Vetenskapsrådet

Delrapport från SKOLFORSK-projektet

# BETYGENS GEOGRAFI

## Forskning om betyg och summativa bedömningar i Sverige och internationellt

VETENSKAPSRÅDETS RAPPORTER 2015

BETYGENS GEOGRAFI – FORSKNING OM BETYG OCH SUMMATIVA BEDÖMNINGAR I SVERIGE OCH INTERNATIONELLT

VETENSKAPSRÅDET
Box 1035
SE-101 38 Stockholm, SWEDEN


ISBN 978-91-7307-284-7

Vetenskapsrådet genomförde under 2014 ett projekt, SKOLFORSK, för att kartlägga befintlig utbildningsvetenskaplig forskning. Arbetet skedde på uppdrag av regeringen för att resultera i kartläggningar av svenska och internationella forskningsresultat med relevans för skolväsendet. Syftet var att skapa en plattform av kunskapsunderlag till det nybildade Skolforskningsinstitutet. Slutsatserna i denna delrapport är författarnas egna. Vetenskapsrådets sammanfattande rapport, Forskning och skola i samverkan, med en beskrivning av projektet och med de frågeställningar, resultat och rekommendationer som redovisats inom delprojekten kan liksom de övriga delrapporterna laddas ner från Vetenskapsrådets webbplats.

# BETYGENS GEOGRAFI

Christian Lundahl, Magnus Hultén, Alli Klapp, Larissa Mickwitz

# FÖRORD

Regeringen gav 2013-11-21 (U2013/6845/S) Vetenskapsrådet i uppdrag att svara för genomförandet av validerade kartläggningar av svenska och internationella forskningsresultat med relevans för skolväsendet. Kartläggningarna skulle utgå ifrån frågeställningar som är relevanta för, och framtagna i samråd med, verksamma i skolan och förskolan. Syftet med kartläggningarna var att utgöra underlag för systematiska sammanställningar av forskningsresultat med relevans för verksamhet inom skola och förskola som Skolforskningsinstitutet skulle få i uppdrag att genomföra. Uppdraget formulerades efter att huvudsekreteraren för Utbildningsvetenskapliga kommittén (UVK) vid Vetenskapsrådet utformat ett förslag till ett antal projekt som under ett år skulle arbeta fram ett underlag till Skolforskningsinstitutet.

Uppdraget från regeringen, med arbetsnamnet SKOLFORSK, har trots den korta tid som stått till buds, resulterat i sexton delprojekt där ett 40-tal forskare från femton olika universitet i Sverige, Norge och USA har medverkat. En välmeriterad forskare med expertkunskaper inom respektive område har varit ansvarig ledare för de olika projekten. Delprojekten, som alla har genomförts under 2014, varierar i tidsomfång - från fyra till elva månader. De kortare studierna syftar till att underlätta den nya myndighetens initiala arbete avseende processer och modeller för kunskapsbildning, och till att skapa gynnsamma förutsättningar för användning av forskningsbaserad kunskap i skolan. De längre projekten är exempel på olika typer av systematiska sammanställningar av forskningsresultat. De visar på olika modeller och metoder för hur forskning avseende lärande i skolan kan systematiseras och synliggöras.

Huvudsekreteraren för UVK, professor Eva Björck samt projektledaren, fil.dr. Cristina Robertson har varit ansvariga för projektet. SKOLFORSK har haft en referensgrupp med olika aktörer som arbetar med att befrämja praktiknära forskning och spridning av forskning. Projektet har haft nära kontakt med den grupp som planerat Skolforskningsinstitutet.

Ett varmt tack riktas till alla forskare som med kort varsel gjort det möjligt att genomföra detta projekt. Ni har berikat skolväsendet och Skolforskningsinstitutet med en gedigen bas att utgå ifrån i fortsatt arbete med skolans vetenskapliga förankring och uppbyggnad av den praktiknära skolforskningen i Sverige till gagn för förskolor, skolor och lärarutbildning.

Skolforskningsinstitutet önskas framgång och lycka med sitt fortsatta arbete!

*Petter Aaasen, ordförande, Utbildningsvetenskapliga kommittén*
*Eva Björck, huvudsekreterare för utbildningsvetenskap, Vetenskapsrådet*

Den här rapporten om betyg och summativa bedömningar är resultatet av ett uppdrag om underlagsrapporter från Utbildningsdepartementet/Vetenskapsrådet inför öppnandet av ett nationellt skolforskningsinstitut (SKOLFORSK).

Arbetet med rapporten har genomförts under ledning av Professor Christian Lundahl, Örebro universitet. Huvudansvarig för kapitel 1 har varit Lektor Alli Klapp, Göteborgs universitet, Larissa Mickwitz fil. lic. och doktorand vid Södertörns högskola har ansvarat för kapitel 2, Docent Magnus Hultén, Linköpings universitet, har haft huvudansvar för kapitel 3. Christian Lundahl har varit ansvarig för kapitel 4. Doktoranden Sverre Tveit vid Oslo universitet har tillsammans med Christian Lundahl ansvarat för sammanställning och analys av betygssystemen i Europa.

# INNEHÅLL

# SAMMANFATTNING

Den här forskningsöversikten om betyg bygger på en läsning av över 6000 abstracts ca 500 artiklar och ett 40 tal avhandlingar. De artiklar vi gått igenom är vetenskapligt granskade och publicerade i vetenskapliga tidskrifter. Våra sökningar och urval har varit systematiska.

Rapporten är uppbyggd kring fyra olika resultatkapitel kopplande till projektets fyra övergripande frågeställningar. I kapitel 1 studeras hur betyg ur ett elevperspektiv påverkar självbild, motivation och lärande. I kapitel 2 har vi sammanställt forskning om betyg ur ett lärarperspektiv, hur och vad lärare betygsätter och hur betyg påverkar undervisning. Kapitel 3 handlar om betyg som styrinstrument på olika nivåer i utbildningssystemet, framför allt nationell nivå. I kapitel 4 har vi beskrivit betyg ut olika komparativa perspektiv och studerat den forskning som finns där jämförelser sker mellan olika betygssystem och betyg och bedömning i ett internationellt perspektiv. Vi gör också en egen jämförelse av hur betygssystemen ser ut i Europa.

Den första delstudien har undersökt forskning om hur summativa bedömningar påverkar elevernas lärande, motivation för lärande och prestationer och vilka resultat den genererat. I delstudien ser vi att resultaten från de granskade artiklarna till viss del är samstämmiga. Vuxna högpresterande studenter verkar påverkas positivt i sitt lärande och prestationer av feedback som innehåller mycket information som kommer i direkt anslutning till uppgiften. Informationen bör också vara positiv. Samtidigt framkommer det att vuxna studenter inte påverkas negativt om feedback kommer i form av betyg. Detta förklaras av att vuxna studenter på universitetsnivå ”kan” systemet och har lång erfarenhet av summativa bedömningar och har utvecklat strategier för att hantera detta system samt att de är högpresterande. Däremot verkar det vara annorlunda för yngre elever och när representativa urval undersöks. En slutsats som kan dras av resultaten från de inkluderade studierna är att betyg generellt differentierar och påverkar äldre och yngre elever och låg- och högpresterande elever på olika sätt. Lågpresterande och yngre elever verkar påverkas mer negativt av betygsättning jämfört med äldre och högpresterande elever. Ålder och erfarenheter av bedömning tycks spela en stor roll för hur elevers lärande, motivation för lärande och prestationer påverkas av betygsättning.

Den andra delstudien handlar om hur och vad lärare betygsätter och hur betyg påverkar undervisning. Vi har studerat internationell respektive svensk forskning för att beskriva skillnader dem. Gemensamt är att validitetsfrågan är central men häri ligger också skillnaden. I svensk forskning är det relationen mellan lärarens betygsättning och *styrdokumenten* som dominerar perspektivet. Utanför Sverige är det framförallt frågan om *vad* läraren bedömer som dominerar, t.ex. elevens kunskaper eller personliga egenskaper.

Att lärares dagliga verksamhet påverkas av betygens inflytande är mer framträdande i den svenska forskning vi funnit. Här är det framförallt godkäntgränsen som problematiseras men även hur betyg tar tid från lärarens pedagogiska arbete. Betygens inverkan på lärarens undervisning är däremot inte centralt i forskningen utanför Sverige. Där dominerar istället kritiken mot ett ökat inflytande av *high-stakes* tester och hur lärare upplever dessa som meningslösa i sin undervisning. Standardisering av betygsättningen och *high-stakes* tester ses som ett problem som kan riskera lärares möjlighet att verka som professionella bedömare. Över huvud taget framkommer i de studier som tar upp betygens dilemman en spänning mellan styrning och kontroll och pedagogiska aspekter av lärarens bedömning.

I den tredje och fjärde delstudien har vi gått mer explorativt tillväga, då det inte funnits internationell forskning som primärt fokuserat betyg ur styrperspektiv. I delstudie tre fann vi tre centrala teman om betyg ur styrperspektiv: 1) Betyg ur rättvise- och jämlikhetsperspektiv, 2) Betyg som kunskaps- och urvalsmått, 3) Betyg som *high-stakes* i bedömnings- och utvärderingssystem. Det tredje temat gjordes till en inramning för de andra två. Den forskning som berörde första temat poängterade bland annat att betygssystem måste sättas in i ett större perspektiv av ett rättvist bedömnings- och utvärderingssystem, med instrument för att följa upp rättviseaspekter i relation till olika elevgrupper m.m. Studierna poängterade vikten av transparens i systemen, så att grunder för bedömning och utvärdering liksom existerande orättvisor blir synliga för systemets aktörer. Kunskapsfrågan lyftes också fram som central, det är lätt att anta att det som står i läroplanen – den kunskap som bedöms – är neutralt, men kunskapen har alltid konsekvenser och olika konsekvenser för olika grupper av elever. När det gällde tema två var ett tydligt resultat att betygens roll i många utbildningssystem reducerats de

senaste decennierna. Samtidigt visar genomgången av betyg ur ett systemperspektiv att betyg är bättre som urvalsinstrument för högre utbildning jämfört högskoleprov och andra liknande tester. I synnerhet kursbetyg på gymnasienivå som ges med stor bredd och i hög frekvens har en god predikativ förmåga. Detta visar att betyg kan fylla viktiga funktioner i ett utbildningssystem och det på ett bättre sätt än andra instrument, och att den utveckling som man sett internationellt mot allt mer centralt administrerade examens- och antagningsprov inte bör anammas okritiskt.

Den fjärde delstudien fokuserar betygen ur olika komparativa perspektiv. Det vi fokuserat på är vad betyg jämför samt hur olika betygssystem jämförs med varandra på nationell och internationell nivå. När vi söker på bedömning och internationella jämförelser ser vi att betyg inte får en särskilt framträdande plats i artiklarna. I huvudsak är det tre områden forskarna fokuserar vid dessa jämförelser: system för *accountability*; kulturella förklaringar till varför bedömnings- och betygssystem ser olika ut i olika länder; variationer mellan olika lärares bedömningar i olika ämnen eller av olika elevgrupper.

Några viktiga iakttagelser i vår genomgång är att det länge funnits en internationell trend mot att upprätta olika system för ökad ansvarsskyldighet (*accountability*) för skolans resultat. Dessa resultat mäts främst i elevprestationer på test eller i betyg. Såväl kritiska forskare som OECD har dock på senare tid noterat, att förhoppningarna om att jämförelser av skolors resultat ska leda till resultatförbättringar har varit överdrivna. De system olika länder har för bedömning och accountability förklarar i princip ingenting av variationen i PISA resultat. Det är snarare vad lärarna gör i klassrummet som har betydelse och lärare ha svårt att dra slutsatser om vad de bör göra utifrån de resultat som tillgängliggörs via *accountability*-modeller. Modellerna har sällan rätt informationsnivå för didaktiska slutsatser.

I kapitel 4 gör vi också en egen jämförelse av betygssystem i Europa i barn- och ungdomsskolan. Det första vi kan konstatera är att informationsläget är väldigt komplicerat. Det finns inte standardiserade data på detta varför alla jämförelser behöver bygga på komplicerat klassificeringsförfarande, där det ibland uppstår tolkningsproblem. Detta är inte bara ett problem för oss utan det finns i alla de jämförelser och hänvisningar till hur det ser ut i andra länder som också görs i den offentliga debatten om betyg. Enkla listor över när betyg ges i ålder eller i hur många skalsteg som används är ganska meningslös information utanför sitt kulturella och strukturella sammanhang.

Baserat på vad vi har fått fram i den här översikten har vi några rekommendationer. Det finns tydliga resultat som åtminstone bör mana till försiktighet om att vidare sänka åldern för betyg. Frågan är också på vilket sätt utblickar mot andra länders betygsstart kan hjälpa oss att ta kloka beslut om när vi ska börja med betyg, där vi efterlyser att man på policy nivå i så fall går mer på djupet och identifierar principer för bedömning som mer har med klassrummet att göra. Flera länder ger exempelvis lärare och skolor stor autonomi över hur bedömningarna i tidiga åldrar ska tillämpas, vilket kan tänkas ha positiva konsekvenser för lärares professionalitet i frågan.

Det är också viktigt att det svenska nuvarande betygssystemet bättre utvärderas på ett nyanserat sätt i förhållande till olika lärare, ämnen och elevgrupper. Betyg fungerar inte lika för alla. Det är också viktigt att fundera över hur vi utvärderar elevers resultat och om det finns möjlighet att kombinera fler modeller med varandra, så att vi bättre kan få data av ”value added”-karaktär samt för att följa kunskapsutvecklingen över tid. Studien visar också på flera olika plan vilka svårigheter det finns med översättning av forskningsresultat och information om utbildningssystem mellan olika länder och kontexter.

Vår studie pekar på att lärarnas autonomi över bedömningssystemen, oavsett hur de ser ut, är det som kanske har störst betydelse. Att lärarna har verktyg som de kan använda i bedömning av elevernas kunskaper och i kommunikationen kring dessa kunskaper som lärarna själva upplever är meningsfulla och som gagnar den pedagogiska processen. Därför är det också av stor vikt att lärare ges möjlighet till fortbildning kring betyg och bedömning och att det kanske blir ett ännu mer markerat inslag i lärarutbildningen.

# SUMMARY

This systematic research review about grades and summative assessments are based on a reading of over 6000 abstracts, 500 articles and about 40 theses. The articles we have read are peer reviewed and published in scientific journals. Our searches and selections have been systematic.

The report is structured around four different chapters linking to the project's four overarching issues. In Chapter 1 we study how grades/marks from a student perspective affects self-image, motivation and learning. In Chapter 2, we have compiled research on grading/marking from a teacher's perspective, how and what teachers think of this and how grading affect teaching. Chapter 3 deals with grades and summative assessment as control instruments at different levels of the education system. In Chapter 4 we describe grading from various comparative perspectives. We also do our own comparison of how the different grading and assessment systems look like in Europe.

In the first Chapter, we see that the results of the studies reviewed are partially coherent. Adults and high-performing students seem positively influenced in their learning and accomplishments from feedback that contains much information that comes directly adjacent to the task and if the information is positive. At the same time, it appears that adult students are not adversely affected if the feedback comes in the form of grades. This is explained by the fact that adult students at the university level and upper secondary education have extensive experience of summative assessments and have developed strategies to cope with this system. However, it seems to be different for younger students and when representative samples are examined. One conclusion that can be drawn from the results of the included studies is that grading generally influence older and younger students and low- and high-performing students in different ways. Underperforming and younger students seem to be more adversely affected by the scores compared with older and high-performing students. Age and experience of assessment appear to play a major role in how students' learning, motivation for learning and performance is influenced by the scores.

The second Chapter is about how and what teachers view of grading and how grading affect teaching. We have studied international and Swedish research to describe differences between them. The issue of validity is central, but in a different way in international and Swedish research, respectively. In the Swedish research, the relationship between the teacher's grading and *policy documents* constitutes a dominant perspective. Outside Sweden, it is mainly the question of *what* the teacher look at when assessing that dominate, e.g. student's skills or personal qualities.

In the third and fourth Chapter, we have used a more exploratory approach since grading isn't that closely linked to governing and control in other countries as in Sweden. Instead external tests are more common. We found however three central themes from a control perspective, that is relevant to the issue of grading: 1) fairness and equality in assessments, 2) grading as merit, as a knowledge and selection measurement, 3) grading as part of a high stakes assessment and evaluation systems. The third theme was made into a setting for the other two. The research that touched the first theme emphasized in particular that grading must be put into a larger perspective of a fair assessment and evaluation system, with instruments to monitor fairness in relation to different student groups, etc. Regarding the second theme we found that the ratings' role in many educational settings, have been reduced in recent decades. But at the same time we see clear tendencies that grades are better as a selection tool for higher education compared to university aptitude tests and other similar tests. This shows that grades can fill important functions in the education system in a better way than other instruments, but are not as useful for other purposes.

The fourth Chapter focuses on grades from different comparative perspectives. When we look at assessment and international comparisons we see that grades doesn't have a particularly prominent place in the international comparative research. Essentially, there are three areas the researchers focused on in these comparisons: systems of accountability; cultural explanations for why the assessment and grading system looks different in different countries; variations between teachers' assessments of various subjects or by different groups of students.

Some key findings of our survey is that there has long been an international trend towards establishing systems for measuring results and to increased accountability in education systems. These results are often

measured as student performance on tests or grades. Both critical scholars as well as the OECD has, however, recently noted that the hopes that comparisons of schools' results will lead to performance improvements have been exaggerated. The systems for assessment and accountability systems in different countries explain almost nothing of the variation in the PISA results. Rather, it is what teachers do in the classroom that are important and teachers find it difficult to draw conclusions about what they should do on those results that are made available through accountability systems. The systems seldom produce the right level of information for didactical implications.

In Chapter 4, we also do our own comparison of grading systems in Europe. The first thing we can say is that the data situation is very complicated. There is no standardized data on this, why all comparisons need to build on a complex classification procedure, where there sometimes are problems of interpretation. This is a problem for all references to how it looks in other countries so common in the public debate on grades in Sweden.

Based on what we found in our overview, we have some recommendations. There are clear results, which at least should lead to caution about further lowering of the age when pupils meet their first grades. It is also important that the Swedish current grading system is better evaluated in relation to different teachers, subjects and groups of pupils. Grades do not work the same for everyone. It is also important to consider how we evaluate students' performance and whether it is possible to combine more models with each other, so that we can get better data of for example "value added" character, and to be able to follow the development of knowledge over time. Our study also shows on several different levels of difficulties in the translation of research findings and information on education between different countries and contexts.

We suggest that teachers' autonomy in assessment systems, no matter what they look like, is perhaps the most important factor for them to work in purpose of support learning and development, at all levels. Therefore, it is also of great importance, not only for the government to pay attention to voice of teachers, but to provide teachers with the possibility to obtain further training on grading and assessment. Equally important, this aspect of teaching should be an even more marked feature of teacher education.

# INLEDNING

Den här rapporten handlar om betyg och om begreppet summativ bedömning som är den vetenskapliga benämningen på den typ av omdöme betyg utgör. En summativ bedömning är en bedömning vid en specifik tidpunkt av elevens kunskap inom ett avgränsat område (Harlen 2004). Bedömningen kan användas som en bedömning av elevens kunskap och eller som ett mått på skolans resultat. Det är i båda dessa betydelser vi griper oss an betyg.

Sverige är sannolikt det land i Europa där frågan om betyg debatterats mest det senaste halvseklet och där betygssystemet genomgått fler förändringar än i andra länder. De senaste stora förändringarna har föregåtts av utredningar och remissförfaranden men påfallande lite forskning. Först några år in på 2000-talet börjar det komma svensk forskning om bedömning. Viveca Lindberg (2005) har i en tidig kunskapsöversikt visat att den mesta svenska forskningen kring bedömning länge var knuten till konstruktionen av standardprov, centralprov och sedermera nationella prov. Den var nära knuten till de institutioner som utvecklade dessa prov och handlade om hur proven skulle kunna bli mer tillförlitliga och vilka slutsatser som kunde dras av resultaten. Dock har det funnits en brist på kunskap om vilka effekter betyg, prov och mer formativ bedömning har för lärandet, för lärarnas arbete och hur den information betyg och provresultat utgör kan användas i styrning, visar Lindberg och Forsberg (2010) i en senare kunskapsöversikt. Lindberg och Forsberg visar emellertid att det under 00-talet växer fram en bredare bedömningsforskning i Sverige.

Det finns hur som helst betydligt mer utländsk forskning om dessa fenomen. Några tidiga översikter är exempelvis Harlen och Deakin Crick (2002) och Harlen (2004). En tydlig tendens i den internationella forskningen om bedömning är att den inte har handlat om betygens vara eller inte vara utan om skillnaden mellan summativ bedömning (vilket betyg brukar betecknas som) och formativ bedömning vilket mer handlar om hur bedömningen används. Att forskningsläget ser ut på detta sätt beror troligtvis på att lärarsatta betyg fått en allt mindre roll i den engelsktalande världen de senaste decennierna och att istället nationellt administrerade tester kommit att ersätta många av de lärarsatta betygens funktioner. Så som forskningsläget sett ut i Sverige och internationellt är det därmed svårt att få svar på de frågor om betygens förtjänster och brister som vi ställer i Sverige. Vi menar dock att mycket av det som skrivits om summativ bedömning, vilket ofta handlar om prov, test och examinationer också kan användas för att säga något om betyg. Vi har därför utöver att systematiskt analysera forskning om betyg också gått igenom forskning utifrån det vidare begreppet summativ bedömning (se vidare kapitel 1).

Syftet med vår studie är att systematiskt kartlägga och redovisa forskningsläget i Sverige och internationellt vad gäller betyg och betygssättning i relation till elevers kunskapsutveckling. Kartläggning har ett kombinerat internt och externt perspektiv och beskriver betygens direkta effekter i lärandet och indirekta effekter som styrmedel för skolan. Systematiska litteraturstudier har som metod utsatts för stark kritik (se MacLure 2005). Inte sällan ger initiala sökningar tusentals träffar som sedan sållas ner till en handfull studier som analyseras på djupet. Vad är det för svar vi får när merparten av studier avfärdas? I vårt anslag har vi försökt att delvis parera för denna kritik genom i högre grad ”fria” än fälla, dvs. inkludera studier även om dessa inte håller måttet när det gäller krav på effektstudier (bl.a. randomiserade urval, kontrollgrupper, att man kontrollerat för relevanta påverkansfaktorer). Det innebär att vi utöver effektstudier även inkluderat relevanta studier av mer kvalitativt slag samt även en del teoretiska bidrag. I respektive resultatkapitel beskriver vi utförligt våra sökstrategier och urvalsstrategier för att så långt som möjligt uppfylla kraven på replikerbarhet.

Med utgångspunkt i vår kartläggning drar vi också slutsatser om konsekvenser av betyg och betygssättning av elever i olika åldrar, i relation till elevers lärande och motivation, samt i förhållande till utvärdering och styrning av skola.

Rapporten är indelad i fyra kapitel. I kapitel 1 studeras hur betyg ur ett elevperspektiv påverkar självbild, motivation och lärande. I kapitel 2 har vi sammanställt forskning om betyg ur ett lärarperspektiv, hur och vad lärare betygsätter och hur betyg påverkar undervisning. Kapitel 3 handlar om betyg som styrinstrument på olika nivåer i utbildningssystemet. I kapitel 4 har vi beskrivit betyg ut olika komparativa perspektiv och studerat den forskning som finns där jämförelser sker mellan olika betygssystem och betyg och bedömning i ett internationellt perspektiv. Vi gör också en egen jämförelse av hur betygssystemen ser ut i Europa.

Rapporten är en så kallad *Systematic Research Review* vilket innebär att vi inte genomför någon egen empirinsamling utan istället systematiskt gått igenom olika forskningsdatabaser och vetenskapliga tidskrifter i syfte att se vad andra forskare har för resultat och analyser kring betyg. En del av kartläggningen är kvantitativ och handlar om att beskriva vilka aspekter av betyg och betygssättning svensk respektive internationell forskning fokuserat på från 2000-talet fram till idag. Vi avgränsade oss till att börja runt 2000-talet dels för att ett par stora kunskapsöversikter genomfördes då som väl täcker upp bilden fram till dess (se kapitel 1), dels för att utvecklingen både av bedömningspraktiker och forskning om bedömning gått snabbt det senaste decenniet och en hel del äldre forskning på området kan betraktas som obsolet. Men vi har låtit studier före 2000 komma in i de fall de bedömts som centrala inom fältet.

I systematiska litteraturstudier ingår att tydligt redogöra för metoder och kriterier för sökning och urval av artiklar. De inkluderade studierna har kodats och tematiserats på ett systematiskt sätt. Tillvägagångssättet har varit lite olika kring våra fyra olika huvudområden och redovisas därför i inledningen till respektive resultatkapitel. Gemensamt är att vi i huvudsak fokuserat åren 2000–2014. Vi har också i kapitel 1 och 2 som har ett elev- respektive lärarperspektiv avgränsat oss till empiriska studier, dvs. där det finns data på konsekvenser av betyg och summativ bedömning. Kapitel 3 och 4 utgår främst från empiriska studier men här finns också några mer filosofiska och begreppsutredande studier med i urvalet. Vi har vidare valt att avgränsa oss till referee-granskade artiklar. Därigenom minskar mängden artiklar att gå igenom men de artiklar vi får fram håller ofta en hög kvalitet. Vi har i huvudsak sökt på engelskspråkiga artiklar, men i kapitel 2 kring lärarperspektivet och i kapitel 3 kring systemperspektivet, har vi också gått igenom forskning presenterad på svenska, norska och danska. Här har det funnits ett större skandinaviskt underlag jämför med områdena i de andra två kapitlen. Den jämförelse vi gör av de europeiska ländernas betygssystem i kapitel 4 utgår från en databas som kallas *EURYDICE*. Det är en informationsdatabas om EU-ländernas utbildningssystem som tillhandahålls av Europeiska kommissionen. Den bygger på självrapporteringsprinciper och håller en något ojämn kvalitet. Vi diskuterar de metodologiska implikationerna av detta vidare i kapitel 4. Huvudsaken här är emellertid att belysa de variationer som finns mellan ländernas betygssystem och ge några rimliga förklaringar till det.

Arbetet med den här rapporten har genomförts under ledning av Professor Christian Lundahl, Örebro universitet. Huvudansvarig för kapitel 1 har varit Lektor Alli Klapp, Göteborgs universitet, Larissa Mickwitz fil lic och doktorand vid Södertörns högskola har ansvarat för kapitel 2, Docent Magnus Hultén, Linköpings universitet, har haft huvudansvar för kapitel 3. Christian Lundahl har varit ansvarig för kapitel 4. Doktoranden Sverre Tveit vid Oslo universitet har tillsammans med Christian Lundahl ansvarat för sammanställning och analys av betygssystemen i Europa. Projektet genomförts under hösten 2014 och artiklar publicerade i sent i december ingår inte i sökningarna. Genom möten i projektgruppen har strategier för sökningar och avgränsningar diskuterats och våra enskilda dokument och filer har gjorts tillgängliga för gruppen via gemensamma kataloger. På så vis har vi gjort avstämningar om var vissa svårklassificerade artiklar hör hemma. I några fall diskuteras samma artikel i fler än ett kapitel, men då ur olika perspektiv.

# BETYGENS EFFEKT PÅ MOTIVATION OCH LÄRANDE

Betyg är en summering och sammanfattning av elevens lärande vid en viss tidpunkt. Ofta sker en summering i slutet av terminen eller i slutet av en kurs på gymnasiet. En summativ bedömning kan även innebära att kortare delmoment av en kurs summeras till exempel genom prov och att resultaten från ett antal prov senare ligger till grund för ett betyg. I detta kapitel studeras hur betyg och summativa bedömningar ur ett elevperspektiv påverkar självbild, motivation och lärande.

Att kunna definiera aspekter av lärande, motivation för lärande och prestationer är komplicerat. Lärande bör optimalt resultera i prestationer som kan mätas av till exempel prov och betyg. Motivation för lärande tar dessutom in aspekter kring elevers socioemotionella kompetenser och hur dessa påverkar motivation och i sin tur prestationer. Det finns skäl att anta att elevers lärande påverkas av deras motivation men att också motivation påverkas av elevens lärande och att dessa reciproka relationer påverkar prestationer. Inom motivationsteorier diskuteras hur olika drivkrafter påverkar elevernas lärande och prestationer (Cameron 2001, Dweck 1992, Deci, Koestner & Ryan 2001, Deci & Ryan 1985, Duckworth & Seligman 2005, Molden & Dweck 2006) och hur prestationer påverkar elevernas motivation. Att ett lärande har ägt rum definieras i skolan ofta utifrån elevers prestationer på olika typer av prov och genom betyg. Andra typer av ”mått” på elevers prestationer kan vara enkäter, observationer eller intervjuer där elevernas lärande kan synliggöras.

