Sensorsamsvar på bacheloroppgaven i dataingeniørutdanning. Når nasjonale kriterier operasjonaliseres lokalt.

Grethe Sandstrak

NTNU, Institutt for datateknologi og informatikk, Trondheim, Norge

Sammendrag Nasjonale karakterbeskrivelser for bacheloroppgaven i ingeniørutdanning er generiske og fortolkes lokalt. Artikkelen beskriver en lokal vurderingsordning i et dataingeniørstudium, måler sensorsamsvaret i den og undersøker hvor variasjonen oppstår. Ordningen knytter hvert kriterium til bestemte deler av leveransen, vekter etter oppgavetype, gir studentene vurderingsgrunnlaget fra semesterstart, kalibrerer sensorene før sensuren og sender oppgaver med usikker karakterplassering til ytterligere vurdering. Materialet er alle 105 bacheloroppgavene fra 2024-2026, vurdert uavhengig av 13 interne og 6 eksterne sensorer. Samsvaret var ICC(1,1) på 0,87 for poengsum og vektet κ på 0,70 for karakterforslag utledet av poengsummene, med samme karakter på 73 % av oppgavene og enighetsgrenser fra $-4{,}4$ til 6,5 poeng av 50. Poengavviket varierte ikke påviselig med oppgavens nivå, men nær en karaktergrense kunne noen få poeng gi ulikt karakterforslag. Interne sensorer, som også hadde veiledet gruppene, ga i snitt 1,1 poeng mer enn eksterne, mens bare en liten del av variasjonen kunne knyttes til den enkelte sensor. Anslagene er usikre. Utvelgelsen til ytterligere vurdering fulgte i langt større grad hvor nær en karaktergrense oppgaven lå enn hvor langt sensorene lå fra hverandre i poeng. Studien har ingen førmåling og kan ikke tilskrive ordningen samsvaret. Bidraget er deskriptivt og tredelt, en dokumentert operasjonalisering, et samsvarsnivå målt i ordinær drift og et målbart skille mellom usikkerhet og uenighet som gjør modereringsordninger etterprøvbare.

Nøkkelord Vurdering · Bacheloroppgave · Ingeniørutdanning · Sensorkalibrering · Sensorsamsvar

1 Innledning

Bacheloroppgaven er den mest omfattende og sammensatte vurderingen i ingeniørutdanningen. I emnet som studeres her, gjennomføres den som et teambasert prosjekt med flere leveranser. Karakterfastsettingen bygger på de nasjonale karakterbeskrivelsene og vurderingskriteriene for ingeniørfag (NRT-kriteriene) [27, kap. 4.4]. Bacheloroppgaver skal sensureres av to sensorer ([1] § 11-7).

Her ligger et paradoks som er lite synlig utenfra. Karakterbeskrivelsene skal bidra til en felles forståelse og bruk av karakterskalaen for bacheloroppgaver i ingeniørfag. De er forankret i Forskrift om rammeplan for ingeniørutdanning,

men er ikke forskriftsfestet og inngår i det retningslinjene omtaler som anbefalt praksis. Siden karakterbeskrivelsene skal dekke alle ingeniørfag, er de nødvendigvis generiske. De angir hvilke kvaliteter som skal vurderes, men ikke hvor disse kommer til uttrykk i en fagspesifikk oppgave, eller hvordan sensorer med ulik faglig bakgrunn skal anvende kriteriene konsistent. Denne fortolkningen skjer derfor lokalt og er som regel lite synlig utenfor sensorkorpset.

Artikkelen undersøker en ordning der de nasjonale kriteriene er koblet til bestemte deler av leveransen, sensorene kalibreres i fellesskap før de leser, og oppgaver med usikker karakterplassering prøves på nytt i en ny kommisjon. Hva den gir av samsvar mellom sensorene, er det overordnede forskningsspørsmålet.

> *Hvilket samsvar oppnås mellom interne og eksterne sensorer i en ordning der nasjonale karakterbeskrivelser er operasjonalisert i en eksplisitt og transparent vurderingsordning for dataingeniørfaglige bacheloroppgaver, og hvor oppstår den gjenværende variasjonen?*

For å belyse både samsvaret og kilder til variasjon er det overordnede forskningsspørsmålet delt i fire underspørsmål.

F1 Hvor stort er samsvaret i poeng, og hvor mye av det overlever karaktergrensene?
F2 I hvilken grad skiller interne og eksterne sensorer seg systematisk fra hverandre?
F3 Hvor mye av variasjonen kan knyttes til den enkelte sensor?
F4 I hvilken grad følger utvelgelsen til ytterligere vurdering usikkerhet om karakteren, og i hvilken grad uenighet mellom sensorene?

2 Bakgrunn og motivasjon

2.1 Vurdering av teambaserte avsluttende prosjekter

Bacheloroppgaven omfatter flere leveransetyper, gjennomføres i grupper og vurderes av flere sensorer. Forskning på teambaserte IT-prosjekter dokumenterer utfordringer knyttet til sammensatte leveranser, individuelle bidrag, ulike sensorroller og konsistent karaktersetting [12,25]. Litteraturen gir imidlertid begrenset kunnskap om hvilket samsvar slike vurderingsordninger faktisk oppnår mellom sensorer som vurderer samme oppgave.

Arbeidene som tallfester forskjeller mellom sensorer, måler ulike forhold. Lopez mfl. [16] finner systematiske karakterforskjeller mellom bedriftsveileder, faglig veileder og kommisjon i dataingeniørutdanning, og at samme fagperson gir høyere karakter som veileder enn som kommisjonsmedlem. De tre har imidlertid ulik informasjonstilgang og bedømmer delvis ulike sider av arbeidet. McQuade mfl. [18] analyserer poengforskjeller mellom to sensorer av samme rapport i et stort materiale fra britisk biofagutdanning. Bonsaksen mfl. [7] rapporterer både eksakt karakterenighet og ICC for 67 bacheloroppgaver i ergoterapi. De måler det samme som i denne studien, men for en ren skriftlig leveranse, og uten å

beskrive hvordan kriteriene er operasjonalisert og kalibrert. Det gjenstår dermed å måle samsvaret i en ordning der operasjonaliseringen og kalibreringen er beskrevet, slik at samsvarsnivået kan holdes opp mot noe bestemt.

