GÖTEBORGS UNIVERSITET
UTBILDNINGSVETENSKAPLIGA FAKULTETEN

Bedömning av nationella prov – Engelska

Gudrun Erickson
Institutionen för pedagogik och didaktik
Enheten för språk och litteratur

April 2009

Bedömning av nationella prov – Engelska

Bedömning av nationella prov – Engelska

Gudrun Erickson
Institutionen för pedagogik och didaktik; Enheten för språk och litteratur

Föreliggande studie utgör en del av en större undersökning av bedömarsamstämmighet i de nationella ämnesproven för årskurs 9 samt ett kursprov i matematik C på gymnasial nivå, initierad av Skolverket våren 2008. En rapport som beskriver bakgrunden till projektet och ger en sammanfattning av resultaten publiceras på Skolverkets hemsida i maj 2009 tillsammans med de fyra delrapporterna.

Ämnesprovet i engelska för årskurs 9 består av tre delprov med fokus på kärnområden i kursplanen, nämligen muntlig interaktion och produktion (*Part A*), receptiv förmåga (*Part B*, uppdelad i *Part B 1*, inriktat mot läsförståelse, och *Part B 2* med fokus på hörförståelse) samt skriftlig produktion och interaktion (*Part C)*. Aspekter av kultur och interkulturalitet integreras i samtliga delprov. Provet har ett övergripande tema, som håller samman delarna och som på olika sätt återspeglas i uppgifterna. Temat för det här aktuella provet, vårterminen 2007, var *Sharing and Caring*.

Ämnesprovet åtföljs av omfattande bedömningsanvisningar med specifikationer av provets delar liksom generella resonemang om bedömning i relation till kursplanen i engelska. Likaså tillhandahålls specifika principer för analys och bedömning av de olika uppgifterna, samt kommenterade elevexempel på olika kvalitativa nivåer. De olika delarna bedöms var för sig och resultaten presenteras i profilform, dvs. så att resultat på enskilda uppgifter och delprov är synliga. Resultaten vägs slutligen samman till ett provbetyg enligt en given modell.

Ämnesprovet för åk 9 utvecklas på basis av principer, såväl för innehåll som för konstruktion, utprövning och analys, som är gemensamma för samtliga nationella provmaterial i språk som produceras vid Göteborgs universitet (Erickson, 2006).

Ämnesprovets delar

Part A, som genomförs av två till fyra elever tillsammans, innehåller uppgifter som prövar elevernas förmåga att samtala med varandra (interaktion) och att tala mera sammanhängande själva (produktion). Provet fokuserar olika typer av muntlig språkanvändning, från berättande/beskrivande framställning till argumentation och diskussion.

Part B präglas av bredd och variation vad gäller innehåll och uppgiftstyper. Texterna spänner från det rent faktuella och rapporterande till det mera berättande och litterära. Bildmaterial används i anslutning till uppgifterna. I det inspelade materialet förekommer såväl monologisk som dialogisk framställning med röster från olika delar av den engelskspråkiga världen. Vad gäller uppgiftstyper används olika varianter av både *selected response* (flervalsuppgifter) och *constructed response* (uppgifter där eleverna själva formulerar svaret, i form av ett enstaka ord eller längre meningar). Användning av flervalsuppgifter i denna del av provet bottnar främst i en strävan att motverka den typ av *bias* det skulle medföra, om en alltför stor andel av elevernas kunskaper bedömdes via deras skrivförmåga. Ett annat sätt att uttrycka detta är att referera till det som brukar rubriceras som ett av de största hoten mot bedömningars validitet, nämligen *construct-irrelevant variance* (Messick, 1989), dvs. att annat än det som står i fokus för den aktuella bedömningen tillåts påverka de slutsatser som dras av resultaten. Slutligen bör nämnas att frågorna i *Part B* fokuserar olika slag av förståelse av texter och avlyssnat material, allt från att uppfatta rena fakta till att läsa/lyssna mellan raderna och dra egna slutsatser, det som av tradition och stundtals har karakteriserats som *reading the lines, reading between the lines* och *reading beyond the lines* (Gray, 1960, diskuterat i Alderson, 2000).

I *Part C* får eleverna välja mellan två olika ämnen med anknytning till provets övergripande tema. Det ena av dessa erbjuder mera strukturstöd, i form av förslag på uppläggning och innehållsliga punkter, medan det andra inbjuder till ett mera fritt skrivande. Ämnestyperna anknyter därmed till målen i kursplanen för engelska i årskurs 9, där såväl förmågan att begära och ge information i skrift som att berätta och beskriva något fokuseras. Det ges i uppgifterna inga angivelser om vare sig minimalt eller maximalt ordantal. Eleverna uppmanas dock att lämna tid för bearbetning av texten mot slutet av den angivna totaltiden för uppgiften (80 min.). Längden på uppsatserna varierar mycket, men den stora majoriteten elever producerar i regel mellan två och fem sidor handskriven text[1].

Bedömning

Bedömningen av *Part A* och *Part C* görs holistiskt, men med stöd av analytiska faktorer baserade på kursplanens skrivningar, inklusive texten Bedömningens inriktning. Dessutom ges ett antal inspelade respektive tryckta s.k. *benchmarks,* dvs. kommenterade och betygsatta, autentiska elevexempel. Dessa exempel är framtagna i en referensgrupp, på basis av omfattande utprövningar och analyser. Detsamma gäller de betygsgränser *(standards)* som ges för *Part B* (se vidare nedan).

[1] En undersökning av längden på de exempeltexter som ges i bedömningsanvisningarna för proven 2005-2008 (sammanlagt 56 stycken), visar på ett genomsnitt av 406 ord, vilket får betraktas som tämligen representativt. Det bör dock påpekas att det här inte ingår texter som bedömts som ej godkända i relation till kursplanens mål för skriftlig produktion och interaktion. Eftersom dessa texter i regel är kortare än övriga, torde det reella medelvärdet för samtliga texter därmed vara något lägre.

Uppgifterna i *Part B* bedöms med poäng, oftast dikotomt (rätt eller fel), men det förekommer också frågor med s.k. *partial credit,* dvs. differentierad poängsättning på basis av svarets innehållsliga kvalitet.

Provbetyget baserar sig på en modell där resultaten på de tre delproven vägs samman. Lika vikt ges här åt de fyra färdigheterna tala/samtala, lyssna, läsa och skriva, vilket innebär att det aggregerade resultatet för *Part B* multipliceras med två.

För att skapa större nyansering används, såväl vad gäller delproven som provbetyget, en tiogradig skala, med distinktion mellan svaga, ordinära och starka prestationer inom betygsstegen Godkänt och Väl godkänt. Vad gäller resultat som ej når upp till målen, liksom mycket väl godkända resultat, används två steg.

Utveckling av provet

Ämnesprovet i engelska utvecklas i en kollaborativ process, i kontinuerligt samarbete med lärare, lärarutbildare, forskare inom olika discipliner, samt inte minst med elever. Efter konstruktion och successiva miniutprövningar av skilda slag genomgår allt material stor utprövning i ett antal slumpvis valda skolor/klasser i landet. Riktvärdet är 400 elever per uppgift. Ankaruppgifter, som i ämnesprovet för årskurs 9 består av tolv mycket reliabla *items*[2], används konsekvent för att möjliggöra jämförelser mellan versioner, samt över tid, dvs. för att öka såväl validitet som reliabilitet. I samband med utprövningarna samlas även synpunkter med hjälp av enkäter in från samtliga medverkande lärare och elever. Såväl kvalitativa som kvantitativa metoder används vid analyserna av det insamlade utprövningsmaterialet, som följaktligen innehåller data både om elevernas prestationer och om deras och deras lärares uppfattningar om de olika uppgifterna. Alla dessa data, inte minst elevernas uppfattningar om olika aspekter av proven, till exempel rörande upplevd svårighetsgrad, bidrar väsentligt till utvecklingen av materialen.

