decision-making as an organization design problem*. J. Organ. Design 10(2021):75. Available: https://link.springer.com/content/pdf/10.1007/s41469-021-00095-2 [14.05.2024]

Putnam, R (2000). *Bowling Alone: The Collapse and Revival of American Community*. New York: Simon and Schuster.

Ramchurn, S. D., Wu, F., Jiang, W., Fischer, J. E., Reece, S., Roberts, S., … Jennings, N. R. (2016). *Human–agent collaboration for disaster response*. Autonomous Agents and Multi-Agent Systems, 30(1), 82–111. https://doi.org/10.1007/s10458-015-9286-4 [14.05.2024]

Ramesh, D., & Sanampudi, S. K. (2022). *An automated essay scoring systems: a systematic literature review*. Artificial Intelligence Review, 55(3), 2495-2527.

Regeringen (2014). *Aftale om Bedre og mere attraktive erhvervsuddannelser. Socialdemokraterne, Radikale Venstre, Venstre, Dansk Folkeparti, Socialistisk Folkeparti, Konservative Folkeparti og Liberal Alliance*. Available: https://www.uvm.dk/-/media/filer/uvm/udd/erhverv/pdf19/190220-aftale-om-bedre-og-mere-attraktive-erhvervsuddannelser.pdf [14.05.2024]

Regeringen (2023). *Forberedt på fremtiden II. Frihed og fordybelse - et kvalitetsprogram for folkeskolen*. Børne- og Undervisninsministeriet. Available: https://www.uvm.dk/-/media/filer/uvm/aktuelt/pdf23/okt/231011-forberedt-pa%CC%8A-fremtiden-2-web.pdf [14.05.2024]

Retsinformation (2015) “Bekendtgørelse om uddannelsen til professionsbachelor som lærer i folkeskolen” Available: https://www.retsinformation.dk/eli/lta/2015/1068 [14.05.2024]

Rigsarkivet. (2017). *Den danske værdiundersøgelse 2017*. Available: https://digidata.rigsarkivet.dk/aflevering/38175 [14.05.2024]

Roersen, Anne Balsby & Anne-Sofie Bøllund Andreasen, Emma Svendsen, Julie Wulff Sachse, Jonas Franksen Nielsen, Maria Laursen og Markus Borleif Jansson (2022). *Tilfældighedernes karaktergivning - et studie om bedømmelsespraksissens betydning for karakterernes troværdighed*. Aalborg Universitet. Available: https://surveyselskab.dk/wp-content/uploads/2023/03/Tilfaeldighedernes-karaktergivning-Bachelor-Prisopgave-2022.pdf [14.05.2024]

Rokani, V., & Kaminaris, S. D. (2020). *Power transformers fault diagnosis using AI techniques*. In AIP Conference Proceedings (Vol. 2307, No. 1). AIP Publishing.

Seawright, Jason & Gerring, John (2008). *Case Selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options*. Political research quarterly, 2008-06, Vol.61 (2), p.294-308. Available: https://www.proquest.com/docview/215333544/fulltextPDF?parentSessionId=SYilCWsB8NsqPplvU%2Bcf%2BtUQqWLCeOSG1SYkGNo1OtE%3D&pq-origsite=primo&accountid=8144&sourcetype=Scholarly%20Journals [14.05.2024]

Scikit-learn (n.d.). *1.5. Stochastic Gradient Descent*. Available: https://scikit-learn.org/stable/modules/sgd.html [14.05.2024]

Serritzlew, S. (2007). *Det politiske eksperiment: hvorfor, hvornår og hvordan?*. Politica, 39 (3): 275-293.

Shehab, A., Faroun, M., & Rashad, M. (2018). *An automatic Arabic essay grading system based on text similarity Algorithms*. International Journal of Advanced Computer Science and Applications, 9(3).

Simon, H. A. (1947). *Administrative behavior; a study of decision-making processes in administrative organization*. Macmillan. Simon and Schuster.

