Högskolan Kristianstad
291 88 Kristianstad
044-20 30 00
www.hkr.se

SJÄLVSTÄNDIGT ARBETE

Våren 2014

Masterprogrammet i utbildningsvetenskap

Kan sambedömning leda till en mer likvärdig bedömning och betygssättning?

Författare
Pia Thornberg

Handledare
Anders Jönsson

www.hkr.se

Kan sambedömning leda till en mer likvärdig bedömning och betygssättning?

Abstract
Att bedömning och betygssättning av elevers prestationer brister i likvärdighet förs ofta fram som ett problem. Olika förslag att åstadkomma en ökad likvärdighet förespråkas i olika sammanhang och ett sådant förslag är sambedömning. Sambedömning handlar om att lärare samverkar kring bedömning och betygssättning. Trots att sambedömning förs fram som ett sätt att få en större likvärdighet vid bedömning och betygssättning så saknas vetenskapliga belägg för sådana antaganden. I denna studie har en litteraturöversikt genomförts i syfte att få kunskap om vilka effekter sambedömning har och om dessa effekter kan leda till en mer likvärdig bedömning och betygssättning Resultaten visar att sambedömning har effekter som på sikt kan leda till en ökad likvärdighet vid bedömning och betygssättning om den får omfatta flera steg i bedömningsprocessen, men det är även en process som kan påverkas av flera olika faktorer.

Ämnesord: sambedömning, likvärdighet, samsyn, samstämmighet, bedömarkompetens

INNEHÅLL

Inledning

Skolverket och Skolinspektionen har i flera granskningar påvisat att bedömningen och betygssättningen i den svenska skolan brister i likvärdighet (ex Skolinspektionen, 2012, 2013; Skolverket, 2012b, 2012d, 2012e). Bland annat har Skolverket (2012d, 2012e) gjort en uppföljning av hur elevers betyg på nationella prov och deras slutbetyg eller kursbetyg förhåller sig till varandra. Resultatet visar på stora skillnader mellan skolor vilket ger signaler om problem med likvärdigheten. Vidare har Skolinspektionen sedan läsåret 2009/2010 genomfört omrättningar av ett antal nationella prov, både i grundskolan och i gymnasieskolan, vilka visar att det finns omfattande och stora avvikelser mellan den bedömning som skolorna utför och Skolinspektionens bedömning (Skolinspektionen, 2012). Dessa avvikelser gäller särskilt delprov med ett mer öppet svarsformat som till exempel uppsatsskrivning i engelska och svenska och är en signal om att olika bedömare inte har ”samma måttstock” vid bedömningen av dessa prov. I en aktuell rapport från Skolinspektionen (2013) har 30 av de skolor, 19 grundskolor och 11 gymnasieskolor, som uppvisade störst avvikelser mellan skolornas egen rättning och Skolinspektionens omrättning och/eller har en stor avvikelse i relationen mellan nationella prov och slutbetyg följts upp. Uppföljningen har skett genom tillsyn där man granskat hur lärare och rektorer arbetar med likvärdig bedömning och betygssättning utifrån skollag och andra författningar. Denna tillsyn visar att rektor och lärare brister i sitt arbete med bedömning och betygssättning genom att lärarna inte samverkar kring bedömning och betygssättning och att rektor inte skapar utrymme för detta sådan samverkan. Vidare visar tillsynen att lärare undervisar och betygssätter utifrån egna tolkningar av kursplanerna samt att det saknas riktlinjer kring hur och i vilken utsträckning de nationella proven ska påverka betygssättningen. Ett ytterligare resultat av tillsynen är att eleverna inte informeras om grunderna för betygssättningen, det vill säga att lärarna på dessa skolor inte på ett tydligt sätt kopplar de uppgifter som eleverna jobbar med till kunskapskraven och att lärarna därmed får svårt att förklara och motivera elevernas betyg.

Att lärares bedömning och betygssättning brister i likvärdighet är problematiskt, eftersom det i skollagen (2010:800) uttrycks att utbildningen ska vara likvärdig i betydelsen att alla elever ska ha lika tillgång till utbildning, oberoende av kön, geografiska, ekonomiska och sociala förhållanden. Utbildningen ska även vara likvärdig i bemärkelsen att den ska ha samma kvalitet oavsett var i landet den anordnas eller vem som anordnar den. Att elever blir bedömda på ett likvärdigt sätt, det vill säga att samma prestation ger samma bedömning och betyg oberoende av vem som bedömer eller hur det bedöms, blir då en del av en likvärdig utbildning. Eftersom elevers betyg används som urvalsinstrument och som konkurrensmedel när eleverna senare söker anställning eller till högre utbildningar kan det innebära betydande konsekvenser för den enskilde eleven om inte bedömning och betygssättning utförs på ett likvärdigt sätt (Korp, 2006).

För att åstadkomma en ökad likvärdighet vid bedömning och betygssättning lyfts olika förslag fram. I en handlingsplan för att öka en rättssäker och likvärdig betygssättning föreslår exempelvis Skolverket (2004a) att de nationella provens roll som stöd för betygssättning ska förbättras. Andra förslag i handlingsplanen är att förstärka lärar- och rektorsutbildningarna samt att tillhandahålla olika stödmaterial så som förtydligande av hur styrdokumenten ska hanteras samt olika former av bedömningsmaterial med exempel på bedömda elevlösningar. Liknande förslag lyfts fram av Gustavsson, Måhl och Sundblad (2012). De menar att det finns tre instanser som kan vidta olika typer av åtgärder för att bättre garantera att betygssättningen blir ”rimligt likvärdig”, det vill säga att ”elever med samma betyg har samma slags

kunskaper/förmågor på ungefär samma kvalitetsnivåer” (sid. 97). De åtgärder som föreslås är (1) att lärarutbildningen vidtar åtgärder så att blivande lärare har goda kunskaper om betygssättning (2) att Skolverket vidtar åtgärder så att anvisningarna i styrdokument och nationella prov blir tydligare samt (3) att rektorer och huvudmän vidtar åtgärder som leder till att lärare samverkar mer kring bedömning och betygssättning.

På senare tid har så kallad sambedömning förts fram som en möjlighet att skapa förutsättningar för lärare att samverka kring bedömning och betygssättning i syfte att denna ska bli mer likvärdig. Sambedömning, i den betydelse den diskuteras i föreliggande studie, handlar om att lärare tillsammans genomför eller diskuterar bedömning och/eller betygssättning. Skolverket (2009a) skriver till exempel i ett missiv till regeringskansliet att ”ökade krav på sambedömning skulle leda till ökad bedömaröverensstämmelse bland lärare och vara ett effektivt sätt att höja kvaliteten på lärares bedömning av nationella prov” (sid. 17) vilket fick till följd att Skolverket tilldelades ett regeringsuppdrag att ”informera lärare och rektorer på vilka olika sätt sambedömning kan genomföras och också sprida goda exempel på sambedömning som har visat sig leda till en mer likvärdig bedömning och betygssättning” (U2011/6543/S, sid. 1).

Sambedömning förs även fram i flera rapporter som ett sätt att nå en ökad samsyn och samstämmighet bland lärare (se till exempel Erickson, 2009; Skolinspektionen, 2013; Skolverket, 2009b). Skolinspektionen (2013) lyfter i sina slutsatser fram sambedömning som ett sätt att uppnå en större samsyn och samstämmighet kring bedömning och betygssättning och grundar denna slutsats dels i intervjuer med lärare från de granskade skolorna och dels i resultatet från de omrättningar som under tre år gjorts av nationella prov i grundskolan och gymnasieskolan. I de genomförda intervjuerna uppger lärare på skolor som har en omfattande sambedömning att de känner sig tryggare i sin bedömning av elevens kunskaper än vad lärare på skolor utan sambedömning uppger. Flera lärare uttrycker också att de samtal som uppstod vid sambedömningstillfällena var utvecklande för deras bedömarkompetens. Analysen utifrån lärarenkäter och omrättningar av de nationella proven som Skolinspektionen (2012) genomfört visar att högstadie- och gymnasielärare som ensambedömer elevernas prov sätter högre provbetyg relativt den externa bedömningen jämfört med lärare som sambedömer proven. Man fann även att flertalet av de systematiska skillnaderna i bedömningen av olika elevgrupper var mindre uttalade när proven ursprungligen sambedömts än när de bedömts av en enskild lärare. Även en rapport från Skolverket (2009b) för fram ökad bedömaröverensstämmelse som en förmodad effekt av sambedömning. I studien, som är en större undersökning av bedömarsamstämmigheten i de nationella proven för årskurs 9 samt ett kursprov i matematik C på gymnasiet har man konstaterat att bedömaröverensstämmelsen är god eller mycket god i vissa ämnen men inte i alla. Den rekommendation man ger för att öka likvärdigheten vid bedömning och betygssättning är att föreskriva om sambedömning av de nationella proven. Det som kan vara värt att notera utifrån dessa, och andra liknande studier, är att de vetenskapliga beläggen för de slutsatser som dras och de rekommendationer som ges är sparsamma och grundar sig i större utsträckning på lärares uppfattningar än på systematiska studier av sambedömning och dess effekter.

Då sambedömning i flera sammanhang lyfts fram som ett sätt att stödja en mer likvärdig bedömning och betygssättning och då detta antagande inte alltid är vetenskapligt underbyggt och förankrat finns ett behov av att studera detta förhållande närmare. Syftet med föreliggande studie blir därför att förstå på vilket sätt sambedömning kan bidra till en mer likvärdig bedömning och betygssättning.

Bakgrund

Begreppet likvärdighet har använts med skiftande innebörd i olika tider. Englund och Quennerstedt (2008) beskriver hur diskussionen om begreppet från 1960-talet och fram till idag har förändrats. De beskriver hur begreppet tidigare hade en omfattande betydelse starkt kopplad till jämlikhet och enhetlighet medan diskussionen idag är mer snäv och handlar om måluppfyllelse och utbildningsresultat, vilket kan ses som en konsekvens av det mål- och resultatstyrda system som infördes under 1990-talet. Likvärdighet i ett sådant perspektiv innebär att samma elevprestation förväntas ge samma bedömning och betyg oberoende av vem som bedömer eller hur prestationen bedöms och innebär i grunden att de bedömningar som utförs är i enlighet med kurs- och ämnesplaners mål och kunskapskrav. Det är ur en sådan snävare syn på likvärdighet som begreppet diskuteras i föreliggande studie. Betonas bör dock att likvärdighet i detta sammanhang även behöver tolkas som att elever ska ges likartade möjligheter att lära sig det som senare ska bedömas och/eller betygssättas. I de fall två olika elevgrupper ska utföra en given bedömningsuppgift räcker det därför inte att olika bedömare är överens om prestationernas kvaliteter. Om vi ska hävda att bedömningen är likvärdig behöver eleverna även ha haft motsvarande möjligheter att lära sig det som ska bedömas.

Reliabilitet

Reliabilitet handlar om tillförlitlighet och i vilken utsträckning som olika bedömningar stämmer överens med varandra. Caroline Gipps definierar reliabilitet vid bedömning så här:

> the extent to which an assessment would produce the same, or similiar, score if it was given by two different assessors, or given a second time to the same pupil using the same assessor. (Gipps, 1994, sid. 2)

Ofta brukar man säga att bedömningar som upprepas vid flera tillfällen och under olika omständigheter och då ger samma resultat är reliabla. För att bedömningar ska kunna ge samma resultat vid varje tillfälle behöver dessa vara både noggranna och felfria, det vill säga att inverkan av slumpen och andra felkällor är minimal. Alla sådana mätningar som utförs, och särskilt kunskapsbedömningar, är behäftade med felkällor som påverkar utfallet. Sådana felkällor kan återfinnas på både på elevnivå, lärarnivå och på uppgiftsnivå. På elevnivå kan det handla om stress, motivation och hur eleven mår vid bedömningstillfället. På lärarnivå finns det flera olika faktorer som kan påverka reliabiliteten vid bedömningar. Det kan till exempel vara att bedömaren tar hänsyn till icke-relevanta aspekter, så som kännedom om vilken elev det är som bedöms och hur denne brukar prestera, eller att bedömningen sker utifrån personliga uppfattningar om vad som är viktigt att bedöma. Sådana uppfattningar kan variera mellan olika bedömare och hur väl två olika bedömares bedömning stämmer överens med varandra brukar benämnas inter-bedömarreliabilitet (Gipps, 1994). Vidare finns det en möjlighet att en lärare som ska bedöma flera elevprestationer inte bedömer de första prestationerna på samma sätt som de senare. Det är även tänkbart att en lärares syn på hur en prestation ska bedömas förändras efter hand som denne bedömer olika elevprestationer utan att läraren själv är medveten om detta. Även faktorer av mer tillfällig art, som till exempel trötthet, kan inverka på en lärares bedömning över tid. I vilken grad lärares bedömning över tid är konsekvent benämns ofta med begreppet intra-bedömarreliabilitet (Gipps, 1994). Det kan även finnas felkällor på uppgiftsnivå som kan påverka tillförlitligheten, reliabiliteten, vid bedömning. Olika uppgifter kan ha olika utformning och kan mäta olika typer av kunskaper. Uppgifter med fasta svarsalternativ, s.k. flervalsfrågor, eller uppgifter där svaren ges i form av ett ord eller ett numeriskt resultat, så kallade ”kortsvarsfrågor”, har fördelen att de jämfört med andra mer öppna uppgifter ger resultat som är lätta att jämföra och de går snabbt att rätta (Cunningham, 1998). Nackdelen med sådana frågor är att de endast ger begränsad

information om elevers kunskaper eftersom de ofta endast har ett eller ett begränsat antal godtagbara svar. Detta är inget problem så länge man är intresserad av fakta- och procedurkunskaper men om man vill utvärdera mer komplexa kunskaper behövs andra uppgiftsformat (Cunningham, 1998). Det har visat sig att flervalsfrågor är det uppgiftsformat som ger högst bedömaröverensstämmelse (Gipps, 1994) medan bedömning av mer omfattande uppgifter, som till exempel uppsatsskrivning, ger sämre överensstämmelse och utrymme för subjektivitet (Murphy, 1982).

Att använda sig av olika former av stöd vid bedömningen, som till exempel bedömningsmatriser har visat sig kunna öka reliabiliteten vid bedömning (Jönsson & Svingby, 2007) eftersom de tydliggör för bedömaren vad denne ska fokusera vid bedömningen av elevers prestationer. Andra sätt att öka reliabiliteten vid bedömning är att jämföra elevers prestationer med på förhand bedömda elevarbeten som representerar olika kvaliteter eller genom bedömarträning (Dunbar, Koretz & Hoover, 1991; Harlen 2004a), vilken kan vara särskilt effektiv om den även innebär att lärare involveras i att diskutera och utveckla bedömningskriterier (Harlen 2004a). Trots tydliga bedömningsanvisningar finns det belägg för att lärare ändå har svårt att bortse från personliga preferenser när enskilda elevarbeten ska tolkas och värderas (Wyatt-Smith, Klenowski & Gunn, 2010) eller att vikta särskilda kriterier högre än andra (Rezaei & Lovorn, 2010).

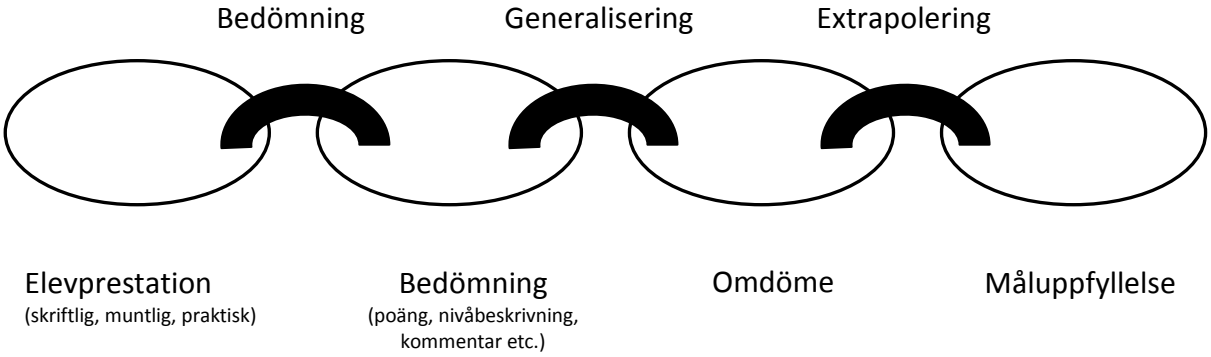
Att bedömningar har låg reliabilitet får konsekvenser för bedömningens likvärdighet och för den enskilde eleven genom att slumpen får stor inverkan på resultaten. Hur stora dessa konsekvenser blir beror på bedömningens syfte. Är bedömningens syfte summativt och utgör underlag för exempelvis elevens betyg kan konsekvenserna bli större än om syftet är formativt, för att stödja lärandet. Vid sådan bedömning kan eventuella felaktigheter enkelt korrigeras i det fortsatta arbetet.

Sammantaget kan man konstatera att hur pass hög eller låg reliabiliteten kommer att vara vid bedömning av en uppgift eller av ett test påverkas av flera olika faktorer och att det är angeläget att sträva efter så låg sådan påverkan som möjligt. Slumpmässiga faktorer som exempelvis om en elev är på dåligt humör vid ett enskilt bedömningstillfälle är svåra att planera för vilket är ett argument för att elever bör ges möjlighet att prövas för enskilda mål vid flera olika tillfällen. Däremot kan mer systematiska felkällor undvikas genom exempelvis tydliga bedömningsanvisningar och bedömarträning.

Validitet

Validitet handlar traditionellt om i vilken utsträckning en uppgift bedömer det som uppgiften var avsedd att bedöma (Gipps, 1994), det vill säga en egenskap knuten till uppgiften eller till provet. Validitetsbegreppet kan även utvidgas och handla om hur elevers resultat tolkas och används (Black, 1998). Validitet är ett komplicerat begrepp eftersom det kan omfatta flera olika aspekter. Det kan till exempel handla om hur väl ett test eller en uppgift stämmer överens med styrdokumenten eller om vilka slutsatser som dras om elevers kunskaper utifrån en bedömningssituation.

En ansats att konkretisera valideringsprocessen har gjorts av Kane, Crooks och Cohen (1999). De har delat upp bedömningsprocessen i flera led, från observationen av en enskild elevprestation till de slutsatser som dras i förhållande till måluppfyllelse eller betyg (Figur 1) och beskriver utifrån en broanalogi hur denna process kräver trovärdighet i flera led för att vara valid.



Figur 1 Modellen är en beskrivning av validitet, anpassad till svenska förhållanden, utifrån en modell av Kane et al. (1999).

Den första bron representerar en *bedömning* av en *elevprestation*. Elevprestationen kan bestå i att eleven genomför ett skriftligt prov med flervalsfrågor, men den kan även utgöras av andra mer öppna frågeställningar som ska redas ut, en bild, en uppsats, en muntlig framställning eller en praktisk handling. Elevprestationen ska sedan tolkas och värderas i förhållande till bedömningsanvisningar eller till någon form av kunskapskrav vilket resulterar i en *bedömning*, antingen i form av poäng, en nivåbeskrivning eller i någon annan form av kommentar som beskriver kvaliteten på prestationen. Notera att i figuren används begreppet bedömning både för att illustrera handlingen (att tolka och värdera elevprestationen) och resultatet (utfallet av handlingen). I detta led beror brons styrka dels på omständigheterna runt bedömningstillfället och dels på i vilken utsträckning som samtliga elevprestationer bedöms på samma sätt. Det får då betydelse vilken typ av uppgiftsformat som används. Flervalsfrågor eller frågor med endast ett eller ett begränsat antal giltiga svar ger som vi tidigare sett en större överensstämmelse vid bedömningen men har nackdelen att de lämpar sig sämre vid bedömning av mer komplexa förmågor. När mer öppna uppgiftsformat används blir det då nödvändigt med tydliga bedömningsanvisningar för att åstadkomma en samstämmig bedömning. Som vi ser är reliabiliteten inbyggd i denna modell som del av validitetsbegreppet.

