



CENTRUM FÖR SPRÅK-
OCH LITTERATURDIDAKTIK

LIKVÄRDIG BEDÖMNING AV ELEVERS SPRÅKFÖRMÅGOR

PRELIMINÄRA RESULTAT FRÅN ETT ÄMNESDIDAKTISKT FORSKNINGSPROJEKT

Michael Tengberg (projektledare), Eric Borgström, Lena Lötmarker, Erica Sandlund,
Gustaf B. Skar, Pia Sundqvist, Michael Walkert & Kristina Wikberg

KAU.SE/CSL

Bakgrund

Att elevers kunskaper ska bedömas likvärdigt och rättvist är en grundpelare i svensk skola. Samtidigt vet vi att sammansatta förmågor som skrivande, muntlighet och läsande medför variation i lärares bedömning. Uppgifter som prövar sådana förmågor kan sällan bedömas utifrån enkla kriterier utan förutsätter lärare med både utvecklad ämneskompetens och utvecklad bedömningskompetens.

När det gäller bedömning av nationella prov har det bland annat föreslagits att uppgifter måste göras mer lätträttade för att stärka likvärdigheten. Ett annat förslag är att flytta bedömningsuppdraget till en expertpanel av externa bedömare. Båda dessa förslag skulle innebära en avlastning för lärare och säkerligen också bidra till ökad reliabilitet i bedömningen. Men eftersom lärare i hela landet fortfarande har ansvar för den klassrumsbedömning som ligger till grund både för fortsatt undervisning och för betyg i ämnet så behöver lärare inte träna mindre på bedömning utan mer. Att ta bort komplexa uppgifter från de nationella proven eller att lyfta bort bedömningsuppdraget från lärare riskerar snarare medföra att den innehållsliga kvaliteten sänks och att samstämmigheten mellan bedömare på sikt försämrats. Detta är förstås ingen hållbar lösning. Istället bör vi arbeta för att stärka kvaliteten och likvärdigheten i lärares bedömning av de slags förmågor och färdigheter som skolan främst ska syfta till att utveckla.

I några tidigare internationella studier har man prövat olika former av bedömarträning för att se i vilken grad de kan bidra till ökad samstämmighet. Resultaten är blandade. Några program fungerar effektivt (se t.ex. Elder et al., 2007), andra är mindre effektiva (se t.ex. Weigle, 1998). Dessa program har dock främst prövats på mindre expertpaneler och inte alls i Sverige. Föreliggande studie däremot syftar till att pröva effektiviteten i ett bedömarprogram avsett att höja samstämmigheten mellan vanliga lärare utan expertträning. I fokus för bedömningen står skrivförmåga i svenska i årskurs sex och muntlig förmåga i engelska, också i årskurs sex. Bedömarträningsprogrammet har utformats med målsättningen att:

- göra lärarna mer medvetna om sin egen bedömaridentitet och sitt sätt att bedöma jämfört med andra bedömare;
- göra lärarna mer förtrogna med bedömningskriterier och bedömareffekter;
- låta lärarna kalibrera sig mot varandra för att i högre grad kunna rangordna och betygsätta prestationer på ett samstämmigt vis

Kombinerat forsknings- och kompetensutvecklingsprojekt

Projektet har genomförts vid Centrum för språk- och litteraturredaktik (CSL) som ett kombinerat forsknings- och kompetensutvecklingsprojekt tillsammans med mellanstadielärare i svenska och engelska i Värmlands- och Örebroregionen. Ambitionen har varit att både bidra till lärarnas professionella utveckling och besvara studiens forskningsfrågor. Undersökningen bygger på ett begränsat dataunderlag och ska därför ses som en pilotstudie. Avsikten med denna pilotstudie är i första hand att prova ut innehållet i bedömarträningen dels som utgångspunkt för en större studie, dels som utgångspunkt för ett skraddarsytt och mer resurseffektivt kompetensutvecklingsprogram att genomföra i större skala.



