Bedömaröverensstämmelse för ett nationellt prov i matematik C

Peter Nyström
Anna Lind Pantzare

Under den senaste tioårsperioden har fyra svenska studier av bedömaröverensstämmelse i de nationella matematikproven presenterats (Boesen, 2004; Lindström, 1998; Olofsson, 2006; Palm, 2008). Den första studien (Lindström, 1998) byggde på ombedömning av 58 elevarbeten från det nationella kursprovet för matematik E våren 1997. En bedömningsgrupp engagerades bestående av tre erfarna lärare från en och samma skola. Varje elevarbete bedömdes av en lärare och därefter sammanträdde granskningsgruppens medlemmar för att diskutera tveksamma fall och slutföra bedömningen, ett förfarande som avsåg att efterlikna det normala rättningsförfarandet på skolan. Bedömningsgruppens poängsättning jämfördes med den poäng som elevens lärare angivit på de inskickade elevarbetena. Resultaten visar på hög korrelation mellan det två bedömningarna (0,99) och att den ursprungliga poängbedömningen är densamma som ombedömningen i cirka en tredjedel av elevarbetena (34 %). I 12 % av fallen ger ombedömningen en eller två poäng mer än den ursprungliga bedömningen. För övriga (drygt hälften) av elevarbeten ger ombedömningen en lägre poäng än den ursprungliga bedömningen. Cirka 23 % har 1-2 poäng lägre och cirka 12 % har 3-5 poäng lägre.

Några år senare gjordes en liknande studie (Boesen, 2004) men denna gång handlade det om ett nationellt provet i matematik B. Syftet med studien var att undersöka bedömarreliabiliteten vid de nationella proven i matematik, med särskilt fokus på den så kallade aspektbedömda uppgiften i den nationella proven i matematik. Aspektbedömning innebär att elevens svar värderas utifrån tre olika perspektiv, med fokus på olika kompetenser. En sådan bedömning kräver en uppgift som är mer omfattande och komplex. Metoden liknar den som användes i Lindströms studie (Lindström, 1998), med fyra erfarna lärare vid en gymnasieskola som bedömer 60 slumpmässigt valda elevarbeten bland de elevarbeten som skickats in från olika skolor i samband med datainsamling avseende nationella prov. Till skillnad från Lindströms studie granskades varje elevarbete av minst två av lärarna men avsikten var även här att komma fram till en konsensusbedömning av varje arbete. Poängsättningen som dessa lärare kom fram till jämfördes därefter med den ursprungliga poängsättning som den lärare gjort som skickat in elevarbetet. I Boesens studie görs ingen sammanfattande bedömning i form av t.ex. korrelationer av hur väl den externa bedömningen överensstämmer med de bedömningar som elevernas lärare gjort. Resultaten visar dock att av de 60 elevarbeten som granskats har 7 bedömts med samma poäng i båda bedömningarna (12 %). Vid ombedömningen har 18 elevarbeten (30 %) fått lägre poäng än vid den första bedömningen och övriga 35 högre (58 %).

Såväl Lindströms som Boesens undersökningar har sitt starkaste fokus på hur olika uppgifter i proven bidrar till variationen i bedömningen. Denna analys är relativt detaljerad med exempel på både uppgifter och elevsvar. Lindström konstaterar att variationen mellan bedömarna var relativt liten, trots att bedömningsanvisningarna kan karakteriseras som relativt oprecisa eller öppna. Till många uppgifter i detta prov gavs bedömningsanvisningen ”Redovisad godtagbar lösning + 1-2 poäng”, vilket skulle kunna ge utrymme för olika tolkningar. Boesen konstaterar att det inte är den aspektbedömda uppgiften som uppvisat störst skillnader i bedömningen (trots att den komplexa och mångfasetterade bedömningen borde kunna ge upphov till variation i bedömningen), och han menar att en möjlig förklaring är att det inte är bedömningsformen i sig som leder till svårigheter. Gemensamt för de uppgifter som uppvisat störst skillnad mellan olika bedömare är att de alla i varierande grad kräver att eleven ska ge en förklaring och argumentera för sin lösning.

Olofssons (2006) studie av bedömaröverensstämmelse utgick från 100 elevarbeten, slumpvis valda bland arbeten som skickades in från landets skolor i samband med resultatrapporteringen för kursprovet i matematik A våren 2001. Ombedömningen gjordes av fyra erfarna lärare från olika gymnasieskolor, och den genomfördes individuellt. Efter den individuella bedömningen samlades bedömarna för att diskutera olikheter och gavs därefter möjlighet att korrigera sin poängsättning. Olofsson analyserar bedömarnas poängsättning såväl före som efter denna korrigering.

Olofsson konstaterar sammanfattningsvis att bedömaröverensstämmelsen var god för drygt 80 procent av de studerade elevarbetena, det skiljer endast något eller några enstaka poäng mellan bedömningarna. En jämförelse av medianen för de fyra externa bedömarnas poäng och den poäng som elevens lärare rapporterat visar på en korrelation på 0,93. Det prov som studerades har liksom dagens nationella prov i matematik uppgifter som bedöms utifrån relativt traditionella bedömningsanvisningar, men även en uppgift som ska bedömas med så kallad aspektbedömning. Olofsson konstaterar i enlighet med Boesen (2004) att aspektbedömningen inte tycks ha sämre bedömaröverensstämmelse än den mindre komplexa bedömningsmodell som används för andra uppgifter. Tre till fyra av de mindre uppgifterna i provet hon studerat kan ge motsvarande eller större variation i bedömningen och påverka slutpoängen i lika hög eller högre grad. Olofsson konstaterar att det i hennes undersökning inte är bedömningsmetoden utan mer uppgifternas utformning som skapar osäkerhet. Uppgifter där eleverna ska göra tolkningar eller ge förklaringar gör bedömning och avgörande om vilka svar som är godtagbara mer komplicerad jämfört med uppgifter där eleverna mer rakt på sak ska göra en beräkning. Detta kan enligt Olofsson till exempel bero på att det är svårare att både beskriva och upptäcka sådana kvalitetsskillnader, eller att matematiklärare genom sin utbildning och erfarenhet har skaffat sig kompetens som gör dem bättre på att bedöma kvaliteten i beräkningsuppgifter.

