PRIM gruppen

Bedömaröverensstämmelse vid bedömning av ämnesprovet i matematik för årskurs 9

Ämnesprovet i matematik

För att kunna bedöma elevens kunskaper i matematik mot kursplanens olika mål och mot betygskriterierna behövs ett så brett bedömningsunderlag som möjligt. Ämnesprovet i matematik omfattar därför olika delar som ska ge eleven möjlighet att visa sina kunskaper på olika sätt. De olika delarna skiljer sig vad gäller kunskapsinnehåll, arbetssätt, redovisnings- och bedömningssätt.

Ämnesproven i matematik för årskurs 9 består av tre delprov. Delprov B och C finns i två versioner där enda skillnaden är att ingående tal är olika i en del av uppgifterna.

Delprov A är ett muntligt delprov som prövar elevens förmåga att muntligt framföra matematiskt grundade idéer samt förmåga att lyssna till, följa och pröva andras förklaringar och argument.

Delprov B består av två delar Del B1 (kortsvar) och Del B2 (problemlösning). Först besvarar eleverna kortsvarsdelen, där miniräknare inte får användas, och de övergår sedan till att lösa en ”mer omfattande” uppgift. Eleverna får själva avgöra när de vill lämna in kortsvarsdelen och börja använda miniräknare.

Del B1 prövar framför allt elevens taluppfattning och grundläggande färdigheter i räkning med naturliga tal, tal i bråk- och decimalform och procent. Några uppgifter prövar elevens kunskaper i grundläggande algebra, geometri och statistik. Exempel på godtagbara svar ges i bedömningsanvisningarna. Endast svaret beaktas.

Del B2 består av en ”mer omfattande” uppgift. Uppgiften kännetecknas av att lösningen är ganska omfattande och kräver motiveringar. Del B2 prövar elevens förmåga att lösa problem, reflektera över och tolka sina resultat samt bedöma deras rimlighet. Där prövas också elevens förmåga att uttrycka sina tankar skriftligt, dra slutsatser och generalisera.

Delprov C består av cirka 10 uppgifter som prövar kunskaper från flera olika kunskapsområden. Uppgifterna är samlade kring ett gemensamt tema. Delprovet prövar elevens förmåga att lösa problem samt reflektera över och tolka sina resultat och bedöma deras rimlighet. Det prövar också elevens förmåga att uttrycka sina tankar skriftligt.

Det är endast Delprov B och Delprov C som ingår i denna undersökning av bedömaröverensstämmelse.

Bedömning av ämnesprovet

G-poäng och vg-poäng

När proven konstrueras görs bedömningar av uppgifternas innehåll och elevlösningarnas kvalitet utifrån kursplanen och betygskriterierna. De olika uppgifterna

kategoriseras och elevlösningar från utprövningen analyseras och bedöms. För att tydliggöra de kvalitativa nivåerna som finns i mål att uppnå och i betygskriterierna, ges vid bedömningen g-poäng och vg-poäng. G-poäng relaterar till kunskaper som kan kopplas till mål att uppnå för årskurs 9 och vg-poäng relaterar till kunskaper som kan kopplas till VG- och/eller MVG-kriterier. Ibland är det subtila skillnader mellan de olika poängkvaliteterna. Bedömningen av vilka poäng som kan anses vara g- och vg-poäng i respektive provdel görs av referensgrupper med bl a yrkesverksamma matematiklärare.

Uppgifter markerade med symbolen ¤

Några uppgifter i provet inbjuder till lösningar och resonemang som indikerar kvaliteter som kan kopplas till betygskriterierna för MVG. Dessa uppgifter har vid utprövningar visat dessa kvaliteter i ett flertal elevarbeten. Det är uppgifter som i sig inte behöver vara särskilt komplicerade. Det är snarare så att dessa uppgifter kan lösas på flera sätt, vilket gör att eleverna kan använda en mer eller mindre generell metod och ett mer eller mindre utvecklat matematiskt uttryckssätt och språk. Uppgifterna är märkta med symbolen ¤.

Uppgifter som ska aspektbedömas med stöd av bedömningsmatris

Några delprov bedöms med stöd av bedömningsmatris. Syftet är att för läraren och eleven dels visa på de olika kunskapsaspekter som kan bedömas, dels att beskriva de olika kvalitativa nivåerna inom varje kunskapsaspekt. Dessa aspekter och beskrivningar är hämtade från kursplan och betygskriterier. Olika uppgifter kan fokusera på olika aspekter i matriserna. I bedömningsanvisningarna finns därför uppgiftsspecifika matriser som ska användas vid bedömningen. Resultatet av bedömningen ger ett antal g- och vg-poäng och eventuellt en kommentar om MVG-kvalitet.

Bedömning av de olika delproven

För *Del B1* gäller att korrekt svar bedöms med 1 g-poäng eller 1 vg-poäng.

För *Del B2* gäller att läraren gör en aspektbedömning med stöd av en uppgiftsspecifik bedömningsmatris och med stöd av exempel på autentiska elevarbeten på olika kvalitativa nivåer. Bedömningen resulterar i ett antal g-poäng och ett antal vg-poäng. Bedömningen grundar sig på hur väl eleven förstår problemet, hur eleven genomför lösningen och analyserar resultatet samt hur klart och tydligt eleven redovisar och använder det matematiska språket.

För *Delprov C* gäller att lösningen bedöms med g-poäng och/eller vg-poäng. Till de enskilda uppgifterna finns korrekta svar och bedömningsanvisningar för delpoäng. Efter varje uppgift anges maximala antalet poäng som en korrekt lösning ger. (2/3) betyder t ex att uppgiften kan ge högst 2 g-poäng och 3 vg-poäng. Elevarbetet ska bedömas med högst det antal poäng som anges i bedömningsanvisningarna. Enbart svar utan motiveringar ger inga poäng. För full poäng krävs korrekt redovisning med godtagbart svar eller slutsats. Vid bedömning av elevens arbete ska positiv poängsättning tillämpas. Utgångspunkten är att eleven ska få poäng för lösningens förtjänster och inte poängavdrag för fel och brister. En elev som kommit en bit på väg får då poäng för det som han/hon har gjort. Redovisningen ska vara tillräckligt utförlig och uppställd på ett sådant sätt att tankegången lätt kan följas. Korrekt metod eller förklaring till hur uppgiften kan lösas ska ge delpoäng även om det därefter följer en felaktighet, t ex räknefel. Om eleven också slutför uppgiften korrekt ger det fler poäng. Till bedömningsanvisningarna för vissa uppgifter finns det också bedömda autentiska elevarbeten på olika kvalitativa nivåer.

