Skolverket

Bilaga

2009-04-14
Dnr 2008:286


# Bedömaröverensstämmelse vid bedömning av nationella prov

# Innehåll

## Bakgrund och syfte

De nationella proven har flera syften. De ska konkretisera mål och betygskriterier, bidra till ökad måluppfyllelse bland eleverna, de ska hjälpa skolan, huvudmannen och staten i arbetet med uppföljning och de ska stödja en likvärdig och rättvis betygssättning.

Likvärdig och rättvis betygsättning är viktigt för elevens rättssäkerhet och det är varje skolas och huvudmans skyldighet att i sitt kvalitetssäkringsarbete eftersträva det. För att de nationella proven ska kunna bidra till att stödja en likvärdig och rättvis betygssättning förutsätts det emellertid att lärarnas bedömningar av elevers svar och prov sinsemellan visar god överensstämmelse.

Mot bakgrund av det nationella provsystemets stödjande funktion för en likvärdig och rättvis betygssättning i ett mål och kriterierelaterat betygssystem ser Skolverket, som ett led i sitt kvalitetssäkringsarbete, ett behov av att följa upp och utvärdera hur pass samstämmiga lärare är i sina bedömningar av elevers svar på de nationella proven. Då den forskning som tidigare har bedrivits på detta område till viss del har blivit inaktuell samt i flera fall även har utgjorts av förhållandevis svag empiri på grund av små urval har generaldirektör Per Thullberg i samband med verksamhetsuppföljningen hösten 2007 uttryckt en önskan om att Skolverket genomför en utvärdering av de nationella provens interbedömarreliabilitet.

Huvudsyftet med denna studie är således att analysera och illustrera samstämmigheten, eller bristen på samstämmighet, i bedömningar av elevers svar på de nationella proven. För ändamålet kommer analyser i engelska, matematik och svenska att göras.[1,2] Dessa analyser kommer sedan att jämföras så att de tillsammans ger en större och vidare bild av hur samstämmigheten i bedömning ser ut på de nationella proven samt hur de i förlängningen bidrar till en likvärdig och rättvis betygssättning. Syftet med denna studie är därmed att förutsättningslöst studera samstämmigheten mellan olika bedömare vid bedömning av de nationella proven.

[1] Svenska som andra språk har av urvalstekniska skäl exkluderats från denna studie.
[2] De muntliga delproven har exkluderats från denna studie i alla ämnen utom engelska. Inom ramen för engelska har en tentativ analys genomförts.

## Validitet, reliabilitet och interbedömarreliabilitet

### Validitet

Då man ämnar mäta eller på annat sätt samla in information om något är det viktigt att det man samlar in och vad det genererar är relevant för ändamålet. Om vi till exempel önskar veta på vilken kunskapsnivå inom ett specifikt område en grupp elever befinner sig är det av yttersta vikt att de test man konstruerar och sedermera använder sig av verkligen mäter det man ämnat mäta samt att de resultat eleverna erhåller går att använda så som det var tänkt. Om man lyckas med detta håller testet en hög validitet. Vid kunskapsmätningar brukar man betrakta begreppet validitet som ett övergripande begrepp[3] som utgörs av flera olika typer av validitet, nämligen innehålls-, kriterie-, begrepps-, face- och konsekvensvaliditet.[4] Begreppet innehållsvaliditet är av särskild betydelse för denna studie då detta begrepp avser hur väl innehållet på ett prov täcker de kunskaper och förmågor som avsetts mätas.

För att man ska kunna mäta en kunskap eller förmåga krävs en teoretisk definition av denna kunskap eller förmåga. När man etablerat en sådan behöver man operationalisera[5] denna teoretiska definition. Ett tests validitet kan uttryckas som relationen mellan den teoretiska och den operationella definitionen. I fallet med de nationella proven är det de nationellt angivna målen i kursplanerna som utgör de teoretiska definitionerna. Det är med andra ord kursplanernas mål som provkonstruktörerna ämnar operationalisera då de med hjälp av olika uppgifter och uppgiftsformat utformar proven.

Innan de kan operationalisera kursplanemålen behöver dock provkonstruktörerna tolka och konkretisera dem. Eftersom kursplanemålen lämnar förhållandevis mycket utrymme för tolkningar kan man förstå att olika provkonstruktörer kan komma att konstruera olika prov beroende på vilka tolkningar de gör av de kunskaper och förmågor som ska mätas. När dessa prov och tillhörande bedömningsanvisningar, vilka också är baserade på tolkningar av målen, bjuds tolkas de i nästa led av lärarna. Prov som på detta vis lämnar öppet för tolkningar i flera led riskerar att hålla en lägre validitet än ett prov som inte gör det.[6]

För att en kunskapsmätning ska ha hög validitet krävs det emellertid inte bara att vi de facto mäter det vi vill mäta, det krävs också att mätningen är reliabel. Det omvända gäller dock inte. En kunskapsmätning kan vara reliabel utan att för den skull vara valid.

---

[3] Messick, S. (1989). Validity. In R.L. Linn (Ed.), Educational measurement (3rd ed., pp. 13-103). New York: Macmillan

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

[5] Med operationalisering menas det sätt på vilket man mäter eller på annat sätt samlar in empiri om ett fenomen.

[6] Haertel & Herman 2005

### Reliabilitet

I de fall där man vill mäta eller på annat sätt samla in information om något är det viktigt att mätningarna som görs är stabila, robusta och att inslaget av slumpmässiga avvikelser är litet. Är de det säger man att mätningarna har hög reliabilitet. Traditionellt brukar man säga att mätningar som ger samma resultat gång på gång då de upprepas är reliabla. För att mätningarna ska kunna ge samma resultat gång på gång krävs det dock att mätningarna inte är behäftade med mätfel. Det vill säga att den enskilda mätningens resultat inte i allt för stor utsträckning påverkas av felkällor. I praktiken finns det dock ingen kunskapsmätning som inte är behäftad med någon form av mätproblematik. Det finns alltid slumpmässig variation som påverkar utfallet på en kunskapsmätning. Som exempel på sådana slumpmässiga felkällor av mera tillfällig karaktär kan nämnas hälsa, motivation, urvalet av uppgifter och stress som påverkar hur eleven presterar på ett visst test. Möjligheten att gissa rätt bidrar också till osäkerhet i resultaten. [7]

Fokus för denna studie är dock inte felkällor som återfinns på elevnivå, utan snarare på lärarnivå eftersom en potentiell felkälla ligger i själva bedömningsmomentet, ett moment som utförs av lärarna. Det är nämligen inte alls säkert att en lärare bedömer en elevs prestation på en uppgift eller ett prov på samma sätt som en annan lärare. Det är därför relevant att studera det som i litteraturen brukar kallas för interbedömarreliabilitet. Det vill säga den delen av en kunskapsmätnings reliabilitet som kan tillskrivas osäkerheten i själva bedömningen och som uttrycks som en variation mellan olika bedömares bedömning av samma elevsvar.

### Interbedömarreliabilitet

De första studierna som kom att behandla interbedömarreliabilitetsproblematiken var studier där man studerade hur två bedömares bedömningar stämde överens med varandra. Begreppet interbedömarreliabilitet har därför kommit att förknippas med samstämmighet i bedömning mellan två bedömare. I denna studie har vi därför valt att inte använda termen interbedömarreliabilitet utan det mer generella begreppet bedömaröverensstämmelse i och med att vi valt att inte enbart fokusera samstämmighet i bedömning mellan två bedömare.

Utrymme för godtycklighet i bedömning påverkar bedömaröverensstämmelsen. Det är emellertid inte enbart utrymmet för godtycklighet som påverkar, utan det finns även andra faktorer vilka har betydelse för vilket uttryck bedömaröverensstämmelsen tar. Lärarna som ska bedöma elevernas resultat på ett visst prov behöver till exempel inte nödvändigtvis ha samma definition av den kunskap som ska mätas som dem som konstruerat provet. Istället kan elevförmåga definieras olika beroende på vad läraren tolkar ska ingå i den berörda elevförmågan, det vill säga vilka olika element som tillsammans utgör den aktuella elevförmågan. Vidare kan det vara så att lärare tolkar kriterierna för de olika skalstegen på

[7] Se t.ex. Crocker & Algina (1986)

betygsskalan olika. Det vill säga de har olika synsätt på vad som krävs för olika prestationsnivåer. I fallet med de nationella proven utformas bedömningsanvisningarna utifrån intentionen att minska dessa utrymmen för godtycklighet i bedömning.

Olika bedömare kan också vara mer eller mindre benägna att använda sig av olika skalsteg vid bedömning. Det är alltså inte bara ”bredden” på betygskalans steg som påverkar bedömarreliabiliteten utan även lärares benägenhet att använda sig av dessa olika skalsteg. Ofta kan denna benägenhet ta sig olika uttryck[8], till exempel kan en lärare konsekvent lägga sig på den nedre delen av betygsskalan (severity error) medan en annan lärare lägger sig konsekvent på den övre delen (leniency error). Ett annat fel som en lärare kan göra är att inte använda hela skalan vid bedömningen, vilket får konsekvensen att de flesta betygen vid bedömning hamnar i mitten på skalan (centraliseringsfel). Lärarens personliga kontakt med och tidigare uppfattning om en elev kan också påverka lärarens bedömning av elevens prestation på ett prov. Denna felkälla i bedömningen brukar kallas för halo-effekten.