Den andra bron representerar en *generalisering* som sker utifrån den bedömning som gjorts av den eller de utvalda uppgifterna vid det specifika bedömningstillfället till att även gälla andra liknande uppgifter. Brons styrka handlar således om i vilken mån läraren hade kunnat välja andra uppgifter för att pröva elevernas kunskaper och i vilken utsträckning dessa andra uppgifter då hade resulterat i samma omdöme. Generaliseringen kan även innebära i vilken utsträckning prestationen kan anses situationsbunden eller om eleven kan antas prestera på motsvarande nivå om uppgifterna ges i ett annat sammanhang eller om de hade bedömts på ett annat sätt, till exempel i en praktisk övning. Generaliseringen resulterar i ett *omdöme*, det vill säga ett uttalande om elevens kunskaper utifrån den eller de utförda uppgifterna. Genom att låta elever lösa olika uppgifter, vid olika tillfällen och genom att använda olika bedömningsformer kan en sådan generalisering blir mer underbyggd. Den tredje och sista bron innebär en *extrapolering*. En extrapolering innebär här att man utifrån elevers prestationer på konkreta bedömningsuppgifterna drar slutsatser om deras mer abstrakta *måluppfyllelse*. I våra styrdokument uttrycks att undervisningen ska resultera i att eleverna utvecklar komplexa kunskaper som till exempel matematisk problemlösningsförmåga. Om bedömningen grundar sig på att elever fått lösa flera olika typer av matematiska problem inom olika matematiska områden är det förhållandevis lätt att extrapolera utfallet ifrån sådana uppgifter till målet, problemlösningsförmåga. Om eleverna däremot enbart har fått lösa

rutinuppgifter eller endast problem inom något avgränsat matematiskt område är det svårare att extrapolera sådana resultat till ett komplext mål som problemlösningsförmåga. I detta tredje led handlar brons styrka alltså om att de slutsatser som dras utifrån elevernas prestationer ska vara valida i förhållande till de mål som anges i kurs- och ämnesplaner. Modellen som beskrivs ovan inkluderar inte explicit att bedömningssituationen är sammankopplad med exempelvis styrdokument. En sådan koppling diskuteras först i förhållande till extrapoleringen i det tredje ledet. Däremot argumenterar Kane et al. (1999) för att det bör finns en nära koppling mellan de uppgifter elever får lösa och de mål som de ska uppnå och menar att om en elev presterar bra vid ett bedömningstillfälle så bör det också innebär att denna elev presterar bra i förhållande till målen, och tvärtom. En aspekt av validitet kan då även vara det finns en tydlig linje som förbinder ett ämnes mål och syfte med undervisning, kunskapskrav och bedömning, så kallad ”alignment” (Biggs & Tang, 2007).

Sammanfattningsvis kan alltså validitet, så som det beskrivs utifrån Kane et al. (1999), uttryckas som styrkan i de tre broarna, givet att det finns en linje (alignment) som förbinder styrdokument, undervisning, bedömning och betygssättning.

I vilken utsträckning en bedömning är valid kan då även relateras till hur väl en bedömningssituation stämmer överens med det som var avsett att bedömas. För en god sådan överensstämmelse krävs både att systematiska fel saknas, eller är minimerade, samt att de slumpmässiga felen är små. Hög validitet förutsätter därför, som tidigare nämnts, även att reliabiliteten är hög och i modellen som beskrivs ovan ses reliabilitet som en delmängd av validitet. Vanligtvis räknas reliabiliteten som överordnad validiteten utifrån argumentet; om det finns en osäkerhet kring om bedömningen är reliabel och om den har påverkats av en mängd faktorer så spelar det mindre roll om uppgiften är valid (Harlen, 2004a). Detta argument tenderar att få till följd att metoder för att öka reliabiliteten eftersträvas vilket generellt innebär att uppgifterna blir mer slutna och att bedömningsmetoder med få felkällor, så som flervalsfrågor, används. Ett mer begränsat bedömningsunderlag minskar i sin tur validiteten (ibid.). Om vi vill att elever ska bedömas utifrån ett bredare underlag i linje med styrdokumentens intentioner menar Gipps (1994) att vi måste se till att lärare istället får en gemensam förståelse för kriterier och för vilka kontexter som framkallar de bästa prestationerna hos eleverna. Om man å andra sidan försöker öka validiteten genom att utöka urvalet av det som bedöms till att omfatta även komplexa förmågor som resonemangsförmåga och problemlösningsförmåga tenderar reliabiliteten att minska eftersom den typen av förmågor inte enkelt låter sig bedömas (Harlen, 2004a). Detta får konsekvenser för den summativa bedömningen när en kompromiss mellan reliabilitet och validitet tvingas göras. När det gäller den formativa bedömningen kan validiteten sägas vara överordnad reliabiliteten då en sådan bedömning går att ändra och justera efter hand. Förhållandet mellan reliabilitet och validitet blir av särskilt intresse i de fall samma bedömning används både i summativt och i formativt syfte. Med vetskap om att validiteten och reliabiliteten inte är oberoende av varandra blir det i praktiken av intresse att studera hur de samverkar (ibid.).

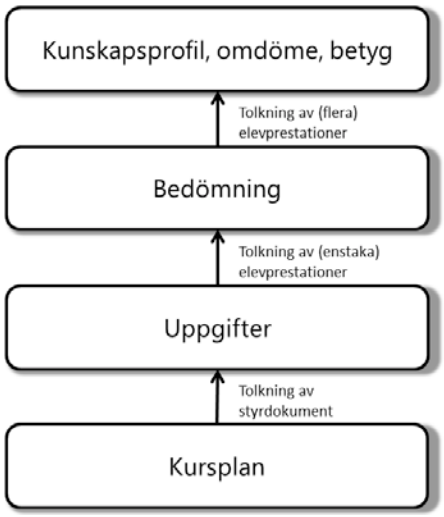
Att ha hög reliabilitet och validitet är alltså två förutsättningar för en likvärdig bedömning, det vill säga att samma elevprestation förväntas ge samma bedömning och betyg oberoende av vem som bedömer eller hur bedömningen går till och innebär även att de bedömningar som utförs är i enlighet med kurs- och ämnesplaners mål och kunskapskrav. Det blir därför centralt att undervisning och bedömning verkligen bygger på att eleverna får visa sitt kunnande utifrån styrdokumenten och att resultaten tolkas och värderas på ett sätt som ger hög validitet. Som nämnts tidigare blir hög reliabilitet en delmängd av hög validitet men det är viktigt att poängtera att en bedömning kan vara reliabel utan att samtidigt vara likvärdig. Lärare kan

nämligen vara väldigt överens om hur en elevprestation ska bedömas utan att bedömningen är i linje med gällande styrdokument.

Reliabilitet och validitet i det svenska betygssystemet

I den svenska skolan idag har vi ett bedömningssystem som är mål- och kriterierelaterat vilket innebär att elevers prestationer jämförs med på förhand uppställda mål och kriterier, det vill säga kunskapskraven i nuvarande kurs- och ämnesplaner. I styrdokumenten (Skolverket, 2011a, 2011b) är det formulerat att undervisningen ska leda fram till att eleverna ges förutsättningar att nå ett antal långsiktiga mål, även kallade förmågor, specifika för varje ämne. Utmärkande för dessa förmågor är att de beskriver elevers kunskaper så som de kommer till uttryck i någon form av handling, till exempel elevers förmåga att föra och följa matematiska resonemang eller att genomföra systematiska undersökningar i biologi. Kunskapskraven, i förhållande till vilka elevers prestationer ska bedömas och betygssättas, utgår i sin tur ifrån de långsiktiga målen, förmågorna, och beskriver hur dessa kan uttryckas. Styrdokumenten ställer alltså krav på att elever ska utveckla komplexa kunskaper och förmågor. Sådana förmågor kan däremot vara svåra att bedöma med uppgifter där svaren kan bedömas som antingen rätt eller fel. Dessa uppgifter är visserligen lättbedömda och ger hög reliabilitet vid bedömning men är svårare att förena med en hög validitet. För en hög validitet krävs därför istället bedömningssituationer där eleverna får möjlighet att visa dessa förmågor, till exempel genom att eleverna får planera och genomföra en systematisk undersökning i biologi. En sådan uppgift kan genomföras av eleverna på flera olika sätt och på olika kvalitativa nivåer vilket då riskerar ge låg reliabilitet vid bedömningen eftersom det lämnar större tolkningsutrymme för bedömande lärare. Även validiteten kan riskera att minska eftersom det visat sig att prov som lämnar öppet för tolkningar i flera led kan få en lägre validitet än prov utan sådant tolkningsutrymme (Haertel & Herman, 2005). Eftersom det är samma förmågor som bedöms under elevernas hela skoltid är det å andra sidan möjligt för lärare att över tid bedöma samma förmåga vid flera olika tillfällen och med olika bedömningsformer. Detta gör att både reliabiliteten och validiteten vid bedömning kan öka (Harlen, 2004a, 2004b).

Att bedöma elevers prestationer på ett likvärdigt sätt utifrån rådande styrdokument ställer stora krav på den som utför bedömningen och kräver att tolkningar görs i flera olika led. I ett första led ska styrdokumenten tolkas och omsättas i undervisning och bedömningsuppgifter. Elevernas prestationer ska sedan tolkas och värderas utifrån bedömningsanvisningar och/eller kunskapskrav och slutligen ska elevers olika prestationer sammanvägas i ett betyg eller i någon form av kunskapsomdöme. Utifrån modellen av validitet, och därmed även reliabilitet, som beskrivs i föregående avsnitt (Kane et al., 1999) så räcker det inte utifrån våra svenska förhållanden att beskriva bedömning och betygssättning som en process som startar i och med att elevers prestationer ska bedömas. Elevers prestationer ska bedömas utifrån kunskapskrav som i sin tur bygger på långsiktiga mål/förmågor. Därför behöver bedömningsprocessen även omfatta såväl tolkningen av vad dessa förmågor innebär som formuleringen av bedömningsuppgifter där dessa förmågor ges möjlighet att uttryckas. Allal (2013) beskriver professionell bedömning som en process vilken omfattar tre olika tolkningsled, se Figur 2.



Figur 2. Modellen beskriver hur en likvärdig bedömning och betygssättning kräver samsyn mellan lärare i flera led. (1) Tolkning av styrdokument för planering av undervisning och för konstruktion av bedömningsuppgifter, (2) tolkning och värdering av enskilda elevprestationer utifrån kunskapskrav och bedömningskriterier samt (3) tolkning av ett brett underlag av elevprestationer i samband med betygssättning eller vid formulering av kunskapsomdömen. Modellen är anpassad med utgångspunkt från Allal (2013).

De olika tolkningsleden samt de utmaningar som varje tolkningsled innebär i förhållande till en likvärdig bedömning kommer att beskrivas närmare i de följande avsnitten.

Tolka styrdokument och konstruera uppgifter

I det första tolkningsledet består tolkningen i att skapa en förståelse för styrdokumenten, främst kursplaner respektive ämnesplaner, och dess innehåll samt i att utifrån dessa planera undervisning och konstruera bedömningsuppgifter. För att bedömning av prov och andra uppgifter ska ske på ett likvärdigt sätt mellan olika lärare, skolor och klasser behöver mål och kriterier tolkas och tillämpas på samma sätt av samtliga lärare, vilket ställer krav på att lärare kommunicerar med varandra. Harlen (2004a) menar att hur konsekventa lärare är vid bedömningen av elevprestationer hör samman med i vilken utsträckning de har förstått styrdokumenten och i vilken utsträckning detta sedan omsätts i undervisning. De kunskaper och förmågor som efterfrågas i dagens styrdokument är mer kognitivt komplexa vilket innebär att de inte enkelt låter sig bedömas med traditionella skriftliga bedömningsuppgifter som ofta har ett fokus på rätt eller fel svar. Istället krävs mer öppna uppgiftsformat med utrymme för eleverna att visa sina kunskaper på olika sätt och i olika situationer. Det innebär även att lärare måste gör tolkningar av vad som är utmärkande för prestationer på olika nivåer och vad som skiljer prestationer på olika nivåer ifrån varandra. Vad innebär det till exempel att en elev i en problemlösningssituation i matematik för ”utvecklade och relativt väl underbyggda resonemang” och hur skiljer ett sådant resonemang sig ifrån ”välutvecklade och väl underbyggda resonemang”? I nästa steg behöver lärare sedan ta ställning till hur man planerar undervisning och konstruerar uppgifter, som ger eleverna möjlighet att visa hur de resonerar. I ett bedömningssystem som detta, som till stor del bygger på tolkningar, har en del bedömare innebörden av kriterier klart för sig medan andra använder egna mer intuitiva kriterier och riktlinjer baserade på till exempel sina egna erfarenheter av att ha blivit bedömda (Yorke, Bridges & Woolf, 2000; Hand & Clewes, 2000). Många lärare upplever själva att uppsatta mål och kriterier är tydliga och att de förstår dem (Skolverket, 2004b) och en sådan

uppfattning blir inte utmanad förrän den konfronteras med någon annan avvikande uppfattning. Ecclestone (2001) argumenterar därför för att bedömare måste vara väl insatta i kriterier och delta i utvecklingen av dessa. En gemensam syn på vad kriterier står för och möjlighet att delta i diskussioner kring dessa påverkar både validitet och reliabilitet vid bedömning (Holroyd, 2000). I vilken utsträckning lärare har en samsyn kring kriterier och utvecklar ett gemensamt språk för att beskriva och bedöma elevers prestationer påverkar därför tillförlitligheten vid bedömningen (Harlen, 2004a).

Som tidigare nämnts räcker det inte med att lärare har en samsyn kring hur styrdokumenten ska tolkas utan de behöver även erbjuda en undervisning som ger eleverna möjlighet att utvecklas mot uppsatta mål. Vilken undervisning elever får kan variera från klassrum till klassrum även om undervisningen utgår ifrån samma styrdokument. I gymnasieskolan har det visat sig att undervisningen i kärnämnena engelska, svenska och matematik, varierar på ett sådant sätt mellan skolor och mellan program att elever ges mycket skilda förutsättningar att nå kursmålen och svara upp mot kriterier för de olika betygen (Korp, 2006). På studieförberedande program förekommer det att undervisningens innehåll och nivå anpassas och har en inriktning mot teoretiskt lärande och analys medan den på yrkesförberedande program har en mer reproducerande och återgivande karaktär (ibid.). Vilken undervisning elever får ger dem i förlängningen olika förutsättningar att svara mot de olika kravnivåer som ställs för olika betyg och är därför en viktig aspekt i en likvärdig bedömning.

Att ha en gemensam tolkning av styrdokumenten får också betydelse när lärare ska skapa valida bedömningsuppgifter att använda för att utvärdera elevers prestationer under eller efter ett arbetsområde. Återigen kommer lärares skilda förmåga att tolka och omsätta kurs- och ämnesplanernas formuleringar till uppgifter i vilka det är möjligt för eleverna att visa sina kunskaper i fokus. Vid formativ bedömning är syftet att stödja elevens fortsatta lärande och då hamnar validiteten i fokus eftersom det blir viktigt att läraren kan ge återkoppling i linje med kurs- och ämnesplaners syfte och kunskapskrav. Självklart är det önskvärt att bedömningen även är reliabel men om syftet är att stödja elevens lärande i den fortsatta undervisningsprocessen kan eventuella felaktigheter i bedömningen enkelt justeras.

Vid summativ bedömning, när elevers kunskaper vid en viss tidpunkt ska kontrolleras och summeras, blir det däremot särskilt viktigt att bedömningen både blir valid och reliabel eftersom sådana bedömningar ofta ligger till grund för elevers betyg eller kunskapsomdömen. Om summativ bedömning ska tjäna sitt syfte på ett effektivt sätt så måste den dessutom spegla en bredd av de mål som satts upp för elevers lärande, så kallade lärandemål (Crooks, 1988; Harlen, 2004a). För att skapa sig en god bild av elevers kunnande behöver lärare därför skaffa belägg för detta utifrån en rad olika aktiviteter som speglar en sådan bredd av lärandemål. Enstaka prov som ges vid något eller några enskilda tillfällen kan däremot endast ge begränsad information om elevers kunnande. Lärare har även möjlighet att använda den information denne får i samband med bedömningstillfället för att hjälpa eleven vidare i sitt lärande, det vill säga att den information som samlas in i samband med till exempel ett prov, kan användas både i formativt och i summativt syfte. De bedömningsuppgifter som lärare använder kan vara av olika slag. I samband med nationella prov, som ges i vissa ämnen och i särskilda årskurser och kurser, är de uppgifter som ges konstruerade och utprövade så att de svarar mot kurs- och ämnesplaner. Dessa prov har ofta hög produktkvalitet, det vill säga att de är valida, välgjorda och relevanta i förhållande till kursplanens mål. Att någon annan konstruerat bedömningsuppgifterna innebär att lärarna inte själva behöver tolka styrdokumenten i detta avseende. Att som lärare själv konstruera uppgifter i enlighet med styrdokumenten är däremot svårt. Boesen (2006) fann i sin avhandling att lärares

egenkonstruerade matematikprov i liten utsträckning liknar de nationella proven avseende produktkvalitet. Medan de nationella proven prövar en bredd av matematiska förmågor som kräver välgrundade matematiska resonemang prövar de lärarkonstruerade proven i stor utsträckning så kallade imitativa resonemang, vilka kan utföras utan matematisk förståelse. Hur lärare uppfattar och förstår styrdokumenten och hur de väljer bedömningsuppgifter blir därmed avgörande för vilken typ av kunskap eleven har möjlighet att visa.

Sammantaget kan man konstatera att lärare tolkar styrdokumenten på olika sätt och därmed genomför undervisning på olika sätt och konstruerar olika uppgifter för att pröva elevernas kunskaper. En likvärdig bedömning och betygssättning, i detta tolkningsled, kräver att lärare tolkar styrdokumenten på ett enhetligt sätt. Det krävs även att lärare har en gemensam linje (”alignment”) som sträcker från tolkningen av styrdokumenten via undervisning och uppgiftskonstruktion till bedömning och betygssättning. Endast en samsyn är inte tillräckligt då de uppgifter som konstrueras för bedömning måste vara valida, det vill säga att de ger eleverna möjlighet att visa sin kunskap i förhållande till uppsatta mål och inte påverkas av andra faktorer. Att lärare samverkar kring detta tolkningsled är enligt Allal (2013) mer vanligt förkommande än att de samverkar i de nästkommande tolkningsleden.

Tolka elevprestationer för bedömning

Det andra tolkningsledet handlar om att tolka och värdera elevers prestationer. När eleverna genomför en uppgift, kommer kvaliteten på deras prestationer inte enbart att bero på kunskaperna i ämnet utan även vara beroende av andra faktorer så som elevens motivation, uppgiftsformatet, vilka hjälpmedel eleverna har tillgång till med mera. Utfallet kan även påverkas av hur läraren förbereder eleverna inför provtillfället och vilken typ av hjälp denne bistår med under bedömningssituationen (Korp, 2006).