2.2 Kriterier gir ikke konsistens av seg selv

En vanlig respons på uenighet mellom sensorer er flere og mer detaljerte kriterier, men forskningen gir begrenset støtte for at det alene gir konsistente vurderinger. Sadler viser at forhåndsdefinerte kriterier er prinsipielt ubestemte i møte med sammensatte arbeider, og at det ikke finnes noen felles forståelse av hva kriteriebasert vurdering innebærer [21,22]. Bloxham mfl. [6] dokumenterer variasjonen empirisk. Blant 24 erfarne sensorer i fire fagdisipliner fikk bare én av 20 besvarelser samme rangering av alle sensorene innen disiplinen. Jonsson og Svingby [13] finner at vurderingsmatriser kan styrke konsistensen når de er analytiske og oppgavespesifikke og kombineres med eksempler og sensortrening, men ikke at de i seg selv sikrer valid vurdering. Gynnild [11] viser det samme i norsk sammenheng. De sju sensorene i et stort bacheloremne brukte sensurveiledningen ulikt, og alle ønsket mer kollegialt samarbeid.

To nyere arbeider peker på hver sin betingelse. Reid mfl. [19] erstattet én generisk vurderingsmatrise med seks oppgavespesifikke og ga bare skriftlige instruksjoner. Samsvaret ble lavere begge årene. Sensortrening, som Jonsson og Svingby stiller som betingelse ved siden av oppgavespesifisering, var nettopp det som manglet. Hos Chong og Romkey [8] var samsvaret høyt mellom de to som hadde utviklet vurderingsmatrisen, men falt tydelig da flere veiledere ble inkludert. Samlet tyder studiene på at kriterier alene ikke er tilstrekkelige. Kriteriene må binde skjønnet uten å oppheve det, og dette forutsetter en felles forståelse av hvordan de skal tolkes og anvendes. Dette prinsippet ligger til grunn for ordningen i kapittel 3.

2.3 Kalibrering, moderering og sensorsamsvar

Moderering omfatter tiltak som skal sikre kvaliteten på vurderinger og karakterer. Bloxham mfl. [5] foreslår å flytte hovedinnsatsen til starten av emnet og la modereringen i etterkant konsentrere seg om vanskelige grensetilfeller. Sadler [23] argumenterer for å flytte hovedvekten til kalibrering av sensorene, fordi justering av karakterer i etterkant korrigerer utfallet uten å endre vurderingsforståelsene som skapte forskjellen.

Kalibrering virker ikke av seg selv. Bjølseth mfl. [3] rapporterer at seks fagdager med drøfting av 25 eksempelbesvarelser ikke ga merkbar bedring i sensorsamsvaret ved en norsk profesjonsutdanning. Enighet om kriteriene på generelt plan sikret ikke lik anvendelse på den enkelte besvarelsen, og forfatterne peker på at effekten av kalibreringsmøter ved starten av sensureringen er lite undersøkt empirisk. Ordningen her kalibrerer mot kriteriene umiddelbart før sensuren, uten felles sensortrening på eksempelbesvarelser, og er dermed smalere enn kalibrering i Sadlers forstand, som bygger på at sensorene sammen vurderer faktiske

studentarbeider. Den kombineres med ytterligere vurdering der karakterplasseringen er usikker.

3 Vurderingsordningen

Ordningen har vært i drift fra 2022 og i nåværende form med uavhengig poengregistrering fra 2024, og de tre gjennomføringene siden da utgjør materialet. Den er utviklet for en gruppebasert bacheloroppgave i et treårig dataingeniørstudium med 33 til 37 oppgaver årlig. Oppgavene spenner fra systemutviklingsprosjekter til kunnskapsbaserte arbeider og vurderes etter ett felles sett kriterier. Hver oppgave vurderes av en intern sensor, som har veiledet gruppen gjennom semesteret, og en ekstern sensor, som møter oppgaven kun gjennom leveransen.

Ordningen består av fem faser, felles kalibrering før oppgavene leses, uavhengig sensurering med påfølgende dialog i kommisjonen, felles sensurmøte, ytterligere vurdering og avsluttende evaluering som danner grunnlag for revisjon før neste gjennomføring. Sensurveiledning, rapportmal, rammedokument for studentene og kalibreringsmateriale er utviklet sammen med veilederne og de eksterne sensorene framfor distribuert til dem, og revideres årlig på grunnlag av emneevalueringen og studentenes referansegrupper. Medvirkningen skal gi eierskap til ordningen, siden sentrale prosesser uten eierskap virker lite motiverende [24].

Kalibreringsmøtet bygger på et kalibreringsmateriale med kriteriene og forrige evalueringsrapport. Materialet advarer mot å begrunne et allerede dannet helhetsinntrykk med kriteriene i etterkant, og inneholder en differensieringsmatrise som skiller A, C og E langs fire dimensjoner, faglig nivå og teori, gjennomføring og struktur, analyse og refleksjon, og bærekraft og samfunnsnytte. Kriteriene gjennomgås og drøftes i fellesskap. Deretter registrerer intern og ekstern sensor poeng fra 0 til 50 hver for seg, uten kjennskap til den andres vurdering. Poengsummen uttrykker sensorens samlede kriteriebaserte skjønn for oppgaven som helhet og settes ikke som en sum av delpoeng per kriterium.

