Samme svar, ulik karakter: konsekvenser av skåringsregel i konfidensbasert vurdering

Omid Mirmotahari and Yngvar Berg

Institutt for informatikk, Universitetet i Oslo
omidmi@uio.no

Sammendrag Konfidensbasert vurdering (CBA) har fått økende oppmerksomhet i høyere utdanning, men valget av skåringsregel behandles ofte som en teknisk detalj. Vi undersøker hva valget gjør, ved å holde besvarelsene fast og variere regelen. Materialet er CBA-besvarelser fra et førstesemesteremne i informatikk fra en obligatorisk innlevering (N = 371) og ordinær eksamen (N = 523), med et samtykkebasert koblet delutvalg (N = 226). Studentene fordeler konfidens i prosent over fem svaralternativer, og hvert eksamensalternativ er klassifisert som riktig, plausibelt, misforståelse eller feil. De samme svarene reskåres kontrafaktisk under sju regler i to familier: fasitbaserte regler (dikotom, lineær, kvadratisk, logaritmisk og sfærisk) og innholdsvektede delpoengmatriser (emnets regel og en alternativ matrise utviklet i emnet). Med CBA-delens faktiske vekt på 14 av 100 poeng er rangeringen etter totalskår tilnærmet uendret (Spearmans $\rho \geq 0,99$). På en konstruert karakterskala skifter derimot mellom 3,4 % av kandidatene (under den mildeste regelen, M1) og 14,3 % (under den mest inngripende, M6) karaktertrinn, avhengig av hvilken regel som velges. Tretten kandidater strøk på maskinvaredelen under emnets regel. Med de samme svarene ville tolv av dem ha bestått under den kvadratiske regelen. Forskjellene følger bestemte svarmønstre. Regler som gir uttelling for spredt konfidens, flytter poeng fra konsentrerte til spredte besvarelser. Et vektsveip viser at utslagene vokser bratt med CBA-delens vekt, og målekvaliteten varierer målbart mellom reglene. Skåringsregel, skalering, håndtering av ugyldige fordelinger og vekt bør derfor behandles som ett samlet vurderingsdesign. Artikkelen gir anbefalinger til emneansvarlige som vurderer å ta i bruk CBA.

Nøkkelord konfidensbasert vurdering, skåringsregler, proper scoring rules, kalibrering, flervalgsoppgaver, informatikkutdanning

1 Innledning

I mange store innføringsemner i informatikk brukes flervalgsoppgaver som en vesentlig del av vurderingen. Formatet skalerer, men har en velkjent svakhet: en student som gjetter riktig, er ikke til å skille fra en student som vet. Konfidensbasert vurdering (CBA) angriper dette ved å la studenten oppgi hvor sikker hun er, og ved å la poengene avhenge av både svar og sikkerhet. Jo mer vekt flervalgsformatet skal bære, desto viktigere blir poengsettingen av det.

I et stort norsk førstesemesteremne i informatikk er det innført CBA-oppgaver i både obligatorisk innlevering og eksamen [15]. Da regelen for poengsetting skulle fastsettes, ga litteraturen mange kandidater, men lite hjelp til å velge. I emnet som studeres her, var to konkurrerende regler i omløp blant underviserne, med ulikt syn på hvor mye delvis kunnskap skal telle og hvor hardt en misforståelse skal straffes.

En tidligere studie dokumenterte innføringen av opplegget og sammenlignet emnets egne poengmodeller på 2023- og 2024-kullene [15]. Denne artikkelen stiller et annet spørsmål, på nytt materiale fra 2025: hva gjør selve regelvalget med poeng, rangering og karakterer når alt annet holdes fast? Vi tar besvarelsene fra 2025-kullet og reskårer de samme svarene under sju regler, og vi følger fire forskningsspørsmål.

RQ1: Hvor mye varierer studentenes poengsum og innbyrdes rangering mellom skåringsreglene?
RQ2: Hvor mange kandidater skifter karaktertrinn ved bytte av regel, og hvor mange krysser bestått-grensen?
RQ3: Hvilke svarmønstre er følsomme for regelvalget?
RQ4: I hvilken grad gjenfinnes mønsteret av forskjeller mellom reglene i en obligatorisk innlevering og på eksamen?

Vi sammenligner sju skåringsregler på de samme besvarelsene fra en obligatorisk innlevering og en eksamen i samme emne. Analysen viser hvordan reglene påvirker poeng, rangering og klassifisering, og hvilke svarmønstre som forklarer forskjellene.

2 Bakgrunn

Ideen om å be studenter tallfeste egen sikkerhet er gammel. De Finetti [4] og Shuford, Albert og Massengill [18] beskrev på 1960-tallet prøveformater der studenten oppgir en sannsynlighetsfordeling over svaralternativene, og viste at bare bestemte poengregler gjør det lønnsomt å rapportere ærlig. En skåringsregel er proper når studentens forventede poengsum maksimeres av å oppgi den fordelingen hun faktisk tror på. Den kvadratiske regelen, kjent fra Briers arbeid med værvarsler [2], den logaritmiske og den sfæriske er standardeksemplene; Gneiting og Raftery [6] gir den moderne oversikten, og Wu mfl. [20] den nyeste beslutningsteoretiske behandlingen av konfidensbasert merking. En lineær regel, der poengene er proporsjonale med andelen lagt på riktig svar, er ikke proper. Da lønner det seg alltid å legge alt på det mest sannsynlige alternativet.

En mye brukt modell i undervisningssammenheng er Gardner-Medwins certainty-based marking ved UCL [5]. Der velger studenten ett svar og ett av tre konfidensnivåer, med poeng 1, 2 eller 3 ved riktig og 0, -2 eller -6 ved feil. Den summative utprøvingen gjaldt oppgaver av typen multiple true/false. Kalibreringslitteraturen har lenge dokumentert at subjektiv sikkerhet ofte overstiger treffraten, særlig i øvre del av skalaen [11].

I informatikkdidaktikken er konfidensvekting lite utforsket. Beasley [1] rapporterte en tidligere utprøving av konfidensvekting i et databehandlingsemne. Kalibrering hos informatikkstudenter er bedre dekket: Murphy og Tenenberg [16] fant moderat kalibrering før prøven og høyere kalibrering etter at studentene hadde sett oppgavene, og Hauswirth og Adamoli [7] undersøkte metakognitiv kalibrering i programmeringsopplæring. Nærmest vårt vurderingsdesign står Simon [19], som argumenterte for delpoeng i flervalgsoppgaver etter hvor gale de enkelte alternativene er – samme prinsipp som klassifiseringen av alternativer i vårt materiale.