Vidare finns det möjlighet att en lärare som ska bedöma flera prov inte bedömer de första elevlösningarna på samma sätt som senare. Det är mycket möjligt att en lärares syn på hur bedömningen ska gå till förändras under bedömningsförloppets gång utan att läraren är eller blir medveten om detta. Det finns även faktorer som är mer av tillfällig art som kan påverka en lärares bedömning över tid. Till exempel kan trötthet påverka en lärares kriterier för bedömning. Hur pass konsistent en lärares bedömning är över tid brukar i litteraturen kallas för intrabedömarreliabilitet och även om intrabedömarreliabiliteten kan komma att påverka bedömaröverensstämmelsen ligger den utanför denna studies ram.

Hur pass hög eller låg bedömaröverensstämmelsen kommer att vara vid bedömning av ett test som till exempel ett nationellt prov påverkas således av flera faktorer. Det är därför önskvärt, vilket är denna studies syfte, att analysera bedömaröverenstämmelsen så som den tar sig i uttryck vid bedömningen av de nationella proven. Detta är emellertid inte så lätt då det kan vara svårt att särskilja vilka olika faktorer som påverkar utfallet. Man behöver därför göra vissa avgränsningar vad det gäller den empiri som ska användas samt bestämma sig för vilka analysmetoder som bör användas.

[8] Nitko & Brookhart 2007 - Educational Assessment of Students 5th ed.

## Metod och urval

### Metod

Det finns flera metoder för att studera bedömaröverensstämmelse. Man behöver därför vara på det klara med vad det är man vill veta utifrån sina analyser samt hur man vill kommunicera de påföljande resultaten. Ett bra tillvägagångssätt för att välja analysmetod är att låta studiens syfte styra valet av metod och för att underlätta valet av metod bör syftet vara väl preciserat. [9] Det teoretiska begreppet interbedömarreliabilitet är nämligen inte vid analys ett alldeles entydigt begrepp. Stemler (2004) menar att interbedömarreliabilitet som begrepp i själva verket innefattar tre typer av interbedömarreliabilitet vilka alla kräver olika metoder för att studera bedömaröverensstämmelse. Enligt Stemler kan interbedömarreliabilitet delas upp i de tre kategorierna, *consensus estimates*[10], *consistency estimates*[11] och *mesurement estimates* vid analys.

Syftet för denna studie är att analysera och illustrera vilka uttryck samstämmigheten, eller bristen på samstämmighet, i bedömning kan ta vid rättning av de nationella proven och därför används metoder vilka fokuserar konsensus och konsistens. De universitetsinstitutioner som konstruerar de nationella proven har på uppdrag av Skolverket genomfört de analyser och illustrationer som denna studie baseras på.

### Urval

Till de institutioner som ansvarar för de nationella proven skickas efter varje provtillfälle in ett antal kopierade elevlösningar av lärarna. I respektive delstudie av varje ämne har populationen utgjorts av de elever vars skriftliga lösningar på ingående delprov har skickats till provinstitutionen och som varit av sådan kvalitet att de kunnat kopieras upp i ytterligare exemplar. Ur denna uppsättning delprov har

---

[9] John Uebersax 2008

[10] Med consensus estimates menar Stemler de mått vilka ämnar mäta hur pass lika bedömare tillämpar bedömningsanvisningar. Det bakomliggande antagandet är att bedömare utifrån samma bedömningsanvisningar bör komma fram till exakt samma bedömning. Vanliga metoder för att analysera bedömaröverensstämmelse ur ett konsensusperspektiv är procentuell överensstämmelse och Cohens kappa.

[11] Consistency estimates utgår från att det inte är helt nödvändigt att bedömare tolkar och tillämpar bedömningsanvisningar lika. Istället syftar konsistensmåtten och dess metoder till att mäta hur pass konsistenta bedömare är i sina bedömningar. Till exempel om en lärare ger några elevlösningar betyget VG och en annan lärare ger exakt samma elevlösningar betyget G inser man att det inte råder konsensus lärarna emellan över hur man ska tillämpa betygsskalan. Däremot kan man säga att lärarna har varit konsistenta i sin bedömning då skillnaden i deras bedömningar på förhand verkar vara förutsägbar. Vanliga metoder för att analysera bedömaröverensstämmelse ur ett konsistensperspektiv är Pearsons korrelationskoefficient, Spearmans rangkorrelationskoefficient och Cronbachs alpha.

sedan respektive provinstitution slumpmässigt dragit 100 elevlösningar som bedömarna har gjort nya bedömningar av.

För studien har utöver ett urval av elevlösningar även ett urval av bedömare gjorts. Urvalet av bedömare har inte varit slumpmässigt, detta för att utöver de kvantitativa aspekterna av bedömaröverensstämmelse även möjliggöra analys av de kvalitativa aspekterna av bedömaröverensstämmelse som till exempel hur olika bedömarprofiler ser ut. Det bestämdes därför i studiens inledande skede att respektive provinstitution skulle anlita tre bedömare vilka alla hade olika bedömarbakgrund. Gruppen bedömare skulle utgöras av en person som ingår i provinstitutionens provgrupp, en lärare som ingår i eller har erfarenhet av provinstitutionens utprövning av provuppgifter samt en lärare med skolerfarenhet.

För att inte skapa någon förvirring kommer de som har genomfört ombedömningen fortsättningsvis att kallas för bedömare och elevens ordinarie rättande lärare för läraren eller lärarna. Bedömarna har inte vid tillfället för deras bedömning haft tillgång till lärarens bedömning och de varken känner eller har på annat sätt kommit i kontakt med eleverna. Viktigt att tänka på i de efterföljande avsnitten är också att de 100 eleverna inte har samma lärare. Det finns en stor möjlighet att elevernas lärare i den här studien består av 100 olika personer vilket gör det svårt att tolka lärarnas bedömningar visavi bedömarnas bedömningar av samma elevlösningar.

## Studiens resultat

De resultat och resonemang som redovisas under detta avsnitt är hämtade från de delstudier som respektive provinstitution genomfört.

I redovisningen som följer inleds varje ämnesavsnitt med en beskrivning av det nationella provet i ämnet samt hur bedömningen går till, dessa texter är hämtade från respektive delstudie. Sedan följer en resultatredovisning samt diskussion som baserar sig på forskarnas delstudier. Disposition och rubrikval följer dem som forskarna valt för sina delstudier.

## Engelska

Ämnesprovet i engelska för årskurs 9 består av tre delprov med fokus på kärnområden i kursplanen, nämligen muntlig interaktion och produktion (*Part A*), receptiv förmåga (*Part B*, uppdelad i *Part B 1*, inriktat mot läsförståelse, och *Part B 2* mot hörförståelse) samt skriftlig produktion och interaktion (Part C).

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

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

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

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

### *Part A - Det muntliga delprovet*

Det muntliga provet som genomförs i par eller i grupper om tre eller fyra elever har inte analyserats på samma sätt som övriga delprov i den här studien. För att i någon mån täcka även det muntliga provmomentet gavs provinstitutionen i engelska uppdraget att kortfattat beskriva hur bedömaröverensstämmelsen ser ut i deras ämne baserat på deras arbete med utprövning av nya uppgifter för kommande prov.

Forskarna har genomfört analyser för de tre senaste årens utfall av deras referensgruppers bedömningar. Man har kunnat se att vissa bedömare varit strängare och andra mildare i sin bedömning. Denna variation har bedömts vara måttlig och på det hela taget tycks bedömaröverensstämmelsen ha varit god då korrelationerna mellan bedömarna över dessa tre år i snitt har legat på strax över 0,90.

*Part B – Receptiv förmåga (Läs- och hörförståelse)*

Överensstämmelsen i bedömningarna anser forskarna generellt sett vara mycket god. Mellan bedömaren[12] och lärarna skiljer det endast i medeltal 0.94 poäng av totalt 90 då alla 100 elever har bedömts. I de allra flesta fall är överensstämmelsen total och de små differenser som finns härrör sig från några enstaka items eller uppgifter. Dessa uppgifter tycks ha det gemensamt att de är av uppgiftstypen *constructed response*, det vill säga uppgifter där eleven själv formulerar svaret.

Även analysen av de 20 elevsvar där ytterligare en bedömare har bedömt tyder på att samstämmigheten är mycket god. Av de totalt 90 poäng som en elev kunnat få har lärarna en genomsnittspoäng på 61,30 sett över dessa 20 elevlösningar. Bedömare 1 har 59,85 och bedömare 2 60,15. Då korrelationskoefficienter beräknas för lärarna, bedömare 1 och bedömare 2 i parvisa konstellationer ges det i alla fallen en korrelationskoefficient på över 0,99.