Sensurskjemaene med begge poengsummer og kommisjonens forslag deles med alle sensorer før sensurmøtet, der hver oppgave drøftes og A- og B-forslag begrunnes særskilt. Oppgaver med usikker karakterplassering, særlig i grenseland mellom to karakterer, identifiseres gjennom faglig drøfting i plenum, ikke ved poenggrenser, og sendes til ytterligere vurdering i en ny kommisjon. Formålet er å prøve plasseringen, ikke beregne et gjennomsnitt. Alle karakterer begrunnes skriftlig.

3.1 Operasjonalisering, vekting og åpenhet

De nasjonale kriteriene er gitt utenfra, mens koblingen til studentenes leveranse gjøres lokalt. Hvert av de seks NRT-kriteriene er knyttet til bestemte kapitler i rapportmalen og til en dimensjon i kalibreringen, og koblingen leses to veier, fra kriterium til kapittel i sensuren og fra kapittel til kriterium når oppgaven skrives. Tabell 1 viser hvordan NRT 2 er operasjonalisert i ordningen. Kriteriet krever et helhetlig systemperspektiv [27, kap. 4.4] uten å si hvor i oppgaven

det skal komme til uttrykk. Ordningen legger kravet inn som en tråd gjennom rapportmalen, slik at sensoren vet hvor det skal letes og studenten hvor det skal skrives.

Tabell 1. Operasjonaliseringen av ett nasjonalt kriterium. For hvert av de seks kriteriene fastsetter ordningen hvor i studentenes leveranse det skal innfris og hentes fra, og hvilken presisering sensorene arbeider etter. NRT 2 er vist som eksempel. Ledd merket (u) inngår bare i utviklingsoppgaver.

NRT-kriterium[a]	Hvor i mal og artefakter	Presisering i sensurveiledningen
2 Ingeniør-faglig innsikt	Kap. 3.2 valg av teknologi og utviklingsmetode. Bærekraft som gjennomgående tråd fra problemstillingen i kap. 1, via rammeverk i kap. 2 og vurderingen i kap. 3.3, til yrkesetisk refleksjon i kap. 5.1 og konklusjonen. Produkt og systemdokumentasjon (u)	Systematisk arbeidsmåte, tekniske og metodiske valg begrunnet mot problemstilling og visjonsdokument (u). Helhetlig systemperspektiv med samfunnsmessige, bærekraftige, økonomiske og etiske hensyn. Rammeverket skal være anvendt metodisk og med reell relevans, ikke bare nevnt. NRT 2 gjelder arbeidsmåten, NRT 5 resultatet

[a] Kriteriet er i tillegg knyttet til én av de fire kalibreringsdimensjonene, her bærekraft og samfunnsnytte.

Vurderingen bygger på tre pilarer, produkt, rapport og prosess. Prosessen teller alltid 20 %, mens produktandelen følger utviklingsinnslaget, fra 40 % i en fullverdig utviklingsoppgave til null i et rent kunnskapsbasert arbeid, og rapporten tar resten.

Kategorivektingen er bindende for sensorens samlede skjønn og fastsatt før arbeidet begynner, mens de nasjonale vektingsintervallene per NRT-kriterium bare gjengis som en veiledende profil per oppgavetype. Faste prosentsatser ville redusert et sammensatt faglig skjønn til en mekanisk beregning [22]. Der det nasjonale sensurskjemaet åpner for at vektingen avtales med kandidatene, binder ordningen den til leveransetypen, slik at lik leveranseprofil gir lik vekting uavhengig av veileder.

Studentene har fullt innsyn i vurderingsgrunnlaget fra semesterstart, men tilgang gir ikke i seg selv forståelse av faglige kvalitetsstandarder [20], og hva studentene får ut av materialet, undersøkes ikke her.

4 Metode

4.1 Design, materiale og variabler

Undersøkelsen er en retrospektiv analyse av sensurdata fra ordinær drift av ordningen beskrevet i kapittel 3. Materialet omfatter alle de 105 oppgavene fra

2024-2026, fordelt med 33 til 37 årlig og vurdert av 13 interne og 6 eksterne sensorer. Uavhengig registrering av intern og ekstern sensors poengsum, avgitt før sensurmøtet, ble innført i 2024, så bare gjennomføringer siden da inngår i studien. Gjennomføringene er slått sammen fordi samsvaret målt i poeng var stabilt, se kapittel 5.

Dataene er hentet fra ordinære sensurskjemaer og er knyttet til oppgaver, ikke kandidater. Ingen studentopplysninger inngår, og sensorene er pseudonymisert. De er informert og kan trekke egne vurderinger. Behandlingen er vurdert av Sikt og hjemlet i personvernforordningen artikkel 6 nr. 1 bokstav e.

Hver oppgave har to uavhengige poengsummer fra 0 til 50, som beskrevet i kapittel 3. Karaktergrensene gjelder den summerte skalaen fra 0 til 100, der A ≥ 90, B ≥ 80, C ≥ 60, D ≥ 40 og E ≥ 20. Sensorene registrerte ikke egne karakterforslag. Forslagene er derfor utledet ved å doble den enkeltes poengsum og lese den mot de samme grensene, mens summen av de to gir avstanden til nærmeste grense (utledet plassering). Kappa er dermed en omregning av poengene og ikke en uavhengig måling. Differansen er intern minus ekstern sensors poengsum og viser hvem som ga mest. Poengavviket er den samme størrelsen uten fortegn og viser hvor langt sensorene lå fra hverandre. Siden karakterintervallene har ulik bredde, uttrykker avstanden nærhet til en grense, ikke faglig nivå. Sensorrolle, år, pseudonymisert sensoridentitet og utvelgelse til ytterligere vurdering er også registrert per oppgave.

4.2 Analyser og begrunnelse for målvalg

De fleste oppgavene ligger høyt på skalaen. Gjennomsnittsbaserte mål forutsetter en jevnere fordeling, så der forutsetningen har betydning, er hvert resultat kontrollert med et rangbasert mål. ICC og κ sier hvor sterkt sensorene følger hverandre, men ikke hvor mange poeng som skiller dem, og bare det siste lar seg holde opp mot bredden på karakterintervallene. Samsvaret er derfor også oppgitt i poeng.

