UPPSALA UNIVERSITET
Institutionen för nordiska språk
Nationella prov i svenska och
svenska som andraspråk
Eva Östlund-Stjärnegårdh

UPPSALA
UNIVERSITET

Bedömaröverensstämmelse – ämnet Svenska

Ämnesprovet i årskurs 9 gäller ämnena Svenska och Svenska som andraspråk, men i föreliggande studie behandlas enbart ämnet Svenska. I ämnesprovet ingår tre delprov, nämligen A Att läsa och förstå, B Muntligt delprov och C Skriftligt delprov. De tre delproven hålls samman av ett gemensamt tema som för det aktuella läsåret (vt 2007) var *Vidgade vyer*. Elevens lärare bedömer de tre olika delarna var för sig och väger sedan ihop dem till ett provbetyg enligt Skolverkets anvisningar.

Ämnesprovet innehåller omfattande bedömningsanvisningar med allmänna resonemang om bedömning i relation till kursplanen, principer för bedömningen av de olika uppgifterna samt kommenterade elevexempel på olika kvalitativa nivåer. Vilket material lärarna har tillgång till redovisas nedan för respektive delprov. För att ge lärarna möjlighet att nyansera sina betyg används vid de enskilda delproven och i det sammanvägda provbetyget en tiogradig skala. I den motsvaras Godkänt av stegen 3, 4 och 5 för att skilja mellan svaga, ordinära och starka prestationer. Motsvarande för Väl godkänt är 6, 7 och 8. Mycket väl godkänt kan nyanseras i 9 och 10. De lösningar som ej nått målen (EUM) kan graderas i 1 och 2, där 2 ligger nära gränsen för Godkänt.

För att göra en ombedömning i syftet att studera bedömaröverensstämmelsen kan enbart delprov A och C komma ifråga. Ett urval elevlösningar av dessa skickas in till Institutionen för nordiska språk och arkiveras. För delprov B rapporterar läraren enbart in betyg.

Urval

Till institutionen har totalt 1670 elevlösningar skickats in efter datumurval från totalinsamlingen av resultat till SCB. Inte alla dessa lösningar är kompletta eller har alla uppgifter om kön, betyg och ämne angivet. Bland de cirka 800 lösningar som har uppgift om elevens kurs, kön och betyg för både A och C har 100 slumpvis plockats ut för studien. En smärre justering har behövt göras för att få samma könsfördelning som i den stora databasen, 47 % flickor och 53 % pojkar, samt för att få ungefär samma betygsfördelning på delprov A som i totalinsamlingen. Vid kopieringen av elevlösningarna befanns tre vara av så dålig kvalitet att de byttes ut mot närmast efterföljande i databasen med samma uppgifter och betyg. Betygen på elevlösningarna har rapporterats in av lärarna med den 4-gradiga skalan: G, VG och MVG eller Ej uppnått målen.

Tabell 1. Andel betyg i 100-urvalet och i totalinsamlingen för äp 9, vt 2007.

		EUM	G	VG	MVG
Urval 100 texter	Delprov A	10	39	42	9
	Delprov C	4	51	36	9
Totalinsamling	Delprov A	9,7	39,3	40,1	11,0
	Delprov C	5,6	50,9	34,2	9,3

Urvalets betygsspridning för delprov A och C visas i tabell 1 liksom den procentuella fördelningen i Skolverkets totalinsamling som bygger på cirka 125 000 elever (Ämnesprovet 2007 i grundskolans årskurs 9).

Bedömare

Till bedömare i studien har utsetts tre personer med lärarutbildning i svenska och lärarerfarenhet från grundskolan. Bedömare 1 är lärare med skolerfarenhet från ett antal år i grundskolan. Bedömare 2 ingår i provinstitutionens projektgrupp som konstruktör för grundskoleproven parallellt med arbete i åk 6–9. Bedömare 3 har erfarenhet av provinstitutionens utprövning av provuppgifter och av textbedömning på olika nivåer i skolan och på högskolenivå.

Varje bedömare fick skrivna instruktioner, aktuellt bedömningsmaterial och kopierade elevlösningar som avidentifierats och där lärarbetyg tagits bort liksom kommentarer i möjligaste mån. Bedömningsarbetet utfördes under sommaren 2008.

Bedömarna har arbetat var för sig och inte diskuterat sitt uppdrag med andra, dvs. inte utnyttjat sambedömningens fördelar så som lärarna ofta gör när de genomför de nationella proven. Betygen har fyllts i på resultatprofiler från provmaterialet, en för varje elev, och eventuella kommentarer till svårigheter och överväganden har bedömarna också skrivit där. Bedömarna har uppmanats att använda den 10-gradiga skalan, dvs. att utnyttja möjligheten att få fram nyanser i bedömningen. Det är också de betygen vi räknat medeltal och spridning på. Även korrelationsberäkningarna redovisas för tiogradig skala. För en jämförelse med lärarnas betyg ”översätter” vi ibland till fyrgradig skala.

De tre bedömarna har också efter sitt arbete medverkat i intervjuer där vi resonerat om vad de uppfattat som förhållandevis lätt eller svårt i bedömningen.

Delprov A: Att läsa och förstå

För delprov A har både texter och frågor prövats ut i flera omgångar i olika klasser över hela landet, och hundratals elever har bidragit till de autentiska elevsvar som till sist kommer med i bedömningsmaterialet. Svaren har diskuterats inom provgruppen och på bedömargruppsmöten, och gränserna för de olika betygen har fastställts inom bedömargruppen.