Reskåring av de samme besvarelsene under alternative regler er undersøkt tidligere. Jaradat og Tollefson [8] analyserte reliabilitet, validitet og karaktersetting under alternativ poengvekting. Rashidi, Rezaei og Meysamie [17] fant grensekryssing i begge retninger ved overgang fra tellende til konfidensbasert skåring, men asymmetrisk. Hos dem gikk 9,1 % fra bestått til stryk og 3,6 % fra stryk til bestått.

Tidligere studier har hovedsakelig rapportert aggregerte forskjeller mellom et mindre antall regler. Her sammenligner vi sju regler, inkludert tre proper scoring rules og to lærerklassifiserte delpoengmatriser, og analyserer hvilke svarmønstre som forklarer poengendringene. En scoping review av 258 kilder og 21 skåringsmetoder [9] avgrenset dessuten eksplisitt mot konfidensbaserte formater.

3 Kontekst og data

Emnet vi bruker data fra, er et stort førstesemesteremne i informatikk med flere hundre studenter og karakterregel bestått/ikke bestått. CBA-opplegget er beskrevet i sin helhet i [15]; her gjengis det nødvendige. Hver CBA-oppgave presenterer fem svaralternativer. Studenten fordeler konfidens på alternativene i trinn på 0, 25, 50, 75 eller 100 %, med instruks om at summen skal være 100. En student som er sikker, legger 100 på ett alternativ. En som står mellom to, kan legge 50 på hvert. Formatet er dermed en diskretisert utgave av sannsynlighetsrapporteringen hos de Finetti og Shuford, og ikke en nivåskala av Gardner-Medwins type.

Vi bruker to datainnsamlinger fra høsten 2025. Den obligatoriske innleveringen (oblig) inneholdt ni CBA-oppgaver om maskinvare og var rent formativ. Den ga ikke poeng mot karakteren, CBA-delen ble godkjent uansett svar, og ingen skåringsregel var fastsatt. Emnet tilbød ulike undervisningsløp, og denne obligen inngikk i ett av dem. Den ble levert av 371 studenter. Obligutvalget er dermed en delmengde av eksamenskullet, avgrenset av hvilket undervisningsløp studentene fulgte. Den ordinære eksamenen ble besvart av 523 kandidater. Den ga til sammen 100 poeng fordelt på mange oppgavetyper, og CBA var én av dem: sju oppgaver à 2 poeng, til sammen 14 poeng. Vekten fulgte av eksamenens sammensetning og ble ikke fastsatt med tanke på regelvalget. For eksamensoppgavene klassifiserte forfatterne – som også er emnets forelesere og laget eksamensoppgavene – hvert alternativ som riktig (R), plausibelt (P), misforståelse (M) eller feil (F), og emnets skåringsregel ga basispoeng per alternativ etter denne klassi-

fiseringen. For obligen finnes bare fasit, ingen klassifisering. De innholdsvektede reglene kan derfor bare beregnes for eksamen.

I obligen samtykket 247 studenter til at besvarelsen kunne kobles til eksamensresultatet. Av disse hadde 244 gyldig kandidatnummer, og 226 gjennomførte eksamen. De utgjør det koblede utvalget for RQ4. Datagrunnlaget er pseudonymisert med kandidatnummer, kobling er kun gjort for studenter som har samtykket, og behandlingen er godkjent i tråd med institusjonens personvernregler.

4 Metode

Det metodiske grepet er å holde besvarelsene fast og variere regelen. Grepet har en innebygd begrensning som må frem før resultatene: reskåringen er *kontrafaktisk*, altså et hva-hvis-regnestykke. Vi kan beregne hva de samme svarene ville gitt under en annen regel, men ikke hva studentene ville svart under en annen regel, siden insentivene er del av regelen. Under en lineær regel lønner det seg å legge alt på ett alternativ. Under en kvadratisk lønner det seg å spre. Resultatene måler altså regelens direkte virkning på poengsettingen av gitt atferd, og de er samlet inn under emnets egen regel. Nettopp derfor er stabilitet interessant. Holder rangeringen seg når regelen byttes under besvarelsene, er regelvalget robust i det leddet der det virker mest direkte.

4.1 Reglene og forbehandlingen

Reglene faller i to familier. De fasitbaserte reglene (M0, M1, M3, M4, M5) bruker bare fasit og studentens konfidensfordeling. De innholdsvektede reglene (M2 og M6) bruker i tillegg faglærernes klassifisering av alternativene, og tar dermed stilling til at en plausibel feil er noe annet enn en misforståelse. Sammenligninger innad i en familie handler om funksjonsform og insentiver; sammenligninger på tvers handler også om hva som skal telle som delvis kunnskap. Emnets regel M2 ble brukt i sensuren; den alternative matrisen M6 ble formulert i undervisningsmiljøet uavhengig av og forut for denne analysen.

Forbehandlingen er felles for alle regler og gjøres før skåring. En oppgavebesvarelse uten avgitt konfidens gis 0 i alle regler. For de fasitbaserte reglene normaliseres allokeringene til vekter $w = a/\Sigma a$. M2 innebygger samme normalisering i sin egen formel, mens M6 slår opp direkte på rånivåene 0/25/50/75/100. Denne forskjellen hører med til forskjellen mellom systemene, og den varieres i sensitivitetsanalysen.

Notasjonen er som følger. a_i er konfidensen (0–100) på alternativ i, w er de normaliserte vektene, r er det riktige alternativet, y er indikatorvektoren for riktig svar, og b_i er emnets basispoeng. Tabell 1 viser reglene, som alle skaleres til $[0,\ 1]$ per oppgave. Emnets egen regel skrives utvetydig som $M_2 = \frac{1}{2}\max\!\big(0,\ \sum_i b_i a_i \,/\, \sum_i a_i\big)$. Divisjonen skjer altså etter summering, og faktoren $\frac{1}{2}$ skalerer mot høyeste basispoeng 2. Gulvet på 0 ble brukt i sensuren. I M5 er $\|w\|$ den euklidske normen $\sqrt{\sum_i w_i^2}$. Basispoengene per oppgave og hele M6-matrisen står i appendiks A.

