Bedömaröverensstämmelse

vid bedömning av elevernas prestationer på de skriftliga delproven i 2017 års nationella prov i matematik för årskurs 9

Charlotte Nordberg, Astrid Pettersson och Samuel Sollerman

MATEMATIKDIDAKTISKA TEXTER
Beprövad erfarenhet och vetenskaplig grund
Del 8

PRIM-gruppen
Institutionen för matematikämnets
och naturvetenskapsämnenas didaktik

Stockholms universitet

Om rapportserien Matematikdidaktiska texter

Denna rapport ingår i en serie publikationer från PRIM-gruppen vid Stockholms universitet. Serien består av texter som behandlar olika aspekter av matematikdidaktik, ofta med fokus på bedömning. Texterna publiceras i rapport- eller i bokform beroende på texternas storlek och syfte. Texterna har sin utgångspunkt i PRIM-gruppens forskning och arbete och baseras på beprövad erfarenhet och vetenskaplig grund. Tidigare utgåvor finns listade i slutet på denna rapport samt på PRIM-gruppens hemsida.

Om PRIM-gruppen

PRIM-gruppen är en forsknings- och provutvecklingsgrupp vars främsta fokus är bedömning av kunskap och kompetens (främst inom området matematik). PRIM-gruppens inriktning inkluderar såväl utveckling, analys och konstruktion av centralt utarbetade prov, nationella utvärderingar och internationella kunskapsmätningar som bedömning med formativ inriktning. Forskningen omfattar såväl kvalitativa som kvantitativa analyser på främst empiriska data. Fortlöpande forskningsverksamhet bedrivs i anslutning till utvecklandet av bedömningsmetoder och mätinstrument kopplade till kunskapsundersökningar. Av speciellt intresse är frågor kring bedömning av kunskaper i matematik i storskaliga nationella och internationella undersökningar. PRIM-gruppen utvecklar bland annat nationella prov i matematik och olika nationella bedömningsstöd samt är ansvariga för den svenska delen av matematiken i de internationella studierna PISA och TIMSS. För kontakt eller mer information om PRIM-gruppen besök hemsidan www.su.se/primgruppen/.

Om författarna till denna rapport

Charlotte Nordberg är provutvecklare och provansvarig för det nationella provet i årskurs 9. Astrid Pettersson är professor i pedagogik med inriktning mot utvärdering och matematikämnets didaktik. Samuel Sollerman är doktor i matematikämnets didaktik med inriktning mot bedömning och föreståndare för PRIM-gruppen Författarna har lång erfarenhet av undervisning på grundskole- och gymnasienivå.

ISSN 1654-0646

Layout: Yvonne Emond och Veronica Palmgren, PRIM-gruppen

Innehåll

Inledning

Två viktiga begrepp vid kunskapsundersökningar är validitet och reliabilitet. Validitet kan övergripande beskrivas som att det innebär att kunskapsundersökningarna ska pröva det som avses att prövas och göra det på ett sådant sätt att resultaten blir så informativa och användbara som möjligt. För att en kunskapsundersökning ska ha hög validitet krävs det emellertid inte bara att vi bedömer det vi avser att bedöma utan det krävs också att bedömningen har hög reliabilitet.

Reliabilitet handlar om att bedömningarna ska vara stabila och inslaget av slumpmässiga avvikelser ska vara litet. Är det så, säger man att reliabiliteten är hög. Målet är att bedömningarna är så reliabla så att de ger liknande resultat, helst samma, om de till exempel genomförs vid upprepade tillfällen eller om någon annan lärare genomför bedömningen av samma elevprestationer. I praktiken innehåller samtliga bedömningar någon form av svårigheter med reliabiliteten (Crocker & Algina, 1986). Denna studie handlar om interbedömarreliabiliteten genom en undersökning av bedömaröverensstämmelsen, d.v.s. den del av en kunskapundersöknings reliabilitet som kan tillskrivas osäkerheten i själva bedömningen och som uttrycks som en variation mellan olika bedömares bedömning av samma elevlösning.

Stemler (2004) menar att interbedömarreliabilitet som begrepp innefattar tre angreppssätt vilka alla kräver olika metoder för att studera bedömaröverensstämmelse. Enligt Stemler kan interbedömarreliabilitet vid analys delas upp i konsensusmått, konsekvensmått och mätningsmått.

Syftet för denna studie är att analysera överensstämmelsen mellan lärares bedömningar. Fokus ligger således på hur pass lika bedömare använder och tolkar bedömningsanvisningar och därför används metoder vilka huvudsakligen fokuserar konsensus. För att även bidra med viss information om konsistensen används även, i en begränsad utsträckning, en metod med konsekvensmått för detta.

Konsensusmått är de mått som används för att mäta hur pass lika bedömare tillämpar bedömningsanvisningar. Det bakomliggande antagandet är att bedömare utifrån samma bedömningsanvisningar bör komma fram till samma bedömning. I denna rapport används konsensusmåtten procentuell överensstämmelse och Cohens kappa.

Konsekvensmått utgår från att det inte är helt nödvändigt att bedömare tolkar och tillämpar bedömningsanvisningar lika så länge som varje bedömare är konsekvent enligt sin egen tolkning av bedömningsanvisningarna. Konsekvensmått och dess metoder används för att mäta hur pass konsistenta bedömare är i sina egna bedömningar. I denna rapport används konsensusmåttet Spearmans rangkorrelation.

Med mätningsmått används all tillgänglig information från alla bedömare när man försöker skapa ett sammanfattande omdöme för varje elevprestation (till exempel en totalpoäng eller ett delprovsbetyg). Varje bedömare anses ge någon unik information som är användbar för att generera det sammanfattande omdömet för en elev. Det är med dessa mått inte nödvändigt att två bedömare har samma tolkning och tillämpning av bedömningsanvisningar, skillnader i bedömarnas stränghet kan beräknas och korrigeras för. Eftersom denna studie inriktar sig på enskilda uppgifter och hur pass lika bedömare använder och tolkar bedömningsanvisningar så fokuseras inte mätningsmått.

Vid två tidigare tillfällen har PRIM-gruppen undersökt bedömaröverensstämmelsen mellan några externa bedömare av nationella provet i årskurs 9 i matematik för de skriftliga delproven. Det har gjorts 2002 och 2007 (Kjellström & Olofsson, 2004; Skolverket, 2009). För gymnasieskolan har motsvarande undersökningar gjorts (Lindström, 1998; Olofsson, 2006 och Lind Pantzare, 2015). När det gäller de skriftliga delproven är överensstämmelsen hög, medan för de muntliga delproven är överensstämmelsen lägre än för de skriftliga delproven (Sollerman, 2016, 2017).

Nationella provet i matematik år 2016/2017

Syftet med de nationella proven 2016/2017 var att stödja en likvärdig och rättvis bedömning och betygssättning och att ge underlag för en analys av i vilken utsträckning kunskapskraven uppfylls på skolnivå, huvudmannanivå och på nationell nivå. De nationella proven kan också bidra till att konkretisera kursplanen och till en ökad måluppfyllelse för eleverna.

De nationella proven i matematik konstrueras utifrån grundskolans läroplan och kursplanen i matematik. I fokus vid konstruktionen står kursplanens syfte, centrala innehåll och kunskapskraven. Proven är konstruerade med fokus på att uppnå både bredd, djup och variation för att eleverna ska ges möjlighet att visa sina kunskaper i matematik på flera olika sätt. Ett nationellt prov är uppdelat i fyra provdelar som tillsammans avser att pröva alla i kursplanen beskrivna förmågorna i matematik.

Delprov A är ett muntligt delprov där avsikten med uppgifterna är att de ska passa bättre för muntlig kommunikation än för skriftlig redovisning samt bjuda in till diskussion mellan eleverna.

Delprov B är ett skriftligt delprov som i huvudsak avser att pröva beräkningar och begreppskunskaper utan hjälpmedel. För de flesta uppgifterna bedöms endast svar.

Delprov C är ett skriftligt delprov som består av en mer omfattande uppgift, där formelblad och miniräknare är tillåtna. Delprovet avser att pröva förmågan att lösa problem, dra slutsatser, generalisera och redovisa tankegångar skriftligt.