Vid arbetet med bedömningsanvisningar till ämnesproven är strävan att göra graden av interbedömarreliabilitet så hög som möjligt. Målsättningen är därför att beskrivningen till varje poäng ska vara så tydlig som möjligt. Avsikten med de bedömda elevlösningarna är att tydliggöra beskrivningen och därmed höja graden av likvärdighet.

Provbetyg

Beskrivningar av kraven för provbetygen Godkänt, Väl godkänt respektive Mycket väl godkänt ges för *ämnesprovet som helhet*. Ett enskilt delprov prövar en alltför begränsad del av målen och betygskriterierna i kursplanen för att kunna betygsättas. Läraren gör sin bedömning av elevernas prestationer enligt de bedömningsanvisningar som finns till varje delprov. Elevens resultat på de olika delproven läggs ihop och bildar då en poängsumma bestående av ett antal g-poäng och ett antal vg-poäng.

För provbetyget *Godkänt* krävs ett minsta antal poäng totalt.

För provbetyget *Väl godkänt* krävs dels att en viss totalpoäng uppnås, dels att ett visst antal av totalpoängen utgörs av vg-poäng.

Bedömningen *Mycket väl godkänt* på provet återspeglas inte bara i en poängsumma. För att en elev ska få detta provbetyg måste hon/han visa både bredd och djup i sina matematiska kunskaper. Bredden visas genom att eleven mer än väl uppfyller kravgränsen för Väl godkänt. Djupet bedöms genom att läraren särskilt studerar elevens arbete med de uppgifter i provet som är markerade med ¤. Här ska läraren söka belägg för att eleven uppvisar sådana kunskapskvaliteter som kan kopplas till betygskriterierna för MVG. För att erhålla provbetyget Mycket väl godkänt ska eleven ha visat prov på flertalet av dessa kvaliteter i sitt arbete.

En utgångspunkt för vårt arbete med beskrivning av kraven för olika provbetyg är hur man internationellt bestämmer kravgränser för olika betyg. Många olika metoder används, men flertalet kännetecknas av att en sammanvägning av olika experters bedömningar görs. I den sammanvägningen ingår tolkning av mål och kriterier, bedömningar av uppgifter mot mål och kriterier samt bedömningar av elevprestationer i förhållande till mål och kriterier. Förutom referensgruppens medlemmar deltar cirka tio verksamma matematiklärare för årskurs 7–9 i arbetet med att beskriva kraven för de olika provbetygen.

Genomförande av ombedömningen

Hur säkra är lärares bedömning av elevernas arbeten på nationella provet i matematik? Syftet med denna undersökning är att kartlägga överensstämmelsen mellan lärares bedömning utifrån de givna bedömningsanvisningarna till ämnesprovet.

Är den bedömning som görs av elevlösningar till uppgifter som bara kräver korta redovisningar säkrare än bedömning av elevlösningar på en större uppgift, som kräver mer omfattande redovisningar? Som bieffekt kanske undersökningen också kan ge exempel på vad i uppgifter eller bedömningsanvisningar som kan misstolkas. En kunskap som är mycket värdefull i det fortsatta arbetet med att konstruera nationella prov och tillhörande bedömningsanvisningar.

Urval av elevarbeten

Efter ämnesprovet 2007 samlade vi in resultat från ett datum-urval av elever. Via webbinsamlingen samlade vi in resultat på uppgiftsnivå för 1100 elever. Dessutom samlade vi in elevarbeten på de skriftliga delproven för ca 350 elever. Etthundra

elevarbeten valdes slumpvis ut från de inskickade elevarbetena. Ett krav på elevarbetet var att det presenterade någon form av lösning till den ”mer omfattande” uppgiften i Del B2. Om lösningen gav några poäng eller ej var inte avgörande. Ett annat krav var att det kunde avgöras antingen från elevarbetet eller från datafiler hur läraren bedömt elevarbetet. Om det fanns lösningar till samtliga mindre uppgifter eller ej togs det ingen hänsyn till då elevarbetena valdes ut.

Varje elevarbete fick sedan ett internt ID-nummer (1–100) och lärarens bedömning, fördelade på g- och vg-poäng, samt bakgrundsdata som kön och version registrerades. All form av bedömning, rättning och kommentarer från lärarna togs bort från de valda elevarbetena innan bedömarna fick materialet. Elevarbetena var alltså avkodade och gjorda av för bedömaren okända elever och någon bedömning kopplat till undervisning kunde inte förekomma.

Eftersom analyserna av Äp 9 2007 visade att det inte fanns några signifikanta skillnader mellan lösningsproportionerna hos version 1 och version 2 användes båda versionerna vid ombedömningen. 100-urvalet hade ungefär samma fördelning mellan både versioner och kön som det större datum-urvalet. Lösningsproportionerna för de enskilda uppgifterna i 100-urvalet avvek inte heller signifikant från vårt datum-urval.

Bedömargrupp

Tre matematiklärare bedömde individuellt de 100 utvalda elevarbetena. Med denna uppläggning gavs alltså möjlighet att jämföra likvärdigheten i bedömning mellan olika lärare, som inte hade någon referens till de enskilda eleverna.

De tre lärare som bedömde etthundra elevarbeten, deltog alla i ämnesprovet med sina egna elever då provet gavs. Två av lärarna har dessutom varit med och arbetat fram ämnesprovet. De var därför väl insatta i bedömningsanvisningarna. Lärarna arbetar på olika skolor i Stockholmsområdet och träffades inte under bedömningsarbetets första fas. Varje lärare fick därför använda sin tolkning av de skriftliga anvisningarna och eventuella tolkningar, som gjorts vid provtillfället på respektive skola. I denna rapport benämns dessa tre lärare som bedömare för att undvika förväxlingen med elevens lärare, som ju också bedömt elevarbetena.