Den fjärde svenska studie av bedömarreliabilitet som presenteras här gäller interbedömarreliabiliteten i ett muntligt prov som utvecklats för gymnasieskolans matematik C (Palm, 2008). Palm konstaterar att muntliga prov av ämneskunskap eller kommunikationsförmåga i allmänhet är behäftade med reliabilitetsproblem, vilket orsakar svårigheter med dess användning. Reliabilitetsstudier om muntliga prov i matematik är dock relativt ovanliga. I Palms studie fick en bedömargrupp på tio lärare lyssna på sex elevers muntliga presentation av sina lösningar till en matematikuppgift. Bedömningen skulle utgå från de generella bedömningsanvisningar som utvecklats för att bedöma såväl kommunikation i matematik som matematisk begreppsförståelse. Resultaten visar på låg reliabilitet (något mer än 50 % överensstämmelse i genomsnitt), även om Palm konstaterar att det inte verkar vara kommunikationsaspekten i sig själv som gör att denna förmåga var svår att bedöma, utan snarare otillräckligheter i den bedömningsmodell som användes. Bedömningsmodellen har senare utvecklats vidare, men det material som nu finns tillgängligt för lärarna har inte studerats utifrån interbedömarreliabiliteten.

Sammanfattningsvis har flera studier av nationella prov för olika kurser i gymnasieskolans matematik visat på en relativt god bedömaröverensstämmelse. Föga överraskande är det dock så att ju mer av tolkningar och förklaringar som eleverna förväntas göra vid lösandet av provuppgifterna och ju mer bedömningen gäller sådant som redovisningens kvalitet, desto större blir variationerna mellan bedömarna. Att detta i ännu högre grad gäller ett muntligt prov som har en större komplexitet är inte heller anmärkningsvärt, men det bör påpekas att den studie som presenterats gäller en modell för bedömning av muntlighet i matematik som sedan utvecklats vidare.

Den allra vanligaste utgångspunkten i studier av bedömaröverenstämmelse är att söka svaret på frågan ”hur olika är bedömningsresultaten från olika bedömare?”. Det är hot av detta slag mot tolkningen av provresultat som i allmänhet avses med termen interbedömarreliabilitet. Bedömares överensstämmelse i absoluta termer kan mätas med en enkel parvis procentuell överensstämmelse, dvs. en beräkning av hur stor andel av bedömningarna som är lika mellan två bedömare. Hur konsekventa lärare är i sin bedömning kan studeras med korrelationer. Observera att en låg korrelation betyder att bedömarna i hög grad är oense om bedömningarna. En hög korrelation behöver inte betyda att två bedömare är överens om enskilda bedömningar, utan visar primärt bara att de två bedömarna har ungefär samma rangordning av de bedömda elevsvaren.

I denna studie av bedömaröverensstämmelse används såväl procentuell överensstämmelse som korrelationer för att söka svaret på frågan om hur olika lärare är i sina bedömningar. För detta syfte är det nödvändigt att låta olika bedömare bedöma ett relativt stort antal elevsvar. I denna studie är det i första hand de tre externa

bedömarna som kan användas. Varje elevsvar har dessutom en bedömning som är gjord av någon lärare på elevens skola. Dessa lärare har ju i de allra flesta fall endast bedömt ett enda elevsvar och kan därför inte användas för interbedömarreliabilitetsstudier, eftersom fokus är på hur olika enskilda bedömare bedömer. Det kan dock vara intressant att betrakta alla dessa bedömningar som om de var gjorda av en lärare, för att undersöka hur denna elevnära bedömning förhåller sig till de övriga (externa) bedömarna. Jämförelsen bör dock göras med försiktighet.

I vår studie har vi i första hand analyserat provbetygen, men vi har även studerat olikheter mellan bedömarna när det gäller provpoäng (totalpoäng) och även poäng på olika uppgifter. Det sistnämnda syftar i första hand till att identifiera uppgifter som tycks särskilt svårbedömda, och eventuellt komma fram till någon tänkbar förklaring till detta utifrån bland annat bedömningsanvisningen.

Ett kompletterande perspektiv på bedömaröverensstämmelse är att ställa sig frågan ”Hur olika blir en elev bedömd?”. Observera att detta är en helt annan fråga som också kräver andra utgångspunkter, metoder och mått. För att besvara den frågan är det nödvändigt att låta ett och samma elevsvar bli bedömt av flera bedömare. Förekomsten av skillnader i bedömningen kan förklaras med att ett visst elevsvar är särskilt svårbedömt eller med att lärarna gör olika värderingar av samma svar.

Det är mindre självklart vilka mått som kan och bör användas här. Ett enkelt mått är variationsbredd, där maximala skillnader mellan bedömarna beräknas för varje elevsvar. Det är också möjligt att bestämma hur eniga de fyra bedömarna var på varje elevsvar, vilket innebär att vi kan säga om alla fyra gav samma provbetyg, om en av bedömarna avviker från de övriga tre, eller om bedömarna är parvis överens.

Det analyserade provet i Matematik C

Det nationella provet i matematik C från våren 2007 består av 16 uppgifter. Vissa av dessa uppgifter består av en eller flera deluppgifter vilket innebär att totalt 26 deluppgifter har analyserats. Deluppgifterna kan ge mellan en till tre poäng. I matematikproven finns g- och vg-poäng, vilka kan kopplas till betygskriterierna. Dessutom finns ett antal uppgifter som är markerade med ¤, vilka ger en större möjlighet än övriga uppgifter att visa på kvaliteter knutna till betygskriterierna för MVG.