Delprov D är ett skriftligt delprov där formelblad och miniräknare är tillåtna. Delprovet avser att pröva förmågan att lösa problem och uttrycka tankegångar skriftligt.

I denna undersökning av bedömaröverensstämmelsen ingår bara de skriftliga delproven, som genomfördes på vårterminen 2017. En motsvarande undersökning har tidigare gjorts beträffande provets muntliga del, delprov A (Sollerman, 2017).

Bedömning av elevernas prestationer på nationella provet

I kursplanen för matematik beskrivs förmågor som undervisningen ska ge eleverna förutsättningar att utveckla. Eftersom kunskapskraven är uppbyggda kring dessa förmågor används en provmodell där kvalitativa förmågepoäng tillämpas. För att tydliggöra de kvalitativa nivåer som finns uttryckta i kunskapskraven används vid bedömningen E-poäng, C-poäng och A-poäng.

Bedömningsanvisningarna i de nationella proven bygger på principen om positiv poängsättning. Det innebär att utgångspunkten är att förtjänster i en elevlösning ska lyftas fram och värderas istället för att det görs poängavdrag för fel och brister. En elev som har kommit en bit på väg mot en lösning kan då få poäng för det han eller hon har visat.

Bedömningen görs på liknande sätt i samtliga uppgifter, men bedömningsanvisningarna skrivs något olika. Till delprov A och delprov C skrivs bedömningsanvisningarna i matrisform och aspekter bedöms på olika nivåer och möjlighet ges att bedöma en aspekt vid flera tillfällen.

Kravgränser för de olika provbetygen sätts enligt vetenskapliga metoder (Angoff, 1971; Impara & Plake, 1997; Sireci, Hambleton & Pitoniak, 2004). I kravgränssättningsgrupperna ingår yrkesverksamma lärare och speciallärare. De har till uppgift att utifrån analys av kursplanen och kunskapskraven genomföra kvalitativa och kvantitativa analyser av provet och föreslå kravgränser för olika betygssteg för provet som helhet. Kravgränserna bygger både på totalpoäng och nivåpoäng. För betygstegen E–A finns krav på totalpoäng och för betygstegen D–A finns också krav på ett minst antal nivåpoäng (på C-nivå och/eller A-nivå).

Genomförande

Syftet med denna studie är att undersöka på vilka sätt och i vilken utsträckning lärares bedömningar skiljer sig åt. Det är t.ex. viktigt att få reda på om kortsvarsuppgifter har större överenstämmelse än uppgifter som kräver redovisning. Ytterligare ett syfte riktar sig till arbetet inom PRIM-gruppen, på vilket sätt kan bedömningsanvisningarna omformuleras, så att dessa blir tydligare och att bedömaröverensstämmelsen kan förbättras. Denna undersökning har ett liknande genomförande som vid tidigare undersökningar.

Urval av elevprov

PRIM-gruppen samlade in cirka 400–450 elevprov från ett slumpmässigt urval av elever födda något av två givna datum. Utifrån dessa valdes 100 elevprov ut

slumpvis. Den information som måste finnas för att ett elevprov skulle kunna vara med i undersökningen var att det fanns ett provbetyg på elevens prov och att lärares bedömning på elevens lösningar på uppgifterna i samtliga delprov B–D var med. Endast elevprov som erhållit provbetyg E–A valdes ut, eftersom dessa elevprov innehåller tillräckligt många elevlösningar för att en undersökning av bedömaröverensstämmelse ska vara relevant.

Elevproven avkodades och den ursprungliga bedömningen doldes. Därefter numrerades elevproven från 1–100 och kopierades. Provbetygsfördelningen för hela provet (inklusie delprov A) i urvalet på 100 elever är E: 30 %, D: 17 %, C: 29 %, B: 13 % och A: 11 %. PRIM-gruppen har samlat in provresultat för drygt 2200 slumpmässigt utvalda elever och betygsfördelningen från den insamlingen är F: 18 %, E: 35 %, D: 14 %, C: 15 %, B: 10 % och A: 8 % (Nordberg & Pettersson, 2017). Reliabiliteten mätt med Cronbach alfa är för delprov B–D sammantaget 0,92, vilket är ett högt värde.

Urvalet till studien begränsades till 100 elevprov med provbetygen E–A. Vid en undersökning av detta urvals representativitet jämfört med det större slumpmässiga urvalet på 2200 provbetyg har en fördelning beräknats där elevprov med provbetyget F är borttagna. Fördelningen blir då E: 43 %, D: 17 %, C: 18 %, B: 12 % och A: 10 %. I urvalet på 100 elevprov är andelen E lägre och andelen C högre, vilket innebär att elevprestationerna i det urvalet är något bättre än i det större urvalet. Det kan bero på att kravet på 100-elevers urvalet är att lärarens bedömning på elevernas lösningar på uppgifter i alla delprov måste vara med.

Urval av bedömare

Förfrågan om att delta i ombedömningen gick ut till några olika lärare som inte tidigare medverkat i arbetet med att konstruera nationella prov, men som tidigare bedömt sina elevers prestationer på de nationella proven i matematik i årskurs 9. Tre lärare valdes ut, två kvinnor och en man, två var från Stockholmsområdet och en var från Örebroområdet. Alla hade lång lärarerfarenhet och erfarenhet av bedömning av elevers prestationer på nationella prov.

Elevproven skickades till de externa bedömarna tillsammans med det aktuella provet och bedömningsanvisningarna. Som stöd till bedömningsanvisningarna finns bedömda elevlösningar till vissa uppgifter. De externa bedömarna förväntades genomföra arbetet på 5–7 dagar och fick betalt för motsvarande den tiden.

Uppskattning av bedömaröverensstämmelsen

Vi har använt tre olika mått för att få en uppskattning av bedömaröverensstämmelsen. Ett mått är den procentuella överensstämmelsen, d.v.s. hur stor andel av poängen respektive uppgifterna för de 100 elevproven som de tre externa bedömarna har gett samma bedömning. Vi har också beräknat den genomsnittliga procentuella överensstämmelsen, d.v.s. ett medelvärde för den parvisa överenstämmelsen för de tre externa bedömarna (Medelvärdet av tre olika parvisa bedömningar, d.v.s. (bedömare 1 x bedömare 2 + bedömare 2 x bedömare 3 + bedömare 1 x bedömare 3)/3). Lie m. fl. (2005) menar att en överenstämmelse mellan två bedömare som är lägre än 85 procent är en dålig överensstämmelse. I vår studie har vi använt tre externa bedömare och vi har också i vissa fall använt lärarnas bedömning och på så sätt får vi fyra olika bedömare av elevproven. För den procentuella överensstämmelsen finns risk för att slumpen kan ge ett visst mått av överensstämmelse. För att förhindra detta använder vi oss också Cohens kappa, som tar hänsyn till slumpinflytande (Cohen 1960, 1968). Cohens kappa kan ha ett värde mellan 0 och 1. Ett riktvärde på minst 0,80 anses vara mycket god överensstämmelse och riktvärden mellan 0,60 och 0,80 som god överensstämmelse (Landis & Koch, 1977). Vid tolkningen av olika värden på Cohens kappa måste man ta hänsyn till hur många poäng en uppgift respektive hela provet kan ge. Ju fler poäng som bedömarna måste ta hänsyn till desto större krav på överensstämmelse krävs för att Cohens kappa ska bli tillfredsställande. Ett tredje mått som vi använder oss av är korrelationer mellan de olika bedömarnas bedömningar. Eftersom poängen på nationella provet grundar sig på det svenska betygssystemet, som är uppbyggt av kvalitativa nivåer, kan poängen anses vara på ordinalskalenivå. Detta leder till att vi som korrelationsmått använder ett icke-parametriskt mått på korrelationen, nämligen Spearmans rangkorrelationskoefficient. Ett acceptabelt värde på korrelation kan sättas till 0,70 (Multon, 2010).

Resultat för bedömaröverensstämmelsen för tre externa bedömare

I tabell 1 och 4 redovisas två av de tre måtten som vi använt för bedömaröverensstämmelsen. Det är den procentuella överensstämmelsen och Cohens kappa. I tabell 5 redovisas korrelationen mellan de tre bedömarna på totalpoängsnivå.