När eleverna har utfört bedömningsuppgiften ska läraren tolka och värdera elevernas prestationer utifrån kunskapskraven och måste då ta ställning till frågor som: ”Vad betyder elevens prestation?”, ”Vad tycks eleven behärska?”, ”Vad behöver utvecklas?”. Sadler (1998) beskriver detta som en process i tre delar som innebär att lärare (1) riktar uppmärksamhet mot elevprestationen (2) värderar denna mot till exempel ett ramverk och (3) ger ett explicit svar, som till exempel att ge prestationen en poäng, ett omdöme eller ett betyg. Sadler (1998) beskriver även ett antal faktorer som utgör de källor lärare förlitar sig på när de bedömer. Dessa är goda ämneskunskaper, djupkunskap om kriterier och kravnivåer i förhållande till det som ska bedömas, bedömningsskicklighet eller erfarenhet av att ha bedömt liknande uppgifter tidigare samt en uppsättning attityder och förhållningssätt till undervisning och elever. Detta ger en bild av bedömningstillfället som en situation som är väldigt beroende av den enskilde bedömaren och dennes tidigare erfarenhet. I denna situation finns flera faktorer som kan påverka bedömningens likvärdighet. Wynne Harlen har genomfört två forskningsöversikter där hon undersökt lärares summativa bedömningar och vad som påverkar reliabilitet och validitet i dessa bedömningar (2004a, 2004b). Ett av resultaten är att lärare tar hänsyn till icke-relevanta aspekter i bedömningen så som uppförande, kön, om elever i behov av stöd eller inte. De låter även elevers muntliga prestationer påverka bedömningen, även i bedömningssituationer som inte är muntliga. Liknande resultat rapporterar även Wyatt-Smith et al. (2010) som visar att lärare vid bedömning använder olika referenser som utgångspunkt som kan påverka bedömningens och betygssättningens likvärdighet. I sådana referenser ingår bland annat tillgängliga artefakter så som bedömningsanvisningar och övrigt stödmaterial, men även deras egen syn på ämnet och deras ämneskunskap. I referenserna ingår även så kallad tyst kunskap, ”tacit knowledge”, av olika slag som till exempel kännedom om den

enskilde eleven, tidigare erfarenheter av bedömning, vilka elever läraren tidigare mött och därmed antaganden om vilken nivå som olika elever ”bör” prestera på (ibid.). Det finns därmed goda skäl att anta att det kan finnas en stor variation i hur olika lärare tolkar och värderar elevers prestationer. Att så är fallet bekräftas även i svenska studier, till exempel i rapporter från Skolverket (2009b) och Skolinspektionen (2012), där man undersökt bedömaröverensstämmelsen i samband med nationella prov i grundskolan och samt för kursproven på gymnasieskolan. Undersökningarna har gjorts genom att provinstitutionerna vid de lärosätena som konstruerar de nationella proven eller Skolinspektionen låtit ombedöma ett urval av de inskickade elevlösningarna och jämfört utfallet av denna ombedömning med den ursprungliga bedömningen. Det man funnit visar bland annat på att det finns omfattande skillnader i hur olika lärare bedömer i uppgifter där eleverna ges ett stort frihetsutrymme så som vid uppsatsskrivning i engelska och svenska och i mer utredande uppgifter i matematik.

Sammantaget kan vi konstatera att lärare tolkar och värderar elevers prestationer på olika sätt och att de tar in olika faktorer i sin bedömning som får konsekvenser för denna. För att bedömningen ska bli likvärdig krävs en samsyn och att lärare värderar liknande prestationer på ett likvärdigt sätt.

Tolka elevprestationer för betygssättning

I det tredje och slutliga tolkningsledet ska lärare sammanväga olika elevprestationer till betyg eller till ett kunskapsomdöme utifrån kurs- eller ämnesplanernas kunskapskrav. Inför betygssättning eller inför utvecklingssamtal med elever och elevers vårdnadshavare ska lärare ha skaffat belägg för elevers kunnande utifrån ett brett underlag och utifrån olika situationer. När lärare ska summera elevers kunskaper i form av ett betyg ska detta ske utifrån de kunskapskrav som finns angivna för varje ämne och de ska sättas oberoende av hur andra elever presterat. För att betygssättningen ska vara likvärdig krävs att kunskapskraven tolkas på samma sätt av olika lärare. I kunskapskraven används ett antal värdeord som i viss utsträckning är gemensamma och återfinns i flera ämnen. Skolverket (2012a) menar att värdeorden till stor del får ”sin betydelse i de sammanhang de används” och att det därför inte är möjligt att göra generella definitioner av dessa. Ett värdeords innebörd kan dessutom skilja mellan olika situationer och mellan olika uppgifter. Det lämnas alltså åt lärarna att göra sina tolkningar av vad som till exempel skiljer ett ”utvecklat resonemang” ifrån ett ”väl utvecklat resonemang”.

Förutom att lärarna ska göra tolkningar av kunskapskraven och dess särskiljande värdeord ska de dessutom göra en sammanvägning av de olika kunskapskraven till ett betyg och detta ska även göras utifrån flera olika bedömningssituationer. Det visar sig att när lärare på gymnasieskolan ska sätta kursbetyg använder de olika modeller för att besluta om vilket betyg eleverna ska ha. Korp (2006) identifierade tre olika modeller som lärare i huvudsak använde sig av vid betygssättning: en analytisk modell, en aritmetisk modell och en blandad- relativ-modell. Den analytiska modellen innebär att elevernas kunskapsprofil analyseras utifrån kursplanens mål och betygskriterier (jämför dagens kunskapskrav) och att de nationella proven (i de ämnen det finns sådana) används som ett av flera underlag. Den aritmetiska modellen bygger på att elevernas betyg räknas fram på basis av resultat på prov och olika kursuppgifter. Även i denna modell användes resultatet på de nationella proven som ett av flera underlag. I den blandade – relativa – modellen bestäms elevers betyg utifrån en blandning av olika överväganden i form av provresultat, allmänna intryck av elevers kunskapsnivå, närvaro, attityd och aktivitet under lektioner. I en sådan modell är det således ett fokus på hur läraren bedömer elevers allmänna kunskapsnivå i ämnet och inte på vilka

specifika kunskaper eleverna har. I denna modell används det nationella provet för att verifiera lärarens egen tidigare bedömning eller som själva utgångspunkten för betygssättningen. I Korps (2006) studie uppträdde de olika modellerna i mer eller mindre renodlad form och enskilda lärares resonemang dominerades ofta av den analytiska eller den aritmetiska modellen och hade inslag av den tredje. Det visade sig även att det tidigare normrelaterade systemet levde kvar på vissa håll. Även Allal (2013) har studerat lärares betygssättningspraktik samt de överväganden som görs vid betygssättning. Hon beskriver hur denna praktik omfattar två huvudsakliga operationer, dels insamlande av information utifrån varierade källor och dels hur denna information tolkas och sammanvägs till ett betyg. Sammanvägningen begränsades ibland till en form av aritmetisk algoritm utifrån det, huvudsakligen, skriftliga underlaget som gick till på liknande sätt för samtliga elever. I de fall elevers olika resultat varierade mycket eller läraren av andra orsaker var osäker på elever förståelse av ett kunskapsområde användes eller insamlades ytterligare underlag. Underlaget kunde till exempel bestå av andra uppgifter eleven utfört men som inte var egentliga bedömningsuppgifter, hur eleven deltagit i klassrumsaktiviteter, inlämnade läxor eller utifrån diskussioner med eleven eller dennes föräldrar. Hur det samlade underlaget sedan användes för att sammanvägas till ett betyg beskrivs av Allal som en tolkningsprocess som kunde se olika ut för olika elever.

Sammantaget kan man se att lärare vid sammanvägning av olika underlag till ett samlat omdöme eller betyg tolkar och viktar olika underlag på skilda sätt och som vi har sett så innebär samtliga tre tolkningsled en osäkerhet i lärares bedömning och betygssättning. Dessa moment av osäkerhets beskrivs ofta i termer av bristande likvärdighet.

Att verka för en likvärdig bedömning

I Sverige har lärare ett stort förtroende när det kommer till bedömning och betygssättning. Ofta är det samma lärare som undervisar eleverna som bedömer dem, rättar deras nationella prov och som sätter deras betyg. Fördelarna med ett sådant system där summativa bedömningar utförs av samma lärare som också undervisar eleverna är flera. Harlen (2005) för fram att bedömningen när den sker som en del av den ordinarie undervisningen ger läraren möjlighet att bedöma utifrån ett bredare underlag och utifrån fler lärandemål än vad enstaka tester utförda vid särskilda tillfällen kan ge. Då bedömningen genomförs som en del av undervisningen kan den även användas för att ge formativ feedback och för att göra elever delaktiga i bedömningsprocessen (ibid.). Lärares bedömningar anses därför ha hög validitet då de har möjlighet att finna belägg för elevers kunnande utifrån många olika situationer (Crooks, 1988).

Lärares bedömningar anses däremot ha lägre reliabilitet, främst eftersom läraren har en förförståelse om enskilda elever samt om den undervisning eleverna fått. Ur ett internationellt perspektiv är det förhållandevis ovanligt att ”high-stake” summativ bedömning utförs av samma lärare som också undervisar eleverna. Flera länder, som till exempel Frankrike och Tyskland, har istället någon form av examensprov där bedömningen utförs externt av ett fåtal särskilt utvalda personer. Dessa särskilt utvalda personer har ibland också fått bedömarträning för att öka reliabiliteten vid bedömningen. Genom att använda externa bedömare som inte har någon tidigare relation med eleven kan så kallade haloeffekter minska, det vill säga att lärares tidigare uppfattning om en elev påverkar bedömningen (Wood, 1991). Detta sätt ger mer reliabla resultat men ger inte lärarna något stöd eller verktyg för att själva tolka styrdokumenten och att genomföra undervisning i enlighet med dessa. Lärarna kan då istället koncentrera sig på att hjälpa varje elev att få så bra resultat som möjligt med den påtagliga

risken att lärarna då väljer att låta undervisningen styras av sådant som man vet kommer att behandlas på testerna (Harlen & Deakin Crick, 2002). Det finns även stöd för att elevers motivation för lärande påverkas negativt av sådana ”high-stake”-prov som får en avgörande betydelse för deras betyg eller framtida möjligheter (ibid.).

Eftersom man i det svenska skolsystemet valt en modell där läraren har ett stort inflytande över både undervisning, bedömning och betygssättning, trots de problem det medför ur ett likvärdighetsperspektiv, blir det angeläget att hitta metoder för att stärka likvärdigheten. Inledningsvis nämndes tre olika instanser som kan vidta olika typer av åtgärder för att åstadkomma ökad likvärdighet vid bedömning och betygssättning: Skolverket, lärarutbildningen och skolhuvudman (Gustavsson et al., 2012).

Skolverket utformar bland annat nationella prov, vilka på senare tid både införts i fler årskurser och i fler skolämnen. Nationella prov kan ge lärare stöd att tolka styrdokumenten genom att de ger exempel på hur bedömningsuppgifter som prövar elevers kunskaper i enlighet med dessa kan se ut och hur de kan bedömas (de två första tolkningsleden i Figur 2). Som nämnts tidigare visar det sig däremot att lärare kan ha svårt att överföra sådana insikter till situationer när de själva ska konstruera bedömningsuppgifter (Boesen, 2006). Vilket stöd de nationella proven ger i betygssättningen, det tredje tolkningsledet, är mer oklart då de provbetyg som ges i de flesta fall är en sammanvägning av flera olika delprov och där delproven även i vissa fall viktas i förhållande till varandra. Skilda provkonstruktörer har dessutom olika eller delvis olika metoder för att sammanväga delprovresultaten till ett provbetyg. Eftersom de nationella proven inte är avsedda att pröva hela kurs- respektive ämnesplanen och det därutöver finns tydliga anvisningar om hur sammanvägning och eventuell viktning av olika delprov ska gå till gör det processen ganska olik den situation lärarna själva befinner sig i när de ska sammanfatta elevernas prestationer och sätta betyg utifrån kunskapskraven. Hur detta provbetyg sedan hanteras i förhållande till annat bedömningsunderlag som lärarna har tillgång till kan variera (se till exempel Korp, 2006). Då genomförandet av proven är obligatoriskt, genomförs i flera årskurser och ämnen samt genom att utfallet av dem är offentligt och används för jämförelser får de stort genomslag. Därmed blir även deras bidrag till en likvärdig bedömning förhållandevis stort eftersom de når ut till och används många lärare. Det är däremot svårare att uttala sig om vilken påverkan det får på lärares bedömning av andra uppgifter eller i ämnen som saknar nationella prov.

En annan åtgärd som Skolverket bidrar med för att stödja en likvärdig bedömning och betygssättning är att tillhandahålla bedömningsstöd av olika slag. Dessa är riktade mot antingen ett enskilt eller mot ett par av de olika tolkningsleden (se Figur 2). I det första tolkningsledet när styrdokumenten ska tolkas och omsättas till uppgifter finns till exempel kommentarmaterial till olika ämnes kursplaner i grundskolan. I en del ämnen finns provbanker och diagnosmaterial med uppgifter som kan användas för undervisning och bedömning. Dessa kan ses som en hjälp i att omsätta kursplanens skrivningar i uppgifter som är i linje med dessa. I ämnen och årskurser som har nationella prov finns det även via provkonstruktörernas hemsidor tidigare nationella prov vars uppgifter kan användas i undervisningen och därmed ses som en operationalisering av kursplanen. I det andra tolkningsledet när enskilda elevers prestationer ska tolkas och värderas i syfte att ge någon form av återkoppling kan delvis samma material användas igen. Diagnosmaterial och tidigare nationella prov med tillhörande bedömningsanvisningar kan vara ett stöd åt lärarna där de ges möjlighet att se hur kvalitativt olika elevarbeten kan värderas. I till exempel matematik finns även ett särskilt framtaget bedömarträningsmaterial med uppgifter, bedömningsanvisningar, diskussionsfrågor och filmer som kan användas för att stötta lärare att göra en rättvis och

likvärdig bedömning. I det tredje och sista tolkningsledet när elevers prestationer ska summeras och sammanvägas till ett betyg finns för en del ämnen i grundskolan kommentarmaterial till kunskapskraven som syftar till att ge en bredare och djupare förståelse för hur dessa är konstruerade. Kommentarmaterialen utgår från verksamma lärares bedömningar av autentiska elevexempel och beskriver hur lärare kan gå till väga för att identifiera bedömningsaspekter utifrån värdeorden.

Utöver Skolverket så föreslår Gustavsson et al. (2012) att lärarutbildningen förbereder blivande lärare så att de har goda kunskaper om (bedömning och) betygssättning när de utexamineras och att rektorer och skolhuvudmän vidtar åtgärder som leder till att lärare i ökad omfattning samverkar kring frågor kopplade till bedömning och betygssättning.

Sambedömning

En strategi som ofta lyfts fram i forskningslitteraturen och i andra sammanhang som ett sätt att få valida och reliabla bedömningar är sambedömning. Sambedömning har till exempel föreslagits av Skolverket (2009a) att ingå som en del i att öka likvärdigheten i lärarnas bedömning och betygssättning, tillsammans med nationella prov och övrigt bedömningsstöd. Sambedömning används till exempel i stor omfattning i länder eller delstater där man inte har nationella prov och där andra sätt att stödja en likvärdig bedömning därmed krävs. Ett exempel är i delstaten Queensland, Australien, där man har mångårig erfarenhet av sambedömning på olika nivåer, både lokalt, regional och på delstatsnivå (Klenowski, 2012).

Ett vetenskapligt begrepp för sambedömning finns inte tydligt etablerat. I forskningsstudier är det vanligt att man undersöker ”inter-rater reliability” genom att låta olika bedömare, oberoende av varandra, göra bedömningar av samma elevprestationer för att se hur dessa överensstämmer med varandra. Så är till exempel fallet i de omrättningar som Skolinspektionen utför när externa bedömare utför en omrättning av ett urval av de nationella proven (se till exempel Skolinspektionen, 2012). Ett annat begrepp är ”co-assessment” som innebär att lärare och elever/studenter samverkar kring frågor kopplade till bedömning. I denna studie ligger fokus inte på samverkan mellan lärare och elever utan på hur lärare samverkar med varandra kring frågor om bedömning.

Ett annat begrepp som återkommer i forskningslitteraturen är ”moderation”. ”Moderation ” kan innebära en extern inspektion där särskilt utvalda inspektörer besöker skolor för att kontrollera hur bedömningen genomförs (Harlen, 1994). En annan variant är så kallad ”statistical moderation” som innebär att lärares bedömning jämförs med något referenstest eller skalas mot något annat bedömningsinstrument (Gipps, 1994). Ett sådant exempel är läsförståelsetest som omvandlas till staninepoäng. Det begrepp som bäst fångar sambedömning i detta sammanhang är den beskrivning av ”moderation” som benämns ”social moderation” eller ”consensus moderation” (se till exempel Gipps, 1994; Linn, 1993) och handlar om att olika bedömare försöker nå konsensus i hur bedömning och betygssättning ska genomföras (Gipps, 1994). Begreppet ”moderation” kommer av ”att reglera” eller ”att plana ut extrema värden” (Sadler, 2013). Denna utplaning kan även åstadkommas genom att medelvärden av flera olika bedömares bedömningar beräknas men då kommer man inte åt de underliggande orsakerna till olikheterna. Syftet med ”social moderation” är att få högre reliabilitet och validitet i lärares bedömningar så att de både stämmer överens med intentioner i styrdokumenten och att de är jämförbara på lokal, regional och nationell nivå (Wilson, 2004). Att genom sambedömning nå ökad reliabilitet hävdas då det dels ger förutsättningar för diskussioner som kan leda till samsyn kring styrdokument och dels ger lärare möjlighet att kalibrera sina bedömningar gentemot andra lärares bedömningar.

> It is in the context of moderation that teachers act as a community of assessors: they talk about actual student work examples and examine how the work matches the expected features as specified in stated standards. And it is through such talk and the classification of the work against the standards that teacher judgement becomes "tuned in" or calibrated to achieve high levels of reliability or inter-rater consistency.
>
> (Klenowski & Wyatt-Smith, 2010b, sid. 115)

Begreppet ”moderation” används i vid bemärkelse i forskningslitteraturen och kan innebära att lärare träffas vid särskilda möten, antingen inom den egna skolan eller mellan olika skolor och har med sig exempel på elevarbeten på olika nivåer. ”Moderation” kan även genomföras ”on-line” vilket gör att lärare från geografiskt skilda skolor kan komma i kontakt med varandra (se till exempel Adie, 2010). ”Moderation” kan utföras före, under och efter själva bedömningstillfället och är även genomförbart som en fortlöpande process (Maxwell, 2002). Vid dessa möten kan detaljerade beskrivningar av kriterier och kravnivåer användas som stöd vid bedömning och lärare ges tillfälle att diskutera olika exempel på elevarbeten och gå igenom hur de motsvarar de specificerade kravnivåerna (standards). Genom sådan diskussion och klassificering ges förutsättningar för lärare att kalibrera sin bedömning mot andra vilket kan leda till en högre reliabilitet (Klenowski & Wyatt-Smith, 2010a).

Klenowski och Adie (2009) beskriver tre olika typer av ”social moderation”:

- ”The calibration model” vilken innebär att ett urval av elevarbeten bedöms individuellt av olika lärare. Lärarna diskuterar sedan sin respektive bedömningar med målet att nå konsensus och gemensam förståelse för kravnivåer (standards).

- ”The conferencing model” som innebär att en lärare individuellt bedömer elevarbeten. Ett urval av elevarbeten som representerar prestationer på olika kvalitativa nivåer väljs sedan gemensamt ut för diskussion med målet att nå konsensus och gemensam förståelse.

- ”The expert model” innebär att lärare bedömer samtliga sina elevarbeten och lämnar dem sedan vidare till en expert. Lärarna får sedan tillbaks feedback på i vilken utsträckning de har uppfattat och använt kravnivåer på det sätt som var avsett.

Med sambedömning i den betydelse som begreppet används i föreliggande studie avses lärare som tillsammans genomför eller diskuterar bedömning och/eller betygssättning och kan sägas omfatta de två första beskrivningarna av ”social moderation”. I båda dessa modeller är målet att nå konsensus kring bedömningen. I definitionen av sambedömning som avses i denna studie behöver målet nödvändigtvis inte vara konsensus utan kan även omfatta att synliggöra skilda synsätt vid bedömning och betygssättning av elevers prestationer.