F1. Samsvaret i poeng er målt med ICC(1,1), en enveis tilfeldig modell for enkeltmålinger [26], som legger all variasjon mellom sensorene, også rolleforskjellen, i feilleddet. Dette er det mest konservative valget. Estimatet gjelder kommisjoner med én intern og én ekstern sensor. Robusthetssjekker uten den laveste oppgaven er oppgitt i kapittel 5. Karakterforslagene er sammenlignet med Cohens vektede kappa med lineære vekter (κ) [9]. Analysen er supplert med Bland-Altman-metodikk [4], som gir gjennomsnittsdifferansen og grensene for 95 % av differansene. Normalitet er vurdert med Shapiro-Wilks test og empiriske 2,5- og 97,5-persentiler, og nivåavhengighet ved å regressere differansen på gjennomsnittet og med Spearmans rangkorrelasjon på poengavviket.

F2. Siden hver sensor vurderer flere oppgaver, er rolleforskjellen anslått i en blandet modell med kryssede tilfeldige effekter for intern og ekstern sensor. Som supplement er forskjellen prøvd med oppgave som analyseenhet, med parvis t-test, Cohens d_z som standardisert effektstørrelse [10] og Wilcoxons test for parvise rangeringer [28], som ikke forutsetter normalitet. Differansene er tilnærmet symmetriske, skjevhet 0,40, slik Wilcoxons test forutsetter.

F3. Poengavviket er dekomponert med sensor som tilfeldig faktor i en enveis modell for hver sensorgruppe, og sensorandelen er sensorvariansen delt på totalvariansen. Residualen omfatter forskjeller mellom oppgavene og øvrig uforklart variasjon. Den blandede modellen i F2 gir tilsvarende komponenter for differansen, med begge sensorgruppene i samme modell. Siden poengavviket alltid er positivt, er normalmodellens forutsetninger bare tilnærmet oppfylt.

F4. De 12 oppgavene som ble sendt til ytterligere vurdering, er sammenlignet med de øvrige 93 med hensyn til avstand til nærmeste karaktergrense og poengavvik ved Mann-Whitneys U-test [17], som ikke forutsetter normalfordeling. Andelen med ulikt karakterforslag er prøvd med Fishers eksakte test [2], siden ett forventet celletall er under 5. Begge poengmålene prøves fordi begge var like tilgjengelige ved utvelgelsen. Hvordan utvelgelsen fordeler seg etter karakterforslag og avstand til grensen samlet, er i tillegg beskrevet uten test. Grensen på to poeng er valgt etter at dataene er sett.

Flere analyser er gjennomført på samme materiale uten innstramming av signifikansnivået. Årene er beskrevet og ikke sammenlignet statistisk, så antallet hypotesetester er uendret. Alle resultater som tolkes som signifikante, har $p \leq 0{,}003$ og tåler Bonferronikorreksjon for de elleve hypotesetestene med oppgitt p-verdi, der grensen blir $0{,}05/11 = 0{,}0045$. Analysene er gjennomført i IBM SPSS Statistics versjon 30.

4.3 Analysens rekkevidde og begrensninger

Materialet omfatter alle oppgavene i de tre gjennomføringene, og førstegangsvurderingene ble avgitt uavhengig før sensurmøtet. Samsvarsnivået, rolleforskjellen og variasjonen knyttet til sensoridentitet kan derfor beskrives. Fem forbehold styrer likevel tolkningen.

- *Samsvaret kan ikke tilskrives ordningen.* Alle tre gjennomføringene ligger etter at ordningen kom i drift, så studien kan ikke fastslå at den har skapt det observerte samsvarsnivået.
- *Samsvar viser ikke om nivået er riktig.* Vurderingene er produsert innenfor samme ordning, så materialet kan ikke prøve karakternivået mot de nasjonale kriteriene. Det er to ulike slutninger, og hver av dem krever sin egen dokumentasjon [14]. Kappa er utledet av poengene og bekrefter dem ikke uavhengig.
- *Observasjonene er ikke uavhengige.* Hver ekstern sensor vurderer 11 til 25 oppgaver og hver intern 2 til 18. Denne ubalansen påvirker usikkerhetsestimatene, og enighetsgrensene har derfor ikke nødvendigvis nominell dekningsgrad. Kommisjonene krysser likevel nok til at de tilfeldige effektene kan skilles, siden hver ekstern sensor har sittet med fire til seks interne og ti av de tretten interne med minst to eksterne.
- *Flere relevante variabler mangler.* Oppgavetype, anvendt vekting, sensorenes erfaring og begrunnelsen for hver utvelgelse er ikke registrert, og den nye kommisjonen kjente de opprinnelige forslagene, så prøvingen av plasseringen var ikke uavhengig.

– *Forskjellene er beskrevet, ikke prøvd.* Varianskomponentene for interne og eksterne sensorer er beregnet i separate modeller uten statistisk sammenligning, og estimatet for eksterne bygger på seks personer.

5 Resultater

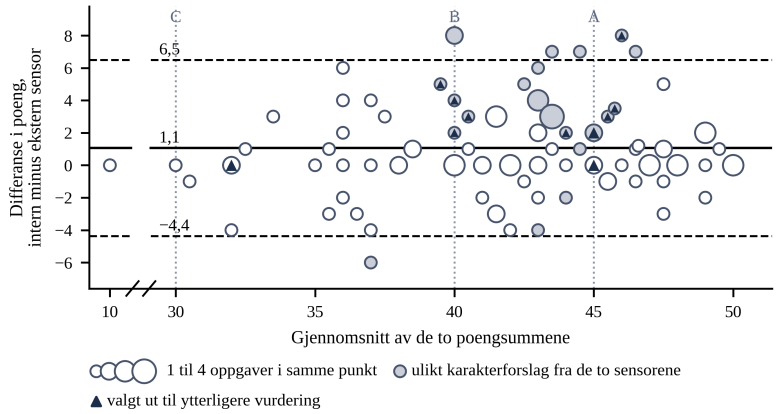
Tabell 2 oppsummerer målingene, og Figur 1 viser differansen mellom intern og ekstern sensor for hver oppgave, på skalaen 0 til 50.