I delprov A prövas elevens förmåga att läsa och förstå texter av olika slag – tidningsartiklar, faktatexter, dikter och skönlitterär prosa. Eftersom kursplanen betonar ett vidgat textbegrepp ingår också bilder i det texthäfte som eleverna

läser och diskuterar i klassen i förväg. (Temat i det texthäftet återkommer i alla tre delproven.) Frågorna i delprov A är av olika format: kryssfrågor, frågor där ett ord eller begrepp räcker som svar, kombinationsfrågor, öppna frågor som kräver läsning på flera ställen i en text och/eller tolkning. Svaren bedöms motsvara olika nivåer i läsningen, från nivå 1 till nivå 3:

- Svar som kan utläsas direkt ur texten och som kan sägas vara *rätt eller fel*. Den sortens kortsvar representeras vanligen av nivå 1 på bedömningsskalan.
- Kortsvar som kan sägas vara *rätt eller fel*, men kräver att eleven överblickar flera avsnitt eller visar ett visst mått av abstrakt tänkande, kan representeras av nivå 2.
- Utförligare svar där *förklaringar, exempel eller flera aspekter på uppgiften* krävs (till exempel både förklaring och motivering) representeras av nivå 2 eller 3. Svaren kan sägas vara *mer eller mindre fullständiga snarare än rätt eller fel.*

Svar som når nivå 1 motsvarar grundläggande läsförmåga som att t.ex. genom sökläsning hitta och välja ut fakta som efterfrågas i uppgiften. I provet 2007 ingick 11 frågor där svaren bara kan nå nivå 1. Svar som når nivå 2 visar att eleven kan hämta information från flera ställen i texten, dra slutsatser utifrån det lästa samt reflektera kring enskilda delar i texten. I provet ingick 9 sådana frågor. Svar som når nivå 3 visar den mer komplexa läsförmågan där eleverna förväntas kunna dra mer långtgående slutsatser av det lästa och att framförallt utifrån hela texter kunna resonera om och reflektera kring det lästa. I provet ingick 3 frågor där svaren kan nå nivå 3. Eleverna kan se till vilken nivå varje enskild fråga sträcker sig och det kan ge dem viss vägledning i hur utförliga de behöver vara i sitt svar. De får även tips om hur omfattande svar som förväntas av dem genom instruktionens utformning och genom hur många rader de får till förfogande för svaret.

Till lärarnas hjälp för bedömningen finns ett detaljerat material där varje fråga gås igenom. Först ges inom hakparentes information om vilket innehåll provkonstruktörerna tänkt att svaret ska ha. Därefter visas exempel på autentiska elevsvar på olika nivåer. Här betonas att det är innehållet och inte den språkliga utformningen som bedöms i delprov A. När varje fråga bedömts väger läraren ihop svaren till ett delprovsbetyg enligt noggranna anvisningar. För betyget Godkänt krävdes 2007 korrekt svar till minst 15 frågor på nivå 1 och minst 3 på nivå 2. Det går alltså inte att få Godkänt på läsförståelseprovet utan att ha svarat på några mer krävande frågor. För betyget Väl godkänt krävdes minst 18 frågor på nivå 1 och minst 7 på nivå 2 och för Mycket väl godkänt krävdes rätt svar på i princip alla nivå 1- och nivå 2-frågor plus minst 2 på nivå 3. Vid betygsättningen har läraren stöd av information om vilka aspekter som är viktigast på respektive betygsnivå:

För *Godkänt* gäller att eleven

- hittar konkreta uppgifter om sakförhållanden som direkt uttrycks i texten eller lätt går att sluta sig till,
- förklarar innebörd på ett begripligt sätt och kan reflektera över textens innehåll,
- väljer exempel ur texten och ger en motivering om än svag.

För *Väl godkänt* gäller *dessutom* att eleven

- ger relevanta och motiverade exempel ur texten när sådana efterfrågas,
- ger självständiga och rimliga förklaringar och argument när egna reflektioner efterfrågas,
- visar förmåga att tolka och dra slutsatser, se sammanhang och/eller göra jämförelser.

För *Mycket väl godkänt* gäller *dessutom* att eleven

- ger tydligt formulerade svar som har en logisk uppläggning (vid längre svar),
- drar nytta av texterna och kopplar sina egna reflektioner till dem,
- sätter argument och förklaringar i relation till hela sammanhanget i texten.

Bedömningshäftet innehåller också fyra helhetslösningar där tveksamma och/ eller svårbedömda svar på olika frågor kommenteras.

Betyget på delprov A förs in på elevens resultatprofil för att sedan ingå i lärarens sammanvägning av delprovsresultaten till ett provbetyg.

Resultat för delprov A i bedömarundersökningen

För de 100 eleverna är medelbetyget från inskickande lärare 5,5 på den fyrgradiga skalan, alltså ett starkt G. Samtliga bedömare ligger i medeltal lägre än den siffran: 4,9 – 5,2 – 5,1. Genomsnittet för deras bedömningar är 5,1. Av de 100 lösningarna har 72 fått samma betyg (på den 4-gradiga skalan) av läraren och bedömarna gemensamt. Bedömarna har ett steg lägre betyg på 22 lösningar, två steg lägre betyg på 1 lösning samt ett steg högre betyg på 5 lösningar. I huvudsak är det de högre betygen som sänks.

Tabell 2. Bedömarnas betyg som medelvärde och standardavvikelse.

Bedömare	Medelvärde	Standardavvikelse
Lärargruppen[1]	5,5	
Bedömare 1	4,9	2,13
Bedömare 2	5,2	2,25
Bedömare 3	5,1	2,13

[1] Lärargruppen kan egentligen inte jämföras med de tre bedömarna, eftersom gruppen består av 100 olika personer.

Korrelationen (Spearman) mellan bedömarna är 0,80–0,86 när deras betyg efter den tiogradiga skalan jämförs.

Tabell 3. Korrelationen mellan bedömarna i delprov A.

Correlations

			Bed1_läs Bed. 1 - läsförståelse	Bed2_läs Bed. 2 - läsförståelse	Bed3_läs Bed. 3 - läsförståelse
Spearman's rho	Bed1_läs Bed. 1 - läsförståelse	Correlation Coefficient	1,000	,863**	,800**
		Sig. (2-tailed)	.	,000	,000
		N	100	100	100
	Bed2_läs Bed. 2 - läsförståelse	Correlation Coefficient	,863**	1,000	,838**
		Sig. (2-tailed)	,000	.	,000
		N	100	100	100
	Bed3_läs Bed. 3 - läsförståelse	Correlation Coefficient	,800**	,838**	1,000
		Sig. (2-tailed)	,000	,000	.
		N	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Skillnader mellan bedömarna på den 10-gradiga skalan

För ett fåtal lösningar skiljer sig bedömarna åt mer än 2 steg på den 10-gradiga skalan. Medianen är 1 steg och medel 1,3 steg.