Fastställande av benchmarks och betygsgränser

Fastställandet av krav- och betygsnivåer för *Part A* och *Part C* bygger på insamling av ett stort antal elevsamtal och texter, som analyseras successivt och till slut föreläggs ca tio referenter. I denna grupp väljs de exempel ut som sedan används för *benchmarking* av olika kvalitativa nivåer av muntlig respektive skriftlig interaktion och produktion. Förfaringssättet beskrivs närmare i avsnittet nedan som behandlar den muntliga delen av provet.

Även vad gäller det poängbaserade *Part B* bestäms kravgränser för de olika betygsstegen i samråd med referenter. Gruppen består i detta fall av ca 15 personer, samtliga lärare i engelska, men med variation i ålder, kön, erfarenhet, ämneskombination, skolform och geografiskt område. Denna, liksom andra grupper som arbetar

---

[2] Uppgiften består av tolv korta, dialoger med en lucka i varje. Eleverna skall fylla i ett ord per lucka. Dialogerna är samlade under en gemensam rubrik men hör inte ihop innehållsmässigt.

med provet, förändras successivt, så att nya personer alltid skall finnas representerade. Den procedur som följs har drag av metoder som *modified Angoff* och *Bookmark* (Kaftandjieva, 2004). Gruppen gör provet, analyserar uppgifter i relation till kursplanen och försöker också definiera vad deras allra svagaste elever på respektive betygsnivå skulle klara. På basis av detta lägger de sedan sina förslag. Dessa diskuteras, bl.a. i relation till data från utprövningen vad gäller såväl uppgifter som uppfattningar, och gruppen enas därefter om förslag till ungefärliga gränser. Det slutgiltiga fastställandet sker internt i projektet, i samråd med Skolverket, på basis av samtliga föreliggande data. Referensgruppens uttalande väger i detta sammanhang mycket tungt men är inte ensamt avgörande för beslutet.

Det muntliga delprovet – *Part A*

Utvecklingen av de två delprov som har fokus på tala/samtala respektive skriva *(Part A* och *Part C)* följer i stort sett samma rutiner, både vad gäller konstruktion, utprövning och fastställande av kravnivåer för olika betygssteg *(benchmarking).* Den enda skillnaden är att vissa strategiska val görs vid utprövningen av det muntliga delprovet, med ambitionen att säkerställa teknisk kvalitet på det material som tas fram och som i slutänden bildar underlag för bedömningsstödet i provet. Det totala antalet lärare och elever som medverkar i utprövningen är också mindre än för det övriga provet (ungefär hälften). Eftersom *Part A* inte ingår i den här rapporterade studien, ges här en något mera ingående beskrivning av utvecklingen av det muntliga delprovet.

Konstruktionen av muntliga uppgifter bygger, liksom vad gäller övriga delprov, på analys av kursplanens skrivningar, liksom på nationella och internationella erfarenheter och relevant forskning. I målen för engelska i årskurs 9 fastslås att eleverna dels skall ”kunna delta aktivt i samtal kring kända ämnen och med hjälp av olika strategier bidra till att kommunikationen fungerar”, dels kunna “muntligt berätta och beskriva något som hon eller han sett, hört, upplevt eller läst samt uttrycka och argumentera för en uppfattning i något för honom eller henne angeläget ämne”. Det handlar således både om interaktion och om produktion, med ett antal olika kontextuella syften med den muntliga framställningen. Uppgiftskonstruktionen utgår från detta och leder till provmodeller med flera steg, ökande komplexitet och med viss valmöjlighet för eleverna vad avser ämnen att tala och samtala om.

Det muntliga provet genomförs i par eller grupper om tre till fyra elever. Det inleds med en sk. *warm-up* del, där eleverna var och en beskriver och berättar något om vardagliga ämnen. Därefter följer en eller två delar med fokus på argumentation och diskussion. En vanlig modell är här att eleverna får välja bland ett antal kort med olika frågeställningar, dels av allmänt slag, dels med anknytning till aktuella, inte sällan samhälleliga och/eller värdegrundsrelaterade frågor. Fler teman än vad som till slut kommer att ingå i provet prövas ut. Vid utprövningarna spelas ett antal parsamtal på olika kvalitativa nivåer in, och elever och lärare besvarar enkäter om

skilda aspekter av den aktuella modellen och om bedömning av muntlig språkfärdighet mera generellt.

När utprövningsmaterialet sänts in sker en successiv bearbetning, som börjar internt i projektet. Ansvariga lyssnar på allt material och tar fram tentativa exempel på olika kvalitetsnivåer (fler än vad som sedan skall skickas till referenter). Två personer i projektet bedömer sedan dessa oberoende av varandra, och graden av överensstämmelse studeras, med avseende på inter- och intrabedömarreliabilitet. På basis av denna urvalsprocess väljs sedan ett femtontal samtal ut, kopieras på cd-skivor och skickas till åtta till tio referenter. Dessa gör oberoende av varandra bedömningar i den tiogradiga skalan och skriver även kommentarer till samtliga exempel. Likaså granskas själva provmodellen i relation till elevernas prestationer. De olika bedömningarna sammanställs därefter, och diverse analyser görs, bl.a. av bedömarsamstämmighet och olika bedömarprofiler. Utfallet diskuteras sedan under ett gemensamt sammanträde, då exempel väljs ut och kommenteras på basis av kursplanen och de vägledande bedömningsfaktorerna. Bedömda och kommenterade elevsamtal på cd inkluderas därefter i bedömningsanvisningarna till det aktuella provet.

Analyser av de tre senaste årens utfall av referensgruppernas bedömningar av de muntliga utprövningarna visar att samstämmigheten mellan bedömarna är god, med genomsnittliga korrelationer på strax över .90. Medelvärdena för enskilda bedömare varierar, inte sällan med tämlig systematik, inom och mellan omgångarna. Detta innebär att vissa bedömare med konsekvens bedömer något lägre respektive högre än andra, vilket är ett exempel på det som rubriceras som individuella bedömarprofiler. Variationen mellan de strängaste och mildaste bedömarna är dock att betrakta som måttlig. Under de tre studerade åren var den genomsnittliga variationsvidden 0.72 på den tiogradiga bedömningsskalan, alltså väsentligt mindre än ett skalsteg.

Slutligen kan påpekas, att utvecklingen av *Part C,* med fokus på skriftlig produktion, som tidigare nämnts, följer samma principer som för *Part A*. Grundmaterialet är här dock betydligt mera omfattande, dels på grund av att elevantalet i utprövningen är större, dels eftersom regelmässigt två ämnen prövas ut. Rutinerna för internt urval och validering är dock desamma, och 60 texter distribueras till de externa bedömarna. Motsvarande analys av de tre senaste årens utfall av referensförfarandet visar på genomsnittliga korrelationer mellan bedömarna på .92. Även i detta fall framträder tydliga och konsekventa bedömarprofiler. Variationsvidden är något större än för det muntliga delprovet, 1.08, detta dock till största delen betingat av en extremt sträng bedömare i en av de analyserade bedömningsomgångarna.