Sammanfattning

I ett mål- och kriterierelaterat bedömningssystem där elevers kunskaper jämförs med på förhand formulerade kravnivåer ställs alltså stora krav på att lärare analyserar, tolkar och värderar kurs- och ämnesplaners formuleringar samt elevers prestationer. Parallellt med detta förväntas lärares bedömning och betygssättning vara likvärdig vilket innebär att man behöver vara överens i tre tolkningled, se Figur 2. Då olika forskningsrapporter och granskningar visar på brister i likvärdigheten krävs olika åtgärder för att åstadkomma en utökad likvärdighet. Det

blir då angeläget att förstå vilka effekter olika ansatser att nå ökad likvärdighet har och på vilka sätt de då bidrar till detta. Sambedömning är en metod som i flera sammanhang förespråkas som ett sätt att göra bedömning och betygssättning mer likvärdig och syftet med denna studie blir därför att förstå om, och i så fall på vilka olika sätt, sambedömning kan bidra till en ökad likvärdighet vid bedömning och betygssättning. Eftersom likvärdighet vid bedömning och betygssättning är en process som kan påverkas i flera led blir det därför även viktigt att förstå var i denna process som sambedömning kan ge effekter och om det finns andra faktorer som kan påverka utfallet.

Metod

Det finns en utbredd föreställning om att sambedömning kan leda till en mer samstämmig bedömning av elevers prestationer och sambedömning föreslås även som en strategi att öka likvärdigheten vid bedömning och betygssättning. Även om de argument som förs fram låter rimliga finns det behov av att undersöka det vetenskapliga stödet för sådana antaganden, vilket inte alltid förs fram i den allmänna diskussionen om sambedömning. Då likvärdighet vid bedömning och betygssättning, som tidigare visats, är en process som kan påverkas i flera led blir det även viktigt att förstå var i denna process som sambedömning kan ge effekter och om det finns andra faktorer som kan påverka denna process, så kallade modererande faktorer. Syftet med föreliggande studie blir därför att undersöka och förstå hur sambedömning kan bidra till en mer samstämmig och likvärdig bedömning och betygssättning. För att få en översikt över det aktuella kunskapsläget kring sambedömning och dess eventuella effekter valdes en systematisk litteraturstudie som vetenskaplig metod för rapporten (jfr Forsberg & Wengström, 2008). Metoden innebär att sökningar efter kvalitativa och kvantitativa studier genomförs, granskas och sedan sammanförs till en ny helhet. Genom att använda resultaten från vetenskapliga studier kan en djupare förståelse och en bredare bild skapas samtidigt som det ger möjlighet att identifiera eventuella kunskapsluckor (ibid.). Frågorna som denna forskningsöversikt försöker besvara blir:

- Vilka effekter kan sambedömning ge?
- Vilka faktorer kan påverka sambedömningsprocessen?

Litteratursökning

Relevanta studier söktes i en primär sökning genom sökmotorn Summon under september månad 2012. Denna sökmotor är kopplad till flera stora databaser, till exempel Academic Search Elite, Science Direct, PsychINFO och ERIC. De kombinationer av sökord som användes var teacher/assessment/teacher-based assessment/ national test/ moderation/ marking / reliability/ consensus/ validity. Eftersom antalet databaser kopplade till summon är stort och då flera av sökordskombinationerna var förhållandevis breda gav dessa sökningar väldigt många träffar. Dessa träffar söktes igenom och relevanta titlar sparades i ett första urval. Dessa studier plockades sedan fram eller beställdes i full-text för en mer grundlig genomläsning och katalogisering och utgjorde sedan ett andra urval. Ett antal studier gick inte att få fram av olika anledningar till exempel så hade ett antal artiklar plockats bort ur databasen ERIC. Ett antal studier som verkade intressanta för studien och som fanns angivna i referenslistor gick inte heller att få tag i då de var bortplockade från de adresser som angavs. Även manuella sökningar utfördes utifrån referenslistor med syfte att utvidga det möjliga resultatet. Under arbetets gång gavs ett temanummer om ”Moderation practice and teacher

judgement" ut av en betydelsefull tidskrift. De ingående artiklarna gicks igenom och i de fall de ansågs relevanta kom även dessa att inkluderas i denna studie.

Urvalskriterier

För att en studie skulle anses som relevant och att inkluderas i det andra urvalet sattes ett antal urvalskriterier upp. Följande urvalskriterier användes i urvalsprocessen:

1. Först och främst skulle studien vara empirisk, det vill säga den skulle ha studerat lärare som sambedömer, antingen som studiens huvudfokus eller som en del av en annan studie. Initialt användes ingen tydligt avgränsad definition av sambedömning så kriteriet för att inkluderas i studien var att lärare skulle genomföra eller diskutera bedömning av elevers prestationer tillsammans. Detta urvalskriterium innebar att studier som endast omfattade att elever var en del av bedömningsprocessen exkluderades.
2. Ett andra urvalskriterium var att studien skulle handla om sambedömning i grundskola, gymnasium eller vuxenutbildning (på grundskole- eller gymnasienivå) eller motsvarande och inte i förskola eller högre utbildning. Vilket skolämne sambedömningen utfördes i var inte begränsande.
3. Ett tredje urvalskriterium var att studien skulle vara publicerad och vetenskapligt granskad (peer-review) vilket till exempel innebar vetenskapliga artiklar, konferenspapers, böcker och doktorsavhandlingar.

Ingen begränsning i vilket utgivningsår studierna hade har gjorts.

Urval

Urvalet av studier genomfördes i olika steg. Under den primära litteratursökningen valdes 86 studier ut varav 24 stycken var dubbletter och sorterades bort direkt. Detta resulterade i att 62 studier ingick i det första urvalet. Författaren läste sedan studierna i det första urvalet i sin helhet för att få en noggrann uppfattning om innehållets relevans för resultatet utifrån urvalskriterierna. De studier som föll ifrån gjorde det främst eftersom det inte handlade om sambedömning i bemärkelsen att lärarna verkligen utförde bedömningen tillsammans eller för att de inte var empiriska. En del av dessa sparades ändå för att användas i bakgrund och i diskussionsdel. Efter en slutlig genomläsning återstod 23 studier som legat till grund för föreliggande studie.

Analysmetod

I denna studie används en kvalitativ ansats som metodiskt tillvägagångssätt. Syftet med studien är att förstå om och på vilket sätt sambedömning kan bidra till en likvärdig bedömning. Då utgångspunkten är att få en förståelse för vilka effekter sambedömning kan ge och vilka faktorer som eventuellt kan påverka denna process är blir det av underordnad betydelse hur starka olika effekter är i förhållande till varandra eller mängden belägg för varje enskild faktor. En anledning till detta ställningstagande är dels att det totala antalet studier är förhållandevis få och att stödet för varje enskild faktor är begränsat. Tillvägagångssättet är därför en kvalitativ innehållsanalys (se till exempel Graneheim & Lundman, 2004). I ett första steg lästes samtliga studier igenom noggrant i syfte att avgöra ifall de uppfyllde urvalskriterierna. I nästa steg plockades relevant information ut och strukturerades i en tabell.

I tabellen kategoriserades informationen under följande rubriker: urvalsgrupp, forskningsfrågor, forskningsdesign, resultat och slutsatser. Slutligen gicks samtliga studier igenom och jämfördes i syfte att upptäcka eventuella mönster.

Resultat

Kontext och forskningsdesign

Av de 23 studierna har sju genomförts i Australien, främst i delstaten Queensland, sju i Storbritannien, sex i USA, främst i Kalifornien, och en studie vardera ifrån Portugal och Nya Zealand. En av studierna har genomförts som en komparativ studie i Australien (Melbourne) och i Hong Kong. Ett av urvalskriterierna var att lärarna som sambedömde skulle undervisa elever i motsvarande grundskola, gymnasium eller vuxenutbildning motsvarande denna nivå. Merparten av studierna (17) har innefattat lärare från olika nivåer skolsystemet med en tonvikt mot grundskolan. Tre studier har omfattat lärare från motsvarande vårt gymnasium, två högstadiet och en studie har enbart omfattat lågstadielärare. Ingen av studierna som ingår i föreliggande studie har studerat vuxenutbildning. Vidare har merparten av studierna varit huvudsakligen kvalitativa och inneburit att lärare som deltagit i sambedömning blivit intervjuade före, under och efter sambedömningsinsatsen. Observationer av sambedömningsmöten har ibland genomförts och ibland har olika typer av dokument och andra underlag samlats in. Vissa studier har kompletterat sådana data med enkäter. Det är lärare och rektorer som är de huvudsakliga informanterna men ibland har andra personer med särskild kunskap eller ansvar för bedömningsfrågor ingått som informanter. Det är betydligt färre av studierna som är experimentella och där effekter av sambedömningsinsatser studerats och mätts på ett systematiskt sätt. Endast tre studier har varit kvantitativa i bemärkelsen att de utifrån ett underlag kunnat avgöra i vilken omfattning sambedömningen leder till en ökad bedömaröverensstämmelse. Gemensamt för dessa studier är att det underlag, på vilket dessa slutsatser dras, inte redovisas. Försök att rekvirera detta underlag lyckades inte heller. Detta innebär att resultaten till betydande del bygger på lärares uppfattningar av sambedömning och i mindre omfattning på en mer objektiv beskrivning av hur det förhåller sig. Till exempel kan lärare uppfatta att likvärdigheten i bedömningen på en skola har ökat utan att man faktiskt har belägg för att det också förhåller sig på detta vis.

Vilka effekter kan sambedömning ge?

Analysen av den forskning som ligger till grund för studien resulterade i att fyra kategorier av effekter kunde urskiljas. Dessa kategorier är:

- Effekter på lärares förståelse för kriterier och kvaliteter i elevers prestationer
- Effekter på reliabiliteten vid bedömning
- Effekter på lärares kompetens
- Effekter på de kollegiala relationerna

Gränserna mellan dessa kategorier är inte skarpa vilket medför att resultat som kan sorteras in under en kategori även kan höra hemma i en annan. Till exempel kan man anta att om en grupp lärare får en djupare förståelse för vad bedömningskriterierna innebär så ökar även kompetensen hos den enskilde läraren att bedöma sina elevers prestationer i förhållande till dessa.

Effekter på lärares förståelse för kriterier och kvaliteter i elevers prestationer

Att lärare tillsammans diskuterar bedömning av elevers prestationer i förhållande till i förväg uppställda mål och kriterier kan resultera i samsyn på vad dessa mål och kriterier innebär och

hur de kan användas (Connolly, Klenowski & Wyatt-Smith, 2012; Hall & Harding, 2002; Reid, 2007; Roberts, Wilson & Draney, 1997).

En aspekt av detta är att olika sätt att uppfatta kriterier och kravnivåer blir synliggjorda när lärare sambedömer. I ett projekt genomfört i England (Black, Harrison, Hodgen, Marshall & Serret, 2011) där lärarna skulle sambedöma elevarbeten blev det en överraskning för lärarna hur olika de bedömde och uppfattade kriterier och kravnivåer. I andra studier genomförda i Queensland, Australien (Adie, Klenowski & Wyatt-Smith, 2012; Klenowski & Adie, 2009) visade det sig att lärarna både hade svårt att skilja begreppen kriterium och kravnivå (standards) ifrån varandra samt att lärarna använde olika metoder för att hantera dessa vid bedömning (Klenowski & Adie, 2009). Det fanns även exempel på att lärare uppfattade kravnivåer som en minimigräns för vad en elev skulle prestera på den aktuella nivån medan andra uppfattade att kravnivåerna representerade en typisk prestation (Klenowski & Adie, 2009). När elevers olika prestationer skulle sammanvägas till ett betyg använde olika lärargrupper olika metoder för att göra detta. En del beräknade ett medelvärde av resultat på olika underlag, andra gjorde en helhetsbedömning av allt tillgängligt underlag medan ytterligare andra viktade olika kriterier mot varandra, ofta utifrån egna preferenser (ibid.).

När lärare har olika uppfattningar om på vilken nivå en elevlösning ska placeras uppstår diskussioner som ger stora möjligheter till lärande (Adie et al., 2012, Black, Harrison, Hodgen, Marshall & Serret, 2010). Det visar sig att det är genom att diskutera och förhandla som lärarna får en gemensam förståelse för de subtila skillnader som skiljer kravnivåerna åt (Adie et al., 2012). Liknande resultat återfinns även i andra studier (Black et al., 2010, 2011). Black et al. (2011) menar att det är utmanande diskussioner som är avgörande för att kunna lyfta fram och synliggöra frågor kring kravnivåer och kvaliteter i elevers arbete och det verkar som att det är just genom att fokusera och analysera elevers prestationer som en gemensam förståelse utvecklas. Falk & Ort (1998) beskriver ett projekt där lärare getts möjlighet att sambedöma och intervjuade lärarna uppger att dialogen med kollegor om bedömning, kravnivåer och elevarbeten var det som var det mest givande och att dessa diskussioner hjälpte dem att utveckla en samsyn, både i förhållande till den egna undervisningen men även som ett stöd i ett vidare sammanhang. Lärarna menade även att diskussionerna gav dem möjlighet att bli medvetna om sina egna uppfattningar, om intentionerna i styrdokumenten samt hur dessa förhöll sig till egna och andras uppfattningar i ett vidare sammanhang (ibid.).

Adie et al. (2012) drar slutsatsen att sambedömning blir viktigt eftersom den synliggör processer och skapar en arena där lärarna tvingas att förklara och rättfärdiga sina beslut i diskussioner kring både kriterier och olika kvaliteter i elevers prestationer. Sambedömningen föranleder då deltagarna att motivera sina ställningstaganden och att ompröva och fördjupa sin förståelse av kravnivåer på ett sätt som därmed kan leda till ändrad bedömning av elevers prestationer.

Ett exempel som visar att sambedömning kan leda till en gemensam förståelse för kriterier och kvaliteter i elevers kunskaper är en komparativ studie (Davison, 2004) där man gjort en jämförelse mellan lärare från Melbourne och lärare från Hong Kong. Det man studerat är lärarnas uppfattningar om bedömning och deras bedömning av elevtexter. I Melbourne har man en utbredd tradition av att sambedöma och lärarna får även träning i att utföra detta. I Hong Kong finns inte denna tradition och det fanns inte heller vid tidpunkten för studien gemensamma bedömningskriterier för skriftligt arbeten eller någon uttalad möjlighet för lärare att delta i sambedömning. I studien fick lärarna bedöma elevtexter på samma sätt de vanligtvis brukade bedöma liknande uppgifter. Efter att ha gjort denna bedömning på egen

hand blev de sedan satta i grupper för att diskutera sin bedömning. Det man fann i studien var att bland Hong Kong lärarna fanns det en betydligt större variation i hur man bedömde, vad man tillmätte betydelse och lärarna var sinsemellan mindre överens än vad Melbourne-lärarna var. Författarna drar slutsatsen att detta dels beror på avsaknaden av gemensamma bedömningskriterier men även på att lärarna har begränsade möjligheter att delta i sambedömning, fortbildning eller andra eller gemensamma aktiviteter kopplade till bedömning. Också lärarna från Melbourne visade på variationer i sina bedömningar. Även om samtliga lärare använde sig av officiella bedömningskriterier fanns olikheter i hur man uppfattade och använde sig av dessa. Det fanns exempel på samma elevtexter som gavs betygen A-D av olika lärare men även exempel på elevtexter som fick samma betyg men av helt olika anledningar. Det fanns även variationer i hur man utförde själva bedömningen med en ytterlighet som bestod av lärare som ”prickade av kriterier” och en annan som förlitade sig mer på helhetsbedömning baserad på sin professionella men även kontext-bundna kunskap och lärare som slets mellan dessa två angreppssätt. I dialogen med andra sökte lärarna vägar att ta professionella beslut. Dessa exempel visar att sambedömning kan öka lärares samsyn kring vad kriterier står för samt hur dessa kan bedömas i förhållande till elevers prestationer. Variationerna, även bland lärarna med erfarenhet av sambedömning, kan dock ses som ett exempel på att även om bedömningar på ett ytligt plan är att betrakta som samstämmiga så kan de underliggande avvägningarna som ligger bakom ändå skilja sig åt.

Att lärare får en samsyn kring kriterier och kravnivåer genom sambedömning kan även leda till att lärare upplever att bedömningen blir mer likvärdig och att lärare känner en större säkerhet i att bedöma elevers prestationer (Connolly et al., 2012). Wyatt-Smith et al. (2010) argumenterar dessutom för att sambedömning är ett särskilt betydelsefullt inslag för att skapa en samsyn både kring kravnivåer och som en väg att utveckla ett gemensamt språk vid bedömning av elevers prestationer i förhållande till dessa nivåer.

Effekter på reliabiliteten vid bedömning

Att lärares bedömningar av elevers prestationer blir mer reliabla som en direkt följd av sambedömning finns dokumenterat i några studier (de Eca & Torres, 2005; Falk & Ort, 1998; Syverson, 2009). I dessa studier menar man att lärares bedömningar överensstämmer till mellan 80 och 90 procent efter att de deltagit i sambedömning. Studierna är dock inte redovisade på ett sådant sätt att data har kunnat verifieras. Att reliabiliteten ökar är däremot något man i andra studier lyfter fram som en *trolig* effekt av sambedömning (Adie et al., 2012; Clarke & Gipps, 2000). Adie et al. (2012) drar slutsatsen att reliabilitet och validitet ökar till följd av att lärare deltar i diskussioner om undervisning och bedömning men hävdar att detta måste verifieras med andra typer av forskningsstudier än de som hittills är genomförda. Clarke och Gipps (2000) har samlat in data från tre stora projekt genomförda med lärare i årskurs 2 och 6 i England. De har bland annat studerat i vilken mån lärares bedömningar överensstämmer med varandra genom att intervjua lärare samt genomföra observationer av sambedömningsmöten. De menar att trots att det finns olikheter mellan hur olika lärargrupper går till väga när de sambedömer är detta en kritisk process för att uppnå en konsekvent bedömning, särskilt om sambedömningen genomförs med deltagare från olika skolor och med lärare med olika bakgrund.

Black et al. (2011) har genomfört en longitudinell explorativ studie över 2,5 år där en del av de genomförda insatserna inkluderade sambedömning. Lärarna i studien uppgav att de upplevde sambedömning som värdefull för att få en förståelse för kriterier och kravnivåer. Denna ökade förståelse ledde i sin tur till att lärarna kände sig säkrare i att hantera frågor

kopplade till likvärdighet på ett sätt som de uppgav även var positivt för eleverna. Studien omfattade däremot ingen systematisk undersökning av om reliabiliteten ökade. På liknande sätt upplever lärare som har deltagit i andra studier, där sambedömning ingår, i stor utsträckning att detta är en process som bidrar till att likvärdighet och rättvisa vid bedömning ökar (Bolt 2011; Connolly et al., 2012). I en studie genomförd av Connolly et al. (2012) uppgav hela 75 procent av de deltagande lärarna att de uppskattade att de fått en samsyn kring kravnivåer och att detta gav dem en känsla av att deras bedömning blev likvärdig och rättvis samt att detta var en av aspekterna av sambedömning som de uppskattade mest. Det fanns även en mindre andel lärare som inte upplevde att sambedömning bidrog till en sådan överensstämmelse. De främsta anledningarna till detta, uppgav de, var inte sambedömningen i sig utan berodde mer på ett missnöje med faktorer kopplade till kravnivåerna (Connolly et al., 2012). Missnöjet handlade om att bedömningskriterierna i sig var formulerade på ett sätt som inte tog hänsyn till sådana faktorer som kan påverka hur elever presterar på bedömningsuppgifter. Exempel på sådana faktorer var olikheter i den kulturella kontexten samt vad lärarna betonat i undervisningen men det fanns även lärare som menade att sambedömning inte gav så mycket i förhållande till den tid som avsattes för detta. Även en studie av Bolt (2011) ger liknande resultat. I samband med genomförandet av ett projekt följdes en grupp lärare (fokus-lärare) på nära håll (n=383) genom intervjuer, observationer och insamling av dokument. Det skickades även ut enkäter till alla lärare som deltog i projektet (n=10 036, svarsfrekvens 26%). Fokus-lärarna upplevde att man var bättre rustade att göra konsekventa bedömningar efter att ha deltagit i sambedömning. Enkät-lärarna däremot hade överlag inte samma positiva upplevelse och menade i större utsträckning att projektet inte bidragit till en ökad samstämmighet. Det framgår inte tydligt vad detta beror på men det kan finnas förklaringar i hur man deltagit i projektet och vilken roll man haft.

