Fra gjetting til refleksjon på eksamen med flervalgsoppgaver

Omid Mirmotahari, Iuliia Perminova, and Yngvar Berg

Department of Informatics, University of Oslo, Norway

Sammendrag Denne studien undersøker implementering av confidence-based assessment (CBA) i flervalgsoppgaver. Gjennom et mixed-methods-design design analyseres data fra over 1100 informatikkstudenter over to år (2023-2024). Resultatene viser at CBA gir et mer nyansert bilde av studentenes kunnskapsnivå ved å redusere effekten av gjetting og fremme metakognitiv refleksjon. Studentene bruker betydelig mer tid på CBA-oppgaver sammenlignet med tradisjonelle flervalgsoppgaver, noe som indikerer dypere kognitiv prosessering. Kvalitative intervjuer avdekket at studenter opplever CBA som utfordrende men verdifull for læring, særlig ved komplekse oppgaver. Ulike poengmodeller blir sammenlignet, og full CBA-poenggivning viser seg å gi mest rettferdige vurderingsresultater. Studien konkluderer med at CBA kan transformere flervalgsoppgaver fra rene kunnskapstester til verktøy som også måler metakognitiv bevissthet, spesielt på høyere nivåer i Blooms taksonomi.

Keywords: confidence-based assessment · Flervalgsoppgaver (MCQ) · Informatikkutdanning · Metakognitiv refleksjon · Blooms taksonomi

1 Innledning

Vurdering i høyere utdanning står overfor et fundamentalt dilemma: Hvordan kan vi effektivt evaluere hundrevis av studenter samtidig som vi gir dem meningsfull tilbakemelding som fremmer læring? Dette spørsmålet blir særlig presserende i store innføringskurs i informatikk, hvor heterogene studentgrupper med svært ulik bakgrunn skal tilegne seg komplekse konsepter innen programmering, dataarkitektur og systemforståelse [Mirmotahari et al., 2003]. Utfordringen forsterkes av at kunstig intelligens nå utfordrer mange av de vurderingsformene vi tradisjonelt har benyttet. Hjemmeoppgaver, prosjektarbeid og andre former for uovervåket arbeid blir stadig mer sårbare for KI-assistert bistand, noe som har ført til en tydelig trend hvor flere institusjoner beveger seg mot eksamen under tilsyn som primær vurderingsform. For emner med flere hundre studenter blir fristelsen stor til å ty til flervalgsoppgaver – ikke primært av pedagogiske hensyn, men fordi automatisk retting både er kostnadseffektivt og fritar institusjonen fra ressurskrevende ordninger med dobbelsensor og håndtering av klagesaker.

Flervalgsoppgaver har forsåvidt lenge vært en praktisk løsning for storskalaundervisning. De muliggjør standardisert testing av flere hundre studenter samtidig, med automatisert retting som reduserer arbeidsbelastningen og subjektiv påvirkning fra sensorer. Likevel har denne vurderingsformen betydelige svakheter som har vært gjenstand for vedvarende kritikk i forskningslitteraturen [Kwon et al., 2023]. En hovedinnvending

er at flervalgsoppgaver primært måler de lavere nivåene i Blooms taksonomi [Bloom et al., 1956] – kunnskap og forståelse – mens de sjelden klarer å teste høyere kognitive ferdigheter som analyse, syntese og evaluering på en pålitelig måte.

Problemet forsterkes av at flervalgsformatet åpner for gjetting. Med fire svaralternativer har studentene 25% sjanse for å gjette riktig, og ved bruk av eliminasjonsstrategier kan denne sannsynligheten øke betraktelig. Som [Gardner-Medwin and Curtin, 2007] påpeker: "A lucky guess is not knowledge, and it is incorrect and inefficient... to mark an assessment as if it were". Dette skaper ikke bare et validitetsproblem for vurderingen, men sender også uheldige signaler til studentene om at overflatestrategier kan erstatte dyp forståelse.

Et annet kritisk aspekt ved tradisjonelle flervalgsoppgaver er at de opererer med en binær logikk – svaret er enten rett eller galt – uten å fange opp graden av konfidens, eng. confidence, bak studentens valg. Historisk sett har nettopp vurdering av kandidatens trygghet og (selv)sikkerhet vært et sentralt element i muntlige eksamener, hvor sensorer naturlig vektlegger om studenten fremstår selvsikker eller usikker i sine resonnementer. Når denne dimensjonen mangler i skriftlige flervalgsoppgaver, mister vi verdifull informasjon om studentenes faktiske kompetansenivå [Novacek, 2017].

Våre tidligere studier har utforsket ulike tilnærminger til disse utfordringene. Arbeid med automatisert tilbakemelding [Mirmotahari et al., 2019b] viste hvordan strukturerte vurderingssystemer kan gi studenter detaljert formativ feedback selv i store kurs. Videre demonstrerte implementering av peer review-systemer [Mirmotahari and Berg, 2018, Mirmotahari et al., 2019a] at studenter kan aktiveres som ressurser i hverandres læring, samtidig som de utvikler metakognitiv bevissthet om vurderingskriterier og kvalitetsstandarder. Disse erfaringene har lagt grunnlaget for å utforske hvordan CBA kan integreres i eksamenssituasjoner.

CBA representerer en lovende utvidelse av tradisjonelle flervalgsoppgaver. Ved å be studentene ikke bare velge et svar, men også angi hvor sikre de er på valget sitt, introduserer metoden en ekstra dimensjon som reduserer effekten av gjetting og fremmer refleksjon over egen kunnskap [Novacek, 2017]. I praksis nærmer dette seg dynamikken i muntlige eksamener, hvor graden av sikkerhet naturlig inngår i vurderingen [Gardner-Medwin, 1995]. Metoden har potensial til å transformere flervalgsoppgaver fra rene kunnskapstester til verktøy som også stimulerer og måler metakognitiv bevissthet.