Provet består av två delar med åtta uppgifter i varje del. Vid lösandet av uppgifterna i Del 1 är det inte tillåtet att använda miniräknare. Miniräknare är däremot tillåten i Del 2, dessutom får eleverna använda formelblad under hela provet. Eleverna har totalt 240 minuter till sitt förfogande när de ska lösa uppgifterna, den rekommenderade provtiden för del 1 är 60 minuter. Del 2 avslutas med en mer omfattande uppgift som är aspektbedömd. De aspektbedömda uppgifterna är normalt något mer komplicerade att bedöma än vanliga uppgifter. Dels består uppgiften av flera delar eller punkter som eleverna ska göra, dels ska bedömningen av uppgiften ske

med avseende på tre aspekter, *Metodval och genomförande*, *Matematiska resonemang* samt *Redovisning och matematiskt språk*." I det aktuella provet har varje aspekt maximalt två kvalitativa nivåer. Tanken är att om eleven når den kvalitativt högre nivån ska underliggande nivå också anses vara uppfylld.

Maxpoängen på provet är 42 poäng fördelat på 22 g- och 20 vg-poäng. Kravgränserna är 12 poäng för Godkänd, 24 poäng varav 6 vg-poäng för Väl godkänd. För provbetyget Mycket väl godkänd ska eleven också ha minst 24 poäng, men här krävs att minst 13 av dessa ska vara vg-poäng. Dessutom ska eleven ha visat minst tre av de MVG-kvaliteter som definierats i bedömningsanvisningen.

Av de 26 analyserade deluppgifterna är det sju uppgifter där endast svar fordras. I övriga uppgifter krävs att eleven redovisar sin lösning. Uppgifterna är en blandning av inommatematiska uppgifter och problemlösningsuppgifter med en kontext.

Urval och andra val

De elevsvar som är bedömda kommer från den insamling av elevlösningar som sker varje termin i samband med inrapporteringen av resultat till de nationella proven. Denna insamling sköts för C-kursprovet av Institutionen för beteendevetenskapliga mätningar och är frivillig. Elevlösningarna som skickas in är lösningar från elever födda ett givet datum. Tanken är att det ska ge ett slumpmässigt urval av elevsvar.

Det nationella provet för matematik C från våren 2007 gavs i två versioner, framförallt för att minska risken för fusk. De insamlade elevsvaren finns alltså i två versioner, och för att förenkla arbetet för bedömarna (och minska risken för skillnader som beror på att bedömaren använt fel bedömningsanvisning till någon uppgift) så beslutades att endast använda en version i denna studie. Alla elevlösningar som ingår i studien är från version 2 av provet. För denna studie skulle 100 elevlösningar väljas ut, och eftersom vi höll oss till en version så fick vi i princip ta alla tillgängliga elevsvar. Det fåtal elevsvar som inte använts i studien var inte möjliga att använda på grund av att kopiornas kvalitet är så dålig att det inte är möjligt att göra en rättvis ombedömning av proven. Det är helt enkelt inte möjligt att se vad eleverna skrivit. Det finns även en del av de utvalda elevlösningarna som är av relativt dålig kvalitet och det är ibland svårt att se om det är läraren som har gjort vissa kommentarer eller om det är eleven som skrivit allt själv. Bland de 100 elevsvar som valdes råkade av misstag ett elevsvar som hörde till version 1 av provet komma med. Detta elevsvar togs bort ur analysen och därför bygger studien på 99 elevsvar från det nationella provet i matematik C våren 2007.

Tre bedömare har genomfört ombedömningen i denna studie. Bedömare 1 är en erfaren lärare som sedan flera år arbetar som provutvecklare vid Institutionen för beteendevetenskapliga mätningar. Bedömare 2 är en erfaren lärare som dessutom

har mångårig erfarenhet som granskare i referensgrupper, dvs. de grupper sammansatta av lärare som granskar de nationella proven. Bedömare tre är en erfaren lärare som aldrig varit inblandad i något arbete med granskning av de nationella proven. Två av bedömarna är män och en är kvinna.

Elevlösningarna avkodades genom att elevernas namn och de ursprungliga bedömningarna raderades. Elevlösningarna skickades ut tillsammans med provet och bedömningsanvisningarna till dem som skulle genomföra ombedömningen. Uppmaningen var att de, i så stor utsträckning som möjligt, skulle rätta på samma sätt som de normalt rättar de nationella proven. I rättningsarbetet ute på skolorna förekommer olika former av samarbete, men här instruerades bedömarna att arbeta enskilt. Däremot uppmanades lärarna att för övrigt rätta på det sätt de normalt gör, t.ex. uppgift för uppgift eller elev för elev. Lärarna hade nästan fyra veckor på sig att genomföra uppdraget.

Resultat

Procentuell överensstämmelse på provbetyg

I jämförelsen mellan de provbetyg som bedömningarna resulterat i ställs de olika bedömarna mot varandra parvis i korstabeller (Tabell 1-6). I tillägg till bedömare 1-3 analyseras här även läraren som inte är en person utan den sammanslagna bedömningen från alla de lärare som gjort den ursprungliga bedömningen av elevsvaren.

Tabell 1 Korstabell som visar överensstämmelse mellan provbetyg från bedömare 1 och 2.

Provbetyg bedömare 1	Provbetyg bedömare 2				
	IG	G	VG	MVG	Totalt
IG	30	0	0	0	30
G	3	43	1	0	47
VG	0	2	12	0	14
MVG	0	0	4	4	8
Totalt	33	45	17	4	99

Tabell 2 Korstabell som visar överensstämmelse mellan provbetyg från bedömare 1 och 3.

		Bed3betyg				Total
		IG	G	VG	MVG	IG
Bed1betyg	IG	24	6	0	0	30
	G	4	41	2	0	47
	VG	0	2	11	1	14
	MVG	0	0	3	5	8
Total		28	49	16	6	99

Tabell 3 Korstabell som visar överensstämmelse mellan provbetyg från bedömare 1 och läraren.

		Lärarebetyg				Total
		IG	G	VG	MVG	IG
Bed1betyg	IG	23	7	0	0	30
	G	1	45	1	0	47
	VG	0	3	11	0	14
	MVG	0	0	3	5	8
Total		24	55	15	5	99

Tabell 4 Korstabell som visar överensstämmelse mellan provbetyg från bedömare 2 och 3.

