Va fan får jag för pengarna, Olof?

Ok! Nedan följer ett försök att sammanfatta vad det egentligen var ni gjorde på det där första (skälvande?) ämneslagsmötet i augusti. För er som redan har förträngt den första arbetsveckans vedermödor: ni bedömde elevuppsatser skrivna som svar på uppsatsdelen från NP i kursen Engelska 5 (från VT 2016). Målet var att ta fram **ankaruppsatser** – elevtexter med ett så rättvist och stabilt betyg som möjligt – som kan användas för att kalibrera min comparative judgement-modell.

Ankaruppsatserna fungerar som fasta referenspunkter när systemet jämför och rangordnar nya elevuppsatser. Processen går till så att "vanliga" uppsatser (utan känt betyg) jämförs mot varandra och med de inblandade ankaruppsatserna vars betyg redan är fastställda. När alla uppsatser har jämförts med varandra kan vi ranka dem genom att analysera vilka uppsatser som vann respektive förlorade mest matcher och hur svåra eller lätta motståndarna i dessa matcher var. Eftersom ankaruppsatsernas betyg är kända kan systemet sedan beräkna det mest sannolika betyget för alla övriga uppsatser baserat på var de hamnar relativt ankarna i rankningen.

Metod för att beräkna konsensusbetyg

För att göra bedömningarna jämförbara används en statistisk modell som tar hänsyn till att olika lärare kan vara mer eller mindre strikta. I stället för att behandla betyg som exakta tal betraktar modellen betygsstegen (F, F+, E, E+, D–, D+, C–, C+, B, A) som en ordning[1]. Modellen kallas en **kumulativ logistisk modell** och fungerar i flera steg:

1. Varje uppsats antas ha en underliggande kvalitetsnivå som vi vill uppskatta genom att kombinera alla bedömningar.
2. Varje bedömare antas ha en personlig stränghet. En sträng lärare sänker betygen jämfört med genomsnittet, en generös lärare höjer dem.
3. Modellen beräknar var gränserna mellan betygsstegen bör ligga baserat på era faktiska bedömningar. Den upptäcker till exempel att steget från C- till C+ är det största i hela skalan[2].
4. Genom att väga samman alla bedömningar och justera för varje bedömares stränghetsprofil får vi ett **konsensusbetyg** för varje uppsats.