Særlig interessant er CBAs potensial til å utvikle det [Schauber et al., 2021] omtaler som kognitiv refleksjon – evnen til å overstyre intuitive, men feilaktige responser til fordel for mer analytiske løsninger. Studenter med høy kognitiv refleksjon er ikke immune mot intuitive feil, men de er bedre rustet til å identifisere og korrigere dem [De Neys and Bonnefon, 2013]. Ved å kreve at studentene eksplisitt vurderer sin konfidens, tvinger CBA fram en metakognitiv prosess hvor intuitive svar må veies mot faktisk kunnskap. Dette representerer et skifte fra raske System 1-responser til mer bevisste System 2-strategier [Kahneman, 2011], noe som er avgjørende for læring på høyere nivåer i Blooms taksonomi.

Til tross for økende interesse for CBA internasjonalt, viser systematiske oversikter at metoden fortsatt er lite utforsket innen informatikk- og teknologifag [Almarzuki et al., 2024]. Dette er overraskende gitt fagets kvantitative natur og behov for presis

problemløsning, hvor evnen til å vurdere egen konfidens kan være særlig relevant. Videre peker nyere studier på at CBA kan bidra til mer rettferdig vurdering ved å redusere kjønnsforskjeller i risikovillighet ved gjetting [Baldiga, 2013], noe som er særlig relevant i teknologifag med skjev kjønnsbalanse.

I denne konteksten presenterer vi en studie gjennomført i emnet "Introduksjon til datateknologi"ved Universitetet i Oslo, hvor vi har implementert og evaluert CBA over to år (2023-2024) med til sammen over 1100 studenter. Studien bygger på prinsipper om constructive alignment [Biggs and Tang, 2011], hvor vurderingsformen søker å harmonere med både læringsmål og undervisningsaktiviteter. Vi undersøker følgende forskningsspørsmål:

1. Kan CBA gi et mer nyansert og detaljert bilde av studentenes kunnskapsnivå sammenlignet med tradisjonelle flervalgsoppgaver?
2. I hvilken grad bidrar CBA til å fremme metakognitiv refleksjon og kognitiv dybdeprosessering hos studentene?
3. Hvordan opplever studenter implementering av CBA i en eksamenssituasjon, og hvilke implikasjoner har dette for fremtidig bruk?

Gjennom en mixed-methods tilnærming som kombinerer kvantitative eksamensdata med kvalitative studentintervjuer, bidrar studien til forståelsen av hvordan innovative vurderingsformer kan implementeres i storskala informatikkundervisning samtidig som både effektivitet og pedagogisk kvalitet ivaretas.

2 Kasus

Denne studien ble gjennomført i et innføringsemne, “Introduksjon til datateknologi”, i informatikk ved Universitetet i Oslo. Emnet er obligatorisk for alle studieprogrammer, tilsammen ca 800 studenter, ved Institutt for informatikk (IFI). Studieprogrammene har forskjellige inntakskrav som bidrar stort til en meget hetrogen studentmasse og sprik i motivasjon. Emnet dekker fagområder som sikkerhet, nettverk, operativsystemer, programmering, maskinkode, dataarkitektur og boolsk algebra – stor bredde og forholdsvis liten dybde. Undervisningen består av fire timer forelesning og to timer gruppeundervisning per uke og for å kvalifisere til eksamen må alle studentene ha bestått tre obligatiske innleveringer i løpet av semesteret. Emnet gir 10 studiepoeng og vurderes med en digital, bestått/ikke bestått skoleeksamen som hovedsakelig består av flervalgsoppgaver, det er ca 30 oppgaver. Denne vurderingsformen er valgt fordi den er skalerbar, effektiv retting (automatisk rettet) og gjør det mulig å teste et bredt pensum på en standardisert måte innen begrenset tid og med flere hundre kandidater. Baksiden med flervalgsoppgaver er gjetting og bruk av eliminasjonsstrategier snarere enn kritisk refleksjon. Derfor har vi utforsket ulike måter å lage automatisk rettede oppgaver flere år, oppgavetyper som “flytt og slipp”, rangere og flere andre muligheter som finnes igjennom eksamensplattformen til Inspera.

For å adressere utfordringen med gjetting og heller snu det til økt refleksjon valgte vi å implementere en CBA-tilnærming som en av oppgavene. I 2023 ble en oppgave på middels nivå i Blooms taksonomi utformet til å kunne vurderes som CBA. Studentene skulle velge et svar og samtidig angi sin egen prosentsikkerhet, samt vurdere

sannsynligheten for at andre alternativer kunne være riktige. Oppgaveteksten var som følger: "*Gitt en ALU som skal gjøre følgende operasjoner: [Addere, NAND, XOR, NOR, XNOR], (a) hvor mange bit må styresignalet være?*" og svaret er en tekstboks som studentene skal skrive sitt svar i. Riktig svar er 3. Utvidelsen av denne oppgaven kommer ved "*(b) Hvor sikker er du på svaret ditt i (a)?* " og her får studentene fem tekstbokser som de skal fordele 100% på.

I 2024 ble prinsippet basert på analyser og resultater fra 2023 videreført med en viktig endring ved å anvendte CBA i en mer kompleks flervalgsoppgave som krevde både anvendelse og analyse av flere av temaene i kurset og som dermed kan plasseres på et høyere nivå i Blooms taksonomi. Denne gangen fikk studentene fire forhåndsdefinerte svaralternativer og de skulle fordele prosentsikkerhet (0%, 25%, 50%, 75%, 100%) mellom dem. En student kunne for eksempel angi 100% på ett alternativ og 0% på de øvrige, eller fordele sikkerheten mellom flere plausible svar.

Målet med disse to oppgavevariantene var å undersøke hvordan CBA kan bidra til å fremme metakognitiv refleksjon, samtidig som vi kunne gi studentene en mer rettferdig poenggivning.

3 Metode og forskningsdesign

Studien bygger på et mixed-mode-design design som kombinerer kvantitativ analyse av eksamensdata med kvalitative fokusgruppeintervjuer. Denne kombinasjonen gjør det mulig både å undersøke mønstre i studentenes besvarelser på aggregert nivå og å belyse hvordan de selv opplevde den aktuelle vurderingsformen.