*Part C – Skriftlig produktion*

I part C skriftlig produktion har ett medelvärde baserat på den tiogradiga skalan beräknats för lärarna och bedömarna. Medelvärdena för de 100 på den tiogradiga skalan elevlösningarna är följande:

| Bedömare | Medel-värde | Standard-avvikelse |
|---|---|---|
| ***Lärargruppen*** | 5.59 | 2.437 |
| **Bed. 1** | 5.00 | 1.912 |
| **Bed. 2** | 5.38 | 2.187 |
| **Bed. 3** | 5.82 | 2.900 |

Skillnaden mellan det högsta och det lägsta medelvärdet är 0,82. Bedömare 3 verkar vara mildare i sin bedömning medan bedömare 1 verkar vara strängare. Går man vidare och tittar på hur bedömarna har fördelat betygen kan man se hur de har utnyttjat betygsskalan. Både genom att titta på standardavvikelsen, som är ett mått på spridning, samt de grafiska framställningarna[13] ser man, enligt forskarna, att det finns betydande skillnader i distributionen av betyg. Bedömare 1, som har det lägsta medelvärdet, är enligt forskarna ett tydligt exempel på centraltendens. Bedömare 3 däremot, som har det högsta medelvärdet, utnyttjar hela betygsskalan.

[12] Baserat på delprovets/uppgiftstypernas karaktär samt tidigare studier, där överensstämmelsen mellan bedömare vad gäller den typ av uppgifter som finns i *Part B* visat sig mycket god (Olsson-Wahlsten, 2002), bestämdes i samråd med Skolverket att ombedömning av de 100 + 100 elevhäftena, med fokus på läs- respektive hörförståelse, skulle göras av en person, och att ytterligare en bedömare skulle bedöma ett slumpmässigt draget, 20%-igt delurval. Den person som gjorde den totala ombedömningen (Bed 1) är en av de ansvariga för utvecklingen av ämnesprovet för åk 9, medan den andra (Bed 2) också arbetar i provprojektet, men inte aktivt med utveckling av provmaterial. Båda ombedömarna har lång lärarerfarenhet i engelska (Bed 1 också i tyska, Bed 2 i svenska och svenska som andraspråk). Båda bedömarna använde strikt de bedömningsanvisningar, inklusive exempel, som bifogas ämnesprovet.

[13] Tillhörande stapeldiagram går att finna i forskarnas delstudie.

Korrelationerna mellan bedömarna och lärarna sträcker sig från 0,86 till 0,93. Korrelationerna mellan bedömarna ligger något högre än korrelationerna mellan bedömarna och lärarna. Detta visar att bedömarna i stor utsträckning är samstämmiga i sina rangordningar av texter och på så sätt, enligt forskarna, mer konsekventa i sina bedömningar. Men det kan då vara värt att påminna om att de rättande lärarna är olika personer.

Vidare har forskarna för engelskan jämfört bedömarna och lärarnas bedömningar av de enskilda elevlösningarna för att se hur många av dessa som är helt eniga samt hur stor differens det kan finnas i bedömningarna.

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

Åtta av elevtexterna har bedömts identiskt av bedömarna och lärarna. Som mest skiljer det sig fyra betygssteg på den tiogradiga skalan, sex elevarbeten har bedömts så olika. Forskarna menar att det inte går att identifiera någon större systematik i vad det är för typer av elevtexter som bedömts olika. Dock tycks det vara så att bedömarna och lärarna är mer överens då elevlösningarna är av lägre kvalitet än då de är av högre kvalitet.

För att öka förståelsen för bedömningsprocessen, bad forskarna bedömarna att skriva ner sina reflektioner kring texter där bedömningen av något skäl fick dem att stanna upp och fundera extra mycket. Bedömarna använde sig olika mycket av möjligheten att skriva ner sina funderingar och kommentarer. Sammanlagt hade 57 elevarbeten kommenterats. I sex av dessa fall hade alla tre bedömarna skrivit ner kommentarer. Det som vållade bedömarna störst problem kan preliminärt summeras i fyra kategorier:

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

Efter att ombedömningen hade genomförts hade forskarna en träff med bedömarna där diskussionen kom att kretsa kring frågeställningar rörande dessa fyra kategorier. Utöver dessa frågeställningar hade även värdet av sambedömning lyfts av bedömarna vid detta tillfälle.

### *Sammanfattande reflektioner och slutsatser*

Forskarna inleder sina sammanfattande reflektioner och slutsatser med att konstatera att studiens resultat är intressanta både som indikation på graden av samstämmighet samt för de nationella provens vidareutveckling.

Bedömaröverensstämmelsen förefaller vara mycket god vad det gäller läs- och hörförståelseuppgifter. Det tycks vara så att uppgifterna fungerar för sitt syfte samt att bedömningsanvisningarna ger gott stöd för enhetlig bedömning enligt forskarna. De säger vidare att bedömaröverensstämmelsen även är god för den skriftliga produktionen. Trots detta menar de att dessa resultat på intet vis kan tolkas som att samstämmigheten i bedömning har nått sin högsta möjliga nivå.

Forskarna poängterar att det av validitetsskäl måste finnas texter och uppgifter som prövar elevernas förmåga till inferens, konklusion, reflektion och interaktion, trots att detta ibland ställer vissa krav på produktion. Studiens resultat visar emellertid att det är möjligt att upprätthålla en hög bedömaröverensstämmelse även för elevproducerade svar samtidigt som det är viktigt att beakta de problem för bedömning, som uppdagades i bedömarkommentarerna, i det fortsatta utvecklingsarbetet.

Huruvida bedömningen ska göras holistiskt eller analytiskt, det vill säga helhetligt eller utifrån särskilda aspekter, tycks vara en evig diskussion. En annan fråga av vikt inför det fortsatta utvecklingsarbetet är om i förväg nivåbestämda skalor och deskriptorer ska utvecklas och bedömningen göras i relation till dessa.[14] Dessutom är det inte bara antalet nivåer om är viktigt utan också hur dessa kommenteras i relation till, och därigenom tydliggör, kursplanernas mål, kriterier och de analytiska bedömningsfaktorerna.

Avslutningsvis konstaterar forskarna att studien på ett uppenbart sätt aktualiserar frågan om sambedömning. De skriver att eftersom samverkan kring bedömning både fyller funktionen att öka samstämmigheten bedömare emellan, och därmed ökar likvärdigheten, samt bidrar till lärares fördjupade tolkning av kursplanerna, förefaller det högst rimligt att den nuvarande rekommendationen om sambedömning förstärks. Vidare anser forskarna att frågor om bedömning måste ges stor uppmärksamhet i lärarutbildningen samt i arbetande lärares planeringsarbete. Till exempel skulle ytterligare material med kompetensutvecklande syfte kunna tillhandahållas inom ramen för det nationella provsystemet.

[14] Bachman & Palmer, 1996

**Matematik äp 9**

Ämnesproven i matematik för årskurs 9 består av tre delprov.

*Delprov A* är ett muntligt delprov. *Delprov B* består av två delar, Del B1 (kortsvar) och Del B2 (problemlösning). *Delprov C* prövar kunskaper från flera olika kunskapsområden, uppgifterna är samlade kring ett gemensamt tema. Delprov B och C finns i två versioner där enda skillnaden är att ingående tal är olika i en del av uppgifterna.

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

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

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

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

*Överensstämmelse mellan de beräknade provbetygen*

Eftersom denna studie endast omfattar de skriftliga delproven B och C har forskarna, med anledning av att det inte finns delprovsbetyg i matematik, enkom för denna studie tagits fram nya kravgränser för alla betygsnivåer baserade på dessa två delprov.[15]

För att se på hur bedömaröverensstämmelsen ser ut mellan de tre bedömarna och lärarna har parvisa jämförelser gjorts mellan varje elevs beräknade provbetyg.

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

För 86 procent av alla elevlösningar är det full samstämmighet i de beräknade provbetygen som baseras på bedömarna och lärarnas bedömningar. I sex fall är det lärarna som avviker från bedömarna, i fem av dessa fall har lärarna ”satt” MVG då bedömarna satt VG. Om man istället endast analyserar bedömarnas betygsättning finner man att överensstämmelsen är på 92 procent. Överensstämmelsen mellan bedömare 1 och 3 visar sig vara hela 98 procent.