Tabell 1. Procentuell andel överenstämmelse för alla tre externa bedömarna, genomsnittlig procentuell överenstämmelse och Cohens kappa på poäng- och uppgiftsnivå.

Uppgift	Maxpoäng	Procentuell andel elevsvar som alla tre bedömarna är överens om	Genomsnittlig procentuell överensstämmelse per poäng	Genomsnittlig Cohens kappa per poäng	Genomsnittlig procentuell överensstämmelse respektive Cohens kappa för uppgiftens maxpoäng
1	1	97	98	0,96	
2	1	100	100	1,00	
3	1	100	100	1,00	
4	1	97	98	0,95	
5	1	100	100	1,00	
6	1	99	99	0,98	
7	1	97	98	0,96	
8	1	100	100	1,00	
9	1	100	100	1,00	
10	1	99	99	0,98	
11	1	96	97	0,93	
12	2	97 resp 95	98 resp 97	0,88 resp 0,93	96 resp 0,95
13	2	94 resp 99	96 resp 99	0,90 resp 0,99	95 resp 0,95
14	1	96	97	0,95	
15	1	92	95	0,88	
16	1	95	97	0,87	
17	1	99	99	0,99	
18	1	98	99	0,96	
19a	1	99	99	0,97	
19b	2	95 resp 94	97 resp 96	0,93 resp 0,89	93 resp 0,92
20	1	99	99	0,98	
21	1	100	100	1,00	
22	1	99	99	0,96	
23	13	Varje poäng redovisas separat i tabell 4.	44 resp 0,37		
24	3	91, 89 resp 95	94, 93 resp 97	0,55, 0,57 resp 0,89	88 resp 0,63
25	2	76 resp 76	84 resp 84	0,67 resp 0,68	78 resp 0,76
26	3	86, 82 resp 90	91, 88 resp 93	0,79, 0,76 resp 0,85	79 resp 0,68
27	3	94, 91 resp 89	96, 94 resp 93	0,91, 0,88 resp 0,81	83 resp 0,77
28	3	87, 92 resp 90	91, 95 resp 93	0,80, 0,88 resp 0,86	83 resp 0,76
29a	1	95	97	0,74	
29b	1	98	99	0,97	
29c	2	72 resp 75	81 resp 83	0,51 resp 0,66	68 resp 0,62
29d	3	78, 90 resp 91	85, 93, 94	0,71, 0,87 resp 0,87	78 resp 0,66
30	4	83, 74, 84 resp 89	89, 83, 89 resp 93	0,77, 0,66, 0,76 resp 0,81	92 resp 0,88
31a	2	89 resp 94	93 resp 96	0,78 resp 0,84	91 resp 0,91
31b	3	86, 86 resp 95	91, 91 resp 97	0,77, 0,74 resp 83	85 resp 0,83
31c	3	91, 92 resp 80	94, 95 resp 87	0,87, 0,86 resp 0,40	77 resp 0,72
32	2	68 resp 91	79 resp 94	0,56 resp 0,77	76 resp 0,73
33	3	83, 97 resp 87	89, 98 resp 91	0,76, 0,95 resp 0,70	79 resp 0,74

För att tydliggöra och problematisera skillnaden på den genomsnittliga procentuella överensstämmelsen för uppgiftens maxpoäng samt de genomsnittliga procentuella överensstämmelserna för vardera poängen i uppgiften diskuteras här uppgift 26 och 30.

För uppgift 26, som maximalt kan ge tre poäng var överenstämmelsen mellan de parvisa externa bedömarna 73 procent, 76 procent respektive 88 procent, vilket ger ett värde för den genomsnittliga procentuella bedömningen på 79 procent ((73 + 76 + 88) / 3 = 79). Intressant är att det oftast är en extern bedömare som skiljer sig från de andra. De två andra är i mycket högre utsträckning överens. På uppgift 26 har en extern bedömare gett två poäng mer (av tre möjliga) än någon av de andra för några av elevlösningarna. Det räcker med att det är en som bedömer annorlunda för att den genomsnittliga procentuella bedömningen på uppgiftens totalnivå ska skilja sig mycket från den procentuella genomsnittliga överensstämmelsen för vardera poängen.

För uppgift 30 som maximalt kan ge fyra poäng, är den genomsnittliga procentuella överensstämmelsen högre för maxpoängen än för de enskilda poängen. En förklaring till detta är att det är en svår uppgift. Nästan hälften av elevlösningarna bedöms med noll poäng och en fjärdedel av elevlösningarna bedöms med fyra poäng. Det finns även här elevlösningar där en extern bedömare skiljer sig från de övriga två i sina bedömningar. Exempelvis har några av elevlösningarna bedömts med fyra poäng av en bedömare och med bara en poäng av de andra. 92 procent är den genomsnittliga procentuella överensstämmelsen för uppgiften som helhet beräknad som ett genomsnitt av de tre parvisa bedömningarna. För uppgiften som helhet är två externa bedömare överens om 96 procent av elevlösningarna (dvs. 96 procent av elevlösningarna har fått samma totala poäng på uppgiften av dessa två bedömare). Om vi ser till de två andra parvisa bedömningarna är den procentuella överenstämmelsen 87 procent respektive 95 procent.

De skriftliga delproven består av 33 uppgifter och deluppgifterna medräknade är antalet 39. Vi tillämpar Lies m.fl. krav på att minst 85 procent av elevlösningarna ska ha samma bedömning av två bedömare för att överensstämmelsen ska vara acceptabel. Men vi använder oss av tre bedömare, alltså

i denna studie har vi använt ett högre krav än Lies m. fl. När det gäller Cohens kappa tillämpar vi två riktvärden, minst 0,80 för mycket god överensstämmelse och värden mellan 0,60 och 0,80 som god överensstämmelse.

Samtliga uppgifter/deluppgifter som kan ge maximalt en poäng uppfyller kravet på minst 85 procent både vad gäller att alla tre externa bedömarna är överens och vad gäller den genomsnittliga procentuella överensstämmelsen. Kravet på ett Cohens kappavärde på minst 0,80 uppfyller samtliga dessa uppgifter/deluppgifter utom uppgift 29a som har ett värde på 0,74, vilket är en god överensstämmelse.

För uppgifter som kan ge fler än en poäng finns värden på den genomsnittliga procentuella överensstämmelsen och Cohens kappa. Kravuppfyllandet redovisas i tabell 2.

Tabell 2. Översikt över kravuppfyllande på uppgiftsnivå för de uppgifter som kan ge fler än en poäng.

Uppgift/ poäng	Minst 85 % genomsnittlig överensstämmelse	Cohens kappa ≥ 0,80 mycket god 0,61–0,79 god
12/2	Ja	Mycket god
13/2	Ja	Mycket god
19b/2	Ja	Mycket god
25/2	Nej	God
29c/2	Nej	God
31a/2	Ja	Mycket god
32/2	Nej	God
24/3	Ja	God
26/3	Nej	God
27/3	Nej	God
28/3	Nej	God
29d/3	Nej	God
31b/3	Ja	Mycket god
31c/3	Nej	God
33/3	Nej	God
30/4	Ja	Mycket God
23/13	Nej	Uppfyller ej kravet

Det är betydligt vanligare att kravet som Cohens kappa ger uppfylls än att kravet på minst 85 procents genomsnittlig överensstämmelse gör det. Alla uppgifter utom uppgift 23 som kan ge 13 poäng uppfyller kravet på minst god överensstämmelse enligt Cohens kappa, medan endast sju uppgifter uppfyller kravet på minst 85 procents genomsnittlig överensstämmelse. Men även här är det intressant att jämföra de tre olika bedömarnas parvisa bedömningar när kravet på minst 85 procents genomsnittlig överensstämmelse inte uppfylls. Ofta då kravet inte uppfylls är det en extern bedömare som skiljer sig från de andra. Lie m.fl. har satt kravet då det bara är två bedömare medan i vår studie är kravet satt högre, att alla tre bedömares procentuella genomsnittliga överensstämmelse ska vara minst 85 procent.

För de uppgifter/deluppgifter som kan ge maximalt två, tre respektive fyra poäng analyserar vi också kravuppfyllandet på poängnivå.