Inspelning av det muntliga provet

Frågan om inspelning av det muntliga delprovet har diskuterats sedan ämnesprovens införande i slutet av 1990-talet. I bedömningsanvisningarna till proven rekommenderas detta starkt. Tre skäl anges: Läraren kan, om inspelning görs, mera

helhjärtat fylla sin funktion av *coach* (inte partner) i samtalet; inspelning möjliggör sambedömning med kolleger, vilket ökar reliabilitet och likvärdighet i bedömningen; inspelning ger möjlighet till dokumentation även av denna del av det obligatoriska ämnesprovet. Indirekt kan inspelning också fylla en positivt styrande funktion på lärande och undervisning i engelska, men detta är inte något som berörs i anvisningarna till provet. Skolverket har dock inte bedömt det möjligt att kräva att inspelning görs, vilket bl.a. har till följd att de muntliga delproven i samtliga ämnen inte kunnat inkluderas i den här aktuella studien.

Frågan om inspelning av det muntliga provet i engelska har ingått i lärarenkäten till ämnesprovet sedan dess introduktion 1998. Första året, då lärare sannolikt tog för givet att det var obligatoriskt att dokumentera även *Part A,* uppgav drygt 40 % att de spelat in elevernas samtal, detta trots att ingen insamling av band gjordes. Denna andel har sedan sjunkit successivt och har de senaste åren legat relativt stadigt på mellan 20 och 25 % (Velling Pedersen, 2007).

Insamling

Vad gäller ämnesproven för åk 9 görs insamling av alla elevers resultat i landet. Dessutom samlas ett urval prov respektive resultatprofiler in för elever födda vissa datum (i fallet engelska fyra datum på året). Det senare vidarebefordras till de institutioner som utvecklar proven. Skolorna ombeds skicka kopior av det aktuella materialet, vilket stundtals vållar problem med läsbarhet. Av skäl som berörts ovan ingår inte muntliga delprov, i engelskans fall *Part A,* i de hela prov som skickas in.

Urvalsförfarande inför studien

De 100 prov som ingår i den här aktuella studien har dragits ur det material som sänts till institutionen på basis av de principer om datumurval som anges i instruktionerna angående SCBs totalinsamling av resultat. Dessa innebär bland annat att hälften av det material som skall skickas in utgörs av hela, kopierade elevprov, medan den andra hälften enbart rapporteras i form av en kopia på resultatprofilen. För 2007 års prov innebär detta att det här aktuella urvalet dragits ur ett totalt antal av 378 insända prov. Urvalet är slumpmässigt, men vissa justeringar har varit nödvändiga. Den övervägande orsaken till detta är att vissa prov inte var möjliga att kopiera på grund av den insända kopians dåliga tryckkvalitet. I några fall var också proven, i synnerhet skrivdelen, bemängda med så många lärarkommentarer och strykningar som inte kunde maskeras, att de bedömdes olämpliga att använda som underlag för ombedömning. Andra orsaker till justering av urvalet var att C-delen saknades och att könsfördelningen var alltför skev visavi totalantalet. Det senare har dock varit av så liten omfattning att det knappast torde påverka utfallet.

Urvalets relation till resultaten i totalinsamlingen respektive de insända resultatprofilerna framgår av Tabell 1.

Tabell 1: Urvalets relation till två större datamängder, Äp 9 engelska, 2007

	Totalinsamling (Antal elever vars resultat föreligger)	**Totalt antal resultatprofiler** (varav 47 % hela prov)	**100-urvalet**
Antal	Ca 128 900[3]	829[4]	100
Andel po/fl	51/49 %	51/49 %	50/50 %
Betygsfördelning (% EUM-G-VG-MVG)			
Provbetyg	5 – 36 – 42 – 17	5 – 37 – 41 – 17	6 – 44 – 37 – 13
Part A	4 – 41 – 38 – 17	4 – 42 – 37 – 17	5 – 42 – 36 – 17
Part B	8 – 33 – 39 – 20	8 – 35 – 38 – 19	11 – 37 – 36 – 16
Part C	6 – 43 – 37 - 14	6 – 45 – 35 – 14	6 – 46 – 31 – 17

Likheten mellan resultaten i totalinsamlingen och de insända profilerna är något större än mellan dessa två datamängder och det 100-urval som gjorts för studien. Detta är föga förvånande, dels med tanke på samplens storlek, dels sett i ljuset av de justeringar som varit nödvändiga i urvalet. Den här aktuella gruppens resultat är något svagare på det receptiva delprovet (B), vilket också återspeglas i det sammanvägda provbetyget (det senare dock med ett större bortfall än i de två övriga datamängderna). Resultatet på det muntliga delprovet (A) är i det närmaste identiskt med de större resultatmängderna, vilket på ett övergripande plan också gäller för skrivdelen (C). I det senare fallet är dock fördelningen av betygssteg något annorlunda i 100-urvalet, med en något större andel av det högsta betygssteget. På det hela taget kan dock det urval som gjorts för studien betraktas som rimligt representativt. Dessutom bör det hållas i minne att syftet med studien är att analysera samstämmigheten mellan lärares bedömningar, vilket inte på något direkt sätt hänger samman med elevresultatens representativitet.

Metod

Materialet till de olika bedömarna kopierades och distribuerades, tillsammans med ett standardiserat informationsbrev, i början av juni 2008. Kopieringen av materialet föregicks av avidentifiering samt, så lång möjligt, av maskering av lärarskrivna kommentarer i de olika provhäftena. Bedömningsarbetet utfördes under sommaren, med slutdatum för inrapportering den 31 augusti.

Part B

Baserat på delprovets/uppgiftstypernas karaktär samt tidigare studier, där överensstämmelsen mellan bedömare vad gäller den typ av uppgifter som finns i *Part B* visat sig mycket god (Olsson-Wahlsten, 2002), bestämdes i samråd med Skolverket att ombedömning av de 100 + 100 elevhäftena, med fokus på läs- respektive hörförståelse, skulle göras av en person, och att ytterligare en bedömare skulle bedöma ett slumpmässigt draget, 20%-igt delurval. Den person som gjorde den totala ombe-

[3] Antalet rapporterade resultat varierar något: Provbetyg: 127 103; *Part A* 128 081, *Part B* 128 918, *Part C* 128 223
[4] Antalet rapporterade resultat: Provbetyg: 780; *Part A* 800, *Part B* 817, *Part C* 806

dömningen (Bed 1) är en av de ansvariga för utvecklingen av ämnesprovet för åk 9, medan den andra (Bed 2) också arbetar i provprojektet, men inte aktivt med utveckling av provmaterial. Båda ombedömarna har lång lärarerfarenhet i engelska (Bed 1 också i tyska, Bed 2 i svenska och svenska som andraspråk). Båda bedömarna använde strikt de bedömningsanvisningar, inklusive exempel, som bifogas ämnesprovet.

Poängdata på fem nivåer finns att tillgå: enskilda frågor/*items* (där vissa ger *partial credit*), uppgifter (t.ex. en text med ett antal frågor), delarna B 1 respektive B 2 (läs- och hörförståelse) samt den aggregerade nivån för hela det receptiva delprovet (*Part B)*. Dessutom finns provbetyg i en tiogradig skala. Samtliga dessa nivåer har analyserats.

Part C

C-delen i ämnesprovet, med fokus på skriftlig produktion, ombedömdes av tre personer: en lärare som inte tidigare haft någon kontakt med provprojektet (A), en lärare som ofta anlitas som referent (B) samt en anställd i provprojektet, med ansvar för ämnesprovet (C). Bedömare A, som är lärare i engelska och svenska, är väsentligt yngre än de övriga och har arbetat ca tre år efter sin utbildning, i huvudsak i en förortsskola i en storstad. Bedömare B har drygt 20 års lärarerfarenhet i engelska och franska från en skola i en mindre stad, medan bedömare C undervisade ca 15 år i engelska och svenska i en landsortsskola, innan hon anställdes vid universitetet. Samtliga tre bedömare är kvinnor.