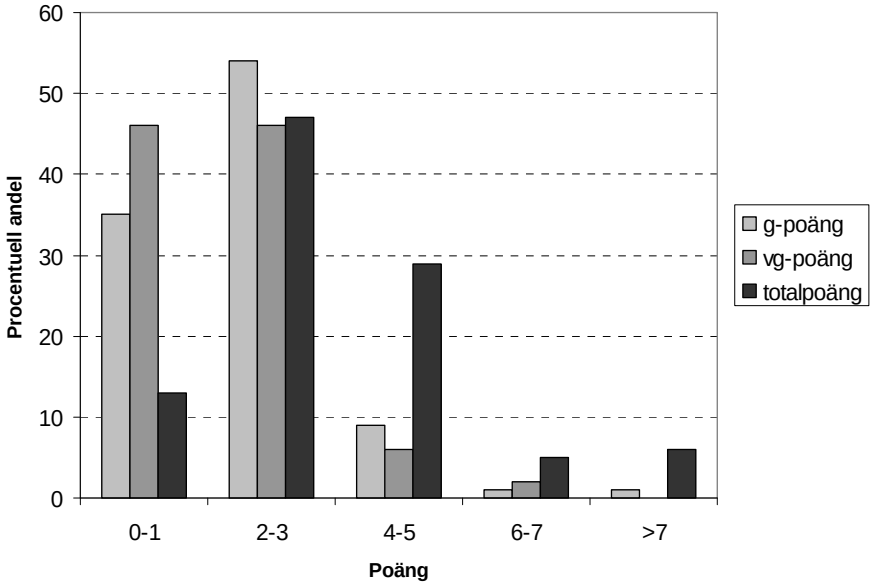
*Överensstämmelse mellan de beräknade summapoängen*

Forskarna konstaterar att korrelationen mellan totalpoängen är mycket hög oavsett vilket par av bedömare som korrelationen beräknats för. Även korrelationen mellan bedömarna och lärarna är hög, om än något lägre (nästintill försumbart lägre) än korrelationen mellan bedömarna. Korrelationerna mellan bedömarna ligger alla på 0,99 medan korrelationerna mellan bedömarna och lärarna sträcker sig från 0,96 till 0,97. Enligt forskarna finns den största överensstämmelsen mellan bedömarna och lärarna på g-poängen medan den lägsta finns på vg-poängen.

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

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

Utöver korrelationer har även variationsbredd som mått använts för att studera bedömaröverensstämmelsen. Trots att de två delproven kunde ge sammanlagt 67 poäng är variationsbredden oftast fem poäng eller mindre. Detta visar på en större överensstämmelse än vid den ombedömning som gjordes 2002.[16]




**Figur 1. Skillnad i poäng på delproven mellan bedömarna och lärarna.**

En jämförelse mellan bedömarnas och lärarnas satta totalpoäng på elevlösningarna visar att läraren i 32 av 100 fall satt högst poäng. Av bedömarna är det bedömare 2 som har bedömt strängast och bedömare 3 som bedömt minst strängt.

### *Resultat på uppgiftsnivå*

För att undersöka hur bedömaröverensstämmelsen ser ut för olika uppgiftstyper har forskarna gjort analyser ur ett uppgiftsperspektiv. I en första analys har de studerat hur de parvisa korrelationerna mellan bedömarna och lärarna fallit ut för bedömningen av delprov B1 innehållandes provuppgifter av typen kortsvar. Korrelationerna ligger storleksmässigt mellan 0,97 och 0,99.

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

Motsvarande analys har gjorts för delprov B2 vilket består av den så kallade aspektsbedömningsuppgiften. Även på detta delprov är korrelationen mycket hög mellan bedömarna, den är dock något lägre då man jämför bedömarna med lärarna, slår forskarma fast. Korrelationerna mellan bedömarna ligger mellan 0,91 och 0,95 och korrelationerna mellan de tre bedömarna och lärarna mellan 0,72 och 0,75.

Även delprov C bestående av uppgifter på temat blandad problemlösning visar på en hög bedömaröverensstämmelse. Korrelationerna mellan bedömarna och lärarna återfinns mellan 0,95 och 0,98. Den höga bedömaröverensstämmelsen till trots så finns det några uppgifter som orsakat mindre överensstämmelse än övriga. Störst skillnader i samstämmighet finns i bedömningen av uppgift 4a. Korrelationerna för denna uppgift är inte lika god som för delprov C i sin helhet men fortfarande mycket god menar forskarna. Korrelationerna mellan bedömarna ligger mellan 0,90 och 0,95 och korrelationerna mellan de tre bedömarna och lärarna mellan 0,79 och 0,83. En förklaring som forskarna anger till att uppgift 4a har den lägsta bedömaröverensstämmelsen i Delprov C är att formuleringen ”lämplig metod” lämnar ett visst tolkningsutrymme till läraren. Dessutom har olika lärare olika krav på vad en klar och tydlig redovisning innebär menar de.

### *Sammanfattande kommentarer*

Avslutningsvis slår forskarna fast att överensstämmelsen mellan bedömarna är mycket god. Anledningen till detta skulle kunna vara att bedömningsanvisningarna som följer med proven är lätta att tolka. Det skulle också kunna vara så att bedömarna har lång erfarenhet och att de varit med i referensgruppen/kravgränsmöten samt deltagit i bedömningskurser. De poängterar också att alla de tre bedömarna undervisar på skolor i Stockholmsområdet som har liknande elevunderlag. Enligt forskarna gör detta troligtvis att bedömarna har ungefär samma referensramar vid bedömning.

Vidare nämner de att bedömarna arbetat under för dem ovanliga förhållanden då de i normala fall har möjlighet att diskutera sin bedömning med kollegor. Bedömarna har också påtalat att de känt sig ensamma vid bedömningen. Tidigare erfarenheter har visat att möjligheten att diskutera sin bedömning med kollegor ger en mycket hög bedömaröverensstämmelse.[17]

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

## Matematik kurs 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. Ett särskilt formelblad är tillåtet hjälpmedel i båda provdelarna men miniräknare behöver och får endast användas i Del 2. 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.

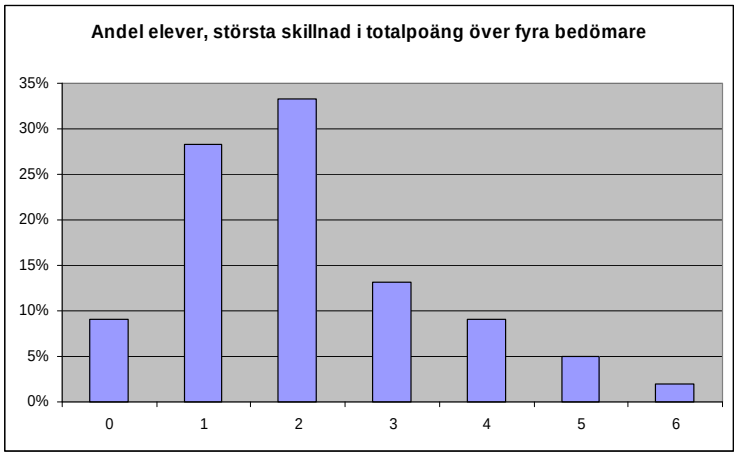
Av de 26 analyserade deluppgifterna är det sju uppgifter där endast svar fordras. I övriga uppgifter krävs att eleven redovisar sin lösning. Uppgifterna är en blandning av inommatematiska uppgifter och problemlösningsuppgifter med en kontext.

### *Procentuell överensstämmelse, korrelationer och variationsbredd*

För att se på hur bedömaröverensstämmelsen ser ut mellan bedömarna och lärarna har forskarna gjort parvisa jämförelser baserade på de erhållna provbetygen. Högst överensstämmelse i betygsättning med 90 procent absolut överensstämmelse uppvisar bedömare 1 och bedömare 2. Lägst överensstämmelse uppvisar bedömare 3 och lärarna med 81 procent. De korrelationer som har beräknats visar på samma bild. Högst korrelation finner man mellan bedömare 1 och bedömare 2, korrelationen dem emellan är 0,95. Den lägsta korrelationen, som är 0,84, finner man mellan bedömare 3 och lärarna.

Motsvarande analyser har även gjorts för provpoängen. För att möjliggöra dessa analyser har man valt att bredda kategorierna. Detta eftersom man har väldigt många kategorier, i detta fall 42. Eftersom en elev kan få 42 poäng kan man inte förvänta sig absolut överensstämmelse. För ändamålet har man därför låtit bedömarna ligga uppemot två poäng från varandra och ändå betraktat det som samma resultat. De parvisa jämförelserna mellan bedömarna sträcker sig från 90 till 92 procent. Korrelationerna som har beräknats på provpoängen visar även de samma bild som motsvarande analyser över procentuell överensstämmelse. Högst korrelation finner man mellan bedömare 1 och bedömare 2. De parvisa korrelationernas storlekar tyder dock på att skillnaderna dem emellan är försumbara då den lägsta korrelationen är på 0,98 och den högsta på 0,99.

Vidare har forskarna jämfört bedömarnas och lärarnas bedömningar av de enskilda elevlösningarna. Man har valt att titta på variationsbredden för att belysa hur olika en elev kan bli bedömd.




**Figur 2. S törsta skillnad i bedömd totalpoäng på en uppgift.**

Skillnaden i antalet poäng varierar mellan noll och sex poäng. För ca. 70 procent av elevlösningarna är differensen inte större än två poäng. Då man istället tittar på provbetygen finner man att bedömarna och lärarna är helt överens i 71 procent av alla elevlösningar. Bland de 77 elevlösningarna där bedömarna sinsemellan är helt överens finns det sju elevlösningar där lärarna avviker i sin bedömning. I fem av dessa fall har läraren gett ett högre provbetyg än bedömarna.