Forskning inom området där relationer mellan betyg och lärande och prestationer har fokuserats har inte varit tydligt definierat och området är multidisciplinärt. I en tidigare översikt om hur summativa bedömningar påverkade elevers motivation för lärande noterade Harlen och Deakin Crick (2002) dels att det saknades empiriska studier av tillräckligt god kvalitet inom området, dels att det saknades utvecklade teorier inom området. Harlen och Deakin Cricks översikt inkluderade studier fram till år 2002. Eftersom intresset för bedömningsfrågor har ökat avsevärt under de senaste 10 åren är det rimligt att anta att antalet studier inom bedömningsfältet också har ökat under denna period. Det första syftet med undersökningen i detta kapitel, har varit att genomföra en systematisk översikt kring summativa bedömningars effekter på elevernas lärande, motivation för lärande och prestationer. Detta har inneburit att en mängd olika aspekter och kombinationer av summativa bedömningar, som betyg, prov, elevers prestationer och lärande har legat till grund för litteratursökningarna. Det har även varit av intresse att kartlägga vilka åldrar som undersökts och vilka forskningsmetoder som använts.

Ett andra syfte med undersökningen i detta kapitel har varit att göra en fördjupad analys av de relevanta inkluderade studierna. Den fördjupade analysen innehåller en sammanfattning av de relevanta studierna och en diskussion om hur resultaten kan tolkas. Forskning inom området fokuserar dels på hur summativa bedömningar generellt påverkar elevers lärande och prestationer, dels hur specifika former av praktiker inom summativ bedömning påverkar elevers lärande och prestationer. Inom forskningsfältet är resultaten i viss mån disparata och visar både på positiva och negativa effekter av summativa bedömningar på elevers lärande och prestationer. Studier med positiva effekter av betygsättning på elevernas lärande och prestationer är få och de disparata resultaten kan förklaras av att studierna undersökt olika elevgrupper med avseende på ålder och förmåga samt att studierna gör olika teoretiska och metodiska antaganden. Det är inte heller alltid klargjort i studierna hur olika elevgrupper påverkas och hur olika individuella karaktäristika, miljöfaktorer som hemmets förutsättningar påverkar relationen mellan summativa bedömningar och elevers lärande och prestationer.

## Resultat av tidigare genomförda översikter inom området

Ett flertal översikter är gjorda med fokus på summativa bedömningar och deras effekter på olika aspekter av elevers utveckling. Crooks (1988) granskade över 200 studier om effekter av bedömning på elevers lärande. Crooks sammanfattade forskningen och menade att bedömning av elevers prestationer för betygsättning överskuggar användandet av bedömning för att stödja elevers lärande. Black och Wiliam (1998) fann i sin ofta citerade översikt att inget är förändrat från Crooks översikt utan menade att forskningen bidrar med ett stort antal studier där resultaten visar att många av de vanligt förekommande betygsättningspraktikerna bidrar till att elever lär sig mindre och presterar sämre. Kluger och DeNisi (1996) genomförde en översikt kring hur

bedömning påverkar skolor och arbetsplatser och av ca 3000 studier fann de att enbart 131 av dessa hade genomförts med tillräcklig kvalitet när det gäller metod, noggrannhet och med tillräckliga detaljer kring tillvägagångssätt presenterade för att anses vara reliabla studier. Av dessa 131 studier var det ett femtiotal som visade att feedback i form av summativ bedömning påverkade människors prestationer negativt. I dessa fall hade det varit bättre om ingen feedback hade getts. Kluger och DeNisi menade att feedback som påverkar människor negativt fokuserar på personen och inte på uppgiften vilket leder till negativa effekter på lärandet. När feedback istället fokuserar på hur uppgiften kan bli bättre, hur eleven kan göra för att förbättra uppgiften, ökar lärandet avsevärt.

Harlen och Deakin Crick (2002) gjorde en systematisk översikt kring hur summativa bedömningar påverkar elevers motivation för lärande. Efter avgränsningar i deras översikt analyserades 19 studier som de fann relevanta för syftet med översikten. De drar slutsatserna att: lågpresterande elevers självkänsla påverkas negativt av summativa bedömningar; vid high-stakes summativa bedömningar anlägger läraren en förmedlande undervisningspraktik som gynnar de elever som lär sig på detta sätt medan elever som lär sig utifrån mer aktiva och elevcentrerad undervisning missgynnas; elever ogillar summativa bedömningar (prov) och speciellt flickor missgynnas av prov; summativa bedömningar gör att elever får ett mer ytligt och prestationsinriktat förhållningssätt till lärande; summativa bedömningar riskerar att leda till att elever utvecklar ytliga lärandestrategier. De fann att äldre elever (över 11 år) hade lättare att förstå innebörden av betyg och sätter större värde på betyg, men de äldre eleverna ansåg att betygen oftare är orättvisa jämfört med vad yngre eleverna tyckte. Äldre elever fokuserar mer på prestationer och resultat än på lärandeprocessen. Lågpresterande äldre elever minskade sin ansträngning att lära sig jämfört med lågpresterande yngre barn och de äldre eleverna utvecklade ett mer cyniskt förhållningssätt till summativa bedömningar. Harlen och Deakin Crick fann även att lågpresterande elever blev dubbelt missgynnade av summativa bedömningar eftersom de blev ”märkta” som misslyckade elever vilket påverkade deras redan låga självkänsla ännu mer, något som i sin tur minskar deras möjlighet att i framtiden kunna anstränga sig och lyckas i skolan. Endast när en lågpresterande elev hade stöd från skolan och hemmet som hjälpte eleven att utveckla strategier för hur den kan lyckas, gick det att komma ur denna negativa spiral. De fann även att summativa bedömningar som är av *high-stakes* karaktär för den enskilda eleven har en särskilt negativ effekt på lågpresterande elever (*high-stakes* innebär långtgående konsekvenser för de som är inblandade). Högpresterande elever är mer uthålliga vid bedömningar, använder mer framgångsrika lärandestrategier, gillar att bli bedömda och har högre uppfattning om den egna kompetensen. Harlen och Deakin Crick menar att ett ökat användande av summativa bedömningar leder till en ökad differentiering mellan elever och ökar gapet och segregationen mellan elever.

Sammanfattningsvis menar författarna av tidigare forskningsöversikter inom bedömningsområdet att summativ bedömning har negativa effekter på lärandet och särskilt för lågpresterande elever. Forskarna är dock tydliga med att det behövs bättre metoder och design på framtida forskningsstudier och att det finns ett behov av teoretisk utveckling kring summativa bedömningars effekter på elevers lärande och prestationer. I översikterna är det i slutändan ganska få studier det gått att dra generella slutsatser från. Hur forskningsläget ser ut idag visar vi i nästa avsnitt.

# Metodbeskrivning

## Betyg och summativa bedömningar – precisering av söktermer

I Sverige används betyg i stor utsträckning som urvalsinstrument för vidare nivå inom utbildningssystemet medan i andra länder är det relativt ovanligt att använda betyg som urvalsinstrument. Betygen kan i andra länder istället vara en garanti för att eleven har genomfört studierna och att eleven har uppnått vissa kunskapskrav (se vidare kapitel 4).

I Sverige används nationella prov för att utvärdera utbildningssystemet, kalibrera lärarnas betygsättning samt för att implementera styrdokumenten. De är dock är inte tänkta att fungera examinerande eller att användas vid urval. I andra länder används olika typer av examinerande prov som urvalsinstrument inom utbildningssystemet. Det svenska betygsystemets konstruktion är relativt ovanligt i internationellt hänseende,

delvis på grund av att betygen i stor utsträckning används som urvalsinstrument. En annan aspekt som behöver beaktas är graden av *high-stakes* i olika summativa bedömningar. Om en bedömning har stora konsekvenser för eleven som till exempel betygen i årskurs 9, kan effekterna av dessa typer av bedömning för elevens lärande, motivation för lärande och prestationer antas vara stora. Eftersom betygen ofta inte är av *high-stakes* karaktär i andra länder bör en forskningsöversikt inom bedömningsområdet söka efter forskning som är av *high-stakes* karaktär, som till exempel olika typer av prov, för att förstå vad den här typen av bedömningar har för effekter. Det har med andra ord varit nödvändigt att använda olika termer för betyg och betygsättning i våra sökningar. Efter diskussioner inom projektgruppen bestämdes att söktermerna skulle identifiera studier som relaterade till summativa bedömningar. En naturlig ingång var då följande söktermer (på svenska och engelska):

Betyg, betygsättning
Prov, test, testning
Internationella, nationella och lokala prov

## Betyg och summativa bedömningar – teoretiska utgångspunkter

Under lång tid har forskning visat att det finns samband mellan summativa bedömningar och elevers prestationer och lärande. I forskningen finns främst två övergripande modeller som används som förklaring till varför summativa bedömningar påverkar elevernas prestationer och lärande. Den första övergripande förklaringsmodellen innebär ett antagande om att alla elever påverkas av bedömning på ett sätt som gör att de, oavsett förutsättningar, blir motiverade att lära sig mer och prestera bättre. Om informationen i ett betyg är lägre än vad en elev förväntat sig, blir eleven ”bestraffad” vilket tänks leda till att eleven bli mer motiverad och därför kommer eleven att prestera bättre. Om informationen i betygen däremot är i linje med eller bättre än elevens förväntningar blir eleven ”belönad” och motiverad att lära sig mer och prestera bättre. Denna förklaringsmodell, som främst har utvecklats inom den ekonomiska forskningsdisciplinen genom studier inom idrotts- och tävlingsområdet (se exempelvis Prendergast 1999 angående *Relative Performance Information* och *Tournament Theory*) innebär att alla elever oavsett förmåga, kön och andra bakgrundsfaktorer och förutsättningar kommer att reagera på summativa bedömningar som belöning eller bestraffning på ett sätt som leder till positiva resultat. En utveckling av denna modell innebär att summativa bedömningar främst hjälper resurssvaga elever med dåliga förutsättningar och med låg social bakgrund eftersom bedömning kan hjälpa dem att företa en ”klassresa” och att bedömningar i skolan hjälper dem att komma vidare inom utbildningssystemet (Azmat & Iriberri 2009, Bandiera, Larcinese & Rasul 2008, 2009, Cameron, 2001, Cameron, Banko & Pierce 2001, Sjögren 2010). Ytterligare en utveckling av denna modell bygger på att alla elever presterar bättre om de kan jämföra sina resultat med varandra och att det skapas en möjlighet till sociala jämförelser i klassen (Azmat & Iriberri 2010, Becker & Rosen 1992, Bandiera et al. 2008, 2009). Denna modell bygger på synsättet att elever motiveras att prestera bättre om de utsätts för jämförelser och kan tävla mot varandra: att tävling sporrar till bättre prestationer. Denna modell bygger främst på empiriska studier inom den ekonomiska disciplinen och ekonomiska teorier utvecklade utifrån forskning kring tävling och lagutveckling (Prendergast 1999).

Den andra övergripande förklaringsmodellen bygger på att *high-stakes* summativa bedömningar som betyg och prov påverkar elever på olika sätt beroende på elevernas förutsättningar och bakgrund. Feedback i form av betyg kan antingen vara positiv eller negativ för eleven och påverkar därför elevens självkänsla, motivation, lärande och prestationer. Inom denna förklaringsmodell är de komplexa relationerna mellan bedömning (negativ och positiv feedback), elevens akademiska och sociala självkänsla, motivation, lärande och prestationer i fokus. Ett kluster av teorier förklarar hur dessa relationer relateras till varandra. *The Conservation of Resource Stress* Theory (Covington 2000, Frydenberg 2008, Hobfoll 1989) förklarar att elever strävar efter att behålla, skydda och utveckla sina egna personliga resurser för att lyckas i skolan. Resurser kan vara personliga förutsättningar som akademisk självkänsla, syn på sig själv som en person som kan lära sig och som tror att förmågor går att utveckla och drivkraft. När dessa resurser hotas, av till exempel misslyckanden och dåliga resultat i skolan, kan förlusten av de personliga resurserna orsaka emotionell stress. Denna stress kan i sin tur leda till att eleven utvecklar ytliga lärandestrategier för att undvika misslyckanden och nedvärderar

betydelsen av skolarbetet. När förluster av resurser ökar kan eleven bli frustrerad och detta kan i sin tur leda till olika destruktiva beteende och uppförandeproblem och risken för skolmisslyckanden ökar. Inom dessa teorier är elevens syn på sig själv central för att förklara konsekvenser av skolmisslyckanden. Inom denna teoribildning fungerar summativa bedömningar differentierande vilket innebär att till exempel prov och betyg påverkar resurssvaga och lågpresterande elevers motivation, lärande och prestationer negativt medan resursstarka och högpresterande elever inte i samma utsträckning påverkas negativt (Butler 1988, Deci, Koestner & Ryan 1999, 2001, Hattie, 2009, Black & Wiliam 1998). Denna modell för att förklara hur summativa bedömningar påverkar elevers lärande och prestationer bygger på olika motivationsteorier som omges av kontroverser mellan forskare (Cameron 2001, Deci & Ryan 1985, Deci, Koestner & Ryan 1999, 2001).

## Litteratursökningar

Under hösten 2014 genomförde vi systematiska sökningar i EBSCO (ERIC och ProQuest) och i LIBRIS. Dessa databaser täcker forskningslitteratur inom ett stort antal discipliner som utbildning, psykologi, sociologi, organisationslära med flera. Den totala sökprofilen är utrymmeskrävande och därför presenteras enbart delar av den i denna översikt. De främsta vetenskapliga tidskrifterna inom bedömningsfältet: *Assessment in Education*, *Educational Psychology*, *Pedagogisk Forskning i Sverige*, *Scandinavian Journal of Education* genomsöktes efter relevanta studier. Vi fann också ett antal studier vid manuell sökning i andra internationella och svenska tidskrifter av betydelse för området.

Vilka studier som skulle bilda ett underlag för översikten följde till en början breda och inte strikta avgränsningar för att inte utelämna studier. Studier som valdes ut skulle vara skrivna på engelska, svenska, danska eller norska. Studierna skulle ha en empirisk design med analys av data, variabler/information skulle fokusera på av summativa bedömningar och någon av variabler som handlar om lärande, motivation för lärande och prestationer. Artiklarna skulle vidare vara granskade (peer-reviewed) och publicerade mellan januari 2002 och december 2014. I det inledande skedet lästes abstracts för 2633 antal studier. Efter denna genomgång fanns 174 studier kvar som på något sätt föll inom de uppsatta kriterierna.

Det inledande arbetet med delprojektet innebar att relevanta söktermer definierades och hur dessa söktermer kunde kombineras. Sökprofilerna bygger på kombinationer av termer som handlar om a) betyg och betygsättning; b) summativa bedömningar c) prov d) motivation och motivation för lärande; e) lärande; f) skolprestationer. Inledningsvis gjordes breda sökningar på termer som ”*grading*”, ”*testing*”, ”*summative assessment*” och ”*achievement*” var och en för sig. Sökningarna gjordes i all text, det vill säga inte enbart i abstrakt och titel. Dessa inledande sökningar gav oss ett stort antal träffar, se tabell 1 och 2 nedan.

***Tabell 1. Sökning i databasen ERIC (EBSCO) granskade (peer-reviewed) (2014-12-15).***

| **Sökord** | **Antal referenser** |
|---|---|
| Grading | 2775 |
| Grades | 26935 |
| Testing | 54652 |
| Tests | 76671 |
| Summative assessment | 384 |
| Achievement | 54824 |
| Learning | 180205 |
| Motivation for learning | 4308 |

*Tabell 2. Sökning i databasen ERIC (ProQuest) granskade (peer-reviewed) (2014-12-15).*

| Sökord | Antal referenser |
|---|---|
| Grading | 2889 |
| Grades | 29283 |
| Testing | 57887 |
| Tests | 85562 |
| Summative assessment | 821 |
| Achievement | 56909 |
| Learning | 184760 |
| Motivation for learning | 7277 |

På grund av det stora antalet referenser vid de inledande sökningarna snävades dels sökprofilerna in, dels kombinerades de olika söktermerna med varandra för att få mer relevanta resultat. Vi avgränsade inte populationen utifrån ålder utan studier med deltagare i alla åldrar inkluderades, på grund av att vi fann att studier där deltagarna var i skolåldern (årskurs 1 till 12) var relativt få.

Ytterligare avgränsningar gjordes när det gällde söktermerna ”*testing*” och ”*achievement*” samt ”*testing*” och ”*learning*” eftersom de gav ett stort antal irrelevanta referenser. Efter att dubbletter tagits bort återstod 2633 referenser som ansågs vara relevanta för projektet och granskades utifrån abstrakt. Om ett abstrakt ansågs vara relevant för översikten sparades abstraktet ner i databasen RefWorks (RW). I de fall där hela artikeln fanns tillgänglig sparades de manuellt i en databas, resterande artiklar beställdes. Detta resulterade i att 174 artiklar har lästs i sin helhet och av dem ansågs 67 studier var relevanta för syftet och vid en närmare genomgång av dem ansågs slutligen 22 studier vara relevanta för att ingå i den fördjupade analysen. Urvalet av artiklar skedde enligt följande principer:

### Deltagare i studierna

De inkluderade studierna skulle undersöka barn, ungdomar och vuxna i utbildningssammanhang.

### Design

De inkluderade studierna skulle vara empiriska undersökningar. Vi hade dock inga restriktioner om vilken typ av empiriska undersökningar det kunde handla om.

### Kriterier för inklusion

Artiklar skrivna på engelska och de nordiska språken i granskade (peer-reviewed) tidskrifter inkluderades. Rapporter och avhandlingar har inte tagits med i den fördjupade analysen. Studier inkluderades om de genomförts från januari 2002 fram till och med december 2014.

### Genomgång av abstrakt

De 2633 abstrakten lästes och sorterades utifrån att någon aspekt av söktermerna var inkluderade i abstraktet eller att innehållet i abstraktet på något viss handlade om summativa bedömningar, lärande och prestationer. Vid tveksamhet kontrollerades artikeln i sin helhet. Av de genomgångna abstrakten ansågs 174 artiklar vara relevanta för översikten och hela artikeln sparades i en fil namngiven med de specifika söktermerna. Dessa artiklar lästes i sin helhet. Artiklarna som valdes bort var inte relevanta för översiktens syfte dels på grund av

att innehållet i artiklarna enbart fokuserade på någon del av lärande, prestationer och summativa bedömningar, dels att utfallsvariablerna inte mätte elevernas lärande eller prestationer.[1]

### Genomgång av artiklar i fulltext

När de 174 artiklarna lästes skapade vi ett protokoll där studiernas karaktärsdrag med avseende på syfte, metod, urval och resultat dokumenterades. Artiklarna var tillgängliga genom RefWorks och via länkar till Göteborgs universitets databaser. Vid en närmare genomläsning av de 174 artiklarna ansågs 107 vara icke relevanta för syftet med översikten. De resterande 67 artiklarna var relevanta för frågeställningen (summativa bedömningars påverkan på lärande och prestationer) men enbart 22 av artiklarna fokuserade på betygsättning i någon form och betygens effekter på elevernas lärande och prestationer. Många av artiklarna fokuserade på prov, testning och *accountability* och effekter av dessa på en mängd olika utfallsvariabler och på aggregerad nivå. För att uppnå syftet med denna översikt bestämdes att fokus i den fördjupade analysen skulle ligga på effekter av betygsättning (feedback i form av poäng och/eller betyg) på elevers lärande och prestationer. I kapitel 3 och 4 har vi däremot ett lite vidare perspektiv på betygens effekter t.ex. i utvärdering och policyimplementering.

### Granskning av artiklarna

Studiernas syften klassificerades vid genomläsningen och dokumenterades i protokollet. Totalt 22 av de 67 studierna hade som syfte att undersöka hur summativ bedömning i form av prov och tester påverkade olika aspekter av elevers lärande, motivation för lärande och prestationer. De resterande 45 studierna hade som syfte att undersöka summativa bedömningar i form av betygsättning, prov och tester för elevers lärande, motivation för lärande och prestationer. Av dess 45 studier var 22 studier fokuserade på effekter av betyg och betygsättning. De resterande 23 studierna fokuserade på mer begränsade och specifika aspekter av betygsättning till exempel betygsskalor och antal skalsteg och dess betydelse för elevernas lärande, motivation för lärande och prestationer. Nedan beskrivs de 22 studierna som ingår i den fördjupade analysen.

## Tematisering av inkluderade studier

Av de 22 utvalda studierna var 8 longitudinella och 3 presenterade analyser på flernivå. Totalt var 5 av studierna komparativa studier och 16 av studierna var experiment eller interventionsstudier. Totalt var 4 studier intervjuer- och/eller observationsstudier. Den vanligast förekommande åldersgruppen i studierna var vuxna studenter på universitetsnivå (N = 14). Totalt 6 studier undersökte elever i åldersspannet 9 till 16 år. Den största delen av studierna genomfördes i USA (N = 7). De resterande är studier från 9 olika länder. Sverige har med tre studier. De inkluderade studierna publicerades mellan januari 2002 och december 2014.

Utifrån den kvantitativa granskningen av litteraturen finner vi således ett antal studier som undersöker hur summativa bedömningar påverkar elevernas prestationer, lärande och motivation för lärande och som vi väljer att närmare studera. De olika typerna av betygsättningspraktiker och hur olika betygssystem påverkar elevernas möjlighet till prestationer och lärande finns som syfte i ett större antal studier jämfört med hur betygsättning och summativ bedömning mer generellt påverkar elevers prestationer, lärande och motivation för lärande. I en mängd studier undersöks inte om summativa bedömningar generellt påverkar elevernas prestationer och lärande utan det är mer specifika betygsättningspraktiker som undersöks till exempel hur olika betygsskalor påverkar elevers motivation och prestationer. Betygsskalor kan vara konstruerade med färre eller fler antal skalsteg till exempel med två skalsteg som godkänt/underkänt eller med fler skalsteg så kallad diskriminerande skala som vår nuvarande betygsskala som går från E till A är ett exempel på.

[1] Sjögrens studie (2010) inkluderades i översikten trots att utfallsvariablerna inte mäter elevernas lärande, motivation för lärande eller prestationer. Att den inkluderats beror på att studiens resultat har använts vid implementering av införandet av betyg i årskurs 6.

Huvudfrågan för detta kapitel är hur summativa bedömningar påverkar elevers lärande, motivation för lärande och prestationer. Resultatet från den första delen av litteraturbearbetningen visade på att det finns en mängd studier som undersöker denna fråga eller närliggande aspekter av frågan. Dock är det tveksamt om det föreligger en empirisk grund som är tillräcklig för att kunna besvara frågan. Innan vi går in i en mer fördjupad analys och tematisering av de 22 relevanta studierna kommer en diskussion kring studiernas metodiska förutsättningar att diskuteras.

### Metodologiska dilemman i inkluderade studier

De metodologiska utmaningarna för att undersöka effekter av summativa bedömningar på elevers lärande och prestationer är stora. Randomiserade experiment är det som av många forskare anses vara den bästa metoden för att kunna göra orsaksanalyser. Ingen av studierna i denna översikt har denna design. Svårigheterna med att genomföra ett randomiserat experiment kan delvis handla om att det är etiskt tveksamt att genomföra ett randomiserat experiment med elever eftersom vissa elever då inte får den behandling som tidigare forskning visat vara effektiv eller att betyg måste ges för att eleven ska kunna söka sig vidare till högre utbildning. Det kan även handla om att det är svårt att skapa förutsättningar för experiment eftersom alla elever i viss årskurs i ett utbildningssystem har samma styrdokument vilket gör det svårt att jämföra grupper av elever eftersom det inte finns variation om och när betyg sätts. Inom samhällsvetenskaplig forskning finns även en stor mängd faktorer som kan orsaka ett visst utfall, faktorer som är svåra att till fullo kontrollera för. Till exempel är det svårt att kontrollera för alla möjliga tänkbara faktorer som påverkar lön i vuxen ålder, en vanlig utfallsvariabel i utbildningsekonomiska studier. Longitudinella studier är en typ av forskningsdesign där det kan vara möjligt att göra orsaksanalyser. Ett fåtal studier i denna översikt har en longitudinell design och bristen på studier med randomiserad, experimentell design med möjlighet att jämföra grupper av elever eller interventioner är stor. På grund av flera orsaker är det i detta läge tveksamt att genomföra en metaanalys, dels för att det finns få studier tillgängliga som använder metoder där det är möjligt att jämföra elevgrupper med hjälp av experiment eller interventioner och som har kvantitativa utfallsmått, dels bristen på tid inom detta projekt. Däremot kan en metaanalys genomföras inom ramen för ett senare projekt. En fördjupad analys som fokuserar på vilka slutsatser och generaliseringar vi kan göra i respektive studier har därför genomförts istället för en metaanalys där genomsnittliga effekter vanligen står i fokus.

### Tematisering av inkluderade studier

Kriterierna för att inkluderas i den fördjupade analysen var att studierna skulle ha som syfte att undersöka hur summativa bedömningar i form av betygsättning (poäng och/eller betyg) påverkar elevernas lärande, motivation för lärande och prestationer. Ett annat kriterium var att studierna skulle ha som utfall mått på prestationer, lärande eller motivation för lärande. De 22 inkluderade studierna var publicerade i 15 olika vetenskapliga tidskrifter. Ett protokoll uppfördes som strukturerade de inkluderade artiklarna utifrån: syfte; relevans; urval och kontext; urvalsstrategi; metod; datainsamling; dataanalys; resultat och sammanfattning och studiens kvalitet.

De inkluderade studierna kan delas in i tre övergripande teman. Det första temat är studier som undersöker hur betyg och poäng som feedback i allmänhet påverkar elevers prestationer och motivation för lärande. Detta innebär att det är *high-stakes* bedömningars påverkan på elevernas prestationer och motivation för lärande som är i fokus.

Det andra temat innehåller studier som har fokus på att undersöka olika typer av betygsättningspraktikers påverkan på elevers prestationer och lärande. Detta innebär jämförelser mellan formativ och summativ bedömningspraktik där den summativa praktiken ofta används som ”business-as-usual” för att kunna utvärdera effekter av en ny formativ bedömningspraktik på elevernas lärande och prestationer. I detta tema finns även 2 studier som undersökt hur högre och lägre betygskrav påverkar elevers prestationer.

Det tredje temat handlar om hur positiv och negativ information i summativa bedömningar påverkar elevers motivation för lärande och prestationer.

Studierna inom dessa teman har flera olika utfallsvariabler. Utfallet mäts i betygsresultat, i elevernas motivation för t.ex. lärande och i ekonometriska utfall som inkomst, utbildningslängd och om eleven har avslutat skolgången.

## Betyg som feedback

De texter som studerats i gruppen Betyg som feedback handlar om hur betyg (betygsättning) påverkar elever och studenters lärande och prestationer. Dessa studier försöker identifiera effekter av betyg och betygsättning på lärande och prestationer.

Utbildningsekonomerna Àrtes och Rahona (2013) genomförde ett experiment med 300 studenter på samma universitetskurs med samma lärare på en kurs i ekonomi för juridikstudenter på Complutense Universitet i Madrid i Spanien. Författarna genomförde ett experiment med en design där varje student både var med i experimentgruppen och i kontrollgruppen genom att på ett examenstillfälle besvara både betygsatta och icke-betygsatta uppgifter. Studenterna fick information om vilka uppgifter som skulle betygsättas innan experimentet startade. Resultatet visade att uppgifter som betygsattes ledde till högre prestationer, i storleksordningen ett betygssteg. Fördelningen var inte jämn utan de resurssvaga studenterna hade större fördel av att få betygssatta uppgifter jämfört med resursstarka elever. Författarna drar slutsatsen att betygssatta uppgifter ledde till bättre prestationer över hela populationen, oavsett förmåga men att det inte går att generalisera dessa resultat bortom den studerade populationen eftersom 1) det var en selekterad grupp universitetsstudenter som var högpresterande; 2) studenterna läste kursen samma termin och därför kan kamrateffekter ha påverkat resultatet; 3) designen av experimentet gjorde att studenterna delades in i en förmiddags- och eftermiddagsgrupp och studenter i eftermiddagsgruppen jobbade oftare deltid och hade lägre intagningspoäng; 4) olika lärare på förmiddags- och eftermiddagskursen hade eventuellt olika karaktäristika (pedagogisk skicklighet: till exempel att kunna förklara problem) vilket kan ha påverkat resultatet. Denna studie hade ett bekvämlighetsurval med studenter på en attraktiv och selekterad universitetsutbildning vilket försvårar möjligheten att dra några generella slutsatser. Det är rimligt att anta att resultaten funna i denna studie kan vara annorlunda för elevgrupper med andra förutsättningar och åldrar.