Analysen av antalet differenssteg ger följande resultat för delprov A:

Steg	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Antal	23	40	24	9	4	0	0	0	0	0	[m=1.3]

Tydliga skillnader finns mellan bedömarna i hur de använder den tiogradiga skalan. Se diagram 1:

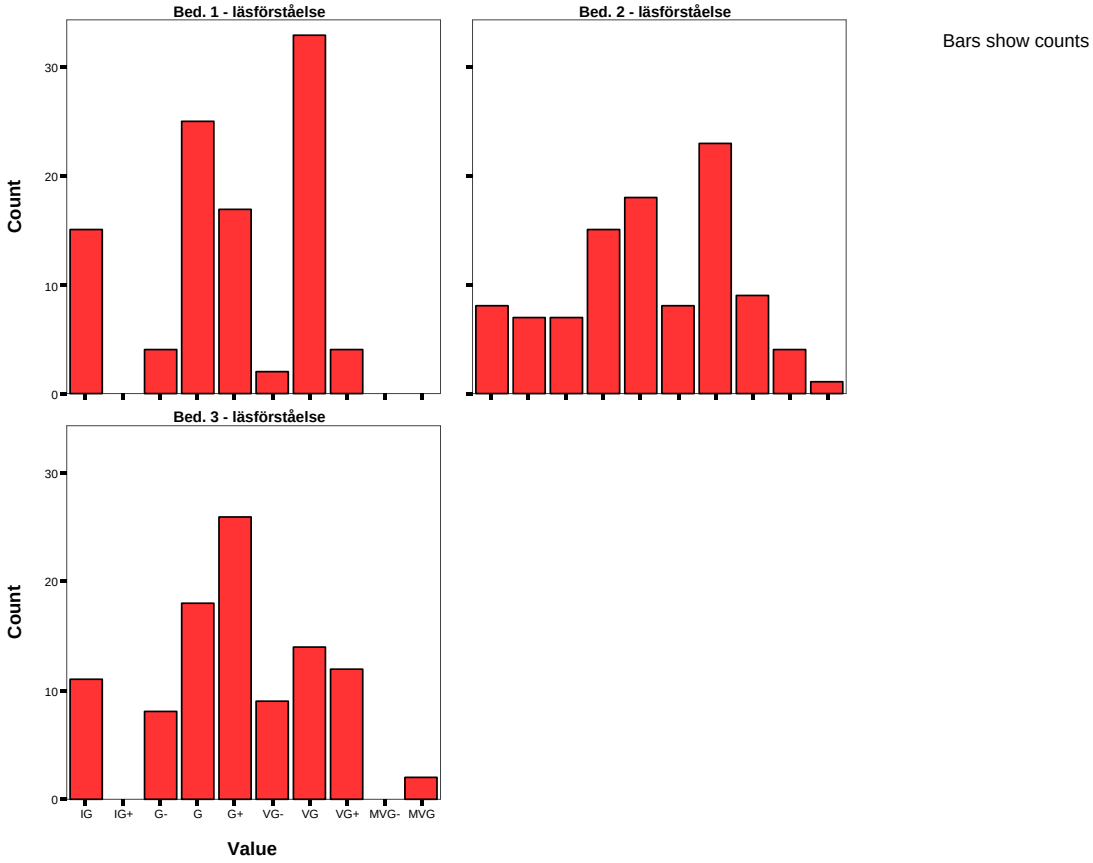

Diagram 1. De tre bedömarnas betyg enligt den 10-gradiga skalan för delprov A.

Bedömare 1 utnyttjar inte alls betygen 9 och 10, dvs. MVG. Detta påverkar naturligtvis medelvärdet för bedömarnas betyg. Bedömare två sprider sina betyg över hela skalan, och bedömare 3 utnyttjar 10 men inte 9. Spridningsmåttet för de tre är 2,13 – 2,25 – 2,13, vilket inte visar på så stora skillnader dem emellan. De underkänner ungefär lika många av lösningarna: bedömare 1 och 2 15 % samt bedömare 3 11 %.

Vanligast betyg hos bedömare 1, som i snitt verkar så sträng, är ändå 7 (VG), liksom hos bedömare 2. För bedömare 3 är det vanligaste betyget 5 (G i övre delen av betygsspannet).

Inom undersökningen finns inte utrymme att studera skillnader i förhållande till kön. En iakttagelse är att det framför allt är flickor som fått MVG av läraren och får VG av bedömarna, till och med i ett fall G.

Enskilda frågor i läsförståelsen

Det är ett fåtal uppgifter som utmärker sig med låg överensstämmelse mellan bedömarna. Främst utgörs de av nivå 2- och 3-frågor men det är dock en av nivå

1-frågorna som har genererat en låg överensstämmelse mellan bedömarna. De två avslutande nivå 3-frågorna har gett en större variation mellan bedömarna än andra frågor. Förklaringen till den lägre korrelationen mellan bedömarna på nivå 3-frågorna kan vara att frågorna är mer komplexa eftersom de mäter en annan aspekt av läsförståelse. Även kravet att eleven själv ska formulera sina tankar i en kortare text kan bidra till skillnaderna i bedömning. Jämför redovisningen av betyg på skrivuppgiften. Med ett undantag är det frågor till skönlitterära texter och bilder som bedömningen varierar för, frågor där elevens tolkning blir avgörande.

Bedömare 1 nämnde en av dessa frågor som särskilt besvärlig, och hon hade gått tillbaka och bedömt om samtliga svar på den frågan. I det sammanhanget sade hon också att hon verkligen saknade möjligheten att diskutera med kolleger när tveksamheter uppstod.

Delprov C: Skrivuppgift

Inför delprov C provas ett antal skrivuppgifter inom provets tema i olika skolor och klasser runt om i landet. Både lärare och elever får lämna synpunkter i enkätform, och provkonstruktörerna har både elevtexterna och dessa enkäter till hjälp när nya versioner av uppgifterna formuleras. Den uppgift som ingår i slutversionen har oftast genomgått ett antal förändringar. Bedömargruppen läser många elevlösningar till varje uppgift och betygssätter först enskilt och sedan gemensamt i en diskussion.