Som primær kvantitativ kilde ble eksamensbesvarelser eksportert fra Inspera i anonymisert form og analysert på aggregert nivå. Analysen tok utgangspunkt i ulike eksamensoppgaver hvor formålet var å studere hvordan studentene arbeidet med oppgavene under ulike vurderingsbetingelser. I 2023 ble en oppgave grundig analysert og evaluert for bruk av CBA.

For begge år ble det gjennomført sammenligninger mellom tradisjonell poengberegning og alternative poengmodeller, for å vurdere hvordan ulike vurderingssystemer vil påvirke sluttresultatene. Videre ble to ulike måter å angi konfidens på testet: I 2023 ble studentene bedt om å oppgi prosentsikkerhet fritt, mens i 2024 fikk de forhåndsdefinerte prosentsatser (0%, 25%, 50%, 75%, 100%) som de skulle fordele på de tilgjengelige alternativene.

Den kvalitative delen av studien baserer seg på semistrukturerte fokusgruppeintervjuer med tilfeldig valgte studenter som hadde gjennomført de aktuelle eksamenene. Fokusgruppeformatet ble valgt for å fremme refleksjon og diskusjon, der studentenes ulike erfaringer kunne nyansere hverandres perspektiver. Denne formen for intervju muliggjør at deltakeres svar skaper nye refleksjoner hos andre, noe som beriker datamaterialet gjennom gruppedynamikken.

Intervjuene, 9 invitasjoner – 7 deltok fordelt på 3 grupper, ble gjennomført med en felles intervjuguide som sikret tematisk konsistens på tvers av gruppene. Guiden besto av åpne spørsmål om studentenes opplevelser av å besvare flervalgsoppgaver med CBA-tilnærmingen, hvordan denne påvirket deres strategier for oppgaveløsning, samt

hvordan de trodde de ville tilpasse læringsstrategiene sine dersom eksamen kun besto av CBA-oppgaver. Kjønnsfordelingen (K/M) var 5/2 og intervjuene varte 20 ± 5 minutter.

Analysen av intervjudata fulgte en induktiv tematisk tilnærming [Braun and Clarke, 2006], hvor mønstre og temaer ble identifisert direkte fra materialet. Det er gjennomgått transkriberingen iterativt hvor tolkninger ble diskutert gjennom flere analyserunder. Dette sikret at de identifiserte temaene var godt forankret i datamaterialet og representerte studentenes faktiske perspektiver fremfor intervjuerens forutinntatte antagelser.

Data fra eksamen 2023 ble analysert i RStudio, mens analysene av 2024 besvarelsene ble utført i Python (Jupyter Notebook). I begge tilfeller ble statistiske analyser og visualiseringer gjennomført ved hjelp av relevante bibliotek for datahåndtering og statistisk modellering. Transkribering av lydopptaket er blitt gjort gjennom autotekst.uio.no.

Studien har godkjenning fra SIKT - Kunnskapssektorens tjenesteleverandør, med referansenummer: 708615. Meldeskjemaet for personopplysninger ble fylt ut og godkjent i forkant av intervjuene.

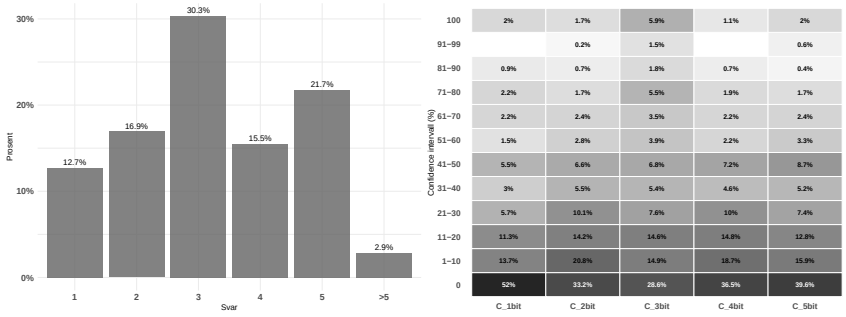
4 Resultat

Vi presenterer først tidsbruksanalyser, deretter svarmønstre og konfidensfordelinger, og til slutt studentenes kvalitative opplevelser av vurderingsformen. Merk at det var først i 2023 CBA ble introdusert.

En sammenligning av tidsbruk mellom tradisjonelle flervalgsoppgaver (før 2023) og CBA-baserte oppgaver (2023) avdekker markante forskjeller, som vist i Tabell 1. Studentene brukte betydelig mer tid på CBA-oppgaver enn på tilsvarende tradisjonelle flervalgsoppgaver. Gjennomsnittlig tidsbruk for tradisjonelle oppgaver var 42 sekunder, mens den for CBA-oppgaver var 470 sekunder – en økning på over ti ganger. Spesielt påfallende er det at hele 81 % av studentene brukte mindre enn ett minutt på tradisjonelle flervalgsoppgaver, mens ingen brukte så kort tid på CBA-oppgaver. I stedet brukte majoriteten (67 %) av studentene mellom 2 og 10 minutter på CBA-oppgavene, med en betydelig andel (24 %) som brukte mer enn 10 minutter.

	< 2023		2023		2024	
Tidsbruk	Antall	%	Antall	%	Antall	%
0-1 min	471	81%	0	0%	11	2%
1-2 min	80	14%	50	9%	82	14%
2-5 min	25	4%	172	31%	265	47%
5-10 min	4	1%	242	43%	176	31%
>10 min	0	0%	137	24%	35	6%
Total	**580**	100%	**561**	100%	**569**	100%
Median	29 sekunder		408 sekunder		281 sekunder	
Gjennomsnitt	42 sekunder		470 sekunder		238 sekunder	

Tabell 1: Tidsbruk på tilsvarende flervalgsoppgave før 2023, 2023 og 2024.



(a) (b)

Figur 1: (a) fordelingen på riktig hovedsvar, (b) detaljer fordeling av konfidens.