Utbildningsekonomerna Azmat och Iriberri (2010) undersökte hur elever påverkades av att få betyg och information om ranking i klassen under ett års tid och hur det påverkade deras senare prestationer. Författarna genomförde ekonometriska analyser och hade ekonomiska teorier om relationen mellan lönearbete och belöningar/bestraffningar och relative performance feedback som utgångspunkt. Ett naturligt experiment i Baskien i Spanien gjorde studien möjlig under skolåret 1990-1991. Totalt deltog 1313 elever på en privat skola i åldrarna 14 till 17 år. Fyra årskullar fick extra information om ranking tillsammans med betygen som de fick 4 gånger under läsåret medan åtta årskullar enbart fick betyg som tidigare utan information om ranking. Författarna hade tillgång till longitudinella data och har därför kunnat följa eleverna under längre tid. Resultatet visade att information om ranking var positivt för alla elevers prestationer och ökade elevernas betyg med 5 procent. När informationen togs bort försvann effekten. Författarna menar att de positiva effekterna kan förklaras med att när elever får information om var de befinner sig i klassen blir de motiverade och kommer därför att prestera bättre. Författarna argumenterar för att elever oavsett förmåga och förutsättningar presterar bättre om de får information om ranking vilket är i linje med forskningsresultat inom tävlings- och idrottsfältet och ekonomiska teorier om relationen mellan arbete och belöningar/bestraffningar. Även i denna studie används ett selekterat urval av elever på en privatskola vilket försvårar möjligheten att dra generella slutsatser och att överföra resultaten till andra elevgrupper.

Cillier, Schuwirth, Adendorff, Herman och van der Vleuten (2010) undersökte effekten av *high-stakes* summativa bedömningar på studenters lärande. Totalt intervjuades 18 studenter som läste medicin på ett universitet i Sydafrika. Deltagarna fick ingen belöning för att delta i studien. Intervjuerna var semi-strukturerade och tog ca 90 minuter per student. Författaren fann ett antal faktorer av betydelse för hur studenterna upplevde att de blev påverkade av summativa bedömningar. Studenterna förhöll sig på två olika sätt: hur stor sannolikheten var att en konsekvens av en bedömning inträffar; hur allvarlig denna konsekvens kunde bli; och om ingen konsekvens förväntades, hur det kunde påverka dem. Studenterna anpassade sina arbetsinsatser och strategier för lärande beroende på tidigare erfarenheter av bedömningar och deras

konsekvenser. Författaren menar att när bedömningar innehåller konsekvenser för att påverka lärandet (t.ex. godkänt eller inte godkänt resultat, högre betyg) kan studenter utveckla mönster och strategier för att undvika misslyckande eller maximera sina möjligheter för att lyckas istället för att höja arbetsinsatsen för att utveckla och förändra sitt lärande. Författaren argumenterar för att bedömningar som åtföljs av en konsekvens (positiv eller negativ) innebär att studenternas lärande påverkas. Denna studie använder sig av ett bekvämlighetsurval med ett fåtal högpresterande vuxna studenter.

Klapp, Cliffordson och Gustafsson (2014) undersökte hur betyg i årskurs 6 påverkade 8558 elevernas resultat ett år senare. Mellan 1969 och 1981 kunde kommuner i Sverige själva bestämma om de skulle betygsätta elever i årskurs 6 eller inte. Detta medförde att i dataregistret Utvärdering Genom Uppföljning (UGU) som är ett nationellt representativt urval om 10 procent av en årskull i Sverige, hade 50 procent av eleverna fått betyg i årkurs sex medan 50 procent inte fått betyg. Detta gjorde det möjligt att genomföra en kvasi-experimentell design. Kontroll gjordes för elevernas kognitiva förmåga, kön och socioekonomiska bakgrund. Oberoende *t*-test genomfördes för att undersöka om det fanns initiala skillnader mellan de två grupperna (fått betyg/inte fått betyg) med avseende på kognitiv förmåga, kön och socioekonomisk bakgrund. Ett antal regressionsanalyser genomfördes med missing data modellering och flernivåanalyser för att ta hänsyn till klustringseffekter av data. Interaktionseffekter mellan de olika oberoende variablerna undersöktes.

Resultatet visade att det inte fanns några generella effekter av betyg på senare prestationer men däremot fanns differentierande effekter: låg- till medelpresterande elever (kognitiv förmåga) fick lägre betyg i årkurs 7 om de fått betyg i årskurs 6, jämfört med låg- till medel presterande elever som inte fått betyg i årskurs 6. Ett användbart och etablerat mått är standardiserade medelvärdesdifferenser, där standardiseringen görs med medeltalet av standardavvikelserna inom grupper. Detta mått betecknas som Cohens *d*. Enligt Cohens *d* räknas *d*-värden runt 0,20 som små, *d*-värden runt 0,50 som medelstora och *d*-värden runt 0,80 som stora. Dock skiljer sig dessa gränser för *d*-värden för olika typer av fenomen och data (Durlak 2009, Hattie 2009). En effektstorlek på $d = 0{,}20$ anses vara av betydelse för policyarbete och reformer när det gäller studier inom utbildningsfältet där analyserar görs på data med olika prestationsmått som provresultat och betyg (Durlak 2009). För denna studie är effektstorlekarna ett mått på styrkan på skillnaderna mellan de två grupperna av elever: betygsatta och inte betygsatta. Effektstorleken på $d = 0{,}30$ kan översättas med att eleverna i gruppen som inte fick betyg presterar 0,30 standardavvikelser högre jämfört med eleverna som fick betyg i årskurs 6. Det fanns en tendens att elever som presterade högt på det kognitiva testet fick högre betyg i årskurs 7 om de hade fått betyg i årskurs 6, jämfört med högpresterande elever som inte fått betyg, men skillnaderna mellan grupperna var låga med *d*-värden nära 0. Könsskillnader identifierades och visade att inom gruppen betygsatta elever fick pojkar en avsevärt sämre betygsutveckling jämfört med flickorna med en effektstorlek på $d = 0{,}51$. Inga skillnader med avseende på elevernas socioekonomiska bakgrund identifierades vilket författarna menar beror på att elever oavsett socioekonomisk bakgrund påverkas på liknande sätt av att få betyg.

Författarna menar att betyg har en differentierande effekt där låg- till medelpresterande elever fick en sämre betygsutveckling om de fick betyg i årskurs 6 jämfört med elever som inte fått betyg i årskurs 6. Dessa resultat kopplas till teorier om elevers akademiska självkänsla och hur riskfyllda situationer i skolan påverkar elevernas uppfattning om sin akademiska förmåga och att betyg därför påverkar elever med olika förutsättningar och bakgrunder på olika sätt. En begränsning kan vara att data är från början av 1980-talet och att de inte helt självklart kan generaliseras till dagens betygssystem. Dock är dagens betygssystem av mer *high-stake* karaktär jämfört med det tidigare betygsystemet (risken att få underkänt resultat F) vilket kan innebära att konsekvenser av betygsättning är än allvarligare för eleverna i dagens system.

Klapp (2014) genomförde en longitudinell uppföljningsstudie av Klapp, Cliffordson och Gustafsson (2014) och undersökte hur betyg i årskurs 6 påverkade elevernas prestationer i årskurs 7, 8 och 9 samt om de gått ut gymnasiet eller inte. Totalt deltog 8558 elever i studien. Kontroll gjordes för kognitiv förmåga, kön och socioekonomisk bakgrund. Oberoende *t*-test, *growth models* (regressionsanalyser) och logistiska regressioner med missing data och flernivåanalyser genomfördes. Resultatet visade på signifikanta (signifikans betyder att resultatet inte berodde på slumpen) negativa effekter av betygsättning i årskurs 6 på elevers senare prestationer med effektstorlekar på $d = 0{,}30$, 0,27 och 0,21 för betyg i årskurs 7, 8 och 9. Inga signifikanta positiva resultat för betygsatta högpresterande elevers senare prestationer. Resultatet för de logistiska regressionerna visade att

låg- till medelpresterande elever som fått betyg i årskurs 6 avgick från gymnasiet i lägre grad, jämfört med elever som inte fått betyg i årskurs 6.

Författaren menar att betyg har en differentierande effekt och att låg- till medelpresterande elever påverkas negativt av betyg i årskurs 6 på deras senare prestationer mätt i betyg upp till och med gymnasiet. De positiva effekterna av betygsättning för högpresterande elever är inte signifikanta och därför argumenterar författaren för slutsatsen att betyg inte leder till bättre prestationer utan att betyg har en negativ effekt på resurssvaga och lågpresterande elevers lärande och prestationer. Resultaten verkar vara robusta över tid men för att kunna generalisera till dagens betygsystem bör studien replikeras på aktuella data.

Utbildningsekonomen Sjögren (2010) genomförde en longitudinell studie och jämförde elever som fått betyg i årskurs 6 med elever som inte fått betyg i årskurs 6 och hur det påverkade deras lön i vuxen ålder, deras utbildningslängd, om de gick ut gymnasiet och sannolikheten för att de tog en universitetsexamen. Studien är inte publicerad i en granskad tidskrift men tas med i denna översikt på grund av att studien är svensk och resultaten har använts i relativt stor omfattning för genomförande av policyförändringar. Författaren analyserade skillnader mellan olika utfall (lön, utbildningslängd, avgått gymnasieskolan och universitetsexamen) för elever som blev exponerade för olika betygsättningspraktiker: om de fått betyg eller inte i årskurs 6. Könsskillnader analyserades och analyser bygger på antaganden om att elever som inte gått ut gymnasiet var lågpresterande elever i åk 6 medan elever som tagit en universitetsexamen var högpresterande elever i åk 6. Inga direkta mått på elevers prestationsförmåga användes. Sjögren kontrollerade för olika kommuneffekter.

Sjögren menar att resultatet visade på generella svagt signifikanta negativa effekter för flickor (i några kohorter) när betygen upphörde. Det handlade om att flickornas utbildningslängd blev ca 2 veckor kortare jämfört med de flickor som behöll betygen i årskurs 6. Inga effekter fanns för pojkarna. När föräldrarnas socioekonomiska bakgrund analyserades i relation till betygseffekterna visade resultatet att för lågpresterande elever (både flickor och pojkar) med lågt utbildade föräldrar fanns det en svag negativ effekt av att betygen upphörde. De kraftfullaste resultaten fanns för pojkar med högutbildade föräldrar som missgynnades både med avseende på utbildningslängd och lön i vuxen ålder av att få betyg i årkurs 6. Sjögren argumenterar för att resultaten i hennes studie innebär att betyg främjar akademiskt svaga och socialt utsatta barn.

Med tanke på de svagt signifikanta resultaten och att ingen direkt information om elevers prestationsförmåga fanns med i analyserna är det svårt att dra några slutsatser. Det tydligaste resultatet gäller för pojkar med högutbildade föräldrar och att de missgynnades av att få betyg i årskurs 6, ett resultat som inte lyfts fram så tydligt i rapportens diskussionsavsnitt.

Trotter (2006) undersökte hur studenter på ett ekonomiprogram på ett universitet i Storbritannien upplevde kontinuerliga summativa bedömningar med möjlighet för studenterna att revidera sina inlämnade uppgifter innan betygsättning. Författaren genomförde en litteraturöversikt, en enkät och intervjuer. Totalt deltog 44 studenter som under kursens gång fick genomföra ett flertal prov av examinerande karaktär men där läraren innan betygsättning gav feedback och möjlighet till revidering. Studenterna kunde revidera sitt prov och fick därefter en bedömning av läraren som var betygsgrundande (10 procent av slutbetyget för kursen). Av de 44 studenterna intervjuades sedan 6-8 studenter. I analysen framkom några olika teman utifrån insamlad data. Resultatet visade på tre teman: att kontinuerlig summativ bedömning stärkte studenternas externa motivation; att kontinuerlig bedömning förbättrade studenternas lärande; att studenterna upplevde att kontinuerlig summativ bedömning med möjlighet att revidera sina arbeten innan betygsättning var positivt för deras lärande. Författaren menar att studenterna var positiva till kontinuerlig summativ bedömning med möjlighet att revidera sina arbeten under bedömningsprocessen. Möjligheten till revidering av en summativ bedömning innan betyget sätts innebar att studenternas fokus på uppgifterna och att den externa motivationen stärktes. Även i denna studie är det få deltagare och urvalsförfarandet är inte beskrivet i detalj.

Van der Kleij, Eggen, Timmers och Veldkamp (2012) undersökte effekten av datorbaserad skriftlig feedback på studenters lärande. De utförde ett experiment med 152 studenter på ett ekonomiprogram på ett universitet i Holland. Studenterna delades in i tre grupper (randomiserat urval) där grupp 1 (52 studenter) fick direkt korrigerande feedback med förklaringar, tänkbara lösningar och referenser. Grupp 2 (48 studenter) fick ”försenad” korrigerande feedback med förklaringar, tänkbara lösningar och referenser. Grupp 3 (52 studenter) fick enbart ”försenad” bedömning på korrektheten i svaren (rätt/fel). Författarna använde sig av följande

instrument för att mäta studenternas lärande: en formativ bedömning; en summativ bedömning inom ett ämne; en enkät; och en tids-logg. Studenterna fick den formativa och summativa bedömningarna direkt efter varandra för att minska inflytandet av annat som kunde påverka deras resultat och lärande. Enkäten delades ut direkt efter att de formativa och summativa bedömningarna gjorts. Författarna utförde variansanalyser (ANOVA, ANCOVA) för att jämföra gruppernas medelvärden på instrumenten. Resultatet visade på skillnader mellan de tre grupperna både med avseende på den formativa och den summativa bedömningen men det fanns inte några signifikanta skillnader mellan de tre grupperna och typ av feedback och resultat på den summativa bedömningen. Resultatet på enkäten visade dock att studenterna upplevde att den direkta feedbacken med förklaringar, tänkbara lösningar och referenser var bättre för deras lärande jämfört med rätt/fel bedömningen. Studenterna föredrog den direkta feedbacken framför att få feedback vid senare tillfälle. Sammanfattningsvis visade denna studie inte på några effekter av typ av feedback på studenternas prestationer. Författarna redovisar vissa problem med instrumenten och menar att i efterföljande studier bör instrumenten prövas ut mer grundligt. Även denna studie har deltagare på universitetsnivå vilket kan innebära att resultatet främst gäller för högpresterande vuxna studenter.

### Sammanfattning

Studierna presenterade i detta första tema kommer fram till olika resultat där Ártes och Rahona (2013) menar att de vuxna studenter som deltar i studien påverkades positivt om de fick betygsatta uppgifter vilket kan förklaras med att vuxna studenter på en selekterad utbildning är relativt drivna när det gäller att prestera. Klapp, Cliffordson och Gustafsson (2014) samt Klapp (2014) undersöker elever i 11-12 års ålder och kommer fram till att betyg har en negativ effekt på resurssvaga och lågpresterande elevernas senare prestationer. I Àrtes och Rahonas (2013) studie deltog vuxna studenter på ett selekterat universitetsprogram vilket är en högpresterande grupp och även de studenter som är lågpresterande i den gruppen är högpresterande i relation till den nationella populationen. I Klapp, Cliffordson och Gustafsson (2014) samt Klapp (2014) undersöktes elever som ingår i ett nationellt representativt urval där alla nivåer av förmåga hos eleverna är representerade. Studierna skiljer sig åt med avseende på ålder på deltagarna och elevernas förmåga vilket är väsentliga faktorer att ta i beaktande. I Azmat och Iriberris (2009) studie deltar en selekterad grupp elever på gymnasienivå som går på en privat skola och fann att *Relative Performance Information* hade en kortvarig positiv effekt för alla elevers prestationer oavsett deras förmåga. Det är möjligt att dessa elever motiveras av att de jämförs med varandra vilket kan göra att de anstränger sig mer för att prestera bättre och få en bättre ”placering” vid nästa betygstillfälle. Att de finner positiva effekter för alla elever oavsett förmåga kan bero på att urvalet inte är representativt utifrån den nationella populationen vilket riskerar att snedvrida resultatet. Sjögrens (2010) använder data för en stor population men har indirekt information om elevernas prestationsförmåga vilket kan vara ett osäkert mått om slutsatser ska dras om hur betyg påverkar låg- och högpresterande elever. Resultaten i Sjögrens studie är svaga och det tydligaste resultatet (att pojkar med föräldrar med hög social status påverkas negativt av att få betyg) diskuteras kortfattat och lyfts inte fram tydligt. I Cilliers et al. (2010), Trotters (2006) och van der Kleij et al. (2012) studier var det ett selekterat mindre urval av studenter på ekonomi- och läkarprogram på universitetsnivå som användes. Även i dessa fall är det inte möjligt att dra några långtgående generella slutsatser. Dock är det intressant att dessa högpresterande studenter menade att deras lärande påverkades positivt av tydlig, direkt och kontinuerlig feedback med möjlighet att revidera arbeten innan betygsättningen skedde (Cillier et al. 2010, Trotter, 2006) vilket är i linje med andra resultat redovisad i denna översikt.

## Jämförelser mellan formativ och summativ bedömningspraktik – fokus på summativ bedömning

De texter som studerats i gruppen Jämförelser mellan formativ och summativ bedömningspraktik handlar om förändringar i bedömningspraktiker och dess betydelse för elevers lärande och prestationer. Dessa förändringar innebär att studierna utvärderar både formativ och summativ bedömningspraktik gentemot elevernas lärande och prestationer.

Abu-Hamour och Mattar (2013) undersökte effekterna av formativ och summativ bedömning på prestationer i matematik. Ca 70 elever i åldern 8-9 år som gick på en privatskola i de centrala delarna av Jordanien deltog. Eleverna valdes ut från en större grupp av elever utifrån att de hade kognitiv förmåga på medelnivå, pratade arabiska som modersmål och inte hade några beteende- och känslomässiga störningar. De 70 eleverna gick i två klasser med vardera 35 elever. De två klasserna användes som experiment- respektive kontrollklass. Ett förprov genomfördes vilket inte visade på några signifikanta skillnader mellan experiment- och kontrollklassen. Eleverna i experimentklassen fick både summativ bedömning och bedömning under processens gång genom curriculum-based-measurements (CBM), som innebär en form av formativ bedömning som sker regelbundet. Inom matematik (M-CBM) använde läraren prov där eleverna fick regelbunden feedback och information om sina svårigheter och vad de kunde göra för att förbättra sig. Läraren fick genom dessa regelbundna prov information om elevernas svårigheter och kunde därför anpassa undervisningen i matematik till elevernas behov. Kontrollklassen fick enbart summativ bedömning. Författarnas hypotes var att elever i experimentklassen som fick M-CBM skulle få högre resultat i ett examinerande prov i matematik jämfört med kontrollklassen som inte fått M-CBM.

Resultatet visade på signifikanta effekter för eleverna i experimentgruppen som fick M-CBM jämfört med elever i kontrollklassen. Effektstorlekar på $r = 0{,}38$ i skillnad mellan klasserna för resultat på det examinerande matematikprovet, till fördel för eleverna i experimentgruppen. Författarna menar att om eleverna får feedback under lärandeprocessen kommer såväl deras lärande att öka som deras motivation att ta sig an utmanande matematikproblem att öka. Författarna menar att det inte är proven som används inom M-CBM i sig som påverkar elevernas lärande positivt utan det är informationen proven bidrar med som gör det möjligt för läraren att anpassa undervisningen till elevernas lärandeprocess som är viktiga för det positiva resultatet. Denna studie genomfördes på en privat skola med ett urval som närmast kan liknas med ett bekvämlighetsurval. Detta gör det svårt att dra slutsatser och författarna själva menar att liknande undersökning bör göras inom den kommunala skolan. Urvalet är lågt och det är endast två klasser som undersökts med samma matematiklärare för båda klasserna.

Bagley (2008) undersökte hur en detaljerad skriftlig feedback istället för betyg, mottogs av elever på en privat skola i USA. Skolan var åldersintegrerad och beskriver sig själv som progressiv. Totalt 115 elever i åldrarna 15 till 18 år besvarade en enkät om hur de upplevde bedömningspraktiken på skolan. Författaren intervjuade sedan 26 elever, sex lärare och gjorde observationer av undervisningen. Studien var longitudinell och varade över två år.

Resultatet av enkäten visade att 59 procent av eleverna gillade att få detaljerad feedback på sitt arbete jämfört med att få betyg, 46 procent av elever tyckte att den detaljerade feedbackens fokus på hur de kunde förbättra sitt arbete var till hjälp för deras lärande, jämfört med att få betyg. Resultatet från intervjuerna visade att eleverna tyckte att detaljerad skriftlig feedback var både positivt och negativt jämfört med att få betyg. De tyckte att deras syn på bedömning förändrats över de två åren. De menade att lärares detaljerade feedback kunde vara alltför detaljerad; att lärares bedömning är subjektiv; och att revideringar av arbetet utifrån den detaljerade feedbacken är alltför utmanande och överväldigande. Vissa elever menade att den detaljerade feedbacken (istället för betyg) ledde till både bra och dålig stress och exempel på positiv stress var att det var möjligt för dem att förbättra sitt arbete och lära sig mer.

Intervjuerna med lärarna visade att den detaljerade feedbacken ledde till en förbättrad relation till eleverna jämfört med att ge betyg. Den detaljerade feedbacken var en väg för lärarna att få bättre kontakt och kännedom om eleverna och deras lärande.

Författaren menar att den största svårigheten med att införa detaljerad skriftlig feedback istället för summativ bedömning är tidsåtgången för lärare att utföra detta med tanke på antal elever per lärare i de flest skolor i USA.

Utbildningsekonomerna Betts och Grogger (2003) undersökte hur högre kunskapskrav för betygen (betygskrav) påverkade elevers prestationer, om de gick ut skolan och lön i vuxen ålder. De använde data från High School and Beyond undersökningen i USA. De menar att skolor med högre betygskrav är mer stringenta i relation till ett objektivt prestationsmått (prov) medan skolor med lägre betygskrav är mindre stringenta i relation till samma mått. Författarna använder elevernas betygsgenomsnitt och deras resultat på provet för att få fram nivån på betygskraven på skolorna. I vissa analyser kontrollerar de för kön, familjestruktur, familjestorlek,

föräldrars utbildningsnivå och inkomst, demografi för boende. Utan kontroll för tidigare prestationer (provresultat i årskurs 10 i USA) och lön visade resultatet en positiv signifikant effekt av högre betygskrav på utbildningsresultat men när kön och tidigare prestationer togs med i analyserna minskar de positiva effekterna av högre betygskrav avsevärt. Minskningen skedde över hela distributionen men en mer positiv effekt av högre betygskrav kvarstod för den övre percentilen (75:e). När det gällde om eleverna har avslutat skolan finner författarna inga effekter av högre betygskrav för hela populationen. Däremot fann författarna att olika etniska grupper av elever påverkades olika av högre betygskrav. De fann negativa effekter av högre betygskrav för Afro-amerikanska och spansktalande (Hispanics) elever när det gällde att avsluta skolan och författarna drar slutsatsen att högre betygskrav riskerar att få minoritetselever att avsluta skolan i förtid (drop-outs). Författarna sammanfattar sina resultat med att högre betygskrav differentierar och menar att resultaten är i linje med en relativ prestations hypotes (*relative performance hypothesis*) som innebär att elever bedömer sina prestationer i relation till klasskamrater och inte i relation till absoluta betygskrav.

Carrillo-de-la-Peña, Baillès, Caseras, Martínez, Ortet och Pérez (2009) undersökte hur studenters lärande påverkades av formativ bedömning jämfört med summativ bedömning. Totalt 548 studenter som läste medicin, psykologi och biologi på fyra universitet i Spanien deltog i studien. Studenterna erbjöds mitt i kursen en frivillig formativ bedömning mellan de ”business-as-usual” summativa bedömningarna. Den formativa bedömningen bestod av ett prov som studenterna fick omedelbar feedback på. De fick information om de korrekta svaren direkt efter provet. Dagen efter fick studenterna diskutera de vanligaste felen de gjort på provet. Alla studenter gjorde det examinerande provet i slutet av kursen inom respektive ämne/disciplin. Deltagandet i den formativa bedömningen var frivilligt men de som deltog i den och som fick ett resultat på minst 5 av 10 på det formativa delprovet fick en liten ökning i slutbetyget för kursen. Författarna jämförde resultaten genom medelvärdesskillnader (*t*-test) på det examinerande provet för studenter som deltagit i den formativa bedömningen med studenter som inte deltagit. De jämförde även resultat på det examinerande provet för studenter som var framgångsrika i den formativa bedömningen med studenter som misslyckades i den formativa bedömningen. Resultatet visade att studenter som deltog i den formativa bedömningen fick signifikant högre betyg i det examinerande provet. En högre andel studenter som deltog i den formativa bedömningen godkändes på det examinerande provet, jämfört med studenter som inte deltog i den formativa bedömningen. Studenter som fick högre resultat på den formativa bedömningen fick även högre betyg på det examinerande provet. Författarna drar slutsatsen att studenter som deltog i den formativa bedömningen presterade bättre på det examinerande summativa provet jämfört med studenter som inte deltog. Misslyckande på den formativa bedömningen predicerade dock inte resultatet på det examinerande provet.

Clymer och Wiliam (2007) intervjuade 19 elever i årskurs åtta i en pilotstudie om hur en ny formativ bedömningspraktik till skillnad mot den tidigare summativa bedömningspraktiken påverkade deras lärande och förhållningssätt till lärande och prestationer. Det nya formativa bedömningssystemet innebar att eleverna fick möjlighet att visa vad de kunde ända fram till att det slutliga betyget sattes. Ett karaktärsdrag för detta system var att läraren dokumenterade elevernas lärande och prestationer på ett sätt som var tydligt för eleven genom regelbundna veckorapporter. Ett annat grundläggande karaktärsdrag var att lärare gav individuell feedback till eleverna på deras arbete i syftet att coacha eleverna till bättre prestationer.

Resultatet av intervjuerna visade att eleverna hade anammat ett förhållningssätt till lärande som var mer inriktat mot ett förhållningssätt där målet för lärandet är en djupare förståelse av innehållet (*mastery*) i undervisningen till skillnad från att eleverna tidigare har haft ett mer prestationsinriktat förhållningssätt (*performance*) där betygen var i fokus. Eleverna menade att de förstod bättre, fokuserade mer på att lära sig viktiga koncept och kände sig mer avslappnade eftersom läraren fokuserade på deras arbete utifrån deras förståelse av det. Atmosfären i klassrummet ändrades på ett sådant sätt att eleverna blev mer engagerade i sitt eget lärande och deltog mer aktivt i undervisningen för att få feedback, både från läraren och från klasskamrater, på hur de kunde förbättra sitt arbete och öka sin förståelse. Eleverna menade att det var en stor fördel att kunna revidera sina arbeten ända fram till att det slutliga betyget sattes till skillnad mot det tidigare systemet där poäng på olika moment under terminens gång räknades samman till ett genomsnitt. Att samla poäng ledde till ett fokus på betyg istället för på lärande, menade eleverna.

Författarna undersökte elevernas resultat på de slutliga examensproven och fann att de fick högre poäng jämfört med det nationella genomsnittet och menar att resultatet är i linje med andra studier som har undersökt

effekten av bedömning för lärande. De fann att det nya formativa bedömningssystemet främst var positivt för låg- och högpresterande elever. Artikeln beskriver en kvalitativ studie där några elever har berättat hur de upplevde ett nytt bedömningssystem. Resultatet är intressant och visar på hur dessa elever påverkades av en förändring från summativt bedömningssystem till ett formativt bedömningssystem. Studiens beskrivande metod är begränsad för att kunna dra slutsatser och generalisera resultatet till andra kontexter men resultatet är i linje med andra liknande studier med fler deltagare och andra analysmetoder.