### *Uppgifter med eventuella reliabilitetsproblem*

De flesta provuppgifter har en absolut bedömaröverensstämmelse på 90 procent eller över. Det finns emellertid tre uppgifter som utmärker sig med en lägre överensstämmelse.

Uppgift 4 i provet är den som har lägst procentuell överensstämmelse, mellan 56,6 % och 78,8 %. Forskarna ger två möjliga förklaringar till varför denna utmärkt sig negativt. Den första är att uppgiften innehåller ett moment som innebär att eleverna förväntas verifiera det svar de kommer fram till. Då bedömningsanvisningarna inte föreskriver vilken metod för verifiering eleverna ska använda hanteras detta olika av olika lärare. Den andra tänkbara förklaringen är att vissa elever inte använt sig av den metod som de förväntas ha lärt sig under den aktuella kursen, utan löser uppgiften med ett mera allmänt resonemang som fungerar här eftersom uppgiftens placering i den miniräknarfria delen kräver enkla beräkningar och användning av heltal. Då bedömningsanvisningarna inte tar upp denna andra metod kan man tänka sig att lärarna hanterar detta olika.

Den andra uppgiften med vissa reliabilitetsproblem är uppgift 13 d. På grund av det matematiska innehållet i denna uppgift kan rätt slutresultat nås även med felaktig metod. I vissa fall har därför full poäng delats ut oavsett vilken metod som använts medan i andra fall, då felaktig metod har använts, inte någon poäng delats ut.

Den tredje uppgiften som uppvisat lägre bedömaröverensstämmelse är den så kallade aspektbedömningsuppgiften. Överensstämmelsen för vg-poängen är relativt bra anser forskarna, den ligger strax under 90 % för alla par av bedömare. Däremot är den procentuella överensstämmelsen något lägre för g-poängen. Den lägsta parvisa överensstämmelsen är 71,7 %. De elevlösningar som tycks orsaka problem för bedömarna är framförallt av tre slag, svar där eleven försöker sig på en lösning av uppgiften men inte för ett resonemang som exemplifieras i bedömningsanvisningen, svar som innehåller korrekta beräkningar utan slutsats, samt svar med felaktiga beräkningar men med godtagbar slutsats utifrån gjorda beräkningar. Dessutom finns vissa skillnader mellan bedömarna när det gäller att bedöma kvaliteten i elevernas redovisning och matematiska språk.

### *Slutsatser och diskussion*

Forskarna inleder sitt avslutande avsnitt med bedömningen, att studien visar på acceptabla eller till och med bra nivåer för bedömaröverensstämmelsen. Provets kortsvarsuppgifter har i princip bedömts enhetligt av dem som bedömt dem och de uppgifter som krävt längre lösningar eller resonerande svar från eleverna har i de flesta fall en procentuell överensstämmelse på över 90 procent.

Vidare säger forskarna att vissa skillnader i bedömning är naturliga då det finns en viss grad av frihet för läraren i sin bedömning. Eftersom lärarna förväntas göra sin bedömning utifrån den undervisning de bedriver så kommer skillnader i bedömning att förekomma. Denna brist i bedömaröverensstämmelse är dock inte att betrakta som ett interbedömarreliabilitetsproblem och skulle eventuellt kunna minskas med än mer styrande bedömningsanvisningar.

Störst problem för bedömaröverensstämmelsen verkar orsakas av uppgifter där eleven ska föra ett resonemang och där läraren ska bedöma elevernas språk. Lärarna har olika synsätt på hur elevlösningarna ska bedömas och således ges poäng. Detta kan ha att göra med hur läraren bedömer eleverna i sin undervisning. ”Har man inte lärt eleverna att göra ett korrekt bevis så är det naturligtvis svårt för eleven att prestera ett sådant på ett prov ” säger forskarna. Fast å andra sidan ”har man höga krav på eleverna så kommer det att avspegla sig även i bedömningen av de nationella proven” säger de vidare och konstaterar att detta torde generera en skillnad i bedömningen

I sin bedömning av provets bedömningsanvisningar säger forskarna att de generellt sett är relativt utförliga. Men för att uppnå en högre bedömaröverensstämmelse anger de som ett förslag, av totalt tre, att man kan specificera bedömningsanvisningarna ytterligare. Risken med detta är dock att man förmodligen begränsar antalet lösningsmetoder som undervisas. Det andra förslaget knyter an till det första om en ytterligare specificering av bedömningsanvisningarna. Bedömningsanvisningarna skulle, utöver att ange vilka kvaliteter som ska ge poäng, även kunna ange vad eleven inte får göra för att få poäng. Det tredje och sista förslaget riktar in sig på själva bedömningsmomentet, de föreslår träning i bedömning för lärarna. I de internationella komparativa studierna till exempel TIMSS och PISA anses träning av bedömarna vara en viktig

del för att öka bedömaröverensstämmelsen. Man föreslår därför att träningsmaterial till lärarna skulle kunna följa med bedömningsanvisningarna vid de nationella proven.

### Svenska

I ämnesprovet ingår tre delprov, nämligen A Att läsa och förstå, B Muntligt delprov och C Skriftligt delprov. De tre delproven hålls samman av ett gemensamt tema som för det aktuella läsåret (vt 2007) var *Vidgade vyer*. Elevens lärare bedömer de tre olika delarna var för sig och väger sedan ihop dem till ett provbetyg enligt Skolverkets anvisningar.

Ämnesprovet innehåller omfattande bedömningsanvisningar med allmänna resonemang om bedömning i relation till kursplanen, principer för bedömningen av de olika uppgifterna samt kommenterade elevexempel på olika kvalitativa nivåer.

I delprov A prövas elevens förmåga att läsa och förstå texter av olika slag – tidningsartiklar, faktatexter, dikter och skönlitterär prosa. Till lärarnas hjälp för bedömningen finns ett detaljerat material där varje fråga gås igenom. Först ges information om vilket innehåll provkonstruktörerna tänkt att svaret ska ha. Därefter visas exempel på autentiska elevsvar på olika nivåer. Här betonas att det är innehållet och inte den språkliga utformningen som bedöms i delprov A.

För det skriftliga delprovet, delprov C, utgörs bedömningsunderlaget av ett kvalitetsschema med kriterier för kvalitet i texter, dels av elevtexter med kommentarer. Kvalitetsschemat används för att göra en helhetsbedömning av elevens text, där hänsyn tas till de fem olika områden som schemat fokuserar: kommunikativ kvalitet, innehållslig kvalitet, sammanhang och uppläggning, språklig kvalitet samt skrivregler. För varje skrivuppgift finns en beskrivning av uppgiften samt bedömda och analyserade elevtexter på alla betygsnivåer.

För att ge lärarna möjlighet att nyansera sina betyg används vid de enskilda delproven och i det sammanvägda provbetyget en tiogradig skala av samma typ som den som används i engelska. I den skalan motsvaras Godkänt av stegen 3, 4 och 5 för att skilja mellan svaga, ordinära och starka prestationer. Motsvarande för Väl godkänt är 6, 7 och 8. Mycket väl godkänt kan nyanseras i 9 och 10. De lösningar som ej nått målen (EUM) kan graderas i 1 och 2, där 2 ligger nära gränsen för Godkänt.

#### *Delprov A: Att läsa och förstå*

För de 100 elevlösningarna på delprov A är det lärarna som har det högsta medelbetyget, 5,5 beräknat utifrån den tiogradiga betygsskalan. Bedömarna ligger lite lägre, mellan 4,9 och 5,2. Av de 100 lösningarna är det 72 som på den fyrgradiga betygsskalan fått samma betyg av lärarna och bedömarna. På den tiogradiga skalan sträcker sig korrelationen bedömarna emellan från 0,80 till 0,86.

| Bedömare | Medelvärde | Standardavvikelse |
|---|---|---|
| Lärargruppen | 5,5 | * |
| Bedömare 1 | 4,9 | 2,13 |
| Bedömare 2 | 5,2 | 2,25 |
| Bedömare 3 | 5,1 | 2,13 |

***Lärarnas bedömningar har inte funnits att tillgå i den tiogradiga skalan.**

Bedömare 1 som tycks vara den mest stränga bedömaren har inte använt betygen 9 och 10, det vill säga MVG och Bedömare 2 som har högst medelvärde sprider sina betyg över hela skalan. Trots detta så skiljer sig inte standardavvikelsen åt nämnvärt mellan dessa två då bedömare 1 har gett betyget EUM i betydligt större utsträckning än bedömare 2.

Det är ett fåtal uppgifter som utmärker sig med låg överensstämmelse mellan bedömarna. Dessa består främst av uppgifter vilka kräver utförligare svar så som förklaringar, exempel eller flera aspekter (till exempel både förklaring och motivering).

Bedömare 1 nämnde en av frågorna som extra besvärlig. Hon sa att hon verkligen saknade möjligheten att diskutera med kollegor när tveksamheter uppstod.