De tre bedömarna gjorde sina bedömningar helt oberoende av andra och varandra, i samma tiogradiga skala som används i provet och på basis av de bedömningsanvisningar och *benchmarks* som bifogas provet. De ombads också att under bedömningsarbetets gång, i de fall de uppfattat någon text som extra svårbedömd, i bedömningsprotokollet skriva ner korta reflektioner kring texten. Tekniken att arbeta med introspektion och *think [aloud] protocols* är väl etablerad i forskningen kring bedömning, bland annat av skriftlig produktion (Lumley, 2005), och har nyligen använts i en studie av de nationella kursproven för Engelska A (Åhs, 2007).

Vidare har bedömarnas reflektioner kring texter och bedömning samlats in vid ett internt möte, då de preliminära resultaten av studien presenterades.

Resultat

Nedan presenteras analyser som gjorts av ombedömningen av de båda delproven.

Part B – Receptiv förmåga (Läs- och Hörförståelse)

Resultaten av ombedömningen av urvalselevernas svar i B-delen i ämnesprovet i engelska kan sammanfattas på följande sätt (Lär = den samlade lärarbedömningen; CR=constructed response, SR=selected response; *Part B 1*, med fyra ingående uppgifter har fokus på läsförståelse, *Part B 2*, med två olika uppgifter, på hörförståelse):

Tabell 2. Ombedömning av *Part B* (n=99)[5], Äp 9 engelska, 2007; en extern bedömare

Uppgift	Svars-format	Max-poäng	m Lär	std	m Bed 1	std
Odd…	SR	16	11.68	3.64	11.80	3.48
Starting…	CR	12	7.86	3.01	7.85	3.07
Gandhi…	CR+SR	5	3.53	1.54	3.41	1.58
Friends…	CR+SR	21	11.39	6.32	10.78	6.30
PART B 1		*54*	*34.46*	*13.20*	*33.84*	*13.11*
Camping…	CR+SR	22	15.02	5.23	14.81	5.38
Environmental…	CR+SR	14	8.18	3.49	8.17	3.54
PART B 2		*36*	*23.20*	*8.25*	*22.95*	*8.43*
Andel CR/SR (%)	53/47					
***Part B* – poäng**		90	**57.66**	20.80	**56.72**	20.95
***Part B* – betyg**		10	**5.71**	2.59	**5.59**	2.62

Överensstämmelsen i bedömningarna är generellt mycket god, med en medelvärdesskillnad i totalsumman på 0.94 poäng av de totalt 90. Vad gäller de olika uppgifter som ingår i *Part B*, framgår att de små differenser som finns i regel härrör sig från uppgifter med till övervägande del *constructed response*, två med fokus på läsförståelse och en på hörförståelse (*Gandhi, Friends* och *Camping*). Det kan dock noteras att den första uppgiften i *Part B 1* tycks ha bedömts aningen mera generöst av ombedömaren än av lärargruppen. Detta kan förefalla besynnerligt, eftersom det här rör sig om en SR-uppgift *(multiple matching).* Vid noggrann analys av de insända provhäftena tycks dock som om vissa svar i några prov har fyllts i efter det att läraren gjort sin bedömning, dvs. med all sannolikhet vid genomgång av provet. Detta förklarar diskrepansen mellan svaren och den totala poängsumma som läraren angivit. Tilläggen syns dock inte i de maskerade kopiorna till ombedömaren, som därmed i vissa fall bedömt en något större andel av svaren som korrekta. I analys av utfallet av ombedömningen kan det vidare noteras att den andra läsförståelseuppgiften i *Part B 1*, *Starting,* som är en produktiv uppgift med medvetet valda, borttagna ord (s.k. *rational deletion),* uppvisar lika god bedömaröverensstämmelse som de uppgifter som innehåller till största delen, eller enbart, flervalsuppgifter.

I de allra flesta fall är överensstämmelsen mellan lärargruppen och ombedömaren i det närmaste total. Vid närmare analys framgår tydligt att det är enstaka frågor/*items* i *Part B* som åstadkommer skillnader i lärarnas och ombedömarens värdering av svar. Det mest tydliga exemplet på detta, och egentligen det enda där skillnaden är annat än marginell, finns i den längre lästexten *Friends,* där eleverna i en fråga med egna ord skall förklara uttrycket *"Many hands make light work".* Kravet i

[5] På grund av ett kopieringsfel har en elevs *Part B* uteslutits ur analysen.

bedömningsanvisningarna är att båda leden i uttrycket, liksom sambandet dem emellan, skall klargöras. Ett flertal autentiska exempel ges också, både på godtagbara och icke godtagbara svar. Lärarnas bedömning tycks här, trots det tämligen omfattande bedömningsstödet, ha varit något mera generös än ombedömarens (en skillnad på i genomsnitt 0.25 poäng).

Analyser av de 20 slumpvis valda *Part B,* som bedömts av ytterligare en person (Bed 2), uppvisar mycket god samstämmighet mellan de tre bedömarna.

Tabell 3. Ombedömning av *Part B* (n=20), Äp 9 engelska, 2007; två externa bedömare

Uppgift	Max-poäng	m Lär	std	m Bed 1	std	m Bed 2	std
Odd…	16	12.15	3.41	12.45	3.48	12.45	3.00
Starting…	12	8.20	2.48	8.25	2.65	8.20	2.61
Gandhi…	5	3.70	1.17	3.55	1.23	3.55	1.23
Friends…	21	13.05	5.63	12.00	5.56	12.30	5.61
PART B 1	*54*	*37.10*	*11.30*	*36.25*	*11.14*	*36.50*	*11.18*
Camping…	22	15.90	4.19	15.40	4.89	15.40	4.88
Environmental…	14	8.30	3.34	8.20	2.61	8.25	3.39
PART B 2	*36*	*24.20*	*7.19*	*23.60*	*7.92*	*23.65*	*7.84*
Andel CR/SR (%)	53/47						
***Part B* – poäng**	90	**61.30**	18.12	**59.85**	18.58	**60.15**	18.50
***Part B* – betyg**	10	6.15	2.30	5.90	2.45	6.00	2.43

Även här är som synes samstämmigheten mellan bedömarna mycket god, med endast marginella skillnader mellan bedömningen gjord av lärarna och de två externa bedömarna. Skillnaden mellan lärarna och Bedömare 1 är på detta urval av prov 1.45 poäng av de 90 totalpoängen. Bedömare 2 ligger 0.3 poäng högre än Bedömare 1 och 1.15 poäng lägre än lärarna. Även i detta fall härrör sig merparten av skillnaden från de tidigare kommenterade uppgifterna.

En beräkning av samstämmigheten mellan lärarnas poängbedömningar och ombedömningarna av hela urvalet, beräknad med en icke-parametrisk rangkorrelation, visar på korrelationer över .99, alltså på synnerligen god överensstämmelse.