Tabell 1 Bedömarnas kön, antal undervisningsår samt erfarenhet av bedömning av nationella prov

		Antal undervisningsår	Erfarenhet av bedömning av nationella prov
Bedömare 1	Man	13	4–9 lärare i Ma/Nv som har deltagit en gång på ett kravgränsmöte
Bedömare 2	Kvinna	10	4–9 lärare Ma/Nv som är med i referensgruppen för Äp 9 sedan 2003
Bedömare 3	Kvinna	9	Högstadielärare i Ma och Fy som arbetar 50 % i PRIM-gruppen med provutveckling av Äp 9

Ombedömningen

De tre bedömarna bedömde var och en för sig alla elevarbeten efter de bedömningsanvisningar som medföljde provet och resultaten bokfördes i en kalkylblankett. Därefter bearbetades kalkylblanketterna så att en jämförelse av de olika bedömarnas poängsättning förtydligades. Bedömarna träffades under en förmiddag och diskuterade sina bedömningar med utgångspunkt från den statistiska bearbetning-

en. Några speciella elevarbeten som var mycket otydligt kopierade diskuterades och bedömarna fick här ändra sin poängsättning när vi enats om vad som stod. Vi fann då vi gemensamt studerade statistiken att de små skillnader som fanns på Del B1 (kortsvar) var rena felskrivningar i kalkylbladet. Denna typ av felskrivningar finns naturligtvis även vid bokföringen av resultaten på uppgifterna i de andra delproven.

Bedömarna beskrev hur de i det praktiska arbetet försökt hålla en hög intrabedömarreliabilitet. Nivån hittar man inte från början utan ”det innebär att bedömningen av de första elevarbetena får göras om en gång”. En annan synpunkt som framkom var att man ser ett mönster i elevlösningarna till vissa uppgifter, då man bedömer så många lösningar. Bedömarna förde anteckningar om hur de bedömt vissa typiska elevarbeten för att var konsekventa genom hela arbetet. Alla tre bedömarna hade känt sig ensamma vid bedömningen. Då de normalt bedömer nationella prov brukar de diskutera bedömningen med sina kollegor.

Resultat och analyser av ombedömningen

Provbetyg

Ämnesprovet 2007 kunde på alla delprov sammanlagt ge maximalt 75 poäng varav 35 vg-poäng. För att få provbetyget Godkänt skulle eleven ha erhållit minst 23 poäng. För att få provbetyget Väl godkänt skulle eleven ha erhållit minst 44 poäng varav minst 14 vg-poäng. För att få provbetyget Mycket väl godkänt skulle eleven ha visat minst 6 MVG-kvaliteter och dessa skulle vara av minst tre olika slag. Dessutom skulle eleven ha erhållit minst 24 vg-poäng för att visa bredd i sina matematikkunskaper.

Eftersom denna ombedömning endast omfattar de skriftliga delproven B och C måste kravgränserna för alla betygsnivåer räknas om. Delprov A, det muntliga delprovet, kunde maximalt ge 4 g-poäng och 4 vg-poäng samt möjlighet att visa 4 olika MVG-kvaliteter. Vid kravgränsmötet inför Äp 9 2007 diskuterades hur många poäng av poängen för de olika kravgränserna som kom från Delprov A för de olika betygstegen. Av resultatfilen på uppgiftsnivå framgår hur många poäng som ”gränseleverna” (dvs de som precis nått en betygsgräns) fått i genomsnitt på det muntliga delprovet. Med stöd av dessa resultat och kravgränsgruppens bedömning av Delprov A beräknades de speciella kravgränser för Delproven B och C tillsammans som används i denna undersökning.

Vid analyserna av ombedömningen användes följande beräknade kravgränser för de olika provbetygen. Delprov B och Delprov C kunde sammanlagt ge maximalt 67 poäng varav 31 vg-poäng. ”Provbetygen” sattes enbart med stöd av totala antalet poäng och antalet vg-poäng som elevarbetena fick. På de flesta elevarbetena fanns tyvärr inga markeringar om vilka MVG-kvaliteter som läraren ansåg att elevarbetet visade. En analys av hur bedömarna använde MVG-kvaliteterna redovisas på sid 10.

- För att få ”provbetyget” Godkänt skulle eleven ha erhållit minst 20 poäng.
- För att få ”provbetyget” Väl godkänt skulle eleven ha erhållit minst 40 poäng varav minst 11 vg-poäng.
- För att få ”provbetyget” Mycket väl godkänt skulle eleven ha erhållit minst 20 vg-poäng.

Överensstämmelse mellan de beräknade provbetygen

Interbedömarreliabiliteten undersöktes genom att jämföra varje elevs beräknade provbetyg för bedömarna parvis.

Bedömare 1 och 2 är överens om ”provbetyget” för 94 % av elevarbetena, bedömare 1 och 3 är överens om ”provbetyget” för 98 % av elevarbetena och bedömare 2 och 3 är överens om ”provbetyget” för 93 % av elevarbetena.

Tabell 2 Procentuell överensstämmelse mellan bedömarna

Samma provbetyg för alla tre bedömarna	92 % av elevarbetena
En av bedömarna avvek ett steg med sitt provbetyg	8 % av elevarbetena

Bedömare 1 har ett steg högre ”provbetyg” på ett elevarbete, Bedömare 3 har ett steg högre ”provbetyg” på två elevarbeten och Bedömare 2 har ett steg lägre ”provbetyg” på tre elevarbeten och ett steg högre på två elevarbeten. Skillnaden i ”provbetyget” beror på en avvikelse på endast 1 eller 2 poäng mellan bedömarna.

Jämförelse med lärarens beräknade provbetyg

De provbetyg som beräknades på lärarens bedömning jämfördes med de tre bedömarnas.

Tabell 3 Procentuell överensstämmelse mellan läraren och bedömarna

	Andel elevarbeten
Alla tre bedömarna var överens med läraren	86 %
Två av bedömarna var överens med läraren	6 %
En av bedömarna var överens med läraren	2 %
Ingen av bedömarna var överens med läraren	6 %

På de sex elevarbeten där lärarens beräknade provbetyg avvek från alla bedömarnas var det fem arbeten som fick provbetyget MVG av läraren men VG av bedömarna och ett som var tvärtom. Provbetyget Mycket väl godkänt beskrivs speciellt på sid 10.

Tabell 4 Överensstämmelse mellan det beräknade provbetyget för läraren och typvärdet för bedömarnas beräknade provbetyg

		Provbetyg Bedömare				
		EUM	**G**	**VG**	**MVG**	**Total**
Provbetyg Lärare	**EUM**	12	0	0	0	12
	G	0	57	0	0	57
	VG	0	*3*	15	*1*	19
	MVG	0	0	*5*	7	12
	Total	12	60	20	8	100

En jämförelse med typvärdet för bedömarnas beräknade provbetyg med de beräknade provbetygen för lärarna visar att de är total överensstämmelse för ”provbetyget” EUM . Tre elevarbeten har fått ”provbetyget” VG av läraren medan typvärdet för bedömarna är G. Fem elever har fått ”provbetyget” MVG av läraren medan typvärdet för bedömarna är VG, men det finns också ett elevarbete som fått VG av

läraren medan typvärdet för bedömarna är MVG. Några av lärarna var generösare med poängen än bedömarna vilket framgår av analysen av poängen.