Tabell 3. Översikt över kravuppfyllande på poängnivå för de uppgifter som kan ge 2, 3 respektive 4 poäng.

Uppgift/ poäng	Minst 85 % överensstämmelse alla tre externa bedömare	Minst 85 % genomsnittlig överensstämmelse	Cohens kappa ≥ 0,80 mycket god 0,61–0,79 god
12/2	Ja, för bägge poängen	Ja, för bägge poängen	Mycket god för bägge poängen
13/2	Ja, för bägge poängen	Ja, för bägge poängen	Mycket god för bägge poängen
19b/2	Ja, för bägge poängen	Ja, för bägge poängen	Mycket god för bägge poängen
25/2	Nej, för ingen poäng	Nej, för ingen poäng	God för bägge poängen
29c/2	Nej, för ingen poäng	Nej, för ingen poäng	God enbart för den andra poängen
31a/2	Ja, för bägge poängen	Ja, för bägge poängen	God för första poängen och mycket god för andra poängen
32/2	Nej, för första poängen Ja, för andra poängen	Nej, för första poängen Ja, för andra poängen	God enbart för andra poängen
24/3	Ja, för alla tre poängen	Ja, för alla tre poängen	Mycket god för tredje poängen, första och andra poängen uppfyller inte kravet god
26/3	Ja, för första och tredje poängen Nej, för tredje poängen	Ja, för alla tre poängen	God för första och andra poängen och mycket god för tredje poängen
27/3	Ja, för alla tre poängen	Ja, för alla tre poängen	Mycket god för alla tre poängen
28/3	Ja, för alla tre poängen	Ja, för alla tre poängen	Mycket god för alla tre poängen
29d/3	Nej, för första poängen Ja, för andra och tredje poängen	Ja, för alla tre poängen	God för första poängen Mycket god för andra och tredje poängen
31b/3	Ja, för alla tre poängen	Ja, för alla tre poängen	God för första och andra poängen Mycket god för tredjepoängen
31c/3	Ja, för första och andra poängen Nej, för tredje poängen	Ja, för alla tre poängen	Mycket god för första och andra poängen Tredje poängen uppfyller inte kravet på god
33/3	Nej, för första poängen Ja, för andra och tredje poängen	Ja, för alla tre poängen	Mycket god för andra poängen God för första och tredje poängen
30/4	Nej, för första, andra och tredje poängen Ja, för fjärde poängen	Ja, för första, tredje och fjärde poängen Nej, för andra poängen	Mycket god för fjärde poängen God för första, andra och tredje poängen.

Då kravet på minst 85 procents överensstämmelse för alla tre bedömarna uppfylls, uppfylls också övriga krav för alla poäng med några undantag. Undantagen gäller de två första poängen i uppgift 24 och den tredje poängen i uppgift 31c. Där uppfylls inte kravet på god överensstämmelse enligt Cohens kappa.

Det kan vara intressant att se om det finns något gemensamt i hur bedömningsanvisningarna är utformade för de uppgifter och poäng där kravet på minst 85 procents överensstämmelse för alla tre bedömare inte uppfylls.

Dessa uppgifter och poäng är: 25 för bägge poängen (2 E-poäng, begrepp och metod), 29c för bägge poängen (1 E-poäng, resonemang och 1 C-poäng, resonemang), 29d för första poängen (1 E-poäng, begrepp) och 31c för tredje poängen (1 A-poäng, kommunikation) samt 32 för första poängen (1 C-poäng, begrepp). Alla dessa uppgifter ingår i delprov D, ett delprov där eleverna måste, med ett par undantag, redovisa sina lösningar. För samtliga dessa uppgifter (25, 29c, 29d, 31c och 32) krävs att eleverna ska redovisa sina lösningar.

I åtta av uppgifterna/deluppgifterna (24, 28, 29c, 29d, 31b, 31c, 32, 33) i delprov D inleds bedömningsanvisningarna med ”Påbörjad lösning, t.ex. …” eller ”Påbörjat resonemang…”. Hälften av dessa uppgifter/deluppgifter har för denna bedömningsgrund, som gav den inledande poängen, inte acceptabla värden för procentuell överensstämmelse. För övriga fem uppgifter/deluppgifter (25, 26, 27, 30, 31a) börjar inte bedömningsanvisningarna med att något ska vara påbörjat i lösningen. I stället används formuleringar som ”visar kunskap om…”, ”beräknar…” och ”teckna algebraiskt uttryck…”. För dessa fem uppgifter är det bara två uppgifter som inte har acceptabla värden på procentuell överensstämmelse för den första poängen. Det kan tyda på att anvisningen ”Påbörjad lösning , t.ex. ...” ger upphov till oenighet i bedömningar. ”Påbörjad lösning, t.ex. ...” är en ganska öppen beskrivning som kan tolkas olika av bedömarna även om exempel ges. Det finns en allmän skrivning om vad som menas med påbörjad lösning i bedömningsanvisningarna: ”… menas att den påbörjade lösningen ska vara relevant och leda framåt. De exempel som skrivs ut är vanliga men det kan också finnas fler sätt att påbörja en relevant lösning av uppgiften”. (Skolverket, 2017b, s. 7). Det kan vara så att det finns färre vägar in i en del uppgifter och då är det lättare att bedöma ”påbörjad lösning”.

Tabell 3 visar att det totalt bara är tre poäng som inte uppfyller kravet på 85 procents överensstämmelse och kravet på minst god överensstämmelse enligt Cohens kappa. Det är en E-poäng i 29c, en A-poäng i 31c och en C-poäng i 32.

Fördjupad analys av de elevlösningar, där avvikelserna i bedömningarna är störst

För fyra uppgifter har inget enskilt par av bedömare en överensstämmelse på minst 85 procent. För uppgift 29c är det sju elevlösningar där spridningen i bedömningarna av de externa bedömarna sträcker sig från noll till maximala två poäng. För uppgift 29d är avvikelsen störst vad gäller vilka elevlösningar som ska ha en poäng. För uppgift 31c är bedömarna relativt överens om vilka elevlösningar som ska ha noll poäng, men mindre överens om elevlösningen ska ha en, två eller tre poäng. Ibland kan det skilja två poäng mellan de olika bedömarna. För uppgift 33 skiljer sig bedömarna vad gäller den första poängen, om den ska delas ut eller ej. Det är därför få elevlösningar som här har fått den första poängen.

Finns det något gemensamt i dessa fyra uppgifter och i bedömningsanvisningarna? Alla uppgifter finns i slutet av delprov D och eleverna ska redovisa sina lösningar. Det gemensamma för dessa uppgifter är att bedömningsanvisningarna för alla uppgifterna inleds med ”Påbörjad lösning, t.ex. ...” och det gäller första poängen. Sammantaget ger dessa uppgifter 2 E-poäng (resonemang och begrepp), 4 C-poäng (1 resonemang, 1 begrepp och 2 problemlösning) 5 A-poäng (1 begrepp, 1 kommunikation, 2 problemlösning och 1 metod). För alla uppgifter finns i bedömningsanvisningarna exempel på bedömda elevlösningar. Det som skiljer är uppgifternas karaktär, t.ex. i krav på redovisning, centralt innehåll och vilka förmågor som prövas.

I uppgift 29 ska eleverna använda tabell och diagram. De två deluppgifter som föregår 29c kräver endast att svar ges.

Uppgift 29c kräver att eleven använder information given i tabell och diagram, gör beräkningar och gör jämförelser samt anger ett svar, som antingen kan vara rätt eller fel. För första poängen, som är en resonemangspoäng på E-nivå, krävs ett påbörjat resonemang. För den andra poängen, som är en resonemangspoäng på C-nivå, krävs ett korrekt svar med underbyggt resonemang. Till bedömningsanvisningarna finns exempel på fyra bedömda elevlösningar, ett 0-poängarbete, där svaret är korrekt men där det saknas redovisad tankegång, ett 1-poängsarbete, där svaret är korrekt och där det finns ett mer allmänt resonemang. Därefter ges exempel på två lösningar som angett korrekta svar men där olika strategier använts. Om vi ser på hur de externa bedömarna har bedömt så antyder resultatet att elevlösningarna från en bedömare kan få poäng där det finns ett korrekt svar men där resonemangen är ofullständiga eller felaktiga, medan en annan bedömare har högre krav på underbyggnad och resonemang även för den första poängen, som är en poäng för påbörjad lösning. Både första och andra poängen innehåller ”Påbörjat resonemang, t.ex. ...”, vilket innebär att läraren måste tolka och ta ställning till vad som kan räcka för poäng. Till hjälp finns exempel på bedömda elevlösningar som alla visar olika strategier vilket kan göra att det är svårt att hitta nivån.