		Bed3betyg				Total
		IG	G	VG	MVG	IG
Bed2betyg	IG	25	8	0	0	33
	G	3	40	2	0	45
	VG	0	1	14	2	17
	MVG	0	0	0	4	4
Total		28	49	16	6	99

Tabell 5 Korstabell som visar överensstämmelse mellan provbetyg från bedömare 2 och läraren.

		Lärarebetyg				Total
		IG	G	VG	MVG	IG
Bed2betyg	IG	24	9	0	0	33
	G	0	43	2	0	45
	VG	0	3	13	1	17
	MVG	0	0	0	4	4
Total		24	55	15	5	99

Tabell 6 Korstabell som visar överensstämmelse mellan provbetyg från bedömare 3 och läraren.

		Lärarebetyg				Total
		IG	G	VG	MVG	IG
Bed3betyg	IG	21	7	0	0	28
	G	3	44	2	0	49
	VG	0	4	11	1	16
	MVG	0	0	2	4	6
Total		24	55	15	5	99

I Tabell 7 redovisas procentuell överensstämmelse mellan de olika bedömarna.

Tabell 7 Provbetyg, procentuell överensstämmelse mellan bedömare

	Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Bedömare 1	100 %	90 %	82 %	85 %
Bedömare 2	90 %	100 %	84 %	85 %
Bedömare 3	82 %	84 %	100 %	81 %
Lärare	85 %	85 %	81 %	100 %

Den procentuella överensstämmelsen är högst mellan bedömare 1 och bedömare 2, dvs. mellan provutvecklaren och referensgruppsläraren. För övrigt är det relativt jämnt på en något lägre nivå. Paret med provutvecklaren och referensgruppsläraren verkar ha en bättre samsyn vad gäller bedömningen av elevsvar än övriga par av bedömare.

Procentuell överensstämmelse på provpoäng

För att kunna jämföra hur överens bedömarna är när det gäller provpoängen så är det nödvändigt att göra en breddning av kategorierna, dvs. tillåta att bedömarna har små avvikelser från varandra. Om man har väldigt många kategorier (i detta fall 42 eftersom det finns 42 poäng att dela ut i bedömningen) så är det inte rimligt att förvänta sig att bedömarna är absolut överens, utan det räcker om de är nästan överens. Om två bedömare tillåts skilja sig åt med upp till två poäng så blir resultatet en överensstämmelse mellan bedömare 1 och 2 på 91 %, mellan bedömare 1 och 3 på 90 %, och mellan bedömare 2 och 3 på 92 %. Dessa värden är i samma storleksordning som motsvarande procentuella överensstämmelse för provbetyg. Här finns dock ingen skillnad mellan de olika paren av bedömare.

Korrelationer

I tabell 8 redovisas parvisa korrelationer mellan de provbetyg som bedömningarna resulterat i.

Tabell 8 Provbetyg, parvisa korrelationer (Spearmans rangkorrelation).

Provbetyg	Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Bedömare 1	1,000	,947(**)	,862(**)	,886(**)
Bedömare 2	,947(**)	1,000	,862(**)	,873(**)
Bedömare 3	,862(**)	,862(**)	1,000	,841(**)
Lärare	,886(**)	,873(**)	,841(**)	1,000

** Korrelationen är signifikant ($p < 0.01$).

Korrelationerna uppvisar samma bild som den procentuella överensstämmelsen. Även här är det paret med bedömare 1 och bedömare 2 som har de högsta värdena. Övriga är relativt jämna på en något lägre nivå.

Motsvarande korrelationer kan beräknas för provpoäng, även om ett annat korrelationsmått används för denna typ av data (Pearsons produktmomentkorrelation). I tabell 9 redovisas korrelationer mellan olika par av bedömare utifrån de poäng de givit de 99 elevsvaren i studien.

Tabell 9 Provpoäng, parvisa korrelationer. (Pearsons produktmomentkorrelation)

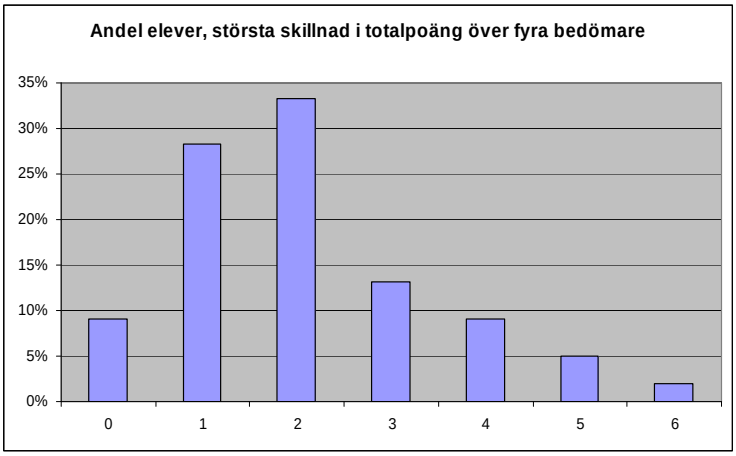
Provpoäng	Bedömare 1	Bedömare 2	Bedömare 3	Lärare
Bedömare 1	1	,991(**)	,988(**)	,982(**)
Bedömare 2	,991(**)	1	,986(**)	,978(**)
Bedömare 3	,988(**)	,986(**)	1	,978(**)
Lärare	,982(**)	,978(**)	,978(**)	1

** Korrelationen är signifikant ($p < 0.01$).

Korrelationer utifrån provpoäng är höga jämfört med motsvarande korrelationer utifrån provbetyg. Detta är ett rimligt resultat eftersom provpoängen är uppdelad i fler steg än provbetygen. En avvikelse med en poäng är liten i förhållande till totalpoängen och därmed påverkas inte korrelationen speciellt mycket.