Överensstämmelse mellan summapoängen

Bedömningen av varje elevarbete har gjorts av fyra personer, tre bedömare samt elevens lärare. Vid denna jämförelse betraktas varje poängtilldelning som fyra separata bedömningar. Detta är dock något missvisande för lärargruppen, vilken ju inte representerar en persons bedömning av alla de etthundra elevarbetena, som de övriga, utan troligen etthundra olika lärares bedömning av ett elevarbete.

Tabell 5 Korrelation mellan bedömares och lärares totalpoäng

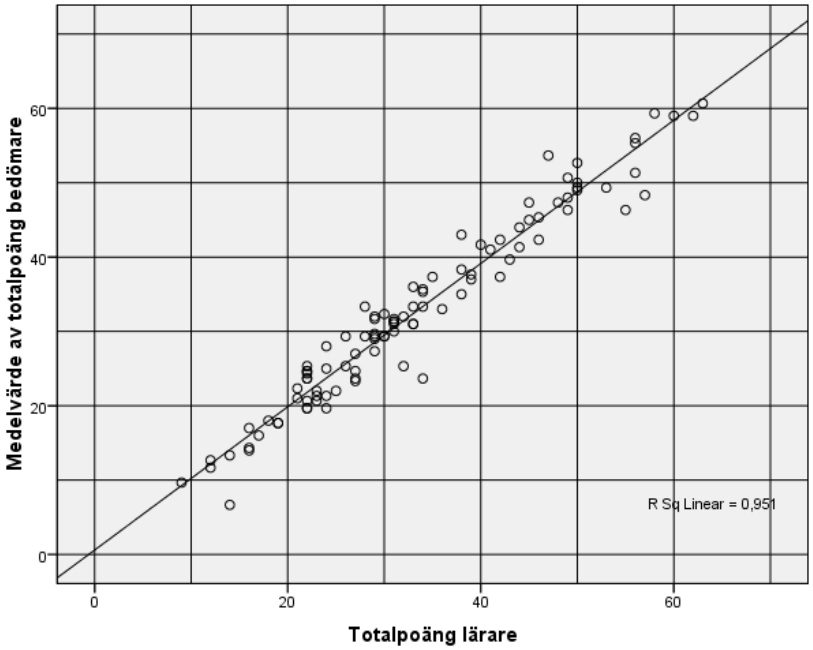
		Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Spearman's rho	Bedömare 1	1,000	**,989****	**,992****	**,965****
	Bedömare 2	,989**	1,000	**,985****	**,961****
	Bedömare 3	,992**	,985**	1,000	**,967****
	Lärare	,965**	,961**	,967**	1,000

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

Korrelationen mellan totalpoängen var mycket hög och lika oavsett mellan vilka par av bedömare som korrelationen beräknades. Även korrelationen mellan läraren och bedömarna var hög om än något lägre än mellan bedömarna. Motsvarande korrelationer beräknades också för g-poängen respektive vg-poängen. Här varierade korrelationerna mellan 0,941 och 0,988. De högsta korrelationerna fanns mellan bedömarna på g-poängen och den lägsta fanns mellan bedömarna och läraren på vg-poängen. Några av vg-poängen faller ut om eleven fullföljt hela uppgiften samt gjort en klar och tydlig redovisning. Här kanske lärarens kännedom om eleven gör att läraren gör en mer positiv tolkning av redovisningens kvalitet.

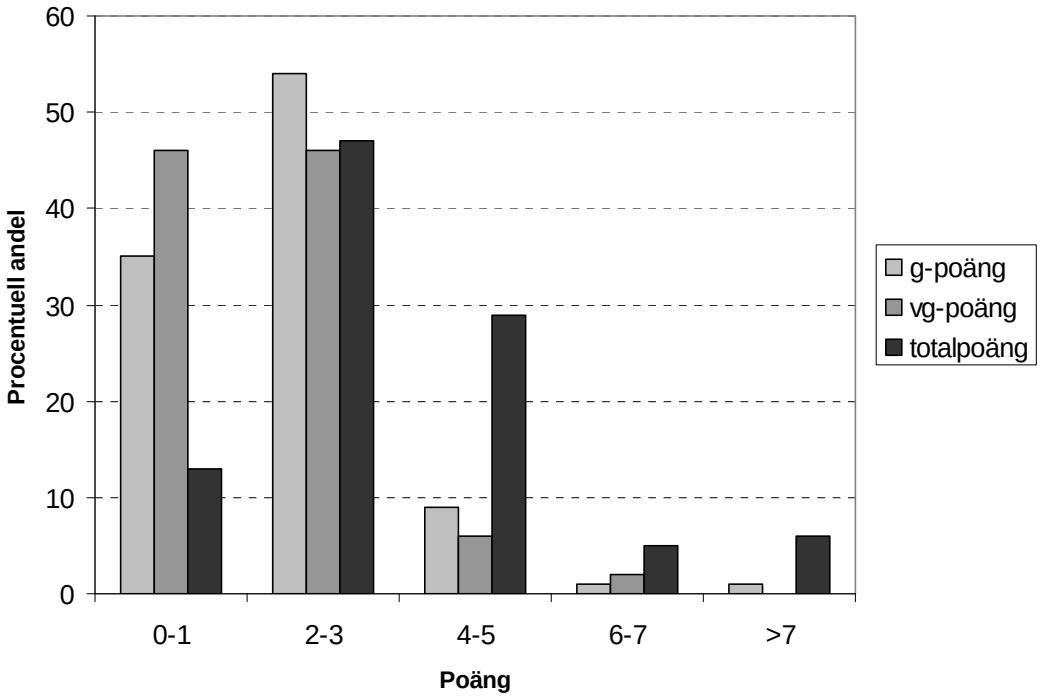
Vid nästa jämförelse (se figur 1) studeras poängsumman för varje elevarbete. Vidare likställs ett resultat med 24 g-poäng och 5 vg-poäng med ett resultat med 28 g-poäng och 1 vg-poäng, dvs antalet poäng betraktas men inte kvaliteten på poängen. I flertalet av elevarbetena är differensen från medelvärdet mindre än 4 poäng. Regressionskoefficienten är ungefär lika med 0,95.

Figur 1 Medelvärdet av bedömarnas totalpoäng som funktion av lärarens totalpoäng



Ett annat sätt att undersöka denna överensstämmelse är att se på variationsbredden.