Eleverna har i delprov C fyra skrivuppgifter att välja mellan. De rör alla provets tema och eleverna kan använda texthäftet som inspiration, men direkt källanvändning som i gymnasieprovet krävs inte. Provtiden är 160 minuter, så eleverna har förhållandevis gott om tid att skriva och bearbeta sina texter. Uppgifterna kan vara av varierande genrer för att motsvara kursplanens mål att eleverna ska kunna skriva ”olika sorters texter”. År 2007 var de aktuella genrerna krönika, debattinlägg, berättelse och artikel. Instruktionen ställer inga krav på visst ordantal, och textlängden varierar också mycket. Tidigare undersökningar på gymnasienivå har visat att ordantal och betyg samvarierar (Hultman & Westman 1977, Östlund-Stjärnegårdh 2002). Längre texter får högre betyg i medeltal, men det gäller inte för den enskilda texten. I äp 9 varierar de flesta godkända elevlösningar mellan 250 och 700 ord.

Bedömningsunderlaget för delprov C utgörs dels av ett kvalitetsschema (s. 24 i bedömningshäftet) med kriterier för kvalitet i texter, dels av elevtexter med kommentarer (s. 25–49 i bedömningshäftet). Kvalitetsschemat används för att göra en helhetsbedömning av elevens text, där hänsyn tas till de fem olika områden som schemat fokuserar: kommunikativ kvalitet, innehållslig kvalitet, sammanhang och uppläggning, språklig kvalitet samt skrivregler. För varje skrivuppgift finns en beskrivning av uppgiften samt bedömda och analyserade elevtexter på alla betygsnivåer. I analysen förs ett resonemang om textens förtjänster och brister i anslutning till kvalitetsschemats olika områden. Vid betygsangivelsen anges om betyget eventuellt ligger nära ett högre eller ett lägre betyg. Lärarna har i enkäter ofta efterfrågat exempel för att få hjälp med gränsfall. De elevtexter som publiceras har bedömts av åtminstone fem lärare

från olika skolor som först satt betyg var för sig och sedan diskuterat ihop sig på ett bedömarmöte.

Resultat för delprov C i bedömarundersökningen

Medelbetyget från inskickande lärare på de 100 skrivuppgifterna är 5,5. Se tabell 3. Jämfört med delprov A ligger bedömarna ännu lägre med ett snittbetyg på 4,6. Bedömare 1 ligger också här lägst med 4,1. Bedömare 2 har snittet 5,2 och bedömare 3 snittet 4,5. Av de 100 elevtexterna får 54 samma betyg på den fyrgradiga skalan av den egna läraren och bedömarna (i snitt), 37 texter får ett steg lägre och 1 text får två steg lägre. Bedömarna höjer betyget ett steg för 8 texter. I huvudsak är det liksom för delprov A de högre betygen som sänks.

Tabell 4. Bedömarnas betyg som medelvärde och standardavvikelse.

Bedömare	Medelvärde	Standardavvikelse
Lärargruppen[1]	5,5	
Bedömare 1	4,1	1,96
Bedömare 2	5,2	2,39
Bedömare 3	4,5	1,67

[1] Lärargruppen kan egentligen inte jämföras med de tre bedömarna, eftersom gruppen består av 100 olika personer.

Det är framför allt flickor som fått MVG av läraren och får VG av bedömarna. En pojkes MVG sänks till G av bedömarnas snitt.

Korrelationen (Spearman) mellan bedömarnas betyg enligt den tiogradiga skalan är låg: 0,36–0,46.

Tabell 5. Korrelationen mellan bedömarna i delprov C.

Correlations

			Bed1_text Bed. 1 - elevtext	Bed2_text Bed. 2 - elevtext	Bed3_text Bed. 3 - elevtext
Spearman's rho	Bed1_text Bed. 1 - elevtext	Correlation Coefficient	1,000	,360**	,397**
		Sig. (2-tailed)	.	,000	,000
		N	100	100	100
	Bed2_text Bed. 2 - elevtext	Correlation Coefficient	,360**	1,000	,455**
		Sig. (2-tailed)	,000	.	,000
		N	100	100	100
	Bed3_text Bed. 3 - elevtext	Correlation Coefficient	,397**	,455**	1,000
		Sig. (2-tailed)	,000	,000	.
		N	100	100	100

**. Correlation is significant at the 0.01 level (2-tailed).

Skillnaden mellan hur bedömarna använder den tiogradiga betygsskalan för delprov C framgår av diagram 2:

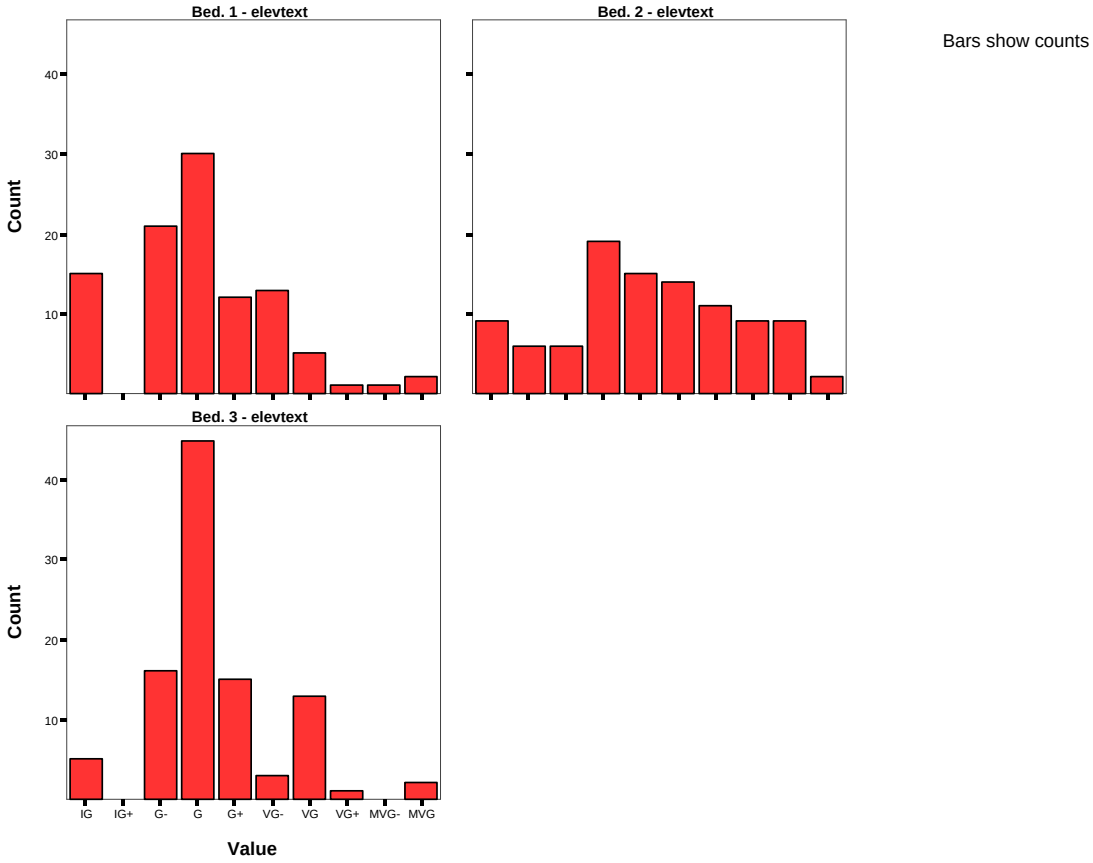

Diagram 2. Bedömarnas betyg fördelade på den 10-gradiga skalan

Bedömare 1 ligger lågt med en tydlig tyngdpunkt på G (3 och 4). Bedömare 2 använder hela skalan även om G (4) är det vanligaste betyget också hos henne. Bedömare 3 tenderar snarast att krympa skalan och verkligen centrera sina betyg på G-nivån. Hon underkänner bara 5 % av texterna, medan bedömare 1 och 2 båda underkänner 15 %. Det är också stor skillnad på MVG-nivån, där bedömare 1 har 3 % av texterna, bedömare 2 11 % och bedömare 3 3 %.

När det gäller delprov C är skillnaden i spridningsmått för de tre bedömarna större än för delprov A: 1,96 – 2,39 – 1,67.

Skillnader mellan de tre bedömarna på den 10-gradiga skalan

För mindre än hälften av lösningarna skiljer sig bedömarna åt mer än 2 steg på den 10-gradiga skalan. Medianen är 2 steg och medel 2,7 steg.

Analysen av antalet differenssteg ger följande resultat för delprov C:

Steg	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	
Antal	1	23	35	16	10	8	4	2	0	1	[m=2,7]

De 24 texter bedömarna är mest överens om domineras av G-texter, 19 av 24. De är också överens om 1 IG, 3 VG och 1 MVG. Ettstegsdifferensen ligger för samtliga inom ett och samma betygssteg på den fyrgradiga skalan. Att bedömare är mer överens om de lägsta betygen har visats i tidigare studier (t.ex. Berge 1996 och Östlund-Stjärnegårdh 2002).

För de 15 texter där bedömarna skiljer sig 5 steg eller mer på den 10-gradiga skalan är det i 10 fall bedömare 1 som ligger lägst. Vid tre tillfällen är bedömare 1 och 3 överens om det lägsta betyget. En text vardera har bedömare 2 och bedömare 3 lägst betyg på. Bedömare 1 framträder också i denna beräkning som en sträng bedömare och som den som oftast avviker från de andra.

Olika typer av skrivuppgifter

I delprov C har eleverna alltså fyra skrivuppgifter med olika genreangivelse att välja på. År 2007 var de aktuella genrerna krönika, debattinlägg, berättelse och artikel. Det tycks som om en del av bedömarvariationen kan höra ihop med vilka genrekrav bedömarna betonar vid bedömningen. Reservation måste göras med tanke på antalet texter som ingår i de olika grupperna. 100-urvalet är inte gjort för att vara representativt på uppgiftsnivå.

Krönikan är en ganska ny genre inom proven och visar sig vara svårbedömd. Eleverna uppfattar den ofta som en rent berättande text, och lärarna/bedömarna ser olika strängt på detta. De 21 inskickade krönikorna får i genomsnitt 5,9 av de egna lärarna medan bedömarna i snitt har satt 4,5. Här får 10 av 21 texter ett lägre betyg på den fyrgradiga skalan av bedömarna.

Debattinlägget är en välkänd genre som eleverna verkar vara säkra på. Bedömningen är också betydligt jämnare för de 10 texterna av denna genre: lärarna har i snitt 6,2 och bedömarna 6,1. Denna uppgift har högst medelbetyg både hos inskickande lärare och hos bedömarna (Jfr samma tendenser i KAL-materialet angående både överensstämmelse och högre medelbetyg för debatterande text än för andra uppgiftstyper. Mer om KAL-materialet under rubriken Diskussion nedan.)

Berättelsen (utan närmare specifikation) lockade flest elever vt 2007. Lärarna har satt 5,5 i snitt och bedömarna 4,4 på de 65 berättelser som ingår i urvalet. Lika betyg av lärare och bedömare får 34 texter, medan 25 får lägre och 6 högre betyg (fyrgradig skala).

Artiklarna är bara fyra till antalet. De följer mönstret för krönika och berättelse, nämligen att bedömarna är strängare än lärarna.

Ett extremfall är en text som fått MVG av elevens egen lärare och MVG av bedömare 2 samt VG av bedömare 3. Det borde vara en riktigt bra text, men bedömare 1 sätter ett svagt G. Skillnaden mellan bedömarna är alltså 7 steg på den 10-gradiga skalan. (Den texten testades i en kurs inom Lärarlyftet, där 5 av 6 lärare satte MVG oberoende av varandra. Det sjätte betyget var VG. Bedömare 1 avviker alltså rejält här. Ytterligare tre texter testades på samma sätt, och de sex lärarna i lärarlyftskursen var i snitt överens med studiens tre bedömare men mer överens inbördes.)

Bedömarprofiler

De tre bedömarna uppvisar olika profiler, som syns i både delprov A och delprov C.

Bedömare 1 är den strängaste av dem; i delprov A sätter hon inte ett enda MVG. Under intervjun tillfrågades hon om detta, och sa att hon tyckte att det hade blivit få höga betyg men att hon inte direkt tänkt på spridningen. Hon anser sig inte vara en särskilt sträng bedömare i jämförelse med sina kolleger. Vid den skolan har man ett välorganiserat samarbete runt bedömningen av både A och C, och bedömare 1 uttryckte också att hon tyckte att det var svårt att inte ha någon att resonera med. Hon menade att det var svårare att bedöma A än C. Vid bedömningen av de skrivna texterna inom delprov C fokuserade hon hur eleverna löst uppgiften och anpassat sig till de tänkta mottagarna. Det är lättare att bedöma okända elevers texter än de egna enligt bedömare 1 som i medel har ett ännu lägre betyg på C än på A.