Det finns även studier som pekar på att sambedömning inte har någon större påverkan på reliabiliteten vid bedömning. Baird, Greatorex och Bell (2004) har gjort en studie där man undersökt hur olika typer av diskussioner påverkar hur lärare bedömer essäer. Tre grupper lärare av lärare studerades varav den ena gruppen endast fick tillgång till elevtexter och bedömningsstöd i form av bedömningskriterier och exempel på redan bedömda texter. En andra grupp fick samma material tilldelat sig och de fick även delta i ett hierarkiskt strukturerat möte där en expertlärare presenterade hur de olika elevarbetena skulle bedömas och hur bedömningsanvisningarna skulle tolkas. Den tredje gruppen fick materialet tilldelat sig under ett möte där de tillsammans skulle diskutera sig fram till en konsensus kring hur elevtexterna skulle bedömas. Efter dessa förberedelser fick varje lärare enskilt bedöma 150 nya elevtexter. Utfallet av bedömningen jämfördes sedan och man fann att skillnaderna mellan hur man i de olika grupperna bedömt elevtexterna var relativt liten. Gruppen av lärare som deltagit i de hierarkiskt strukturerade mötena var de som hade lägst spridning i sina bedömningar. Efter bedömningssituationen intervjuades lärarna och samtliga uppgav då att de satte stort värde på alla former av koordination. Lärarna i den första gruppen, som inte deltagit i något möte innan bedömningen, uppgav att man saknat möjligheten till diskussion. Lärarna från samtliga grupper uppgav även att man ansåg att bedömningsanvisningarna var mer användbara än de bedömda elevarbetena. Argument om att icke-hierarkiska diskussioner leder till större lärande och till en mer reliabel bedömning verifierades alltså inte i denna studie utan det visade sig att bedömningsanvisningar har en stark standardiserande effekt även utan sambedömning. Morgan och Watson (2002) har gjort liknande erfarenheter i en studie av hur lärare bedömer elevarbeten i matematik. I studien ingick både lärare som hade erfarenhet av sambedömning och lärare som inte hade någon erfarenhet av detta. Det man fann var att förhållandevis lika elevlösningar uppfattades på olika sätt av olika lärare och att de bedömningar lärare med erfarenhet av sambedömning gjorde inte skilde sig från de

bedömningar lärare som inte hade denna erfarenhet gjorde, varken i fråga om betyg eller i hur man uppfattade kravnivåer. Skillnaderna bestod till största delen av hur man uppfattade det matematiska innehållet. Också Connolly et al. (2012) har genom en ”blind-review” visat att lärare utan erfarenhet av sambedömning gjorde bedömningar som var jämförbara med bedömningar utförda av med sådan erfarenhet.

Effekter på lärares kompetens

Sambedömning kan även ge effekter i form av ökad kompetens hos den enskilde läraren som deltar i denna och i förlängningen därmed även för elever. Reid (2007) fann att lärare som medverkat i sambedömning kände en större säkerhet när de bedömde elevers prestationer. Detta i sin tur fick dem att bli tydligare i sin kommunikation med elever om hur goda prestationer kunde se ut genom att visa exempel och göra kopplingar mellan bedömningskriterier och målen för undervisningen. Det finns även stöd för att lärare som deltar i sambedömning får en ökad bedömnings- och undervisningskompetens eftersom de kollegiala diskussionerna bidrar till att fokusera och reflektera kring faktorer som är centrala i lärares yrkesutövning (Adie, 2012; Bolt, 2011; Falk & Ort, 1998; Malone, De Lucchi & Long, 2004; Roberts et al., 1997). Det kan även handla om att lärare som ett resultat av de fördjupade diskussionerna om elevers prestationer i förhållande till kriterier och kravnivåer själva får ökade ämnes- och ämnesdidaktiska kunskaper. Sambedömning kan även bidra till att lärare bättre kan förstå sig själv som bedömare (Adie, 2012).

När lärare sambedömer eller diskuterar elevers prestationer erbjuds möjlighet till reflektion kring olika frågor kopplade till undervisning, lärande och bedömning, både formativ och summativ (Black et al., 2011; Reid, 2007). Genom att i sambedömning fokusera elevers prestationer kan lärare bli bättre på att hitta belägg för hur elevers kunnande svarar mot uppsatta mål och kriterier och att använda detta för att bättre planera, och genomföra undervisning samt använda bedömning på ett sätt som svarar mot elevers behov (Bolt, 2011; Falk & Ort, 1998). Malone et al. (2004) menar att detta sker eftersom att man i sambedömningen uppmuntras att sätta ord på vad som är nästa steg i elevernas lärande. Även effekter så som insikten att en lektions värde främst består i hur den bidrar till elever förståelse snarare än hur läraren känner att lektionen gått kan ses (Malone et al., 2004). Clarke och Gipps (2000) har som en del i en större studie, där de undersöker hur lärare bedömer och hur konsekventa deras bedömningar är, genomfört intervjuer med lärare. Lärarna i studien uppfattar att sambedömning är en tidskrävande men nödvändig process då den har en direkt påverkan på elevers lärande och lärarnas egen undervisning. Adie et al. (2012) beskriver hur lärare när de började med sambedömning hade olika uppfattningar om detta medan man efter hand som projektet fortskred mer kom att se sambedömning som en process som startar redan vid uppgiftskonstruktionen.

Ett exempel på att sambedömning kan ge lärare en ökad bedömarkompetens är en studie rapporterad av Roberts et al. (1997). I studien beskrivs ett system, SEPUP Assessment System, som omfattar både undervisning och bedömning. Ett antal skolor fick använda sig av båda delarna av systemet medan andra endast använde sig av undervisningsdelen. Som kontrollgrupp fanns även lärargrupper som fortsatte att undervisa så som de brukade. I studien undersökte man hur lärarna ändrade sin praktik avseende bedömning, kollegialitet och undervisning. När det gällde undervisning kunde man inte se några skillnader mellan de båda grupperna, däremot när det gällde bedömningspraktik och kollegialitet. Resultaten i förhållande till lärares bedömningspraktik, där man undersökte i vilken utsträckning lärarna använde sig av alternativa strategier för bedömning samt hur användbara de uppfattade att dessa strategier var, var de mest intressanta. Båda grupperna uppfattade vid projektets start att

alternativa bedömningsstrategier var användbara. Vid projektets slut ett år senare kunde man däremot se signifikanta skillnader mellan de båda grupperna. I de grupper som fått stöd både i undervisningsfasen och i bedömningsfasen, vilken bland annat inkluderade sambedömning, använde man sig i ökad omfattning av öppna frågeställningar vid bedömning av elever och mindre av slutna frågeställningar, medan det var tvärtom i den andra gruppen. I den andra gruppen hade lärarna vid läsårets slut en fortsatt positiv inställning till alternativa bedömningsformer, även om man inte använde sig av dessa i någon större utsträckning, medan man i den första gruppen hade en något mindre positiv inställning. Denna minskning var dock inte signifikant. Resultaten indikerar att de lärare som endast använde sig av undervisningsdelen i systemet inte var tvungna att ta itu med de dilemman som uppstod till följd av den ändrade bedömningspraktiken och att de därmed höll kvar en positiv inställning till denna. Just möjligheten till möten och diskussioner kring bedömning lyftes fram som den huvudsakliga faktorn för att lösa dessa dilemman och ge varandra stöttning att fortsätta utvecklas. Man såg även att sambedömningen spelade en nyckelroll i att bygga förståelse för bedömning och för att använda bedömningsanvisningar på ett effektivt sätt. Att endast ge lärare bedömningsanvisningar räckte inte för att lärarna skulle få en förståelse för vad en förändrad bedömningspraktik innebär. Genom att delta i sambedömning där lärarna fick möjlighet att utveckla en samsyn lärde sig lärarna även vad en förändrad undervisnings- och bedömningspraktik i linje med rådande styrdokument innebar. Konsensus blev att fortlöpande stöttning genom sambedömning var en kritisk faktor för att få lärare att använda både undervisnings- och bedömningssystemet och att sambedömningen gav dem det nödvändiga stöd de behövde för att förändra sin undervisning och bedömningspraktik.

Ett annat exempel på att sambedömning kan bidra till en ökad bedömarkompetens är en studie genomförd av Black et al. (2010). Deras resultat visar att sambedömningsträffarna bidrar till att avslöja hur pass användbara de bedömningsuppgifter lärarna använder är. Det visade sig när lärarna skulle bedöma elevarbetena gemensamt, utifrån mål och bedömningsanvisningar, att de upptäckte att en del av de uppgifter de använt var begränsande i den bemärkelsen att de inte gav eleverna möjlighet att visa sitt kunnande i tillräckligt stor utsträckning medan andra var så pass svåra att de av den anledningen inte bjöd in eleverna att visa sitt kunnande. Vikten av att verkligen använda lämpliga uppgifter som ger möjlighet att särskilja olika förmågor vid bedömning blev synliga först i samband med sambedömningen när de skulle bedömas på ett sätt som krävde enighet med kollegor (Black et al., 2011).

Som tidigare redovisats kan lärare som deltagit i sambedömning utveckla en samsyn kring de kriterier och kravnivåer som elevers prestationer ska bedömas utifrån (se till exempel Connolly et al., 2012). Detta i sig ger därmed även den enskilde läraren en ökad bedömarkompetens. Det finns även stöd för att lärare, genom sambedömning, får en ökad kompetens att tolka mål och kriterier, att bedöma elevers prestationer och att på ett tydligare sätt kommunicera kravnivåer och bedömning med elever (Falk, 1998; Falk & Ort, 1998; Roberts et al., 1997).

Sambedömning ger även lärare möjlighet att fördjupa sina ämneskunskaper och ämnesdidaktiska kunskaper. Malone et al. (2004) beskriver en studie där lärare sambedömde elevers arbete i naturvetenskap. Diskussioner om elevers förståelse av centrala begrepp bidrog till att lärarna utvecklade insikter som fördjupade deras egen förståelse av dessa. Liknande erfarenheter har gjorts av både Limbrick och Knight (2005) och Reid (2007). I Reids (2007) studie deltog lärare både från tidigare skolår och från senare skolår i sambedömning av elevers skrivna texter. Båda grupperna fick ökade kunskaper till följd av sambedömningen men av olika slag. Lärarna från de tidigare skolåren ökade sina ämneskunskaper genom

diskussionerna med lärarna från de senare skolåren medan dessa i sin tur utvecklades i att kunna uttrycka denna ämneskunskap samt lärde sig om undervisning av yngre barn och vad som kännetecknar denna. I en annan studie (Falk & Ort, 1998) gavs lärarna möjlighet att genom sambedömning se hur det som uttrycktes i kravnivåer tog sig uttryck i autentiska elevarbeten. Detta gav dem även möjlighet att både utveckla en gemensam förståelse och ett gemensamt språk för att diskutera viktiga frågor knutna till deras undervisningsämnen.

Effekter på de kollegiala relationerna

Att samarbeta med andra lärare genom sambedömning kan också ge effekter på de kollegiala relationerna på olika sätt. Det kan handla om allt ifrån att man känner tillhörighet inom gruppen och att man får stöd i att lösa dilemman som uppstår i det dagliga arbetet till att det kan bidra till en känsla av professionalitet. Det finns även studier som lyfter fram det faktum att sambedömning kan medföra att enskilda lärare istället för gemenskap känner ett utanförskap. Utanförskapet kan vara en följd av att man har en avvikande uppfattning gentemot sina kollegor eller att lärare som deltagit i sambedömning tillsammans med lärare från andra skolor har svårt att kommunicera den nya kunskap och de nya erfarenheter man gjort när man kommer tillbaks till den egna skolan

Ett exempel på en studie i vilken man sett kollegiala effekter av sambedömning är rapporterad av Connolly et al. (2012). I denna studie uppger lärarna att samarbete med andra var välgörande för de kollegiala relationerna samtidigt som det gav självförtroende, en känsla av tillhörighet och acceptans samt bidrog till en känsla av professionalitet. Ett annat exempel är den amerikanska studie av Roberts et al. (1997), som beskrivits tidigare. I SEPUP-projektet ingick sambedömning som en del av den bedömningsprocess som en del av lärargrupperna inom projektet fick använda sig av. Det visade sig att de lärargrupper som använde sig av både undervisnings- och bedömningsdelen uppvisade en högre kollegialitet jämfört med övriga grupper som inte använt sig av projektets bedömningsdel. Att dessa grupper i större utsträckning än övriga höll kvar vid de nya strategierna för undervisning och bedömning tillskrevs just möjligheten de haft att samverka och diskutera frågor om bedömning. Sådana kollegiala effekter var mer uppenbara bland de grupper som kom från samma distrikt, där samarbete skedde även mellan sambedömningstillfällena och där man hade ett starkt ledarskap i gruppen. Forskarna betonar även att de grupper som utgjorde sambedömningens kärna hade en viktig roll som kollegialt stöd under hela implementeringen av det nya systemet och att stark kollegialitet i sin tur var något som bidrog till större framgång med sambedömningen.

Betydelsen av om lärarna i de grupper som sambedömer kommer från samma skola eller inte verkar vara oklar och i olika studier dras olika slutsatser. I de studier där man intervjuat lärare om deras uppfattningar om sambedömning uppger de ofta att det är mest värdefullt med sambedömning tillsammans med lärare på den egna skolan eller i det egna skolområdet (Bolt, 2011; Hall & Harding, 2002; Roberts et al., 1997). Till exempel visar Bolt (2011) att lärarna var mer positivt inställda till sambedömning inom den egna skolan än till sambedömning mellan skolor. Hon menar även att de grupper som etablerats i samband med genomförandet av ett projekt var svagare än de grupper som var etablerade redan innan. Det finns dock andra studier där man argumenterar för värdet av att sambedömning sker mellan lärare från olika skolor eller distrikt eftersom det innebär ökade möjligheter att tillföra nya perspektiv i diskussionerna(Adie, 2010, 2012, 2013; Falk & Ort, 1998; Reid, 2007). Clarke och Gipps (2000) menar att sambedömning mellan lärare från olika skolor och stadier är nödvändig om målet är en likvärdig bedömning. Adie (2010, 2012, 2013) har även studerat sambedömning i en on-line kontext. Resultatet från dessa studier ligger i linje med vad man sett i andra studier

men hon betonar att online-kontexten ger möjlighet att komma i kontakt med lärare bortom den egna skolan.

Oavsett hur gruppen är sammansatt så verkar de kollegiala relationerna i gruppen vara av avgörande betydelse för hur lärare upplever sambedömningen. Hall och Harding (2002) studerade sex skolor i England under två år med fokus på deras bedömningspraktik och fann två konceptuellt olika ansatser, ett med fokus på samarbete och ett mer individualistiskt. I vissa skolor fanns en mer uttalad gemenskap och samsyn kring styrdokument, undervisning och bedömning och där lärarna utförde flera arbetsuppgifter tillsammans, bland annat använde de sig av sambedömning inom skolan, medan lärarna på andra skolor arbetade mer individuellt med dessa frågor. Hall och Harding (2002) argumenterar för att kvaliteten på undervisningen i klassrummet starkt påverkas av kvaliteten på de professionella relationer lärarna har med sina kollegor utanför klassrummet och menar att det finns större förutsättningar för detta i en miljö som präglas av samarbete. Detta stöds även av Malone et al. (2004) som menar att den kollegiala interaktionen till följd av sambedömning ”spiller över” på den dagliga verksamheten.

Little, Gearhart, Curry och Kafka (2003) har studerat tre projekt som har sambedömning av elevers prestationer som ett centralt inslag. De drar slutsatsen att sambedömning ger lärare ökade möjligheter att lära och odla en kultur som präglas av professionalitet. De främsta anledningarna till detta är (1) att lärare tillsammans fokuserar elevers lärande och sin egen undervisningspraktik, (2) att konkreta elevprestationer hamnar i fokus och blir föremål för diskussion och (3) att konversationen struktureras. I de olika projekten kom strukturen till uttryck på olika sätt. Det kunde vara genom tydlig samtalsordning, genom fokus på vad man verkligen såg i elevernas arbeten snarare än vilket betyg en prestation var värd och genom möjligheten att ge varandra feedback, både positiv feedback men även ifrågasättande genom de frågor som ställdes.

En avgörande faktor för att sambedömning ska upplevas som värdefull verkar vara känslan av tillhörighet i gruppen. Det bör dock noteras att detta inte är en självklarhet. Ett exempel är en studie från Nya Zealand av Limbrick och Knight (2005) som rapporterar om ett antal lärare som upplevde sambedömningsprocessen frustrerande och att den inte främjade kollegialitet. En del lärare upplevde ett tävlingsmoment mellan olika deltagare och en ibland aggressiv stämning med deltagare som inte ville kompromissa. Andra negativa synpunkter handlade om olika förväntningar på träffarna och synpunkter på gruppstorlek och gruppsammansättning. Även om många lärare hade positiva erfarenheter av sambedömningsprocessen och upplevde den som en möjlighet till värdefulla kollegiala diskussioner så är denna och liknande studier en viktig påminnelse om att olika lärare kan uppleva samma diskussioner på olika sätt.

Vilka faktorer kan påverka sambedömningsprocessen?

Som vi kunnat konstatera kan sambedömning bland annat leda till att lärare får en gemensam syn på styrdokumenten och vad de innebär samt att lärare får en ökad bedömarkompetens. Det är rimligt att anta att sådana effekter inte är något som uppstår av sig själv endast utifrån det faktum att lärarna sambedömer utan att det är en process som sker gradvis och kan påverkas av flera andra faktorer. Studiens andra forskningsfråga handlar om att identifiera sådana faktorer som kan påverka sambedömningsprocessen. Vid den genomgång av forskning om sambedömning som genomförts i föreliggande studie har ett antal sådana faktorer identifierats. Dessa faktorer kan delas in i tre huvudsakliga grupper: (1) faktorer som beror på den yttre strukturen och organisationen; (2) faktorer som beror på och är kopplade till bedömningssituationen och (3) faktorer kopplade till deltagarna i sambedömningsprocessen.

Struktur och organisation

Faktorer som verkar påverka utfallet av sambedömningsprocessen är sådant som är kopplat till den struktur och organisation som finns runt själva sambedömningssituationen. En sådan betydelsefull faktor är om det finns en ledare i gruppen samt hur denne tar sig an en sådan roll. Roberts et al. (1997) har rapporterat om implementeringen av ett projekt i Kalifornien, USA, där man införde ett helt nytt system som inkluderade både undervisning och bedömning och där sambedömning var en del. I grupper med ett starkt ledarskap kunde man även se de största kollegiala effekterna och stark kollegialitet var även en faktor som var avgörande för utfallet av sambedömningen. Little et al. (2003) menar även att en god struktur och ett tydligt ledarskap är en särskilt central faktor för att hålla fokus på elevarbetena och för att fördjupa diskussionerna kring undervisning och elevers lärande. De menar att i de fall man såg exempel på grupper som lyckades få till öppna, kritiska och konstruktiva diskussioner så var detta ett resultat av ett strategiskt och skickligt ledarskap. De lyfter även fram betydelsen av protokoll, vikten av att följa de riktlinjer man satt upp och betonar därmed vikten av att någon sätter ramarna för och håller ihop diskussionerna för att dessa ska leda till utveckling.