Figur 2 Variationsbredden mellan totalpoäng (max 67), g-poäng (max 36) och vg-poäng (max 31) för de fyra bedömningarna



Trots att dessa två delprov kunde ge sammanlagt 67 poäng var variationsbredden oftast fem poäng eller mindre. Detta visar på en större överensstämmelse än vid den ombedömning som gjordes 2002 (se sid 15).
I tabell 6 är det den totala poängsumman för varje elevarbete som jämförs. Dessa fyra totalsummor för varje elevs resultat har jämförts och högsta respektive lägsta

värde har markerats. För att en bedömning ska noteras som högsta eller lägsta bedömning måste detta resultat avvika från övriga. Det är alltså endast då bedömaren är ensam om att ha ett högsta eller lägsta värde, som avvikelsen har noterats. I tabellen redovisas också de större avvikelserna från medianvärdet.

Tabell 6 Antal elevarbeten där poängsumman är högst respektive lägst av alla fyra bedömningarna (maximalt 67 poäng)

	Bedömare 1	Bedömare 2	Bedömare 3	Läraren
Högst poängsumma av alla	17	12	14	32
Lägst poängsumma av alla	9	27	10	20
2–3 poäng över median	10	8	5	15
2–3 poäng under median	6	18	5	12
4 poäng eller mer över median	0	0	0	7
4 poäng eller mer under median	0	1	1	4

Av denna framställning framgår att av bedömarna är det bedömare 2 som bedömt strängast och bedömare 3 som bedömt minst strängt. Att lärargruppen har bedömt ett elevarbete högst av alla fler gånger än övriga bedömare är kanske inte förvånande. Dels har bedömningen troligen gjorts av etthundra olika personer, dels har dessa personer en referens till eleven vars arbete de bedömer. Att elevernas lärare ger den lägsta poängtilldelningen till så många elevarbeten är kanske mer förvånande. Haloeffekten, dvs att lärarens tidigare uppfattning om elevens kunskaper påverkar bedömningen, kan innebära att kvaliteterna i elevarbetet både överskattas och underskattas.

Provbetyget Mycket väl godkänt

I de bedömningsanvisningar som medföljer proven redovisas i tabellform vilka MVG-kvaliteter som respektive uppgift erbjuder möjligheter att visa.

De fem olika MVG-kvaliteter som finns beskrivna i tabellen kommer från betygskriterierna för betyget Mycket väl godkänt.

Tabell 7 På de ¤-märkta uppgifterna i ämnesprovet 2007 kunde eleven visa följande MVG-kvaliteter (markerat med ○)

MVG-kvalitet	Uppgift (¤-märkt)					
	Dp A	Del B2	Dp C			
			6a	8c	10	11b
Visar säkerhet i problemlösning och beräkningar		○	○	○	○	○
Formulerar och utvecklar problemet, använder generella strategier vid problemlösningen	○	○				
Tolkar och analyserar resultat, jämför och värderar olika metoders för- och nackdelar	○			○		
Använder matematiska resonemang, tar del av andras argument och för diskussionen framåt	○	○		○		
Redovisar strukturerat med lämpligt/ korrekt matematiskt språk	○	○	○	○		○

På ämnesprovet 2007 kunde eleven visa MVG-kvalitet på sex olika uppgifter. För provbetyget Mycket väl godkänt krävdes att elevarbetet uppfyllde minst sex av kvaliteterna (markerade med ○) och dessa sex var av minst tre olika slag.

Provbetyget Mycket väl godkänt på denna ombedömning är svår att jämföra med den bedömning som läraren gör på det nationella provet. Dels saknas ett delprov (Delprov A) där eleverna kunde visa fyra olika MVG-kvaliteter, dels finns inte lärarnas MVG-bedömning med på alla elevarbeten. De beräknade provbetygen för Mycket väl godkänt utgår därför bara från antalet vg-poäng. På det nationella provet är antalet MVG-kvaliteter mer avgörande än antalet vg-poäng för det högsta betygssteget.

Våra bedömare fick fylla i MVG-tabellen då de bedömde elevarbetena. De gjorde detta för alla elevarbeten som visade någon MVG-kvalitet oavsett vilken poängsumma som elevarbetet uppvisade. Vid diskussionerna med bedömarna visade det sig att de brukar göra på detta sätt då de bedömer nationella prov. Elever som inte når provbetyget MVG blir uppmuntrade av att få veta att de har visat MVG-kvalitet på någon uppgift.

Tabell 8 Sammanställning över beräknade provbetyg som också visade godtagbar MVG-kvalitet

	Antal elevarbeten		
	Bedömare 1	Bedömare 2	Bedömare 3
MVG enligt vg-poäng	8	8	9
MVG enligt MVG-tabell	7	5	12
MVG enligt både vg-poäng och MVG-tabell	4	3	8

Av tabellen framgår att det finns elevarbeten som har tillräckligt många vg-poäng men inte tillräcklig hög kvalitet i lösningarna, men att det också finns det motsatta förhållandet. Denna sammanställning visar också att bedömare 3 använder MVG-

kvaliteterna mest och att bedömare 2 använder dem minst. Bedömare 3 är den som är mest insatt i bedömningsanvisningarna.

På fem elevarbeten fick eleverna det framräknade provbetyget MVG av läraren medan alla bedömarna hade VG. Detta visar att dessa lärare är mer generösa med vg-poäng kanske beroende på att de har kunskaper om elevens matematikkompetenser som inte syns i provet. Vi vet dock inte om läraren har bedömt dessa elevarbeten med Mycket väl godkänt eftersom vi inte har haft tillgång till lärarens bedömning av MVG-kvaliteter.

Resultat på uppgiftsnivå

I de analyser som redovisat hittills studeras resultatet ur ett elevperspektiv. Det är också intressant för en provkonstruktör att analysera resultatet på uppgiftsnivå. Vilka uppgiftstyper och vilken typ av bedömningsanvisningar visar på störst överensstämmelse vid bedömningen?

Tabell 9 Korrelation mellan bedömares och lärares totalpoäng på Del B1: kortsvar

		Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Spearman's rho	Bedömare 1	1,000	**,992****	**,972****	**,987****
	Bedömare 2	,992**	1,000	**,976****	**,991****
	Bedömare 3	,972**	,976**	1,000	**,975****
	Lärare	,987**	,991**	,975**	1,000

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

På detta delprov finns det bara flervalsuppgifter och kortsvarsuppgifter. Endast svaret ska bedömas. Att korrelationen inte är 1,000 beror troligen på avskrivningsfel vid bokföringen i kalkylbladet.