Tabell 2. Sensorsamsvar og variasjonskilder. Positiv differanse betyr at intern sensor ga høyest poengsum. Tallene i F3 er andelen av variasjonen som følger sensoridentitet. F4 sammenligner de 12 utvalgte oppgavene med de øvrige 93. Hakeparentes angir 95 % konfidensintervall.

	Mål	Resultat
F1	ICC(1,1), poengsum	0,87 [0,82 til 0,91]
	Vektet κ, utledet plassering	0,70 [0,60 til 0,80]
	Grenser for enighet	−4,4 [−5,3 til −3,4] og 6,5 [5,6 til 7,4]
	Nivåavhengighet, retning	$R^2 = 0{,}010$; $p = 0{,}31$
	Nivåavhengighet, spredning	Spearmans $\rho = -0{,}10$; $p = 0{,}29$
	Poengavvik ved ulik plassering	4,3 mot 1,3 poeng av 50
F2	Differanse, blandet modell	1,09 poeng [−0,03 til 2,21]; $p = 0{,}054$
	Differanse, per oppgave	1,06 poeng [0,53 til 1,60], $p < 0{,}001$, $d_z = 0{,}38$
	Positiv differanse	14 av 19 sensorer
F3	Eksterne, differanse	6,9 % (blandet modell)
	Interne, differanse	0,2 % (blandet modell)
	Eksterne, poengavvik	14,8 % [3 til 58 %], $p = 0{,}002$
	Interne, poengavvik	5,0 % [0 til 27 %], $p = 0{,}18$
F4	Avstand til grense	1,0 mot 4,3 av 100, $p < 0{,}001$, $r = 0{,}41$
	Ulik plassering	10 av 28 mot 2 av 77, $p < 0{,}001$, $\varphi = 0{,}46$
	Poengavvik	2,9 mot 2,0 av 50, $p = 0{,}11$, $r = 0{,}15$

ICC og vektet κ er 1 ved full overensstemmelse. Grensene for enighet ventes å romme 95 % av differansene. R^2 og ρ nær null betyr at differansen og poengavviket ikke følger oppgavens nivå. Cohens grenser for liten og moderat effekt er 0,2 og 0,5 for d_z og 0,1 og 0,3 for r og φ [10].

F1. De modellbaserte enighetsgrensene går fra −4,4 til 6,5 poeng, et spenn på 10,9. Differansene avviker svakt fra normalfordelingen, $W = 0{,}955$, $p = 0{,}001$, på grunn av en markert opphopning i null, der 33 sensorpar ga samme poengsum. De empiriske persentilene 2,5 og 97,5 er −4,0 og 8,0 poeng, et spenn på 12,0, så den nedre grensen treffer, mens den øvre er noe lav og grensene bør leses som et minimumsanslag. De to sensorene fikk samme karakter på 77 av 105 oppgaver. Materialet ligger høyt på skalaen, med 97 av 105 oppgaver på 35 poeng eller mer i gjennomsnitt, og den summerte poengsummen gir 34 A, 44 B, 26 C og én E. Per år er poengavviket 2,1, 1,9 og 2,2 poeng og andelen med



Figur 1. Differanse mot gjennomsnitt etter Bland-Altman for de 105 oppgavene. Heltrukket linje er gjennomsnittsdifferansen, stiplede linjer grensene for enighet og loddrette linjer nedre karaktergrense for C, B og A. De utvalgte oppgavene ligger ved de loddrette linjene, ikke ved de stiplede.

samme karakterforslag 69, 76 og 76 %, altså stabilt over de tre gjennomføringene. Uten den laveste oppgaven, med 10 poeng fra begge sensorer mot 30 for den nest laveste, er ICC 0,83 mot 0,87 og vektet κ 0,69 mot 0,70, begge innenfor konfidensintervallene, mens enighetsgrensene er uendret. Enigheten om denne oppgaven utvider spredningen mellom oppgavene og løfter dermed koeffisientene. Toveismodeller gir 0,876 og 0,889, mot 0,875 i enveismodellen.

F2. Intern sensor gir i snitt om lag ett poeng mer enn ekstern, 1,09 poeng i den blandede modellen, med et konfidensintervall som så vidt omslutter null. Med oppgave som analyseenhet er anslaget 1,06 poeng og intervallet vesentlig smalere, fordi de 105 oppgavene da regnes som uavhengige. Gjennomsnittsdifferansen er positiv for 14 av de 19 sensorene og null for én, og ingen sensor med mer enn to vurderte oppgaver har en negativ differanse større enn 0,4 poeng. Wilcoxons test gir samme konklusjon, $p < 0{,}001$. Som standardisert effektstørrelse ligger forskjellen mellom liten og moderat, $d_z = 0{,}38$.

F3. I den blandede modellen, der begge sensorgruppene inngår samtidig, svarer sensoridentitet for 6,9 % av variasjonen i differansen blant de eksterne og 0,2 % blant de interne. Enveismodellene på poengavviket, én for hver gruppe, gir 14,8 % og 5,0 %. Anslagene er usikre, intervallet for eksterne går fra 3 til 58 %, og de to gruppene er ikke sammenlignet statistisk.