Tabell 4. Korrelationer mellan lärarbedömningar och ombedömningar av *Part B*, Äp 9 engelska, 2007

			Lär sum *Part B*	**Bed 1** sum *Part B*	**Bed 2** sum *Part B*
Spearman's rho	**Lär** sum *Part B*	Correlation Coefficient Sig. (2-tailed) N	1,000 , 100	**.994**** ,000 99	**.994**** ,000 20
	Bed 1 sum *Part B*	Correlation Coefficient Sig. (2-tailed) N	.994** ,000 99	1,000 , 99	**.998**** ,000 20
	Bed 2 sum *Part B*	Correlation Coefficient Sig. (2-tailed) N	,994** ,000 20	,998** ,000 20	1,000 , 20

** Correlation is significant at the .01 level (2-tailed).

Slutligen kan nämnas att delprovsbetyget på *Part B* beräknas utifrån summerade poäng i en tiogradig skala. Inga ytterligare överväganden, som till exempel viktningar av vissa uppgifter, görs i samband med detta. Överensstämmelsen mellan totalsumman på delprovet och delprovsbetyget kan följaktligen förväntas vara mycket hög, eller total. Detta bekräftas i studien, där korrelationerna ligger över .99 i samtliga konstellationer.

Part C – Skriftlig produktion

Liksom vad gäller *Part B,* kommer lärarna i analyserna av skrivdelen att behandlas som en enhet. Detta har naturligtvis sina svagheter, och det bör hela tiden hållas i minne att det här, med all sannolikhet, handlar om 100 olika individer.

Bedömningarna kommer med ett undantag att redovisas i den tiogradiga skalan, eftersom denna använts både av lärarna och av ombedömarna. Analyser har visserligen gjorts även i en fyrgradig skala, men det förefaller inte helt korrekt att anta att de båda skalorna har uppfattats som identiska, eller är utbytbara från analys- eller redovisningssynpunkt. Enda undantaget från detta är att spridning i ett fall kommer att illustreras med grafer baserade på fyra steg, eftersom det i detta specifika sammanhang har bedömts bidra till större tydlighet kring tendenser i bedömningen, som också är av mera generellt intresse.

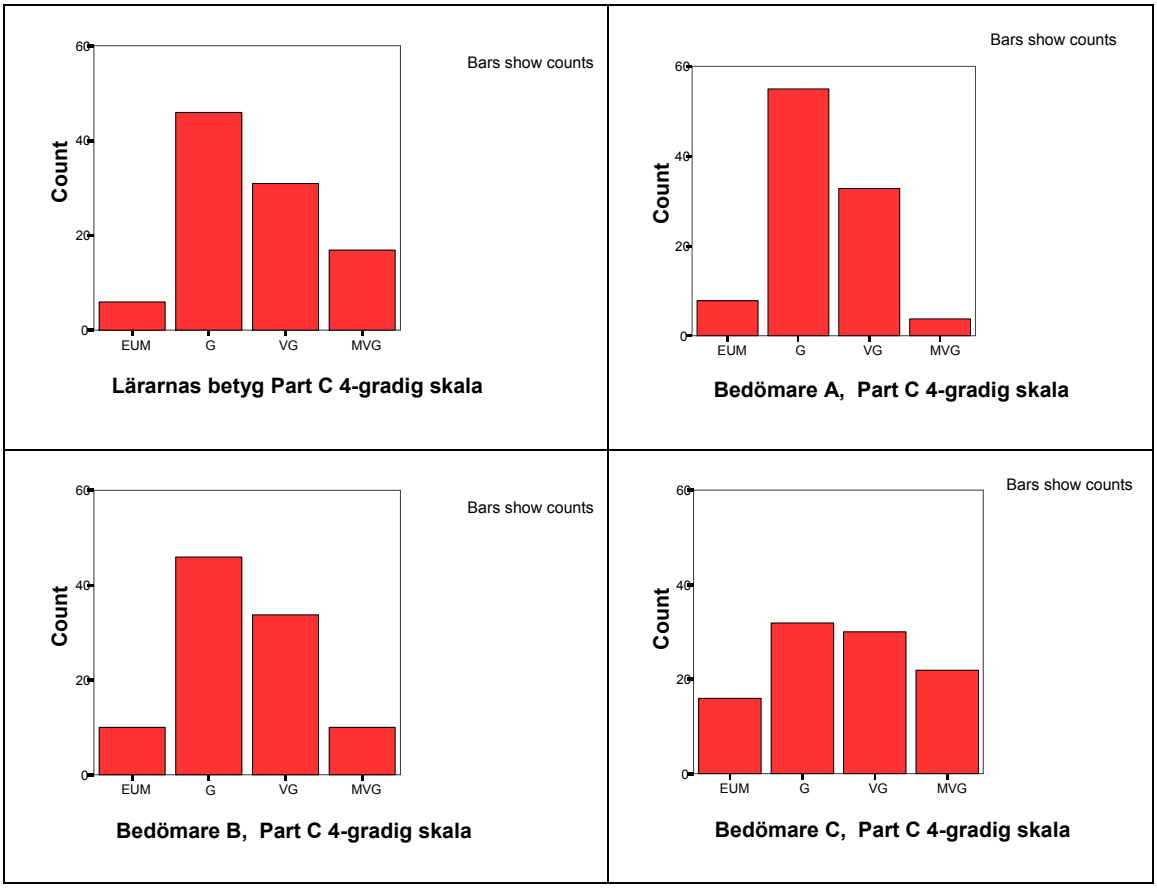
Medelvärdena av bedömningarna av de 100 texterna i urvalet är följande:

Tabell 5: Medelvärden av bedömningar av *Part C,* Äp 9 engelska, 2007; tiogradig skala

Bedömare	Medel-värde	Standard-avvikelse
Lärargruppen	5.59	2.44
Bed. A	5.00	1.91
Bed. B	5.38	2.19
Bed. C	5.82	2.90

Skillnaden mellan det lägsta och högsta värdet är 0.82 på den tiogradiga skalan. Värdena håller sig inom ett steg på denna skala. Det lägsta medelvärdet, 5.00, motsvarar ett resultat i det övre skiktet av Godkänt medan det högsta, 5.82, ligger betydligt närmare det nedre skiktet av Väl godkänt. Dessa två ytterkantsvärden hänför sig till två av ombedömarna (A och C), medan medelvärdena för lärargruppen och Bedömare B ligger mellan dessa två, på 5.59 respektive 5.38.

Måtten på standardavvikelse visar att bedömarna sprider sina bedömningar på klart olika sätt, att de alltså utnyttjar skalan i olika hög grad. Detta blir ytterligare tydligt i en grafisk representation, här baserad på den 4-gradiga skalan.



Figur 1: Lärarnas och de tre ombedömarnas användning av olika betygssteg i en fyrgradig skala; Äp 9 engelska, 2007, *Part C*

Trots att graferna visar på betydande skillnader i distributionen av de olika stegen, kan också vissa likheter noteras. Betygssteget Godkänt är till exempel det mest frekventa i samtliga bedömningar, följt av VG. I övrigt är det emellertid tydligt att den fördelningsmässigt största skillnaden finns mellan bedömare A och C, alltså mellan de två som också uppvisade den största differensen i medelvärde. Grafen för Bedömare A är ett tydligt exempel på s.k. centraltendens, alltså att de två mellersta stegen (G och VG) utnyttjas i mycket stor omfattning, medan de två yttre (EUM och MVG) är betydligt mindre frekventa. I synnerhet tycker sig denna bedömare ha svårt att identifiera texter som uppfyller kraven för det högsta bedömningssteget. Bedömare C däremot utnyttjar samtliga steg i påtaglig grad och identifierar både väsentligt fler texter som inte når upp till godkänd nivå och sådana som fyller kraven för MVG. Profilen för de samlade lärarbedömningarna visar att bedömningssteget G, följt av VG, dominerar tydligt, vilket liknar utfallet för Bedömare A och B. Lärarnas bedömningar har emellertid också ett gemensamt drag med Bedömare C vad gäller andelen MVG, som är mera generös än hos Bedömare A och B.