*Delprov C: Skrivuppgift*

Även för de 100 elevlösningarna på det skriftliga delprovet är det lärarna som har det högsta medelbetyget, 5,5 beräknat utifrån den tiogradiga betygsskalan. Bedömarna ligger också här lite lägre, mellan 4,1 och 5,2. Av de 100 lösningarna är det 54 som på den fyrgradiga betygsskalan fått samma betyg av lärarna och bedömarna. På den tiogradiga skalan sträcker sig korrelationen bedömarna emellan från 0,36 till 0,46.

| Bedömare | Medelvärde | Standardavvikelse |
|---|---|---|
| Lärargruppen | 5,5 | * |
| Bedömare 1 | 4,1 | 1,96 |
| Bedömare 2 | 5,2 | 2,39 |
| Bedömare 3 | 4,5 | 1,67 |

***Lärarnas bedömningar har inte funnits att tillgå i den tiogradiga skalan.**

Liksom för delprov A är det bedömare 1 som i genomsnitt har varit strängast i sin bedömning och bedömare 2 som varit mildast. Bedömare 2 har återigen i störst utsträckning använt sig av hela skalan, vilket man också ser genom att titta på bedömare 2:s standardavvikelse visavi de andra bedömarnas. Bedömare 3 har på det skriftliga delprovet i större utsträckning än de andra centrerat sina betyg.

| **Steg**[18] | **0** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Antal** | 1 | 23 | 35 | 16 | 10 | 8 | 4 | 2 | 0 | 1 | [m=2.7] |

Av de 24 texter där bedömarna varit mest överens är det hela 19 stycken som på den fyrgradiga betygsskalan har fått betyget G. Detta, att bedömare är mer överens om de lägsta betygen, har visats i tidigare studier.[19] Bland de 15 elevlösningar där det råder störst diskrepans mellan bedömarnas bedömningar ligger bedömare 1 i tio av fallen lägst.

Forskarna bedömer vidare det som sannolikt att en del av bedömarvariationen hör ihop med vilka genrekrav bedömarna betonar vid bedömningen. I delprov C kan eleverna välja mellan fyra skrivuppgifter med olika genreangivelse och de ger härvid exempel på hur krönikan, som är en ganska ny genre inom proven, visar sig vara svårbedömd. De 21 inskickade krönikorna får i genomsnitt 5,9 på den tiogradiga skalan av de egna lärarna medan bedömarna i snitt har satt 4,5.

### *Bedömarprofiler*

De tre bedömarna i denna studie uppvisar olika bedömarprofiler. Bedömare 1 är den strängaste av de tre. Hon tyckte att det var svårt att genomföra bedömningen då hon inte hade någon att resonera med. Bedömare 2, som ligger närmast de av lärarna inskickade betygen, är den som utnyttjar hela betygsskalan. Hon betonar helhetsbedömningen för den skriftliga produktionen samt hur eleverna lyckats med genreträffen. Bedömare 3 tenderar att centrera betygen, eller som forskarna säger det, hon ”överanvänder” betygssteget G. Hon har likt bedömare 1 gett uttryck för att det hade varit bra om man hade fått diskutera svårbedömda fall med någon.

### *Diskussion*

Lärarna gör i genomsnitt en mer positiv bedömning än bedömarna för både delprov A och delprov C. Forskarna tror att det beror på att lärarna förmodligen väger in övrig kännedom om vad eleven kan och vad hon presterat på övriga delprov. De säger också att lärarna kanske betygsätter proven med tanke på det stundande slutbetyget samt bättre förstår vad eleven menar ”när uttryckssättet ibland är suddigt och tankegången oklar”[20] då ett tests pålitlighet är låg när det är enstaka elevtexter som bedöms.[21] Vidare säger forskarna att en förklaring till bedömningsdifferensen mellan lärarna och bedömarna kan bero på en ”hellre fria än fälla”-inställning från lärarna. Det kan dock också vara så att bedömarna vill visa på att de minsann vet vad som krävs för de olika betygen menar forskarna.

---

[18] Följande tabell är inte direkt jämförbar med den för engelskans skriftliga delprov, matematik äp9 eller matematik kurs C då lärarna inte tagits med i svenskans beräkningar.

[19] Till exempel Berge 1996 och Östlund-Stjärnegårdh 2002.

[20] Björnsson, C.H., 1960: Uppsatsbedömning och uppsatsskrivning. Stockholm.

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

Forskarna bedömer det som ett måste att, efter denna studies genomförande, diskutera frågeformaten. Kan en ytterligare skärpning av uppgiftsformaten och bedömningsanvisningarna ge en större bedömaröverensstämmelse frågar de sig. Samtidigt konstaterar de att det behövs olika uppgiftsformat som mäter olika aspekter och som är av olika svårighetsgrad då det lättbedömbara inte ger tillräckligt besked om elevens kunskaper och förmågor.

I sin diskussion väljer forskarna, för att sätta in denna studies resultat i ett vidare perspektiv, att göra en internationell utblick. Man tar därför avstamp i den redovisning av bedömaröverensstämmelsemått som redovisats inom ramen för det så kallade KAL-projektet.[22] Forskarna konstaterar, utifrån KAL-projektets redovisning, att denna studies resultat på den skriftliga produktionen är lågt. De menar att en korrelation på 0,6 hade varit att förvänta istället för den på ca 0,4 som utfallet blev. Störst samstämmighet uppträder i de storskaliga internationella undersökningarna där bedömarna är specialtränade.

Emellertid är det så att forskarna vid tidigare ombedömning av ämnesprovet i årskurs 9 erhållit betydligt större bedömaröverensstämmelse. När skrivuppgifterna har diskuterats vid bedömargruppsmöten och lärarna som ingått i bedömargruppen satt sina betyg oberoende av varandra, har bedömaröverensstämmelsen varit högre än i den här studien. Vid det senaste bedömarmötet för ÄP9 hade varje bedömare 12 texter att betygsätta. Korrelationen mellan de i gruppen ingående lärarna låg då mellan 0,75 och 0,92. I ljuset av detta, menar forskarna, framstår samstämmigheten för delprov C som förvånande låg.

En möjlig föklaring till den låga bedömaröverensstämmelsen som forskarna lyfter är att de nationella proven i större utsträckning än tidigare tryckt på genrekunskap och specifika genredrag. Enskilda lärare betonar genreaspekten olika och krönikan tycks vara den mest problematiska genren om man ser till bedömaröverensstämmelsen. De menar att provkonstruktörerna efter denna studies resultat måste disktuera vilka genrer som ska vara med på det skriftliga provet i fortsättningen.

En annan möjlig förklaring som forskarna nämner är antalet elevlösningar som bedömarna fått ombedöma. Så många som 100 stycken är ingen svensklärare van vid. Det skulle vara intressant att göra om studien med färre texter fast med fler bedömare menar de.

Avslutningsvis konstaterar forskarna att sambedömning är oerhört viktigt, särskilt för det skriftliga delprovet. Skolverket borde, enligt forskarna, föreskriva sambedömning, inte enbart rekommendera det. Vidare anser de att lärare överlag behöver träning i bedömning. Kanske ett framtida nationellt provsystem ska tillhandahålla träningsmaterial i bedömning för lärarna?

---

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

## Diskussion

I tre av de fyra delstudier denna rapport grundar sig på värderar forskarna sina resultat som att bedömaröverensstämmelsen är god eller till och med mycket god. Den delstudie som avviker från detta resultatmönster är den som är genomförd inom ramen för ämnet svenska. Forskarna i svenska säger att resultaten är lägre än förväntat och att de är lägre än vad de tidigare erfarit utifrån praktiken med utvecklingen av de nationella proven.

Utifrån den resultatbild som forskarna till största del värderat positivt infinner sig emellertid ett antal frågor. Hur ska man se på den erhållna bedömaröverensstämmelsen, dels i sig självt dels som del av de nationella provens reliabilitet och i förlängningen validitet? I vilken utsträckning fungerar de som stöd för likvärdig och rättvis betygsättning? Hur är det möjligt att i det fortsatta arbetet förbättra bedömaröverensstämmelsen för de nationella proven?

### Validitet och reliabilitet – är bedömaröverensstämmelsen god?

Forskning har visat att de uppgiftsformat som ger minst utrymme för godtycklighet i bedömning och därmed ger högst bedömaröverensstämmelse är flervalsfrågor. I andra änden av skalan befinner sig de uppgiftsformat som ger mer utrymme till godtycklighet i bedömning, dessa är i regel de uppgiftsformat som är mera omfattande och som bedöms genom holistisk bedömning.[23] Denna studies resultat stödjer denna bild av uppgiftsformat visavi bedömaröverensstämmelse. Till exempel så visar studien på att flervalsfrågor med strikta bedömningsanvisningar har en högre bedömaröverensstämmelse än mer öppna uppgifter så som aspektbedömningsuppgiften i matematik eller den skriftliga produktionen i svenska där man tillämpar holistisk bedömning.