Tabell 10 Korrelation mellan bedömares och lärares totalpoäng på Del B2: mer omfattande uppgift

		Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Spearman's rho	Bedömare 1	1,000	**,913****	**,905****	**,749****
	Bedömare 2	,913**	1,000	**,945****	**,727****
	Bedömare 3	,905**	,945**	1,000	**,723****
	Lärare	,749**	,727**	,723**	1,000

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

Även på detta delprov är korrelationen mycket hög mellan bedömarna men den är något lägre då man jämför med lärarens bedömning. Mellan bedömarna är variationsbredden aldrig mer än 2. På 42 elevarbeten är bedömarna helt överens, på 51 elevarbeten avviker en av bedömarna med 1 poäng och på 7 elevarbeten är variationsbredden 2 poäng.

Tabell 11 Variationsbredden mellan de fyra bedömningarna av 100 elevarbeten till Del B2 (max 10 poäng)

Variationsbredd	0	1	2	3	4
Antal elevarbeten	19	48	25	6	2

De flesta lärarbedömningarna följer också de bedömningar som bedömarna gjort. Det är dock åtta lärare som avviker med fler än 2 poäng.

Tabell 12 Korrelation mellan bedömares och lärares totalpoäng på Delprov C: blandad problemlösning

		Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Spearman's rho	Bedömare 1	1,000	**,980****	**,983****	**,953****
	Bedömare 2	,980**	1,000	**,979****	**,952****
	Bedömare 3	,983**	,979**	1,000	**,958****
	Lärare	,953**	,952**	,958**	1,000

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

Även på detta delprov är korrelationen hög om man jämför totalpoängen på delprovet. På några uppgifter i Delprov C var skillnaderna mellan bedömarna något större än på andra uppgifter. Detta gällde framför allt uppgift 4a. Här är korrelationen inte lika hög men fortfarande mycket god.

Tabell 13 Korrelation mellan bedömares och lärares totalpoäng på uppgift 4a på Delprov C

		Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Spearman's rho	Bedömare 1	1,000	**,895****	**,946****	**,794****
	Bedömare 2	,895**	1,000	**,903****	**,828****
	Bedömare 3	,946**	,903**	1,000	**,816****
	Lärare	,794**	,828**	,816**	1,000

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

Uppgift 4a prövade elevernas kunskaper om begreppen medelvärde och median.

4. På Nya Zeeland kan man vaska guld. En dag var det 12 personer som vaskade. Efter en timme vägde de hur mycket guld de hade lyckats vaska per person. Resultatet ser du i tabellen.

0,15 g	2,96 g	0,23 g	0,62 g	0,43 g	0,36 g
0,16 g	0,28 g	0,32 g	0,19 g	0,26 g	0,30 g

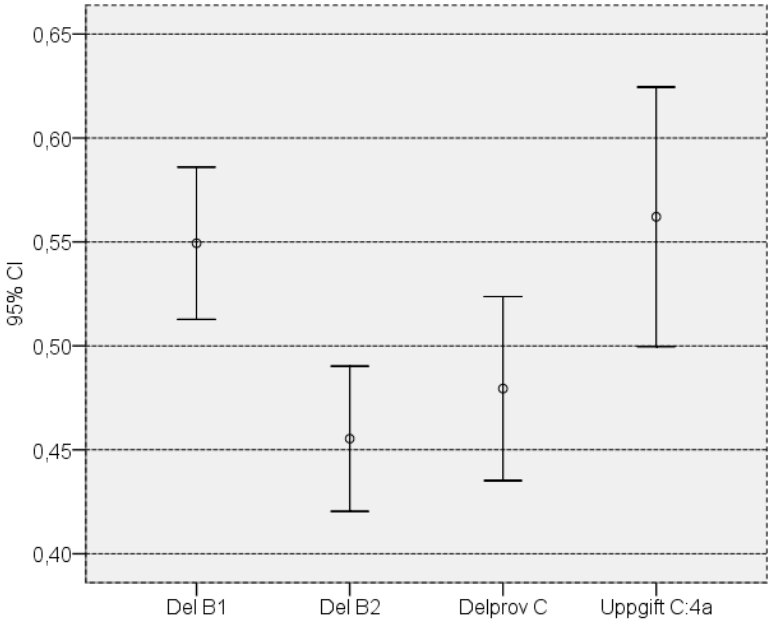
a) Bestäm medelvärde och median för hur mycket guld de vaskade på en timme. (2/1)

Bedömningsanvisningarna såg ut på följande sätt.

4. a)	**Medelvärde: 0,52 g Median: 0,29 g**	**(Max 2/1)**
	Redovisar lämplig metod för beräkning av medelvärdet	+ 1g
	Redovisar lämplig metod för beräkning av medianen	+ 1g
	Klar och tydlig redovisning med korrekta svar	+ 1 vg

En förklaring till att uppgiften har den sämsta korrelationen av uppgifterna i Delprov C är troligtvis formuleringen ”lämplig metod” som lämnar ett visst tolkningsutrymme till läraren. Dessutom har olika lärare olika krav på vad en klar och tydlig redovisning innebär. Detta såg vi i den undersökning som gjordes 2002 (Olofsson 2006) och detta framkom också i diskussionerna med bedömarna i undersökningen 2008.

Figur 4 Lösningsproportioner med felmarginal (95 % konfidensintervall)



För varje bedömare summerade vi poängen för de 100 eleverna per uppgift och per delprov. Därefter beräknade vi lösningsproportionen per uppgift/delprov för varje bedömare genom att dela med uppgiftens maxpoäng och 100. Uppgift 4a på Delprov C var också den uppgift som hade störst spridning av alla uppgifter på provet.

Figur 4 visar att felmarginalen var lika stor för de tre skriftliga delarna i provet. Både Del B1 och Delprov C består av många olika uppgifter. Del B2 består enbart av en uppgift. Felmarginalen för uppgift 4a i Delprov C som kunde ge 2 g-poäng och 1 vg-poäng var större än uppgiften i Del B2 som kunde ge 4 g-poäng och 6 vg-poäng.