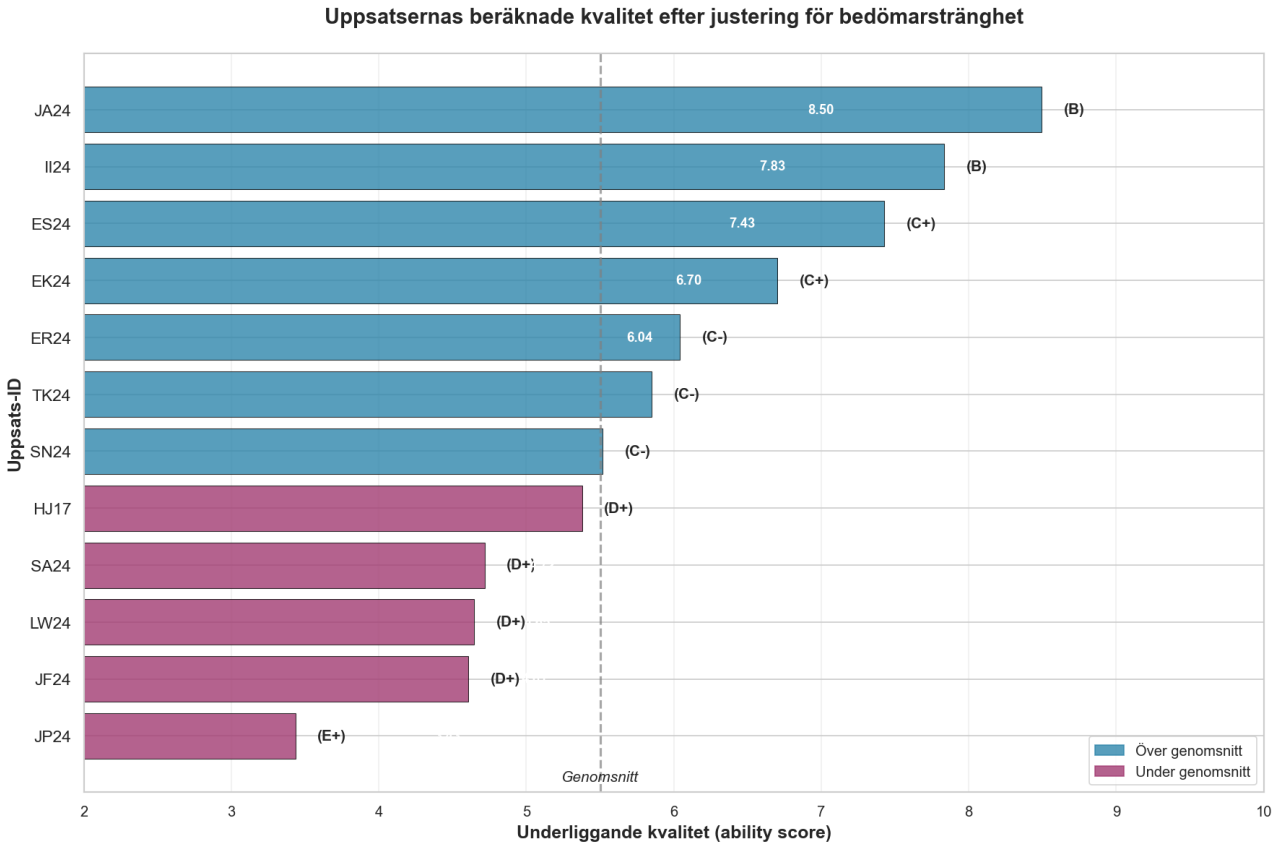
På så sätt får varje uppsats ett betyg som representerar dess mest sannolika kvalitetsnivå baserat på alla bedömningar, efter att ha rensat bort effekten av individuella bedömares stränghet. Man kan se det som det betyg uppsatsen troligen skulle få av en "neutral" bedömare – eller som panelens samlade bedömning efter justering för era olika bedömarprofiler. Dessa justerade betyg används sedan för att rangordna uppsatserna och välja ut ankaruppsatserna för AI-systemet.

---

[1] **Ordning:** Modellen bryr sig bara om rangordningen mellan betygen (F är lägre än E, som är lägre än D, och så vidare). Den antar inte att avståndet mellan F och E är lika stort som mellan D och C – avstånden får vara olika stora, vilket våra resultat också visar.

[2] **Betygsgränserna anpassas efter data:** Istället för att anta att alla betygssteg är lika stora låter modellen era bedömningar avgöra hur stora kvalitetshoppen mellan betygen faktiskt är. Via vårt förfarande kan vi skönja ett intressant mönster – steget från C- till C+ är det största i hela skalan, vilket jag och säkert många med mig känner igen från vår undervisning.

Resultat: Uppsatsernas rangordning



Figur 1: Uppsatsernas beräknade kvalitet efter justering för bedömarstränghet. Staplarna visar den underliggande kvaliteten (ability score) sorterat från lägst till högst. Blå staplar indikerar uppsatser över genomsnittet (5.5), lila staplar under. Konsensusbetyget visas i parentes efter varje uppsats-ID. JA24 ligger klart högst med värde 8.50 (B), följt av II24 med 7.83 (B). JP24 har lägst värde med 3.43 (E+).