Korrelationerna mellan lärarnas bedömningar och ombedömningen i den tiogradiga skalan, visar på följande:

Tabell 6: Korrelationer mellan lärarbedömningar och tre externa ombedömningar, Äp 9 engelska, 2007, *Part C*; tiogradig skala

			Lär bed *Part C*	**Bed A** bed *Part C*	**Bed B** bed *Part C*	**Bed C** bed *Part C*
Spearman's rho	**Lär** bed *Part C*	Correlation Coefficient	1,000	**.864****	**.855****	**.890****
		Sig. (2-tailed)	,	,000	,000	,000
		N	100	100	100	100
	Bed A bed *Part C*	Correlation Coefficient	.864**	1,000	**.882****	**.907****
		Sig. (2-tailed)	,000	,	,000	,000
		N	100	100	100	100
	Bed B bed *Part C*	Correlation Coefficient	,855**	,882**	1,000	**.929****
		Sig. (2-tailed)	,000	,000	,	,000
		N	100	100	100	100
	Bed C bed *Part C*	Correlation Coefficient	,890**	,907**	,929**	1,000
		Sig. (2-tailed)	,000	,000	,000	,
		N	100	100	100	100

** Correlation is significant at the .01 level (2-tailed).

Korrelationerna mellan de olika bedömningarna sträcker sig mellan .86 och .93. Samtliga är signifikanta på enprocentsnivå. Den lägsta korrelationen är mellan de samlade lärarbedömningarna och Bedömare B (.855), den högsta mellan Bedömare B och C (.929). Det senare har sannolikt, åtminstone delvis, att göra med att dessa två personer bedömt gemensamma texter förut, nämligen i den referensgrupp som

tar fram anvisningarna till ämnesprovet. Den samlade lärarbedömningen visar störst samstämmighet med Bedömare C (.890), men skillnaderna är på det hela taget små. Eftersom det ingalunda är oproblematiskt att behandla lärarbedömningarna som om de emanerade från en person, när i själva verket varje lärare bara bedömt en text, kan det vara intressant att också enbart studera de tre bedömare (A, B och C), som var och en, utan samråd med andra eller varandra, bedömt samtliga 100 texter. Dessa personer, alla lärare i engelska, men med tämligen olika bakgrund och nuvarande professionell kontext, använder, som visats tidigare, skalan på olika sätt. De är också olika vad avser grad av stränghet, emellertid inom ramen för väsentligt mindre än ett skalsteg på den tiogradiga skalan. Korrelationerna mellan deras bedömningar är tämligen höga, från .882 (A och B), via .907 (A och C) till .929 (B och C). Detta visar att bedömarna i stor utsträckning är samstämmiga i sina rangordningar av texter och på så sätt konsekventa i sina bedömningar, om än representerande något olika bedömarprofiler.

Beräkningar har också gjort på basis av den fyrgradiga skalan. Korrelationerna blir här självklart något lägre, eftersom skalstegen är färre (mellan .78 och .82). Som tidigare nämnts, förefaller det dock mindre lämpligt att gå till väga på detta sätt, med en i efterskott applicerad skala, även om det ytligt sett resulterar i att fler texter hamnar på samma steg. Både lärarna och ombedömarna har gjort sina överväganden och fattat sina beslut utifrån den tiogradiga skalan, och man kan inte veta vilka gränsbeslut de skulle ha tagit om inga mellansteg mellan betygsstegen funnits. De högre korrelationerna för den tiogradiga skalan är också en påminnelse om att precision till viss del hör samman med antalet skalsteg.

Eftersom korrelationsberäkningar av det traditionella slaget enbart kan göras mellan två bedömare i taget, har även en sk. generaliserbarhetsanalys genomförts, med hjälp av programmet Genova, introducerat av Crick och Brennan, 1983, och därefter successivt vidareutvecklat (Brennan, 2001). Denna typ av analys bygger på en skattning av varianskomponenter med hjälp av variansanalys och tar hänsyn till variationen, inklusive mätfelen, mellan flera bedömare. Enkelt uttryckt skulle det kunna beskrivas som en generaliserad korrelationsanalys, som ger ett mått på reliabiliteten i den totala bedömningen. Analysen, på den tiogradiga skalan, resulterade i det här aktuella fallet, med lärarnas bedömning behandlad som en variabel, i en sk. *generalizability coefficient* på .85, vilket får betraktas som ett högt värde i den här typen av studie, med en heterogen grupp av icke samtränade bedömare.

Ett ytterligare sätt att se på samstämmighet är att analysera graden av överensstämmelse mellan bedömarna med utgångspunkt i bedömningen av de enskilda elevtexterna, dvs. att anlägga ett mera horisontellt perspektiv. Ett sätt att göra detta är att studera antalet ”differenssteg” på den tiogradiga skalan, där 0 innebär total överensstämmelse mellan alla fyra bedömarna (läraren och bedömare A, B och C) och 9 betyder att både det lägsta och det högsta skalsteget använts för samma text.

Analysen av antalet differenssteg ger följande resultat:

Steg	0	1	2	3	4	5	6	7	8	9	
Antal	8	30	39	17	6	0	0	0	0	0	[m=1.83]

Åtta av de 100 texterna bedömdes identiskt av läraren och de tre externa bedömarna. Den största differensen i bedömning, 4 skalsteg, gäller sex texter. Medelvärdet för differensen räknat på den tiogradiga skalan är 1.83 steg. En analys dels av de identiskt bedömda texterna, dels av dem där meningarna går mest isär kan sammanfattas på följande sätt:

Tabell 7: Analys av identiskt respektive maximalt olikt bedömda texter, Äp 9 2007, *Part C;* tiogradig skala[6]

	0 stegs differens (n = 8)	**4 stegs differens** (n = 6)
Strukturerat ämne	4 texter	3 texter
Friare ämne	4 texter	3 texter
Betygssteg på texterna	*2:* 2 texter; *3:* 2; *4:* 3; *10:* 1 text	*3/7* (2 texter); *4/8; 5/9; 6/10* (2 texter)
Kön	7 po, 1 fl	4 po, 2 fl
Sammanvägt provbetyg	*2:* 1 text; *3:* 2; *4:* 2; *5:* 1; *10:* 1; *saknas*: 1 text	*4:* 1 text; *6:* 1; *7:* 1; *8:* 1; *9:* 2 texter

Det kan knappast hävdas att analysen visar på någon större systematik, och eleverna i respektive kategori är framför allt för få för att några egentliga slutsatser skall kunna dras. Några försiktiga iakttagelser kan ändå göras. De båda ämnena, som har delvis olika karaktär vad gäller graden av struktur respektive frihet, förekommer i lika stor omfattning, både i de identiska och i de mest divergerande bedömningarna. Detta skall ses i ljuset av att eleverna, både i 100-urvalet och i det större, insända materialet, betydligt oftare valt det strukturerade ämnet[7]. Vad gäller könsfördelning, kan noteras att andelen pojkar är betydligt större i den identiskt bedömda kategorin och även något större i gruppen där olikheterna var som tydligast.

Likaså representerar i detta urval de texter där enigheten var total, med ett undantag, en något lägre kvalitativ nivå än dem där meningarna gick kraftigt isär. En analys av medelvärdet för samtliga fyra bedömares bedömning av texterna per differenssteg visar på följande:

[6] Bedömningssteg: 1 = ej uppnått målen; 2 = övre skiktet av ej uppnått målen; 3 = nedre skiktet av Godkänt osv. – 9 = nedre skiktet av Mycket väl godkänt; 10 = Mycket väl godkänt

[7] Ca 70 % valde det strukturerade ämnet, 30 % det friare. Inga stora skillnader kan noteras i dessa två grupper, till exempel vad gäller könsfördelning eller resultat på provet.