Lärares interbedömarreliabilitet vid två olika undersökningar

Bakgrund

2002 genomförde PRIM-gruppen en liknande ombedömning av det nationella ämnesprovet i matematik för årskurs 9. Då undersöktes särskilt om det fanns någon skillnad mellan överensstämmelsen mellan poängbedömning av mindre uppgifter och aspektbedömning av öppna, ”mer omfattande” uppgifter. För att ta reda på detta lät vi fyra matematiklärare bedöma etthundra elevarbeten från ämnesprovet för årskurs 9 som gavs 2001. Gemensamt för dessa lärare var att de deltog med sina egna elever då proven gavs och att de därför var relativt väl insatta i bedömningsanvisningarna. Bedömarna valdes ut så att vi fick bedömare av olika kön, med olika lång erfarenhet som lärare och undervisning i olika typer av skolor. En motsvarande ombedömning av det nationella provet för kurs A finns beskriven i en rapport av Gunilla Olofsson, Likvärdig bedömning?, från 2006.

Urval av uppgifter

Först valde vi ut de uppgifter som skulle bedömas. Från ämnesprovet för årskurs 9 valde vi två aspektbedömda uppgifter, Delprov A och Del B2, samt några uppgifter från Delprov C. Vi valde båda de aspektbedömda uppgifterna eftersom ett av syftena med undersökningen var att undersöka bedömaröverensstämmelsen för denna typ av bedömning. Vid urvalet av uppgifter från Delprov C kategoriserade vi bedömningsanvisningarna till dessa från vad vi ansåg vara lättbedömda till mer svårbedömda och valde uppgifter från båda dessa kategorier. Dessutom tog vi med uppgifter där elevarbetet kunde visa MVG-kvaliteter. Ett tredje krav var att summan av poängen från Delprov C skulle vara densamma som poängen för en aspektbedömd uppgift. Vi kunde då jämföra bedömaröverensstämmelsen mellan aspektbedömning och bedömning med vanlig bedömningsanvisning med poäng.

Urval av elevarbeten

Vi valde slumpvis ut etthundra elevarbeten. Ett krav på elevarbetet var dock att alla delprov fanns med och att vi kunde avgöra, antingen från elevarbetet eller från våra datafiler, hur läraren bedömt elevarbetet.

Ombedömning

Ombedömningen genomfördes på samma sätt som årets ombedömning. De fyra matematiklärarna bedömde var och en för sig alla elevarbeten efter de bedömningsanvisningar som medföljde provet och bokförde resultaten på ett kalkylblad. Resultaten av denna bedömning bearbetades och vi undersökte vilka avvikelser det fanns i lärarnas bedömning.

Efter detta träffades matematiklärarna och diskuterade de olikheter i bedömningen som hade framkommit. Bedömarna fick, om de ville, ändra sin bedömning efter det gemensamma mötet.

Erfarenheter från mötet med ombedömargruppen för Äp 9

Orsaker till olikheter i bedömningen kan delas upp i tre grupper:

- Olika tolkning av uppgiften vilket leder till att man tolkar elevarbetet olika
- Olikheter i analysen av elevarbetet som medför att lärare kan fokusera på olika saker som förståelse, räknefel eller svar
- Olika tolkning av bedömningsanvisningarna

Skrivningar i bedömningsanvisningarna som kan leda till olika tolkning: Ansats till lösning t ex…, kan få läraren att tro att det som står som exempel är det enda rätta sättet att inleda uppgiftens lösning. Det är svårt att avgöra vad som menas med acceptabelt matematiskt språk och med lämplig metod.

Fel i inledningen av lösningen men sedan rätt (s.k. följdfel) ger inte alltid poäng för alla lärare. Felaktiga beräkningar eller missbruk av likhetstecken gör att läraren ibland dömer ut lösningen helt även om eleven till synes har förstått uppgiften.

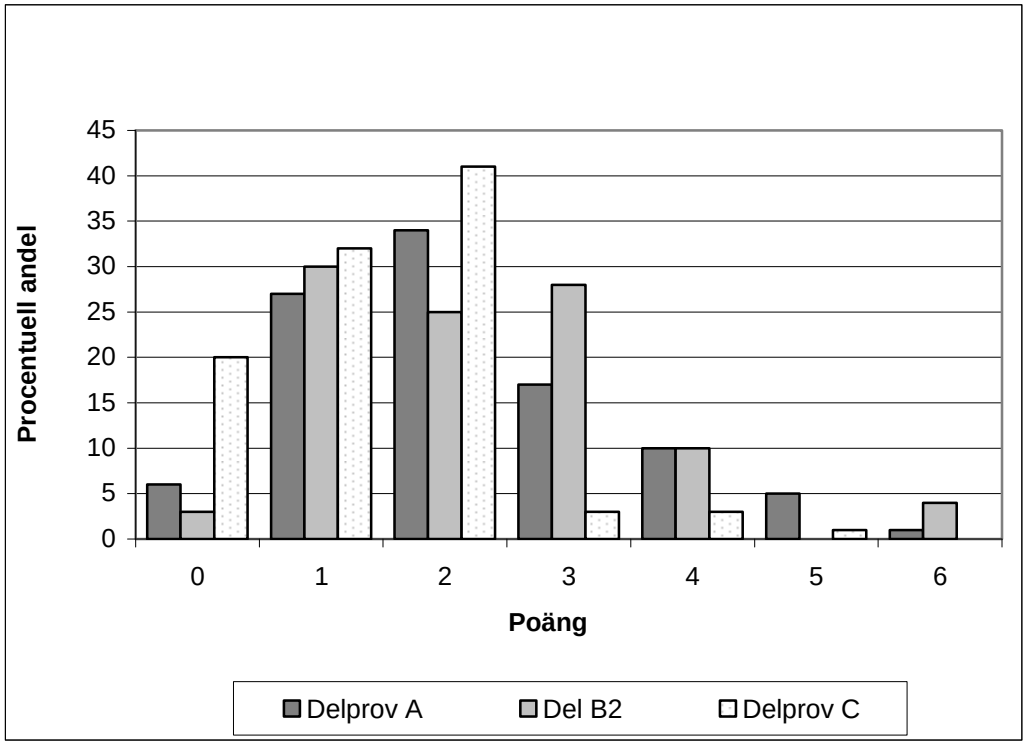
Elevarbeten som publiceras som stöd till bedömningsanvisningarna bör vara typiska och inte ”udda”. Det är ibland svårt att avgöra skillnaden mellan den sista vg-poängen och MVG-kvalitet.

Erfarenheterna från ombedömningen 2002 har gjort att vi numera har tydligare bedömningsanvisningar, fler bedömda autentiska elevarbeten och en MVG-tabell för att stödja bedömningen av MVG.

Några resultat från ombedömningen 2002

Vid jämförelser med den ordinarie lärarens bedömning användes bedömarnas reviderade bedömning. Olikheter i bedömningen är ungefär lika stora på summan av ett antal kortare uppgifter som på en aspektbedömd uppgift.