F4. De utvalgte oppgavene lå klart nærmere en karaktergrense enn de øvrige, i snitt 1,0 mot 4,3 poeng på den summerte skalaen, $r = 0{,}41$. Blant de 28 oppgavene med ulikt karakterforslag ble 10 sendt videre, mot 2 av de 77 med likt forslag. Alle de ti oppgavene med ulikt forslag som ble sendt videre, hadde en sum innenfor to poeng fra en grense, mot 5 av de 18 med ulikt forslag som ikke ble sendt videre. Av de 16 oppgavene med likt forslag og sum like nær en grense ble én sendt videre. Poengavviket skilte de to gruppene av oppgaver

langt svakere, i snitt 2,9 mot 2,0 poeng, $r = 0{,}15$, $p = 0{,}11$. Utfallet er registrert for alle tolv. De to med likt forslag ble opprettholdt, og ni av de ti med ulikt forslag endte på intern sensors trinn. Blant de 18 som ikke ble sendt videre, gikk det motsatt vei, 12 mot 6 i eksterns favør. Sluttkarakteren var lik den utledede karakteren fra intern sensors poengsum på 92 av 105 oppgaver og lik eksterns på 90.

6 Diskusjon

6.1 Sammenligning med andre studier

Sensorene i studien er i stor grad enige. Bloxham mfl. [6] viste at kriterier alene ikke gir lik vurdering, og her er kriteriene gjort fagspesifikke, knyttet til bestemte deler av leveransen og gjennomgått i fellesskap før lesing. Siden kalibreringen her er smalere enn den Bjølseth mfl. prøvde, kan operasjonaliseringen, rapportmalen og det smale karakterspennet ha bidratt like mye.

Det finnes ingen etablerte normer for godt samsvar ved sammensatte avsluttende arbeider, men tre studier gir et sammenligningsgrunnlag. Nærmest ligger Bonsaksen mfl. [7], der intern sensor også hadde veiledet studentene. De rapporterer ICC 0,81 for gjennomsnittet av de to sensorene, i en modell som holder den systematiske forskjellen mellom rollene utenfor feilleddet. Regnet på karakterskalaen gir materialet her fra 0,81 til 0,90, alt etter om tallet gjelder én sensor eller gjennomsnittet. Modellvalg og analyseenhet flytter dermed tallet like mye som sammenligningen med deres studie, og intervallene overlapper, så forskjellen er ikke sikker. Sensorene deres foreslo samme karakter på om lag halvparten av 67 oppgaver, mot nær tre av fire her, men de satte bokstavkarakter direkte, mens forslagene her er utledet av poengsummene. Chong og Romkey [8] fant vektet κ på samme nivå som her, men på enkeltkriterier og for 12 oppgaver vurdert av de to fagpersonene som selv hadde laget vurderingsmatrisen, så verdiene måler ikke det samme. McQuade mfl. [18] fant at sensorene skilte med mer enn ti prosentpoeng på 22,8 % av 897 bachelorrapporter. Med samme grense, mer enn 5 av 50 poeng, gjelder det 8,6 % her.

Bare sammenligningen med Bonsaksen mfl. er gjort på noenlunde like vilkår, og både koeffisientene og andelen over en fast grense avhenger av hvilke oppgaver som inngår og hvor stor del av skalaen som brukes. Samsvar bør derfor oppgis med modell, skala og om tallet gjelder én sensor eller gjennomsnittet, og helst med uenighet i skalaens egen enhet.

6.2 Fra poeng til karakter

Samsvaret er lavere for karakter enn for poeng, vektet κ 0,70 mot ICC 0,87. Noe av forskjellen følger av at poengene deles inn i karakterer. Uenigheten i poeng varierer ikke påviselig med oppgavenes nivå, men konsekvensene av den gjør det. Ligger oppgaven nær en grense, kan et poengavvik på et par poeng gi ulike karakterforslag, mens et større poengavvik midt i et intervall ikke trenger

å endre noe. Sensorer som endte på ulikt forslag, skilte i snitt 4,3 poeng av 50, mot 1,3 der forslagene var like (Tabell 2). Ulike forslag skyldes dels at sensorene vurderte oppgaven ulikt, og dels at en moderat forskjell falt over en grense.

Hvor mye dette betyr i praksis, avhenger av hvor brede intervallene er. Intervallet for B er ti poeng bredt på den summerte skalaen og rommer flest oppgaver. Standardavviket for differansene er 2,8 poeng, så hadde den ene sensoren vurdert som den andre, ville den summerte poengsummen anslagsvis i 95 % av tilfellene endret seg med mindre enn om lag 5,4 poeng, drøyt halvparten av intervallet.

Lauvås mfl. [15] finner at de intervjuede IT-studentene foretrekker gradert skala, blant annet fordi den gir konkret tilbakemelding på mestringsnivå. Resultatene her viser hva graderingen koster ved en sammensatt leveranse, siden små poengforskjeller kan gi ulike karakterforslag når oppgaven ligger nær en grense. Om gradert skala er egnet, blir dermed et spørsmål om vurderingsdesign og ikke bare om sensorenes vurdering.

6.3 Rolleforskjellen og dobbeltrollen

At veileder som intern sensor gir noe høyere karakter, er kjent fra tidligere studier [7,16]. Det materialet her tilfører, er størrelsen i en ordning der kriteriene er operasjonalisert og kalibrert. Om lag ett poeng av 50 er lite målt mot karakterintervallene, men slår likevel ut i fordelingen av karakterforslag. Intern sensors poengsum tilsvarer A på 47 av de 105 oppgavene, ekstern på 30, mens ekstern til gjengjeld foreslo B og C oftere. Designet identifiserer ikke årsaken. Intern sensor har tilgang til prosesskunnskap som ekstern bare møter gjennom dokumentasjonen, har veiledet gruppen gjennom semesteret og kjenner det generelle nivået i kullet, noe som kan gi mer vekt på normrelatert sammenligning [7]. Sensorgruppene har dessuten ulik faglig sammensetning, så mildhet knyttet til veilederrollen er bare én mulig forklaring. Retningen er likevel ikke særegen for dette materialet, siden ekstern sensor hos Bonsaksen mfl. foreslo høyere karakter i bare ett av 34 tilfeller med uenighet, mot tre av 28 her.