Bedömare 2 är den som ligger närmast de inskickade betygen, även om hon också i snitt sätter lägre betyg än lärarna. Hon är den som i både A och C utnyttjar hela betygsskalan med en spridning/standardavvikelse på 2,25 för A och 2,40 för C. Medelbetyget för A och C ligger lika hos bedömare 2. Hon betonar helhetsbedömningen för de skrivna texterna men talar också en del om hur eleverna lyckas med genreträffen, något som är viktigt för högre betyg.

Bedömare 3 ligger för bägge delproven i snitt mellan bedömare 1 och 2. Hon tenderar i bägge delproven att centrera sina betyg och ”överanvända” G. I intervjun säger hon att det på ett sätt är enklare att bedöma A tack vare det detaljerade bedömningsmaterialet. Att bedöma skrivna texter är knepigare, och där är hon mycket intresserad av textmönster. Hon menar att om det finns ansatser till rätt genre ska det bedömas positivt. Det är också mycket viktigt hur skrivinstruktionen är formulerad. Att få diskutera med någon vid svårbedömda fall hade varit bra, men bedömare 3 tyckte sig inte ha några större problem att utföra arbetet.

Diskussion

Lärarna gör i genomsnitt en mer positiv bedömning än bedömarna för både delprov A och delprov C. Samma förhållande visas också i bedömnings-undersökningen av gymnasietexter i Östlund-Stjärnegårdh 2002. Där diskuteras ett par möjliga skäl till att elevens lärare tenderar att sätta högre betyg än en oberoende bedömare. Det första är att lärarna förmodligen väger in övrig kännedom om vad eleven kan och vad hon presterat i övriga delprov. Provlösningarna kanske också betygsätts med tanke på det förestående slut-betyget. Liknande resonemang förs också av Björnsson som menar att den egna läraren bättre förstår vad eleven menar ”när uttryckssättet ibland är suddigt och tankegången oklar” (1960:56). Om en förklaring till betygsskillnaden är att elevens egen lärare hellre friar än fäller är en annan att de bedömare som medverkar i en forskningsundersökning vill visa att de minsann vet vad som krävs för de olika betygen.

Delprov A brukar ge upphov till fler frågor till provkonstruktörerna än delprov C. Från olika källor (lärarenkäter, telefonfrågor, studiedagar) drar vi

slutsatsen att delprov C oftare sambedöms, och då löser man förmodligen eventuella problem gemensamt i svensklärargruppen. De tre bedömarna som ingår i denna studie är mycket mer överens om delprov A än om delprov C. Bedömare 3 sa också i intervjun att A var lätt att bedöma tack vare de detaljerade anvisningarna.

Frågeformaten måste efter studien diskuteras, något som i och för sig alltid görs. Kan ytterligare skärpning av frågor och anvisningar ge än större samstämmighet i bedömningen av delprov A, vore det naturligtvis en vinst. Det gäller dock att bibehålla frågor av olika svårighetsgrad och olika aspekter av läsning i delprovet. Det lätt bedömbara ger inte tillräckligt besked om elevens läsförmåga, särskilt inte när det gäller skönlitteratur. En viss del kvalitativ bedömning måste ingå också i delprov A.

Kvalitativ bedömning är givetvis en komplicerad uppgift där många faktorer spelar in. Att bedöma en elevtext handlar inte bara om att se till språklig korrekthet utan fler faktorer ingår i det kvalitetsschema lärarna har till sin hjälp. Se beskrivning under Delprov C.

Tidigare studier av bedömning och bedömaröverensstämmelse har oftare handlat om gymnasieskolans texter, t.ex. Hultman & Westman 1977 och Östlund-Stjärnegårdh 2002. Hultman & Westman beskriver i *Gymnasistsvenska* den tidens centralprov på treåriga linjer. Bedömningen är huvudsakligen språkligt inriktad. Författarna resonerar om ”en gemensam skolnorm” (1977:24f.), dvs. de ingående lärarna är sinsemellan någorlunda överens om betygen. Detta hävdas trots att författarna också skriver att varannan elev förmodligen ”skulle få ett annat betyg om hans uppsats bedömdes av en annan lärare än den som för tillfället är hans svensklärare”. I *Godkänd i svenska?*, (Östlund-Stjärnegårdh 2002) behandlas gymnasietexter från det nationella provet i Svenska B med fokus på texter i spannet IG–G. Huvudresultatet är att de oberoende bedömarna i snitt sätter betydligt lägre betyg än elevens egen lärare. Bedömarna är mest överens om vilka texter som ska ha IG, medan de uppvisar större variation i G-texterna. Ingen text av de 60 får dock större betygsspridning än 1–5 på den 10-gradiga skalan.

Variation i bedömningen är också ett stort ämne i Berges avhandling *Norsksensorenes tekstnormer og doxa* (1996) om examensprov i den norska gymnasieskolan. Han diskuterar bedömarprofiler som ”snille” och ”strenge” bedömare och en tredje grupp som rangordnar annorlunda och därför inte ingår i bedömargemenskapen. Också i Berges undersökning är bedömarna mer överens om vad som är dåliga texter. Ett kapitel presenterar bedömarnas inställning till uppgiftstyper och bedömningskriterier. Där är Berge något kritisk mot bedömarna som inte förändrat sina normer trots att examensuppgifter ändrats till mer genrespecifika. Det allmänna språkkriteriet ”språkføring” är högt värderat av bedömarna (”sensorene”).

För ett jämförelsematerial till äp 9 är det lämpligt att vända sig till den stora norska undersökningen KAL (Kvalitetssikring av læringsutbyttet i norsk skriftlig), där Berge med flera undersökte flera år av examensprov från den norska grundskolan för att se om nya läroplaner påverkat elevernas skrivande och betygsättningen. Studierna är presenterade i volymerna *Ungdommers skrivekompetanse I–II* (2005). Utifrån egna och andras undersökningar skriver Berge i *Ungdommers skrivekompetanse I,* s. 102: ”Det er en etablert

kjensgjerning at påliteligheten på skriveprøver alltid er svært dårlig når vi tar utgangspunkt i vurderinger av enkelttekster. I studie etter studie har det vist seg at det er umulig å oppnå en reliabilitet på et nivå som er selvsagt når kompetanser og ferdigheter i matematikk og i lesing prøves og vurderes." Detta påstående besannas i föreliggande studie där samstämmigheten är mycket högre i läsförståelseprovet än i skrivuppgiften där eleven bara skriver en enstaka text.