Uppsatser med högst och lägst kvalitetsvärden

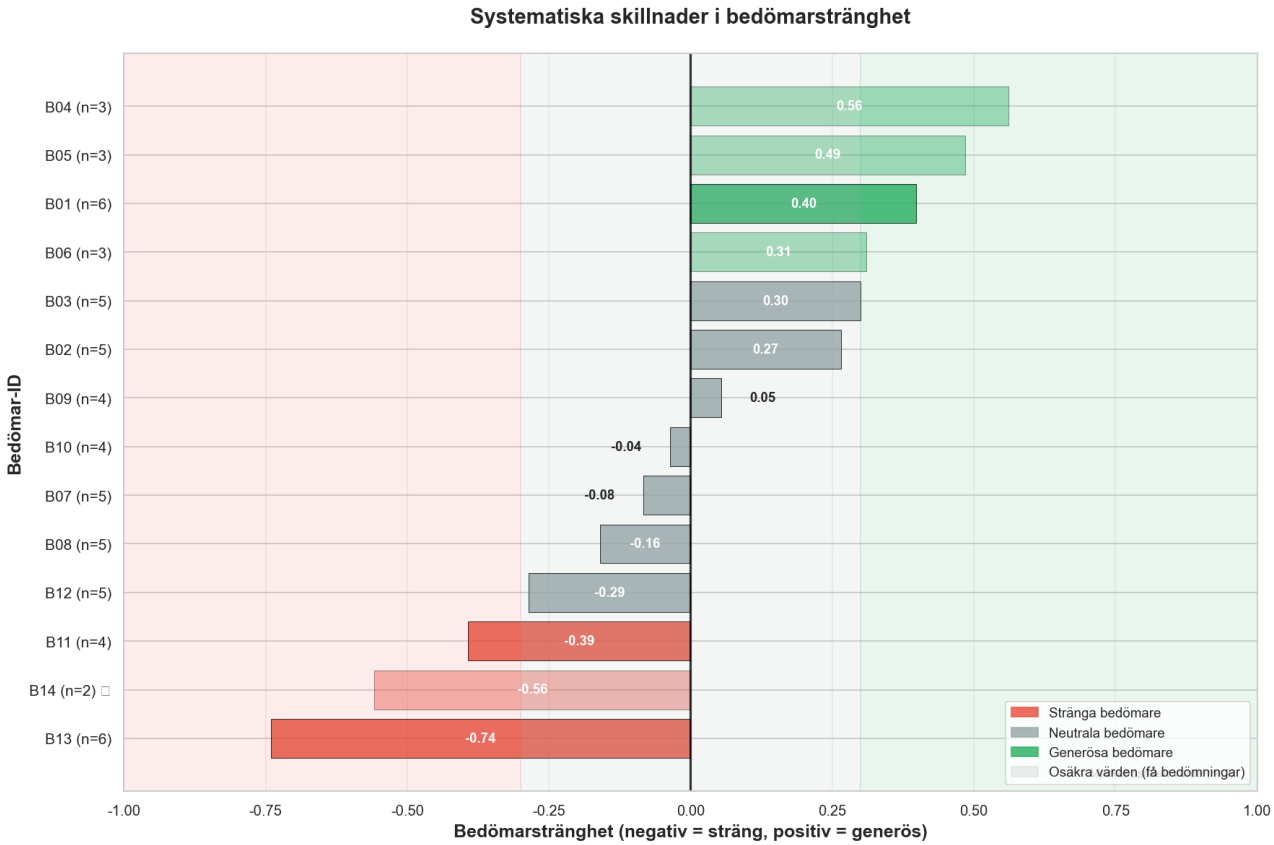
Uppsats-ID	Konsensusbetyg	Underliggande kvalitet[3]	Konfidens[4]
Högst rankade			
JA24	**B**	8.50	0.39
II24	**B**	7.83	0.37
ES24	**C+**	7.43	0.31
Lägst rankade			
LW24	**D+**	4.65	0.23
JF24	**D+**	4.61	0.29
JP24	**E+**	3.43	0.12

[3] **Underliggande kvalitet (ability score):** Ett mått på uppsatsens kvalitetsnivå där genomsnittet ligger kring 5.5. JA24:s värde 8.50 betyder att uppsatsen ligger mycket högt över genomsnittet, medan JP24:s värde 3.43 ligger betydligt under.

[4] **Konfidens:** Visar hur säker modellen är på konsensusbetyget. Högre värde indikerar större säkerhet. Värden runt 0.30 anses generellt inte som tillförlitliga och beror på att vi saknar uppsatser vid samtliga trösklar, för att få ett mer rättvisande värde på tillförlitligheten kan vi slå ihop exvis D- och D+, vilket gör att tillförlitligheten hamnar på >0.6 för de flesta konsensusbetyg, vilket är acceptabelt.

Resultat: Bedömarnas stränghet

Analysen visar att ni bedömer olika strängt – vilket är helt normalt! Det viktiga är att vi kan ta hänsyn till detta i modellen.