Tabell 1. De sju skåringsreglene. * Proper før trunkering; trunkeringen trengs fordi log-skåren er minus uendelig når riktig alternativ har fått 0 %. YB-matrisen står i appendiks A.

Modell	Familie	Skår per oppgave (skala 0–1)	Proper?
M0 Dikotom (velg-maks)	Fasitbasert	1 hvis a_r er entydig størst, ellers 0	–
M1 Lineær andel	Fasitbasert	w_r	Nei
M2 Emnets regel	Innholdsvektet	$\frac{1}{2}\max\big(0,\ \sum_i b_i a_i / \sum_j a_i\big)$	Nei
M3 Kvadratisk (Brier)	Fasitbasert	$1 - \frac{1}{2}\sum_i (w_i - y_i)^2$	Ja
M4 Logaritmisk (trunkert)	Fasitbasert	$(\ln \max(w_r, \varepsilon) - \ln \varepsilon)/(-\ln \varepsilon)$, $\varepsilon = 0{,}025$	Ja*
M5 Sfærisk	Fasitbasert	$w_r / \lVert w \rVert$	Ja
M6 Alternativ matrise	Innholdsvektet	$\max(0,\ \mathrm{YB}(\text{kategori}, \text{nivå}))/8$	Nei

Tabell 2. Skår per oppgave for fire kanoniske svarmønstre. * M2 og M6 er beregnet for en typisk oppgave (basispoeng 2 / 0,5 / 0 / −0,5 / −2) med den mest plausible distraktoren som andrevalg. Regneeksempel for M6 og mønsteret «25 % på fire»: 25 % på riktig gir +2, 25 % på hver av to plausible gir +1 hver, og 25 % på misforståelsen gir 0; summen 4 av maksimum 8 gir 0,50.

Svarmønster	M0	M1	M2*	M3	M4	M5	M6*
100 % på riktig	1	1	1	1	1	1	1
50/50, riktig + plausibel	0	0,50	0,63	0,75	0,81	0,71	0,75
25 % på fire (én riktig)	0	0,25	0,25	0,63	0,62	0,50	0,50
100 % på feil (F)	0	0	0	0	0	0	0

To forbehold gjelder på tvers. Det ene gjelder skalaens grovhet. Med trinn på 25 % og krav om sum 100 finnes det bare 70 mulige måter å fordele konfidensen på, så en student som egentlig er 60 % sikker på ett alternativ, må runde av til 50 eller 75. Garantien i proper-reglene – at ærlig rapportering lønner seg – forutsetter at studenten kan oppgi troen sin eksakt, og gjelder derfor bare tilnærmet. Hvor fine trinn studentene får, er dermed en designbeslutning av samme slag som regelvalget.

Det andre forbeholdet er at reglene reagerer ulikt nær null. Den logaritmiske er langt mer følsom for små sannsynligheter enn den kvadratiske og den sfæriske. Det gir en presis forventning til sensitivitetsanalysen. Tabell 2 viser hva reglene gir for fire kanoniske svarmønstre. Spennet er stort. En besvarelse som sprer 25 % på fire alternativer der ett er riktig, får alt fra 0 poeng under M0 til 0,63 under M3. Karakterutslagene i kapittel 5 skyldes derfor både funksjonsform og hvilket nullpunkt hver regel forankrer skalaen i.

4.2 Totalskår, karaktertrinn og analysemål

Totalskåren under en regel beregnes i to deler. Poengene fra resten av eksamenen, maksimalt 86, beholdes som de er. CBA-delen byttes ut med poengene fra regelen, skalert til maksimalt 14. Hver regel settes altså inn i den samme eksamenen.

Emnet bruker bestått/ikke bestått, og den reelle bestått-mekanismen er delbasert. Eksamenen består av fire deler à 25 poeng, og hver del må bestås med mer enn 6 poeng. CBA-oppgavene utgjør 14 av de 25 poengene i maskinvaredelen. Hva regelvalget gjør ved denne reelle delgrensen, analyseres i 5.3. Karaktertrinnene i analysen er derimot konstruerte. Vi bruker UH-sektorens veiledende grenser ($A \geq 89$, $B \geq 77$, $C \geq 65$, $D \geq 53$, $E \geq 40$, $F < 40$) med en estimert helhetlig bestått-grense på 40. I tillegg rapporteres et normbasert regime, der karakterandelene under M2 holdes fast som kvoter og de øvrige reglene tildeler karakterer etter rangplass mot de samme kvotene. Karakteranalysen skal leses som et hva-hvis-regnestykke for emner med gradert skala. Valget mellom gradert skala og bestått/ikke bestått er selv et aktuelt spørsmål i IT-utdanningen [10].

Analysemålene følger forskningsspørsmålene. For RQ1 rapporterer vi poengforskjeller i originalenheter og Spearman-rangkorrelasjon med bootstrap-konfidensintervall. For RQ2 rapporterer vi andelen som skifter karaktertrinn og andelen som krysser bestått-grensen, i begge regimer, med perturbasjon av grensene som robusthetssjekk. For RQ3 dekomponerer vi poengendringen mot to trekk ved svarmønsteret. Det ene er treffsikkerhet, altså andelen oppgaver der størst konfidens lå entydig på riktig alternativ. Det andre er konsentrasjon, altså den gjennomsnittlige andelen av konfidensen som er lagt på ett alternativ. For RQ4 undersøker vi om mønstrene gjenfinnes i det koblede utvalget.