Ytterligare en faktor som kan kopplas till organisationen och strukturen kring sambedömningsprocessen är vad syftet med diskussionerna är. Vid diskussion om olika elevers prestationer kan syftet vara att nå konsensus, det vill säga att man enas om hur en prestation ska bedömas. Ett annat syfte kan vara att synliggöra olika sätt att uppfatta och tolka en prestation. Adie (2012) menar att de situationer när lärarna är oeniga kring hur en prestation ska bedömas var de situationer som erbjöd de största möjligheterna till lärande.

Susan Bolts (2011) forskning, som tidigare beskrivits, lyfter bland annat fram det faktum att lärare var mer positivt inställda till sambedömning inom den egna skolan än till sambedömning mellan olika skolor. Samtidigt finns det andra studier där man argumenterar för värdet av att sambedömning sker mellan lärare från olika skolor då det innebär större möjligheter att tillföra nya perspektiv och ger en bredare förståelse för styrdokument och bedömning av elevers prestationer (Adie, 2013; Falk & Ort, 1998). Att lärare är mer positivt inställda till sambedömning inom den egna skolan är något som inte får stöd i samtliga studier. Reid (2007) har funnit att lärare mest uppskattar att diskutera bedömning med lärare från andra skolor eller skolområden.

Ett sätt att komma i kontakt med lärare från andra skolor eller skolområden är att sambedömningen sker online vilket studerats av Adie (2010, 2012, 2013). Denna kontext är inte helt annorlunda jämfört med att träffas fysiskt även om vissa faktorer kopplade till denna kontext kan få större eller mindre betydelse (Adie, 2013). Även om sambedömning med lärare utanför den egna skolan kan ge nya perspektiv, och därmed ökad förståelse, menar Black et al. (2010) att de erfarenheter man gör när sambedömning sker mellan lärare som inte jobbar på samma skola kan vara svåra att kommunicera när man kommer tillbaks till den egna skolan. Detta är en faktor som även lyfts fram i andra studier (Bolt, 2011).

Bedömningssituationen

Hur själva bedömningstillfället utformas och vilka hjälpmedel som används som stöd för bedömningen kan påverka utfallet av sambedömningsprocessen. En faktor som kan påverka sambedömningen, är vilken typ av elevprestationer det är som bedöms. Lesley Reid (2007) har i en studie visat att om de elevarbeten som bedöms är en färdig produkt så tenderar bedömningen och samtalet att få en summativ karaktär. Om sambedömningen istället utgår ifrån elevarbeten som ännu inte är färdiga, som till exempel ett första utkast till en uppsats,

kan effekten bli att diskussionen får en mer formativ och framåtsyftande prägel vilket kan hjälpa lärarna att sammankoppla den formativa och den summativa bedömningen. Falk och Ort (1998) menar även att när elevarbetena är valda på ett sådant sätt att de representerar en stor bredd av prestationer och elever med olika bakgrund så får lärare erfarenheter som går utanför den egna lokala kontexten som hjälper dem att göra mer nyanserade bedömningar.

Vilka hjälpmedel som finns att tillgå som stöd för bedömningen och hur dessa används kan påverka utfallet av sambedömningen. Exempel på sådana hjälpmedel kan vara de kravnivåer som arbetena bedöms mot (jfr till exempel dagens kunskapskrav), bedömningsanvisningar, bedömningsmatriser och på förhand bedömda elevarbeten som exemplifierar en viss nivå av kunnande. Studier utförda av en forskargrupp i Queensland, Australien, (Klenowski & Adie, 2009; Wyatt-Smith et al., 2010) visar att lärare vid sambedömning kan använda hjälpmedel på olika sätt och att dessa riskerar att konkurrera om lärarnas uppmärksamhet och till och med komplicera bedömningen. Forskargruppen såg exempel på lärare som föredrog att använda de tillhandahållna hjälpmedlen: bedömningsanvisningar, på förhand bedömda elevarbeten och kravnivåer, som utgångspunkt för bedömningen. Lärarna genomförde bedömningen och gick sedan till kravnivåerna för att diskutera olika uppfattningar innan man till slut enades om ett betyg. I dessa fall var kravnivåerna själva utgångspunkten för diskussionen. Forskarna såg även exempel på att de på förhand bedömda elevarbeten användes som ett definitivt exempel på hur en lösning på en viss nivå skulle se ut. Lärarnas egna elevlösningar jämfördes sedan mot dessa istället för att föra en diskussion kring varför detta exempel svarade mot den aktuella nivån. Forskarna såg också att lärarna verkade kämpa för att finna en gemensam och konsekvent hållning som stöd för att kunna enas i sina beslut.

Om på förhand bedömda elevarbeten används som stöd vid bedömningen spelar det även roll hur dessa är utvalda. Val Klenowski och Leonore Adie (2009) visar att en del lärare uppfattar ett sådant exempel som en beskrivning av en typisk prestation på den aktuella nivån medan andra lärare uppfattar samma exempel som en minimigräns för en prestation på denna nivå. Att kunna använda det tillgängliga bedömningsmaterialet på ett flexibelt sätt visar sig även vara en faktor som hjälper lärare att hålla fokus på elevprestationerna och att fördjupa diskussionerna (Little et al., 2003). Forskarna har studerat tre projekt där man på olika sätt introducerat verktyg och riktlinjer för att stödja lärare att skapa en miljö för samtal där undervisning och elevers lärande hamnar i fokus. Forskarna menar att i de fall där lärare kunde använda verktygen på ett mer flexibelt sätt och anpassa dem efter egna syften, så uppstod längre och mer intensiva diskussioner (ibid.). Att lärare använder tillgängliga hjälpmedel vid bedömning på olika sätt medför att de gör olika tolkningar av elever prestationer på ett sätt som därmed påverkar likvärdigheten vid bedömningen.

Vilka hjälpmedel lärare sätter värde på verkar däremot variera. Bolt (2011) visar i sin studie att det är de på förhand bedömda elevarbeten som lärare värdesätter mest medan Wyatt-Smith et al. (2010) menar att det är individuellt vilka hjälpmedel som lärarna fäster störst vikt vid. Det finns även en studie som visar att det inte finns några signifikanta skillnader i bedömaröverensstämmelse mellan grupper som haft tillgång till bedömda elevarbeten som hjälpmedel vid bedömning och de som inte haft det (Baird et al., 2004). Wyatt-Smith et al. (2010) framhäver också att det inte är en självklarhet att det tillhandahållna stödmaterialet leder till en samsyn och gemensam förståelse och tolkning av hur materialet ska användas.

Deltagarna

Den tredje och sista gruppen faktorer är kopplad till deltagarna i sambedömningen. I Australien, Queensland, har en forskargrupp studerat lärare (åk 4-9) som betygsatt elevarbeten

utifrån beskrivningar av olika kunskapsnivåer från A-E (Wyatt-Smith et al., 2010). Efter att lärarna bedömt elevprestationer individuellt genomfördes sambedömning där lärarna diskuterade och enades om ett slutgiltigt betyg. Studien visar att lärarna vid sambedömningen använde sig av, och tog stöd i, både skriftliga artefakter, som bedömningsanvisningar, kravnivåer och exempel på bedömda elevarbeten, men även av så kallad ”tyst kunskap” (”tacit knowledge”). Den tysta kunskapen innebar de egna ämneskunskaperna, kunskaper om kursplanen, tidigare erfarenheter av bedömning, kunskap om enskilda elever och uppfattningar om vad ”en genomsnittlig elev” i en given årskurs förväntas kunna samt sociala processer, som dialog och förhandling. Resultat från samma forskningsprojekt visar även på att lärarna i sambedömningsprocessen använde sig av olika sätt att enas om bedömningen som inte alltid var i linje med de anvisningar som getts (Adie, 2012; Klenowski & Adie, 2009). Det som kom fram var till exempel att olika kriterier betygsattes var för sig och att ett medelvärde sedan räknades ut och avgjorde helhetsbetyget, vissa gjorde en helhetsbedömning utifrån kriterierna och i vissa fall sågs något eller några kriterier som överordnade andra. Man såg även exempel på att personliga kriterier användes, kriterier som enskilda lärare värdesatte. Även vetskapen om vilken elev det var som bedömdes kunde vägas in. Forskarna drar slutsatsen att kriterier spelar en roll vid bedömningsprocessen men att formella dokument och riktlinjer bara utgör en begränsad del av de resurser som lärare använder för att värdera och betygsätta elevers prestationer. Även lärare själva uppger att de väger in faktorer i bedömningsprocessen som inte finns omnämnda i bedömningsanvisningar (Adie et al., 2012). Sammantaget innebär detta att personliga uppfattningar och erfarenheter får betydelse för lärarnas bedömning vilket i sin tur påverkar möjligheten till en likvärdig bedömning och betygssättning.

En annan faktor som kan påverka sambedömningsprocessen är om den ursprungliga bedömaren är med i gruppen och deltar i sambedömningen eller inte. I en studie av Adie et al. (2012) visade det sig att när den ursprunglige bedömaren var närvarande och kunde motivera bedömningen utifrån kunskaper om den enskilde eleven och kring situationen i vilken bedömningsuppgiften utfördes så blev bedömningen en annan än när samma elevarbete bedömdes av en grupp lärare där den ursprunglige bedömaren inte var med. Denna faktor och dess eventuella påverkan på sambedömningsprocessen kan även ses som en organisationsfråga och skulle då höra hemma under rubriken ”Struktur och organisation”. Eftersom kunskap om enskilda elever även kan ses som en del av lärares ”tysta kunskap” redovisas det som en faktor under denna rubrik.

Det kan även ha betydelse för sambedömningsprocessen att de lärare som deltar har olika stor erfarenhet, både av att undervisa och av att bedöma elevers prestationer. Lärares olika erfarenhet kan bland annat påverka hur de använder kravnivåer och andra tillgängliga hjälpmedel vid sambedömning. Adie et al. (2012) fann att lärare i början, när de var ovana vid att bedöma utifrån kravnivåer, var mer benägna att använda dessa som ett facit men att man senare i större utsträckning använde dessa som ett stöd och som ett exempel på en elevprestation på en viss nivå, medvetna om att det fanns andra sätt att visa motsvarande kunskap på. Det verkar även som att en ny eller oerfaren lärare har särskilt stort utbyte av sambedömning (Baird et al., 2004; Crisp, 2013) eftersom det kan vara ett sätt att få stöd i att utföra bedömningar och att få ta del av mer erfarna kollegors kunskaper. Erfarna lärare är dessutom mer benägna att använda kriterier på ett flexibelt sätt vid bedömningen medan mindre erfarna i större utsträckning använder dessa som ett facit (Klenowski & Adie, 2009). Det kan däremot uppstå konflikter kopplade till erfarenhet vilket Adie (2012) visar på. En oerfaren lärare som har en avvikande uppfattning kan ha svårt att göra sin röst hörd i förhållande till mer erfarna kollegor och feedback från andra, positiv och negativ kan påverka

hur man deltar och agerar vid fortsatta möten. Det finns studier som lyfter fram att en negativ upplevelse kan resultera i att lärare inte vill fortsätta att delta i sambedömning (Adie, 2012,). Eftersom det framför allt är diskussioner där deltagarna tycker olika, som visat sig produktiva och utvecklande (Adie et al., 2012; Black et al., 2010) kan skillnader i lärarnas erfarenhet därför påverka sambedömningsprocessen negativt. Av denna anledning kan sambedömning mellan lärare från olika skolor vara att föredra (Black et al., 2010). Ett sätt att komma i kontakt med lärare från andra skolor är att sambedömningen sker online, vilket tidigare redogjorts för. I en online-kontext har lärare möjlighet att vara anonyma vilket kan medföra att tidigare nämnda sociala faktorer kan elimineras. Lärare som i vanliga fall tar stort utrymme och utgör en auktoritet tvingas att på samma villkor som andra att motivera och argumentera för sin bedömning. På samma sätt kan andra sociala faktorer som vetskapen om hur ”strängt” eller ”snällt” lärare i vanliga fall bedömer sättas ur spel. Adie (2013) argumenterar därför för att den anonyma online-kontexten tvingar bedömare att omförhandla sin uppfattning om sig själv som bedömare.

Vilket stadium lärarna undervisar på kan även inverka på sambedömningsprocessen. Clarke och Gipps (2000) rapporterar erfarenheter från tre stora projekt i England där man studerat införandet av bedömning utförd av lärare. De såg en skillnad i hur lärare som undervisar i yngre respektive äldre elever tog sig an och utförde sambedömningen. Medan lärarna i för de äldre eleverna, som hade större erfarenhet av att sätta betyg, använde sambedömningen till att jämföra provresultat och betyg så var lärarna för de yngre eleverna mer fokuserade på att analysera elevarbeten och att få en samsyn i hur man uppfattade och värderade olika prestationer.

Resultatdiskussion

I denna studie har syftet varit att förstå på vilket sätt sambedömning kan bidra till en likvärdig bedömning och betygssättning av elevers prestationer genom att söka svar på vilka dokumenterade effekter sambedömning kan ha samt hur denna process kan påverkas av olika faktorer. Utifrån resultaten framgår det att sambedömning har potential att bidra till en ökad likvärdighet främst eftersom det kan leda till att lärare får en samsyn kring tolkning av styrdokument och kring bedömning av elevers prestationer men även eftersom lärare kan få en ökad kompetens beträffande undervisning och bedömning.

En likvärdig bedömning och betygssättning kräver, som tidigare beskrivits, en samstämmighet i flera led: (1) i hur styrdokumenten tolkas och omsätts i undervisning och i bedömningsuppgifter; (2) i hur enskilda elevprestationer tolkas och värderas i förhållande till bedömningsanvisningar och kunskapskrav samt (3) i hur resultatet från flera olika bedömningsunderlag sammanvägs till ett betyg eller till ett kunskapsomdöme (se Fig. 2). Denna modell innebär således att likvärdigheten vid bedömning och betygssättning ökar när ett eller flera tolkningsled stärks. Har då sambedömning potential att ge en ökad samstämmighet i förhållande till dessa tolkningsled och på så vis bidra till en ökad likvärdighet? Svaret på denna fråga kommer här att diskuteras genom att föra ett resonemang kring studiens resultat i förhållande till de tre tolkningsleden och därigenom försöka att dra en mer övergripande slutsats om huruvida sambedömning kan leda till en mer likvärdig bedömning och betygssättning samt vilken påverkan olika faktorer kan antas ha i denna process.

Ett av argumenten för att likvärdigheten i lärares bedömning och betygssättning ökar till följd av sambedömning är att resultaten pekar på att det medför en ökad bedömarkompetens hos de

deltagande lärarna. Sambedömningen ger lärare utrymme och tid för diskussion och reflektion kring frågor kopplade till undervisning och bedömning utifrån konkreta elevprestationer. Vid diskussioner kring kvaliteter i elevers prestationer tvingas lärarna att sätta ord på vad som skiljer de olika prestationerna från varandra, vad som är respektive elevprestations styrkor respektive svagheter och därmed identifiera vad som är nästa steg i elevernas lärande. Detta i sin tur kan innebära att lärarna blir bättre på att finna belägg för elevers kunnande och att de därmed kan använda denna kunskap som utgångspunkt vid planering av undervisning (Bolt, 2011; Falk & Ort, 1998; Malone et al., 2004). Ökad bedömarkompetens i kombination med att sambedömning också kan bidra till att lärare utvecklar insikter som stödjer undervisningen (Bolt, 2011; Clarke & Gipps, 2000; Falk & Ort, 1998) innebär därmed att både det första och det andra tolkningsledet stärks vilket i sin tur bidrar till en ökad likvärdighet. Resultatet av denna forskningsöversikt ger däremot inga tydliga belägg för att lärares bedömningskompetens beträffande sammanvägning av olika prestationer till ett kunskapsomdöme eller betyg, det vill säga tolkningsled tre, ökar.

Ett annat resultat som talar för att sambedömning kan leda till en ökad likvärdighet vid bedömning och betygssättning är att det verkar skapa en samsyn mellan lärare kring vad mål, kriterier och kravnivåer innebär (Connolly et al., 2012; Hall & Harding, 2002; Reid, 2007; Roberts et al., 1997). Att lärare får en gemensam syn kring styrdokumenten beror bland annat på att olika sätt att uppfatta dessa blir synliggjorda och föremål för diskussion utifrån konkreta elevprestationer. Adie et al. (2012) menar att en gemensam förståelse bland lärare för de begrepp som används och vad dessa begrepp står för även är en förutsättning för att kunna diskutera och sambedöma. De konkreta elevprestationerna tvingar lärarna att sätta ord på sina uppfattningar så att dessa blir tydliga och medvetandegjorda, både för den enskilde läraren men även för kollegor vilket kan leda till att lärare omförhandlar sin bedömning (ibid.). Enligt Sadler (2009) är sådana kravnivåer som används i våra styrdokument abstrakta och för att utveckla en förståelse för dessa behövs just konkreta exempel för att illustrera olika kvaliteter vilket blir möjligt att göra vid sambedömning. De gemensamma diskussionerna kring konkreta exempel blir då betydelsefulla också för att utveckla en gemensam förståelse för innebörden av kravnivåerna och för utvecklandet av en känsla för vad som utmärker olika kvaliteter, så kallad tyst kunskap. Harlen (2004a) menar även att hur pass konsekventa lärare är i vid bedömningen av elevprestationer hör ihop med hur man tolkat styrdokumenten och hur dessa omsatts i undervisning. Att låta sambedömningen omfatta även det första tolkningsledet blir då av betydelse för en likvärdig bedömning och betygssättning eftersom en samsyn både påverkar både reliabiliteten och validiteten vid bedömning (Harlen, 2004a; Holroyd, 2000).

Att lärare genom sambedömning får en ökad bedömarkompetens och skapar en samsyn kring styrdokument, undervisning och bedömning ger en grund för en likvärdig bedömning och betygssättning, men det finns å andra sidan saker som talar emot detta. Bland annat har det visat sig vara svårt för lärare att omsätta kunskaper om styrdokumenten i undervisnings- och lärandesituationer trots att de uppmärksammat hur dessa inverkar och stöttar dem i att fokusera på centrala områden att bedöma (Klenowski & Wyatt-Smith, 2010a). Att lärare har svårt att omsätta sådana insikter och kunskaper i undervisning och i bedömningsuppgifter kan innebära att validiteten, och därmed likvärdigheten, vid bedömning och betygssättning påverkas negativt trots att lärarna har en samsyn kring hur styrdokumenten ska tolkas. Ett annat problem kan vara att även om lärare enas om hur en elevprestation ska bedömas eller om hur styrdokumenten ska tolkas innebär det inte med självklarhet att en sådan tolkning är korrekt, vilket kan få till följd att även om lärares bedömningar har hög inter-bedömarreliabilitet så är de ändå inte likvärdiga i bemärkelsen att de är i linje med

kursplanernas mål och kunskapskrav eller med tolkningar som gjorts av andra lärare på andra skolor. Risken är då att vi når en likvärdig bedömning på en ytlig nivå utan att för den sakens skull komma åt de underliggande värderingarna på vilka bedömningen vilar och som gör att erfarenheterna tas med till andra bedömningssituationer. Ett sätt att komma runt detta vid bedömning skulle kunna vara att enbart utgå från till exempel nationella prov vid bedömning och betygssättning, eftersom olika provkonstruktörer här gjort tolkningen av styrdokumenten åt lärarna. Det ger å andra sidan inte lärare den kunskap och kompetens som krävs för att tolka elevprestationer i nästa tolkningsled (2) eller att genomföra undervisning som ger eleverna förutsättningar att lyckas med sådana uppgifter. Genom att låta sambedömningen omfatta även det första tolkningsledet kan man däremot förvänta sig att lärare utvecklar en samsyn kring styrdokumenten som i sin tur kan ge förutsättningar för en mer valid och reliabel bedömning också i andra bedömningssituationer än vid bedömningen av de nationella proven.