4.1 Svarfordeling og konfidensanalyse

Figur 1a viser svarfordelingen på flervalgsoppgaven fra 2023, det korrekte svaret = 3. Som figuren illustrerer, valgte 30,3 % av studentene dette alternativet. Interessant nok valgte også en betydelig andel av studentene (21,7 %) alternativ 5, noe som kan indikere en spesifikk misforståelse av det faglige innholdet.

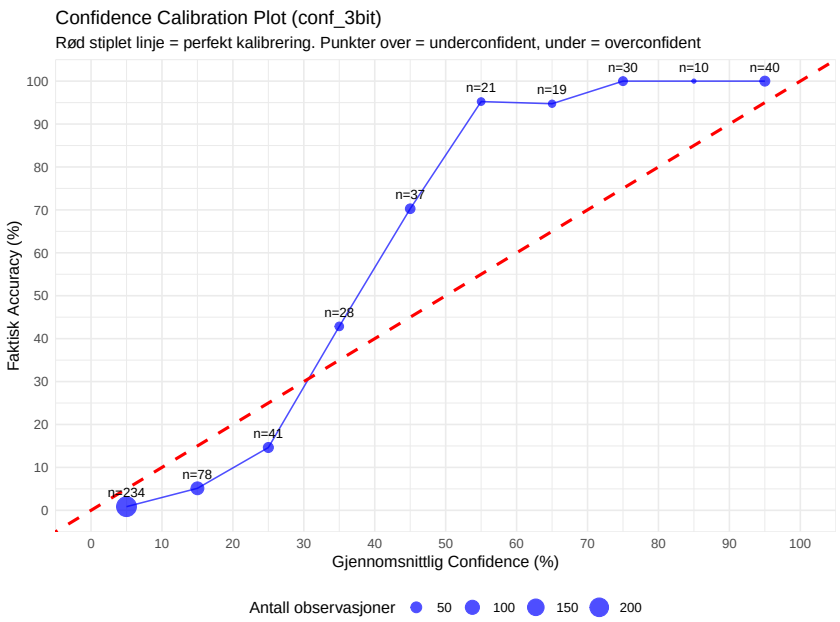
For å få en dypere forståelse av studentenes kunnskapsnivå, analyserte vi hvordan de fordelte sin konfidensgrad på de ulike svaralternativene. Figur 1b presenterer denne fordelingen i form av et heatmap. Et viktig funn er at selv studenter som ikke valgte det korrekte svaret (alternativ 3) som sitt hovedsvar, likevel tildelte dette alternativet en viss grad av konfidens. Dette tyder på at studentene gjenkjente delvis korrekte løsninger, selv når de ikke var sikre nok til å velge dem som sitt primære svar.

Heatmapet i Figur 1b avdekket også et interessant mønster i hvordan studentene fordelte sin konfidensgrad. Vi observerte en tendens til at studentene konsentrerte sine vurderinger rundt verdiene 0 %, 25 %, 50 %, 75 % og 100 %. Denne observasjonen dannet grunnlaget for utformingen av forhåndsdefinerte konfidensintervaller i 2024-eksamen.

For å undersøke forholdet mellom studentenes konfidens og faktiske prestasjoner, gjennomførte vi en mer detaljert analyse av svaralternativ 3 (korrekt svar). Figur 2 på neste side illustrerer dette forholdet gjennom et kalibreringplot. Den røde stiplede linjen representerer perfekt kalibrering, der konfidensnivå og faktisk prestasjon samsvarer nøyaktig. Punkter over denne linjen indikerer underkonfidens (studenter som presterer bedre enn de tror), mens punkter under linjen representerer overkonfidens (studenter som overvurderer sin kunnskap).

Analysen avdekker et interessant mønster: Studenter som svarte riktig på hovedspørsmålet (Faktisk Accuracy = 100 %) viste moderat til høy konfidens, typisk mellom 75 % og 95 %. Dette tyder på at de fleste studenter som hadde korrekt forståelse, også var relativt sikre på svaret sitt, men ikke perfekt sikre. I motsetning til dette viste studenter som svarte feil (Faktisk Accuracy = 0 %) ofte høy konfidens i at alternativ 3 ikke var korrekt. Dette demonstrerer den velkjente Dunning-Kruger-effekten.

De fleste flervalgsoppgaver har tradisjonelt benyttet et binært poengsystem, der korrekt svar gir full uttelling mens feil svar gir null poeng. Vi analyserte dataene fra



Figur 2: Detaljer for hvordan studentenes hovedsvar og relasjonen til konfidens for 3 bit er.

2023-eksamen med tre alternative poengsystemer: tradisjonelt, hybrid og full CBA. I det hybride systemet beregnet vi poengene som: P_{hybrid} = hovedsvaret$_{poeng}$ $\times$ C_3bit %. I full CBA-systemet derimot, beregnet vi poengene utelukkende basert på studentens konfidensfordeling for det korrekte alternativet ($P_{fullCBA}$ = Max_Poeng $\times$ C_3bit %). Tabell 2 presenterer en sammenlignende analyse av disse tre poengsystemene. Det tradisjonelle systemet gav et gjennomsnitt på 0,625 poeng, men med en median på 0, noe som indikerer en skjev fordeling der mange studenter fikk null poeng. Standardavviket var høyt (0,928), tilsvarende nesten 50% av maksimal poengsum. Full CBA-systemet viste seg å gi en mer balansert poengfordeling med et gjennomsnitt på 0,590 og en median på 0,4. Spesielt interessant er at standardavviket på 0,643, noe som indikerer en jevnere fordeling av poeng blant studentene. Dette systemet gir uttelling for delvis kunnskap – studenter som gjettet riktig men var usikre fikk lavere poeng.

System	$\bar{x}$[0,2]	median	σ
Tradisjonell	0.625	0.0	0.928
Hybrid	0.434	0.0	0.698
Full CBA	0.590	0.4	0.643

Tabell 2: Statistikk for de tre ulike poengsystemene for flervalgsoppgaven i 2023.