Utbildningsekonomerna Figlio och Lucas (2004) undersökte hur lärares högre kunskapskrav för betygen (betygskrav) påverkade elevernas resultat på standardiserade prov i matematik och läsning Florida Comprehensive Assessment Test (FCAT) och Iowa Test of Basic Skills (ITBS) i Florida, USA. Lärares betygskrav mättes av den genomsnittliga samstämmigheten mellan ett standardiserat prov och de betyg som lärarna satte, ett slags likvärdighetsmått. De använde longitudinella data för provresultat och betyg för elever i årskurs tre till fem över fyra år (1995-96 till 1998-99). Som beroende variabel användes ITBS-provet som är ett nationellt prov på kunskaper och lärande inom olika ämnesdomäner som eleverna tar varje år. Författarna analyserar förändring på proven för eleverna över fyra år. Författarna kontrollerar för etnicitet, kön, grattis lunch (indikator på socioekonomisk bakgrund), talang och funktionsnedsättning. Kontroll för lärarkaraktäristika som utbildningsnivå, arbetserfarenhet och lärosäte gjordes. Resultaten visade på positiva effekter av lärares högre betygskrav på elevernas prestationer på de standardiserade proven, framför allt för högpresterande elever. Författarna har några olika förslag till förklaringar till resultaten: 1) att eleverna blir mer motiverade av högre betygskrav och därför presterar bättre på proven; 2) att föräldrar till barn som har lärare med högre betygskrav engagerar sig mer i barnets läxor vilket gör att eleven presterar bättre. Precis som Betts och Grogger (2003) finner författarna att lärares betygskrav differentierar, det vill säga påverkar elever med olika förutsättningar på olika sätt. De menar att lärares högre betygskrav leder till att elever anstränger sig och lär sig mer. Resultaten av regressionsanalyserna visade att lärares utbildningsnivå påverkade resultatet på så sätt att lärare med högre betygskrav i större utsträckning hade en Masterexamen jämfört med lärare som hade lägre betygskrav. Författarna refererar till ekonomen Hanusheks (1986) forskningsresultat och menar att lärares utbildningsnivå inte har betydelse för elevernas prestationer i skolan. Författarna nämner i konklusionerna att det finns en risk att resultaten i studien beror på lärarkaraktäristika som de inte tagit med i analyserna.

Meyer, Wijekumar, Middlemiss, Higley, Lei, Meier och Spielvogel (2010) undersökte effekter av olika typer av feedback på elevers läsförståelse i årskurs 5 och 7 i USA. Totalt deltog 56 elever i årskurs fem och 55 elever i årskurs sju i studien. Ett stratifierat randomiserat urval användes med för- och efterprov. Interventionen bestod av att eleverna arbetade med uppgifter inom läsförståelse och som de fick feedback och handledning på via ITSS. En experimentgrupp av elever fick formativ (elaborated) feedback via en webbaserad handledning (ITSS) och en kontrollgrupp av elever fick business-as-usual summativ feedback. Eleverna i experimentgruppen fick elaborated feedback under 90 minuter i veckan under sex månader. Efterprovet i form av ett standardiserat läsförståelseprov (GSRT) genomfördes av eleverna dels direkt efter interventionen, dels sex månader efter avslutad intervention. Resultatet visade på stora skillnader mellan grupperna, speciellt för de lågpresterande eleverna på provet i läsförståelse. Eleverna i båda grupperna förbättrade sin läsförmåga över tid oavsett typ av feedback. Elever som deltog i interventionen förbättrade dock sin läsförståelse i högre grad med en effektstorlek på $d = 0{,}55$ mellan förprov och efterprov medan elever som fick summativ bedömning förbättrade sin läsförståelse i mindre grad, $d = 0{,}15$. För elever med dålig läsförmåga fick de ännu större förbättringar mellan för- och efterprov, $d = 0{,}73$ medan elever med bra läsförmåga inte förbättrade sin läsförmåga i samma grad, $d = 0{,}27$. Författarna menar att resultatet är i linje med annan forskning (Hattie & Timperley 2007, Shute 2007) som visar att typ av feedback påverkar elevers lärande och prestationer. Feedback som fokuserar på hur eleverna kan förbättra sina fel och som ger klara direktiv för hur det kan ske ökar elevernas lärande jämfört med feedback som enbart fokuserar på rätt och fel.

Ross (2005) undersökte effekterna av formativ och summativ bedömning för 2215 japanska universitetsstudenter (18-20 år) på ett universitetsprogram inom stadsplanering och internationell utveckling. Studenterna var antagna till ett privat, selekterande universitet. Studien var longitudinell och sträckte sig över åtta år där fyra kohorter (de första fyra åren) hade summativ bedömningspraktik med traditionella slutprov medan de efterföljande fyra kohorterna hade en formativ bedömningspraktik utan summativa bedömningar. För dessa studenter omräknades resultat på de formativa bedömningarna till ett betygsgenomsnitt. Alla studenter

genomförde tre prov som mätte deras nivå på engelska (English for Academic Purposes, EAP). Författaren använde tre olika metoder: dokumentanalys; tillväxtmodeller; och jämförelser av de två gruppernas medelpoänger på EAP-proven. Studenternas GPA och EAP-proven användes som utfallsvariabler.

Resultatet visade att den formativa bedömningspraktiken ledde till positiva effekter på studenternas utveckling av språkkunskaper. Den övergripande bilden visade att den främsta positiva effekten fanns på studenternas förhållningssätt som påverkar deras uppmärksamhet och deltagande i aktiviteter vilket ledde till bättre hörförståelse. Författaren menar att studenternas mer aktiva deltagande påverkar hur bedömning av språkkunskaper kan definieras. De positiva effekterna av en formativ bedömningspraktik var dock inte tydlig för studenternas läsförmåga och författaren menar att fler analyser krävs. Denna studie har ett stort antal deltagare på universitetsnivå som gick på ett privat, selektivt universitet.

### Sammanfattning

Gemensamt för studierna är att implementeringen av ett nytt bedömningssystem (på lokal nivå) innebar en möjlighet att undersöka skillnader mellan ett summativt och formativt bedömningssystem. I stort sett alla författare till studierna i denna grupp visar resultat där den summativa bedömningspraktiken ledde till sämre lärande, sämre motivation för lärande och prestationer medan formativ bedömning ledde till positiva resultat (Abu-Hamour & Mattar 2013, Carrillo-de-la-Peña et al. 2009, Clymer & Wiliam 2007, Meyer et al. 2010, Ross 2005). Vissa resultat visar på positiva effekter på studenternas förhållningssätt till sitt lärande när den summativa bedömningspraktiken försvann (Clymer & Wiliam 2007, Ross 2005). Abu-Hamour och Mattar (2013) visade att det var positivt för 8-9-åringar att få kontinuerlig formativ feedback framför allt för att undervisningen då anpassades bättre till elevernas behov. I Bagleys (2008) studie menar författaren att den ökade stressen som studenterna upplevde med den formativa bedömningen berodde på utmanande och detaljerad feedback som de var ovana vid. Den ökade arbetsbelastningen för lärarna med användandet av formativ bedömning istället för summativ bedömning lönade sig eftersom det påverkade studenternas prestationer positivt i längden. Betts och Grogger (2003) fann att högre betygskrav differentierar: high-school minoritetselever påverkas negativt av högre betygskrav medan Figlio och Lucas (2004) fann att högre betygskrav var positivt framför allt för högpresterande high-school elever dock fanns det effekter av att mer högutbildade lärare hade mer positiva resultat av högre betygskrav vilket kan bero på att viktiga faktorer som betydelse av lärarnas kompetens har utelämnats i analysen. I Carrillo-de-la-Peña et al. (2009) fann att de 548 studenter på några olika universitetsutbildningar fick högre resultat på ett examinerande prov om de hade deltagit i en formativ bedömningsprocess jämfört med studenter som enbart fått summativ bedömning. Både Meyers et al. (2010) och Ross (2005) fann positiva effekter av formativ bedömning på språkkunskaper jämfört med summativ bedömning.

## Effekter av positiv och negativ feedback på lärande, motivation för lärande och prestationer

De texter som studerats i gruppen Effekter av positiv och negativ feedback på lärande, motivation för lärande och prestationer handlar om hur olika typ av information i summativa bedömningar påverkar elevernas lärande, motivation för lärande och prestationer.

Bies-Hernandez (2012) undersökte hur negativa och positiva bedömningar påverkade studenter på ett psykologprogram på ett universitet i USA. Två experiment genomfördes. I det första experimentet undersöktes 76 studenter med en genomsnittsålder på 21 år. De fick kompensation i form av *credit points* (poäng som ligger till grund för slutligt betyg i kursen) för att delta i studien. Studenterna fick ta ställning till fiktiva exempel på olika betygssystem där ett betygssystem innebar att studenterna börjar med 100 poäng vid kursens start och sedan dras poäng av under kursens gång beroende på studenternas prestationer. Det andra betygssystemet som studenterna skulle ta ställning till innebar att studenterna börjar kursen med 0 poäng och sedan får poäng för sina prestationer under kursens gång. Studenterna fick besvara frågor kring de två betygssystemen, som till exempel ”jag skulle prestera bra med detta betygssystem” på en skala 1-7 (man använde en s.k. Likertskala). Variansanalys genomfördes (medelvärdesskillnader) för de olika frågorna vilken visade att studenterna menade

att de skulle vara mer motiverade, få högre betyg och att kursen skulle vara mer tillfredsställande och lättare för dem med ett betygssystem där poäng adderades jämfört med om poäng drogs bort. Författaren menar att betygssystemet där studenterna förlorar poäng kan påverka studenternas förhållningssätt till kursen på ett negativt sätt.

I det andra experimentet undersökte författaren om dessa två betygssystem påverkar studenternas lärande. Tre lärare som undervisade i två olika moment på en introduktionskurs i psykologi deltog. Varje lärare hade två grupper av studenter i varje moment av kursen. Kursen pågick under en termin och 181 studenter deltog. Varje lärare applicerade ett av de två betygssystemen i de två olika grupperna. Lärarna använde samma material i de båda grupperna men applicerade ett betygsystem där studenterna förlorar poäng i den ena gruppen och ett betygssystem där poäng adderas i den andra gruppen. Studenternas examensbetyg för kursen användes som utfallsvariabel i variansanalyserna (ANOVA). Resultatet visade att studenterna i gruppen som förlorade poäng under kursens gång hade sämre resultat på examensbetygen på kursen. Inga samband (interaktioner) mellan betygssystemen och lärarna fanns. Även denna studie använder ett selekterat urval av studenter på universitetsnivå vilket gör det svårt att generalisera resultaten till andra grupper av elever. Denna studie undersöker hur förlust av poäng i en summativ bedömning påverkar studenternas prestationer, motivation och lärande och handlar om hur negativ och positiv information/feedback påverkar studenternas förhållningssätt till lärande. Det som är intressant i resultaten från denna studie är att även en selekterad högpresterande grupp av studenters prestationer påverkades negativt av negativ feedback. Att negativ feedback skulle leda till ökad motivation och ökade prestationer stöds inte av dessa resultat.

Dlaska och Krekeler (2013) undersökte hur olika typer av feedback påverkade vuxna studenters lärande inom språk (morfologi och syntakter). Total deltog 226 studenter på ett introduktionsår på ett universitet i Tyskland. Introduktionsåret innebar intensiva studier i det tyska språket och i en mängd andra ämnen. Studenterna kom från 12 olika undervisningsgrupper och medelåldern var 22 år. Deltagarnas språkkunskaper i tyska var tillräckligt bra (över medel på ett språktest alla studenter skriver för att studera i Tyskland). Studenterna delades in i tre grupper som fick skriva tre texter vilka de sedan fick olika feedback på: 1) feedback; 2) feedback och betyg 3) betyg.

Oberoende *t*-test visade att det inte fanns någon skillnad mellan de tre grupperna med avseende på morfologi medan det fanns skillnader mellan grupperna med avseende på syntaktisk förmåga; eleverna i feedback gruppen hade lägre syntaktisk förmåga. Resultat från analyserna visade på generella förbättringar med avseende på morfologi och syntaktisk kompetens mellan utkast av texter och de reviderade texterna (en direkt effekt av feedback på revidering). Dock varken förbättrades eller försämrades studenternas morfologiska och syntaktiska kompetens i de olika grupperna (feedback; feedback och betyg; betyg).

Författarna menar att det inte fanns några effekter på studenternas språkutveckling (morfologi och syntaktisk kompetens) om feedback gavs med eller utan betyg. De fann inga signifikanta skillnader mellan de tre grupperna med avseende på betydelsen av feedback och de menar att studenterna drog fördel av att få feedback oavsett om de fick betyg eller inte. För studenterna i denna studie var betygen på texterna en del av deras slutbetyg för kursen (av *high-stakes* betydelse för studenterna) vilket kan ha inneburit att studenterna var motiverade att lära sig oavsett om de fick betyg eller inte. Författarna menar att resultatet är i kontrast till Ruth Butlers ofta citerade studie från 1988 vilket kan bero på flera saker: att studenterna i Dlaska och Krekelers studie är äldre och därför mer drivna av yttre målorientering; att studenterna i studien ser sig själva som högpresterande vilket påverkar deras förhållningssätt och motivation oavsett typ av feedback. Författarna menar att feedback är central för lärandet men att för vuxna studenter verkar inte betygen underminera effekten av feedback så som Butler tidigare visat för mellanstadieelever (1988).

Docan (2006) undersökte hur studenter på ett universitet i USA påverkades av att starta med maximala poäng på en kurs och få avdrag under kursens gång eller börja med noll poäng och erhålla poäng under kursens gång. Tre lärare undervisade en klass var med adderade poäng och en klass var med avdragna poäng. Undervisningen i övrigt var lika mellan grupperna. Författaren kontrollerade för skillnader innan interventionen startade och fann inga skillnader mellan lärarnas undervisning. Betygssystemet (erhålla eller förlora poäng) var den oberoende variabeln medan en enkät (*The Student Motivation Scale*) som mätte studenternas motivation användes som beroende variabel (utfall) i den kvantitativa analysen. Författaren inkluderade en öppen fråga i enkäten vilken redovisades som ett kvalitativt resultat. Hypotesen var att de studenter som fick negativa

incitament (poäng drogs av) skulle få ökad motivation medan de studenter som fick positiva incitament (poäng adderades) skulle få försämrad motivation. Oberoende *t*-test genomfördes vilket marginellt stödde hypotesen. Den kvalitativa analysen visade att hälften av studenter tyckte att det betygssystem de hade var bra. De studenter som fick avdrag på sina poäng tyckte systemet var obekant viket kan ha gjort att de var mer observanta på sina poäng, menar författaren. De flesta studenterna i båda betygssystemen tyckte att betyg är stressande. Väldigt få studenter menade att de fick positiva känslor när de tänkte på betyg. Studenter inom båda betygssystemen tyckte att det var en styrka att få information om deras utveckling. Författaren menar att betyg i de båda systemen stressar studenterna vilket kan leda till att studenterna utvecklar en yttre motivation för lärande. Studien har få deltagare och de är studenter på universitetsnivå. Studenterna visste om att de deltog i ett ”experiment” med fokus på betyg vilket kan ha påverkat resultatet.

Elliot, Shell, Henry och Maier (2005) undersökte effekter av elevers motivation på deras prestationer i skolan. Tre experiment genomfördes. Det första experimentet undersökte direkta effekter av motivation eller prestationsmål (*performance-approach*, *performance-avoidance*, *mastery*) på elevernas prestationer på ett matematikprov. Totalt deltog 101 elever som var 17 år gamla och som gick på motsvarande gymnasienivå på en kommunal skola i Tyskland. Oberoende variabler var typ av motivation/prestationsmål (*performance-approach*, *performance-avoidance*, *mastery*) medan kön och betygsmedelvärde var kovariater. Resultatet på matematikprovet var beroendevariabel. Eleverna delades in i tre grupper: 1) *performance-approach*; 2) *performance-avoidance*; 3) *mastery*. Eleverna fick under en ordinarie lektion genomföra ett matematikprov och de tre grupperna fick innan provet olika information angående hur deras prestationer på provet skulle bedömas.

Eleverna i *performance-approach* och *mastery* grupperna fick information om att provet gav dem möjlighet att visa att de var exceptionellt bra problemlösare medan eleverna i *performance-avoidance* gruppen fick information om att forskning visat att elever löser problem på liknande sätt men att vissa elever utmärker sig på grund av deras dåliga förmåga att lösa problem. Matematikprovet kan därför hjälpa eleverna att visa att de inte är dåliga på problemlösning. Alla eleverna i de tre grupperna fick information om att de skulle få personlig feedback efter provet: eleverna i *performance-appproach* gruppen fick information om att de skulle få feedback om hur väl eleven presterat jämfört med de andra eleverna, i *performance-avoidance* gruppen fick eleverna information om att de skulle få feedback om hur dåligt eleven presterat jämfört med de andra eleverna och i *mastery*-gruppen fick eleverna information om att de skulle få feedback om eleven lärt sig att lösa problem på ett bra sätt.

Oberoende *t*-test (medelvärdesskillnader) genomfördes som visade att elevernas motivation/prestationsmål var i linje med den information de fick i experimentsituationen (*performance-approach*, *performance-avoidance* och *mastery*). Analyser av variansen mellan grupperna av elever visade på en signifikant direkt effekt av typ av information de fick i experimentsituationen på resultatet på matematikprovet. Eleverna i gruppen som fick information utifrån ett *performance-avoidance* förhållningssätt (negativ information) presterade sämre på matematikprovet jämfört med de två andra grupperna (*performance-approach* och *mastery*). Könsskillnaderna existerade till pojkarnas fördel.

Det andra experimentet genomfördes på liknande sätt som det första experimentet. Totalt deltog 36 elever på gymnasienivå som var 17 år gamla. Skillnaden mellan det första och andra experiment var uppgiften som eleverna skulle lösa. I det andra experimentet var det ett lexikalt språkprov. Oberoende *t*-test visade att elevers prestationsmål var i linje med den information studenterna fick i experimentsituationen. Variansanalyserna visade en signifikant direkt effekt av experimentsituationen på resultatet på språkprovet: eleverna i *performance-avoidance* gruppen presterade signifikant lägre jämfört med elever i de två andra grupperna.

Det tredje experimentet replikerade de två första experimenten men med en annan provuppgift, en problemlösningsuppgift (scrabbles) och med en manipulation. Totalt deltog 61 amerikanska studenter på collegenivå på en introduktionskurs i psykologi. De deltog frivilligt men fick *credit points* som påverkade deras slutbetyg på kursen. Manipulationen innebar att vissa elever fick information om att de var tvungna att klara ett visst antal uppgifter för att de skulle kunna få fler *credit points* (villkor). Resultatet av oberoende *t*-test visade att prestationsmål korresponderade med villkoren i experimentsituation och för manipulationen. Författarna menar att det finns 1) en signifikant effekt för studenternas prestationsmål på prestationer över de villkorade situationerna där elever i gruppen *performance-avoidance* presterade lägre på ”scrabbles”; 2) en signifikant interaktionseffekt för typ av prestationsmål och villkorade situation; 3) elever i *performance-approach* gruppen

presterade signifikant bättre med en villkorad situation (extra *credit points*), jämfört med i en icke villkorad situation, 4) elever i *performance-avoidance* gruppen presterade signifikant sämre i den villkorade situationen jämfört med studenter i den icke villkorade situationen.

Författarna menar att resultaten utgör kraftfulla bevis för att studenter med *performance-avoidance goals* får sämre prestationer jämfört med de andra två grupperna. Författarna menar att tidigare forskning har antagit att elever med *mastery goals* presterar bättre och att elever med olika typer av performance goals (*performance-approach* och *performance-avoidance*) har når sämre prestationer. Resultaten från dessa experiment visar att prestationsmål inte har en generell negativ effekt på prestationer utan det är en viss typ av prestationsmål hos eleven som påverkar prestationerna negativt: *performance-avoidance goals*. Den övergripande slutsatsen från denna studie är att betyg påverkar elevernas prestationer negativt eftersom betyg gör att eleverna utvecklar yttre motivation som leder till ytliga lärandestrategier. Dessa experiment genomfördes med gymnasieelever och studenter på högskolenivå men eftersom det var yrkeselever med olika sociala bakgrunder och förutsättningar är resultaten mer generaliserbara än om det varit ett selekterat urval som användes. Resultaten är i linje med andra studiers resultat presenterade i denna översikt som visar att betyg differentierar.

Pulfrey, Buchs och Butera (2011) undersökte hur betyg påverkade studenters motivation. Författarna använde sig av motivationsteori som visar att människor utvecklar olika typer av mål i olika situationer till exempel vid bedömning. *Performance-approach goals* och *mastery goals* innebär att studenter utvecklar strategier som kan leda till ett mer fördjupat lärande vilket i sin tur kan leda till bättre prestationer medan *performance-avoidance goals* innebär att studenter utvecklar strategier för att undvika misslyckande vilket kan leda till negativa konsekvenser för lärande och prestationer. Utifrån denna teori undersökte författarna följande två hypoteser: 1) om studenter förväntar sig att bli betygsatta kommer *performance-avoidance goals* att öka jämfört med studenter som inte blir betygsatta; 2) relationen mellan förväntningen av att få betyg eller inte och *performance-avoidance goals* hos studenterna kan förklaras av lägre nivåer av autonomi för studenterna i betygsatta situationer.

Tre experiment genomfördes. I alla experimenten deltog studenter på olika yrkesprogram på en Schweizisk yrkesskola. I det första experimentet deltog 115 studenter som var 20 år gamla. Experimentet genomfördes under lektioner där studenterna läste engelska för invandrare (EFL) och klasserna som fick olika behandling under experimentet valdes ut på ett randomiserat sätt. Studenterna fick innan en lektion i hörförståelse fylla i en enkät med frågor om deras motivation. Enkäten bestod av frågor som till exempel ”*It is better for me to do better than others*” där studenterna kunde svara från 1 till 7 där 1 innebar att studenten inte alls höll med utsagan. I hälften av klasserna fick studenterna reda på vid lektionsstart att deras hörförståelseuppgift skulle betygsättas medan i de andra klasserna fick studenterna samma hörförståelseuppgift, men de fick information om att inga betyg skulle ges på denna uppgift. Författarna undersökte skillnaderna mellan klasserna genom medelvärdesanalyser. Resultatet visade att studenter som förväntade sig att få betyg på uppgiften hade högre nivåer av *performance-avoidance goals* jämfört med de studenterna som inte förväntade sig att få betyg.

Experiment 2 innebar ett liknande scenario som det första experiment men med tillägget att studenternas uppfattning om sina kompetenser också undersöktes. Totalt 122 studenter som var 17 år gamla deltog. Skillnaderna mellan det första och andra experimentet var att i det andra experimentet fick studenterna information om betygsättning en vecka i förväg, de fick besvara en enkät om deras uppfattning om deras kunskapsnivå, samt att betygsättningen delades in i tre tillvägagångssätt, studenterna fick antingen enbart betyg, betyg och kommentarer, eller enbart kommentarer. Resultatet visade att studenter som förväntade sig att få betyg på uppgiften utvecklade mer *performance-avoidance goals* jämfört med studenter som inte förväntade sig att få betyg. Bedömning som enbart utgjordes av kommentarer producerade lägst nivåer av *performance-avoidance goals* jämfört med bedömningar med enbart betyg och betyg och kommentarer.

I det tredje experimentet deltog 96 studenter som var 19 år gamla. Studenterna delades in i tre grupper där de blev bedömda utifrån följande tillvägagångssätt: enbart betyg; betyg och kommentarer; enbart kommentarer. Analyserna visade liknande resultat som för de två tidigare experimenten. De medierande modellerna visade att studenterna hade erfarit minskad autonomi i de betygsatta situationerna jämfört med de icke betygsatta situationerna, vilket kan förklara varför studenterna utvecklade en högre nivå av *performance-avoidance goals* i de betygsatta situationerna. Även om studiens design var tre olika experiment med ett randomiserat urval för de tre experimentgrupperna var deltagarna studenter på universitetsnivå, ett selekterat urval. I linje med andra

resultat redovisade i denna fördjupade analys verkar även högpresterande studenter prestera bättre om de fått feedback av formativ karaktär, till exempel i form av kommentarer.

Vaden-Goad (2009) undersökte hur frekvensen av betyg och poäng i en matematikkurs på universitetsnivå påverkade studenters prestationer. Författaren genomförde en longitudinell intervention där studenter i en grupp bedömdes med hög frekvens genom att de genomförde ”*quizzes*” varannan vecka, prov var sjätte vecka och ett slutligt större prov. Utöver en hög frekvens av summativa bedömningar kunde studenterna själva välja att byta ut resultatet på en *quizz* mot resultatet på ett senare prov och ett resultat på ett prov kunde bytas ut till resultatet på ett delprov i det slutliga examensprovet, studenterna kunde alltså styra vilka resultat som skulle ingå i betygsgenomsnittet för kursen genom att byta ut tidigare provresultat mot senare. I den låg-frekventa gruppen fick studenterna tre prov och ett avslutande större prov som räknades samman till ett betygsgenomsnitt för kursen. Totalt deltog 1447 studenter på en introduktionskurs i matematik på ett statligt universitet i USA. Totalt 90 procent av studenterna var mellan 18 och 24 år gamla.

Prestationer mättes med elevernas resultat på kursen, mellan 0-100 poäng som senare omvandlas till betyg. För låg-frekvensgruppen räknades de tre proven och det slutliga större provet samman till ett kursbetyg medan för hög-frekvensgruppen räknades de fem *quizzen*, två prov och det större slutprovet samman till ett kursbetyg. Författaren genomförde en variansanalys som visade att frekvensen av bedömningar inte påverkade studenternas prestationer medan däremot möjligheten att byta ut *quizzes* mot senare prov och prov mot delar av det slutliga större provet. Studenterna som hade möjlighet att byta ut bedömningar fick i genomsnitt ett betygssteg högre betyg jämfört med studenter som inte kunde byta ut sina bedömningar. Författaren redovisar en effektstorlek på 10.05 men redovisar inte vilken typ av effektstorlek det rör sig om. Författaren menar att om studenter byter ut lägre provresultat mot senare högre provresultat (när studenterna lärt sig mer) blir betygsgenomsnittet för kursen högre. Genom att beräkna genomsnittsbetyg för kursen med och utan byte av provresultat var det möjligt att kontrollera för effekter av högre betyg i enskilda prov (om studenternas provresultat i gruppen där de kunde byta ut sina tidigare resultat mot senare är beräknade utan denna möjlighet, var effekten mindre (författaren redovisar en effektstorlek på 3,67 men preciserar inte vilken typ av effektstorlek det rör sig om eller hur uträkningen har gjorts). Författaren argumenterar för att det finns bevis för att, om möjligheten finns för studenter att byta ut tidigare summativa bedömningar mot senare summativa bedömningar för beräkning av deras betygsgenomsnitt, påverkas deras lärande, motivation för lärande och prestationer positivt. Författaren menar att genom att inte använda summativa bedömningar tidigt i studenternas utbildning ökar möjligheten för dem att utveckla sin motivation, akademiska självkänsla och lärande på ett positivt sätt. Studien är dock delvis odetaljerad och utelämnar information om metod och analyser vilket gör det svårt att generalisera resultaten. Det som dock är intressant är att det är en selekterad grupp högpresterande studenter som deltog i studien och att dessa studenter påverkas negativt om summativa bedömningar kommer tidigt i utbildningen.

## Sammanfattning

Fler av studierna presenterade i denna grupp har genomfört experiment med studenter på universitetsnivå. Deltagarna har valts ut genom bekvämlighetsurval i någon form och sedan har studenterna utifrån randomisering delats in i undergrupper som fått olika typer av feedback. Syftet för flera av dessa studier var att undersöka hur olika typer av feedback påverkar studenternas motivation för lärande och prestationer. Biez-Hernandez (2012) fann negativa effekter av negativ feedback på studenters prestationer och Pulfrey et al. (2011) fann att studenter utvecklade ett ytligt förhållningssätt till sitt lärande i betygsatta situationer. De drar slutsatserna att studenterna påverkas av typ av information i de summativa bedömningarna: negativ information och negativa incitament påverkar studenternas prestationer negativt jämfört med positiv information och positiva incitament. Docan (2006) fann ett svagt stöd för sin hypotes att negativ feedback ökar studenternas motivation. Däremot visade resultatet att studenter tyckte att betyg stressade dem och att det var bättre att få kontinuerlig information om sin utveckling. Dlaska och Krekeler (2013) menar att de finner stöd för att vuxna studiemotiverade studenter inte påverkas negativt av betyg. Vid jämförelser av feedback med och utan betyg finner de inga skillnader i deras prestationer. Elliot et al. (2005) fann att elever (17 år gamla) fick sämre prestationer på ett matematikprov om de fått negativ feedback tidigare. Författaren menar att det finns kraftfulla

bevis för att elever som utvecklar strategier för att undvika misslyckanden i sitt lärande får sämre prestationer. Vaden-Goad (2009) menar att studenternas motivation för lärande ökar om de blir bedömda senare i lärandeprocessen och att detta kan leda till bättre prestationer.