Även om sambedömning leder till en ökad samsyn bland lärare betyder detta inte heller per automatik att likvärdigheten vid bedömning och betygssättning ökar. Detta eftersom samsyn och samstämmighet inte är samma sak. Samstämmighet innebär att lärare ger samma prestation samma poäng eller betygsbelägg och är därmed en aspekt av reliabiliteten. Det finns få studier som styrker att sambedömning ger ökad samstämmighet vid bedömning och de studier som återfunnits ger dessutom motsägelsefulla resultat. Å ena sidan finns det studier som redovisar en överensstämmelse på uppemot 80-90 procent (ex. Falk & Ort, 1998) men å andra sidan finns det studier som visar att reliabiliteten inte alls påverkas (Baird et al., 2004) eller att lärare utan erfarenhet av sambedömning gör bedömningar som är jämförbara med dem som har sådan erfarenhet (Connolly et al., 2012).

Sadler (1998) beskriver den delen av bedömningsprocessen där lärare ska tolka och värdera elevers prestationer som ett förlopp i tre steg: (1) uppmärksamhet riktas mot elevprestationen; (2) elevprestationen värderas mot till exempel ett ramverk och (3) poäng, betygsbelägg eller dylikt utdelas. Utifrån en sådan beskrivning av lärares bedömning kan samsyn innebära att lärare är överens i steg 2 om hur en elevprestation ska tolkas och värderas, det vill säga vad som utifrån bedömningskriterierna är elevprestationens styrkor och vad i denna som är mindre bra. En sådan samsyn betyder inte nödvändigtvis att lärare i nästa steg (3) är överens och ger elevprestationen exakt samma poäng eller betygsbelägg.

Även det omvända förhållandet kan råda, det vill säga att lärare i steg 2 värderar olika kriterier på olika sätt i förhållande till varandra men ändå gör en likartad helhetsbedömning (steg 3) vid sammanvägningen av de olika kriterierna (Davison, 2004). Också en svensk studie genomförd av Olofsson (2006) pekar på detta. I studien har fyra lärare bedömt elevarbeten från det nationella provet i matematik, först enskilt och därefter ännu en gång efter att ha sambedömt där emellan. Resultatet visar att lärarna efter sambedömning ändrade sina bedömningar i viss utsträckning så att de blev mer samstämmiga men att lärarna sedan var olika benägna att revidera sina bedömningar, bland annat beroende på vilka kriterier de hade olika uppfattningar kring. Detta kan få till följd att olika elevers prestationer, även om man sambedömer, kan få en likartad bedömning men att denna baseras på olika kriterier. Om detta är att betrakta som godtagbart eller inte beror på i vilken utsträckning man kräver samstämmighet och på vilka grunder. Måste lärares bedömning och betygssättning vara exakt lika eller är det tillräckligt att vara överens om på vilka grunder elevprestationerna bedöms? Ett sätt att få till stånd samstämmighet är förstås att kräva konsensus. En nackdel med detta, förutom att det är väldigt tidskrävande att bedöma samtliga elevprestationer tillsammans, är dock att det finns en risk att olikheter i uppfattningar blir svåra att hantera och att betydelsen

av enskilda lärares status i förhållande till varandra får stor inverkan när målet med diskussionerna är att nå konsensus. Detta kan få till följd att målet med sambedömningen blir att komma överens snarare än att skapa en samsyn vilket bland annat diskuteras både av Davison (2004) och av Moss & Schutz (2001).

> …. we argue that dissensus is an essential natural resource that should be acknowledged and nurtured along-side the search for consensus or agreement. Although it does not ensure a fair and inclusive dialogue, the representation and exploration of dissensus helps to protect us from the false assurance of an articulated consensus that may underrepresent, misrepresent, or exclude groups of voices within the community.
>
> (Moss & Schutz, 2001, s. 65)

Det är därför av betydelse att det skapas goda förutsättningar för ett öppet och tillåtande diskussionsklimat om sambedömningen ska ha möjlighet att leda till samsyn kring styrdokument och bedömning av elevprestationer.

Att olika uppfattningar kommer upp till ytan och diskuteras kan även få till följd att betydelsen av vem som bedömer elevernas prestationer kan minska och därmed påverka både reliabilitet och validitet vid bedömning. Det har visat sig att lärare även när de sambedömer (se till exempel Wyatt-Smith et al., 2010) tar hänsyn till sådana faktorer vid bedömning som inte finns angivna i bedömningskriterier så som personliga uppfattningar om vad som är väsentligt att kunna och vetskapen om vilken elev det är som bedöms och hur denne brukar prestera. Ett sätt att minimera den påverkan som personliga preferenser kan ha vid bedömning kan vara att avidentifiera de elevarbeten som väljs ut för sambedömning. En annan möjlighet är att den ursprunglige bedömaren inte är med och sambedömer de egna elevernas prestationer då det visat sig att när den ursprunglige bedömaren är med och kan motivera sin bedömning blir bedömningen en annan än när denne inte är närvarande (Adie et al., 2012).

Ytterligare ett problem med en likvärdig bedömning är kopplat till det tredje och sista tolkningsledet när elevers olika prestationer ska sammanvägas till ett kunskapsomdöme eller ett betyg utifrån kunskapskraven i kurs- eller ämnesplanerna. För en likvärdig bedömning i detta led krävs att kunskapskraven tolkas på samma sätt av olika lärare och att sammanvägningarna till omdöme och betyg sker enligt läroplanernas föreskrifter. I en studie av Allal (2013) uppgav lärare att de ofta samarbetade i det första tolkningsledet, med att planera undervisning och att konstruera bedömningsuppgifter medan det var mindre vanligt förekommande med samarbete när elevernas prestationer skulle tolkas och värderas i de nästkommande leden. Konsekvensen kan bli att även om lärare är överens om vad som ska bedömas och att de enskilda bedömningarna är både reliabla och valida så blir sammanvägningen av olika resultat en process som inte hanteras på ett likvärdigt sätt. Allal (2013) visar till exempel att sammanvägningen av elevers resultat till ett betyg sker på ett sätt där olika informationskällor tas i beaktande och ges olika tyngd för olika elever. Lärarna i studien uppgav att deras huvudsakliga information vid betygssättningen var elevers resultat på skriftliga prov men när det gällde elever på gränsen mellan två betyg togs större hänsyn även till andra informationskällor. Även till exempel Korp (2006) visar utifrån svenska förhållanden att lärare vid betygssättning använder olika metoder för att sammanväga olika elevprestationer till ett betyg vilket talar för att även betygssättningen och formerna för denna behöver diskuteras och göras transparent för att den ska bli likvärdig. Allal (2013) argumenterar dessutom för att sambedömning i det tredje ledet ställer högre krav på transparens i hur den enskilde läraren går till väga jämfört med de två andra leden för att leda till ökad likvärdighet.

Utifrån resultaten i föreliggande studie kan man se att sambedömning kan leda till en samsyn kring tolkningen av styrdokumenten vilket talar för att förutsättningarna för en likvärdig betygssättning ökar om även det tredje tolkningsledet omfattas av sambedömningen. Däremot finns inga belägg för att sambedömning leder till att lärare använder modeller för sammanvägningen till kunskapsomdöme eller betyg som är i linje med styrdokumentens intentioner. Att sambedömning förmår att skapa förutsättningar för likvärdighet även kring denna process kan inte uteslutas men, utifrån denna studies urval, saknas mer explicit forskning i detta avseende. Det finns däremot andra erfarenheter som talar för att likvärdigheten i det tredje tolkningsledet kan stödjas genom sambedömning, exempelvis erfarenheter ifrån Queensland, Australien. Här har man inga nationella prov utan istället ett väl utbyggt system för sambedömning som sker i flera steg, både på lokal, regional och på nationell nivå och som leds av personer och grupper av personer som blivit utsedda och utbildade för ändamålet av Queensland Studies Authority (QSA) (Maxwell, 2007). Sambedömningen sker utifrån ett urval av elevportföljer som innehåller ett brett urval av uppgifter insamlade som en följd av den ordinarie undervisningen. I en rapport från QSA där man granskat ett slumpmässigt urval av 2649 elevportföljer, fann man en överensstämmelse mellan skolornas och granskningsgruppens bedömning i 92 procent av fallen (QSA, 2012) vilket tyder på att sambedömning kan stödja också betygssättningen, även om det som i detta fall kräver en omfattande organisation. För att kunna tala om en likvärdighet vid bedömning och betygssättning behöver därför sambedömningen omfatta samtliga tre tolkningsled. Att sambedömningen omfattar samtliga led ger förutsättningar för en linje, alignment, i hela bedömningsprocessen, från tolkningen av styrdokumenten via undervisning och uppgiftskonstruktion till bedömning och betygssättning. Ett sådant synsätt att skapa alignment genom att låta sambedömning omfatta samtliga tre tolkningsled, förs även fram i annan forskning (se till exempel. Smith, 2012) som ett argument för att få valida och reliabla bedömningar.

Att sambedömning på det sätt som diskuterats här kan leda till en mer utvecklad bedömarkompetens och till ökad samsyn och därmed i förlängningen också till en ökad likvärdighet vid bedömning och betygssättning borde dock rimligtvis endast gälla på lokal nivå. Att lärare på enskilda skolor eller på närliggande skolor får en ökad samsyn i att tolka styrdokument och att tolka och värdera elevers prestationer är inget som får effekter för likvärdigheten på nationell nivå. Det finns studier där lärare träffats över skolgränser och ibland också över internet, vilket kan skapa förutsättningar för nationell likvärdighet. Ett utvecklat exempel på detta är återigen erfarenheterna ifrån Queensland, Australien. I ett sådant system där man på olika nivåer har en organisation för att genomföra sambedömning finns förutsättningar en likvärdig bedömning och betygssättning på nationell nivå. Detta kräver dock en struktur som i dagsläget inte finns i Sverige. Betydelsen av att diskutera med lärare från andra skolor är dock något som lyfts fram i flera studier som ett sätt att få andra perspektiv på bedömningen.

Även om lärares egna uppfattningar i stor utsträckning handlar om värdet av diskussioner inom den egna skolan eller mellan närliggande skolor krävs att dessa erfarenheter diskuteras i ett vidare sammanhang om målet ska vara en likvärdig bedömning och betygssättning (Clarke & Gipps, 2000). Ett sätt att komma i kontakt med lärare bortom den egna skolan är att sambedöma i en online-kontext, vilket har studerats av Adie (2010, 2012, 2013). Som tidigare nämnts visar den forskning som rapporterats att detta i stort ger liknande effekter som när deltagarna träffas fysiskt, även om vissa faktorer kopplade till denna kontext kan få större eller mindre betydelse. Ett exempel är betydelsen av vem som bedömer. Skillnader i erfarenhet och auktoritet mellan lärare kan medföra att viktiga argument inte tas hänsyn till

medan andra får för stort utrymme. Lärare som har svårt att göra sin röst hörd i förhållande till kollegor riskerar att inte bidra med sina perspektiv eller att gå med på en bedömning de egentligen inte håller med om. I en online-kontext finns möjligheten att vara anonym, vilket kan bidra till att samtliga deltagare deltar på liknande villkor vilket kan kompensera för den typen av effekter (Adie, 2013). Det finns däremot resultat som tyder på att lärare kan ha svårt att ta med sig de erfarenheter som gjorts i grupper som är konstruerade för detta särskilda ändamål (Black et al., 2010; Bolt, 2011.) och att grupper som är redan är etablerade fungerar bättre (Bolt, 2011). Att organisera för sambedömning i en on-linekontext kräver dessutom en hel del i form av tillgång till teknik och samordning, vilket även är centrala faktorer när sambedömningen sker genom fysiska träffar.

De resultat som hittills diskuterats kan antas ge förutsättningar för att sambedömning leder till att likvärdigheten vid bedömning och betygssättning ökar på lång sikt. Flera av de genomförda studierna har däremot genomförts när sambedömning introducerats som en del i ett större projekt eller när deltagarna av andra orsaker varit nybörjare på området. I en studie genomförd av Saunders och Davis (1998) till exempel fick de deltagande lärarna vid två tillfällen bedöma en avhandling utifrån gemensamma kriterier. Det visade sig då att lärarna vid det andra tillfället när de redan hade skaffat sig erfarenheter av bedömningskriterierna och de hade en gemensam syn på vad dessa stod för kände en större säkerhet vid användandet av dem. Tidigare erfarenhet av det ämnesområde som ska bedömas verkar även medföra att en sådan samsyn kan nås snabbare (Black et al., 2010). Dessa resultat ger stöd åt ett antagande att sambedömning kan ge mer långsiktiga och effektiva resultat när den genomförs under en längre tid och inte på kort sikt som ett enstaka inslag vid något eller några tillfällen. Sådana mer långsiktiga effekter av sambedömning är därför något som ytterligare behöver belysas genom forskning.

Sammanfattningsvis kan sägas att sambedömning kan leda till en mer likvärdig bedömning och betygssättning eftersom lärare utvecklar sin bedömarkompetens och skapar en samsyn kring de delar av bedömningsprocessen som ingår i sambedömningen. Det finns också andra faktorer som kan påverka sambedömningsprocessen och därmed indirekt bedömningens och betygssättningens likvärdighet. En sådan faktor är den variation och bredd i elevprestationer som väljs ut som underlag för sambedömning. Att elevarbetena som diskuteras representerar en bredd av prestationer och elever med olika bakgrund är något som lyfts fram som centralt för att lärare ska kunna skapa sig en nyanserad uppfattning om olika kvaliteter i elevernas prestationer och för att kunna ta med sig sina erfarenheter i nya situationer (Falk & Ort, 1998). En annan sådan faktor är hur diskussionerna i gruppen förs. Adie (2012) menar att oerfarna lärare kan ha svårt att göra sin röst hörd gentemot mer erfarna kollegor vilket kan medföra att alla argument och infallsvinklar som skulle kunna tillföra ytterligare perspektiv inte kommer fram. Sådana faktorer blir mindre tydliga i till exempel en anonym on-line kontext (Adie, 2013). Ytterligare en faktor som kan påverka är om elevarbetena redan är bedömda när sambedömningen sker. Att någon redan bestämt sig för en bedömning, kan fungera hämmande för diskussionen (Adie et al., 2012; Klenowski & Adie, 2009). Garry, Mc Cool och O'Neill (2005) visade att om lärare bedömde elevarbeten oberoende av varandra blev skillnaderna i deras bedömningar större än när de bedömde elevarbeten som redan var bedömda, och där lärarna fick tillgång till den ursprungliga bedömningen. Resultatet bekräftade författarnas antagande som innebar att om det finns en bedömning att utgå ifrån, ett ankare, så görs den andra bedömningen med den första som utgångspunkt medan en förutsättningslös bedömning inte har samma referens att utgå ifrån. Vid sambedömning skulle det innebära att om elevarbetena redan är bedömda, av till exempel undervisande lärare, när sambedömningen sker så sker diskussionerna med utgångspunkt ifrån den redan gjorda

bedömningen och att denna påverkar och begränsar de diskussioner som förs. Om elevarbetena däremot inte är bedömda på förhand finns bättre förutsättningar för en förutsättningslös diskussion där olika aspekter och argument kan föras fram. Nackdelen med att elevarbetena inte är bedömda på förhand är att möjligheten att göra ett medvetet urval av elevarbeten försvinner. Också de hjälpmedel som lärare har som stöd vid bedömningen så som bedömningsanvisningar, kunskapskrav, redan bedömda elevprestationer och liknande, kan påverka bedömningsprocessen, dels eftersom tydliga bedömningsanvisningar ger ökad reliabilitet och dels eftersom flexibel användning av sådant bedömningsstöd ger större förutsättningar för fördjupade diskussioner (Little et al., 2003). Slutligen är det av vikt att beakta behovet av ledning, struktur och organisation som förutsättning för genomförandet av sambedömning.

Adie (2012) menar att lärande till följd av sambedömning inte bör ses som något som uppstår av sig självt utan det är beroende av flera faktorer och kan motverkas av att deltagarna får negativa upplevelser av sambedömningen. Det finns studier som på olika sätt betonar att ett positivt utfall av sambedömningsprocessen är kopplat till en god organisation och ett tydligt ledarskap. Little et al. (2003) betonar att ett starkt ledarskap är särskilt centralt för att hålla fokus på elevarbetena och en förutsättning för en öppen, kritisk och konstruktiv diskussion. Roberts et al. (1997) visar även att ett starkt ledarskap kan kopplas till att sambedömningsgrupperna utvecklar en stark kollegialitet. Även om kollegialitet i sig inte är en effekt som direkt påverkar bedömningens och betygssättningens likvärdighet kan det ses som en indirekt påverkansfaktor eftersom stark kollegialitet är en faktor som kan ses som avgörande för utfallet av sambedömningen (Little et al., 2003).

Metoddiskussion

Syftet med denna studie var att försöka förstå om och i så fall på vilket sätt sambedömning kan bidra till en ökad likvärdighet vid bedömning och betygssättning genom att söka svar på studiens två övergripande frågor. Detta krävde en genomgång av befintlig forskning. Ett initialt problem var att hitta lämpliga sökord eftersom det inte finns ett etablerat vetenskapligt begrepp för sambedömning. Valet av sökord och sökrutiner kan därför ha medfört att underlaget för studien är ofullständigt.

Sökningarna resulterade vidare i ett stort antal träffar och urvalet av studier utifrån dessa sökningar har genomförts av endast en person vilket gör att rutinerna i samband med sökning och urval endast i viss utsträckning kan anses kvalitetssäkrade. Det som ändå talar för studiens tillförlitlighet är att urvalet av de ingående studierna genomförts systematiskt. Urvalet är resultatet av ett metodiskt arbete där relevanta studier valts ut och inga studier är inkluderade på basis av rekommendationer eller andra mer slumpartade urvalskriterier. Urvalet har dessutom diskuterats och utförts i samråd med forskare med stor erfarenhet av liknande studier. Det är däremot sannolikt att det finns studier som missats av olika anledningar och en del studier som var menade att ingå har inte gått att få tag i. Det är även troligt att det finns genomförda studier av relevans som inte är rapporterade på ett sätt som gör att de ger träffar vid de utförda databassökningarna.

En annan begränsning i metoden är att de vetenskapliga studier som har bildat underlag för resultatet innehåller resultat och slutsatser som i sin tur tagits fram i annat syfte än det som föreligger i denna studie och som redan bearbetats och tolkats i förhållande till författarnas egna syften. Då syftet med de olika vetenskapliga studierna som utgör underlag för denna studie har varit ett annat än syftet med denna studie, innebär det att de resultat som förts fram

i de enskilda studierna har fått tolkas utifrån de frågor som ställts här. Denna process medför därmed en viss osäkerhet.