Figur 5 Variationsbredden mellan de fem bedömarna

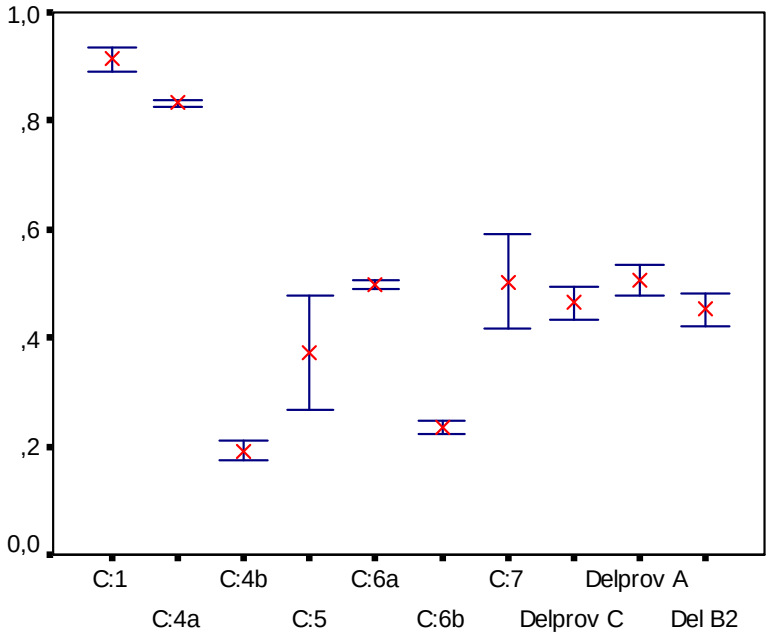


Vid ombedömningen av Äp 9 2002 var variationsbredden något större för de två aspektbedömda uppgifterna (Delprov A och Del B2) än för summapoängen av fem uppgifter i Delprov C.

Delprov A och Del B2 kunde vardera ge 6 g-poäng och 7 vg-poäng vilket var detsamma som uppgifterna från Delprov C kunde ge tillsammans. För varje bedömare summerade vi poängen för de 100 eleverna per uppgift och per delprov. Därefter beräknade vi lösningsproportionen per uppgift/delprov för varje bedömare genom att dela med poängantalet och 100.

Figur 6 ger en annan bild av samstämmigheten vid bedömningen av olika typer av uppgifter. Här är spridningen i bedömningen av summan av de mindre uppgifterna på Delprov C lika stor som spridningen på Delprov A och Del B2, som var de båda aspektbedömda uppgifterna. Den största spridningen visar två mindre uppgifter i Delprov C.

Figur 6 Lösningsproportioner med felmarginaler (95 % konfidensintervall)



Sammanfattande kommentarer

Överensstämmelsen mellan bedömarna är mycket god (korrelationen mellan bedömarnas totalpoäng är 0,99) i denna undersökning av Äp 9 2007. En slutsats är att bedömningsanvisningarna som medföljer proven är tydliga och lätta att tolka. En annan förklaring kan vara att bedömarna har lång erfarenhet och att de varit med i referensgruppen/kravgränsmöten samt deltagit i bedömningskurser.

Våra tre bedömare undervisar alla på skolor i Stockholmsområdet som har ett liknande elevunderlag. Detta gör troligtvis att de har ungefär samma referensramar då de bedömer. Bedömare 2 är något strängare i sin bedömning medan Bedömare 3 är den som är minst sträng. Bedömare 3 har varit med och arbetat fram bedömningsanvisningarna och är därför mest insatt i tankarna bakom dessa.

Våra bedömare har dock arbetat under förhållanden som är ovanliga för dem. När de normalt bedömer nationella prov brukar de diskutera sin bedömning med sina kollegor. Ombedömningen från 2002 av både Äp 9 och Kurs A visade att diskussioner med kollegor under bedömningsarbetet ger en mycket hög bedömaröverensstämmelse (Olofsson 2006).

De beräknade gränserna för ”provbetygen” i undersökningen av Äp 9 2007 kan inte jämställas med de riktiga provbetygen. Elevernas möjlighet att muntligt visa matematiska kunskaper finns inte med i denna undersökning. De beräknade provbetygen kan dock ge en indikation på bedömaröverensstämmelsen. Våra bedömare arbetade dessutom utan att känna till några kravgränser för provbetygen vilket de annars gör. Eftersom denna undersökning bara omfattar Delprov B och Delprov C stämde inte kravgränserna för Äp 9 2007 som helhet.

Merparten av de 100 lärarna är överens med bedömarna om antalet totalpoäng på respektive elevarbete. Endast ett fåtal avviker med flera poäng. Även för de beräknade provbetygen är överensstämmelsen mycket god. Bedömarna är överens om 92 % av de beräknade provbetygen och de är också överens med läraren i 86 % av

elevarbetena. Skillnaden i provbetyg beror i de flesta fallen på en avvikelse på endast en eller två poäng av 67 möjliga.

Överensstämmelsen mellan bedömarna var god för det beräknade provbetyget Mycket väl godkänt men sämre då hänsyn togs till antalet MVG-kvaliteter. Vid bedömning av provbetyget Mycket väl godkänt brukar lärare noga gå igenom MVG-kvaliteterna för de elever som når poänggränsen för MVG. Bedömarna kände inte till den gränsen. Undersökningar av tidigare ämnesprov visar att en större andel elever når poänggränsen än den andel som också har visat tillräckligt djup för MVG-kvalitet.

Jämförelsen mellan ombedömningarna 2002 och 2008 visar att överensstämmelsen mellan bedömarna har blivit större för de aspektbedömda uppgifterna. Detta ser man framför allt på att variationsbredden i poäng har minskat. En förklaring till detta är att bedömningsmatriserna blivit tydligare och att fler bedömda autentiska elevarbeten ingår i bedömningsanvisningarna. Det finns även stöd för detta i lärarenkäten från 2007.

Referenser

Hallén S. & Kjellström K. (2007). Matematik. *Ämnesprovet 2007 i grundskolans årskurs 9*. Skolverket

Kjellström K. & Olofsson G. (2004) *Pålitligheten i lärares bedömning av nationella prov – presentation av en undersökning.* Matematikbiennalen 2004. Malmö Högskola

Olofsson, G. (2006). *Likvärdig bedömning?* En studie av lärares bedömning av elevarbeten på ett nationellt prov i matematik kurs A. Stockholm: Lärarhögskolan i Stockholm: PRIM-gruppen.