Variationsbredd i bedömning av enskilda elevsvar (provpoäng)

För varje elevsvar har skillnaden mellan bedömarnas högsta och lägsta poäng beräknas. Denna skillnad varierar från 0 till 6 poäng, och frekvensen för de olika skillnaderna redovisas i figur 1 nedan.



Figur 1 Andel elever med variationsbredd i totalpoängen över de fyra bedömarna. För varje elev har den maximala skillnaden beräknas**.**

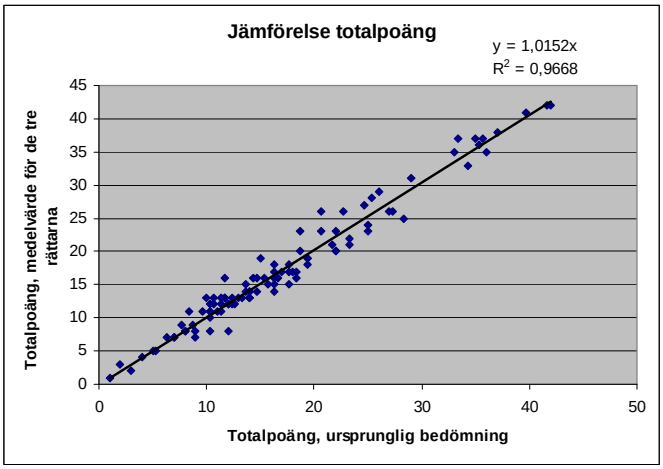
För nästan 70 % av elevsvaren så skiljer de fyra bedömarna sig åt med maximalt två poäng, vilket är samma skillnad som accepterades i den tidigare redovisade analysen av procentuell överensstämmelse på poäng. Denna skillnad kan anses vara tillräckligt liten för att anse att eleverna bedöms lika av olika bedömare.
Andelen elever med en måttlig skillnad mellan olika bedömare (tre eller fyra poäng) är cirka 22 procent och andelen elever med en avsevärd skillnad mellan högsta och lägsta poängbedömningen (fem eller sex poäng) är sju procent.

Bedömaröverensstämmelse på enskilda elevsvar

På 29 av de 99 elevsvaren finns det avvikelser mellan bedömarna när det gäller provbetyg. Skillnaden mellan högsta och lägsta bedömningen är aldrig mer än ett betygssteg. Det betyder att för 71 % av elevsvaren så var alla fyra bedömarna överens om provbetyget. För 20 % av elevsvaren var det en bedömare som avvek från de övriga, och för 9 % av elevsvaren var bedömarna överens två och två.

Bland de 77 elevsvar där bedömare 1-3 var helt överens, så finns sju elevsvar där läraren avviker i sin bedömning. För fem av dessa har läraren gett ett högre provbetyg och för två har läraren gett ett lägre provbetyg än bedömarna. Av dem som fått högre provbetyg i lärarens ursprungliga bedömning gäller tre provbetyget Godkänd och två provbetyget Väl godkänd.

I Figur 2 redovisas en analys där totalpoängen i lärarens ursprungliga bedömning ställs mot medelvärdet för bedömarna.



Figur 2 Sambandet mellan medelvärdet för bedömarna och totalpoängen i lärarens ursprungliga bedömning.

Figuren visar inga systematiska skillnader mellan lärarens ursprungliga bedömning och den ombedömning som gjorts i denna studie. Det finns alltså inga signaler i detta material om att lärare systematiskt ger högre provbetyg än en extern bedömare.

Uppgifter med eventuella reliabilitetsproblem

För att undersöka om det finns enskilda uppgifter där bedömningen skiljer sig åt har procentuell överensstämmelse mellan bedömare beräknats på uppgiftsnivå. För de allra flesta av uppgifterna ligger bedömaröverensstämmelsen på 90 % eller högre. Det finns dock tre deluppgifter som har en lägre överensstämmelse, nämligen uppgift 4, uppgift 13d och den aspektbedömda uppgiften, uppgift 16.

Uppgift 4 är den uppgift som har lägst procentuell överensstämmelse, mellan 56,6 % och 78,8 %. Det är en standarduppgift som på ett eller annat sätt alltid finns med i C-kursproven. Uppgiften ligger på den räknarfria delen.

Utifrån elevlösningarna finns det två förklaringar till de relativt låga procentuella överensstämmelserna. Den första förklaringen är att uppgiften innehåller ett moment som innebär att eleverna förväntas verifiera att det svar de kommer fram till. Kravet på verifiering är implicit det vill säga inget som anges i uppgiften men det står i bedömningsanvisningen. Bedömningsanvisningen föreskriver däremot inte vilken metod eleverna ska använda, vilket innebär att detta krav på verifiering hanteras olika av olika lärare. Beroende på lärares syn på detta krav skulle en olikhet kunna uppstå.

Den andra förklaringen är att vissa elever inte följer den metod som förväntas i uppgiften. Den specifika uppgiften i detta prov har en karaktär som uppenbarligen uppmuntrat eleverna att använda en metod som fungerar i detta fall, men som inte

är den typiska som eleverna bör ha lärt sig i kursen. På grund av att den dessutom ligger på den räknarfria delen i provet så är de beräkningar eleverna ska utföra relativt enkla och det slutgiltiga svaret är ett heltal. Eleverna kommer därmed fram till ett korrekt slutresultat oavsett vilken metod de använder. Bedömningsanvisningen täcker inte det eleverna gör om de använder den icketypiska metoden. Bedömarna har av den anledningen själva fått tolka bedömningsanvisningen, vilket kan ge upphov till vissa skillnader.

Nästa uppgift med vissa reliabilitetsproblem är uppgift 13 d. I denna uppgift är det en av bedömarna som skiljer sig åt gentemot övriga bedömare och läraren. Uppgiften karakteriseras av att ett korrekt slutresultat kan uppnås med en felaktig metod, på grund av det matematiska innehållet. Bedömare 3 har i denna uppgift delat ut full poäng oavsett vilken av metoderna eleverna använt sig av vilket bedömare 1 och 2 inte gjort, de har bara gett poäng för den korrekta metoden. Vad gäller läraren så har vissa gett poäng och andra inte.