Skapandet av kategorier och/eller teman är en viktig del i en kvalitativ innehållsanalys (Graneheim & Lundman, 2004). I denna studie inleddes innehållsanalysen genom att de studier som föll inom ramen för urvalskriterierna lästes igenom ett flertal gånger för att få en känsla för helheten. Därefter valdes meningsbärande enheter, *meaning units,* ut och fördes in i en tabell. Meningsbärande enheter är enligt Graneheim och Lundman. (2004) ord, meningar eller delar av texter som innehåller aspekter vilka relaterar till varandra genom innehåll och kontext. Att välja ut sådana enheter är en kritisk process (ibid.) och i denna studie utfördes denna process i två steg. Inledningsvis markerades omfattande delar av texterna och först efter ytterligare genomläsningar valdes enheterna ut och fördes in i en tabell. Kategorier skapades sedan genom att de meningsbärande enheterna kodades och grupperades i gemensamma teman. Krippendorff (1980) menar att de kategorier som skapas måste vara både uttömmande och ömsesidigt uteslutande vilket innebär att inga data ska uteslutas på grund av att de inte passar in i någon kategori. Det innebär även att inga data ska falla mellan kategorier eller passa in i mer än en kategori. I föreliggande studie skapades i ett första steg kategorier som sedan diskuterades i flera omgångar med en forskare och reviderades därefter. De kategorier som skapades uppfyller i stort sett de kriterier som Krippendorf (1980) menar behöver uppfyllas för att skapa tillförlitlighet till forskningsstudier. I något fall innebär kategoriseringen att data kan passa in i mer än en kategori. Graneheim och Lundman (2004) menar dock att det inte alltid är möjligt att skapa sådana ömsesidigt uteslutande kategorier när de texter studien baseras på innefattar människors erfarenheter, vilket är fallet i denna studie.

Det bör vidare betonas att all forskning som ligger till grund för resultatet genomförts i andra länder än Sverige i en kontext som i flera avseende är annorlunda. Då den svenska kontexten ur ett internationellt perspektiv är förhållandevis unik med lärare som oftast både undervisar, bedömer och betygsätter eleverna, går det inte att med säkerhet att avgöra om dessa resultat är relevanta även ur ett svenskt perspektiv. Däremot har många av de ingående vetenskapliga studierna utförts i länder med utbildningssystem som i flera avseenden liknar det svenska eller genom projekt som syftar till att prova sådana system. Till exempel så sker den mesta bedömningen utifrån kriterier och i samtliga studier är det lärare och inte särskilt utvalda externa bedömare som diskuterar eller bedömer elevprestationer. Att elevers betyg i många länder avgörs på grundval av resultatet på prov som bedöms av externa bedömare kan även vara en bidragande förklaring till det begränsade antalet studier som särskilt studerat hur sambedömning påverkar bedömarreliabiliteten och i vilken utsträckning till exempel lärare ger samma poäng eller betygsbelägg för samma elevprestationer.

Slutsatser

Kan sambedömning bidra till en likvärdig bedömning? Svaret på denna fråga har visat sig vara komplext eftersom befintlig forskning inte ger ett entydigt svar och eftersom utfallet av sambedömningsprocessen kan påverkas av ett flertal olika faktorer. Utifrån diskussionen som förts kan man dra slutsatsen att sambedömning har potential att stödja en likvärdig bedömning och betygssättning eftersom det finns belägg för att sambedömning kan bidra till en samsyn och till en ökad bedömarkompetens bland lärare. En förutsättning för en sådan slutsats är däremot att man inte bör förvänta sig att en samsyn är det samma som att lärare ger samma poäng eller betygsbelägg för samma elevprestation utan att man nöjer sig med att de grunder utifrån vilka sådana ges ges är likartad. Ytterligare en förutsättning för att sambedömning ska kunna leda till en ökad likvärdighet vid bedömning och betygssättning är att bedömning ses

som en process som omfattar samtliga tolkningsled och att samtliga dessa inkluderas i sambedömningsprocessen. De tre tolkningsleden består av (1) tolkning av styrdokument för undervisning och konstruktion av bedömningsuppgifter (2) tolkning och värdering av elevers prestationer i förhållande till kravnivåer och (3) sammanvägning av flera olika elevprestationer till ett betyg. Ju fler av dessa tolkningsled som sambedömningen omfattar desto större är förutsättningarna att likvärdigheten vid bedömning och betygssättningen ökar.

En ökad likvärdighet till följd av sambedömning tycks vidare vara en lokal process som sker på lång sikt och med behov av en tydlig struktur och ledning. Att likvärdigheten ökar på lokal nivå är däremot inte en garanti för att det samma sker på regional eller nationell nivå om inte dessa nivåer på något sätt kopplas samman. Detta kräver en välutbyggd struktur och omfattande satsningar. Erfarenheter från Queensland, Australien, har visat att detta är möjligt och dessutom med hög likvärdighet som följd. Ett sätt att åstadkomma denna sammankoppling, som man även har erfarenheter av i Queensland, är genom att utnyttja de möjligheter som den moderna tekniken har att erbjuda.

Andra enskilda faktorer som är av särskild att betydelse för att sambedömning ska leda till en ökad likvärdighet är svårare att identifiera. Däremot tycks variationen vara central för att sambedömningen ska vara framgångsrik och därmed blir den indirekt en faktor som påverkar likvärdigheten genom: (1) att olika åsikter och perspektiv beaktas, snarare än sökande efter konsensus, för att skapa förutsättningar för fördjupade diskussioner; (2) att urvalet av elevprestationer för bedömning speglar en bredd och variation av prestationer samt (3) att lärare kommer ifrån olika kontexter och ges möjlighet till sambedömning i olika konstellationer.

Vid genomgång av den befintliga forskningen om sambedömning och dess effekter blev det uppenbart att en betydande del av denna forskning bygger på intervjuer med lärare eller på observationer av lärare när de sambedömer. Det medför att forskningen i stor utsträckning baseras på lärares *uppfattningar* om sambedömning och dess effekter. Till exempel kan lärare uppfatta att deras bedömningar blir mer samstämmiga utan att ha belägg för att det verkligen förhåller sig på detta sätt. Av denna anledning finns det ett stort behov av studier som på ett systematiskt sätt undersöker om sambedömning verkligen leder till ökad reliabilitet och validitet i lärares bedömning. Det finns även en avsaknad av forskning som studerat sambedömning i det tredje tolkningsledet, när elevers prestationer ska sammanvägas till ett kunskapsomdöme eller till ett betyg. Allal (2013) lyfter detta tolkningsled som särskilt viktigt för att stödja en likvärdig bedömning och betygssättning då forskning pekar på att detta är en process som hanteras på olika sätt av olika lärare, vilket även stöds av svensk forskning (Korp, 2006). Det finns även en avsaknad av studier som undersökt vilken påverkan sambedömning kan ha inte enbart på bedömningens likvärdighet utan även på elevers lärande. Denna möjliga påverkan blir av intresse eftersom det visat sig att lärare har svårt att omsätta mål och kriterier i undervisnings- och lärandesituationer trots att de uppmärksammat hur dessa inverkade och stöttade dem i att fokusera på centrala områden att bedöma (Klenowski & Wyatt-Smith, 2010a).

Referenser

Adie, L. E. (2010). *Developing shared understandings of standards-based assessment: On-line moderation practices across geographically diverse contexts*. Doktorsavhandling, Queensland University of Technology, Australien.

Adie, L. (2012). Learning as identity and practice through involvement in online moderation. *Educational Assessment, Evaluation and Accountability, 24*, 43-56.

Adie, L. (2013). The development of teacher assessment identity through participation in online moderation. *Assessment in Education: Principles, Policy & Practice, 20*, 91-106.

Adie, L. E. Klenowski, V. & Wyatt-Smith, C. (2012). Towards an understanding of teacher judgement in the context of social moderation. *Educational Review, 64*, 223-240.

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice, 20*, 20-34.

Baird, J-A., Greatorex, J. & Bell, J. F. (2004). What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice, 11*, 331-348.

Biggs, J.B. & Tang, C.S. (2007). *Teaching for quality learning at university: what the student does*. (3. ed.) Maidenhead: McGraw-Hill/Society for Research into Higher Education.

Black, P. (1998). *Testing: Friend or foe? Theory and practice of assessmentand testing*. London: RoutledgeFalmer.

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice, 17*, 215-232.

Black, P., Harrison, C., Hodgen, J., Marshall, B. & Serret, N. (2011). Can teachers' summative assessments produce dependable results and also enhance classroom learning? *Assessment in Education: Principles, Policy & Practice, 18*, 451-469.

Boesen, J. (2006). *Assessing mathematical creativity: comparing national and teacher-made tests, explaining differences and examining impact*. Diss. Umeå: Umeå universitet, 2006. Umeå.

Bolt, S. (2011). Making consistent judgments: Assessing student attainment of systemic achievement targets. *Educational Forum, 75*,157-172.

Clarke, S. & Gipps, C. (2000). The role of teachers in teacher assessment in England 1996–1998. *Evaluation and Research in Education, 14*, 38-52.

Connolly, S., Klenowski, V. & Wyatt-Smith, C. M. (2012). Moderation and consistency of teacher judgement: teachers' views. *British Educational Research Journal, 38*, 593-614.

Crisp, V. (2013). Criteria, comparison and past experiences: how do teachers make judgements when marking coursework? *Assessment in Education: Principles, Policy & Practice, 20*, 127-144.

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research, 58*, 438–481.

Cunningham, G. (1998). *Assessment in the Classroom, Constructing and Interpreting Tests.* London och Washington D.C.: The Falmer Press.

Davison, C. (2004). The contradictory culture of teacher-based assessment: ESL teacher assessment practices in Australian and Hong Kong secondary schools. *Language Testing, 21*(3), 305-334.

de Eca, P. & Torres, M. T. (2005). Using portfolios for external assessment: An experiment in Portugal. *International Journal of Art & Design Education, 24*, 209-218.

Dunbar, S. B., Koretz, D. M. & Hoover, H. D. (1991). Quality Control in the Development and Use of Performance Assessments. *Applied Measurement in Education, 4*(4), 289-303.

Ecclestone, K. (2001). 'I know a 2:1 when I see it': Understanding criteria for degree classifications in franchised university programmes. *Journal of Further and Higher Education 25*, 301–13.

Erickson, G. (2009). *Nationella prov i engelska – en studie av bedömarsamstämmighet.* Hämtad 10 april 2013, från http://www.nafs.gu.se/publikationer/

Englund, T. & Quennerstedt, A. (2008). *Vadå likvärdighet? Studier i utbildningspolitisk språkanvändning*. Göteborg: Daidalos.

Falk, B. (1998). Testing the Way Children Learn. Principles for Valid Literacy Assessments. *Language Arts, 76*, 57-66

Falk, B. & Ort, S. (1998). Sitting down to score. Teacher learning through assessment. *Phi Delta Kappan, 80*, 59-64.

Forsberg, C. & Wengström, Y. (2008). *Att göra systematiska litteraturstudier: värdering, analys och presentation av omvårdnadsforskning*. (2., [uppdaterade] utg.) Stockholm: Natur & Kultur.

Garry, J., McCool, M. A. & O'Neill, S. (2005). Are Moderators Moderate?: Testing the 'Anchoring and Adjustment' Hypothesis in the Context of Marking Politics Exams1. *Politics, 25*(3), 191-200.

Gipps, C. V. (1994). *Beyond testing: towards a theory of educational assessment*. London: Falmer.

Graneheim, U. H. & Lundman, B. (2004). Qualitative content analysis in nursing research: concepts, procedures and measures to achieve trustworthiness. *Nurse Education Today, 24*(2), 105-112.

Gustavsson, A., Måhl, P. & Sundblad, B. (2012). *Betygssättning: en handbok*. (1. uppl.) Stockholm: Liber.

Haertel, E.H. & Herman, J.L. (2005). A historical perspective on validity arguments for accountability testing. In J. L. Herman & E. H. Haertel (Ed.), *Uses and misuses of data for educational accountability and improvement.* Chicago: National Society for the Study of Education, 1-34.

Hall, K. & Harding, A. (2002). Level descriptions and teacher assessment in England: towards a community of assessment practice. *Educational Research, 44*(1), 1-16.

Hand, L. & Clewes, D. (2000). Marking the difference: An investigation of the criteria used for assessing undergraduate dissertations in a business school. *Assessment and Evaluation in Higher Education, 25*, 5–21.

Harlen, W. (Ed.) (1994) *Enhancing Quality in Assessment*. BERA Policy Task Group on Assessment, Paul Chapman Publishers.

Harlen, W. (2004a). A systematic review of the evidence of reliability and validity of assessment by teachers used for summative purposes. In: *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2004b) A systematic review of the evidence of the impact on students, teachers and the curriculum of the process of using assessment by teachers for summative purposes. *Research Evidence in Education Library*. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Harlen, W. (2005). Teachers' summative practices and assessment for learning - tensions and synergies. *Curriculum Journal, 16*(2), 207-223.

Harlen, W. & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students' motivation for learning. In: *Research Evidence in*

Education Library. London: EPPI-Centre, Social Science Research Unit, Institute of Education.
Holroyd, C. (2000). Are assessors professional? *Active Learning in Higher Education 1* (1), 28–44.
Jönsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130–144.
Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice, 18*, 5-17.
Klenowski, V. & Adie, L. (2009). Moderation as judgement practice: Reconciling system level accountability and local level practice. *Curriculum Perspectives, 29*, 10-28.
Klenowski, V. & Wyatt-Smith, C. (2010a). Standards-Driven Reform Years 1-10: Moderation an Optional Extra? *Australian Educational Researcher, 37*(2), 21-39.
Klenowski, V. & Wyatt-Smith, C. (2010b). Standards, teacher judgement and moderation in the contexts of national curriculum and assessment reform. *Assessment Matters, 2*, 107–131.
Klenowski, V. (2012). Raising the stakes: the challenges for teacher assessment. *The Australian Educational Researcher, 39*(2), 173-192.
Korp, H. (2006). *Lika chanser i gymnasiet? En studie om betyg, nationella prov och social reproduktion*. Doktorsavhandling, Malmö högskola.
Krippendorff, K. (1980). *Content Analysis. An Introduction to its Methodology*. The Sage Commtext Series, Sage Publications Ltd., London.
Limbrick, L. & Knight, N. (2005). Close reading of students' writing: What teachers learn about writing. *English Teaching: Practice and Critique, 4*(2).
Linn, R. L. (1993). Linking Results of Distinct Assessments. *Applied Measurement in Education, 6*(1), 83-102.
Little, J. W., Gearhart, M., Curry, M. & Kafka, J. (2003). Looking at student work for teacher learning, teacher community, and school reform. *Phi Delta Kappan, 85*, 185-192.
Malone, L., De Lucchi, L., & Long, K. (2004). All Things in Moderation. *Science and Children, 41*(5), 30-34.
Maxwell, G. S. (2002). Progressive assessment: Synthesising formative and summative purposes of assessment. In F. Ventura and G. Grima (eds), *Contemporary issues in educational assessment: Proceedings of the ACEAB Second International Conference, Malta 2002*, pp. 220–234. Msida, Malta: MATSEC Examinations Board.
Maxwell, G. (2007). *Implications for moderation of proposed changes to senior secondary school syllabuses.* Paper commissioned by the Queensland Studies Authority. Brisbane: Queensland Studies Authority.
Messick, S. (1989). Validity. I R. L. Linn (Ed.), *Educational Measurement* (Vol. 3, sid. 13-103). New York: American Council on Education.
Morgan, C. & Watson, A. (2002). The interpretative nature of teachers' assessments of students' mathematics: Issues for equity. *Journal for Research in Mathematics Education, 33*, 78-110.
Moss, P. A. & Schutz, A. (2001). Educational standards, assessment, and the search for consensus. *American Educational Research Journal, 38*, 37-70.
Murphy, R. J. L. (1982). A further report of investigations into the reliability of marking of GCE examinations. *British Journal of Educational Psychology, 52*(1), 58-63.
Olofsson, G. (2006). *En studie av lärares bedömning av elevarbeten på ett nationellt prov i matematik kurs A.* Stockholms universitet, PRIM-gruppen.
Queensland Studies Authority. (2012). *Random sampling project 2012. Report on random sampling of assessment in Authority subjects*. Brisbane: Queensland Studies Authority.

Reid, L. (2007). Teachers talking about writing assessment: valuable professional learning. *Improving Schools, 10*, 132-149.
Roberts, L., Wilson, M. & Draney, K. (1997). The SEPUP assessment system: An overview *BEAR Report Series* (Vol. SA-97-1). Berkeley: University of California.
Rezaei, A. R. & Lovorn, M. (2010). Reliability and validity of rubrics for assessment through writing. *Assessing Writing, 15*(1), 18-39.
Sadler, D. R. (1998). Formative Assessment: revisiting the territory. *Assessment in Education: Principles, Policy & Practice, 5*(1), 77-84.
Sadler, D. R. (2009). Grade integrity and the representation of academic achievement. *Studies in Higher Education, 34*(7), 807-82.
Sadler, D. R. (2013). Assuring academic achievement standards: from moderation to calibration, *Assessment in Education: Principles, Policy & Practice, 20* (1), 5-19.
Saunders, M. N. K. & Davis, S. M. (1998). The use of assessment criteria to ensure consistency of marking: some implications for good practice. *Quality Assurance in Education, 6*(3), 162-171.
Skolinspektionen (2012). *Lika för alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år.* Stockholm: Skolinspektionen.
Skolinspektionen (2013). *Tillsyn av bedömning och betygssättning.* Stockholm: Skolinspektionen.
Skolverket (2004a). *Handlingsplan för en rättssäker och likvärdig betygssättning*. Stockholm: Skolverket.
Skolverket (2004b). *Det nationella provsystemet i den målstyrda skolan: omfattning, använding och dilemman*. Stockholm: Skolverket.
Skolverket (2009a). *Redovisning av regeringsuppdrag att ge förslag på hur det nationella provsystemet bör utvecklas och utformas*. Stockholm: Skolverket.
Skolverket (2009b). *Bedömaröverensstämmelse vid bedömning av nationella prov.* Dnr 2008:286. Stockholm: Skolverket.
Skolverket (2011a). *Läroplan för grundskolan, förskoleklassen och fritidshemmet 2011*. Stockholm: Skolverket.
Skolverket (2011b). Läroplan, examensmål och gymnasiegemensamma ämnen för gymnasieskola 2011. (2011). Stockholm: Skolverket.
Skolverket (2012a). *Betygsskalan A-E.* Hämtad 4 april 2013, från http://www.skolverket.se/kursplaner-och-betyg/betyg
Skolverket (2012b). *Likvärdig utbildning i svensk grundskola? En kvantitativ analys av likvärdighet över tid. Rapport nr. 374*. Stockholm: Skolverket.
Skolverket (2012c). *Om nationella prov.* Hämtad 3 april 2013, från http://www.skolverket.se/prov-och-bedomning/nationella-prov/nat-prov-1.111291
Skolverket (2012d). *Redovisning av uppdrag om avvikelser mellan provresultat och betyg i grundskolan årskurs 9*. Dnr 75-2012:311. Stockholm: Skolverket.
Skolverket (2012e*). Redovisning av uppdrag om avvikelser mellan provresultat och kursbetyg i gymnasieskolan. Dnr 75-2012:311. Skolverket: Stockholm.*
Smith, C. (2012). Why should we bother with assessment moderation? *Nurse Education Today, 32*(6), e45-e48.
Sverige. (2010). *Skollagen (2010:800): med Lagen om införande av skollagen (2010:801).* Stockholm: Norstedts juridik.
Syverson, M. A. (2009). Social justice and evidence-based assessment with the Learning Record. *Forum on Public Policy*, 1-27.
U2011/6543/S. Stockholm: Utbildningsdepartementet.
Wilson, M. (Ed.). (2004). *Towards coherence between classroom assessment and accountability*. Chicago: University of Chicago Press.

Wood, R. (1991). *Assessment and testing: a survey of research.* Cambridge, Cambridge University Press).

Wyatt-Smith, C., Klenowski, V. & Gunn, S. (2010). The centrality of teachers' judgement practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice, 17*, 59-75.

Yorke, M., Bridges, P. & Woolf, H. (2000). Mark distributions and marking practices in UK higher education. *Active Learning in Higher Education 1*, 7–27.