# Diskussion och slutsatser

Av de 22 inkluderade studierna i den fördjupade analysen genomfördes 11 av studierna på vuxna studenter på universitetsutbildningar. I alla dessa studier deltog studenter på utbildningar med högpresterande studenter som antagits från en selekterad grupp sökanden. I flera fall använde sig författarna av bekvämlighetsurval. Bristen på studier med randomiserat urval är stor och påverkar möjligheten att dra generella slutsatser utifrån studierna i denna översikt. I sju av studierna var deltagarna motsvarande högstadie- och gymnasienivå medan i fyra studier var deltagarna yngre elever. I ett fåtal studier användes en interventionsdesign, kvasi-experimentell eller experimentell design med ett representativt urval. I ett antal studier används olika typer av experiment för att undersöka hur bedömningar påverkar elevernas lärande, experiment som i vissa fall kan anses vara av *high-stakes* karaktär för deltagarna, medan i andra studier är experimenten inte av *high-stakes* karaktär.

Om vi bortser från de brister som finns i de redovisade studierna kan vi dra några övergripande slutsatser. För det första är resultaten från studierna till stor del samstämmiga. Vuxna högpresterande studenter såväl som yngre elever verkar påverkas positivt i sitt lärande, motivation för lärande och prestationer av feedback som innehåller mycket information som kommer i direkt anslutning till uppgiften för bedömningen och information bör vara positiv, jämfört med summativa bedömningar med lite information. Samtidigt framkommer det i en studie att vuxna studenter inte verkar påverkas negativt om feedback kommer i form av betyg till skillnad från yngre elever. Detta förklaras av att vuxna studenter på universitetsnivå ”kan” systemet och har lång erfarenhet av summativa bedömningar och har utvecklat strategier för att hantera och lyckas inom detta system. Dessa vuxenstudenter har antagits till universitetsutbildningar i konkurrens med andra och är redan en selekterad och relativt högpresterande grupp. Däremot verkar det vara annorlunda för yngre elever och när representativa urval används där hela distributionen av förmåga finns representerad. En slutsats som kan dras av resultaten från de inkluderade studierna är att summativa bedömningar generellt differentierar och påverkar framför allt låg- och högpresterande elever på olika sätt när det gäller lärande, motivation för lärande och prestationer i skolan och på universitet. Det som framstår som speciellt intressant i resultaten från de inkluderande studierna är att summativ bedömning och negativ feedback påverkar lärande, motivation för lärande och prestationer negativt över åldrar: även vuxna högpresterande studenter påverkas mer positivt av formativ bedömning jämfört med summativ bedömning. Dock verkar lågpresterande och yngre elever påverkas avsevärt mer negativt av betygsättning jämfört med äldre och högpresterande elever när det gäller deras lärande och prestationer. Ålder och erfarenheter av bedömning samt förmåga (resurssvaga/resursstarka) tycks spela en stor roll för hur elevers lärande, motivation för lärande och prestationer påverkas av summativ bedömning.

## Teoretiska brister i de inkluderande studierna

Ett annat intressant resultat från denna systematiska genomgång handlar om de metodiska och teoretiska antaganden som görs i de olika studierna. Generellt visar resultat från studier genomförda av utbildningsekonomer (exempelvis Ártes & Rahona 2013, Azmat & Iriberri 2010, Betts & Grogger 2003, Docan 2006, Figlio & Lucas 2004, Sjögren 2010) på mer positiva resultat av summativa bedömningar på senare utfall som utbildningslängd och prestationer jämfört med studier genomförda av forskare inom andra discipliner såsom psykologi och pedagogik. Det verkar som att teorierna som ligger till grund för studierna påverkar vilken information som analyseras och vilka variabler som används, utfallsvariabler som till exempel prestationer och motivation eller lön. Endast i Sjögrens studie (2010) används lön och utbildningslängd som enda utfallsvariabler vilket innebär att en viss typ av slutsatser kan dras, slutsatser som främst kan kopplas till ekonomiska teorier och inte till teorier om hur bedömningar påverkar elevers lärande och prestationer i skolan. Till exempel har det visats finnas stora skillnader mellan vilka faktorer som predicerar prestationer i skolan (betyg) och vilka som predicerar om elever går ut skolan och vilken lön en individ får i vuxen ålder (Levin 2012). I de utbildningsekonomiska studierna bygger de teoretiska antagandena på teorier som utvecklats inom

idrotts-, tävlings- och konkurrensstudier (Prendergast, 1999). Deltagare i dessa studier har varit olika idrottslag, individer i en tävlingskontext eller individer anställda på företag som har getts stimuli i form av belöningar för att prestera bättre. Vanligt fokus i denna typ av studier är med vilka hjälpmedel och verktyg arbetsledningen kan ”styra” arbetarna i rätt riktning för att de ska producera mer. I dessa studier har det främst varit statiska modeller där till exempel skillnader mellan arbetstagare med kontrakt där lönen till stor del är styrd av resultat (bonuslön) och arbetstagare med kontrakt utan bonuslön som har analyserats. På grund av tillgänglighet på viss typ av data har dessa teorier utvecklats från studier som har analyserat yrken som chefer, golfare, trädfällare där det har varit möjligt att relativt enkelt konstruera objektiva utfallsmått (Prendergast 1999).

Dessa resultat är givetvis viktiga framför allt inom forskning och policyutveckling för arbetsmarknaden men de flesta yrken är ofta mer komplexa än en del av de yrken som analyserats i de tidigare studierna, och utvärderas därför på mer subjektiva sätt till exempel genom arbetsgivarens och chefens utvärdering. Att överföra konkurrens och tävlingsteorier till en skolkontext inom utbildningsområdet innebär förmodligen en risk för förenklingar både med avseende på vilken metod och design som valts för studien samt vilka analyser som gjorts. Det handlar om att det är rimligt att anta att det finns skillnader mellan hur arbetare i ett arbetslag och idrottslag inom en idrottsgren påverkas av belöningar i form av bonusar, ökad lön och priser och hur elever i skolan som är i en utvecklings- och lärandeprocess påverkas av bedömningar. Inom bedömningsforskning är lärande- och motivationsteorier centrala för att förklara varför summativa bedömningar påverkar elever på olika sätt. Elever i skolan har olika sociala bakgrunder, förutsättningar och olika kognitiva och socio-emotionella förmågor vilket är faktorer som bör analyseras i relation till hur summativa bedömningar påverkar deras lärande, motivation för lärande och prestationer.

## Metodiska brister i de inkluderande studierna

Endast i tre studier var urvalet av elever nationellt representativt och det var i de tre svenska studierna (Klapp et al. 2014, Klapp 2014, Sjögren 2010). Detta kan bero på att Sverige under lång tid har byggt upp databaser med information för hela populationer. Att urvalet av deltagare är representativt är av stor vikt för att kunna dra generella slutsatser. I flera av studierna deltar elever och studenter som går på privata och selektiva skolor vilket innebär att resultaten är svårt att generalisera till andra grupper av elever. Dock kan antalet studier med liknande resultat visa på samband mellan exempelvis summativa bedömningar och elevers prestationer där vissa slutsatser kan dras. En annan brist i flera av de presenterade studierna är avsaknad av diskussion om urvalet och de möjliga konsekvenser det för med sig. Dock drar flera av studiernas författare relativt långtgående slutsatser av resultaten vilket kan få konsekvenser för policyutveckling och reformarbete. Studier kan misstolkas genom att ge intryck av att till exempel betygsättning påverkar lärande och prestationer positivt eller negativt trots att resultaten inte kan generaliseras bortom det urval som studerats. Bristande metodisk kvalitet, avsaknad av vetenskapliga bevis och bristen på detaljerade redovisningar av forskningsdesign riskerar att påverka policy och allmänhetens uppfattning om vad som påverkar elevers lärande och prestationer. Trots att forskare inser begränsningarna med sina egna studier och varnar för att dra generaliserande slutsatser används resultaten i media och vid policyförändringar (Raymond & Hanushek 2003). Att genomföra randomiserade studier med elever ställer dock höga krav på etiskt förhållningssätt och det finns svårigheter med tillgänglighet av data i vissa länder. Detta framförs av vissa författare som problematiskt och att det begränsar möjligheten att designa studier inom fältet som kan ge mer användbara resultat.

En annan brist handlar om att det är få studier som analyserar betydelsen av elevernas bakgrunder och förutsättningar. Ett fåtal studier kontrollerar för elevens förutsättningar som kognitiv förmåga. Könsskillnader redovisas i lite större omfattning, framför allt för de yngre eleverna. Endast ett fåtal studier kontrollerar för elevernas sociala bakgrund. En annan skillnad mellan studierna som kan förklara de disparata resultaten är de utfallsvariabler som används. Inom det utbildningsekonomiska fältet används ofta utbildningslängd och inkomst som utfallsvariabler. Inom andra discipliner används mått på lärande (betyg eller resultat på prov) eller mått på motivation som utfall. Dessa olika typer av utfallsvariabler betyder olika saker. Ett exempel kan vara att de kausala sambanden mellan att få betyg i de tidiga skolåren och lön i vuxen ålder är svåra att bevisa. Inom utbildningsvetenskap finns en mängd fenomen som kan orsaka att elever lär sig och presterar i skolan och detta kan i sin tur påverka utfall i vuxen ålder på en mängd olika sätt. Att kontrollera för alla tänkbara orsaker till låg

eller hög lön i vuxen ålder är alltså problematiskt. Ytterligare en annan aspekt handlar om att utbildningsforskning bör ta hänsyn till klustereffekter vilket är effekter som handlar om att elever på en skola har mer gemensamt med varandra än med elever på andra skolor: elever på samma skola har ”samma” skolklimat, samma rektor och samma lärare. Att ta hänsyn till dessa klustringseffekter är viktigt för att den metodiska kvaliteten ska bli hög och resultatet inte missvisande.

När resultaten ska användas för policyändamål bör följande aspekter tas i beaktande: inom vilken disciplin författaren skriver (till exempel pedagogik/psykologi eller ekonomi); vilka teoretiska utgångspunkter som används; vilka variabler som används; samt den metodiska kvaliteten.

Mot bakgrund av denna översikt visar resultaten av de genomgångna studierna att:

- Summativa bedömningar har en generell differentierande effekt: betyg påverkar olika elevgrupper på olika sätt beroende på prestationsförmåga och kön. Lågpresterande elever och pojkar får en negativ utveckling i sitt lärande och sämre prestationer med summativa bedömningar jämfört med högpresterande elever och flickor.
- Summativa bedömningar som betyg påverkar elevers prestationer negativt jämfört med formativ bedömning, över åldrar.
- Högpresterande äldre studenter (universitetsnivå) påverkas positivt av färre och senarelagda summativa bedömningar för sitt lärande, motivation för lärande och prestationer, jämfört med hög frekvens av summativa bedömningar och tidigt förekommande summativa bedömningar.
- Negativ feedback påverkar elever negativt: feedback som är summativ och som visar på elevers bristande kunskaper och svaga resultat verkar inte hjälpa elever att ”skärpa sig” utan elever presterar sämre med negativ feedback och de presterar bättre om de får feedback med mycket och kontinuerlig ”positiv ”information om hur de kan förbättra sitt arbete.
- Ett fåtal studier undersöker betygens effekter på yngre elevers lärande och prestationer (före årskurs 6) och dessa visar att feedback som ger mycket information om hur elever kan förbättra sina prestationer är bättre för deras lärande och prestationer jämfört med summativa bedömningar.

# FORSKNING OM BETYG UR ETT LÄRARPERSPEKTIV

I det här kapitlet presenteras vår sammanställning av betygsforskning utifrån ett lärarperspektiv. Sammanställningen visar att den största andelen forskning om betyg med ett lärarperspektiv fokuserar lärarens betygsättande praktik (hur och vad man betygsätter). Hur lärarens betygsättning undersöks, och vad som är studieobjektet, skiljer sig vidare till stor del åt mellan svensk forskning och forskning utanför Sverige. I Sverige är relationen mellan styrdokumenten och lärarens betygsättning central medan fokus i forskning utanför Sverige snarare ligger på relationen mellan lärarens betygsättning och elevers kunskap, framförallt validitetsaspekten i olika betygssammanhang såsom elevunderlag och ämne.

I kapitlet presenteras initialt hur forskning om betyg ur ett lärarperspektiv definieras i forskningsöversikten och hur sökningen av forskning om betyg ur ett lärarperspektiv har avgränsats. Sedan presenteras först forskning utanför Sverige som här benämns internationell forskning varefter svensk forskning presenteras.

## Metodbeskrivning

Hur forskning om betyg utifrån ett lärarperspektiv ska avgränsas är inte självklart. Vi har dock valt att avgränsa oss till forskning om lärarens bedömarkompetens och hur betygen påverkar lärares undervisning. Att bestämma vilken forskning som faller in under dessa kategorier kan sedan göras olika. Under den första frågeställningen har vi valt att placera in den forskningen som i Sverige ofta bedrivs i ett ämnesdidaktiskt perspektiv: forskning om lärarens betygsättande praktik, hur lärare betygsätter och vad de betygssätter. Frågan om lärarens bedömarkompetens infinner sig inte sällan även här. Men vi har även noterat en aspekt av betygsforskning som fokuserar läraren som betygsättare som intresserar sig för lärares upplevelser av, och attityder till, betyg och betygsättning. Denna forskning genomförs ofta genom intervjuer med lärare där den betygsättande lärarens beskrivningar och åsikter utgör empirin. Här framkommer också betygens inverkan på lärarens dagliga arbete i skolan.

Det går således att formulera två övergripande kategorier av forskning om betyg i ett lärarperspektiv:

- Lärares betygsättande praktik: I denna kategori har vi placerat forskning som studerar hur och vad lärare bedömer när de betygsätter. Det handlar om lärarens summativa bedömning och innefattar både prov och slutbetyg.

- Lärares upplevelser och attityder av betyg och betygsättning: I denna kategori har vi forskning som i sitt lärarperspektiv utgår från läraren och inte primärt betygsättningens praktik. Det handlar om lärares attityder och hur de resonerar och problematiserar betyg och betygsättning.

Inom dessa områden genomförs givetvis forskning med angränsande och mångfacetterade aspekter av betyg och betygsättning. I Tabell 3 har vi sammanställt den kvantitativa fördelningen av perspektiven i svensk och internationell forskning. Eftersom några artiklar och avhandlingar förekommer både i kategorin *lärares betygsättande praktik* och *lärares upplevelser och attityder till betyg och betygsättning* blir slutsumman inte överensstämmande med det totala antalet artiklar, vilket har synliggjorts genom att totalsumman placerats inom parentes.

Vi har genomfört systematiska sökningar i den internationella databasen ERIC, utvalda relevanta tidskrifters egna databaser och i Libris. I resultaten saknades dock forskning som vi sedan tidigare kände till. Det visar att allt inte kommer upp som träffar vid sökningar i databaser eftersom sökning via nyckelord, titlar och abstract inte täcker hela innehållet. Således har vi även genomfört manuella sökningar i relevanta avhandlingars och artiklars referenslistor.

*Tabell 3. Kategorisering utifrån studieobjekt*

| | Lärarens betygsättning: forskning om betygsättningens praktik | Lärares upplevelser/attityder om betyg och betygsättning | Summa |
|---|---|---|---|
| **Svensk forskning** | 25 | 8 | 33 (29) |
| **Internationell forskning** | 39 | 6 | 45 (40) |
| **Summa** | 64 | 14 | 78 (69) |

# Internationell forskning om betyg i ett lärarperspektiv

Den internationella forskningen om betyg i ett lärarperspektiv har framförallt sökts genom den internationella databasen ERIC som i sin tur bygger på flera olika databaser med relevans för forskning inom pedagogik och utbildningsvetenskap. De sökord som angavs är ”teacher” och ”grading”. När ordet *grades* användes blev urvalet 37 363 träffar och vid en överskådlig genomsökning noterade vi att sökordet *grades* inte genererade rätt sorts träffar. *Grades* betyder även årskurs på engelska vilket genererade forskning där termen förekommer i den bemärkelsen. I den forskning där *grades* betyder betyg handlade forskning ofta om betyg i en annan betydelse än utifrån ett lärarperspektiv eller om lärares betygsättning och då förekom även ordet *grading*. Det medförde att vi valde att koncentrera oss på de sökträffar som genererats med söktermerna ”grading” och ”teacher” både i ERIC och i tidskrifters sökmotorer. Vidare avgränsades sökningen i ERIC till referee-granskade artiklar.

Ytterligare avgränsningar som gjorts är att forskning om högre utbildning valts bort eftersom denna litteraturöversikt riktar in sig på skolans betygsättning. Avgränsningar har även gjorts utifrån studieobjekt (vilket benämns ämnesgrupper i ERIC). Efter en genomgång av de ämnesgrupper ERIC sorterar in forskningen under, valdes de ämnen som indikerade student/elev-perspektivet eller ett systemperspektiv bort eftersom dessa aspekter redan undersöks i andra delar av vår forskningsöversikt. Då genererades 1136 sökträffar.[2] Även då visar det sig att mycket handlar om effekter av betygsättning, såsom elevers upplevelser och attityder och relationer mellan elev och lärare, vilket således valdes bort. Vi har även valt bort sådan forskning som i huvudsak fokuserar andra aspekter som faller utanför frågeställningarna exempelvis artiklar utan empirisk förankring som räknar upp råd för en bättre betygsättning. Även forskning som visserligen omnämner läraren som betygsättare men där studieobjektet är något annat än betyg och betygsättningen har valts bort.

Efter avgränsning utifrån studieobjekt som faller utanför ramen för forskningsöversiktens uppdrag genererades 529 träffar. Efter läsning av abstract bedömdes 94 artiklar vara relevanta av de internationellt publicerade artiklarna, av vilka 6 stycken var svensk forskning. Dessa 94 artiklar har sedan lästs i sin helhet. I granskningen visar det sig att flertalet inte fokuserar betygsättning i ett lärarperspektiv eller att de inte utgörs av empirisk forskning. Samtidigt tillkom artiklar efter manuella sökningar i tidskrifters databaser och referenslistor. Det har resulterat i att 40 artiklar (32 från ERIC och 8 från manuell sökning) om betygsättning i ett lärarperspektiv har bedömts som relevanta för forskningsöversikten. Vi har sedan genomfört en tematisk kategorisering av dessa och placerat in dem i de övergripande perspektiven som presenterades i inledningen av kapitlet (se Tabell 3).

[2] Värt att nämna i sammanhanget är att när sökorden ”assessment” och ”teacher” användes genererades 32 895 träffar. Det visar hur stort forskningsfältet om bedömning är i jämförelse med betyg. I de fall där både *assessment* och *grading* (eller *marking*) förekommer i artiklarna påträffades dessa texter även i de fall sökorden var ”grading” och ”teacher”.

## Betygsättningens praktik

Majoriteten av den forskning som vi funnit om summativ bedömning utanför Sverige fokuserar i första hand inte på slutbetygen. Internationell forskning om betygsättning i ett lärarperspektiv fokuserar istället ofta på validitetsfrågan, även om betygsättningens reliabilitet också undersöks (se exempelvis Congon & McQueen 2002, Klein 2002).

Den dominerade andelen forskning om betyg i ett lärarperspektiv undersöker lärarens betygsättande praktik. Här är en aspekt lärarens betygsättning i relation till olika *standards* och nationella kunskapstester (Black, Harrison et al. 2010, Cooksey, Freebody & Wyatt-Smith 2007, Martinez, Stecher & Borko 2009, Russell & Austin 2010, Tierney, Simon & Charland 2011, Welsh, D'Agostino & Kaniskan 2013). I dessa påtalas att betygsättning i ökad utsträckning blivit reglerad och att lärarna har att förhålla sig till olika typer av *standards* i sin betygsättning. Det har medfört att provresultatet blivit en alltmer dominerande aspekt av bedömningen. I en amerikansk studie av hur betygsättningen i så kallade *report cards*, information till elever och föräldrar om elevens betyg utifrån olika ämnesspecifika aspekter, överensstämmer med resultaten på standardiserade kunskapstester i matematik och skriv- och läsförmåga visade det sig dock att betygen i framförallt läs- och skrivförmågan skilde sig från testresultaten (Welsh, D'Agostino & Kaniskan 2013). Samtidigt framhålls i flera av studierna att lärarna behöver väga in elevens individuella förutsättningar och behov för en valid bedömning. Det framhålls i det sammanhanget att betygsättning är en komplex praktik där exempelvis kollegiala samtal, och användandet av olika typer av stödmaterial i bedömningsarbetet utgör viktiga aspekter av bedömarkompetensen (Allal 2013, Cox 2011, Wyatt-Smith, Klenowski & Gunn 2010).

I vår genomlysning av internationell forskning om betygsättningens praktik framkommer flera studier om lärarens betygssättning som undersöker huruvida lärare betygsätter olika beroende på elevers individuella egenskaper. I dessa menar man att aspekter såsom kön, etnicitet och socioekonomiska förutsättningar påverkar lärarnas betygsättning (Willingham, Pollack & Lewis 2002, Hanna & Linden 2012, Sprietsma 2013, Van Ewijk 2011). Exempelvis visar en studie från Tyskland hur samma uppsats bedöms hårdare med ett lägre betyg som resultat, om lärarna tror att det är en elev med turkiskt ursprung som bedöms jämfört med om det är en tysk elev (Sprietsma 2013). Här ifrågasätts relationen mellan betyg och elevernas kunskapsnivå. Liknande resultat påvisades i en studie från Indien där elever från lägre kast bedömdes hårdare (Hanna & Linden 2012). Men det påtalas även att dessa faktorer kan ha mer indirekt påverkan på vilka betyg läraren sätter, exempelvis att lärares lägre förväntningar på elever med annan etnicitet än majoritetssamhällets medför att elever presterar sämre, en sorts självuppfyllande profetia, och därigenom får lägre betyg (Van Ewijk 2011).

I de internationella artiklar som vi funnit med ett lärarperspektiv är det, liksom påpekats ovan, framförallt praktiken och huruvida betygsättningen är tillförlitlig eller ej som undersöks. Relationen mellan gemensamma standards (standardiserade tester, kursplaner och kriterier) och lärarens betygsättning förekommer i det avseendet som forskningsobjekt men framförallt har vi funnit studier som intresserar sig för premisserna för lärarens betygsättning. I denna forskning undersöks vilka elevrelaterade faktorer det är som påverkar lärarens betygsättning. Individuella egenskaper som kön, etnicitet och socioekonomisk bakgrund nämndes ovan, men aspekter som handlar om elevbeteende lyfts också fram. Det handlar om elevens motivation, attityd och beteende i klassrummet samt elevernas förhållande till läxor. Samtidigt framhålls studieresultat (framförallt på prov) vara den mest avgörande faktorn vid lärares betygsättning (Cox 2011, McMillan, Myran & Workman 2002, Randall & Engelhart 2009, 2010, Resh 2009). I det fallet framhåller en studie att det är när eleven är på gränsen mellan att bli godkänd och underkänd som lärarna tar hänsyn till elevens individuella aspekter (Zoeckler 2007). Ett undantag här är en studie från Kina som visar att lärare i engelska i 20 skolor i norra Kina visserligen lägger vikt vid elevers studieresultat men att de framförallt tar hänsyn till ansträngning, hemarbete och studiebeteende när de sätter betyg (Sun & Cheng 2014). Här är alltså beteendeaspekten mest avgörande.

Ytterligare en aspekt av hur och vad lärare betygsätter är lärarens ämnestillhörighet. I flertalet studier framkommer att lärare resonerar olika beroende på vilket ämne de undervisar i med avseende på om de fokuserar studieresultat och prov eller beteendeaspekter som att hemarbete lämnas in, att eleven deltar i aktiviteter i klassrummet och anstränger sig i studierna (Biberman-Shalev, Sabbagh et al. 2011, McMillan, Myran & Workman 2002, McMillan 2001, Resh 2009). Det framkommer att framförallt lärare i matematik i huvudsak fokuserar studieresultat, och då främst utifrån provresultat, medan språklärare i större utsträckning

väger in beteendeaspekter. Forskning om musiklärares betygsättning pekar dessutom på att musiklärare tenderar att betona beteendeegenskaper i sin betygsättning och att betygsättningen inte utgår från gemensamma standards, exempelvis i kursplaner (Russel 2010).

En annan underkategori som framträder i forskningen om lärarens betygsättande praktik är den specialpedagogiska. Vi har funnit tre stycken relevanta artiklar inom detta område (Kurth, Gross et al. 2012, Mastergeorge, Martinez & Jose Felipe 2010, Silva, Munk & Bursuck 2005). Dessa studier visar framförallt hur lärare bedömer olika när de betygsätter elever med särskilda behov. I en av artiklarna beskrivs utifrån fallstudier i skolor hur lärarna kan betygsätta elever med särskilda behov mer korrekt och rättvist (Silva, Munk & Bursuck 2005). Silva och kollegor presenterar i artikeln ett system för betygsättning där denna individualiseras i samrådan med eleven och föräldrarna.

Faktorer som inte är direkt elevrelaterade som exempelvis klasstorlek, förekomsten av standardiserade tester, skolreformer och föräldrars förväntningar framhålls också påverka lärares arbete, inte minst betygsättningen (Duncan & Noonan 2007). En studie påvisar dessutom att i vilken ordning lärare rättar examinationer påverkar betygsättningen genom att lärare tenderar att sätta högre betyg ju fler examinationer de rättat. Mental trötthet framhålls vara en central orsak härvidlag (Klein 2002).

En norsk studie som analyserat hur lärares betygsättning kan öka elevers lärande framhåller att lärare betygsätter olika strängt beroende på elevunderlag. Inte bara individuella aspekter såsom kön och socioekonomisk bakgrund lyfts fram utan även att elevgruppen, om eleven befinner sig i en hög- eller lågpresterande elevgrupp, påverkar lärarens betygsättning där ett högpresterande sammanhang tenderar medföra att lärare är mer generösa i sin betygsättning (Bonesrønning 2004).

Forskning om hur bedömningsverktyg kan utvecklas för att öka tillförlitligheten i lärares betygsättning är ytterligare en aspekt av forskningen om lärarens betygsättande praktik. Det handlar bland annat om utveckling av *standards* i form av lokala kursplaner och utvecklingen av ett lärargemensamt förhållningssätt till betygsättning i syfte att öka tillförlitligheten i bedömningen. Det kan även handla om utveckling av rättnings- och betygssättningssystem, gemensamma riktlinjer för omprov och lägsta godkäntnivå (Cox 2011) och datoriserade system (Friedler, Tan et al. 2008). Hur summativ och formativ bedömning kan komplettera varandra framhålls också. Black, Harrison et al. (2011) visade i ett projekt med 18 lärare hur validiteten i lärarnas summativa bedömning kunde öka samtidigt som elevers lärande förbättrades genom att introducera portfolios och mer kollegiala samtal. Arbetet med portfolios medförde även att samtalen mellan lärare och elev ökade, vilket ytterligare bidrog till elevers lärande. I Hall & Hardings (2002) studie av sex skolor i England som pågick under två år, poängteras också vikten av lärares kollegiala samtal i samband med bedömning och att skolorna skapar något de benämner *assessment community of practice*.

Att kollegiala samtal skulle öka reliabiliteten stöds dock inte i en experimentstudie där lärare i historia fick rätta examensprov med eller utan kollegiala samtal och exempeltexter. Däremot hade alla grupper gemensamma betygskriterier (*mark schemes*) (Baird, Greatorex & Bell 2010). Utgångspunkten var tanken att betygskriterierna standardiserar betygsättningen medan lärarsamtalen och exempeltexterna kan koordinera och skapa medvetenhet om betygskriteriernas betydelse. Ingen av grupperna kunde dock påvisa mer eller mindre överensstämmelse i betygsättningen, något Baird och kollegor menar motsäger tidigare forskning om betydelsen av *community of practice* i betygssammanhang. De framhåller dock att lärarna som ingick i studien redan hade ett starkt kollegialt samarbete och flerårig erfarenhet att betygsätta utifrån standardiserade betygskriterier. Det senare menar Baird m fl. kan medföra att betygskriteriernas standardiserade effekt kvarstår med eller utan lärarmöten för sambedömning. De drar slutsatsen att vikten av pedagogiska kollegiala samtal troligtvis är av större betydelse i de fall där lärare inte redan har inarbetade rutiner för bedömning i *communities of practice*. I studien beskriver dock lärarna att de upplever exempeltexter och kollegiala samtal som viktiga i betygssammanhang.