Basert på analysene fra 2023 videreutviklet vi CBA-tilnærmingen for 2024-eksamen. I stedet for å be studentene angi prosentsikkerhet fritt, innførte vi forhåndsdefinerte konfidensgrader (0 %, 25 %, 50 %, 75 %, 100 %) som studentene skulle fordele mellom svaralternativene. Oppgaven ble også utformet for å teste høyere nivåer i Blooms taksonomi, med fokus på analyse og anvendelse fremfor ren gjenkjenning.

For 2024-eksamen erstattet vi også det hybride poengsystemet med et differensiert system som gav ulik poenguttelling basert på svaralternativenes faglige nivå. Svar som reflekterte høy forståelse og analyse gav maksimalt 2 poeng, mens delvis korrekte svar på lavere taksonomisk nivå gav henholdsvis 1 og 0,5 poeng. Feilsvar gav 0 poeng. Vi eksperimenterte også med negative poeng (straff") for åpenbart feilaktige svar.

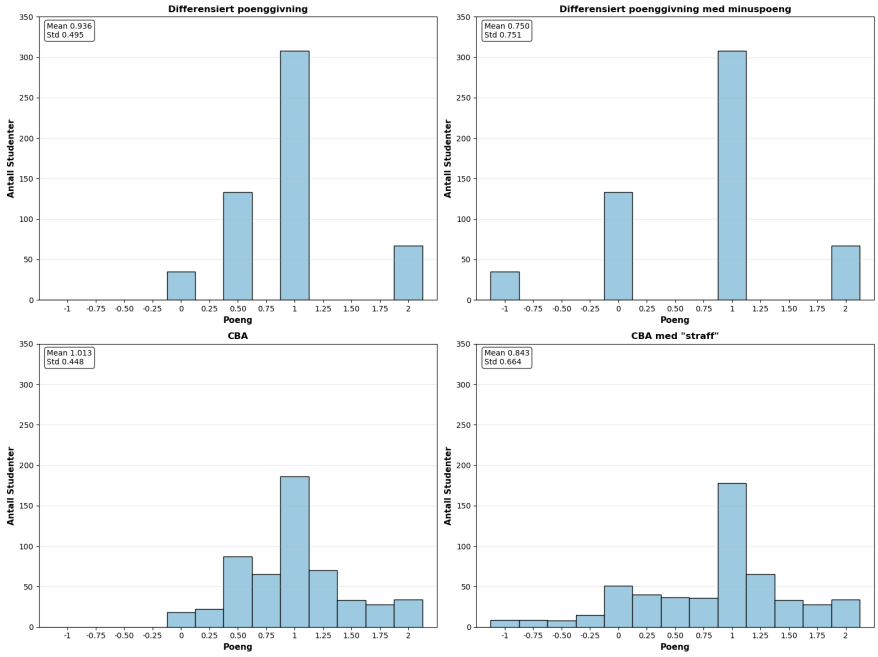
Tabell 3 presenterer den deskriptive statistikken for disse ulike poengsystemene. Det tradisjonelle systemet resulterte i et lavt gjennomsnitt (0,247) og en median på 0, noe som gjenspeiler igjen den binære tilnærmingen. Det differensierte systemet gav et gjennomsnitt på 0,936 og en median på 1,0, mens CBA-systemet resulterte i et gjennomsnitt på 1,013 og en tilsvarende median på 1,0. Begge disse alternative systemene reduserte standardavviket betraktelig sammenlignet med det tradisjonelle systemet, hvilket indikerer en mer nyansert vurdering av studentenes kunnskapsnivå.

4.2 Sammenligning av tidsbruk mellom 2023 og 2024

Figur 3 på neste side illustrerer fordelingen av poeng under de ulike poengsystemene for 2024-eksamen. Her fremkommer det tydelig at både det differensierte systemet og CBA-systemet gav mer nyanserte poengfordelinger sammenlignet med det tradisjonelle systemet. Særlig interessant er det at CBA-systemet med “straff” viste den mest differensierte fordelingen, hvilket tyder på at denne tilnærmingen best evner å skille mellom ulike grader av forståelse. Tidsbruken for CBA-oppgaver viste en interessant utvikling fra 2023 til 2024, som presentert i Tabell 1 på side 5. Selv om studentene fortsatt brukte betydelig mer tid på CBA-oppgaver i 2024 sammenlignet med tradisjonelle flervalgsoppgaver før 2023, observerte vi en markant reduksjon fra 2023. Median tidsbruk falt fra 408 sekunder i 2023 til 281 sekunder i 2024, mens gjennomsnittlig tidsbruk ble redusert fra 470 til 238 sekunder. Dette kan indikere at studentene ble mer fortrolige med formatet, eller at de forhåndsdefinerte konfidensintervallene gjorde oppgavene mindre tidkrevende å besvare.

System	$\bar{x}$[0,2]	median	σ
Tradisjonell	0.247	0.0	0.658
Differensiert	0.936	1.0	0.495
CBA	1.013	1.000	0.448
Med minuspoeng	$\bar{x}$[-1,2]	median	σ
Differensiert	0.750	1.0	0.751
CBA	0.843	1.000	0.664

Tabell 3: Deskriptiv statistikk for poenggivning (eksamen 2024)



Figur 3: Visualisering for de ulike poengsystemene for flervalgsoppgaven i 2024.

Figur 4 på neste side gir ytterligere innsikt ved å vise gjennomsnittlig tidsbruk for de ulike svaralternativene i 2024. Studenter som primært valgte svaralternativet "minne" brukte mest tid, noe som indikerer at studenter som valgte det faglig mest korrekte svaret også investerte mer tid i oppgaveløsningen. Dette underbygger antakelsen om at CBA fremmer dypere kognitiv prosessering hos studentene.