I uppgift 29d prövas begreppsförmågan genom att eleven ska ange två olika formler utifrån information given i tabell och diagram. Första poängen på E-nivå ges för påbörjad lösning och exempel på vad det kan vara ges i bedömningsanvisningarna och i en bedömd elevlösning. För den andra poängen som är på C-nivå krävs två korrekta uttryck eller en korrekt formel medan den tredje poängen på A-nivå ges för två korrekta formler. Exempel på bedömda elevlösningar finns för alla poängnivåer. Den första poängen i denna uppgift ges för ”Påbörjad lösning, t.ex. …”. Två exempel på påbörjad lösning finns i exemplen på bedömda elevlösningar och utifrån det måste bedömaren tolka vilka fler strategier som kan anses vara påbörjade lösningar som kan leda framåt. Poäng två och tre är mer konkret beskrivna. Vid analys av bedömningen av denna uppgift framgår också att bedömningsanvisningen

kan tolkas så att det kan finnas överhopp, d.v.s. att i det här fallet har den andra poängen delats ut men inte den första. Så kan det vara i vissa uppgifter men det är ovanligt. Här har bedömarna tolkat det olika varvid överensstämmelsen för vissa elevlösningar blivit lägre.

I uppgift 31c behöver eleven använda kunskaper i geometri och kunna välja lämplig formel för att lösa uppgiften. Den första poängen, som är en problemlösningspoäng på C-nivå, ges för en påbörjad lösning. Bedömningsanvisningarna ger exempel på de två vanligaste strategierna. Den andra poängen är en problemlösningspoäng på A-nivå och ges för korrekt svar med redovisade beräkningar. Den tredje poängen är en kommunikationspoäng på A-nivå och ges om redovisningen dessutom är lätt att följa med lämpligt matematiskt språk. Till bedömningsanvisningarna finns exempel på fem bedömda elevlösningar. De två första bedömda lösningarna ger en poäng och visar på de två vanligaste strategierna för att påbörja en lösning av uppgiften. Därefter följer en bedömd elevlösning som getts två poäng, där svaret är korrekt men redovisningen är lite otydlig och kommunikationen inte helt korrekt. De två sista bedömda elevlösningarna får alla tre poäng och visar de båda tidigare nämnda strategierna för att lösa uppgiften. Om vi ser på hur bedömarna bedömt elevlösningarna kan man när det gäller den andra poängen, som är en A-poäng, fundera över om det även här tolkas in ett kvalitetskrav på redovisningen/beräkningen även om det inte står uttryckt i bedömningsanvisningarna. Det visar sig vid analys av elevlösningarna där eleven kommit fram till korrekt svar och det finns en redovisning, men redovisningen inte är lätt att följa och vissa delar är utelämnade. Tredje poängen ges om de två första delats ut och redovisningen dessutom är lätt att följa med lämpligt matematiskt språk. Här finns en möjlighet till tolkning av ”lätt att följa” och ”lämpligt matematiskt språk”. De exempel på bedömda elevlösningar som visas i bedömningsanvisningarna berättar att en felaktig användning av likhetstecknet inte betraktas som lämpligt matematiskt språk. De två exemplen på bedömda elevlösningar som ges full poäng är fullständiga och anger också vilka formler som använts och de håller mycket hög kvalitet.

I uppgift 33 behöver eleven använda kunskaper i algebra eller aritmetik för att lösa uppgiften. Bedömningsanvisningarna anger att första poängen är en problemlösningspoäng på C-nivå och ges för påbörjad lösning. Två exempel på bedömda elevlösningar anges. För den andra poängen, som är en problemlösningspoäng på A-nivå, krävs lösning av problemet med korrekt svar. För den tredje poängen som är en metodpoäng på A-nivå, krävs att en effektiv metod (aritmetisk eller algebraisk) har använts. Till bedömningsanvisningarna finns exempel på fem bedömda elevlösningar. De två första bedömda elevlösningarna ger en poäng och visar exempel på hur en lösning av uppgiften kan påbörjas, en med algebraisk metod och en med aritmetisk metod. Den tredje bedömda elevlösningen är ett 2-poängsarbete som verifierat svaret. Elevlösning fyra och fem får tre poäng och visar fullständiga lösningar med korrekt svar, en med algebraisk metod och en med aritmetisk metod. I den här uppgiften är det bedömningen av påbörjad lösning som kan vara anledningen till att bedömningen skiljer sig åt. Några olika tolkningar behöver göras av bedömaren. Vad menas med en påbörjad lösning i denna uppgift? I bedömningsanvisningarna ges två exempel på bedömda elevlösningar som båda kräver viss tolkning. Det är upp till bedömaren att jämföra andra varianter mot dessa för att avgöra om de kan motsvara samma nivå. Utöver det kan elever ha påbörjat en lösning med någon annan metod, till exempel prövning. I bedömningsanvisningarna står det inte utskrivet och inget exempel på bedömd elevlösning visar heller prövning. Däremot finns ett exempel på en bedömd elevlösning som verifierat sitt svar och som kan ha prövat sig fram på ett kladdpapper som inte lämnats in.

Sammanfattningsvis visar den fördjupade analysen att

- Det är i stort sett bara en formulering som kan vara orsak till att bedömaröverensstämmelsen inte är så hög. Det är formuleringen att eleven ska få poäng om hen påbörjat en lösning eller ett resonemang. Även om det till så gott som alla dessa bedömningsanvisningar finns exempel på bedömda elevlösningar så är bedömaröverensstämmelsen inte alltid tillfredställande.
- Vid bedömning av resonemang antyder resultatet att elevlösningarna från en bedömare kan få poäng där det finns ett korrekt svar men där resonemangen är ofullständiga eller felaktiga, medan en annan bedömare har högre krav på underbyggnad och resonemang även för inledande poäng som motsvarar påbörjad lösning. Skillnader i bedömningen hos de olika bedömarna framkom också i tolkningen av bedömningsanvisningarna: ”Redovisningen är dessutom lätt att följa med lämpligt matematiskt språk.”. Här finns en möjlighet till tolkning av ”lätt att följa” och ”lämpligt matematiskt språk”. De exempel på bedömda elevlösningar som visas i bedömningsanvisningarna berättar att en felaktig användning av likhetstecknet inte betraktas som lämpligt matematiskt språk. De två exemplen på bedömda elevlösningar som ges full poäng är fullständiga och anger också vilka formler som använts och de håller mycket hög kvalitet.
- Bedömningsanvisningarna kan tolkas så att det kan finnas överhopp, det vill säga att den andra poängen delas ut men inte den första. Så kan det vara i vissa uppgifter men det är ovanligt. Här har bedömarna tolkat det olika varvid överensstämmelsen vid bedömningen av vissa elevlösningar blivit lägre.

Fördjupad analys av resultaten på uppgift 23

Uppgift 23, som utgjorde hela delprov C, var den mer omfattande uppgiften som ska bedömas med en bedömningsmatris med olika aspekter. Totalt kan uppgiften ge 13 poäng.

Tabell 4. Procentuell andel överensstämmelse för alla tre externa bedömarna, genomsnittlig procentuell överensstämmelse och Cohens kappa på poängnivå för uppgift 23.