I Berges diskussion av tillförlitligheten i de norska proven redovisar han korrelationsmått från ett antal olika bedömningsundersökningar. Tabell 4 är hämtad från KAL-projektets redovisning i *Ungdommers skrivekompetanse* (Berge 2005:106, tabell 3.2).

[Tabell 4.] Jamførbare korrelasjoner i ulike skriveferdighetsprøver og skriveferdighetsstudier.

Kilde	r_p	tekstkategori	bedømmere
Westman 1974	0.27	tekster skrevet av yrkesskribenter	yrkesskribenter
Diderich, French, Carlton 1961	0.31	elevtekster	personer med ulik bakgrunn uten bedømmererfaring
Breland et al. 1987	0.52	elevtekster	lærere
Stalnaker 1937	0.55	elevtekster	lærere
Berge 1996, 2002	0.56	eksamen i vg. skole	utvalgt gruppe lærere
Larsson 1984	0.62	elevtekster (klassetrinn 4-11)	lærere
Eneskär 1990	0.62	sentralprøven i 11. klassetrinn	utvalgt gruppe lærere
KAL	0.69	avgangsprøve i norsk hovedmål 10. klassetrinn	utvalgt gruppe lærere
Löfqvist 1990 (IEA)	0.75	elevtekster	skolerte bedømmere
Gorman, Purves & Degenhart 1988 (IEA)	0.87	stratifiserte elevtekster/mønstertekster	skolerte bedømmere

Sammanställningen visar att lärare bedömer elevtexter med mer samstämmighet än vad som förekommer i andra undersökningar som de två första i tabellen. Störst samstämmighet uppträder i undersökningar där bedömarna är specialtränade ("skolerte") som de stora IEA-undersökningarna. KAL-projektet har i sin kategori stor överensstämmelse medan den studie vi nu gjort på äp 9-

material ligger lågt i jämförelse med andra elevtextundersökningar. En korrelation runt 0.6 vore att förvänta i stället för cirka 0.4 som blev utfallet. För *Gymnasistsvenska* som inte finns med i tabellen var medelkorrelationen för lärarna 0.56.

En jämförelse med KAL-materialets intensivurval (171 texter) är intressant när det gäller texter som bedömarna är mycket oense om. Termen ”sprikteksters” används om bedömarna skiljer sig 3(–4) steg på rådande betygsskala (5, senare 6 steg). Så många som 26 % av texterna befanns vara ”sprikteksters”. I denna svenska undersökning ligger 25 % av texterna på fyra stegs differens eller mer på den 10-gradiga skalan. Även om det finns svårigheter med att jämföra olika skalor, ser andelen texter med spretande betyg inte ut att vara större i denna studie än i den norska.

En mindre studie av bedömning av elevtexter i åk 9 har gjorts av Ciolek-Ciastek (2008). Hon undersöker bedömaröverensstämmelse på 40 elevtexter från det nationella provet 2006. De texter som ombedömts är alla berättande texter och varje text har bedömts av tre av varandra oberoende bedömare. Varje bedömare fick 5 texter att betygsätta. Resultaten visar att de undervisande lärarna i 14 fall av 40 är överens med de externa bedömarnas medelbetyg på texten. I 13 fall av 40 har läraren bedömt elevtexten motsvara ett högre betyg än de externa bedömarna och i 13 fall av 40 har läraren bedömt elevtexten motsvara ett lägre betyg än de externa bedömarna. Ciolek-Ciastek kommenterar att ”det inte nödvändigtvis behöver vara så att en undervisande lärare riskerar att bedöma utanför en bedömargrupps koncensus oftare än en extern bedömare” (2008:13). Intressant är att i den studien finns alltså inte tendensen att lärarna i genomsnitt sätter högre betyg än externa bedömare som i föreliggande undersökning och i Östlund-Stjärnegårdh 2002.

Tidigare efterbedömning av ämnesprovet i årskurs 9 har gett betydligt större bedömaröverensstämmelse. I ljuset av sådana kontroller framstår samstämmigheten för delprov C som förvånande låg. När skrivuppgifterna diskuteras vid bedömargruppsmöten och lärarna som ingår i bedömargruppen har satt sina betyg oberoende av varandra, är samstämmigheten också högre än i den här undersökningen. Vid senaste bedömarmötet för äp 9 hade varje bedömare 12 texter att betygsätta. Korrelationen mellan de i gruppen ingående lärarna låg mellan 0,75 och 0,92. De som i flera år ingått i bedömargruppen är att betrakta som tränade bedömare; nya lärare introduceras dock regelbundet för att gruppen inte ska cementeras. För ingen text skilde mer än fyra steg på den 10-gradiga skalan, för de flesta ett eller två steg. Om korrelationen skulle vara så låg som runt 0.4, skulle provkonstruktörerna dra tillbaka uppgiften, stryka den helt eller förändra den inför en ny utprövning. Den arbetsgången har gällt också för skrivuppgifterna i 2007 års prov.

I de tidigare nationella proven hade eleverna ofta stor valfrihet i hur de ville utforma sina texter. Uppgiften kunde ge valet mellan att skriva en novell eller en artikel om det ämne som intresserade. Då valde de allra flesta, 8 av 10, att berätta. Den norska undersökningen (KAL) visar samma berättarlust hos eleverna. Ungefär 65 % av examenstexterna är berättande oavsett vilken genre som finns angiven i uppgiften.

Efter kursplanerevisionen 2000 har de nationella proven i kursplanens anda tryckt mer på genrekunskap och specifika genredrag. Både uppgifts-

instruktionerna och bedömningsanvisningarna har för den aspekten blivit mer detaljerade. Det räcker inte längre att bara berätta för att lösa en utredande eller argumenterande uppgift som krönika och debattinlägg. Här finns en möjlig källa till bristen på överensstämmelse. Enskilda lärare betonar genreaspekten olika mycket. Bedömarna i studien är mest överens om den argumenterande uppgiften, där både elever och bedömare tycks ha god genrekunskap. Krönikan är den mest problematiska skrivuppgiften, och bedömarna skiljer sig mycket åt där, vilket troligen kan förklaras med hur de ser på rent berättande texter som ett svar på uppgiften att skriva en krönika. Genrekrav och textdrag hos de 13 texter där bedömarna skiljer sig två steg på den fyrgradiga skalan kommer att undersökas närmare och presenteras i en senare studie.