Metod

Design

För att undersöka om bedömarträningen bidragit till ökad samstämmighet har studien utformats som ett experiment med förtest, intervention och eftertest. Interventionen bestod av feedback på förtestet samt en dags träning där teori, individuell bedömning och gruppdiskussioner ingick. En del av gruppdiskussionerna har spelats in för att möjliggöra närmare analyser av hur bedömningar motiveras och förhandlas mellan lärarna. Vid både förtest och eftertest fick svensklärarna ($n=12$) bedöma 30 elevtexter och engelsklärarna ($n=12$) bedöma 10 muntliga, inspelade elevprestationer. Såväl texter som muntliga prestationer bedömdes dels på aspektnivå, dels som ett sammanfattande delprovsbetyg. Svensklärargruppen indelades i en grupp som fick träning och en kontrollgrupp som istället för träning ägnade sig åt sambedömning enligt Skolverkets (2013) rekommenderade modell.¹

I rapporten använder vi tre olika sätt för att redovisa förändring av samstämmighet:

- *Konsensus*, dvs. andelen betyg som bedömare är helt överens om. Konsensus anges både i procent och med ett mått som kallas Cohen's kappa. Kappa skiljer sig från procentberäkning genom att kontrollera för andelen slumpmässig samstämmighet. Kappavärdet anger (liksom ICC-värdet nedan) styrkan i en korrelation där 1,0 uttrycker total samstämmighet och 0 ingen samstämmighet alls utöver den som uppstår av slumpen.
- *Konsistens*, dvs. i vilken utsträckning lärarna rangordnat prestationer på liknande sätt. Konsistens mäts med Intra Class Correlation (ICC) på en liknande skala där 1,0 innebär total samstämmighet och 0 ingen samstämmighet alls.
- *Relativ stränghet*, dvs. spridningen de mellan olika bedömarens grad av stränghet. Detta mäts med G-index, vilket anger antalet grupper av stränghet bland bedömare.

¹ I engelsklärargruppen gjordes ingen uppdelning dels eftersom antalet bedömningar som gjordes var färre och en uppdelning skulle göra det statistiska underlaget för svagt, dels eftersom vi redan studerat sambedömning av muntliga prestationer i engelska (se Sandlund & Sundqvist, 2016).

Resultat

Resultat från studien visar att samstämmigheten mellan lärarna generellt är väldigt låg. Som framgår av tabell 1 ligger kappa-värden både på för- och eftertest under 0,30. För bedömningar med praktiska konsekvenser räknas kappa-värden i intervallet 0,20–0,30 vanligen som otillfredsställande (Cicchetti, 1994). Bedömarträningen har emellertid fungerat väl och bland både svensklärarna och engelsklärarna ökar såväl konsensus som konsistens (se tabell 1 och 2). I svensklärargruppen observerar vi också att bedömarträningen bidragit till en större ökning av samstämmigheten än vad sambedömningen gjort.

Hos både svensklärare och engelsklärare är det framför allt konsistensen som ökat. Det betyder att lärarna blivit bättre på att rangordna prestationer på ett likartat sätt. De är helt enkelt mer överens än tidigare om vilka prestationer som är bättre respektive sämre i förhållande till bedömningskriterierna. Däremot har andelen elevprestationer som lärarna sätter samma betyg på endast ökat i liten grad. Det betyder att många texter och muntliga prestationer fortfarande bedöms olika beroende på vem som bedömer.

TABELL 1. Samstämmighet bland svensklärarna

	Förtest	Eftertest
Proc. samstämmighet		
<i>Bedömarträning</i>	47,8%	52,5%
<i>Sambedömning</i>	53,5%	54,0%
Kappa-värden		
<i>Bedömarträning</i>	0,16	0,21
<i>Sambedömning</i>	0,24	0,27
ICC		
<i>Bedömarträning</i>	0,54	0,67
<i>Sambedömning</i>	0,67	0,69

TABELL 2. Samstämmighet bland engelsklärarna

	Förtest	Eftertest
Proc. samstämmighet	36,0%	39,5%
Kappa-värden	0,17	0,23
ICC	0,58	0,83

Som framgår av tabell 2 har förmågan att rangordna prestationer på ett likartat sätt ökat väsentligt i engelsklärargruppen. Detta är mycket positivt och brukar inom bedömningsforskningen ses som en förutsättning för att också kunna höja konsensus i en grupp av bedömare. Att lärarna blivit mer överens om rangordningen av prestationer utan att andelen bedömningar de är helt överens om ökat lika mycket kan hänga samman med att variationen i stränghet mellan bedömarna samtidigt tycks öka något. G-index, som redovisas i tabell 3, anger hur många grupper av stränghet bland bedömarna som kan urskiljas statistiskt.

Ett lågt G-värde betyder alltså att bedömarna i hög grad är lika stränga medan ett högt värde betyder att det är stor variation mellan stränga och mindre stränga bedömare.