Uppgift/ poäng-nivå/ förmåga[1]	Procentuell andel elevsvar på poäng-nivå som alla tre externa bedömarna är överens om	Genomsnittlig procentuell överensstäm-melse	Genomsnittlig Cohens kappa	Antal kriterier uppfyllda (minst 85 % på överensstäm-melserna och ett Cohens kappa på minst 0,60)
23EB	98	99	0,66	3
23EM	89	93	0,83	3
23ER	98	99	0,96	3
23EP	85	90	0,24	2
23EK	78	85	0,52	1
23CM	91	94	0,87	3
23CP	86	91	0,82	3
23CR	94	96	0,92	3
23CK	79	86	0,72	2
23AM	82	88	0,52	2
23AP	89	93	0,81	3
23AR	96	97	0,57	2
23AK	99	99	0,77	3

[1]Första bokstaven i kolumn ett anger betygsnivån, E, C respektive A. Andra bokstaven anger den förmåga som avses (B = Begrepp, M = Metod, R = Resonemang, P = Problemlösning, K = Kommunikation)

För åtta av de 13 poängen är alla krav uppfyllda, 3 poäng är på E-nivå, 3 är på C-nivå och 2 poäng är på A-nivå. Om vi analyserar de poäng som inte uppfyller kravet på minst 85 procents överensstämmelse så är det tre poäng, 1 E-poäng, kommunikationsförmåga, 1 C-poäng också kommunikationsförmåga och 1 A-poäng, metodförmåga. Det är fyra poäng som inte uppfyller kravet på god eller mycket god överensstämmelse enligt Cohens kappa. I två av dessa fall (23EP, 23AR) beror de låga värdena på att en av bedömarna skiljer sig mer än de andra i sin bedömning. I de två andra fallen (23EK, 23AM) är det mindre överensstämmelse mellan alla bedömarna. Bedömningsanvisningarna för denna uppgift har i två fall texten ”Påbörjar ett algebraiskt resonemang…” respektive ”Påbörjar en algebraisk förenkling…”. I det första fallet finns en alternativ bedömning så att eleverna inte behöver ha påbörjat ett algebraiskt resonemang för att få poäng (1 C, resonemang). I detta fall är överensstämmelsen mellan de tre externa bedömarna mycket god för alla de tre kraven. I det andra fallet där eleven skulle påbörja en algebraisk förenkling uppfyller endast ett poäng kravet på överensstämmelse (1 A, metod). Det kan återigen indikera att ordet påbörja kan tolkas på olika sätt av bedömarna.

Uppgift 23 bedömdes med hjälp av bedömningsmatris med tre olika aspekter, Begrepp och Metod (max 4 poäng), Resonemang och Problemlösning (max 6 poäng) och Kommunikation (max 3 poäng). Den genomsnittliga procentuella överensstämmelsen, för poängen i respektive aspekt är 89 procent för Begrepp och Metod, 67 procent för Resonemang och Problemlösning och 71 procent för Kommunikation. Motsvarande värden för Cohens kappa är 0,84, 0,58 respektive 0,68.

Resultat på totalpoängsnivå

Vi har hittills använt två mått för att beskriva bedömaröverensstämmelsen, den procentuella överensstämmelsen och Cohens kappa. De måtten har varit på uppgiftsnivå men framförallt på poängnivå. Det tredje måttet är korrelationen mellan de tre bedömarna när det gäller totalpoängen. Tabell 5 redovisar Spearmans rangkorrelationskoefficient.

Tabell 5. Korrelationen mellan de tre externa bedömarna (totalpoängen).

Bedömare	1	2	3
1	1,00	0,98	0,99
2		1,00	0,99
3			1,00

Multon (2010) anser att 0,70 är ett acceptabelt värde på korrelation när det gäller bedömaröverensstämmelse. Som framgår av tabell 5 är korrelationen i totalpoängen mellan de tre bedömarna mycket hög. Ett högt värde på en korrelation innebär att det är ett starkt samband mellan bedömares totalpoäng. I detta fall betyder det att en elevprestation som hos en bedömare resulterat i en viss nivå på totalpoängen, med mycket stor sannolikhet har resulterat i samma nivå hos en annan bedömare.

Resultat för bedömaröverensstämmelsen för tre externa bedömare och läraren

Eleverna, vars prov har analyserats, har haft olika lärare eftersom eleverna kommer från olika skolor. Det är inte säkert att det alltid är elevens lärare som har bedömt elevens prestationer. Det kan t.ex. vara en annan lärare på skolan eller i kommunen. Om vi tar ett genomsnitt för alla uppgifter i de skriftliga delproven är 83 procent av elevlösningarna bedömda på samma sätt av lärarna som av de tre externa bedömarna. Till 91 procent ger lärarna samma bedömning som minst två externa bedömare. I fem procent av fallen är ingen extern bedömare överens med lärarna. Överenstämmelsen är större för enpoängsuppgifter än för flerpoängsuppgifter. Korrelationen mellan lärarens bedömning och de tre externa bedömarnas på totalpoängsnivå är mycket hög, 0,98 (för en extern bedömare) och 0,99 (för två externa bedömare).

Provbetyg

Det nationella provet för 2016/17 gav totalt för alla delproven (Delprov A–D) maximalt 91 poäng, varav 38 E-poäng, 32 C-poäng och 21 A-poäng. Eftersom denna undersökning enbart omfattar de skriftliga delproven (med totalpoängen 77 poäng, varav 34 E-poäng, 27 C-poäng och 16 A-poäng) har vi räknat om samtliga kravgränser och i denna analys använder vi följande provbetygsgränser:

Provbetyg E: Minst 16 poäng
Provbetyg D: Minst 29 poäng varav minst 6 poäng på lägst nivå C
Provbetyg C: Minst 40 poäng varav minst 13 poäng på lägst nivå C
Provbetyg B: Minst 52 poäng varav minst 4 poäng på nivå A
Provbetyg A: Minst 61 poäng varav minst 8 poäng på nivå A.

Överensstämmelse mellan de tre externa bedömarnas provbetyg

Det beräknade provbetyget grundar sig på totalpoängen för provbetyget E och för provbetygen A–D även på nivåpoängen.

Tabell 6. Procentuell överenstämmelse för provbetygen av de tre externa bedömarna.

Samma provbetyg för alla tre bedömarna	74
En av bedömarna avvek ett steg med sitt provbetyg från övriga två	26

För bedömarparet med högst överensstämmelse stämde provbetyget överens i 90 procent av fallen. Paret med lägst överensstämmelse hade en 78-procentig överensstämmelse. Nästan tre av fyra elevprov är bedömda med samma provbetyg av alla tre externa bedömarna och i övriga fall skiljer sig provbetygen med ett betygssteg. Om vi ser till skillnaden i totalpoäng för de prestationer där provbetygen skiljer sig åt varierar den mellan 1 och 8 poäng, där en skillnad på två poäng är det vanligaste.

Jämförelse mellan lärarens provbetyg och de tre externa bedömarnas

Det beräknade provbetyget grundar sig på totalpoängen för provbetyget E och för provbetygen A–D även på nivåpoängen.

Tabell 7. Procentuell överenstämmelse för provbetygen av de tre externa bedömarna och läraren.

	Andel prov i procent
Alla tre externa bedömarna är överens med läraren	69
Två av de externa bedömarna är överens med läraren	14
En av de externa bedömarna är överens med läraren	12
Ingen av de externa bedömarna är överens med läraren	5

Drygt två elevprov av tre har fått samma provbetyg av samtliga tre externa bedömarna och lärarna och i övriga fall skiljer sig provbetygen med ett betygssteg. Skillnaden i totalpoäng mellan de fyra bedömarna (inklusive lärarna) där provbetygen skiljer varierar mellan 2 och 14 poäng, där en skillnad på 3 poäng är det vanligaste. Skillnaden mellan provbetyg och poäng kan illustreras av de fem fall där ingen av de externa bedömarna har samma provbetyg som lärarna.

Tabell 8. Översikt över betyg och poäng för de elevprov, där den bedömning läraren har gjort har lett till ett annat provbetyg än de tre externa bedömarnas.

Bedömare	Prov-betyg	Poäng	Prov-betyg	Poäng	Prov-betyg	Poäng	Prov-betyg	Poäng	Prov-betyg	Poäng
Läraren	A	67	A	61	B	52	D	30	E	18
Bedömare 1	B	53	B	58	C	49	E	23	F	15
Bedömare 2	B	56	B	58	C	46	E	27	F	13
Bedömare 3	B	59	B	60	C	50	E	28	F	15
Största skillnad i poäng		14		3		6		7		5

Tabell 8 visar att i ett fall behövs bara 3 poängs skillnad för att provbetyget ska bli olika mellan de externa bedömarna och den bedömning som läraren har gjort, i ett annat fall är skillnaden 14 poäng. Det visar hur de skarpa kravgränserna slår mot provbetygen.