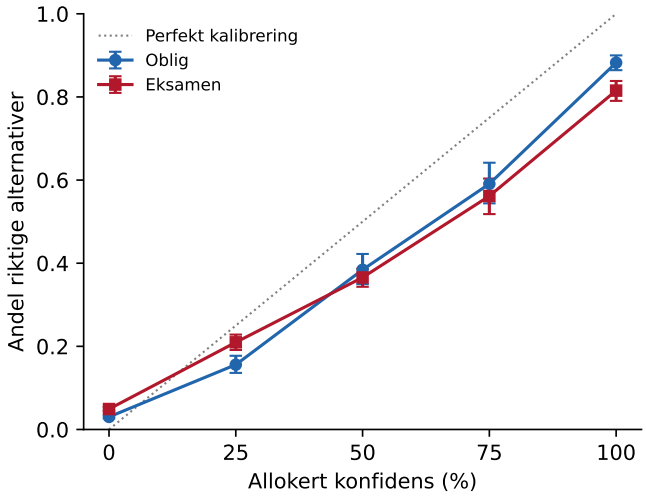
Poengendringen er en deterministisk funksjon av de samme svarene som prediktorene. RQ3-dekomponeringen har derfor ikke noe feilledd, og koeffisientene rapporteres uten p-verdier. I tillegg rapporteres intern konsistens, målt som Cronbachs alfa over de sju oppgavene, og samsvar med resten av eksamen per regel. Bootstrapintervallene resampler kandidater. De uttrykker dermed usikkerhet over kandidater gitt akkurat disse oppgavene, ikke over nye oppgavesett. Som validering ble emnets egen poengeksport reberegnet fra råallokeringene, med eksakt samsvar for alle 523 kandidater. Hele analysen er skriptet, og alle tall og figurer genereres direkte fra skriptene.

5 Resultater

5.1 Svaratferd og kalibrering

Studentene brukte konfidensformatet aktivt: på eksamen fordelte 43 % av oppgavebesvarelsene konfidensen på mer enn ett alternativ, i obligen 31 %. Instruksen om at summen skal være 100, ble brutt i 5,0 % av besvarelsene på eksamen og 10,2 % i obligen. Bruddene går nesten utelukkende én vei: 93 % av eksamensbruddene og 82 % av bruddene i obligen summerte til mer enn 100 %. Vi har forbehandlet en normaliseringen for å retter opp dette. Konfidensene skaleres om slik at de alltid summerer til 100, uansett hva studenten faktisk oppga. Uten den ville en besvarelse som feilaktig summerte til mer enn 100 %, kommet bedre ut enn en besvarelse som fulgte instruksen.

Figur 1 viser kalibreringen, med bootstrapintervaller på kandidatnivå siden allokeringer fra samme student ikke er uavhengige. Studentene skiller godt mellom alternativene: alternativer gitt 0 % var riktige i bare 3–5 % av tilfellene. På



Figur 1. Kalibrering. Andel riktige alternativer per allokert konfidensnivå, med 95 % bootstrapintervaller (resampling på kandidatnivå).

de lave nivåene er kalibreringen god. Alternativer gitt 25 % var riktige i 21 % av tilfellene på eksamen. Toppnivået preges derimot av overkonfidens. Alternativer gitt 100 % var riktige i 81,5 % av tilfellene på eksamen og 88,2 % i obligen. Forskjellen mellom oblig og eksamen kan ikke tilskrives innsatshøyden alene, siden også oppgavesett og tidspunkt er forskjellige.

5.2 RQ1: Poeng og rangering

De tre proper-reglene (M3, M4 og M5) og den alternative matrisen (M6) er gjennomgående mildere enn emnets regel. De gir 1,1–1,6 poeng høyere CBA-snitt av 14 mulige, og for hver tjuende kandidat er forskjellen mer enn 3–4,5 poeng. Rangeringen etter totalskår er likevel svært stabil. Alle regelpar har Spearman-korrelasjon minst 0,993, med bootstrapintervaller som ikke strekker seg under 0,994. Tabell 3 samler tallene.

CBA-delen utgjør 14 av 100 poeng, og de øvrige 86 poengene er ikke omfavnet av denne analysen. Isolerer vi CBA-delen, faller korrelasjonene til mellom 0,84 og 0,99. I en eksamen som i sin helhet ble skåret med to regler, fant Rashidi mfl. [17] en korrelasjon på 0,885. Den høye totalskårkorrelasjonen skyldes dermed delvis vekten, ikke bare likhet mellom reglene. Vekten analyseres for seg i avsnitt 5.7.

5.3 RQ2: Karaktertrinn og bestått-grense

Mellom 3,4 % (M1) og 14,3 % (M6) av kandidatene får et annet karaktertrinn i det kriteriebaserte regimet. I det normbaserte er spennet 3,2 til 9,6 %. Flytting på mer enn ett trinn forekommer praktisk talt ikke. Mellom 0,6 og 1,7 % havner

Tabell 3. Hver regel sammenlignet med emnets regel M2, eksamen (N = 523). CBA-snitt og avvik er i poeng av 14; karakterflytting er oppgitt for kriteriebasert (krit.) og normbasert (norm.) regime, og bestått-kryssing gjelder den konstruerte grensen på 40 poeng.

Modell	CBA-snitt (SD)	Avvik mot M2	ρ total-skår	ρ CBA-delen	Skifter trinn krit.	Skifter trinn norm.	Krysser bestått
M0 Dikotom	8,99 (3,48)	−0,17	0,996	0,922	9,6 %	9,4 %	1,3 %
M1 Lineær	9,31 (3,04)	+0,14	0,999	0,981	3,4 %	3,2 %	0,6 %
M2 Emnets regel	9,17 (3,30)	–	–	–	–	–	–
M3 Kvadratisk	10,75 (2,33)	+1,59	0,997	0,929	13,2 %	8,8 %	1,7 %
M4 Logaritmisk	10,66 (2,54)	+1,49	0,997	0,923	12,4 %	9,6 %	1,7 %
M5 Sfærisk	10,26 (2,75)	+1,09	0,998	0,945	10,9 %	7,3 %	1,5 %
M6 Alt. matrise	10,75 (2,62)	+1,58	0,997	0,919	14,3 %	8,0 %	1,5 %

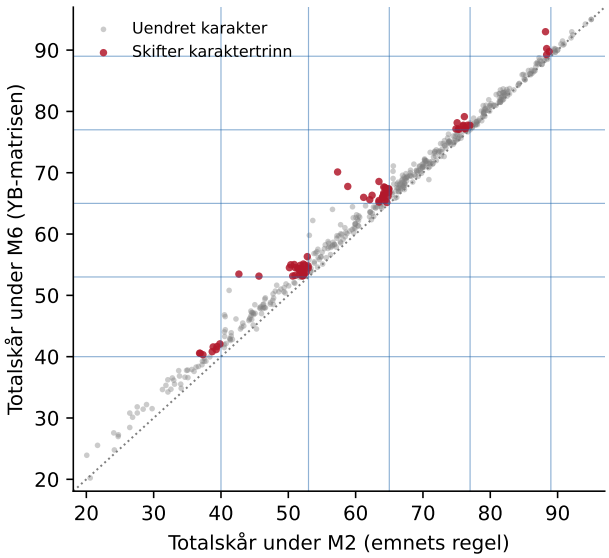
på motsatt side av den konstruerte bestått-grensen på 40, med identiske svar. Forskyves alle grenser ±2 poeng, varierer andelen som skifter trinn mellom 10 og 15 % for M3 og M6.