TABELL 3. Variation i stränghet mellan bedömarena

G-index	Förtest		Eftertest	
Svensklärare				
<i>Bedömarträning</i>	1,6		3,6	
<i>Sambedömning</i>	2,2		2,7	
Engelsklärare	2,4		3,5	

Förändring av G-index indikerar alltså en ökad spridning mellan lärarna. Denna framgår även av mått på spridningen i genomsnittligt betyg (om betygen räknas om till en poängskala) för bedömarena före och efter träningen. I kvalitativa analyser av enskilda bedömares bedömarstatistik har vi observerat att några lärare, särskilt de som på förtestet låg i den ena eller andra ytterkanten av skalan över stränghet, påverkats mycket av bedömarträningen. Detta kan eventuellt förklara att spridningen inom gruppen ökat.

TABELL 4. Spridning i genomsnittlig stränghet (endast bedömarträning)

	Förtest		Eftertest	
	Min.	Max.	Min.	Max.
Svensklärare	2,33	2,68	2,23	2,82
Engelsklärare	2,40	3,50	2,88	4,99

a Skalan översätts F=1, E=2, D=3, C=4, B=5 och A=6.

Sammanfattningsvis kan vi konstatera att både när det gäller bedömning av elevtext i svenska och av muntliga prestationer i engelska har träningen i någon grad bidragit till att höja konsensus mellan bedömarena och i hög grad bidragit till att höja konsistensen. Samtidigt tycks emellertid spridningen mellan de strängaste och de mest generösa bedömningarna ha ökat.



Slutsatser och kommande aktiviteter

Baserat på den föreliggande pilotstudien kan vi dra slutsatsen att den bedömarträningsmodell som vi använt oss av möjliggör en ökad samstämmighet mellan lärarna och därmed ökad likvärdighet för eleverna. Några av resultaten är fortfarande motsägelsefulla och ytterligare undersökningar kommer att bli nödvändiga. Vi vet till exempel inte helt och hållet varför träningen i huvudsak påverkar konsistensen och inte konsensus. Vi behöver också ta reda på mer om vilka slags elevprestationer som orsakar större svårigheter för lärarna att bli eniga. För det tredje kommer vi att behöva analysera data från gruppdiskussionerna för att få veta mer om innehållet i dessa samt om huruvida diskussionerna kännetecknas av en ökad grad av harmonisering mellan lärarnas olika perspektiv. Dessa frågor är viktiga både för att förstå de statistiska resultaten och för att kunna utveckla bedömarträningsmodellen.

Samstämmighet i bedömning av komplexa uppgifter kan inte ersättas av prov som innehåller enklare och mer lättträttade uppgifter. Skolan måste utveckla professionell kompetens att både undervisa om och bedöma det slags förmågor som kursplanerna i förstone syftar till och som vi tror att elever kommer att behöva i framtida utbildning, yrkesliv och privatliv. Till dessa förmågor hör att kunna formulera sig i tal och skrift i varierade sammanhang och till olika mottagare liksom att kunna läsa, förstå och använda texter av varierande slag och svårighetsnivåer. Att höja samstämmigheten i lärares bedömningar av den här sortens förmågor och därmed bidra till både professionell utveckling för lärarna och till likvärdighet för eleverna är därför högst angeläget. Därför är den här studien viktig och bör läggas till grund för mer ingående undersökningar av hur bedömarträning kan optimeras och i vilka former den kan distribueras på ett pedagogiskt och resurseffektivt sätt. Sådana undersökningar bör genomföras på ett större dataunderlag för att säkerställa representativitet och statistisk styrka. De bör också genomföras i nära samarbete med huvudmän, skolledare och lärare för att säkerställa såväl legitimitet i systemet som praktisk användbarhet.

Referenser

- Cicchetti, D. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instrument in psychology. *Psychological Assessment*, 6(4), 284-290.
- Elder, C., Barkhuizen, G., Knoch, U., & von Randow, J. (2007) Evaluating rater responses to an online training program for L2 writing assessment. *Language Testing* 24(1), 37-64.
- Sandlund, E., & Sundqvist, P. (2016). Doing versus assessing interactional competence: Contrasting L2 test interaction and teachers' collaborative grading of a paired speaking test. The 4th Nordic Interdisciplinary Conference on Discourse and Interaction (NORDISCO), Oslo and Akershus University College of Applied Sciences, Oslo, Norway, 23-25 November, 2016.
- Skolverket (2013). *Sambedömning i skolan – exempel och forskning*. Stödmaterial. Nr. 13:1395. Stockholm: Skolverket
- Weigle, S. (1998). Using FACETS to model rater training effects. *Language Testing* 15(2), 263-287.



För frågor om studien, kontakta:

Michael Tengberg (projektledare)

michael.tengberg@kau.se

054 - 700 18 78

KAU.SE/CSL