Hur hög ska en bedömaröverensstämmelse vara för att betraktas som god? Har vi en hög eller låg bedömaröverensstämmelse i denna studie? Majoriteten av de forskare som varit anlitade för den här studien bedömer att bedömaröverensstämmelsen för deras ämne är god. Men vad som är en god bedömaröverensstämmelse är inte av naturen given, det beror helt på hur man värderar de resultat som den kvantitativa analysen ger. Ur ett strikt reliabilitetsperspektiv kan man säga att ett prov med en högre bedömaröverensstämmelse är bättre än ett prov med lägre bedömaröverensstämmelse allt annat lika. I alla situationer där bedömare i någon utsträckning är involverade är det viktigt att bedömaröverensstämmelsen är så hög som möjligt i och med att bedömaröverensstämmelsen har betydelse för de bedömda elevresultatens validitet.[24]

Om man endast skulle vilja fokusera reliabilitetsaspekten av ett prov skulle man kunna använda sig uteslutande av enklare lättbedömda uppgiftsformat i och med

---

[23] Gipps 1994
[24] Stemler 2004

att de leder till en högre bedömaröverensstämmelse. Men ett sådant förfarande skulle inte vara förenligt med kursplanernas intentioner då de lättbedömda uppgiftsformaten inte mäter de mera kvalificerade kunskaper och förmågor som kursplanerna anger som mål. Därför behöver man använda olika uppgiftsformat som mäter olika aspekter och som är av olika svårighetsgrad för att få tillräcklig information om elevens kunskaper och förmågor. Vid utvecklingen av provuppgifter måste hela tiden målet vara att de ska vara av mycket hög kvalitet och ha bedömningsanvisningar som stöder lärarna så att bedömningen överstämmer i så hög grad som möjligt.

Hur hög behöver då bedömaröverensstämmelsen vara för att man ska betrakta den som tillräckligt god? Hur långt kan man gå i att strama upp uppgiftsformaten för att öka på reliabiliteten innan det stjälper provets validitet genom ett begränsande av dess innehållsvaliditet. Det nationella provuppdraget[25] så som det är utformat ger här ingen vägledning. Vid en tillbakablick kan man emellertid utläsa hur synen förändrats över tid. I mitten av 1990-talet då det nuvarande nationella provsystemet sjösattes betonades validiteten och den lokala bedömningen för att över tid allt mer fokusera på reliabilitet och nationell likvärdighet.[26]

I denna rapports inledande avsnitt nämndes det att ett provs reliabilitet påverkar dess validitet. Man bör därför vid konstruktionen av ett prov sträva efter att ha en så hög reliabilitet som möjligt. Man skulle ur detta perspektiv kunna säga att det är önskvärt att ha en så hög reliabilitet som möjligt givet att man också har en hög validitet. Anledningen till detta är att validiteten är att betrakta som överordnad reliabiliteten vid kunskapsmätningar, för om provet inte mäter det man avsett det att mäta spelar det inte så stor roll om det är reliabelt eller ej. Därför kan uppgiftsformat vilka uppvisar en låg bedömaröverensstämmelse men som motiveras genom hög innehållsvaliditet vara nödvändiga.

### Stöd för likvärdig och rättvis betygsättning

När regeringen gav Skolverket det nya provuppdraget 1994 var det framförallt de nationella provens stödjande funktion som betonades. De nationella proven skulle i första hand syfta till att vara ett stöd för likvärdig betygsättning i ett system med stor lokal frihet att tolka mål och betygskriterier samt att välja innehåll i undervisningen.[27] Men efterhand kom det nuvarande provsystemet att tillskrivas ytterligare syften som att främja nationell likvärdighet samt vara underlag för kvalitetssäkring och resultatkontroll.[28] Som Skolverket tidigare har påpekat är det tveksamt om dessa senare tillkomna syften är förenliga med de pedagogiskt stödjande och utvecklande syftena som betonades initialt.[29] En ökad betoning av provresultat på lärarens samlade bedömningar innebär att provets egenskaper och bedömningen av elevens provresultat getts ökad uppmärksamhet. Det är också i

[25] Uppdrag till Statens skolverk avseende det nationella provsystemet 2004
[26] Lundahl, C. (2009) Varför nationella prov? Lund: Studentlitteratur (under utgivning)
[27] Bilaga till regeringsbeslut 1994-04-21, dnr U94/1031/Gru
[28] Lundahl, C. (2009) Varför nationella prov? Lund: Studentlitteratur (under utgivning)
[29] Skolverket 2003 – Det nationella provsystemet

det sammanhanget som den ökade uppmärksamheten på bedömaröverensstämmelse ska ses.

Resultaten i svenska är mot denna bakgrund inte så lätta att förhålla sig till. Forskarna i svenska säger i anslutning till analyserna av den skriftiga produktionen, vilken i den här studien visar på den lägsta bedömaröverensstämmelsen, att likvärdig bedömning av en enskild text är svårt. För att säkra likvärdigheten bör varje elev skriva flera texter i olika genrer. Men trots de lägre resultaten för bedömaröverensstämmelse så kan det nationella provet även i svenska uppfylla sitt syfte om att vara ett stöd för likvärdig och rättvis betygsättning, om de nationella proven bidrar till att lärare emellan diskuterar bedömningsfrågor. På så sätt hamnar bedömningsfrågor på lärarens agenda och kan därigenom fungera som stöd för likvärdig och rättvis betygsättning.

Bedömarna har inom ramen för denna studie fått bedöma elevlösningarna på egen hand. Flertalet av bedömarna har dock under studiens genomförande påtalat att de känt sig ensamma i bedömningsfasen då de i normala fall samråder med kollegor. Under dessa omständigheter är det rimligt att anta att bedömaröverensstämmelsen, och därmed likvärdigheten i bedömning, är lägre än vad den hade varit om det varit möjligt att genomföra denna studie som en för bedömarna autentisk bedömningssituation.

### Sambedömning

Forskarna i denna studie är alla överens om att det vore önskvärt med sambedömning vid bedömningen av de nationella proven. De har i flertalet fall uttryckt att det vore önskvärt att Skolverket föreskrev om obligatorisk sambedömning. Framförallt skulle förmodligen de mer svårbedömda uppgiftsformaten få en högre bedömaröverensstämmelse om lärarna utövade sambedömning.

Bedömarna har som tidigare nämnts uttryckt att de känt sig ensamma i den bedömning de gjort inom ramen för detta arbete. De säger att de till vardags rådgör med kollegor i bedömningen, åtminstone då det gäller svårbedömda fall. Tidigare studier indikerar att detta är ett vanligt sätt att förhålla sig till bedömningen av de nationella proven.[30] I och med att möjligheten till att gemensamt diskutera och problematisera bedömningar förmodligen ger en ökad likvärdighet i bedömning finns det skäl att tro att bedömaröverensstämmelsen hade varit högre om det funnits möjlighet att rådgöra med kollegor. Möjligtvis har bedömningen i svenska blivit extra lidande av bristen på sambedömning. De öppna uppgiftsformaten och de längre krävande uppgifterna som bedöms genom holistisk bedömning torde vara mera utsatta än andra uppgiftsformat som har mer strama bedömningsanvisningar[31] där möjligheten till sambedömning inte är aktuell.

Det är angeläget att sambedömningen av de nationella proven ökar, både inom och mellan skolor. Det gäller i synnerhet de delar av proven där man behöver göra en

[30] Ämnesprovet 2008 i grundskolans åk 9 och specialskolans åk 10 - En resultatredovisning
[31] Gipps 1994

mer holistisk bedömning. Skolverket behöver se över möjligheten att föreskriva om sambedömning av de nationella proven. En föreskrift om obligatorisk sambedömning skulle även kunna bidra till att bedömningsfrågor hamnar högre upp på lärarens agenda. Det skulle också kunna stimulera till ett ämnesdidaktiskt samtal i vidare mening på skolan genom att i ökad utsträckning bidra till att det ute på skolorna diskuteras vad det är för kunskaper och förmågor de nationella proven prövar samt hur dessa kunskaper och förmågor är konstituerade. Detta positiva mervärde av obligatorisk sambedömning ska dock ställas mot att lärarens arbetsbelastning skulle kunna komma att öka.

### Det kontinuerliga utvecklingsarbetet

Om bedömaröverensstämmelsen är låg kan det antingen vara bedömarna eller bedömningsanvisningarna som brister. Oavsett om det är bedömarna som brister i sin bedömningsförmåga eller om det är bedömningsanvisningarna som inte lever upp till vad som förväntas av dem så bör man i det kontinuerliga utvecklingsarbetet arbeta för att ytterligare förbättra uppgiftsformaten, bedömningsanvisningarna samt även öka lärarnas bedömarkompetens.