Figur 2 viser hvor flyttingene skjer. Kandidatene som skifter trinn, ligger klemt inntil grensene, der et regelskifte på under ett poeng er nok. Ved overgang fra M2 til den alternative matrisen M6 skiftet 75 av 523 kandidater trinn, alle oppover. De to reglene som var i omløp i emnet, skiller seg altså ikke bare i detaljene. Den ene er gjennomgående mildere.

Emnets strykmekanisme gjør disse tallene til mer enn et regnestykke. Hver av de fire delene må bestås med mer enn 6 av 25 poeng, og CBA-oppgavene utgjør 14 av de 25 poengene i maskinvaredelen. Tretten kandidater, 2,5 %, strøk faktisk på maskinvaredelen under emnets regel M2. Når de samme besvarelsene reskåres mot den samme reelle delgrensen, ville tolv av disse ha bestått under den kvadratiske regelen og under den alternative matrisen; matrisen ville samtidig strøket én kandidat som besto. Den dikotome regelen trekker motsatt vei og ville strøket 8 kandidater som besto under M2. Ved en reell grense får altså 1–3 % av kullet et annet kontrafaktisk bestått/stryk-utfall avhengig av regel.

5.4 RQ3: Hvilke svarmønstre bærer forskjellene

Poengendringene følger to trekk ved svarmønsteret, og Tabell 5.4 viser dekomponeringen. Merk at dette er en deskriptiv dekomponering. Endringen er en deterministisk funksjon av de samme svarene som prediktorene, og de to prediktorene er dessuten sterkt korrelerte (r = 0,73). Koeffisientene må derfor leses samlet, ikke hver for seg. For de tre proper-reglene (M3, M4 og M5) dominerer konsentrasjonen: besvarelser som legger alt på ett alternativ, taper på overgangen fra M2, mens besvarelser med spredt konfidens vinner. Den dikotome M0 trekker motsatt vei og gir uttelling for treffsikkerhet alene, siden delvis kunnskap ikke teller. M6 skiller seg fra begge: endringen er negativt assosiert med treffsikkerhet, fordi delpoengene for plausible svar først og fremst løfter besvarelser med svakere treff.



Figur 2. Totalskår under M2 mot M6. Røde punkter skifter karaktertrinn; blå linjer er de konstruerte karaktergrensene.

Modell	β treffsikkerhet	β konsentrasjon	R^2
M0 Dikotom	+0,77	−0,61	0,28
M1 Lineær	+0,08	−0,50	0,20
M3 Kvadratisk	+0,06	−0,95	0,82
M4 Logaritmisk	+0,06	−0,87	0,68
M5 Sfærisk	+0,29	−0,99	0,64
M6 Alt. matrise	−0,36	−0,34	0,43

5.5 Målekvalitet per regel

Skåringsregelen påvirker også hvor godt CBA-delen måler. Emnets egen regel hadde høyest observerte verdi på begge mål (α = 0,73, r = 0,72), tett fulgt av den lineære. De tre proper-reglene (M3, M4 og M5) hadde ikke høyere verdier, og kvadratisk og logaritmisk hadde de laveste alfaverdiene. Tabell 4 viser intern konsistens over de sju oppgavene og samsvaret med de 86 poengene fra resten av eksamen, som er skåret uavhengig av regelvalget.

Forskjellene er ikke testet, så rangeringen bør leses som observert i dette materialet. En mulig forklaring er at besvarelsene ble avgitt under emnets regel, slik at variasjonen i materialet er tilpasset den. Også det er en side av det kontrafaktiske forbeholdet. Med bare sju oppgaver og uten en ekstern gullstandard skal tallene tolkes varsomt. De viser likevel at insentivriktighet og målekvalitet er to forskjellige egenskaper.

Tabell 4. Målekvalitet per regel, eksamen (N = 523). Samsvar målt med Pearsons r mot poengsummen utenom CBA-delen.

Modell	Cronbachs α (7 oppgaver)	r mot resten av eksamen
M0 Dikotom	0,60	0,67
M1 Lineær	0,69	0,72
M2 Emnets regel	0,73	0,72
M3 Kvadratisk	0,56	0,68
M4 Logaritmisk	0,56	0,69
M5 Sfærisk	0,60	0,70
M6 Alt. matrise	0,65	0,67

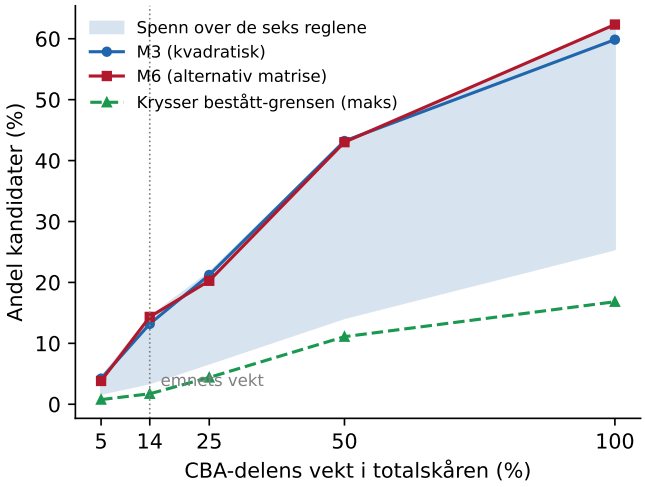
5.6 RQ4: Gjenfinnes mønsteret i obligen?