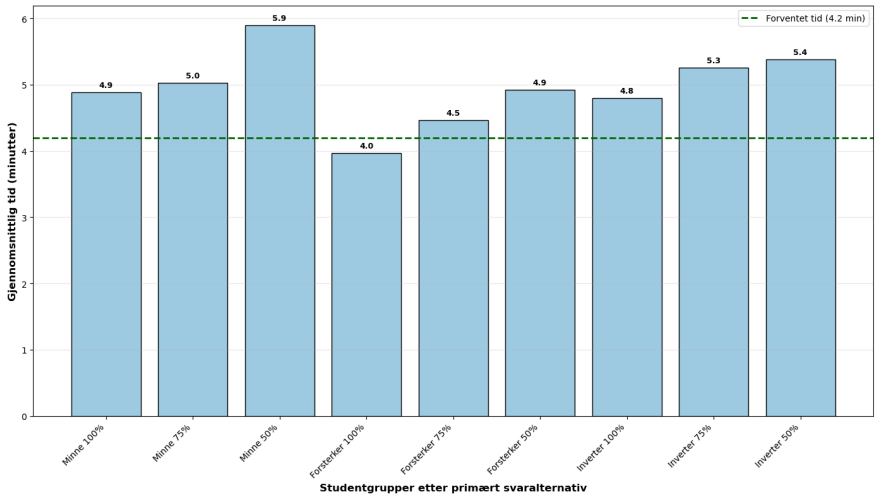
De kvalitative intervjuene gav verdifull innsikt i studentenes opplevelser av CBA oppgavene. En gjennomgående observasjon var at CBA-formatet tvang studentene til å reflektere grundigere over oppgavene:

> "*Hadde jeg ikke måttet si hvor sikker jeg var på det, så tror jeg ikke jeg ville tenkt så mye over oppgaven som jeg gjorde. Da hadde det nok bare blitt sånn: 'Det er sikkert den der,' og så hadde jeg gått videre. Men her måtte jeg jo – i stedet for å trykke ett klikk – bruke fire. Da må man jo tenke litt mer. Det tar jo lengre tid.*" (student 1)

Flere studenter beskrev oppgaven som faglig utfordrende, men samtidig interessant og motiverende:

> "*Det var litt vanskelig, men samtidig morsomt. Så selv om det var krevende, ble det også mer motiverende.*" (student 1)

> "*Jeg syntes også det var et vanskelig tema, men også et morsommere tema enn de andre. Det som gjorde oppgaven vanskelig var at jeg måtte sette den i kontekst med de andre temaene.*" (student 2)



Figur 4: Gjennomsnitt tidsbruk for de ulike svaralternativ i 2024. Det siste alternativet er ikke inkludert her da vi anser den som irrelevant for denne visualiseringen.

Studentene påpekte også at CBA-formatet, til tross for at det initialt virket fremmed, gav dem mulighet til å uttrykke nyansene i sin fagforståelse - særlig fremhevet de at formatet ga mulighet til å synliggjøre usikkerhet og vise hvordan de tenkte underveis, i stedet for bare å velge ett alternativ:

> "*Fordelen (red.ed. med oppgaveformen) er at de kanskje får sett litt mer hva folk tenker... og ikke bare krysse av på ett svar per oppgave.*" (student 3)

> "*Og altså, man har jo fortsatt mulighet til å bare velge ett alternativ og si seg fornøyd med det. Sånn sett så er det jo bare en mulighet for større muligheter, da... kunne svare annerledes og vise hva man tenker.*" (student 1)

Interessant nok reflekterte studentene også over for hvilke typer oppgaver CBA formatet er mest hensiktsmessig:

> "*Jeg tenker det er greit å bruke sikkerhet (red.ed. konfidens) på de vanskelige oppgavene, der det faktisk er rom for usikkerhet... og det er bra at man kan velge flere alternativer. Men på de enkleste oppgavene blir det litt overflødig å sitte og dele ut prosenter, når det er mye raskere å bare velge ett svar. Eksamen er jo stressende nok fra før med tidspress og alt, så da er det unødvendig å gjøre den vanskeligere der man ikke trenger det.*" (student 4)

5 Diskusjon

Denne studien avdekker flere sentrale aspekter ved implementering av konfidens-basert vurdering, confidence-based assessment (CBA), i informatikkutdanning. Våre funn belyser både utfordringer og muligheter ved denne vurderingsformen, særlig relatert til

metakognitiv refleksjon, dybdelæring og kunnskapsvurdering. I det følgende vil vi diskutere hovedfunnene i lys av våre forskningsspørsmål og eksisterende litteratur.

Vårt første forskningsspørsmål undersøkte hvorvidt CBA kan gi et mer nyansert bilde av studentenes kunnskapsnivå sammenlignet med tradisjonelle flervalgsoppgaver. Resultatene gir betydelig støtte til denne antakelsen. De komparative analysene av poengsystemene viser at CBA-basert poenggivning produserer en mer differensiert fordeling med lavere standardavvik og høyere median enn tradisjonell binær poenggivning (se Tabell 2 og 3). Dette indikerer at metoden effektivt fanger opp grader av forståelse som ellers ville blitt kategorisert som enten helt korrekt eller helt feil.

Særlig interessant er funnet at konfidenskalibreringsplottet (Figur 2 på side 7) avdekker nyanser i studentenes metakognitive bevissthet som er utilgjengelige i tradisjonelle vurderingsformer. At studenter med korrekt svar viser moderat til høy, men ikke perfekt konfidens (75-95 %), indikerer en sofistikert selvvurdering der selv kunnskapsrike studenter erkjenner muligheten for feil. Dette samsvarer med [Gardner-Medwin and Curtin, 2007] som argumenterer for at vellykket læring innebærer både fagkunnskap og evnen til å vurdere påliteligheten av denne kunnskapen.

Videre avdekket analysen tegn på Dunning-Kruger-effekten, der studenter med feilaktige svar ofte utviste uforholdsmessig høy konfidens i sine vurderinger. Dette fenomenet, først beskrevet av [Kruger and Dunning, 1999], representerer en sentral utfordring i høyere utdanning: Studenter med de svakeste faglige forutsetningene er ofte også de som minst presist kan vurdere sin egen kompetanse. Ved å eksplisitt måle konfidens avdekker CBA dette misforholdet og gir dermed et mer komplett bilde av studentenes faktiske kompetansenivå – ikke bare hva de vet, men også hva de vet at de vet.