Tabell 8: Medelvärde för bedömning av texter per differens-kategori; Äp 9 2007, *Part C;* tiogradig skala

Differens-steg	Medelvärde fyra bedömare
0 (n=8)	4.00
1 (n=30)	4.70
2 (n=39)	5.28
3 (n=17)	7.49
4 (n=6)	6.46

Ett tämligen tydligt mönster framträder, nämligen att samstämmigheten tycks vara större i texter som bedömts som mindre avancerade. Oenigheten tilltar följaktligen vad gäller texter med ökande kvalitet, dock med undantag av den mest divergerande kategorin, med en differens på fyra skalsteg (det senare sannolikt påverkat av det begräsade antalet texter). En analys av övriga provdata visar att tendensen bekräftas av de individuella betygen på såväl *Part A* som *Part B* samt av det sammanvägda provbetyget. Det kan slutligen nämnas att samma fenomen, med större enighet i bedömning av svagare texter, rapporteras för ämnesprovet i svenska i den här aktuella studien.

Det bör slutligen betonas att de här gjorda iakttagelserna måste tolkas med stor försiktighet. Undersökningar i betydligt större skala behövs för att avgöra om de mönster i bedömarsamstämmighet i relation till kvalitativa nivåer som framträder har någon mera övergripande relevans.

Det kan vidare noteras att en text av de 100 som analyserats har bedömts med en skillnad som sträcker sig över två betygssteg, nämligen mellan det övre skiktet av Godkänt *(5)* och det nedre skiktet av Mycket väl godkänt *(9)*. Differensen härrör sig från bedömare A och C, som också är de som har de tydligaste och mest konsekventa profilerna av något strängare respektive mildare bedömning – det som i litteraturen brukar refereras till som *severity* respektive *leniency.* Både elevens lärare och Bedömare B har i detta fall bedömt texten som ett rent VG, och elevens sammanvägda provbetyg var också Väl Godkänt. Intressant att notera är också att ingen av de tre ombedömarna markerat texten som särskilt svårbedömd.

Slutligen har de texter som resulterat i störst differens (3 eller 4 steg) studerats närmare. Bilden av distinkta bedömarprofiler blir här mycket tydlig. Det kan dock framför allt noteras att det i dessa totalt 23 texter endast i åtta fall är elevens lärare som ensam representerar någon av ”ytterkantsbedömningarna”, och då i hälften av dessa som den som gjort den strängaste bedömningen. Detta är intressant, inte minst i ljuset av den stundtals artikulerade uppfattningen att lärare ofta ”snällbedömer” sina egna elever.

Bedömarnas kommentarer och reflektioner

För att bidra till förståelse kring bedömningsprocessen, inte minst orsaken till såväl differenser som överensstämmelser, ombads bedömarna att skriva ner reflektioner kring texter där bedömningen av något skäl fick dem att stanna upp och fundera extra mycket. Baserat på erfarenheter både från det reguljära arbetet med *benchmarking,* som beskrivits tidigare, och från studier där introspektions- och protokollstekniker använts (bl.a. Åhs, 2007), vet vi att detta förfaringssätt på ett positivt sätt bidrar till provutveckling, inte minst vad gäller uppgiftsformuleringar och bedömningsanvisningar.

Bedömarna använde sig i olika omfattning och på olika sätt av möjligheten att lämna kommentarer. Antalet texter som kommenterats av de tre individuella bedömarna varierade från 15 till 42, och längd och ”stil” skilde sig mycket åt. Totalt kommenterades 57 av de hundra texterna, i sex fall av alla tre bedömarna, i 17 fall av två och i 34 fall bara av en. Av de sex texter som uppenbarligen vållade mest funderingar kan fem rubriceras som *cliff hangers,* dvs. bedömarna tvekade mycket i det allra nedersta skiktet av Godkänt, medan den sjätte är en av de totalt sex texter där differensen mellan bedömarna är störst, med en variation från ett resultat i det nedre skiktet av G till VG. Samtliga kommentarer i det senare fallet fokuserar elevens [bristande] ämnesbehandling.

En preliminär analys av de olika kommentarerna visar relativt stor enighet mellan bedömarna om vad som vållar mest problem vid bedömningen. Fyra tentativa kategorier kan identifieras, och dessa återfinns hos alla tre bedömarna, alltså oavsett mängd och typ av kommentarer de lämnat. De fyra kategorierna, som även uppvisar viss likhet med den nyckelordsanalys som redovisas i Åhs’ studie (2007) av kursproven i Engelska A, är följande:

- Relation/balans mellan innehåll, textstruktur och språklig form
- Begriplighet/språklig form
- Ämnesbehandling/*task fulfillment*
- Textmängd/längd

Frågeställningar runt samma teman artikulerades också vid det möte med bedömarna som genomfördes strax efter det att de avslutat sitt arbete och de första, preliminära resultaten var klara. Vid detta tillfälle diskuterades också både individuella bedömarprofiler och eventuella yttre kontexter, ramar och kulturer som påverkar enskilda lärares sätt att definiera kvalitet och att bedöma elevarbeten. Värdet av sambedömning framhölls starkt, även om bedömarna inte hade upplevt bedömningen av de 100 texterna under den angivna tiden (två och en halv månad) som särskilt betungande. Samtliga beskrev ett förhållningssätt med en initial bedömning och sedan återkomst till materialet vid senare tillfälle(n). Under mötet diskuterades också några av de texter som givit upphov till stor variation, vilket ytterligare förtydligade de komplicerande faktorer vid bedömning som kommenterats under be-

dömningsarbetets gång och som redovisas ovan. Samstämmigheten mellan de tre bedömarna kring de aktualiserade texterna ökade också påtagligt under resonemangets gång.

Sammanfattande reflektioner och slutsatser

Resultaten av den gjorda studien är intressanta, inte bara som en indikation på graden av samstämmighet mellan bedömare i de nuvarande nationella proven utan också för vidareutveckling på flera plan. Detta gäller givetvis principer för och konstruktion av uppgifter och bedömningsanvisningar inom respektive ämne, men också mera övergripande, policyrelaterade frågor.

En kort sammanfattning av resultaten vad gäller engelska är att bedömarsamstämmigheten generellt förefaller vara mycket god vad gäller läs- och hörförståelseuppgifter, trots att enstaka *items* ger upphov till viss variabilitet. Uppgifterna tycks alltså fungera för sitt syfte och bedömningsanvisningarna ge gott stöd för en rimligt enhetlig bedömning. Detta stämmer också väl överens med de svar som över tid lämnats i lärarenkäterna som medföljer proven och också med data som genererats i utprövningarna. Även vad gäller skriftlig produktion är samsynen mellan bedömare god, även om korrelationerna här naturligt nog är lägre än för de receptiva uppgifterna. Sett ur ett allmänt och internationellt perspektiv, kan dock konstateras att interbedömarreliabiliteten i *Part C* är hög, i synnerhet i ljuset av att uppgifterna är tämligen fria, eleverna skriver relativt långa texter och bedömarna inte aktivt samtränat(s) för sitt arbete. Detta innebär dock på intet sätt att samstämmigheten bedömare emellan skulle ha nått sin möjliga högstanivå, eller att metoder för att ytterligare bidra till gemensamma tolkningar och därmed ökad likvärdighet inte behöver vidareutvecklas. Det bör här också återigen framhållas att lärargruppen utgörs av 100 olika lärare, och stabiliteten i deras individuella bedömningar givetvis kan variera betydligt.