Strukturen i regelforskjellene gjenfinnes i obligen. Korrelasjonsmatrisene mellom de fasitbaserte reglene er nesten identiske i de to vurderingssituasjonene, med største parvise differanse på 0,02. M0 mot M3 er for eksempel 0,93 i obligen og 0,91 på eksamen. Også konsentrasjonsmønsteret fra RQ3 går igjen. Fordi obligen bare omfattet ett undervisningsløp, er de to matrisene beregnet på ulike populasjoner. Gjenfinningen skyldes likevel ikke populasjonsforskjellen. Regnet på de 226 koblede kandidatene alene avviker eksamensmatrisen med høyst 0,008 fra matrisen for hele kullet, og innenfor det koblede utvalget er største differanse mellom oblig og eksamen 0,014.

Den individuelle rangstabiliteten er derimot lav uansett regel. Spearmankorrelasjonen mellom en students plassering i obligen og på eksamen ligger mellom 0,11 og 0,24 i det koblede utvalget (N = 226). Noe av dette er målestøy, for med ni og sju oppgaver er reliabilitetstaket lavt. Korrigert for målefeil i begge delskalaene stiger den høyeste korrelasjonen fra 0,24 til om lag 0,34, som fortsatt er lavt. To forbehold begrenser tolkningen. Utvalget er selektert, siden de koblede kandidatene skårer bedre enn de øvrige (67,2 mot 60,9 poeng under M2, og bestått-andel 93 mot 89 %). Og de to vurderingssituasjonene skiller seg på oppgaver, tidspunkt og innsatshøyde samtidig, så en årsak kan ikke pekes ut. Det designet kan si, er at regelvalgets virkemåte gjenfinnes på tvers, mens den enkeltes plassering ikke gjør det.

5.7 Sensitivitet: parametere, forbehandling og vekt

Valg innenfor en regel har også konsekvenser. Det tydeligste eksemplet er trunkeringsparameteren ε i den logaritmiske regelen. Med disse trinnene er den minste positive vekten på riktig alternativ større enn ε. Trunkeringen binder derfor bare når studenten har lagt null på riktig alternativ, noe som gjaldt 16,7 % av eksamenssvarene. Parameteren avgjør med andre ord hvor hardt det skal straffes å være helt sikker på feil svar, og det er et rendyrket pedagogisk spørsmål. Varieres ε i et rimelig intervall, skifter mellom 9,6 % (ved $\varepsilon = 0{,}1$) og 14,0 % (ved $\varepsilon = 0{,}005$) av kandidatene karaktertrinn. Det stemmer med at den logaritmiske regelen legger langt større vekt på svært små sannsynligheter enn den kvadratiske og den sfæriske.



Figur 3. Karakterflytting mot M2 (kriteriebasert) og bestått-kryssing som funksjon av CBA-delens vekt i totalskåren.

Emnets to egne grep er mindre, men målbare. Å fjerne gulvet mot negative oppgavepoeng flytter 3,4 % av kandidatene, og å fjerne normaliseringen flytter 0,8 %. Også rene konvensjoner slår ut. Gis M0 brøkdelspoeng når riktig svar deler ledelsen, flytter 6,5 % av kandidatene.

Til slutt vekten. Figur 3 viser karakterflytting og bestått-kryssing når CBA-delens vekt varieres fra 5 til 100 %, med resten av eksamen skalert tilsvarende. Ved 5 % vekt flytter regelvalget 2–4 % av kandidatene. Ved emnets faktiske 14 % er spennet 3–14 %, ved 25 % opptil 22 %, og ved 50 % over 40 %. Ved halv vekt havner mer enn hver tiende kandidat på feil side av den konstruerte bestått-grensen.

6 Diskusjon

6.1 Hva reglene operasjonaliserer

Skåringsreglene operasjonaliserer ulike typer evidens for kunnskap. M0 belønner entydig valg av riktig svar, proper-reglene belønner gradert sannsynlighetsrapportering, og M2 og M6 skiller i tillegg mellom typer feilalternativer, slik at en misforståelse med full styrke teller verre enn en plausibel feiltolkning. Det er prinsippet Simon [19] argumenterte for, og det er beslektet med begrunnelsen for den harde straffen for sikre feil i Gardner-Medwins skala [5]. Det kan være pedagogisk ønskelig å skille systematiske misforståelser fra usikkerhet, men de to operasjonaliseringene av dette standpunktet ga ulik karakter til hver sjuende kandidat i vårt materiale, og den ene var konsekvent mildere. Hos Messick [12]

inngår konsekvensene av tolkning og bruk i validitetsargumentet. Benchmarktabellen gjør det samme poenget konkret – hvor mye en spredt besvarelse er verdt, spenner fra ingenting til to tredjedeler av oppgavepoenget, avhengig av regel.

6.2 Hvem valget angår

Dekomponeringen i RQ3 viser hvem regelvalget angår, med det forbeholdet at vi beskriver svarmønstre i denne prøven, ikke stabile studenttyper. Besvarelser med høy treffsikkerhet kommer godt ut under alle regler. Det er kandidatene nær karaktergrensene som berøres, og særlig to mønstre. Konsentrerte besvarelser taper på overgang til proper-regler, og spredte besvarelser taper på dikotomisering. Alternativer gitt 100 % var riktige i 81,5 % av eksamenssvarene (figur 1). Denne overkonfidensen i øvre del av skalaen er i tråd med kalibreringslitteraturen [11], og nesten hvert femte helt sikre svar var dermed feil. Det er disse svarene regler med sterk straff for sikre feil rammer. Kalibreringsstudier i informatikk viser dessuten at studentenes sikkerhet avhenger av hva som vurderes og når [16,7]. I et innføringsemne i digital kretsdesign overvurderte studentene seg på konseptuelle dimensjoner og undervurderte seg på åpent, kreativt arbeid [14]. Hvem en regel favoriserer, kan altså variere med oppgavetypen. Er det pedagogisk ønskelig å straffe sikre feil hardt, bør intensjonen gjøres eksplisitt overfor studentene framfor å ligge i verktøyets standardinnstilling.

6.3 Et beslutningsgrunnlag for emneansvarlige

Regelvalget bør knyttes til formålet med vurderingen. Er hovedmålet å gjøre ærlig rapportering av usikkerhet rasjonell, er den kvadratiske regelen et enkelt, avgrenset og insentivriktig alternativ uten log-regelens følsomme trunkeringsparameter, og med presedens i et absolutt karaktersystem [3]. Våre data viser samtidig hva man gir fra seg: kvadratisk skår skiller ikke plausible feil fra misforståelser, og den hadde ikke høyere observert målekvalitet enn emnets egen regel. Insentivriktighet gjelder innsamlingen, ikke målekvaliteten ved reskåring.