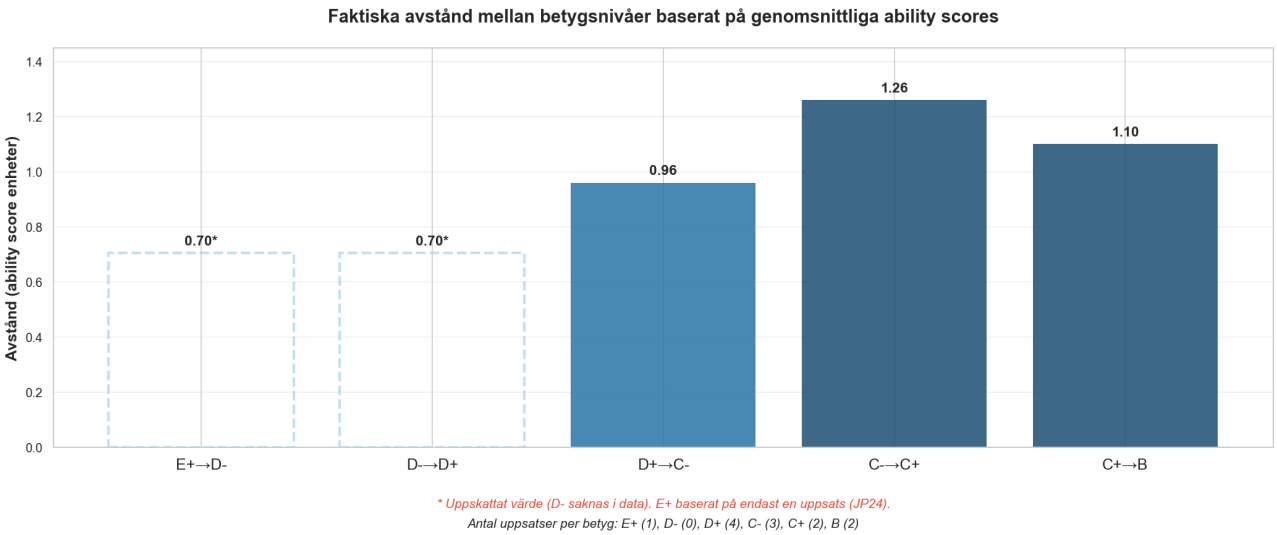
Figur 2: Systematiska skillnader i bedömarstränghet. Värdet 0 representerar genomsnittlig stränghet. **Negativa värden (röda staplar) indikerar stränga bedömare** som tenderar att ge lägre betyg än genomsnittet. **Positiva värden (gröna staplar) visar generösa bedömare** som ger högre betyg. Antalet bedömda uppsatser visas i parentes. B13 är strängast med -0.74 medan B04 är generösast med +0.56. Observera att B14 (markerad med ⚠) endast har bedömt 2 uppsatser vilket gör värdet extremt osäkert. Bedömare med färre bedömningar visas med lägre transparens.

Bedömar-ID	Stränghet	Antal bedömda	Tolkning
B13	-0.74	6	Sträng
B14	-0.56	2	Sträng*
B11	-0.39	4	Något sträng
B12	-0.29	5	Något sträng
B08	-0.16	5	Något sträng
B07	-0.08	5	Neutral
B10	-0.04	4	Neutral
B09	0.05	4	Neutral
B02	0.27	5	Något generös
B03	0.30	5	Något generös
B06	0.31	3	Något generös
B01	0.40	6	Något generös
B05	0.49	3	Något generös
B04	0.56	3	Generös

*B14 har endast bedömt 2 uppsatser vilket gör värdet mycket osäkert

Avstånden mellan betygsstegen

Modellen visar att betygsstegen inte är jämnstora. Vissa övergångar kräver stora kvalitetshopp medan andra är små – en observation som stämmer väl med många lärares erfarenhet - och med grundtanken bakom betygsskalor, som antar att betyg är normalfördelade.



Figur 3: Hur stora är stegen mellan betygen? Övre panelen visar betygsgränserna som färgzoner. Varje färg representerar ett betygsområde från F (mörkröd) till A (mörkgrön). Nedre panelen visar storleken på stegen mellan betygsnivåerna. Gröna staplar = små steg, orange = medelstora, röda = stora steg. De största hoppen är från D+ till C- och från C- till C+. För att ta sig över C-gränsen måste eleverna kraftigt förbättra sin skrivförmåga – något som stämmer väl med många lärares erfarenhet att just denna gräns är särskilt svår att passera.