Simonsen, Amalie Rokkedal (2024). *Dansk Erhverv står i spidsen for udviklingen af en dansk version af ChatGPT.* DR. Available: https://www.dr.dk/nyheder/viden/teknologi/dansk-erhverv-staar-i-spidsen-udviklingen-af-en-dansk-version-af-chatgpt

Singh, Vibudh. (2023). *A Guide to Controlling LLM Model Output: Exploring Top-k, Top-p, and Temperature Parameters*. Medium. Available: https://ivibudh.medium.com/a-guide-to-controlling-llm-model-output-exploring-top-k-top-p-and-temperature-parameters-ed6a31313910 [14.05.2024]

Sniderman, Paul M. (2011). *The Logic and Design of the Survey Experiment in* Druckman, James N, Greene, Donald P. & Kuklinski (2011). *Cambridge Handbook of Experimental Political Science*. Cambridge University Press, 1. udgave. Available: https://books.google.dk/books?hl=en&lr=&id=ES5PUvvShF8C&oi=fnd&pg=PA102&dq=method+social+science+survey+experiment&ots=WEdDfwJVb1&sig=7_zxbPa0C-QCPvhgBs-KWdeXlro&redir_esc=y#v=onepage&q=random&f=false [14.05.2024]

Suresh, KP (2011). An overview of randomization techniques: An unbiased assessment of outcome in clinical research. Journal of Human Reproductive Sciences 4(1):p 8-11, Jan–Apr 2011. | DOI: 10.4103/0974-1208.82352. Available: https://journals.lww.com/jhrs/fulltext/2011/04010/an_overview_of_randomization_techniques__an.3.aspx

Svendsen, Gert Tinggaard (2012) *Tillid.* Aarhus Universitetsforlag. Available: https://unipress.dk/media/17128/tillid.pdf [14.05.2024]

Taris, T. W. (2000). *Longitudinal data and longitudinal designs*. In A primer in longitudinal data analysis (pp. 1-16). SAGE Publications Ltd. Available: https://www-doi-org.zorac.aub.aau.dk/10.4135/9781849208512 [14.05.2024]

Tariq, A.; Luo, M.; Urooj, A.; Das, A.; Jeong, J.; Trivedi, S. & Banerjee, I. (2024). *Domain-specific LLM Development and Evaluation–A Case-study for Prostate Cancer*. MedRxiv, 2024-03.

Tavakol, Mohsen & Dennick, Reg. (2011). *Making sende of Cronbach's alpha*. International Journal of Medical Education. 2011; 2:53-55. Available: https://www.researchgate.net/publication/270820426_Making_Sense_of_Cronbach's_Alpha [14.05.2024]

Tavakol, M., & Wetzel, A. (2020). F*actor Analysis: a means for theory and instrument development in support of construct validity*. International journal of medical education, 11, 245–247.

Troelsen, S. (2020). *Forkastet eller anerkendt? Et eksplorativt studie i folkeskoleelevers afgangseksamen i 'Dansk, skriftlig fremstilling'*. Institut for Kulturvidenskaber, Syddansk Universitet.

Uddannelses- og forskningsministeriet (2020). *Evaluering af optagelsessystemet til de videregående uddannelser*. Available: https://moreinfo.addi.dk/2.11/more_info_get.php?lokalid=48909949&attachment_type=856_a&bibliotek=870970&source_id=870970&key=5e81ef0008ae68bf3440 [14.05.2024]

Ueno, T., Sawa, Y., Kim, Y., Urakami, J., Oura, H., & Seaborn, K. (2022, April). *Trust in human-ai interaction: Scoping out models, measures, and methods*. In CHI Conference on Human Factors in Computing Systems Extended Abstracts (pp. 1-7).

UVM (2023). *Censorvejledning for beskikkede censorer ved folkeskolens prøver*. Available:https://www.uvm.dk/-/media/filer/uvm/udd/folke/pdf23/marts/230303-censorvejledning-2023.pdf [14.05.2024]

Wilson, H. J., & Daugherty, P. R. (2018). *Collaborative intelligence: Humans and AI are joining forces*. Harvard Business Review, 96(4), 114-123.

World value survey (n.d.) *World Value Survey Wave 7 (2017-2022)*. Available: https://www.worldvaluessurvey.org/WVSDocumentationWV7.jsp [14.05.2024]

Wu, Y.; Rabe, M. N.; Hutchins, D. & Szegedy, C. (2022). *Memorizing transformers*. Available: https://arxiv.org/abs/2203.08913 [14.05.2024]

Wærn, Y. and Ramberg, R. (1996). *People's perception of human and computer advice*. Computers in Human Behavior, 12, pp. 17–2.

Zeberg, R. H. (2021). Verdensmester i offentlig digitalisering. Available: https://tidsskrift.dk/samfundsokonomen/article/view/125540 [14.05.2024]

Zhou, Yuhang; Zhu, Yiyang & Wong, Weng Kee (2023) *Statistical tests for homogeneity of variance for clinical trials and recommendations*. Contemporary Clinical Trials Communications. 2023 Jun; 33: 101119. Availble https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10151260/ [14.05.2024]