Hur betygsättningen bör förankras i pedagogiska kollegiala samtal i samband med att regleringen av lärarpraktiken ökar, understryks i några studier (Cox 2011, Wyatt-Smith, Klenowski & Gunn 2010). Vikten av att betygsreformer förankras i lärarkåren framhålls av Susan Brookhart (2011) som även poängterar vikten av den formativa bedömningen och att ett ökat fokus på summativa bedömningar påverkar den formativa aspekten av bedömningen. Hur inflytandet av ansvarsutkrävande praktiker (*accountability*), reformer och standardiserade tester i USA påverkar lärarens summativa bedömning är centralt i hennes forskning. Dessa praktiker relaterar

hon till ett ökat misstroende mot lärarens betygsättning (Brookhart 2011, 2013a). Se även kapitel 3 och 4 i vår översikt.

Samma resonemang går att finna i forskning om lärares summativa bedömning i ljuset av olika betygsreformer och formella regleringar. I en studie av Pope, Johnson & Mitchell (2009), vilken även faller in under kategorin *lärares upplevelser och attityder*, framkommer att lärare upplever etiska dilemman i mötet mellan formell reglering och pedagogisk praktik i bedömningsarbetet. I denna studie analyseras olika dilemman som uppstår i spänningen mellan betygsättningens institutionella krav (såsom exempelvis standardiserade tester och kravet på godkänt för att få flytta upp en klass) och elevens behov och förutsättningar (Pope, Johnson & Mitchell 2009). De visar hur de institutionella faktorerna sätter lärarens professionella betygsättning på spel.

## Lärares upplevelse och attityder till betyg och betygsättning

Forskning om lärares attityder och upplevelser av betyg och betygsättning är mindre frekvent. Men vi har funnit några artiklar som framförallt intresserar sig för lärares erfarenheter och åsikter. Det framkommer exempelvis att lärarna upplever bedömning, inte minst den summativa, som problematisk. I Pope, Johnson & Mitchells (2009) studie, som även kan sägas falla in under forskning om lärares betygsättande praktik, ombads lärare beskriva en situation de upplever som etiskt problematisk i relation till bedömningsarbetet. Lärare berättar hur institutionella faktorer, exempelvis införandet av standardiserade tester, och något som benämns *score pollution*, att betyget inte motsvarar elevens ämneskunskaper på grund av att irrelevanta aspekter påverkar betygen, skapar etiska dilemman i deras bedömarpraktik. I bedömningsarbetet är det inte minst betygsättningen (på prov och som slutbetyg) som av lärarna beskrivs som problematisk. Krav på höga resultat på proven gör att lärare tränar eleverna innan provtillfället och på så sätt minskar betygens tillförlitlighet vilket är ett exempel på *score pollution*, föräldrars krav på höga betyg en annan. Men det kan även vara krocken mellan elevernas behov och institutionella krav på betygsättning som skapar dilemman, exempelvis när lärare upplever att ett underkänt betyg skulle missgynna eleven samtidigt som riktlinjerna i skolans betygspolicy kräver att läraren ger eleven ett underkänt betyg.

Krocken mellan institutionaliserade betygsrapporteringssystem och lärarnas syn på bedömningsarbetet framträder även i Zoecklers (2007) intervjustudie med lärare i engelska. Även här framkommer det att lärarna upplever att betyg kan manipuleras genom att underkända betyg ”fixas” (något som kan falla in under begreppet *score pollution*). Det framkommer också att lärarna misstror betygens möjlighet att visa elevens kunskapsnivå, framförallt med tanke på att kunskap förändras över tid.

I två studier från Kanada undersöktes lärares betygsättning i ljuset av en ökad reglering (Simon, Tierney et al. 2010, Suurtamm & Kochs 2014). I den ena studien följer vi matematikläraren ”Anne” i en fallstudie utifrån hennes betygsättning (Simon, Tierney et al. 2010). ”Anne” beskriver hur hon försöker hantera spänningen mellan en ökad standardisering av betygsättningen och praktikens förutsättningar. Centralt är att olika policy, på nationell och lokal nivå, tenderar att hamna i konflikt med varandra. Hon beskriver även dilemmat med att lärare betygsätter olika trots standardiseringar samt hur olika aspekter av elevbeteende hanteras vid betygsättningen. I Suurtamm & Kochs (2014) intervjustudie med 42 lärare i matematik i två skolor i Kanada analyseras framförallt lärarnas arbete med den formativa bedömningen. Även om det i första hand är den formativa bedömningen som ligger i fokus är studien relevant att ta upp eftersom den synliggör lärarnas dilemma i mötet mellan skolpolitikens fokus på betyg och standardiserade bedömarverktyg och lärarnas arbete med formativ bedömning. I betygsättningen krävs det förutom att lärarna följer föreskrivna mallar för betyg, att färdigformulerade fraser måste användas i rapporteringen. Dessa fraser, exempelvis *with considerable effectiveness*, menar lärarna går förbi föräldrars förståelse och resulterar i att de inte läser de omdömen som följer med betygen. Ett annat problem är att betygsrapporteringen i så kallade *achievment cards* består av färdigformulerade kunskapskategorier. Lärarna upplever att de behöver anpassa sin undervisning till dessa kategorier men att dessa snarare är kontraproduktiva. Suurtamm och Koch framhåller slutligen ett övergripande dilemma som handlar om krocken mellan den bedömning lärarna upplever som meningsfull och den bedömning som snarare handlar om ansvarsutkrävning (*accountability*). Se vidare kapitel 4.

I Hall och Hardings (2002) studie som diskuterats ovan, är lärarnas upplevelse av betygsättningen i fokus. I deras studie beskriver lärarna det problematiska med en politisk press att öka skolresultaten. Att skolors resultat

offentliggörs är en annan aspekt som lärarna upplever som ett problem i betygsättningen. De beskriver hur de upplever att lärarens arbete med bedömning fått allt mindre status i relation till de standardiserade kunskapstesterna. Hall och Harding menar att lärarnas professionalism hotas när deras professionella bedömning inte tillmäts samma status som resultaten på de nationella testerna.

I en stor studie i Holland där 260 skolor deltog framkommer det att lärare inte är intresserade av att införa standardiserade normrelaterade kunskapstester (Blok, Otter & Roeleveld 2002). Detta menar Blok m.fl. är kopplat till lärarnas syn på kunskapstesternas funktion. Lärarna, men även rektorerna, som deltog i surveystudien framhåller dock kunskapstesters lärandepotential. Det framkommer att det är i relation till elevers lärande som lärarna i studien uppskattar kunskapstester. När syftet är jämförelse och ansvarsutkrävande, vilket vid tidpunkten för studiens genomförande förordades från skolpolitiskt håll, ses kunskapstester som meningslösa. Sammantaget framkommer att lärarna och rektorerna är positivt inställda till de summativa bedömningarna och att de inte känner sig hotade av *high-stakes* tester, vilket författarna till studien menar kan ändras om dessa testers inflytande ökar i skolan.

En generell iakttagelse i internationell forskning är att det i stort saknas förankring till teorier om kunskap och lärande kopplade till forskningsbidragen i de båda kategorierna, även om flertalet artiklar om lärarens betygsättande praktik har en teoretisk-metodologisk förankring till sin empiriska analys. Däremot *diskuteras* betygsättningen i relation till lärande och hur den summativa bedömningen kan bidra till lärande och tydliggörande av elevens kunskapsutveckling och behov. Det är således inte den kunskap som bedöms som diskuteras eller hur betygen och andra summativa bedömningar påverkar lärares kunskapssyn. Framförallt är det betygens validitet och hur betygsättningen kan göras mer tillförlitlig som framstår som centralt när relationen betyg och kunskap är i centrum.

## Svensk forskning om betyg i ett lärarperspektiv

Vi presenterar den svenska forskningen på samma sätt som den internationella, utifrån om det är lärarens betygsättande praktik som är i fokus eller om det är lärares upplevelser och attityder till betygsättningen som är det huvudsakliga studieobjektet. Även i svensk betygsforskning med ett lärarperspektiv är det forskning om den betygsättande praktiken som dominerar. En stor del av forskningen genomförs dessutom inom en ämnesdidaktisk disciplin.

Det är framförallt licentiat- och doktorsavhandlingar men även en del vetenskapliga artiklar vi funnit som är relevanta för perspektivet (se tabell 4). Artiklarna är både publicerade i svenska och internationella tidskrifter. De svenska artiklar som publicerats internationellt tillhör både den internationella betygsforskningen och den svenska. Vi har dock valt att placera dem inom svensk betygsforskning. Vad gäller sammanläggningsavhandlingar där artiklar publicerats i internationella tidskrifter, redovisas dessa här som avhandling.

Svensk forskning om betyg i ett lärarperspektiv har vi framför allt hittat genom sökningar via Libris och manuellt i referenslistor. I sökmotorn Libris användes söktermen ”betyg*” med en avgränsning till böcker, avhandlingar och tidskrifter. Det var framförallt avgränsningen avhandlingar som genererade relevanta träffar. 35 licentiat- och doktorsavhandlingar samt 40 artiklar/kapitel i böcker påträffades. När det gäller artiklarna blev träffsäkerheten låg. Ett stort antal sökträffar var konferensbidrag från samma bedömningskonferens vid Linköpings universitet 2010 och kapitelbidrag i framförallt två böcker inom området (Lindström, Lindberg & Pettersson 2011, Lundahl & Folke-Fichtelius 2010). De böcker och kapitelbidrag som kom fram i sökningen var främst sådant vi redan kände till eller som föll utanför denna forskningsöversikt område. Efter en granskning av träffarna i Libris och av manuell sökning ansåg vi 18 licentiat- och doktorsavhandlingar samt 11 artiklar och kapitelbidrag vara relevanta för översikten.

*Tabell 4. Sökresultat, antal studier utifrån textgenre och kategori*

| | **Lärarens betygsättning: forskning om betygsättningens praktik** | **Lärares upplevelser/attityder om betyg och betygsättning** | **Summa** |
|---|---|---|---|
| **Doktorsavhandling** | 10 | 3 | 13 (11) |
| **Licentiatavhandling** | 6 | 3 | 9 (7) |
| **Artikel/kapitelbidrag** | 9 | 2 | 11 |
| **Summa** | 25 | 8 | 33 (29) |

Intresset för betyg som forskningsobjekt synes tydligt ha ökat i Sverige. En förklaring till det ökade intresset för hur lärare betygsätter kan rimligen förstås i ljuset av ett allmänt ökat fokus på betyg och på betygens likvärdighet. Skolverkets kvalitetsgranskning (Skolverket 2000) kan i detta avseende sägas utgöra en startpunkt för ett antal studier som fokuserar lärarens betygsättning i relation till styrningen av betygsättningen. I denna påvisades brister i hur lärare betygsätter i relation till styrdokumenten. Betygens funktion som urvalsinstrument ifrågasattes då betygen inte framstod som rättvisa och likvärdiga (Selghed 2010). Under perioden presenteras också forskning som påvisade betygsinflation och hur betygen brister som kunskapsmått (Cliffordsson 2004, Wikström 2005a).

Under 2000-talet växer en betygsforskning fram som fokuserar lärarens betygsättning i relation till styrningen och betygens funktion som urvalsinstrument och de problem, dilemman och praktiker betygsättningen omges av (Korp 2006, Lindberg 2002, Selghed 2004, Tholin 2006). Denna inriktning av betygsforskning handlar om lärarens betygsättande praktik, men här i relation till styrningen av skolans verksamhet. Denna typ av forskning faller in under den första kategorin *lärarens betygsättande praktik*.

Den andra kategorin, *lärares attityder och upplevelser av betyg och betygsättning*, genererade även i svensk betygsforskning ett mindre antal studier. Under denna kategori placerar vi forskning som intresserar sig för lärares föreställningar, erfarenheter och hur de resonerar kring betyg och betygsättning.

Det förekommer också en betygsforskning som intresserar sig för summativa bedömningar och hur de används som styr- och kontrollmedel av skolan. Dessa handlar, när det gäller forskning i ett lärarperspektiv, om hur lärares arbete påverkas av betygens ökade inflytande (Forsberg 2008). Det är egentligen en systemkritisk forskning men faller till viss del in i den kategori av forskning om betyg i ett lärarperspektiv då den ökade styrningen av lärares summativa bedömning relateras till den ökade kontrollen och styrningen av bedömning i skolan. Bland annat lyfts frågan hur lärares undervisning kan komma att påverkas av ett ökat fokus på resultat (betyg och resultat på standardiserade kunskapstester). Det hävdas att risken är stor att det är det som går att mäta som fokuseras i undervisningen och att andra aspekter av kunskap får stå tillbaka (Forsberg & Lundahl 2006, Forsberg 2008, Lundahl 2010, Englund, Forsberg & Sundberg 2012). Denna forskning faller dock för det mesta utanför denna översikt då den inte baseras på empiriska undersökningar. Det finns dock policystudier som diskuterar betygens betydelse för lärarens arbete och hur en förskjutning av skolans kunskapsbedömningar skett mot att användas i styrningen av skolan även om lärarperspektivet inte kan sägas vara det centrala (se exempelvis Forsberg & Lundahl 2006, Forsberg 2008). De utgör exempel på forskning i ett kritiskt perspektiv som framhåller att ett ökat fokus på elevers kunskapsresultat, bland annat mätt genom betygen (även andra summativa bedömningar såsom PISA och nationella prov), får konsekvenser för lärares möjlighet att verka. Införandet av godkäntgränsen och det målrelaterade betygssystemet 1994 är en sådan aspekt (Hultqvist 2011, Kroksmark 2002, Lindberg 2002), men även hur ett ökat fokus på kunskapsmätningar överlag inverkar på lärares undervisning och hur kunskap värderas (Forsberg & Lundahl 2006).

# Betygssättningens praktik: svenska doktors- och licentiatavhandlingar

Hur kunskapsbedömningar, och däribland betygen, påverkar och sätter ramar för lärarens pedagogiska verksamhet synes vara ett växande kunskapsområde. Under 2011 kom flera licentiatavhandlingar som på olika sätt berörde bedömningsfrågan (se exempelvis *Forskarskolan för lärare i historia och samhällskunskap* vid Karlstad universitet). Bland dessa finns flera som intresserar sig för prov- och betygsfrågan i relation till lärarens arbete.

## Betygsättningen i spänningsfältet mellan styrning och praktik

I Sverige har främst under 2000-talets första sekel betygen debatterats och kopplats till en krisdiskurs där skolans likvärdighet ifrågasatts (Mickwitz 2011, Selghed 2010, se även Englund, Forsberg & Sundberg 2012). Läroplansteorin har haft stort inflytande i svensk pedagogisk forskning och använts i flertalet undersökningar av relationen mellan skolans formella reglering och lärarens betygsättning. En stor andel av den svenska forskningen om betygssystemet och hur den omsätts i praktiken visar hur relationen mellan teori och praktik brister (Bergman 2007, Korp 2006, Selghed 2004, Tholin 2006, Wedin 2007). Det framkommer exempelvis att olika skolor resonerar olika när de sätter betyg, inte minst i relation till nationella prov (Bergman 2007, Korp 2006).

I Helena Korps doktorsavhandling *Lika chanser på gymnasiet? En studie om betyg, nationella prov och social reproduktion* (2006) framkommer det att olika skolor resonerar olika när de sätter betyg i relation till nationella prov i gymnasieskolan. Att få ett godkänt betyg i en skola behöver inte betyda att man skulle fått det i en annan, menar Korp. Det är framförallt mellan yrkesprogram och studieförberedande program denna skillnad föreligger. Lärarna i svenska, engelska och matematik tenderar att i de yrkesförberedande programmen fokusera form och reproduktion, medan eleverna i de studieförberedande programmen får undervisning i teoretisk kunskap och analytisk förmåga. Och här uppstår ett problem, menar Korp, eftersom eleverna sedan mäts utifrån samma krav i de nationella proven. Hon visar också att lärare resonerar olika i skolorna i fråga om betygsättningens relation till de nationella proven. På de yrkesförberedande programmen är det viktigaste att lärarna får eleverna över G-gränsen. Lärare ”friar hellre än fäller” och letar poäng i provresultaten ifall eleven inte når upp till godkänt (Korp 2006, s. 205). Elever som riskerar IG kan också ”räddas” ifall de får godkänt betyg på nationella provet. Däremot gäller inte det omvända. Ett icke-godkänt betyg på det nationella provet medför inte att en elev som hittills fått godkänt kommer att få ett underkänt betyg. Lärare som arbetar i skolor med generellt sett låga meritvärden, beskriver dessutom hur de pressas att sätta höga betyg av skolledningen. Studierna visar hur det finns ett glapp mellan regleringen av betygen och lärarnas betygsättning i skolan.

Även i Ann-Sofie Wedins avhandling (2007) framkommer att lärarna försöker balansera mellan å ena sidan det reglerande systemet i form av betygskriterier och kursmål och praktikens förutsättningar å andra sidan. Med avseende på betygsättning visar Wedin på samma tendens till pragmatiska lösningar i en kaotisk vardag som en stor del av svensk betygsforskning under perioden. Lärarna har olika strategier för att få sin vardag att fungera. Lärarna önskar att eleverna ska få så bra betyg som möjligt vilket visar sig i en praktik där lärarna får jaga oinlämnade G-uppgifter för att kunna sätta betyg. Wedin menar att lärarna tar stort ansvar för att eleverna ska få de betyg de eftersträvar. Lärarna låter exempelvis elever som misslyckats på det nationella provet göra omprov. När läraren bedömer att eleven har kunskaper motsvarande G men inte lyckats visa det på det nationella provet, ges nya chanser.

Flera av avhandlingarna studerar lärarens betygsättning i relation till styrdokumenten. Jörgen Tholin menar i sin avhandling (2006) att de lokala betygskriterier skolorna i 1994 års betygssystem var tvungna att formulera, skiljde sig markant från skola till skola både gällande utformning och innehåll, vilket riskerar likvärdigheten. Inom 14 av de 93 skolor han studerat saknades lokala betygskriterier i ett eller två ämnen.[3] Tholin visade också

[3] Notera att kravet att upprätta lokala betygskriterier nu mera är avskaffat.

i sin analys av lokala kursplaner ute på skolorna, att det fanns en tydlig skillnad mellan ideal och verklighet i betygsammanhang.

Bengt Selghed (2004) har i större utsträckning ett lärarperspektiv genom att han intervjuat 30 lärare i matematik, svenska och engelska om deras betygsättning. Lärarna berättar om hur de upplever betygssystemet och hur de arbetar med betygsättning. Men det är inte lärares attityder eller upplevelser av betygens inflytande över undervisningen han studerar utan framförallt hur de grundar sin betygssättning. Selghed menar att lärarnas inställning till betygsreformen 1994 inte överensstämmer med de intentioner som reformen uttrycker. Enligt Selghed är det få av lärarna som förstått betygssystemet och den kunskapssyn som ska ligga till grund för betygsättningen. Lärarna väger bland annat in andra faktorer i sin bedömning och jämför även eleverna sinsemellan. Selghed menar i sin avhandling, i linje med Korp, att lärare brister i förståelsen av det målrelaterade betygssystemet och att ”studiens utfall är nedslående, inte minst ur rättssäkerhetssynpunkt för eleven.” (s. 198).

Selghed menar att en förklaring till att lärarna inte följer det reglerande systemet i sin betygsättning är att de inte vill framstå som ”dåliga pedagoger”. Han, liksom Tholin, påtalar att en orsak till att lärarna sätter godkänt trots att eleven inte nått upp till de nationella målen kan vara det merarbete som ett underkänt betyg medför. Även Wedin diskuterar den ökade tidsåtgång som kravet på godkänt betyg för alla elever medfört. Men hon visar hur lärare utarbetat strategier i syfte att få eleverna godkända i sina kurser och inte hur de löser kravet att godkänna eleverna genom att sätta G trots att de inte når upp till målen.

Lotta Bergmans (2007) avhandling handlar om svensklärares arbete i fyra gymnasieklasser. Centralt i studien är vad lärarna ser som viktigt och hur lärare betygsätter. Ett av hennes resultat är att svenskundervisningen framförallt riktas mot formell färdighetsträning. Hon kan också visa att lärarnas bedömning är inriktat på uppnåendemålen. Lärarna uttrycker det komplicerade i att dels individanpassa undervisningen i svenska, dels att betygen i de olika elevgrupperna bygger på samma betygskriterier. Kravet på betygens likvärdighet i relation till styrdokumenten krockar med kunskapsuppdraget. Lärarna uttrycker det som att ”de gemensamma målen och betygskriterierna begränsar lärares frihet att välja olika innehåll för olika grupper, eftersom rättvisa betyg ska sättas” (s. 129). Bergman konstaterar att svenskundervisningen i stor utsträckning tenderar att fokusera formell färdighetsträning. Kursplanens övergripande mål är vidare underordnade i lärarens undervisning. Istället utgår lärarna från betygskriterier och mål att uppnå när de planerar sin undervisning. Lärarna ger elever på olika program som har olika intressen och förutsättningar samma kunskapsmässiga innehåll men mer eller mindre omfattande. Lärarnas fokus på betygskriterierna och den kommande betygsättningens krav på likvärdighet går att koppla till detta resultat. Men detta, menar Bergman, resulterar i en ”lägre” svenska för de yrkesförberedande programmen och en ”högre” för de studieförberedande. Hon ser alltså, liksom Korp, att eleverna på de yrkesförberedande programmen får en annan undervisning än eleverna på de studieförberedande programmen. Samtidigt framkommer det också att betygens påverkan på undervisningen gått så långt att inslag som inte är betygsgrundande anses mindre viktiga.

Ytterligare en ämnesdidaktisk avhandling är Anne Dragemark Oscarson (2008) studie. Den fokuserar i huvudsak på formativ bedömning och på elevers självbedömning men har även som delsyfte att undersöka om dessa praktiker kan göra bedömningen mer rättvis. Hon följer fyra gymnasieklasser i engelska. Eleverna ska bedöma skrivuppgifter där en av utgångspunkterna är betygskriterierna, en annan de betyg lärarna satte på det nationella provet. Lärarna beskriver hur eleverna blir mer medvetna om hur de kan arbeta med kriterierna och att en vinst är att diskussionerna om grunderna för betygssättningen minskar när lärarna arbetar formativt. Dragemark Oscarsson menar även att lärarnas betygsättarkompetens ökade.

Spänningen mellan styrning och verksamhet är centralt i Helge Räihäs avhandling *Lärares dilemman* (2008). Räihä studerar lärarnas olika dilemman i den dagliga verksamheten. Det är i spänningsfältet mellan globala systemstyrmedier, t.ex. lagparagrafer, tid och pengar, och lokala förväntningar som han undersöker olika dilemman. Räihäs studie visar hur lärares språkliga kompetens är centralt i hur de hanterar olika dilemman i sin dagliga verksamhet. Men även här framkommer problem, vilket synliggörs i ett exempel som är aktuellt att lyfta här. I lärares arbete med att formulera lokala mål och betygskriterier för svenska och engelska uppstår ett dilemma. Lärarna måste förhålla sig till läroplanens och kursplanens text när de formulerar de lokala målen och kriterierna trots att dessa ofta upplevs motsägelsefulla i praktiken. Räihä påpekar att det inte är frivilligt för lärarna att förhålla sig till styrdokumenten utan det krävs att de utgår från dessa, vilket gör att lärarna ofta bara

kopierar läroplanstexten istället för att formulera egna lokala mål och kriterier. Det resulterar dock i att lärarna inte kan leva upp till kravet på lokal anpassning av läroplanen.

En stor andel av de avhandlingar vi funnit studerar lärarens betygsättning i spänningsfältet mellan styrning och verksamhet. Dessa visar framförallt hur olika strukturella och organisatoriska faktorer i skolans dagliga verksamhet riskerar påverka lärarnas pedagogiska arbete med betyg och betygsättning.

## Bedömarverktyg och betygens validitet

*Hur* lärare bedömer och *vad* de bedömer undersöks i flera studier. Här är de ämnesdidaktiska studierna framträdande. Inte minst har vi funnit flera ämnesdidaktiska licentiatavhandlingar om betyg och bedömning.

Christina Odenstad (2010) analyserar samhällslärares skriftliga prov och jämför dem med hur ämnet skrivs fram i kursplanen. Hon har även ambitionen att jämföra hur prov utformas på gymnasiets yrkesförberedande- och studieförberedande program. Odenstad menar att styrdokumenten över lag implementeras i lärarnas bedömningsarbete men att det främst är de generella förmågor styrdokumenten formulerar (till exempel att eleven ska kunna redogöra för, analysera och diskutera) som testas. Undersökningen pekar på att det finns en del mål i styrdokumenten som inte behandlas i proven. Framförallt är det aspekter som mångkulturella frågor, miljömässiga och etiska perspektiv samt frågor om mänskliga rättigheter. Ytterligare ett resultat är att de förmågor som framhålls i kursplanen 1994 och som återfinns i strävansmålen i högre grad testas på de studieförberedande programmen. De yrkesförberedande programmens prov fokuserar också i högre grad kortfattade faktafrågor och riktas inte i samma utsträckning som för eleverna på de studieförberedande programmen mot de högre betygen. Odenstad menar att proven ger eleverna olika bilder av ämnet samhällskunskap.

I Tobias Janssons (2011) ämnesdidaktiska licentiatavhandling är validitetsfrågan vid provrättning i fokus. Han undersöker utifrån intervjuer med sex lärare, samt med hjälp av analyser av deras prov i samhällskunskap, i vilken grad prov testar de kunskaper som styrdokumenten uttrycker att eleverna ska uppnå för de olika betygen. Jansson menar att det visserligen råder en överensstämmelse mellan lärarnas tolkning av de olika betygen och vad som uttrycks i betygskriterierna men att det saknas en tydlig koppling mellan kursmål (mål att uppnå) och prov. Kursmålen, menar Jansson, ställer högre krav på elevernas kognitiva förmågor än vad som uttrycks i betygskriterierna. Lärarna bedömer även elevens kunskap utifrån kriterierna olika. Jansson påpekar att det finns problem med validiteten i de skriftliga proven och att det finns en del att utveckla för att nå en bättre samstämmighet, vilket tyder på att han menar att reliabiliteten också brister.

Ytterligare en ämnesdidaktisk licentiatavhandling är Izabela Segers (2014) studie av betygsättningen i ämnet idrott och hälsa. Hon har intervjuat lärare och undersöker hur de betygsätter, men även hur de upplever betyg och betygsättning, vilket gör att hennes studie både faller in under kategorin *lärares betygsättande praktik* och *lärares upplevelser och attityder till betyg och betygsättning*. Lärarna i studien anser att alla förmågorna som beskrivs i kursplanen måste ingå i betygsättningen. Samtidigt visar Seger att de är osäkra på hur de ska bedöma alla förmågorna. Inte minst beror detta på kunskapskravens olika värdeord som av lärarna upplevs svåra att tolka. Lärarens egen tolkning har fortfarande stor betydelse i betygsättningen, vilket är ett problem menar Seger. Det leder oundvikligen till olika bedömningsgrunder och därmed brister i likvärdighet.

Betygsättningen på yrkesprogrammen studeras av Helena Tsagalidis i avhandlingen *Därför fick jag bara Godkänt: Bedömning i karaktärsämnen på HR-programmet* (2008). Genom intervjuer med yrkeslärare urskiljer hon sju nyckelkvalifikationer och fem kategorier av specifika yrkeskunskaper som anses centrala i betygsättningen i karaktärsämnen. Tsagalidis menar att lärarnas värdering i sin bedömning avspeglar deras yrkeskulturella kunskap där erfarenhet väger tungt. Av den orsaken kan det för en oerfaren elev vara svårt att uppnå högre betygsnivåer i karaktärsämnen.