Sammenligningen av de ulike poengsystemene demonstrerer videre at full CBA poenggivning skaper den mest rettferdige vurderingen, ved å:

- Straffe gjetting ved å redusere poengene for korrekte svar avgitt med lav konfidens
- Belønne delvis kunnskap ved å gi poeng for korrekte konfidensvurderinger selv ved feilaktige hovedsvar
- Differensiere mer presist mellom ulike grader av forståelse, som reflektert i den jevnere poengfordelingen

Dette adresserer direkte kritikken mot tradisjonelle flervalgsoppgaver som [Kwon et al., 2023] fremhever: At formatet belønner gjetting og overflatestrategier fremfor dyp forståelse.

Vårt andre forskningsspørsmål undersøkte i hvilken grad CBA bidrar til å fremme metakognitiv refleksjon og kognitiv dybdeprosessering. Tidsbruksanalysene gir overbevisende empirisk støtte for at CBA faktisk stimulerer til dypere kognitiv engasjement. Økningen i gjennomsnittlig tidsbruk fra 42 sekunder (tradisjonelle flervalgsoppgaver) til 470 sekunder (CBA, 2023) representerer en tidobling, noe som indikerer en fundamental endring i studentenes tilnærming til oppgaveløsning.

Denne kvantitative observasjonen underbygges av studentenes egne beskrivelser i intervjuene, hvor de eksplisitt beskriver hvordan formatet tvang dem til grundigere refleksjon: ‘‘*Hadde jeg ikke måttet si hvor sikker jeg var på det, så tror jeg ikke jeg ville tenkt så mye over oppgaven som jeg gjorde*” (student 1). Dette samsvarer med [Kahneman, 2011] distinksjon mellom System 1 (raske, intuitive) og System 2 (langsomme,

analytiske) kognitive prosesser. CBA ser ut til å fremtvinge et skifte fra System 1 til System 2-tenkning ved å kreve eksplisitt metakognitiv vurdering av egen sikkerhet.

Særlig interessant er observasjonen fra Figur 4 på side 10 som viser at studenter som valgte det faglig mest korrekte svaret ("minne") brukte mer tid enn de som valgte det intuitivt plausible men faglig mindre presise alternativet ("forsterker"). Dette indikerer at CBA ikke bare øker tidsbruken generelt, men spesifikt stimulerer til den type dybderefleksjon som er nødvendig for å overkomme intuitive misforståelser til fordel for presis faglig forståelse. Dette samsvarer med [Schauber et al., 2021]s funn om at metakognitiv bevissthet er spesielt viktig for å kunne overstyre intuitive, men feilaktige responser.

Reduksjonen i tidsbruk fra 2023 til 2024 (Tabell 1 på side 5) er også interessant i denne sammenhengen. Selv om studentene fortsatt brukte betydelig mer tid enn på tradisjonelle flervalgsoppgaver, ble tiden nesten halvert. Dette kan tolkes som et tegn på at studentene ble mer fortrolige med formatet, eller at de forhåndsdefinerte konfidensintervallene (0 %, 25 %, 50 %, 75 %, 100 %) forenklet den metakognitive vurderingsprosessen. Dette antyder at den kognitive belastningen ved CBA kan reduseres gjennom bevisst oppgavedesign, uten å kompromittere de metakognitive fordelene – en observasjon som har viktige implikasjoner for videre implementering.

Et annet sentralt funn er at CBA ser ut til å være særlig verdifullt for oppgaver som tester høyere nivåer i Blooms taksonomi [Bloom et al., 1956]. Studentene identifiserte selv dette i intervjuene, hvor de påpekte at konfidens-vurdering gir mest mening "*på de vanskelige oppgavene, der det faktisk er rom for usikkerhet*" (student 4). Dette samsvarer med vårt forskningsdesign, der vi i 2024 bevisst utformet CBA-oppgaver for å teste analyse og anvendelse fremfor ren gjenkjenning.

Dette antyder at CBA kan adressere en av de mest vedvarende kritikkene mot flervalgsformatet: At det primært tester lavere kognitive nivåer. Ved å integrere metakognitiv vurdering transformeres oppgavene fra rene kunnskapstester til verktøy som også måler studentenes evne til å vurdere kompleksitet og usikkerhet – nøkkelkompetanser for fremtidige informatikere som vil operere i domener preget av tvetydighet og raske endringer.

Vårt tredje forskningsspørsmål undersøkte hvordan studenter opplever implementeringen av CBA i eksamenssituasjoner. De kvalitative intervjuene avdekket et interessant mønster. Studentene opplevde initialt formatet som utfordrende og tidkrevende, men utviklet gradvis en verdsettelse for dets pedagogiske verdi. Denne spenningen mellom umiddelbar opplevelse og retrospektiv verdsettelse representerer et viktig pedagogisk dilemma ved implementering av innovative vurderingsformer.

Studentene beskrev eksplisitt hvordan CBA-formatet gjorde oppgavene mer kognitivt utfordrende, men samtidig mer faglig engasjerende: "*Det var litt vanskelig, men samtidig morsomt. Så selv om det var krevende, ble det også mer motiverende*" (student 1). Dette tyder på at den økte kognitive belastningen ikke nødvendigvis oppfattes negativt, men kan stimulere til dypere faglig engasjement når den er godt integrert i oppgavedesignet. Dette samsvarer med [Biggs and Tang, 2011]s prinsipp om "constructive alignment", der økt kognitiv utfordring kan fremme dybdelæring når den samsvarer med læringsmålene.

Særlig interessant er studentenes refleksjon rundt formatets evne til å synliggjøre tankeprosesser: "*Fordelen er at de kanskje får sett litt mer hva folk tenker... og ikke bare krysse av på ett svar per oppgave*" (student 3). Dette indikerer at studentene selv erkjenner begrensningene ved tradisjonelle flervalgsformater og verdsetter muligheten til å kommunisere nyansene i sin forståelse. Denne metakognitive bevisstheten er særlig verdifull i informatikkfaget, hvor komplekse problemer sjelden har enkle, binære løsninger, og hvor erkjennelse av usikkerhet er en sentral profesjonell kompetanse.