Praktisk betydelse

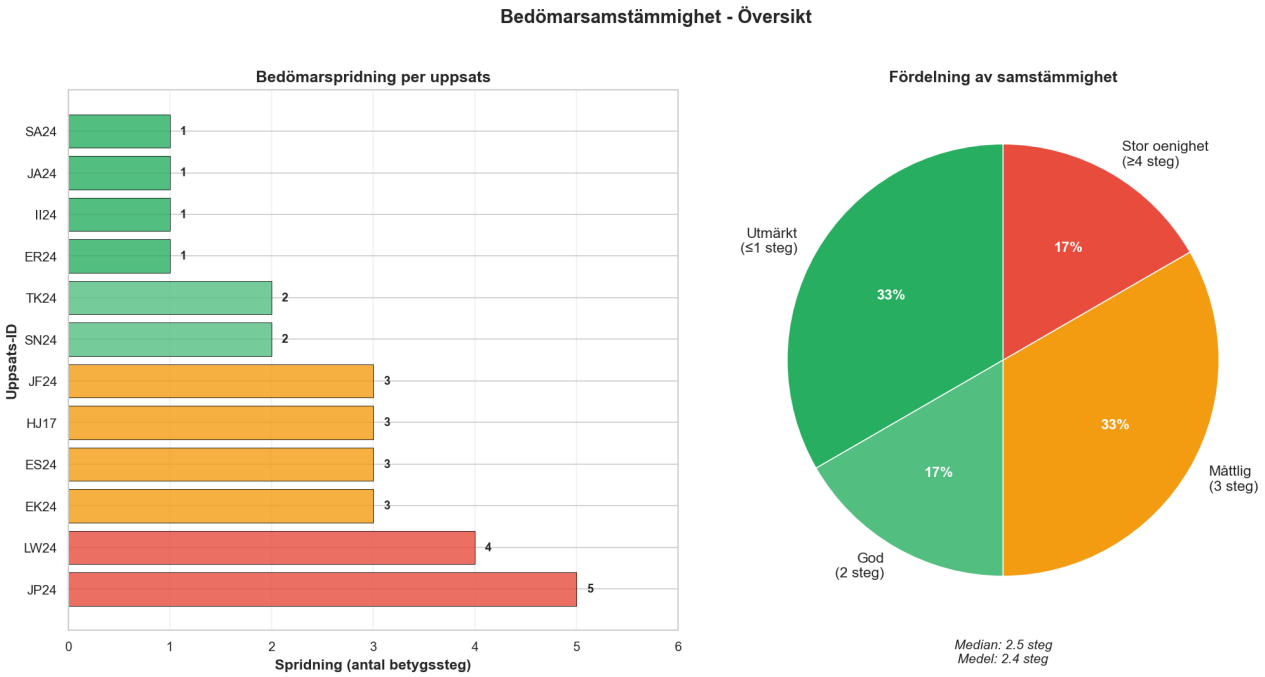
Analysen av faktiska avstånd mellan betygsnivåer visar att:

- Steget från D+ till C- är det minsta: 0.96 enheter
- Steget från C- till C+ är betydande: 1.26 enheter
- Steget från C+ till B är medelstort: 1.10 enheter

OBS: Endast en uppsats fick E+ (JP24) samtidigt som ingen uppsats fick betyget D-, vilket gör det svårt att dra säkra slutsatser om dessa betygssteg.

Bedömarsamstämmighet

Analyses av hur väl ni överensstämmer i era bedömningar.



Figur 4: Bedömarspridning per uppsats. Vänstra panelen visar spridningen i bedömningar för varje uppsats, sorterad från högst till lägst. Gröna staplar = god samstämmighet (≤2 betygssteg), orange = måttlig oenighet (3 steg), röda = stor oenighet (≥4 steg). JP24 sticker ut med 5 betygssteg i spridning – här är bedömarna verkligen oense. LW24 har också stor spridning med 4 steg. Högra panelen visar fördelningen över alla uppsatser. Cirka 50% har god samstämmighet medan 33% har måttlig oenighet. Detta är faktiskt helt normalt för uppsatsbedömning – internationell forskning visar liknande mönster.