Provkonstruktörerna måste efter studiens resultat diskutera vilka genrer som ska vara med i provet i fortsättningen. Kursplanens mål att eleverna ska kunna skriva ”olika sorters texter” måste prövas, samtidigt som bedömningsmaterialet ska ge lärarna verktyg för en likvärdig bedömning. Kan provet ligga ”före” lärarna när det gäller genreundervisning? Det nyligen publicerade diagnosmaterialet *Språket på väg* för åk 6–9 kommer att ge lärare och elever större kunskaper om olika genrer om det får ordentligt genomslag. Den enda skrivna text som betygsätts inom provet måste i vilket fall kompletteras med andra texter och genrer inför lärarens slutbetyg. Likvärdig bedömning av en enskild text är som sagt svårt (se citat från Berge ovan). Om kommande provs betyg ska knytas närmare slutbetyget borde varje elev skriva två eller tre texter inom provets ram för att säkra likvärdigheten.

En annan möjlig orsak till den bristande överensstämmelsen för delprov C är mängden texter som varje bedömare skulle bedöma. Så många som 100 skrivna elevtexter i samma omgång är ingen ”vanlig” svensklärare van vid. De lärare som medverkat i tidigare undersökningar av texter från de nationella proven har haft betydligt färre texter att betygsätta; Östlund-Stjärnegårdh (2002) gav 10 texter till varje bedömare och Ciolek-Ciastek (2008) 5 texter. Provkonstruktören, i den här studien bedömare 2, arbetar oftare med större mängder elevtexter i utprövningsomgångar med hundratals elever inblandade. Det skulle vara intressant att göra om den oberoende bedömningen med fler lärare som fick färre texter var, motsvarande en vanlig klass. Att direkt kritisera bedömarna är inte vår avsikt. De vanliga lärare som ställer upp i sådana här undersökningar mot ett ringa arvode är inte vem som helst utan lärare som är så intresserade av sitt yrke och av bedömning att de avsätter tid på sommarlovet för att medverka.

Förutom genrefrågor och textantal kan ytterligare en faktor spela in i resultaten för bedömare 1 och bedömare 3, alltså de som inte arbetar i provgruppen. Det kan vara så att man som deltagare i en undersökning vill visa att man följer instruktioner och anvisningar noga och av den anledningen blir mycket strängare än man skulle vara mot sina egna elever. Sådana tendenser diskuteras angående gymnasietexters bedömning i Östlund-Stjärnegårdh 2002. Det är dock inte hela sanningen, eftersom bedömarna i snitt också höjer betyget för cirka en tiondel av texterna. Den stränghetstendensen finns inte heller i Ciolek-Ciasteks studie runt äp 9 (2008).

Slutsatser

Att sambedömning är oerhört väsentlig, särskilt för delprov C, blir uppenbart. Skolverket borde föreskriva det, inte bara rekommendera. Lärarnas inskickade betyg är, om vi ska döma av lärarenkäterna, väldigt ofta tillkomna med hjälp av sambedömning i synnerhet för svårare fall.

Från lärarenkäten 2008 kommer följande uppgifter (2007 fanns inga frågor om sambedömning): 38 % av lärarna uppger att de sambedömer svårbedömda elevlösningar medan 13 % bedömer själva men diskuterar alla lösningar med en kollega. 16 % svarar att de sambedömer alla elevlösningar medan 15 % av lärarna bedömer uppgifterna helt ensam. Endast 2 % uppger att en annan lärare har bedömt uppgifterna och ytterligare 16 % att de löser bedömningen på annat sätt eller kombinerar de tidigare nämnda sätten.

Om ett framtida provsystem får mer karaktär av examensprov och/eller provresultaten jämförs direkt med slutbetygen, borde varje elev skriva mer än en text som bedöms inom ramen för provet. Då ökar möjligheten att elevens skrivförmåga blir synlig för bedömande lärare. Kursplanens ”olika texter” kan ändå inte helt täckas inom provet, men likvärdigheten borde bli större.

Lärare behöver överlag träning i bedömning. Under ett antal år var momentet nästan helt borta ur lärarutbildningen vid många lärosäten. Därför finns ett stort behov av studiedagar runt bedömningsfrågor. Kanske ett framtida provsystem också ska innehålla underlag för diskussion och träning i lärargrupper. Ett nytt betygssystem behöver implementeras i linje med vad som nu skett med målen för årskurs 3. Sådant förekom i alltför ringa utsträckning när den nuvarande 4-gradiga skalan introducerades.

Litteratur

Berge, Kjell Lars, 2005: Skriveprøvenes pålitelighet. I: *Ungdommers skrivekompetanse. Bind I.* Red. Kjell Lars Berge m.fl. Oslo

Berge, Kjell Lars, 1996: *Norsksensorenes tekstnormer og doxa. En kultursemiotisk og sosiotekstologisk analyse.* Dr. art. avhandling. Institutt for anvendt språkvitenskap. Trondheim.

Björnsson, C.H., 1960: *Uppsatsbedömning och uppsatsskrivning*. Stockholm.

Ciolek-Ciastek, Beatrice, 2008: Lärares bedömning av elevers berättande skrivande i åk 9. I: Språkinlärning, språkdidaktik och teknologi. Rapport från ASLA:s höstsymposium i Lund 2007. Red. av Jonas Granfeldt m.fl. Lund.

Hultman, Tor G. & Westman, Margareta, 1977: *Gymnasistsvenska.* (Skrifter utgivna av Svensklärarföreningen 167.) Lund.

Ungdommers *skrivekompetanse. Bind I–II.* Red. av Kjell Lars Berge m.fl. Oslo.

Östlund-Stjärnegårdh, Eva, 2002: *Godkänd i svenska? Bedömning och analys av gymnasieelevers texter.* (Skrifter utgivna av Institutionen för nordiska språk vid Uppsala universitet 57.) Uppsala.