Flera av forskarna tar upp möjligheten till förbättringar och ytterliga specificeringar av bedömningsanvisningarna. Provkonstruktörerna bör således även fortsatt sträva efter att förbättra de nationella provens bedömningsanvisningar samt kontinuerligt utvärdera dem. I detta utvecklingsarbete med bedömningsanvisningarna finns det skäl att prioritera det fortsatta arbetet med utveckling av *scoring rubrics* för att minska diskrepansen i bedömning mellan lärare.[32] Huruvida det är analytiska eller holistiska scoring rubrics[33] som ska utvecklas vidare är emellertid en fråga för provkonstruktörerna då det är de som har i uppdrag att tolka kursplanerna och utifrån dem utveckla proven.

I detta sammanhang bör man också tillägga att bedömningsanvisningarna kommer att påverkas av de nya kursplaner som ska tas fram och träda i kraft 2011/12. I regeringens proposition (2008/09:87) är uppdraget till Skolverket att föreslå tydligare mål och kunskapskrav. En nyhet är angivelser av ett centralt innehåll i varje ämne. Detta bör leda till större möjligheter för provkonstruktörerna att göra än tydligare bedömningsanvisningar till de nationella proven.

Om man tar avstamp i den här studiens resultat kan man vidare säga att det finns skäl till att erbjuda lärarna träning i bedömning i någon form då forskningen har visat att bedömaröverensstämmelsen ökar om lärarna genomgår träning.[34] Det finns i huvudsak tre möjliga tillvägagångssätt för träning i bedömning för lärarna. Det första är att lärarutbildningarna i större utsträckning än tidigare uppmärksammar bedömningsproblematiken.[35] Det andra tillvägagångssättet inbegriper det nationella provet och dess bedömningsanvisningar. Olika

[32] Moskal & Leydens 2000 - Scoring Rubric Development: Validity and Reliability
[33] Barbara M. Moskal 2000 – Scoring Rubrics: What, When and How?
[34] Se t.ex. Gipps 1994 och Alderson m.fl. 1995
[35] SOU 2008:109 En hållbar lärarutbildning

träningsmaterial skulle till exempel kunna inkluderas i de utskick till skolorna som görs inför de nationella proven. Dessa träningsmaterial skulle då vara avsedda att användas före genomförandet av bedömningen av de nationella proven för att på så sätt höja lärarnas bedömningskompetens. Det tredje sättet för att få tillstånd en träning i bedömning är att nationella kompetensutvecklingsinsatser initieras. Sådana insatser skulle till exempel kunna bestå av interaktiva kurser på webben.

### Konsekvenser av undersökningens design

Avslutningsvis bör det påpekas att denna studie är behäftad med vissa problem som troligtvis begränsar möjligheten till generaliseringar. För det första är antalet bedömare i denna studie lågt, endast tre stycken per delstudie. För det andra är dessa inte dragna som ett slumpmässigt obundet urval. Urvalet bedömare i denna studie är därmed inte att betrakta som representativt för lärarpopulationen.

Studiens resultat ligger emellertid i linje med vad tidigare internationell[36] och nationell[37] forskning visat även om resultaten i ämnet svenska ligger lägre än förväntat. Detta utfall, att det är lägre än förväntat, ska dock tolkas med försiktighet då detta kan vara ett utslag av studiens utformning.

Det är också viktigt att poängtera att den bedömningssituation de utvalda/anlitade bedömarna står inför inte är autentisk. Bedömarna är vid genomförandet av bedömningen medvetna om att deras rättning kommer att studeras och analyseras. Den icke-autentiska bedömningssituationen leder därmed förmodligen till att bedömarna bedömer mer noggrant och därmed mer strikt jämfört med hur de skulle ha bedömt i vanliga fall. Eller som forskarna i svenska uttrycker det, man ska minsann visa vad som krävs för olika betyg.

Vidare kan man anta att det inte är vilka lärare som helst som tar sig an ett sådant här bedömningsuppdrag. Man kan argumentera för att dessa lärare förmodligen har ett större intresse och engagemang för bedömningsfrågor. Denna positiva selektion av lärare kan leda till selektionseffekter vilka påverkar utfallet vid bedömarstudier.

Som redan nämnts så har bedömarna i den här studien uttryckt att de känt sig ensamma i sin bedömning då de till vardags i viss utsträckning genomför bedömningen tillsammans med kollegor. I första hand är det i svårbedömda fall de säger sig ta hjälp av kollegor för att diskutera bedömningen. Denna brist på möjlighet att rådgöra bör sannolikt ha påverkat bedömaröverensstämmelsen i denna studie i negativ riktning.

Vidare ska det nämnas att bedömarna inte är vana vid att bedöma så många som 100 elevlösningar. Det höga antalet elevlösningar gör att bedömarna med största sannolikhet har haft svårt att hålla en hög intrabedömarreliabilitet genom hela bedömningen vilket kan ha påverkat denna studies bedömaröverensstämmelse negativt.

---

[36] Se t.ex. Gipps 1994 och Berge, Kjell Lars, 2005: Skriveprøvenes pålitelighet. I: Ungdommers skrivekompetanse. Bind I. Red. Kjell Lars Berge m.fl. Oslo

[37] Se t.ex. Lindström 1998, Boesen 2004, Olofsson 2006, Olsson-Wahlsten 2002, Åhs 2007, Östlund-Stjärnegårdh 2002 och Ciolek-Ciastek 2008.

Frågan om bedömaröverensstämmelse är även fortsättningsvis aktuell och det finns behov av att göra ytterligare studier. Erfarenheter från denna studie är att man bör tänka på följande aspekter när man planerar för nya analyser:

- För det första bör man försöka minimera effekten av de faktorer vilka har påverkat den här studiens möjligheter till generaliseringar negativt.
- För det andra finns det analysområden vilka inte har varit möjliga att belysa inom ramen för det här arbetet men som skulle vara intressanta att undersöka i framtida studier. Exempel på sådana analysområden är om pojkar och flickors elevlösningar bedöms olika, om manliga och kvinnliga lärare skiljer sig åt vid bedömning, finns det skillnader mellan skolor med kommunal eller fristående huvudman, hur ser bedömaröverensstämmelsen ut vid olika prestationsnivåer?
- För det tredje skulle det vara intressant att studera hur bedömaröverensstämmelsen ser ut inom- respektive mellan skolor. För att detta ska vara möjligt behövs dock en annan form av insamlingsförfarande för att empirin ska hålla för analys. Detta eftersom datuminsamlingarna endast genererar enstaka elevlösningar per skola.
- För det fjärde vore det intressant att studera vilka uttryck bristen på bedömaröverensstämmelse kan ta samt hur dessa påverkar elevernas resultat. För när man vet vilka uttryck bristen på bedömaröverensstämmelse tar samt vad de har för effekter underlättas beslut till insatser för att öka likvärdigheten i bedömning.

## Referenslista

Alderson m.fl. 1995. *Language Test Construction and Evaluation.* Cambridge University Press.

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

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

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

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

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

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

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

Gipps, C. 1994. *Beyond Testing.* London: The Falmer Press.

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

Lundahl, C. (2009) *Varför nationella prov?* Lund: Studentlitteratur

(under utgivning).

Messick, S. 1989: "Validity" i: R.L. Linn (Ed.). *Educational measurement* (3rd ed., pp. 13-103). New York: Macmillan.

Moskal, B. & Leydens, J. 2000. "Scoring Rubric Development: Validity and Reliability"[Elektronisk version]. *Practical Assessment, Research & Evaluation*, 7. Hämtad mars 20, 2009 från http://pareonline.net/getvn.asp?v=7&n=10.

Moskal, B. 2000. "Scoring Rubrics: What, When and How?"[Elektronisk version]. *Practical Assessment, Research & Evaluation*, 7. Hämtad mars 20, 2009 från http://pareonline.net/getvn.asp?v=7&n=3.

Nitko A. & Brookhart S. 2007. *Educational Assessment of Students* (5th ed.) New Jersey: Pearson Education Inc.

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

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

Proposition 2008/09:87. *Tydligare mål och kunskapskrav - nya läroplaner för skolan.*

Skolverket 2003. "Det nationella provsystemet – vad, varför och varthän?" i *Skolverkets bedömning av dagens system med nationella prov med avseende på kvalitet och kostnadseffektivitet*. Regeringsuppdrag. Redovisning av uppdrag avseende resultatinformation. Del E. www.skolverket.se

Skolverket 2009. *Ämnesproven 2008 i grundskolans åk 9 och specialskolans åk 10.* www.skolverket.se

SOU 2008:109. *En hållbar lärarutbildning.*

Stemler, S. E. 2004. "A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability"[Elektronisk version]. *Practical Assessment, Research & Evaluation*, 9. Hämtad mars 20, 2009 från http://PAREonline.net/getvn.asp?v=9&n=4.

Uebersax, J. 2008. http://ourworld.compuserve.com/homepages/jsuebersax/agree.htm#recs

Utbildningsdepartementet 1994: "Ett uppdrag till Statens skolverk om att utarbeta och tillhandahålla nationella prov". Dnr U94/1031/Gru

Utbildningsdepartementet 2004. ”Uppdrag till Statens skolverk avseende det nationella provsystemet”. U2004/5293/S

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

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