Studentenes differensierte perspektiv på når CBA er mest hensiktsmessig – "*på de vanskelige oppgavene, der det faktisk er rom for usikkerhet*" (student 4) – samsvarer med våre kvantitative funn. Dette tyder på at CBA bør implementeres selektivt, med fokus på oppgaver som tester kompleks fagforståelse og høyere kognitive nivåer, snarere enn som en universell erstatning for alle flervalgsoppgaver. Dette har viktige implikasjoner for ressurseffektiv eksamensdesign, hvor CBA kan reserveres for de mest faglig sentrale spørsmålene.

Våre funn gir flere praktiske implikasjoner: (1) CBA bør introduseres gradvis gjennom semesteret før det anvendes i eksamenssituasjoner, da tidsbruken reduseres med erfaring. (2) Full CBA-poenggivning gir mest differensiert vurdering, men også enklere differensierte poengmodeller representerer betydelig forbedring sammenlignet med binære systemer. (3) Forhåndsdefinerte konfidensintervaller (0 %, 25 %, 50 %, 75 %, 100 %) reduserer kognitiv belastning uten å kompromittere metodens pedagogiske fordeler. (4) CBA bør primært anvendes på oppgaver som tester høyere kognitive nivåer, hvor legitim usikkerhet er et relevant aspekt av fagkompetansen.

5.1 Begrensninger og videre forskning

Tidligere forskning [Baldiga, 2013, Riukula, 2023] dokumenterer at kvinnelige studenter ofte viser større motvilje mot å gjette under usikkerhet – en adferd som tradisjonelt straffes i flervalgsformater. CBA adresserer potensielt denne skjevheten ved å belønne kalkulert usikkerhet fremfor ukritisk gjetting, noe som kan bidra til mer rettferdige vurderingsresultater i fag med skjev kjønnsbalanse. Fremtidige studier bør systematisk undersøke denne effekten.

Studiens primære begrensninger inkluderer datagrunnlag fra kun ett emne ved én institusjon, manglende innsikt i langtidseffekter på læringsstrategier, og begrenset undersøkelse av demografiske forskjeller. Fremtidig forskning bør fokusere på implementering i ulike fagkontekster, longitudinelle studier av metakognitiv utvikling, integrasjon med andre innovative vurderingsformer, og kartlegging av hvilke oppgavetyper innen informatikk som er best egnet for CBA-tilnærming.

Referanser

Almarzuki et al., 2024. Almarzuki, H. F., Abu Samah, K. A. F., Riza, L. S., and Nordin, S. (2024). Research trends in confidence assesment: a systematic litereture review. *Journal of Theoretical and Applied Information Technology*, 102:1227–1239.

Baldiga, 2013. Baldiga, K. (2013). Gender differences in willingness to guess. *Management Science*, pages 434–448.

Biggs and Tang, 2011. Biggs, J. and Tang, C. (2011). *Teaching for quality learning at university*. McGraw-Hill/Society for Research into Higher Education, Maidenhead.

Bloom et al., 1956. Bloom, B., Engelhart, M., Furst, E., Hill, W., and Krathwohl, D. (1956). *Taxonomy of educational objectives: The classification of educational goals*. Longmans, Green.

Braun and Clarke, 2006. Braun, V. and Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3(2):77–101.

De Neys and Bonnefon, 2013. De Neys, W. and Bonnefon, J.-F. (2013). The 'whys' and 'whens' of individual differences in thinking biases. *Trends in Cognitive Sciences*, 17:172–178.

Gardner-Medwin, 1995. Gardner-Medwin, A. (1995). Confidence assessment in the teaching of basic science. *Association for Learning Technology Journal*, 3:80–85.

Gardner-Medwin and Curtin, 2007. Gardner-Medwin, T. and Curtin, N. (2007). Certainty-based marking (cbm) for reflective learning and proper knowledge assessment. *REAP International Online Conference on Assessment Design for Learner Responsibility*, pages 1–7.

Kahneman, 2011. Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux, New York, USA.

Kruger and Dunning, 1999. Kruger, J. and Dunning, D. (1999). Unskilled and unaware of it: How difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of Personality and Social Psychology*, 77(6):1121–1134.

Kwon et al., 2023. Kwon, H.-J., Chae, S. J., and Park, J. H. (2023). Educational implications of assessing learning outcomes with multiple choice questions and short essay questions. *Korean Journal of Medical Education*, 35(3):285.

Mirmotahari and Berg, 2018. Mirmotahari, O. and Berg, Y. (2018). Structured peer review using a custom assessment program for electrical engineering students. In *IEEE Global Engineering Education Conference (EDUCON)*, pages 999–1006.

Mirmotahari et al., 2019a. Mirmotahari, O., Berg, Y., Fremstad, E., and Damsa, C. (2019a). Student engagement by employing student peer reviews with criteria-based assessment. In *IEEE Global Engineering Education Conference (EDUCON)*, pages 1152–1157.

Mirmotahari et al., 2019b. Mirmotahari, O., Berg, Y., Gjessing, S., Fremstad, E., and Damsa, C. (2019b). A Case-Study of Automated Feedback Assessment. In *IEEE Global Engineering Education Conference (EDUCON)*, pages 1190–1197.

Mirmotahari et al., 2003. Mirmotahari, O., Holmboe, C., and Kaasbøll, J. (2003). Difficulties learning computer architecture. In *Proceedings of the 8th annual conference on Innovation and technology in computer science education*, pages 247–247.

Novacek, 2017. Novacek, P. F. (2017). Exploration of a confidence-based assessment tool within an aviation t viation training pr aining program. *Journal of aviation/aerospace education & research*, 26:65–88.

Riukula, 2023. Riukula, K. (2023). Gender differences in multiple-choice questions and the risk of losing points. *Journal of the Finnish Economic Association*, 10:42–44.

Schauber et al., 2021. Schauber, S. K., Hautz, S. C., Kämmer, J. E., Stroben, F., and Hautz, W. E. (2021). Do different response formats affect how test takers approach a clinical reasoning task? an experimental study on antecedents of diagnostic accuracy using a constructed response and a selected response format. *Advances in Health Sciences Education*, 26:1339–1354.