Peter Nyström undersöker i en sammanläggningsavhandling (2014) olika aspekter av validitet i lärares bedömning i matematik. Det är framförallt den summativa bedömning såsom exempelvis de nationella proven som undersöks. En viktig slutsats är att även konsekvenser av bedömningen måste beaktas, vilket kräver att lärare klargör provets syfte, läroplanens mål samt den kunskapssyn som eftersträvas. Eftersom läroplanen definierar utbildningens mål, och därmed vad som ska bedömas hävdar Nyström slutligen att en viktig egenskap hos bra bedömningar är att de är samstämmiga med läroplanen.

Även Wikström (2005a) intresserar sig för betygens reliabilitet och validitet. Huvudfokus i hennes sammanläggningsavhandling är dock betygens alltmer betydande roll vid antagning till högre utbildning och utvärderingar av skolor. Höga betyg ger inte bara studenter möjlighet till vidare studier utan används också som kvalitetsindikatorer för skolor, framhåller hon. Detta har, menar Wikström, lett till en betygsinflation. Betygens validitet kan ifrågasättas framförallt på grund av dess dubbla funktion av att fungera som urval och som tecken på skolkvalitet. Ur ett lärarperspektiv är det inte lärarens betygsättning i sig hon studerar utan istället visar Wikström hur lärares betygsättning kan ifrågasättas utifrån olika aspekter. Hon kan genom sin kvantitativa studie påvisa systematiska skillnader i betygssättningen kopplad till enskilda skolor och ser att framförallt mindre skolor och friskolor tenderar att ge eleverna högre betyg än större skolor och kommunala skolor. Slutsatsen är att det finns brister i betygens tillförlitlighet.

## Betygssättningens praktik: artikel- och kapitelbidrag

Betygsättningens validitet, exempelvis att lärare väger in andra aspekter i sin summativa bedömning än de som styrdokumenten framhåller, är ett vanligt studieobjekt både i avhandlingar och i artiklar. Liksom i genomgången av de internationellt publicerade artiklarna kan vi i svensk forskning notera att validitets- och reliabilitetsfrågan är aktuell. Forskning har påvisat att personliga egenskaper hos eleverna kan påverka lärares betygsättning genom att lärare exempelvis ger högre betyg till flickor än pojkar. En orsak som framhålls härvidlag är att flickor påvisar större motivation till lärande (Klapp Lekholm & Cliffordsson 2008, Klapp Lekholm 2010). Alli Klapp Lekholm menar dessutom att externa faktorer kan påverka lärarnas betygsättning så som föräldrars inflytande i skolan, politiska beslut för den pedagogiska verksamheten och externa kunskapsmätningar. En hel del forskning som intresserat sig för vad betygen mäter, pekar mot att andra aspekter än det ämnesspecifika spelar in när lärare sätter betyg. Exempelvis kan skillnaden mellan lärares betygsättning och elevers resultat på de nationella proven förklaras med lärares dilemma i den specifika klassrumssituationen. Här möts pressen på lärarens objektivitet i bedömningen med den mer subjektiva relationen mellan elev och lärare där andra aspekter såsom elevens sociala förmåga och beteende har betydelse (Annerstedt & Larsson 2010, Klapp Lekholm & Cliffordsson 2008). Klapp Lekholm och Cliffordsson (2008) framhåller att skillnader mellan provresultat och betygsresultat reflekterar de dilemman lärare möter i sin betygsättande praktik. Exempelvis att lärare ska motivera eleverna och ha en god relation till dem, samtidigt som de ska genomföra objektiva kunskapsbedömningar.

Vad lärarna fokuserar i sin betygsättning och hur olika ämnestillhörighet påverkar bedömningspraktiken återfinns också som ett studieobjekt i svensk betygsforskning. Språklärares grunder för betyg studeras av Mats Oscarsson och Britt-Marie Apelgren (2011) i en kvantitativ studie baserad på en enkätundersökning med 605 lärare som kompletterats med 20 intervjuer. Det visar sig att traditionella prov och inlämningsuppgifter dominerar som betygsunderlag. För lärare i moderna språk och engelska dominerar även muntlig prestation i klassrummet. För samtliga lärare visade sig metoder såsom portföljmetodik och självbedömning vara ovanliga som bedömningsunderlag. I deras studie, liksom i Claes Annerstedt och Staffan Larssons (2010) studie av lärare i idrott och hälsa, påvisas i likhet med internationell forskning, att lärare även tar med beteendeegenskaper i sin bedömning såsom att elever visar intresse och engagemang och gör sina läxor. Annerstedt och Larsson intresserar sig för validitets- och reliabilitetsaspekten av betygsättning. De menar att lärare i idrott och hälsa inte förmår explicitgöra sina bedömningsgrunder utan har egna utgångspunkter i sin betygsättning. Det medför att betygen varierar mellan skolor och lärare. Det visar sig också att individuella beteendeaspekter såsom visat intresse, positiv inställning, att vara i tid och ombytt och att göra sitt bästa spelar stor roll i lärarnas betygsättning. Annerstedt och Larsson menar att dessa outtalade grunder för betygsättningen leder till brister i reliabiliteten och validiteten i betygsättningen. Intressant är att det i en annan studie av lärare i idrott och hälsa (Redelius et al. 2009) framkommer att dessa främst grundar sin betygsättning på elevers förmåga och idrottsresultat även om positiv inställning och visat intresse har betydelse. Redelius och kollegor menar att det är anmärkningsvärt då styrdokumenten inte framhåller personlig förmåga och idrottsresultat utan att eleverna ska utveckla ett livslångt lärande och förståelse för vikten av fysisk aktivitet för hälsa och välbefinnande. Förmågor som styrka, snabbhet, aktivitet och att vara en bra ledare premieras av lärarna. Även om detta inte självklart kan förklara de skillnader som påvisats i hur killar och tjejer upplever ämnet (flickor är

mer negativa till idrott och hälsa än killar), menar Redelius et al. att dessa egenskaper traditionellt sätt anses manliga. Att idrottslärare inte följer styrdokumenten får stöd i ytterligare en studie (Svennberg, Meckbach & Redelius2014) där det framhålls att lärare i idrott och hälsa snarare följer sin magkänsla än betygskriterierna när de sätter betyg.

I Jan-Erik Gustafssons och Gudrun Ericksons (2013) studie är de nationella proven i fokus. De undersöker det faktum att lärare sätter högre provbetyg än Skolinspektionen gör när de låter omrätta de nationella proven. Detta har, menar författarna, lett till att lärares betygsättning av de nationella proven misstros. Gustafsson och Erickson menar dock att de externa rättningarna av de nationella proven inte utgick från samma förutsättningar som lärarnas rättningar, vilket framhölls som en delförklaring till skillnaden i betyg. Bland annat visade sig de kopior på texter som Skolinspektionen lät omrätta vara i dåligt skick och svårlästa, något Gustafsson och Erickson framhåller kan bidra till lägre betyg. Samtidigt skilde sig själva betygsskalan åt mellan rättningstillfällena vilket torde medföra att betygen faller olika ut.

I ett kapitelbidrag i en antologi om kemiundervisningen i svenska och finlandssvenska skolor bidrar Viveca Lindberg och Ragnhild Löfgren (2011) med en studie om vad lärare bedömer i kemiundervisningen. Det är framförallt prov som analyseras men även intervjuer med lärare har genomförts. Det framkommer att grunderna i bedömningen i styrdokumenten överensstämmer relativt väl mellan länderna. Däremot kunde Lindberg och Löfgren iaktta flera skillnader i hur proven konstrueras och hur bedömningen genomförs. Framförallt framkom att lärarna i Finland inte använder betygskriterierna i styrdokumenten såsom lärarna i Sverige. De finska lärarnas betygsättning utgick vidare utifrån en normalfördelningskurva där betygen justerades i relation till provresultat så att en jämn fördelning av betygen skulle uppnås. De svenska lärarna bedömde proven utifrån kvalitativa skillnader där G-frågor bestod av vad-hur frågor och VG och MVG-frågorna krävde förklaringar. Däremot var en likhet mellan lärarna i Finland och Sverige att lärarna inte bara tog hänsyn till studieresultat i sin bedömning utan även beteendeaspekter, något vi kunnat se i flera forskningsresultat om lärares bedömning.

## Lärarens attityder till betyg och betygsättning

I de fåtal studier som i huvudsak fokuserar lärarens upplevelser och attityder till betyg och betygsättning är det ofta lärarnas dilemman som framhålls. En ämnesdidaktisk licentiatavhandling som fokuserar lärares dilemman är Annika Karlssons studie av sju samhällskunskapslärares förhållningssätt vid betygsättning (2011). Där framkommer tre dilemman som alla handlar om relationen mellan ämnet, styrdokumenten och lärarnas egna övertygelser. Det handlar framförallt om huruvida lärarna ska fokusera på betyg alternativt att fokusera på elevens lärande och utveckling, motsättningen mellan lärarens egna övertygelser och tolkningar i relation till andra krav i den konkreta betygsättningen samt relationen mellan vad läraren kan bedöma och vad som ska bedömas utifrån ämnets krav. Karlsson menar att när lärare i ökad utsträckning reflekterar och diskuterar tillsammans kan betygsättningens validitet öka och lärarna få hjälp att hantera dilemman i sin betygsättning, något som även lyfts i internationell forskning.

I Larissa Mickwitz (2011) licentiatavhandling beskriver lärare i fokusgruppsintervjuer hur den betygsättande aspekten av arbetet konkurrerar med deras pedagogiska arbete med elevers lärande. Att skolledningen efterfrågar höga betyg i en alltmer konkurrensutsatt skola medför också att lärare upplever press att sätta högre betyg. Samtidigt är rätt betyg i enlighet med styrdokumenten ett viktigt kriterium för professionalism. Det framkommer att lärarna strävar efter att sätta ett korrekt betyg utifrån betygskriterierna samtidigt som de känner press från elever, föräldrar och rektorer på att sätta ett högt betyg. De externa och interna faktorer som påverkar lärarnas betygsättande praktik medför att lärare måste hävda sin professionalism och medvetandegöra hur de bedömer, på vilka grunder och hur det påverkar elevers lärande för att hävda sitt rätta betyg (se även Klapp Lekholm 2010, Mickwitz 2015). Betygsättningen framstår som en motstridig praktik. Även Helge Räihäs studie (2008) pekar ut bedömning och betygssättning som en av de arbetsuppgifter lärare upplever som motstridiga. Dessa dilemman menar Räihä ofta beror på motsättningar mellan olika styrsystem och lokala förväntningar.

I Segers licentiatavhandling (2014), som även diskuterats tidigare, framkommer att lärarna möter kravet på insyn i deras betygsättning (och att den ska kunna granskas) med att alltmer dokumentera betygsgrunderna. Detta dilemma handlar om att ”ha ryggen fri” som lärarna uttrycker det i studien. Lärarna uppger att de utför

bedömningar dagligen och uttrycker en oro för att det kan inverka negativt på elevernas spontanitet och rörelseglädje. Men samtidigt menar lärarna att de när de dokumenterar och utvecklar matriser, får en möjlighet till formativ bedömning genom att enklare kunna ge feedback till eleverna. Seger ser dock en risk med att lärarna så ”slaviskt” utgår från styrdokumenten i sin betygsättning och dokumentation. Hon påtalar risken med att lärarnas professionalism förminskas och att arbetet reduceras till att reproducera vad som beskrivs i kursplanerna.

Illona Rinne (2015) studerar i sin avhandling genom en analys av 149 inspelade betygskonferenser på gymnasiet, hur betyg förstås och förklaras av lärare och elever. Dessa betygskonferenser delas in i samtal två grupper av samtal där å ena sidan eleven och läraren kommer överens och å andra sidan där en gemensam syn inte uppnås. I en hermeneutisk analys utkristalliserar hon olika teman i samtalen. Rinne framhåller några motsättningar i betygssamtalen. Bland annat är ofta lärarens utgångspunkt att eleven ska ha ett så högt betyg som möjligt vilket inte är fallet för eleverna i samma utsträckning. Aspekter av betygsättningen som inte framkommer i styrdokumenten är också synliga, bland annat att elevens välbefinnande är central för läraren. Det medför att betyg kan ges för att uppmuntra eleven och lärarna undviker att göra elever besvikna. Hon menar att de existentiella aspekterna av betygsättningen borde ha större utrymme i hur vi förstår denna praktik.

Godkäntgränsens införande 1994 har uppmärksammats i betygsforskningen. Elisabeth Hultqvist (2011) har intervjuat 21 lärare i syfte att studera yrkets förändring. Lärarnas betygsättning är en central del i studien då lärarna beskriver hur de skolpolitiska reformerna under 1990-talet har medfört ett ökat fokus på prestation och resultat där meritvärdet blivit ett allt viktigare instrument i konkurrensen om eleverna. Lärarna beskriver hur kravet att alla eleverna ska nå målen, minst få betyget godkänt, blir problematisk i praktiken. Även i Wedins avhandling som diskuterats ovan, framkommer förutom hur lärarnas betygsättande praktik ser ut, hur godkäntgränsen upplevs som problematisk. Lärarna beskriver hur kravet på att alla elever ska få godkänt tar mycket tid i anspråk i form av stöd och uppföljning. Men ett ökat fokus på betyg beskrivs även ta tid från lärarnas mer prioriterade pedagogiska arbete med eleverna, genom ökad administration. Denna upplevelse återspeglas i flertalet av de studier som infaller under denna kategori.

Det finns ett antal antologier om betyg och bedömning där flera forskare bidrar med kapitel (Lindström, Lindberg & Pettersson 2011, Lundahl & Folke-Fichtelius 2010, Skolverket 2002). I dessa resonerar författarna om betyg och betygsättning utifrån olika aspekter av utbildning och elevers lärande. Flertalet är dock inte empiriska forskningsbidrag och redovisas därför inte i denna sammanställning. I Skolverkets publikation *Att bedöma eller döma* (2002) som innehåller flera artiklar om betygsättning, presenteras en undersökning av betydelsen av godkäntgränsen som infördes efter 1994-års betygsreform, baserad på intervjuer med 31 lärare på grund- och gymnasieskolan. Viveka Lindberg, som genomfört studien, menar att man kan säga att reformen medförde att läraruppdraget förändrades. Fokus har i ökad utsträckning flyttats från undervisningen mot elevens lärande. Lärarna uppger att betygsdiskussioner har ökat, både lärare emellan och mellan elever och lärare. På grundskolenivå uppger lärarna att även kommunikationen med föräldrarna ökade. Vad gäller kunskapskraven framkommer att lärarna i ökad utsträckning fokuserar godkäntnivån när de planerar sin undervisning. Det framkommer också att lärarna tenderar att erbjuda vissa elever möjlighet att nå de högre betygen medan andra elever främst erbjuds undervisning mot godkäntnivån. Lärarna i studien uttryckte också en oro för att tolkningen av betygskriterierna varierade, bland annat för att kriterierna var otydliga. De menade att elever genom det riskerar att bedömas på olika grunder

I svensk forskning är de skolpolitiska reformer som genomfördes under 1990-talet centrala. Framförallt förhåller sig flera undersökningar till betygsreformen 1994, då ett målrelaterat betygssystem infördes. Studierna i båda kategorierna är även förankrade i de mer övergripande förändringar som en målstyrning av skolan medfört. Betygens politiska betydelse förskjuts i och med läroplansreformen 1994 mot att användas som mått på skolkvalitet. Den goda skolan, och i förlängningen den goda läraren, definieras därigenom utifrån elevernas måluppfyllelse. I denna betydelse har betygens funktion förskjutits mot att mäta skolkvalitet. Men när betygen fortfarande i praktiken används som mått på elevens kunskapsnivå blir det ett problem när kravet på måluppfyllelse ska genomföras. Hultqvists studie, liksom flera andra (Bergman 2007, Korp 2006, Selghed 2004, Tholin 2006), har visat att lärare upplever att de får ”dra eleverna över godkäntgränsen”. Forskning har också visat att lärare sätter högre betyg utan att elevernas kunskapsnivå egentligen tillåter det (Bergman 2007, Korp 2006, Mickwitz 2011, Selghed 2004, Tholin 2006). När lärarnas betygsättning blir alltmer granskad

synliggörs praktikens ”lösning” på dilemman i relation till betygssystemet. Värt att notera är att det är det förra betygsystemet som infördes 1994 som ligger som grund i de svenska studierna. Men flera problem kvarstår trots att lokala betygskriterier inte längre krävs och trots att kunskapsmålen inte längre delas upp i uppnåendemål och strävansmål. I dagens betygssystem har lärarna att förhålla sig till olika kvalitativt skilda förmågor och viktade B- och D-betyg. Betygens betydelse i granskningen av skolor kvarstår och de nationella provens betydelse i betygsättningen har förstärkts.

# Diskussion och slutsatser

Sammanfattningsvis är det några aspekter av betygsforskning med ett lärarperspektiv som vi vill rikta uppmärksamheten mot. Som redan fallhållits är det framförallt lärarens betygsättande praktik som det forskas om. Det är alltså hur betygsättningen går till, hur och vad som betygsätts som röner störst intresse. Hur lärare över huvud taget ser på betyg och betygens inflytande och funktion i skolan framkommer i vår genomgång vara ett område som det inte forskas om i lika stor utsträckning.

I fråga om studieobjekt finns det både likheter och skillnader mellan svensk och internationell forskning. Gemensamt är att validitetsfrågan är central. Men häri ligger också skillnaden. I Svensk forskning är det relationen mellan lärarens betygsättning och styrdokumenten som dominerar perspektivet. Utanför Sverige finns också ett intresse för olika former av reglering men det är framförallt frågan om *vad* läraren bedömer som dominerar. I internationell forskning studeras lärarens summativa bedömningspraktik utifrån frågan om bedömningen sker på andra grunder än utifrån elevernas studieresultat på prov och inlämningsuppgifter. Att lärare i sin betygsättning tar hänsyn till beteenden och personliga egenskaper framhålls som ett problem för betygsättningens reliabilitet och validitet, även om vissa beteendeaspekter såsom motivation och ansträngning framhålls som viktiga i prognossyfte (Biberman-Shalev, Sabbagh et al. 2011, McMillan, Myran & Workman 2002, McMillan 2001, Resh 2009, Willingham, Pollack & Lewis 2002, Hanna & Linden 2012, Sprietsma 2013, Sun & Cheng 2014, Van Ewijk 2011, Zoeckler 2007). Därmed är det också logiskt att det i internationell forskning påträffas studier som syftar till att utveckla verktyg för att öka dessa aspekter av lärarens betygsättning.

Det finns även internationell forskning som analyserar lärarens betygsättning i relation till styrningen av den betygsättande praktiken. Det är framförallt de standardiserade kunskapstesten och betygsreformer som avser öka styrningen av betygsättningen som diskuteras. Här poängteras vikten av att lärarna samarbetar och diskuterar sin betygsättning med varandra när standardiseringen av betygen och inflytandet av *high-stakes* tester ökar (Black, Harrison et al. 2011, Brookhart 2011, Cox 2011, Hall & Harding 2002, Wyatt-Smith, Klenowski & Gunn 2010).

En iakttagelse är att studierna är mer kritiska gentemot standardisering och ansvarsutkrävning när perspektivet är lärares upplevelser och attityder. Bland annat framkommer etiska dilemman och spänning mellan pedagogiska avväganden och krav i policy (Pope, Johnson & Mitchell 2009). I två studier från Kanada (Simon, Tierney et al. 2010, Suurtamm & Kochs 2014) kritiseras också en alltmer styrd bedömning genom olika standards vid betygsättning, vilket lärare menar styr deras undervisning och snävar in deras möjlighet till formativ bedömning. Även i Sverige anläggs ett mer kritiskt perspektiv till en ökad reglering av betygsättningen när det är lärares upplevelser som ligger till grund för studien (Hultqvist 2011, Mickwitz 2011, Seger 2014). I de studier som har den betygsättande praktiken i sig som studieobjekt framstår oftast betygsättningens bristande validitet som det överhängande problemet även i svensk empirisk forskning.

Betygens inverkan på lärarens undervisning är inte centralt i forskningen utanför Sverige. Däremot framhålls kritik mot ett ökat inflytande av *high-stakes* tester och hur lärare upplever dessa meningslösa i undervisningen. Standardisering av betygsättningen och *high-stakes* tester ses som ett problem som kan komma att riskera lärarens möjlighet att verka som professionell bedömare (Hall & Harding 2002, Pope, Johnson & Mitchell 2009, van Ewijk 2011). Att lärares dagliga verksamhet på olika sätt påverkas av betygens inflytande är mer framträdande i den svenska forskning vi funnit inom området. Här är det framförallt godkäntgränsen som problematiseras men även hur betyg tar tid från lärarens pedagogiska arbete. Över huvud taget framkommer i de studier som tar upp betygens dilemman en spänning mellan styrning och kontroll och de pedagogiska aspekterna av lärarens bedömning.

# BETYGEN UR ETT SYSTEMPERSPEKTIV

> It is the political balancing act of keeping the language of policy making in accordance with what is in the child's best interests as well as satisfying the needs of government to 'know' a population so that it can act in ways which are considered to be appropriate to state building, that keeps assessment practices on the agenda of governmentality. (Meadmore 1995, s. 9)

Bedömnings- och utvärderingssystem befinner sig i centrum för frågan om hur skolan ska styras menar Meadmore (1995). I detta kapitel ska vi diskutera betyg ur ett systemperspektiv genom att sätta in betygsfrågan i det större perspektivet av bedömning och utvärdering av skolans verksamhet. I nästa kapitel, som också har ett systemperspektiv, kommer vi anlägga ett komparativt perspektiv på dessa frågor.

Systemperspektivet kan belysas utifrån flera olika nivåer såsom utifrån skolklass, skola, kommun och stat. Fokus i denna del av översikten ligger på den statliga/nationella nivån. Genom de litteraturstudier som genomförts hoppas vi kunna lyfta fram centrala insikter i relation till betygsfrågan samtidigt som vi visar på frågans komplexitet såsom de svåra avvägningar som alltid ligger bakom utformningen av ett givet betygssystem och där olika behov inom ett skolsystem kan kollidera. Ur ett styrperspektiv handlar det om att finna en utformning av ett betygssystem som på ett godtagbart sätt uppfyller de funktioner som de styrande givit det. Det kan exempelvis handla om att skapa ett system som på ett godtagbart sätt kan ge såväl återkoppling till eleven och dess föräldrar, information till styrande som fungera som fungera som urvalsgrund till högre utbildning.

Komplexiteten i frågan om betyg ur ett systemperspektiv ligger bland annat i att mätningar av prestanda i ett system alltid får en inverkan på systemet i sig (se t.ex. Ecclestone 2004, Hopmann 2003). Att varje mätning av ett system samtidigt innebär en påverkan på själva systemet är känt från naturvetenskapen, man kan aldrig samtidigt bestämma både läge och hastighet av en partikel. Effekten av mätningar på ett system bli än mer komplex när den tillämpas på samhälleliga fenomen. Sådana mätningar påverkar inte bara systemet på ofta oförutsedda sett, vi får även återkopplingseffekter, det vill säga aktörerna i systemet – elever, föräldrar, skolor, nationer osv – agerar på nya sätt givet hur de uppfattar resultaten av mätningarna (Hacking 1995, s. 369). De systemaktörer som ansvarar för mätningar och utvärderingar kan därför aldrig styra hur resultat som de betyg elever ges, testresultat från mätningar som PISA eller Skolinspektionens rapporter ska tolkas och därmed kan dessa aktörer inte heller styra vilka effekter mätningarna kommer att få på skolsystemet. Några av de utmaningar som finns och hur forskningen angripit dessa kommer att behandlas i detta kapitel.

## Metodbeskrivning

Betyg är ett vanligt förekommande inslag i utbildningssystem världen över. Detta betyder som framgått av tidigare kapitel inte att man över allt menar samma sak när man talar om betyg. För att ta USA som jämförelse brukar man grovt säga att det finns ett tvådelat resultatsystem, ett som baseras på standardiserade tester och som är riktat till administratörer och beslutsfattare, och ett som baseras på betyg och är riktat till elever och deras föräldrar. Det komparativa perspektivet belyses framför allt i nästa kapitel.

Studier av betyg ur ett systemperspektiv inom internationell forskning är få, i alla fall om man avgränsar sig till de senaste decennierna. Merparten av forskningen handlar därtill om betyg i högre utbildning, som vi även såg i kapitel 1. Sökningar på ”grading systems” i databaser som ERIC ger nästan enbart träffar som berör avvägningar och jämförelser kring betygssystems utformning i högre utbildning, och där majoriteten av studier är från 1970-talet och där man kan se en kontinuerlig minskning av antal studier inom området fram till idag. För betyg i högre utbildning är förutsättningarna andra och resultaten kan inte med enkelhet överföras på det obligatoriska skolväsendets styrning. Även om denna typ av studier innehåller viktiga kunskaper kring jämförelser mellan olika typer av betygssystem så har vi valt att avgränsa oss från studier på högre utbildning. För att närmare illustrera den förhållandevis perifera roll som forskning kring betyg spelar i nutida bedömningsforskning kan nämnas att en sökning efter betyg (”grad*”) i titeln i två av de ledande

bedömningstidskrifterna *Assessment in education: principles, policy & practice* (1994-2014) och *Educational Research and Evaluation* (1990–2014) gav åtta träffar. Hälften av artiklarna var skrivna av svenskar, de övriga fyra av författare från engelskspråkiga länder (USA, England och Australien). Resultatet justeras något när även artiklar med ”mark*” i titeln tas med. Som vi nämnt i tidigare kapitel är ”grade” eller ”grading” de termer som oftast används i internationell forskning även om det finns de som använder termerna ”mark” eller ”marking”. Termerna ”grade” och ”mark” är att betrakta som synonyma (Brookhart 2013b, s. 72). Detta säger något om betygsfrågan i Sverige och internationellt, framför allt i relation till den engelskspråkiga världen där lärarsatta betyg allt fått en allt mer begränsad funktion senaste decennierna i samband med att olika typer av centralt utformade tester fått en ökad betydelse i dessa delar av världen. Det innebär också att undersökningsobjektet för denna delstudie måste justeras för att kunna sätta in frågan om betyg ur systemperspektiv i ett internationellt sammanhang.

I syfte att bestämma ett lämpligt undersökningsobjekt och tillhörande sökord för sökningarna i den internationella forskningsdatabaserna bestämdes att vi skulle utgå från några tydligt identifierbara ledord såsom grade, mark, assessment, policy, system, equity, justice och till dessa relaterade begrepp och med hjälp av dessa manuellt gå igenom samtliga årgångar av den ledande tidskriften på området *Assessment in Education*. I första hand var det titlarna som vi gick på och verkade titeln intressant valdes artikeln ut för närmare granskning. Efter hand fann vi att även artiklar som berörde prov och tester ur ett systemperspektiv hade relevans och därför lades ett urval av dessa till listan. I detta urval gavs företräde till *editorials* som behandlade frågor relaterat till vårt undersökningsobjekt samt studier som citerats många gånger. Genomgången av *Assessment in Education* gav 44 artiklar, inklusive *editorials*. En slutsats vi kunde dra av genomgången var att flertalet av de frågor med relevans för ett systemperspektiv på betyg behandlas under begrepp som ”assessment system”. Genomgången av *Assessment in Education* kan summeras enligt följande:

- Tydligt att tidskriften tillkommer i en tid när ”test- och utvärderingskulturen” har etablerats i styrning av skolan och att tidskriften generellt belyser negativa konsekvenser av det tilltagande testandet.
- Tidskriften följer utvecklingen av internationella storskaliga tester såsom PISA och TIMSS mycket noga.
- Betyg utgör ett perifert område i tidskriften, bara en handfull studier har betyg som huvudfokus.
- Även om betyg inte spelar stor roll i den diskussion som förs i tidskriften kan man säga att synen på betyg bland utbildningsforskare i Sverige speglar synen på high-stakes tester och ”accountability” inom den internationella utbildningsforskningen. Man återfinner med andra ord samma typ av argument men i relation till ett annat objekt (tester istället för betyg).
- Tidskriften publicerar kontinuerligt genomgångar av bedömningssystem i länder världen över. Här kan man få bra inblick i policyfrågor och vilken roll betyg spelar i bedömnings- och utvärderingssystem runt om i världen (mer om detta i nästa kapitel).

Vi valde efter genomgången att justera undersökningsobjektet från ensidigt fokusera betyg ur ett systemperspektiv till att mer övergripande belysa bedömnings- och utvärderingssystem och den roll betyg kan ges i dessa. Detta innebar att vi istället för att sammanställa all befintlig forskning i relation till några specifika frågor om betyg ur systemperspektiv fått sänka ambitionen till att ge en överblick över viktiga perspektiv inom området. Detta gäller dock enbart den forskning som behandlar förhållanden i andra länder än i Sverige, när det gäller forskning på svenska förhållanden så har vi kunnat ge en heltäckande sammanställning av den forskning som finns på området.