Sammanfattande mått

- **Krippendorffs alpha:** 0.56 (måttlig samstämmighet)
- **Medianspridning:** 2 betygssteg
- **Uppsatser med stor oenighet (≥3 steg):** 50%
- **Uppsatser med mycket stor oenighet (≥4 steg):** 16%

Ett alpha-värde på 0.56 ligger under den gräns (0.667) som Krippendorff rekommenderar för att dra säkra slutsatser. Det är dock viktigt att komma ihåg att vi inte genomfört någon gemensam kalibreringsträning innan bedömningen – ni fick uppsatserna och bedömde dem individuellt utifrån era egna tolkningar av betygskriterierna. Med tanke på detta är 0.56 faktiskt inte så illa. Efter kalibreringsträffar brukar samstämmigheten öka markant.

Några problem och hur vi bäst löser dem

Problem: Ojämn bedömarmatris

Flera bedömare har för få uppsatser:

- B14: endast 2 uppsatser (behöver minst 3-4 till)
- B04, B05, B06: endast 3 uppsatser (behöver 2-3 till)
- B09, B10, B11: endast 4 uppsatser (behöver 1 till)

Med fem bedömningar per person får vi tillförlitliga bedömarprofiler – både för varje individ och för gruppen som helhet. Detta ger oss underlag för framtida kalibrerande sambedömningsträffar.

Mitt förslag på kompletterande bedömningar

Vem bedömer vad

Första omgången:

- **B14: EK24** och **ER24**

Båda ligger på bara fyra bedömningar.

Andra omgången:

- **B04: JP24**
- **B05: SA24**
- **B06: JA24**

Då har alla bedömare minst fyra uppsatser. JP24 och SA24 får sin femte bedömning.

Tredje omgången – om vi vill ha fem uppsatser per person:

- **B14: SN24** eller **TK24**
- **B04: II24**
- **B05: ES24**
- **B06: LW24**

Med fem uppsatser per bedömare får modellen tillräckligt underlag för att skilja mellan verklig bedömarstränghet och slumpmässig variation. Då vet vi om någon verkligen är strängare eller om det bara såg ut så på grund av vilka uppsatser hen råkade få.

Tidslinje

1. **Vecka 39-47:** Kompletterande bedömningar enligt ovan
2. **Vecka 48:** Omkörning av analysen med fullständiga data
3. **December/januari:** Pilottest med elevernas mid-terms

4. **Januari/februari:** Eventuell finjustering baserat på pilottestets resultat
5. **April/maj:** Om allt fungerar kör vi om processen på riktigt i skarpt läge direkt efter NP, Del C: Writing (Eng 5 och om vi orkar: Eng 6)

Avslutande kommentarer

Analysen visar att ni bedömer olika – vissa är konsekvent strängare (negativa värden) medan andra är mer generösa (positiva värden) i sin bedömning. För vissa uppsatser skiljer det upp till fem betygssteg mellan er. Det är inget konstigt utan helt normalt! Poängen med den bayesianska modellen är att den kan räkna ut ett konsensusbetyg för varje uppsats, justerat för era individuella bedömarprofiler. De uppsatser som får stabila konsensusbetyg (hög konfidens) blir våra ankaruppsatser.

Nästa steg blir att samla in de kompletterande bedömningarna enligt schemat ovan. Därefter kör vi om analysen och får förhoppningsvis mer tillförlitliga siffror för alla bedömare. I pilottestet får vi se om ankaruppsatserna verkligen fungerar som stabila referenspunkter när AI-systemet ska rangordna och betygsätta elevuppsatserna.

Tack och hej, leverpastej,

Olof

En oerhört parentetisk fotnot: Analysen baseras på en ordinal kumulativ logit-modell som skattar underliggande uppsatskvalitet (θ), bedömarstränghet (ρ) och betygsgränser (τ) samtidigt. Modellen använder nollsummerestriktion ($\Sigma\rho = 0$) för identifiering och monotont ökande betygsgränser. Inget antagande om lika avstånd mellan betygsstegen görs – stegen tillåts vara olika stora, vilket data bekräftar (särskilt de stora hoppen vid C-gränserna). Genomförd i Python/PyMC med 2000 MCMC-dragningar över 4 kedjor.