Den tredje uppgiften med lägre procentuell överensstämmelse är den mer omfattande uppgiften som också bedöms med så kallad aspektbedömning (uppgift 16). Överensstämmelsen för vg-poängen är relativt bra, den ligger strax under 90 % för alla par av bedömare. Däremot för g-poängen är procentuella överensstämmelsen något sämre. Den lägsta parvis överensstämmelsen är 71,7 %.

Bedömningsmatrisen, som normalt består av tre aspekter, är i detta fall lite extra komplicerad eftersom den på den andra aspekten, matematiskt resonemang, är uppdelad i två delaspekter. Resonemangen kan därmed följa två olika spår, vilka båda kan ge poäng. En fullständig lösning av uppgiften kommer att inkludera båda delaspekterna.

De tre bedömarna i studien har alla redovisat sin bedömning uppdelat på de tre aspekterna. Däremot har läraren i ursprungliga bedömningarna relativt sällan redovisade i de tre aspekterna vilket omöjliggör en jämförelse med dem. För att bättre analysera vari skillnaderna består har de tre bedömarnas poängsättning för varje aspekt kontrollerats. Det är för totalt 29 elevlösningar som det finns någon skillnad i bedömningen, övriga 70 elevlösningar har av de tre bedömarna bedömts på exakt samma sätt. De skillnader som finns uppstår inom aspekterna *Matematiska resonemang* samt *Redovisning och matematiskt språk*. För aspekten *Metodval och genomförande* är skillnaderna små. De elevlösningar som gett upphov till olikheter i bedömningen är sådana där eleven försöker sig på en lösning av uppgiften, gör en del resonemang men inte exakt sådana som exemplifieras i bedömningsanvisningen. Det är därmed en skillnad i hur lärare tolkar vad som ska krävas för att få g-poängen i det matematiska resonemanget.

Den andra delaspekten vad gäller resonemang innebär att eleven till att börja med måste göra vissa beräkningar för att sedan dra en slutsats. Beräkningarna ska göras

med derivata eller annan liknande metod, dvs. det är möjligt att lösa uppgiften grafiskt med hjälp av sin räknare. Det är i 14 av elevlösningarna skillnad mellan bedömarna och det är g-poängen som verkar vara svårbedömd.

I de elevlösningar där skillnader mellan bedömarna uppstår så finns två typer av elevsvar. Elevsvar som innehåller korrekta beräkningar utan slutsats respektive felaktig beräkning med godtagbar slutsats utifrån gjorda beräkningar, visar skillnader mellan bedömarna. Den första typen innebär att vissa bedömare ger poäng enbart för beräkningen även om bedömningsanvisningen dessutom kräver en slutsats. Den andra typen går att betrakta som ett följdfel. I de allmänna instruktionerna till bedömningsanvisningen står det att om en elev gör ett fel som inte avsevärt underlättar beräkningen så ska efterföljande poäng bedömas som korrekta.

Bedömningen av aspekt nummer tre, redovisning och matematiskt språk, skiljer sig åt i sju av elevlösningarna. Bedömare 1 har i alla dessa fall delat ut en poäng för redovisningen, bedömare 2 har gett två av de sju elevlösningarna redovisningspoängen och bedömare 3 har i två av de sju elevlösningarna delat ut en redovisningspoäng. Bedömningen av redovisning och matematiskt språk är den del där lärarna får minst hjälp i sin bedömning. Förvisso finns det bedömda elevlösningar men det är ändå svårt att finna elevlösningar som anger den nedre gränsen för poängen. Dessutom finns klausulen i bedömningsanvisningarna om att läraren ska bedöma elevernas lösningar utifrån den undervisning som föregått provet. Det innebär att en viss skillnad i bedömningen kan uppträda bara på grund av detta.

MVG-bedömning

I de nationella proven i matematik finns till de ¤-märkta uppgifterna en speciell tabell innehållande bedömningsanvisningar för MVG-bedömning. Tabell 10 är en generell tabell. De MVG-kvaliteter, 1-5, som finns redovisade i tabellens vänstra kolumn är en tolkning av kursplanens betygskriterier. Till varje ¤-märkt uppgift görs en uppgiftsspecifik bedömningsanvisning som infogas i tabellens högra kolumn.

Tabell 10 Tabell innehållande de MVG-kvaliteter som kan vara möjliga att visa.

MVG-kvalitet	visar eleven i denna uppgift genom att:
1. Formulerar och utvecklar problem, använder generella metoder/modeller vid problemlösning	
2. Analyserar och tolkar resultat, drar slutsatser samt bedömer rimlighet	
3. Genomför bevis och analyserar matematiska resonemang	
4. Värderar och jämför metoder/modeller	
5. Redovisar välstrukturerat med korrekt matematiskt språk	

Vad gäller möjligheten att granska de bedömningar av MVG-kriterier som gjorts är vi begränsade till de tre bedömarna eftersom alla lärare inte har specificerat i elevernas svarshäften vilka MVG-kriterier som uppfyllts. En anledning till detta kan vara att lärarna använder det kopieringsunderlag för MVG-bedömning som följer med bedömningsanvisningen som sedan inte bifogas när de skickar in elevlösningen.

Det är totalt 17 elevlösningar som enligt en eller flera av de tre bedömarna klarat något eller några av de MVG-kriterier som anges i provet. Av dessa är det 12 elevlösningar som enligt alla tre bedömarna klarat något eller några MVG-kriterier. Tre av de 17 elevlösningarna har enligt två av bedömarna klarat något eller några MVG-kriterier och två av elevlösningarna har enligt en av bedömarna klarat något eller några MVG-kriterier.