Vad gäller uppgifter som avser bedömning av receptiv förmåga (läs- och hörförståelse) måste bredden av texter, läsarter och svarsformat upprätthållas och vidareutvecklas. Av validitetsskäl måste det t.ex. finnas texter och uppgifter som även prövar elevernas förmåga till inferens, konklusion, reflektion och interaktion, trots att detta ibland ställer vissa krav på produktion. Det är dock samtidigt viktigt att inte skrivandet tar överhanden, utan att elever också ges möjlighet att visa förståelse på olika plan utan att behöva skriva. Resultaten i studien visar också på att det är fullt möjligt att upprätthålla hög reliabilitet även i bedömningen av uppgifter med elevproducerade svar, detta givet att bedömningsanvisningarna är tydliga och generösa vad gäller exemplifiering. Arbetet måste därför envist fortsätta med att skapa en rimlig balans mellan uppgifter med, i vid bemärkelse, olika utformning, till förmån både för bedömningens validitet och reliabilitet.

Det rimmar väl med de svenska kursplanerna i språk att ge både ett mera, och ett mindre strukturerat ämne som utgångspunkt för skrivande, och responsen från

lärare och elever kring detta är också positiv. Det är dock väsentligt att via utprövningar säkerställa att dessa ämnen, just på grund av sin olikhet, inte introducerar *bias* genom att appellera särskilt till, eller faktiskt gynna eller missgynna, enskilda grupper av provtagare. Andra frågor kring skrivprovet som behöver diskuteras från olika utgångspunkter är effekten av ordantal, och huruvida minimi- och maximigränser skall anges för elevernas texter. Likaså är aspekter *task fulfillment* väsentliga, dvs. elevers närhet respektive distans till uppgiften, och vad som kan krävas i relation till instruktionerna för att ett ämne skall anses behandlat. Alla dessa aspekter av bedömning diskuterades av *Part C*-bedömarna som exempel på faktorer som kan påverka bedömningen av enskilda texter.

En till synes evig diskussion vad gäller bedömning av muntlig och skriftlig produktion och interaktion är huruvida denna skall göras holistiskt eller analytiskt, dvs. med ett övergripande omdöme eller via separata bedömningar av definierade delkomponenter, som i slutänden (eventuellt) vägs samman till ett omdöme. Det finns uppenbara fördelar med båda tillvägagångssätten (Cushing Weigle, 2002), men i ett betygsstödjande, nationellt prov torde ett helhetsomdöme vara mest adekvat. Frågan är då snarare hur stark den analytiska prägeln skall vara, dvs. vilken funktion de angivna bedömningsfaktorerna skall fylla. En ytterligare fråga är om à priori nivåbestämda skalor och deskriptorer skall utvecklas och bedömningen göras i relation till dessa (Bachman & Palmer, 1996). Ett närliggande exempel på denna typ av system är den europeiska referensramen för språk, *Common European Framework of Reference för Languages* (Council of Europe, 2001), till vilken de svenska kursplanerna successivt närmas. Slutligen är inte bara antalet *benchmarks* viktigt utan också hur dessa kommenteras i relation till – och därmed tydliggör – mål, kriterier och analytiska bedömningsfaktorer.

Studien aktualiserar på ett uppenbart sätt frågan om sambedömning. I en fråga som ställdes i samtliga lärarenkäter till de nationella proven våren 2008 framkom att det framför allt är vid bedömning av *Part C,* med fokus på skrivande, som lärare i engelska söker hjälp av varandra. Här uppger 22 % att samtliga texter medbedöms och en lika stor andel att många texter bedöms även av en kollega. Intressant är också att jämföra de olika delproven med avseende på svaret ”Samtliga bedöms av mig ensam”. Detta alternativ valdes av 64 % för det muntliga provet, 46 % för läs- och hörförståelse men bara 14 % vad gäller skriva *(Part C).* Eftersom samverkan kring bedömning både fyller funktionen av att öka samstämmighet bedömare emellan, och därmed öka likvärdighet, och att bidra till lärares fördjupade tolkning av kursplanerna, förefaller det högst rimligt att den nuvarande rekommendationen om sambedömning förstärks. Viktigt är dock att det som fokuseras inte (bara) är svårbedömda texter, utan minst lika mycket de fall där man som lärare känner sig säker. Det finns god anledning att tro att det i dessa förgivettaganden om kvalitativ nivå döljer sig mycket variation, vilket till viss del även framkommer i den här aktuella studien.

Sist men inte minst måste frågor om bedömning i vid bemärkelse ges stor uppmärksamhet i utbildning av blivande lärare, men också i planering av verksamhet för dem som redan verkar i professionen. Detta skulle bland annat kunna innebära att det inom ramen för det nationella provsystemet, för att bidra till ökad likvärdighet, regelmässigt tillhandahålls material att använda i kompetensutvecklande syfte, enskilt och inom skolor, men också mellan skolor och kommuner. Ett exempel på sådant material i språk är DVD-produktioner kring bedömning av muntlig språkfärdighet i engelska åk 5, och franska, spanska, tyska Steg 3. Dessa producerades, på uppdrag av Skolverket, vid Göteborgs universitet och distribuerades kostnadsfritt till alla skolor under vårterminen 2006.

Referenser

Alderson, C. (2000). *Assessing Reading.* Cambridge: Cambridge University Press.

Bachman, L. F. & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests.* Oxford: Oxford University Press.

Brennan, R.L. (2001). *Generalizability Theory*. New York: Springer Verlag.

Council of Europe (2001). *Modern Languages: Learning, Teaching, Assessment. A Common European Framework of Reference.* Cambridge: Cambridge University Press.

Cushing Weigle, S. (2002). *Assessing Writing.* Cambridge: Cambridge University Press.

Erickson, G. (2006). Bedömning av och för lärande - En kollaborativ ansats i arbetet med nationella prov i språk. I U. Tornberg (red.), *Mångkulturella aspekter på språkundervisningens kommunikativa praktiker. En konferensrapport* (s. 187-207). Örebro: Örebro Universitet. Hämtad 27 oktober 2008 från http://www.oru.se/templates/oruExtNormal.aspx?id=8719

Gray, W. S. (1960). The Major Aspects of Reading. I H. Robinson (ed.) *Sequential Development of Reading Abilities* (Vol. 90, s. 8-24). Chicago: University of Chicago.

Kaftandjieva, F. (2004). Standard setting. I Council of Europe, *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages* (Section B). Hämtad 27 oktober 2008 från http://www.coe.int/t/dg4/linguistic/Manuel_EN.asp - TopOfPage

Lumley, T. (2005). *Assessing Second Language Writing: The Rater's Perspective*. Frankfurt am Main: Peter Lang (Language Testing and Evaluation series, volume 3).

Messick, S. A. (1989). Validity. I R. L. Linn (ed.), *Educational Measurement* (Third edition, s. 13-103). New York: American Council on Education/Macmillan.

Olsson-Wahlsten, C. (2002). *Öppna svar – hur funkar det? Elever svarar och lärare bedömer i en läsförståelseuppgift i ett nationellt prov i engelska.* (D-uppsats i pedagogik med didaktisk inriktning). Göteborg: Göteborgs universitet, Institutionen för pedagogik och didaktik.

Velling Pedersen, D. (2007). Engelska. I Skolverket *Ämnesprovet 2007 i grundskolans årskurs 9.* Hämtad den 27 oktober 2008 från http://www.skolverket.se/sb/d/306/a/1881

Åhs, M. (2007). *Bedömning av fri skriftlig produktion i engelska – Teori, procedur, process. En studie av de nationella proven.* (Mastersuppsats i ämnesdidaktik). Göteborg: Göteborgs universitet, Institutionen för pedagogik och didaktik.