Bonsaksen mfl. peker på muligheten for at veiledere ikke bør være interne sensorer på oppgaver de selv har veiledet. Det fjerner rolleforskjellen, men også prosesskunnskapen. Ordningen registrerer i stedet forskjellen hvert år, slik at den kan følges over tid.

6.4 Variasjon hos den enkelte sensor

Bare en liten del av variasjonen i poengavviket kan knyttes til sensoridentitet, og andelen er usikker i begge gruppene. Siden poengavviket deles av sensorene på en oppgave, kan andelen for en ekstern sensor delvis gjenspeile hvilke interne sensorer og oppgavetyper vedkommende fikk. De interne arbeider med kriteriene gjennom semesteret, mens de eksterne først møter dem i kalibreringsmøtet, til gjengjeld med flere oppgaver hver (17,5 mot 8,1). Estimatet peker likevel mot mer variasjon blant de eksterne. Med forbeholdene i kapittel 4.3, og siden flere observasjoner per ekstern sensor gjør reelle forskjeller lettere å oppdage, er dette et mønster som bør undersøkes videre og ikke et resultat.

6.5 Usikkerhet eller uenighet

Utvelgelsen var ikke blind, og usikker plassering nær en grense var det uttalte kriteriet, så grensenærheten overrasker ikke. Mer opplysende er hva utvelgelsen ikke fulgte. Av de tre opplysningene som lå like tilgjengelige (grensenærheten, karakterforslagene og poengavviket), er poengavviket det eneste som direkte uttrykker hvor ulikt sensorene leste arbeidet, og det skilte gruppene langt svakere, $r = 0{,}15$ mot 0,41. Kollegiet reagerte mer på at plasseringen var vanskelig enn på avstanden mellom sensorene.

Hva usikkerheten besto i, presiseres av hvordan utvelgelsen fordelte seg etter karakterforslag og grensenærhet (F4). Verken ulikt forslag eller grensenærhet var nok alene. Det som skilte de utvalgte, var kombinasjonen, at sensorene havnet på hver sin side av en grense og summen lå så nær den at det var uklart hvilken karakter som var riktig. De to unntakene hadde identisk poengsum fra begge sensorene, den ene presis på grensen til A. Kriteriet ble ikke anvendt som en regel, men i praksis som en kombinasjon av de to forholdene.

Dette skiller ordningen fra en terskelbasert kontroll som hos McQuade mfl. [18], der poengavviket styrer hvilke oppgaver som prøves om igjen, og gir en måte å undersøke om moderering i etterkant faktisk konsentreres om vanskelige grensetilfeller, slik Bloxham mfl. [5] foreslår. De to veiene løser uenighet i hver sin retning, og siden den nye kommisjonen kjente begge forslagene, kan retningen ikke skilles fra rolleforskjellen i F2.

7 Konklusjon

De to sensorene ga samme poengsum på 31 % av oppgavene og samme karakter på 73 %, og de skilte med fire poeng eller mindre av 50 på 93 av de 105 oppgavene. Uenigheten følger ikke oppgavens nivå, men slår ulikt ut, siden et par poeng ikke endrer noe midt i et karakterintervall og kan avgjøre karakteren nær en grense. Intern sensor, som også hadde veiledet gruppen, ga gjennomgående om lag ett poeng mer enn ekstern. Forskjellen følger altså rollen og er liten målt mot bredden på karakterintervallene, mens den enkelte sensors identitet forklarer lite av hvor langt de to lå fra hverandre. Retningen er tydelig, mens tallene er usikre og beskriver dette sensorkorpset, ikke sensorroller generelt. Oppgavene ble sendt videre når sensorene havnet på hver sin side av en karaktergrense og de to poengsummene til sammen lå tett på den, og i langt mindre grad når de lå langt fra hverandre. Utvelgelsen sporet altså usikkerhet om plasseringen framfor uenighet om kvaliteten, og det skillet lar seg måle og viser om moderering i etterkant treffer grensetilfellene.

Ordningen kan tas i bruk i andre emner uten at de nasjonale kriteriene endres, men må tilpasses egne leveranser og forhold, og samsvarsnivået er målt i ordinær drift og kan sammenlignes med. Den binder hvor i leveransen hvert kriterium hentes fra, men ikke hva som er god kvalitet, så det faglige skjønnet ligger fortsatt hos veilederne. Gjennom medvirkning i utviklingen får de også eierskap til ordningen. Nettopp fordi skjønnet forblir lokalt, må samsvaret måles jevnlig, ellers går endringer i vurderingspraksis upåaktet hen.

Takk Takk til de eksterne og interne sensorene, som gjennom fem gjennomføringer har bidratt i utvikling av vurderingsordningen sammen med forfatteren.

Interessekonflikt Forfatteren er emneansvarlig for emnet som studeres, men har ikke vært sensor for noen av oppgavene i materialet.