Uppgift 8b gav eleverna möjlighet att visa belägg för kriterierna tre och fem tabell 10. Uppgiften är ett bevis som görs via beräkningar. Det krävs inget resonemang. För denna uppgift har elevsvaren bedömts lika av alla tre bedömarna, bortsett från en elevlösning där bedömare 1 inte ansett att eleven uppfyller kriteriet för redovisning och språk (kriterium 5). Uppgiften är matematisk i sin utformning. Eleven förväntas tolka informationen i uppgiften och omsätta detta till en beräkning. Det finns därmed ett tydligt slutresultat vilket skulle kunna underlätta bedömningen. En fullständig redovisning blir inte speciellt omfattande, vilket även det kan göra bedömningen enklare.

I uppgift 15 hade eleverna möjlighet att visa belägg för kriterierna 2 och 5. MVG-bedömningen av elevsvaren uppvisar ganska stora skillnader mellan bedömarna.

Uppgiften kräver ett resonemang, där eleven med ord och en figur ska avge sitt svar. Bedömare 1 har bedömt att fem elever uppfyller ett eller båda kriterierna i uppgiften. Bedömare 2 har bedömt att endast en elev uppfyller ett av kriterierna och bedömare 3 har bedömt att nio elever uppfyller ett eller båda kriterierna. Intressant att notera är att bedömare 3 har för tre av eleverna endast angett kriterium 5, redovisning och matematiskt språk. Det står förvisso inte i bedömningsanvisningen att det inte får göras, men tanken med kriterium 5 är att eleven ska ha visat andra MVG-kvaliteter för att redovisningen ska kunna bedömas.

Uppgiftens bedömningsanvisning är relativt strukturerad och tydlig, vilket borde tala emot skillnader mellan bedömare. Å andra sidan innehåller anvisningen formuleringar som är tolkningsbara, t.ex. när har eleven gjort en godtagbar redogörelse? Det kan ge upphov till skillnader i bedömningen.

Uppgift 16, där eleverna hade möjlighet att visa belägg för kriterierna 2, 3, 4 och 5 visar resultaten på en skillnad mellan bedömare 3 kontra bedömare 1 och bedömare 2. Bedömare 3 har, till skillnad från uppgift 15, bedömt att färre elever uppnår någon av MVG-kriterierna på denna uppgift. Totalt är det 15 elever som bedömts uppfylla något av kriterierna. Bedömare 3 har angett att tio elever uppfyller minst ett MVG-kriterium på uppgift 16, bedömare 2 har angett 15 elever och bedömare 1 har angett 13 elever.

I tabell 12 redovisas antal elever som respektive bedömare bedömts uppfylla respektive kriterium.

Tabell 12 Antal elever som fått de olika MVG-kvaliteterna på uppgift 16.

	Bedömare 1	Bedömare 2	Bedömare 3
MVG-kriterium 2	7	5	4
MVG-kriterium 3	13	14	10
MVG-kriterium 4	4	6	3
MVG-kriterium 5	9	8	5

Endast tre av elevlösningarna till uppgift 16 har MVG bedömts på exakt samma sätt av alla tre bedömarna. För fyra av eleverna innebär skillnaden i bedömning av MVG-kriterierna i uppgift 16 en skillnad i provbetyg.

Sammanfattningsvis vad gäller MVG-bedömningen så finns det uppgifter som verkar vara relativt enkla att bedöma och andra som är betydligt mer komplicerade. Utifrån resultaten i denna studie så verkar det vara synen på kvaliteten hos och presentationen av resonemang som orsakar de största skillnaderna mellan bedömarna.

Uppgifter med ett tolkningsutrymme för läraren

I bedömningsanvisningarna så står i de allmänna riktlinjerna att vissa typer av bedömningar ska lämnas till lokala beslut. Lärarna ska därmed ha samma krav på motiveringar, redovisning, avrundningar, följdfel m.m. som de tillämpar under lektioner och vid andra provtillfällen. Denna typ av skillnader kan inte ses som en brist i bedömaröverensstämmelse utan betraktas som en naturlig följd av att lärarna har ett sådant friutrymme.

Erfarenheten från arbetet med att bedöma de elevlösningar som ska inkluderas i bedömningsanvisningen till de nationella proven visar på att olika lärare har olika krav på utförligheten i elevernas lösningar. Vissa lärare kräver att eleverna motiverar allt de använder sig av medan andra inte har samma krav. Detta visar sig i den aspektbedömda uppgiften, både för poängen och för bedömningen av mvg-kvaliteter. För uppgift 15, som är en ¤-märkt uppgift, finns en skillnad i bedömningen av kvaliteten för redovisning och matematiskt språk. I bedömningsanvisningen bifogas elevlösningar som ska vara ett stöd för lärarna i deras bedömning, men det är inte alltid så enkelt att överföra bedömningsanvisningens exempellösning på den verkliga situationen.

Verifieringen av extrempunkter i max-min problem är också en bedömning där ett friutrymme uppstår. Här förekommer skillnader i bedömningen där vissa lärare kräver att verifieringen ska ske på ett visst sätt, med hjälp av ett teckenschema eller med andraderivata, medan andra lärare anser att en sådan verifiering inte är så viktig, det räcker att eleven finner derivatans nollställen och väljer rätt rot. Kursplanen ger inget stöd i denna fråga. Den elev som aldrig behövt göra verifiering via teckenschema eller dylikt på lektionerna och på de lokala proven kommer troligtvis inte heller göra det på de nationella proven. Denna skillnad framträder i denna studie i uppgift 4 och delvis i uppgift 16.

Det finns ett antal uppgifter där eleverna inte får ett exakt svar, där det skulle kunna uppstå skillnader på grund av att bedömarna kräver olika noggrannhet i avrundningen. Det är däremot inget som har visat sig i denna studie. Har eleven gjort rätt så har alla bedömare i studien delat ut poäng även om eleven inte avrundat korrekt, bedömaröverensstämmelsen i denna studie har inte påverkats av detta. Det är däremot inte möjligt att utifrån empirin i denna studie att dra några slutsatser om hur det ser ut generellt.