Sammanfattande diskussion

Detta är den första studien av bedömaröverenstämmelse för de skriftliga nationella delproven i matematik för årskurs 9 enligt Lgr 11. Två tidigare studier har gjorts för samma årskurs och för de skriftliga delproven, men då utifrån en annan kursplan och med färre antal betygssteg än de sex vi har nu. (Kjellström & Olofsson, 2004; Skolverket 2009). Dessa studier liksom denna visar på en mycket hög bedömaröverensstämmelse.

I denna studie har vi använt oss av tre olika mått på bedömaröverens-stämmelse vad gäller de tre externa bedömarna:

- procentuell överensstämmelse mellan bedömare (görs på två olika sätt) på poängnivå, uppgiftsnivå och provbetygsnivå
- Cohens kappa på uppgiftsnivå och poängnivå
- sambanden mellan bedömarna på totalpoängsnivå, mätt med Spearmans rangkorrelationskoefficient.

Det mått som är mest krävande på poängnivå och uppgiftsnivå är att alla tre bedömarna ska vara överens (minst 85 procents överensstämmelse). Det är mer krävande än måttet genomsnittlig procentuell överensstämmelse och Cohens kappa.

Resultaten visar att bedömaröverensstämmelsen, oavsett mått, är hög eller mycket hög mellan de tre externa bedömarna, men också mellan dessa och respektive lärare som bedömt elevprovet. Korrelationen mellan de olika bedömarna inklusive lärare är minst 0,98 vid en parvis jämförelse på total-poängsnivå.

De beräknade gränserna för provbetygen har inte tagit hänsyn till elevernas prestationer på det muntliga delprovet. Elevernas möjlighet att visa sina kun-skaper muntligt finns alltså inte med i denna undersökning.

För knappt tre av fyra elevprov är alla bedömare inklusive läraren överens på provbetygsnivå. Om vi här tar Lies m.fl. krav på minst 85 procents överens-stämmelse för två bedömare och överför det till provbetygsnivå, finner vi av tabell 7 att provbetyget uppfyller detta krav eftersom för 95 procent av elev-proven sammanfaller lärarens provbetyg med minst en extern bedömare.

Vid arbetet med denna studie är det två förhållanden som vi särskilt vill ta upp förutom den höga bedömaröverensstämmelsen som tidigare har berörts. Det är dels de externa bedömarnas bedömningar dels bedömningsanvisningarna.

Våra analyser visar att många uppgifter fungerar utmärkt att bedöma lika. Det gäller särskilt de uppgifter som har ett begränsat antal korrekta lösningar. Denna studie bekräftar tidigare studier att uppgifter med flera poäng riskerar att bli bedömda olika av olika bedömare.

En av de externa bedömarnas bedömning var inte alltid i linje med de övriga två eller med lärarna. Skillnaden på elevprovets totalpoäng och lärarnas bedömning var som störst 14 poäng och för de två andra externa bedömarna var det som störst 8 poäng för båda bedömarna.

I några tidigare undersökningar har de externa bedömarna diskuterat sina bedömningar och därefter fått bedöma på nytt. Olofsson (2006) visade att de skillnader i bedömningar som kvarstod efter diskussioner var relaterade till att bedömarna har olika formella krav på elevernas lösningar. Något sådant upplägg med diskussioner och ombedömningar har vi inte haft för denna undersökning. De externa bedömarna kände inte varandra och diskuterade inte sina bedömningar med varandra, som man gör vid medbedömning eller sambedömning för att få bedömningen mer likvärdig.

En aktuell fråga handlar om extern bedömning av de nationella proven, alltså att elevens lärare inte ska bedöma sina egna elevers prov utan det ska en extern bedömare göra. Data till denna studie ger möjlighet att få kunskap om hur många elever som i denna studie skulle få ett annat provbetyg än lärarens (och då antar vi att läraren är elevens lärare, vilket det inte alltid behöver vara). Vi jämför därför lärarens givna provbetyg med var och en av de externa bedömarnas. Antalet elever som inte får samma provbetyg av respektive extern bedömare som av läraren varierar naturligtvis och varia-tionen ligger mellan knappt tio till drygt 20 för de hundra elevproven. På en klass med cirka 25 elever skulle det betyda att mellan två och fem elever skulle få ett annat provbetyg som följd av de externa bedömarnas bedöming. Det intressanta är att om vi även låter lärarens bedömning ingå som en extern bedömare skulle variationen i stort sett vara densamma. Likvärdig-heten kanske bättre garanteras med medbedömning/sambedömning än med extern bedömning och då kan kravet vara att två lärare ska på poängnivå vara överens till minst 85 procent.

När det gäller bedömningsanvisningarna var det i stort sett bara en formu-lering som skulle kunna vara orsak till att bedömaröverensstämmelsen inte var så hög. Det var formuleringen att eleven skulle få poäng om de påbörjat en lösning eller ett resonemang. Även om det till så gott som alla dessa bedömningsanvisningar fanns exempel på bedömda elevlösningar så var bedömaröverensstämmelsen inte alltid tillfredställande. Formuleringen med att påbörja en lösning eller resonemang är ganska öppen och ger möjlighet för flera tolkningar. Även andra undersökningar (Lind Pantzare, 2015), som undersökte bedömaröverensstämmelse för gymnasieskolans skriftliga delprov, såg att formuleringen påbörjad lösning skulle kunna ställa till problem för bedömaröverensstämmelsen. Vad kan då göras? Om vi slutar med ”t.ex.” och istället försöker ange de varianter som vi sett i utprövningarna, får vi då säkrare bedömaröverensstämmelse? Nackdelen är att vi måste ange alla varianter som finns, vilket dels är omöjligt, dels leder till långa beskrivningar. Vi kan annars utesluta ovanliga varianter med följden att det finns en risk

att elever som löst uppgiften med en ovanlig strategi kan bli utan poäng om läraren tolkar att lösningen inte motsvarar beskrivningen i bedömnings-anvisningarna.

Denna studie visar att bedömaröverensstämmelsen är god på poängnivå, uppgiftsnivå och provbetygsnivå. När det gäller provbetygsnivå är det på individnivå ett problem med de strikta kravgränserna, att en poängs skillnad på totalnivå kan medföra att ett provbetyg ändras. Det talar för att det vore bättre att kravgränserna inte är så strikta utan att de ges som ungefärliga eller som ett intervall. De skiljaktigheter i bedömningarna som finns på uppgiftsnivå beror till en liten del på oklarheter i bedömningsanvisningarna och i en någon större del på olikheter i bedömarnas bedömningar där det finns ett tolkningsutrymme. Det är därför viktigt att lärare som ska bedöma elevers lösningar, oavsett om det är lärarens egna elever eller andra lärares elever, får möjlighet att diskutera olika elevlösningar och göra gemensamma bedömningar, men också att de får utbildning i bedömning av elevlösningar på uppgiftsnivå. Det skulle innebära att bedömaröverensstämmelsen skulle bli bättre och därigenom också likvärdigheten.

Referenser

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike (Ed.), *Educational Measurement.* 2nd ed., 508–560.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20 (37), 37-46.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin, 70(4)*, 213-220.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York: Holt, Reinhart and Winstorn.

Impara, J.C. & Plake, B.A. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement*, 34, 355–368.

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

Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorial data. *Biometrics, 33(1)*, 159.

Lie, S., Hopfenbeck, T.N., Ibsen, E. & Turmo, A. (2005). *Nasjonale prover på ny prove. Rapport fra en utvalgsundersökelse for å analysere og vurdere kvaliteten på oppgaver og resultater til nasjonale prover våren 2005*. Institutt for lärarutdanning og skoleutvikling. Universitetet i Oslo.

Lind Pantzare, A. (2015). *Bedömaröverensstämmelse på skriftliga delprov. En studie av interbedömarreliabiliteten vid bedömning av skriftliga nationella prov i matematik i gymnasieskolan*. Institutionen för tillämpad utbildningsvetenskap. Umeå universitet.