Er hovedmålet derimot diagnostikk av misforståelser, kreves en innholdsvektet matrise med dokumentert klassifiseringsprosess, og matrisen bør kalibreres mot en referanseregel slik at mildheten er villet og ikke tilfeldig; vår alternative matrise løftet 75 kandidater uten at noen gikk ned.

Hvilket utfall som er ønskelig, avhenger av hva emnet vil oppnå med CBA. Her ble formatet innført for å redusere gjetting og fremme refleksjon over egen sikkerhet [15]. Målt mot det formålet er M0 og M1 dårlige valg. M0 gir null for en ærlig fordeling der riktig svar deler ledelsen, og M1 belønner å legge alt på ett alternativ. CBA kan forstås som egenvurdering knyttet til hver oppgave, og egenvurdering har gitt målbar forbedring i et innføringsemne i informatikk [13]. Skal CBA støtte slik refleksjon, bør regelen gi uttelling for godt kalibrert usikkerhet og ikke bare for treff. Om reglene faktisk endrer svaratferden, kan designet vårt ikke besvare. Wu mfl. [20] fant at risikoholdninger får studenter under CBM til å avvike fra det poengmaksimerende valget, så samme regel kan virke ulikt på ulike studenter. Rashidi mfl. [17] fant, som vi, at grensekryssingen

går begge veier, men ikke symmetrisk. Den pedagogiske vurderingen må derfor ta med hvem som flyttes, ikke bare hvor mange.

Brukes CBA summativt, må regel, skalering, gulv, normalisering og vekt være fastsatt og kommunisert før besvarelsene foreligger. Vekten bør velges med figur 3 i mente: jo mer vekt CBA-delen får, desto mer avhenger karakteren av hvilken regel som velges. Ved 5 % vekt varierer andelen som skifter trinn mellom reglene fra 2 til 4 %. Ved emnets faktiske vekt på 14 % er spennet mellom reglene fra 3 % (for M1) til 14 % (for M6). Ved 25 % vekt kan opptil 22 % av kandidatene skifte trinn. I vårt emne fulgte vekten av eksamenens sammensetning, og vektsveipet viser at regelvalget ville slått langt sterkere ut med høyere vekt. Vil man prøve seg frem, bør eksperimenteringen skje formativt eller med lav, forhåndsfastsatt vekt. Sensurregelen skal ikke eksperimenteres frem etter eksamen.

Skåringsregelen må også kunne etterprøves. Universitets- og høyskoleloven krever skriftlig sensorveiledning og at en karakterbegrunnelse redegjør for de generelle vurderingsprinsippene som er lagt til grunn (§§ 11-7 og 11-8). En resultatavgjørende regel bør derfor være fastsatt før eksamen og dokumentert slik at den er tilgjengelig ved ordinær sensur og klagesensur, ikke bare i koden. Instruksbrudd bør dessuten håndteres med teknisk tvang mot sum 100 i besvarelsesløsningen, for en stilltiende normalisering i etterkant skjuler valget. I vårt materiale var 93 % av bruddene oppblåste allokeringer som normaliseringen skalerte ned.

7 Begrensninger

Reskåringen er kontrafaktisk. Studentene ville trolig svart annerledes under en annen regel, og alle resultater gjelder poengsettingen av atferd formet under emnets egen regel. Det gjelder også målekvalitetssammenligningen i 5.5. Karaktertrinnene er konstruerte, siden emnet bruker bestått/ikke bestått; nivåtallene for trinnflytting illustrerer mekanismen og er ikke anslag for et bestemt gradert emne, selv om grenseperturbasjonen viser at de ikke står og faller med grenseplasseringen. Delgrense-analysen bruker emnets reelle grense, men bare utfallet under M2 er observert sensur; utfallene under de øvrige reglene er kontrafaktiske.

Sammenligningen mellom oblig og eksamen blander oppgavesett, tidspunkt og innsatshøyde, og obligen omfattet bare ett av emnets undervisningsløp. Det koblede utvalget er selektert mot sterkere kandidater, og med ni og sju oppgaver er reliabilitetstaket for rangstabilitet lavt. Klassifiseringen av alternativer som plausible eller misforståelser er gjort av forfatterne uten formalisert prosedyre, og de innholdsvektede reglene finnes bare for eksamen. Analysen gjelder ett emne ved én institusjon. Hvordan studentene ville tilpasset seg en kjent regel, krever et design der regelen er kunngjort før besvarelsen, for eksempel ved å sammenligne kull med ulike regler.

8 Konklusjon

Vi reskåret besvarelser fra en obligatorisk innlevering (N = 371) og en eksamen (N = 523) under sju skåringsregler, med svarene holdt fast. Resultatene viser regelens direkte virkning på poengsettingen av gitt svaratferd. De viser ikke hva som ville skjedd om studentene hadde kjent en annen regel på forhånd. Innenfor denne rammen var rangeringen etter totalskår stabil, mens andelen som skiftet konstruert karaktertrinn, varierte fra 3,4 % (M1) til 14,3 % (M6). Ved emnets reelle delgrense ville tolv av de tretten som strøk under emnets regel, ha bestått under M3 og M6 med de samme svarene. Forskjellene hang sammen med besvarelsenes treffsikkerhet og konsentrasjon, og de økte med CBA-delens vekt. Tilpasser studentene seg en ny regel, kan utslagene bli både mindre og større. Funnene tilsier likevel at regel, forbehandling og vekt bør fastsettes og dokumenteres før besvarelsene foreligger, fordi regelvalget alene kan avgjøre utfallet for kandidater nær grensene.