Referanser

1. Lov om universiteter og høyskoler (universitets- og høyskoleloven) (2024), https://lovdata.no/lov/2024-03-08-9
2. Agresti, A.: A Survey of Exact Inference for Contingency Tables. Statistical Science **7**(1), 131–153 (1992). https://doi.org/10.1214/ss/1177011454
3. Bjølseth, G., Havnes, A., Lauvås, P.: Lavt sensorsamsvar – kan det bedres? Uniped **34**(4), 4–16 (2011). https://doi.org/10.18261/issn1893-8981-2011-04-01
4. Bland, J.M., Altman, D.G.: Statistical Methods for Assessing Agreement between two Methods of Clinical Measurement. The Lancet **327**(8476), 307–310 (1986). https://doi.org/10.1016/S0140-6736(86)90837-8
5. Bloxham, S., Hughes, C., Adie, L.: What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. Assessment & Evaluation in Higher Education **41**(4), 638–653 (2016). https://doi.org/10.1080/02602938.2015.1039932
6. Bloxham, S., den Outer, B., Hudson, J., Price, M.: Let's stop the pretence of consistent marking: exploring the multiple limitations of assessment criteria. Assessment & Evaluation in Higher Education **41**(3), 466–481 (2016). https://doi.org/10.1080/02602938.2015.1024607
7. Bonsaksen, T., Thørrisen, M., Sveen, U., Kjeken, I., Aas, R., Lund, A.: Grade correspondence between internal and external examiners of occupational therapy students' bachelor theses. Uniped **41**(3), 319–330 (2018). https://doi.org/10.18261/issn.1893-8981-2018-03-11
8. Chong, A., Romkey, L.: Testing inter-rater reliability in rubrics for large scale undergraduate independent projects. Proceedings of the Canadian Engineering Education Association (CEEA) (2016). https://doi.org/10.24908/pceea.v0i0.6465
9. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychological Bulletin **70**(4), 213–220 (1968). https://doi.org/10.1037/h0026256
10. Cohen, J.: Statistical Power Analysis for the Behavioral Sciences. Lawrence Erlbaum Associates, Hillsdale, NJ, 2 edn. (1988). https://doi.org/10.4324/9780203771587
11. Gynnild, V.: Jeg tror jeg brukte den som taus kunnskap, om du skjønner meg: En studie av sensurveiledning på bachelornivå. Uniped **45**(1), 39–52 (2022). https://doi.org/10.18261/uniped.45.1.5
12. Hooshangi, S., Shakil, A., Riddle, S., Aydin, I., Nasir, N., Parupudi, T., Rehman, A., Scott, M.J., Vahrenhold, J., Weerasinghe, A., Wu, X.: Evaluating Assessment Practices in Team-Based Computing Capstone Projects. In: Proceedings of the 2025 Working Group Reports on Innovation and Technology in Computer Science Education. pp. 277–312. ACM, Nijmegen Netherlands (2025). https://doi.org/10.1145/3760545.3783974
13. Jonsson, A., Svingby, G.: The use of scoring rubrics: Reliability, validity and educational consequences. Educational Research Review **2**(2), 130–144 (2007). https://doi.org/10.1016/j.edurev.2007.05.002

14. Kane, M.T.: Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement **50**(1), 1–73 (2013). https://doi.org/10.1111/jedm.12000
15. Lauvås, P.J., Mirmotahari, O., Sandtrø, T., Berg, Y., Lindsjørn, Y.: Gradert karakterskala eller Bestått/Ikke bestått? Refleksjoner fra studenter i høyere IT-utdanning. Norsk IKT-konferanse for forskning og utdanning (NIKT) **37**(4) (2025). https://doi.org/10.5324/bf1td211
16. Lopez, A., Castaño, M.A., Ibáñez, M.V., Sanz, I., Museros, L., Grangel, R.: Improving the Assessment of Capstone Projects in the Bachelor's Degree in Computer Engineering. ACM Transactions on Computing Education **25**(2), Article 14 (2025). https://doi.org/10.1145/3722230
17. Mann, H.B., Whitney, D.R.: On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. The Annals of Mathematical Statistics **18**(1), 50–60 (1947). https://doi.org/10.1214/aoms/1177730491
18. McQuade, R., Kometa, S., Brown, J., Bevitt, D., Hall, J.: Research project assessments and supervisor marking: maintaining academic rigour through robust reconciliation processes. Assessment & Evaluation in Higher Education **45**(8), 1181–1191 (2020). https://doi.org/10.1080/02602938.2020.1726284
19. Reid, K.J., Chiavaroli, N.G., Bilszta, J.L.C.: Assessing a capstone research project in medical training: Examiner consistency using generic versus domain-specific rubrics. Journal of Medical Education and Curricular Development **9** (2022). https://doi.org/10.1177/23821205221081813
20. Rust, C., Price, M., O'Donovan, B.: Improving Students' Learning by Developing their Understanding of Assessment Criteria and Processes. Assessment & Evaluation in Higher Education **28**(2), 147–164 (2003). https://doi.org/10.1080/02602930301671
21. Sadler, D.R.: Interpretations of criteria-based assessment and grading in higher education. Assessment & Evaluation in Higher Education **30**(2), 175–194 (2005). https://doi.org/10.1080/0260293042000264262
22. Sadler, D.R.: Indeterminacy in the use of preset criteria for assessment and grading. Assessment & Evaluation in Higher Education **34**(2), 159–179 (2009). https://doi.org/10.1080/02602930801956059
23. Sadler, D.R.: Assuring academic achievement standards: from moderation to calibration. Assessment in Education: Principles, Policy & Practice **20**(1), 5–19 (2013). https://doi.org/10.1080/0969594X.2012.714742
24. Sandstrak, G., Klefstad, B.: Referansegrupper som verktøy for kontinuerlig kvalitetsforbedring. NIKT: Norsk IKT-konferanse for forskning og utdanning (2021)
25. Scott, M.J., Garcia, R., Lovellette, E., Russell, S., Wu, X., Alshaigy, B., Keuning, H., Schulz, S., Mitchell, A., Gutica, M., Parthasarathy, P.D.: Exploring Assessment Approaches for Student Group Projects in Undergraduate Computing Education. ACM Transactions on Computing Education (2026). https://doi.org/10.1145/3822407
26. Shrout, P.E., Fleiss, J.L.: Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin **86**(2), 420–428 (1979). https://doi.org/10.1037/0033-2909.86.2.420
27. UHR-MNT: Nasjonale retningslinjer for ingeniørutdanning. Vedtatt 25. november 2020, oppdatert mai 2023. Tech. rep., UHR-MNT (2023)
28. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin **1**(6), 80–83 (1945). https://doi.org/10.2307/3001968