Slutsatser och diskussion

Granskningen av bedömaröverensstämmelsen visar acceptabla eller till och med bra nivåer för det prov i matematik C som ingår i studien. I provet ingår en hel del kortsvarsuppgifter där bedömningen är entydig och där bedömare i princip är helt eniga. Dessutom är bedömningen relativt överensstämmande för det uppgifter som kräver en längre lösning eller ett mer resonerande svar, de flesta uppgifterna har en

procentuell överensstämmelse på över 90 %. Generellt sett är matematikens bedömningsanvisningar relativt utförliga. Det finns oftast en beskrivning för varje poäng även om den kan vara ganska generell. Det beror dock oftast på att det finns flera olika lösningsmetoder som är acceptabla.

Vissa skillnader är naturliga på grund av att det finns en frihet i bedömningen. I och med att lärarna förväntas bedöma utifrån den undervisning som skett så kommer det att uppstå vissa skillnader. Det innebär att om överensstämmelsen skulle vara för hög så finns inget sådant friutrymme. Denna brist i överensstämmelse kan inte anses vara ett interbedömarreliabilitetsproblem eftersom vi har det system vi har. En mer styrande bedömningsanvisning kan troligen öka bedömaröverensstämmelsen, men kommer samtidigt att minska lärarens möjligheter att göra lokala tolkningar av målen.

Skillnader i bedömningen vad gäller elevernas användning av räknare är inget som har varit möjligt att granska i denna studie eftersom det inte funnits någon elevlösning där eleven tydligt använt sin räknare. Det är däremot något som diskuterats ingående i samband med provkonstruktionen. Till vissa uppgifter är det naturligtvis helt i sin ordning att eleverna använder sitt hjälpmedel i form av räknare. Det finns däremot inga principer för hur eleverna ska redovisa detta. Det här är något som kommer att bli mer intressant när andelen elever med de mer avancerade symbolhanterande räknarna kommer att öka. Det blir naturligtvis ingen orättvisa inom varje enskild klass men mellan klasser kan det uppstå skillnader.

Det som överlag verkar orsaka störst problem att bedöma är uppgifter där eleven ska göra någon form av resonemang och där läraren ska bedöma elevernas språk. Det finns bland lärare ganska olika synsätt på vad eleven måste göra för att få poäng och anses uppfylla MVG-kriterier för sitt resonemang och språk. Ofta finns det till denna typ av uppgifter bedömda elevsvar med kommentarer men det är inte alltid helt lätt att hitta lösningar som ”ligger på gränsen”. Dessutom finns det till vissa uppgifter nästan lika många formuleringar av svar som det är elever, vilket gör att det i slutändan är upp till läraren att göra en tolkning av bedömningsanvisningen och relatera den till den undervisning som föregått provet. Har man inte lärt eleverna att göra ett korrekt bevis så är det naturligtvis svårt för eleven att prestera ett sådant på ett prov. Å andra sidan har man höga krav på eleverna så kommer det att avspegla sig även i bedömningen av de nationella proven. Detta kommer att generera en skillnad mellan bedömarna. Kursplanerna ger inte heller här något stöd för vad som är det korrekta utan det är upp till de lokala tolkningar som gjorts. För att uppnå högre överensstämmelse vad gäller resonemang samt redovisning så är det säkert möjligt att försöka förtydliga bedömningsanvisningarna vilket vi arbetar med hela tiden. Dessutom kan säkert informationen i de bedömda elevlösningarna förbättras ytterligare. Om dessa åtgärder är tillräckliga är däremot osäkert. Problemet med uppgifter som genererar ett resonerande svar är att beroende på den lokala tolkning läraren gjort av kursplanen så kommer olikheter att uppstå.

Vad ska man då göra om man vill öka överensstämmelsen mellan olika bedömare i matematik? För det första kan man som angett ovan specificera bedömningsanvisningarna ytterligare men då uppstår en risk att det sker en begränsning i de lösningsmetoder som undervisas. För det andra skulle sannolikt skillnader i bedömningen vad gäller elevernas metoder vid lösandet av uppgifterna kunna minskas om bedömningsanvisningarna i vissa fall inte bara beskrev vilka kvaliteter som ska ge poäng, utan också specificerade vad eleven inte får göra. För det tredje skulle en träning av bedömarna kunna genomföras. Stemler (2004) nämner vid ett flertal tillfällen i sin artikel vikten av att träna bedömarna. Detta är någonting som är en viktig del i t.ex. bedömningen av de stora internationella komparativa studierna så som TIMSS och PISA. Där får bedömarna träna på att bedöma ett antal elevhäften innan den egentliga bedömningen startar, allt för att öka överensstämmelsen i de bedömningar som görs. Det skulle vara möjligt att göra något motsvarande till de nationella proven. Till bedömningsanvisningen skulle man kunna bifoga ett antal bedömda elevlösningar som lärarna förväntas träna på och diskutera med övriga lärare inom ämnet innan den egentliga rättningen startar. Det skulle sannolikt öka överensstämmelsen i bedömningen. Det kräver dock betydligt mer utförligt bedömda elevlösningar än vad som är fallet i dagens bedömningsanvisningar.

Referenser

Boesen, J. (2004). *Bedömarreliabilitet. Med fokus på aspektbedömningen i det nationella B-kursprovet i matematik våren 2002* (No. 195). Umeå: Enheten för pedagogiska mätningar.

Lindström, J.-O. (1998). *Rättvis rättning i nationella prov* (Pm No. 144). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.

Olofsson, G. (2004?). *Likvärdig bedömning? En studie av lärares bedömning av elevarbeten på ett nationellt prov i matematik kurs A*. Stockholm: Stockholms universitet: PRIM-gruppen.

Palm, T. (2008). Interrater reliability in a national assessment of oral mathematical communication. *Nordic Studies in Mathematics Education, 13*(2).

Skolverket. (2000). *Naturvetenskapsprogrammet. Programmål, kursplaner, betygskriterier och kommentarer* (2000:14). Stockholm: Fritzes

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability [Electronic Version]. *Practical Assessment, Research & Evaluation*, 9. Retrieved January 18, 2008 from http://PAREonline.net/getvn.asp?v=9&n=4.