Referanser

1. Beasley, R. E. (2012). The effects of confidence weighting on test scores and confidence weighting perceptions in the computing curriculum. Journal of Computing Sciences in Colleges, 28(2), 91–102.
2. Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. Monthly Weather Review, 78(1), 1–3. https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2
3. Dees, R. A., Gilliam, K. S. og Kwinn, M. (2008). Strictly proper scoring rules in an absolute grading environment – preliminary findings. ASEE Zone 1 Conference. https://doi.org/10.18260/1-2-1153-53745
4. de Finetti, B. (1965). Methods for discriminating levels of partial knowledge concerning a test item. British Journal of Mathematical and Statistical Psychology, 18(1), 87–123. https://doi.org/10.1111/j.2044-8317.1965.tb00695.x
5. Gardner-Medwin, A. R. og Gahan, M. (2003). Formative and summative confidence-based assessment. I J. Christie (red.), Proceedings of the 7th International Computer-Aided Assessment Conference, s. 147–155. Loughborough University.
6. Gneiting, T. og Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. Journal of the American Statistical Association, 102(477), 359–378. https://doi.org/10.1198/016214506000001437
7. Hauswirth, M. og Adamoli, A. (2017). Metacognitive calibration when learning to program. I Proceedings of the 17th Koli Calling Conference on Computing Education Research, s. 50–59. ACM. https://doi.org/10.1145/3141880.3141904
8. Jaradat, D. og Tollefson, N. (1988). The impact of alternative scoring procedures for multiple-choice items on test reliability, validity, and grading. Educational and Psychological Measurement, 48(3), 627–635. https://doi.org/10.1177/0013164488483006
9. Kanzow, A. F., Schmidt, D. og Kanzow, P. (2023). Scoring single-response multiple-choice items: Scoping review and comparison of different scoring methods. JMIR Medical Education, 9, e44084. https://doi.org/10.2196/44084

10. Lauvås, P., Mirmotahari, O., Sandtrø, T. A., Berg, Y. og Lindsjørn, Y. (2025). Gradert karakterskala eller Bestått/Ikke bestått? Refleksjoner fra studenter i høyere IT-utdanning. NIKT: Norsk IKT-konferanse for forskning og utdanning, 37(4). https://www.ntnu.no/ojs/index.php/nikt/article/view/6554
11. Lichtenstein, S., Fischhoff, B. og Phillips, L. D. (1982). Calibration of probabilities: The state of the art to 1980. I D. Kahneman, P. Slovic og A. Tversky (red.), Judgment under Uncertainty: Heuristics and Biases (s. 306–334). Cambridge University Press. https://doi.org/10.1017/CBO9780511809477.023
12. Messick, S. (1995). Validity of psychological assessment. American Psychologist, 50(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741
13. Mirmotahari, O. og Berg, Y. (2024). Trenger vi læringsassistenter? Hvordan kan hverandre-vurdering og/eller egen-vurdering gi god læringsutbytte. NIKT: Norsk IKT-konferanse for forskning og utdanning, nr. 4. https://doi.org/10.5324/nikt.6234
14. Mirmotahari, O. og Berg, Y. (2026). Students judge different things differently: dimension-specific self-assessment calibration in introductory digital circuit design. Computer Science Education. https://doi.org/10.1080/08993408.2026.2705528
15. Mirmotahari, O., Perminova, I. og Berg, Y. (2025). Fra gjetting til refleksjon på eksamen med flervalgsoppgaver. Norsk IKT-konferanse for forskning og utdanning (NIKT/UDIT). https://doi.org/10.5324/a5zvhg53
16. Murphy, L. og Tenenberg, J. (2005). Do computer science students know what they know? A calibration study of data structure knowledge. Proceedings of ITiCSE, 148–152. https://doi.org/10.1145/1067445.1067488
17. Rashidi, F., Rezaei, M. og Meysamie, A. (2025). Comparison of certainty-based marking (CBM) and number right scoring (NRS) in MCQ assessments. Advances in Medical Education and Practice, 16, 2133–2145. https://doi.org/10.2147/AMEP.S564289
18. Shuford, E. H., Albert, A. og Massengill, H. E. (1966). Admissible probability measurement procedures. Psychometrika, 31(2), 125–145. https://doi.org/10.1007/BF02289503
19. Simon (2011). Wrong is a relative concept: Part marks for multiple-choice questions. I J. Hamer og M. de Raadt (red.), Proceedings of the Thirteenth Australasian Computing Education Conference (ACE 2011), CRPIT 114, s. 47–53. Australian Computer Society.
20. Wu, Q., Vanerum, M., Agten, A., Christiansen, A., Vandenabeele, F., Rigo, J.-M. og Janssen, R. (2021). Certainty-based marking on multiple-choice items: Psychometrics meets decision theory. Psychometrika, 86(2), 518–543. https://doi.org/10.1007/s11336-021-09759-0

A De innholdsvektede reglene i full bredde

Tabell A1 viser basispoengene b_i i emnets regel (M2) per eksamensoppgave, med faglærernes klassifisering i parentes. Tabell A2 viser hele poengmatrisen i den alternative regelen (M6); oppgaveskåren er summen av oppslagene for alle fem alternativer, gulvet på 0 og delt på 8.

Tabell 5. Basispoeng per alternativ i emnets regel (M2), i alternativenes originalrekkefølge. Poeng per oppgave: $\sum_i b_i a_i / \sum_i a_i$, gulvet på 0.

Oppgave	Alt. 1	Alt. 2	Alt. 3	Alt. 4	Alt. 5
2.1 Cache-miss	−2 (F)	+2 (R)	0 (P)	−0,5 (M)	0 (P)
2.2 Halv-adder	0 (P)	+2 (R)	0 (P)	+0,5 (P)	−2 (F)
2.3 ALU	+2 (R)	+0,5 (P)	0 (P)	−0,5 (M)	−2 (F)
2.4 Program Counter	+2 (R)	+0,5 (P)	0 (P)	−2 (F)	−0,5 (M)
2.5 Universelle porter	+0,5 (P)	−0,5 (M)	+2 (R)	−1 (M)	−2 (F)
2.6 Pipeline	+0,5 (P)	0 (P)	0 (P)	+2 (R)	−2 (F)
2.7 Ripple Carry Adder	+0,5 (P)	+2 (R)	0 (P)	0 (P)	−2 (F)

Tabell 6. Den alternative matrisen (M6): poeng per (kategori, allokeringsnivå).

Kategori	0 %	25 %	50 %	75 %	100 %
R (riktig)	0	+2	+4	+6	+8
P (plausibel)	0	+1	+2	+3	+4
M (misforståelse)	0	0	−4	−8	−12
F (feil)	0	−1	−6	−11	−16