Lindström, J-O. (1998). *Rättvis rättning i nationella prov*. Pedagogiska mätningar nr 144. Umeå universitet.

Multon, K. (2010). Interrater reliability. In N. Salkind (Ed.), *Setting performance standards: Concepts, methods and perspectives* (pp. 627-629). Thousand Oaks: SAGE Publications, Inc.

Nordberg, C. & Pettersson, A: (2017). Ämnesprovet i matematik i årskurs 9, 2017. PRIM-gruppen. https://www.su.se/polopoly_fs/1.360637.1512640915!/menu/standard/file/Rapport%20%C3%84p9%202017.pdf

Olofsson, G. (2006). *Likvärdig bedömning? En studie av lärares bedömning av elevarbeten på nationella prov i matematik kurs A*. Rapport nr 23 från PRIM-gruppen. Lärarhögskolan i Stockholm.

Sireci, S.G., Hambleton, R.K. & Pitoniak, M.J. (2004). *Setting Passing Scores on Licensure Examinations Using Direct Consensus*. CLEAR Exam Review, 21–25.

Skolverket. (2009). *Bedömaröverensstämmelse vid bedömning av nationella prov. Bilaga till redovisning av regeringsuppdrag*. Dnr. U2009/1671/S.

Skolverket. (2017a). *Skolverkets systemramverk för nationella prov*. Skolverket.

Skolverket. (2017b). *Ämnesprov, läsår 2016/17. Matematik Årskurs 9. Bedömningsanvisningar 2. Delprov B, C och D*. Skolverket.

Sollerman, S. (2016). *Bedömaröverensstämmelse på muntliga prov. En studie av interbedömarreliabiliteten vid bedömning av nationella muntliga delprov i matematik gymnasieskolan*. Opublicerat manus. Stockholms universitet.

Sollerman, S. (2017). *Bedömaröverensstämmelse på muntliga prov. En studie av interbedömarreliabiliteten vid bedömning av nationella muntliga delprov i matematik årskurs 9*. Opublicerat manus. Stockholms universitet.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9* (4).

Rapporter i serien Matematikdidaktiska texter

L Alm, L Björklund Boistrup, T Englund (red), E-S Källgren, E Norén, G Olofsson, K.A Paulsson, A Petterson (red), T Tambour (red), B Ulin & T t Vehn. (2007). *Matematikdidaktiska texter, Beprövad erfarenhet och vetenskaplig grund.* Del 1.

T Englund (red), L Engström, I Ingemansson, N Larsson, K.A Paulsson, I O Persson, J Petterson, A Pettersson (red) & T Tambour (red). (2007). *Matematikdidaktiska texter, Beprövad erfarenhet och vetenskaplig grund.* Del 2.

I-M Parzyk. (2009). *Är det jag eller matteläraren som inte fattar?* Unga vuxna väletablerade och internermed utländs familjebakgrund berättar om matematikens betydelse för självkänslan i (skol)livet. Matematikdidaktiska texter. Del 3.

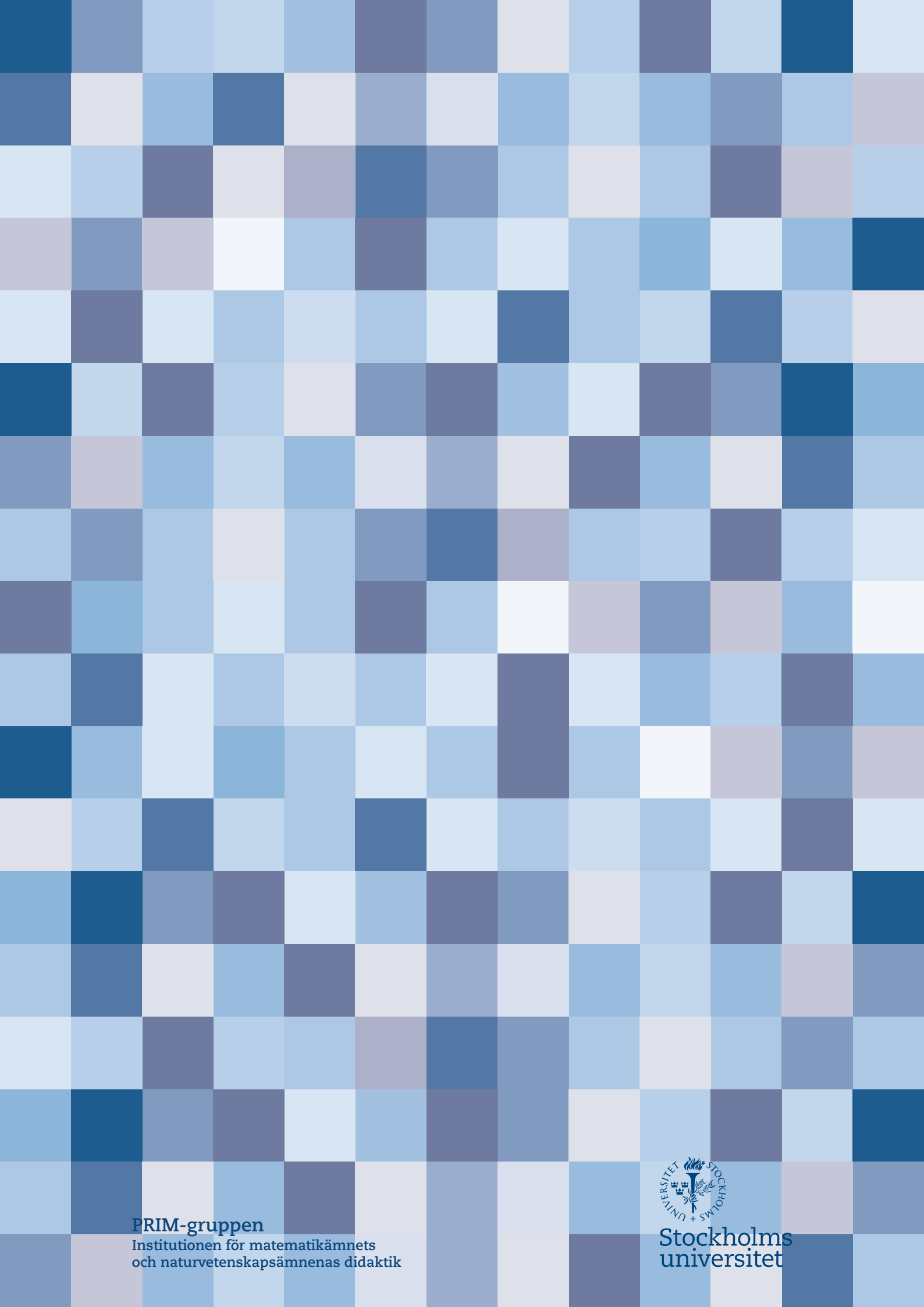
A Petterson, G Olofsson, K Kjellström, I Ingemansson, S Hallén, L Björklund Boistrup & L Alm. (2010). *Bedömning av kunskap – för lärande och undervisning i matematik.* Matematikdidaktiska texter, Beprövad erfarenhet och vetenskaplig grund. Del 4.

M Enoksson. (2014). *Innehåll i behov av särskilt stöd.* Erfarenheter från lesson/learningstudies i matematik. Matematikdidaktiska texter. Del 5.

I Ridderlind. (2014). *Elevperspektiv på bedömning för lärande.*Matematikdidaktiska texter. Del 6.

M Nordlund & A Petterson (red). (2019). *Bedömning i matematik – i lärandets och undervisningens tjänst.* Matematikdidaktiska texter, Beprövad erfarenhet och vetenskaplig grund. Del 7.

C Nordberg, A Pettersson & S Sollerman. (2021). *Bedömaröverensstämmelse vid bedömning av elevernas prestationer på de skriftliga delproven i 2017 års nationella proven i matematik för årskurs 9.* Matematikdidaktiska texter, Beprövad erfarenhet och vetenskaplig grund. Del 8.



PRIM-gruppen
Institutionen för matematikämnets
och naturvetenskapsämnenas didaktik

Stockholms universitet