*Tabell 5. Sammanställning av sökning och antal utvalda artiklar.*

| **Sökningar databas/söksträng** | **Sökningen/ genomgången gav** | **Därav utvalda (abstract och titel)** |
|---|---|---|
| Genomgång av *Assessment in education : principles, policy & practice*, samtliga årgångar (1994–2014) | 44 | 44 |
| ERIC/”effect* grading education”<br>Avgränsning: scholarly journals; abstract. | 52 | 6 |
| ERIC/”effect* mark* assessment education*”<br>Avgränsning: scholarly journals; abstract. | 134 | 9 |
| ERIC/GPA<br>Avgränsning: GPA i titeln samt endast tidskrifter med “education*” i titeln. | 52 | 2 |
| ERIC/”’grade* OR mark*’ AND ’policy OR system’ AND ’justice OR equality’”<br>Avgränsning: scholarly journals; abstract. | 184 | - |
| ERIC/”’high AND stakes’ AND ’assessment OR examination’ AND effect* AND ’learning OR knowledge’”<br>Avgränsning: scholarly journals; abstract. | 180 | 4 |
| ERIC/”(’examination system’ OR ’assessment system’ OR ’grading system’) AND education NOT ’higher education’”<br>Avgränsning: scholarly journals; abstract. | 171 | 11 |
| ERIC/”’assessment OR grade* OR examination*’ AND ’system OR policy’”<br>Avgränsning: scholarly journals; abstract. | 765 (gick igenom 100 första) | 5 |
| ERIC/”’assessment OR grade*’ AND ’system OR policy OR effect*’ AND ’equity OR justice OR equality OR fairness’ NOT ’higher education’”<br>Avgränsning: scholarly journals; abstract. | 438 | 0 |
| LIBRIS/”likvärd* betyg*” (21 träffar) samt ”rättvis* och betyg*” (8 träffar) | 29 | 29 |
| **Totalt utvalda** | | **109** |

**Sammanställning av använda sökord och synonymer:**

education
effect*
grading OR grade* OR mark*
assessment OR examination
gpa
policy OR system
justice OR equality
high stakes
learning OR knowledge
”examination system” OR ”assessment system” OR ”grading system”
equity OR justice OR equality OR fairness

På basis av de utvalda artiklarnas referenslistor tillkom ytterligare studier. Efter sökningarna var genomförda hanterades alla artiklar, böcker och rapporter som ett och tematiserades baserat på deras innehåll. Följande tre huvudteman och tillhörande underteman identifierades.

1) High-stakes och mer allmänna resonemang om bedömnings- och utvärderingssystem.
2) Betyg ur rättvise- och jämlikhetsperspektiv – principiella övervägande (underteman: Betygssystem i ljuset av teorier om rättvisa; Kunskapsfrågan i relation till sociala kategorier; Lagar och regler i ett rättvist betygssystem).
3) Betyg som kunskaps- och urvalsmått i svensk skola (underteman: Betyg som kunskapsmått ur systemperspektiv; Betyg och urval till högre utbildning; Betyg som förutsägelse av studieavhopp respektive studieframgång).

***Tabell 6. Sammanställning över de studier som ingår i den slutliga analysen, uppdelat per tema.***

| | **Antal studier (varav svenska studier)** |
|---|---|
| Betyg ur rättvise- och jämlikhetsperspektiv | 14(2) |
| Betyg som kunskaps- och urvalsmått i svensk skola | 19(12) |
| *High-stakes* och mer allmänna resonemang om bedömnings- och utvärderingssystem | 36(9) |
| **Totalt** | **69(23)** |

De artiklar som hör till *high-stakes* temat bedömdes inte relevanta för översiktens huvudfokus utan de har huvudsakligen arbetats in i de inledande avsnitten, för att ge en bättre kontext till frågan om betyg ur systemperspektiv samt för tydliggöra centrala distinktioner inom området. Av de nio svenska studier som finns med i detta tema är åtta statliga utredningar och rapporter som är viktiga för systemperspektivet på betyg, men som inte faller inom våra urvalskriterier (forskning). Den nionde texten är Lundahls bok *Varför nationella prov?* från 2009. De studier som hör till de översta två teman i tabellen beskrivs närmare under respektive avsnitt. Som framgår är det framför allt frågan om betyg som kunskaps- och urvalsmått som varit i fokus för svensk forskning om betyg ur systemperspektiv. Detta betyder inte att frågor om rättvisa och jämlika betyg inte hanterats av svensk forskning – snarare har det varit en mycket stor fråga som berörts i flera rapporter och utredningar senaste dryga decenniet och där det konstaterats att betyg kan variera från skola till skola och från lärare till lärare – men nästan inga av de studier som finns på detta område har satt frågorna i ett teoretiskt perspektiv, där övergripande frågor om rättvisa och jämlika bedömnings- och utvärderingssystem hanterats. Det är denna typ av mer teoretiskt och kvalitativt inriktade studier som behandlas under temat ” Betyg ur rättvise- och jämlikhetsperspektiv” och som vi upplever viktiga att lyfta fram då de är ett perspektiv som saknas i nuvarande forskning på svenska betygssystem. Den mer begränsade frågan om likvärdiga betyg behandlas i avsnittet ”Betyg som kunskaps- och urvalsmått i svensk skola”.

# Betyg ur ett systemperspektiv – centrala distinktioner

Den som har kontroll över bedömning och utvärdering i ett skolsystem har också makt över vad som sker i undervisningen. Samtidigt har skolan visat sig vara ett svårstyrt område. Hopmann (2003) menar att det funnits två olika traditioner när det gäller utvärdering och kontroll i västerländska utbildningssystem. Den ena traditionen har dominerat det kontinentala Europa och där har fokus legat på processutvärdering och processkontroll, kontroll genom läroplaner, regler och dylikt, det vill säga av processer och ramar för verksamheten snarare än utfall. Den andra traditionen finner man inom den engelskspråkiga världen där det varit fokus på utbildningssystemets resultat och i linje med det en utvärdering och kontroll av dessa genom kunskapsstandarder, examens- och inträdesprov m.m. Vad som skett de senaste decennierna, menar Hopmann

(2003, s. 475), är att den engelskspråkiga världen infört mer av processkontroll medan det kontinentala Europa infört mer av resultatkontroll. Enligt Hopmann var det under 1980-1990-talen som båda dessa traditioner hamnade i kris och de kom att rikta blickarna mot varandra. Istället för att överge det ena för det andra så har vi fått utbildningssystem med både och. Vad detta kommer leda till är ännu för tidigt att säga, menar Hopmann.

För svensk del talas det ofta om övergången från regel till målstyrning, som om vi gick från den ena traditionen till den andra, men detta bör alltså nyanseras. Vad det handlar om är snarare att process- och regelstyrningen mildrats något medan resultatkontrollen ökat. Det svenska systemet är att betrakta som ett mellanting av dessa båda traditioner enligt Waldow (2014).

Sammansmältningen av de båda styrtraditionerna är en del av utvecklingen av skolsystem som kan iakttas internationellt, en annan är en kvalitativ förändring i kontrollen och utvärderingen som gäller framför allt resultatsidan och där har vi sett en utveckling mot ökad användning av *high-stakes* tester och att dessa blivit *high-stakes* för fler typer av aktörer (Amrein och Berliner 2002, Mehrens 1998). En ny *high-stakes* företeelse är PISA och liknande internationella tester, de är *low-stakes* för de elever som tar dem men *high-stakes* för politiker och regeringar (Stobart och Eggen 2012). Denna utveckling har inom utbildningsforskningen följts av en parallell ökning av forskning på bedömning- och utvärdering med uppkomst av nya tidskrifter såsom *Assessment in education: principles, policy & practice*. Forskningen har bland annat riktat stark kritik mot att detta inneburit en implementering av en ny typ av ”Taylorism” i utbildningsväsendet (Au 2011, Hopmann 2013). Det kanske tydligaste exemplet är *No Child Left Behind* (NCLB) som lanserades 2001 i syfte att skapa ett system för ständig förbättring i USAs skolväsende. Systemet har därtill fått en stor internationell spridning (Hopmann 2013). Australien är ett av de länder som implementerat liknande system (Klenowski och Wyatt-Smith 2012). Samtidigt är fenomenet inte nytt utan det har funnits liknande system tidigare, exempelvis i Kanada (Jacob 2005). Analyser av *high-stakes* skolsystem är ett hett område inom den internationella utbildningsforskningen (se t.ex. Au 2007, Braun 2004, Nichols, Glass et al. 2006, Plank och Falk Condliffe 2013, Wyse och Torrance 2009). I samband med analyser av denna utveckling har vikten av att även undersöka den kultur som omger *high-stakes* betonats (se t.ex. Daugherty 2008). När det kommer till effekterna av NCLB är detta ett stort forskningsområde där resultaten pekar i olika riktningar, att det kan ha skett utveckling inom vissa områden men inte inom andra. Det verkar som NCLB har lett till en ökning av elevers akademiska prestationer i matematik för vissa elevgrupper, men inga mätbara effekter på läsning för någon grupp (Dee och Jacob 2011). NCLB tycks även lett till minskad motivation bland elever och ökat fusk inom utbildningssystemet. Systemet i sig är därtill kostsamt. Se vidare kapitel 4.

Hittills har vi tecknat det större sammanhang i vilket varje betygssystem måste förstås när de behandlas ur ett styrperspektiv. I det följande ska vi peka på några viktiga dimensioner i begreppet betygssystem. När man talar om betygssystem brukar man lyfta fram tre olika typer av system (se t.ex. Gustafsson 2006, Neumann, Trautwein et al. 2011, SOU 1977:9): individrelaterat; normrelaterat; målrelaterat. Valet av betygssystem stannar dock inte vid ett val mellan dessa tre utan innefattar en mängd andra avvägningar. De Luca (1994, 8) menar att det bakom varje givet betygssystem ligger en mängd beslut gällande exempelvis:

- Om det ska vara norm- eller målrelaterat;
- Om det ska skötas externt eller baseras helt på lärares bedömningar eller vara en mix av de båda;
- Vilka typer av hjälpinstrument som ska finnas i systemet (diagnostiska prov, standardiserade prov, bedömningsstöd, inspektioner mm);
- Vilka syften betygssystemet ska uppfylla.

De olika syften som brukar nämnas i relation till betyg är:

1) Urval/antagning/uppflyttning till nästa årskurs (antagningskrav eller urvalskriterium till utbildning);
2) Information (till framför allt elever och deras föräldrar);
3) Motivation (eller disciplinering);
4) Diagnostisk (belysa styrkor och svagheter);
5) Certifiering (avklarad utbildning eller uppnådd kunskapsnivå);
6) Information på lokal (klass, skola, och kommun) och nationell nivå om utbildningens resultat.

Här bör man notera att syften både kan överlappa och motverka varandra (Gustafsson 2006, Neumann, Trautwein et al. 2011). Till detta kan läggas att betyg och motsvarande *high-stakes* tester ofta används som verktyg av styrande i syfte att nå en effektivare implementering av en reform, men att det är svårt att utvärdera effekterna av detta antagande (McDonnell 1994; Spillane 1999).

Andra centrala dimensioner i betyg utifrån systemperspektivet är betygens betydelse för elev, lärare, skola eller skoldistrikt. Betydelsen kan vara: *low-stakes* eller *high-stakes* för någon eller flera av dessa. Och på vilken nivå makten över betygssättning ligger: lokalt (den enskilde läraren) eller centralt (i de fall betyg sätts utifrån resultat på nationella prov).

I vissa länder finns det en spänning mellan lokal, privat och central styrning av skolformer och där kan makten över och regelverket kring skolformer skifta från en nivå och aktör till en annan. Som exempel på detta kan man ta USA där man sett en utveckling där företag driver fram tester i huvudsak utformade för att passa antagning och urval till högre utbildning men som får stor inverkan på vad lärarna undervisar om i det obligatoriska och lokalt styrda skolsystemet. Då testerna normalt inte har en tydlig korrelation till den lokala läroplanen gör det att fokus i elevers lärande i många fall hamnar på annat än det som står angivet i läroplanen (se t.ex. Au 2007, Stobart 2009). Effekterna av ett betygssystem är alltså ingalunda begränsade till den skolform de utformats för, utan systemen finns alltid i en större kontext av exempelvis mer eller mindre uttalade krav från närliggande skolformer vilket påverkar hur väl betygssystemet fungerar som styrinstrument.

## Betyg ur rättvise- och jämlikhetsperspektiv – principiella överväganden

I Sverige är frågan om likvärdig bedömning och betygssättning central och har varit så en längre tid. Ofta likställs frågan om likvärdig bedömning med frågan om samma kunskaper ger samma betyg oavsett vilken lärare eleven har, vilken skola han eller hon går på eller oavsett elevens kön, klass och etnicitet. Men för att överhuvudtaget kunna diskutera likvärdighet i relation till ovanstående aspekter så har man gjort en mängd antaganden vad gäller betygssystemets utformning ur rättvisesynpunkt och det är denna typ av antaganden som belyses i detta avsnitt. Närmare bestämt kommer vi att närma oss frågan om rättvis och jämlik betygssättning ur ett teoretiskt perspektiv. Huvuddelen av den forskning som redovisas i avsnittet tar sin utgångspunkt i rent teoretiska resonemang men några baserar även sina resultat på empiriskt material, exempelvis enkätstudier i syfte att visa på olika uppfattningar om rättvis betygssättning inom systemet, eller analyser av tester, läroplaner och läromedel i syfte att belysa hur de kunskaper som värderas kan gynna eller missgynna olika grupper.

En huvudslutsats vi drar, och som vi framhöll även i relation till lärarperspektivet i kapitel 2, är att varje rättvis och likvärdig betygssättning balanserar mellan två närmast motstridiga krav: opartisk bedömning (som erhålls av exempelvis standardiserade prov med litet tolkningsutrymme i analysen av svaren, såsom flervalsprov) och mångsidig och individualiserad bedömning (som åstadkoms lärarnära och därmed kan skilja sig beroende på lärarens erfarenheter, ålder, etnicitet och kön samt den bedömda elevens kön, klass, ålder, etnicitet och eventuella funktionsnedsättning). Dessa motstridiga krav har diskuterats åtminstone sedan 1700-talet (Lundahl 2006). Båda polerna har sina svagheter: standardiserade prov ger sällan en riktigt god bedömning av den enskilde individen, som kan ha en bra eller dålig dag, eller som kanske har sina styrkor inom andra områden än de som just tas upp av det aktuella testet. Även standardiserade prov kan gynna vissa grupper framför andra. Ett rättvist och likvärdigt betygssystem måste därför försöka hitta en balans mellan dessa båda poler och ser man runt om i världen så löses detta på olika sätt. Många system har, precis som Sverige, inslag av såväl lärares bedömningar som normerande inslag i form av centralt administrerade tester. Men det kan se väldigt olika ut hur exakt denna balansgång löses. För att ta ett exempel i Alberta, Kanada, så baseras slutprov till 50 procent på lärarbedömningar (vilka i sin tur stöds på olika sätt) och till 50 procent på någon form av centralt administrerat examensprov (Scott, Webber et al. 2014). En viktig slutsats är att inget system är perfekt, utan att varje system måste kännetecknas av öppenhet/transparens och ständig genomlysning och omprövning/förbättring för att frågor om rättvisa inte ska hamna i skymundan:

> The best defense against inequitable assessment is openness. Openness about design, constructs and scoring, will bring out into the open the values and biases of the test design process, offer an opportunity for debate about cultural and social influences, and open up the relationship between assessor and learner. These developments are possible, but they do require political will. (Gipps 1999, s. 385)

Utvecklingen mot allt högre grad av privata aktörer inom området utvärdering och bedömning är ett tydligt hot mot detta, och alltså mot möjligheten att skapa rättvisa system. Lika så den överstatliga utvecklingen med PISA, som bara ger begränsad insyn i utformning av sina tester och därmed begränsade möjligheter till kritisk belysning av dessa utifrån deras eventuellt systematiska fel i relation till kön, klass, etnicitet och funktionsnedsättning.

”Equity” och ”Justice” är vanligt förekommande ord i de artiklar som analyserats rättvis bedömning. Resultaten presenteras i tre underavsnitt. Vi återkommer emellertid till frågan om rättvisa bedömningar även i kapitel 4 men då ur ett mer komparativt perspektiv.

**Tabell 7. Sammanställning av analyserade studier temat ”Betyg ur rättvise- och jämlikhetsperspektiv”.**

| **Effektstudier eller forskningsöversikter som baseras på effektstudier** | **Andra studier**<br>*** markerar att studien baseras på empiriskt underlag** |
|---|---|
| - | Baker och O'Neil Jr, 1994; Chilisa 2000*; Cumming 2008*; Deutch 1979; Gewirtz, 1998; Gipps 1995; Gipps 1999; Korp 2006*; Nieuwenhuis 2010; Resh 2009*; Scott, Webber et al. 2014*; Sikes och Vincent 1998; Stobart 2008; Waldow 2014* |

## Betygssystem i ljuset av teorier om rättvisa

Tidskriften *Assessment in Education: Principles, Policy & Practice* har ägnat området bedömning och rättvisa stort utrymme genom åren. Stobart (2008, s. 121) skriver i inledningen till ett specialnummer på temat att rättvisa kan ses som en grundläggande premiss i varje bedömning: ”One of the key requirements of any assessment is that it seeks to be fair.” Samtidigt innebär inte detta att varje bedömning blir rättvis. Sett till systemnivån kan man konstatera att varje konstruktion av ett bedömnings- och utvärderingssystem får konsekvenser för individer och grupper av individer. Gipps (1995) menar att det inte finns någon neutral konstruktion utan att varje konstruktion av ett sådant system innebär potentiellt negativa konsekvenser för vissa elevgrupper. Av denna anledning måste rättviseaspekter alltid bevakas och diskuteras och justeringar göras i syfte att förbättra systemet.

Styrningen av ett skolsystem, liksom dess själva utformning, bör överhuvudtaget ta sin grund i teorier om rättvisa hävdar forskare (t.ex. Gewirtz 1998). Sikes och Vincent (1998) menar därtill att ständiga förändringar i betingelser för utbildning såsom nya aktörer, nya ekonomiska premisser, globalisering m.m. gör att frågan om rättvisa hela tiden får ny aktualitet och måste rekonceptualiseras (jfr Nieuwenhuis 2010). När det gäller rättvisa och betygssystem mer specifikt så har det varit i fokus för ett flertal studier (t.ex. Deutsch 1979, Korp 2006, Resh 2009, Südkamp. Kaiser & Möller 2012, Waldow 2014).

En ofta förekommande egenskap som nämns i relation till allmän utbildning är att den ska fungera som ett verktyg för social förändring, det vill säga syfta till att utjämna skillnader mellan människor med avseende på social klass, genus, kön, etnicitet och funktionsnedsättning (Deutsch 1979, Korp 2006, Lundahl 2006). Olika betygs- och bedömningssystem har olika egenskaper i relation till rättvisa. I de fall betyg är *high-stakes* för elever, så som i Sverige, kan man se betyg som utbildningssystemets grundläggande valuta genom vilket meriter uttrycks (Deutsch 1979). Här bör man skilja mellan hur rättvis den slutliga fördelningen av betyg är samt hur rättvis själva utformningen av betygssystemet i sig är, såsom vilka medverkande aktörer som finns och vilka deras roller är samt hur processerna är utformade m.m. Man brukar i detta sammanhang skilja mellan

distributiva och proceduriella perspektiv på ett rättvist betygssystem (Deutsch 1979, Waldow 2014). Proceduriell rättvisa söker rättvisa procedurer, grovt: om alla procedurer som ingår i fördelningen av betyg kan anses rättvisa så anses betygen som resulterar vara rättvisa. Det distributiva perspektivet fokuserar främst på att resultatet, betygen och betygsfördelningen, ska vara rättvis.

Waldow (2014) använder perspektivet proceduriell rättvisa när han jämför examinationssystem i England, Tyskland och Sverige. Han tar hjälp av Hoffmans distinktion mellan process- och resultatkontroll som vi beskrev tidigare i kapitlet, och menar att England är ett tydligt fall av resultatkontroll, Tyskland ett fall av process medan Sverige intar ett mellanting. Waldow drar slutsatsen att medan samtliga dessa tre länder utger sig för att vara meritokratier så ger de uttryck för tre mycket skilda uttolkningar av proceduriell rättvisa. Följaktligen kommer rättvisa betyg innebära tre olika saker i dessa system, vilket accentueras ytterligare då dessa länders bedömningshistoria och deras sociala kontext tas med i analysen (Waldow 2014, s. 339, se även kap. 4). När det gäller rättviseaspekter så betonas alltså även vikten av att det enskilda landets betygssystem analyseras i sitt historiska och sociala sammanhang i forskningen. Några principiella frågor att ställa sig i relation till ett rättvist betygssystem är (Deutch 1979, s. 400; jfr Stobart 2005a):

- Vilket innehåll, kvalitet och kvantitet av ”ont” respektive ”gott” distribueras i systemet?
- Vilka människor och aktörer är involverade? Till vem distribueras betyg och av vem?
- Hur och när sker distributionen? Öppet eller slutet? Är den bedömde införstådd i de konsekvenser bedömningen kan få?
- Vilka grundläggande värden speglar distributionen av det distribuerade?
- Vilka effekter får distributionen?

Stobart (2005a) nyanserar några av ovanstående frågor i relation till multikulturella samhällen:

- Hur beaktas kulturell och språklig mångfald?
- Hur speglar innehållet i det som bedöms olika gruppers erfarenheter?
- Hur svarar bedömningsmetoder på kulturell mångfald bland de bedömda?
- Hur följs de olika gruppernas prestationer upp och hur återkopplas denna information till systemet?

Belysningen av rättviseaspekter av ett betygssystem måste även inbegripa studier av hur lärare, elever, föräldrar, rektorer, beslutsfattare med flera uppfattar frågor om rättvis bedömning, inte minst för att dessa uppfattningar samspelar med systemet (Scott, Webber et al. 2014). Det visar sig vid studier av olika aktörers uppfattningar att det ofta föreligger olika uppfattningar om huruvida ett system är rättvist eller inte. Det finns spänningar i frågor om rättvisa och likvärdighet i relation till inklusion, särskilda behov och begåvning (ibid., s. 53). En kanadensisk storskalig enkätstudie av Scott, Webber et al. (2014) visade exempelvis att på frågan om skolor gör ett bra jobb i att bedöma begåvade elever så svarade endast 52,8 procent av lärarna, 49,2 procent av eleverna och 28,8 procent av föräldrarna att de höll med (ibid., s. 57). Scott, Webber et al. (2014) drar bland annat slutsatsen att det krävs en ökad medvetenhet hos aktörer inom ett utbildningssystem för de konsekvenser bedömningar får för individuella elever och för deras familjer. Om vi drar detta ett steg längre kan vi notera att det finns forskare som menar att utformningen av systemet i sig måste vara sådant att det odlar en känsla för rättvisa och likvärdighet hos systemets aktörer:

> It may suggest that a pre-condition for adjustment or change in teachers' pedagogical practices is not necessarily a "change of hearts" (teachers' views, beliefs and attitudes), but rather the creation of systemic conditions that encourage certain behaviors, which may also motivate attitudinal changes. (Resh 2009, s. 323)

## Kunskapsfrågan i relation till sociala kategorier

Kunskapsfrågan har framhållits som central för diskussionen om ett rättvist utbildningssystem (Gipps 1995). Länge trodde man att objektiva bedömningar skulle ta bort alla former av orättvisor (ibid., s. 273). Detta är fel väg att gå menar Gipps (1995), istället för en tro på objektiv bedömning bör man sträva efter att på systemnivå synliggöra hur de kunskaper som bedöms gynnar eller missgynnar olika elevgrupper, skolor etc., menar hon. På så sätt skapas en transparens i systemet som kan ligga till grund för en utveckling mot högre grad av likvärdig och rättvis bedömning:

> There is no such thing as a fair test, nor could there be: the situation is too complex and the notion simplistic. However, by paying attention to what we know about factors in assessment, administration and scoring, we can begin to work towards tests that are more fair to all the groups likely to be taking them, and this is particularly important for assessment used for summative and accountability purposes. (Gipps 1995, s. 279).

Kunskapsfrågan har i forskningen belysts utifrån frågor om sociala kategorier som genus, klass, etnicitet, funktionsnedsättning (se t.ex. Baker och O'Neil Jr 1994, Chilisa 2000). I det följande redogörs för några mer allmänna aspekter av denna typ av överväganden baserat på Gipps (1995).

En första fråga man bör ställa sig är om det är själva mätinstrumentet som ger upphov till skillnader mellan grupper eller om det är faktiska skillnader mellan grupper som observeras. Enligt Gipps (1995, s. 280) är det sannolikt så att varje betygssystem innehåller en blandning av båda dessa aspekter och för att nå ett rättvist betygssystem bör man minimera det förra samtidigt som man lär sig förstå och förklara orsakerna till det senare. A och O bör vara att synliggöra den kunskap som faktiskt mäts i syfte att tydliggöra spelreglerna. Detta kan ske på olika sätt, genom standardiserade prov eller genom att förespråka målrelaterade system. Ett tydligt exempel från historien är det faktum att flickor länge presterade sämre än pojkar på naturvetenskapliga prov. Detta försökte man finna förklaringar till i form av kognitiva brister hos flickor. Det var inte förrän man insåg att det var uppgifternas utformning som var orsaken till skillnaderna som man fick rätt verktyg för att hantera detta, testerna hade nämligen varit utformade så att de gynnade vita medelklasspojkar (Gipps 1995, s. 274). Minoritetsgrupper och barn med funktionsnedsättning är särskilt viktiga att uppmärksamma. Det kan med tanke på dessa grupper vara viktigt att erbjuda bedömningshjälp utifrån; andra lärare eller tester som kan ge nya perspektiv på individens kunskap och som kan utmana de fördomar som enskilda individer bär på (Gipps 1995). I relation till dessa frågor är det även viktigt att behandla frågan om vilka vägar till kunskap som kan eller inte kan betraktas som likvärdiga.

Mer allmänt kan resultat på nationellt standardiserade prov skilja sig från lärarens bedömningar. Detta behöver inte betyda att det ena eller det andra är rätt, men måste ge upphov till reflektioner kring vad som bedöms och på vilka grunder. Exempelvis kan man genom detta upptäcka om vissa grupper eller enskilda elever systematiskt missgynnas av endera lärares bedömningar eller externa prov. Utvecklingen av dessa rättviseaspekter i bedömnings- och utvärderingssystem måste följas noggrant eftersom förutsättningarna hela tiden ändras (Gipps 1995). Slutligen, orättvisor till följd av *high-stakes* bedömningar såsom betyg slår troligtvis olika hårt i olika stadier av utbildningssystemet, exempelvis kan man räkna med att olika gruppers bakgrund slår igenom starkare i tidiga skolår, med konsekvenser för rättvis bedömning – dvs den bör utformas annorlunda i tidiga år jämfört med sena år (Gipps 1995, s. 274).

## Lagar och regler i ett rättvist betygssystem

Cumming (2008) har undersökt rättsfall i USA, Australien och England som berör rättvisefrågor i relation till bedömning. Hon menar att kraven på den rättsliga apparaten står i relation till konsekvenserna för den bedömde (ibid., s. 132). Ett exempel hon tar gäller de ökade kraven på de kunskaper som elever minst ska ha uppnått under sin obligatoriska utbildning, och som är relevant för svensk del när det gäller gränser för godkända prestationer. Kunskapskraven har ökat i många länder på senare tid och Cumming menar att ökade krav på elevers minsta godtagbara kunskapsnivå ställer likaledes ökade rättsliga krav på att systemet (skolan/staten/kommunen/läraren) tillhandahållit eleven de resurser han eller hon skäligen har rätt till

(Cumming 2008, s. 127). Detta kan röra grundläggande förhållanden i skolor liksom elevers tillgång till extra resurser samt system för att säkerställa att varje elev får ta del av det han eller hon har rätt till givet de kunskapskrav som ställs. Cumming (2008, s. 125) identifierar följande rättsliga utmaningar i relation till elevens möjligheter att lära sig:

- Att eleverna i sin undervisning har tagit del av det innehåll de ska bedömas på.
- Att eleverna har fått tillräcklig information om eventuella ändringar i bedömningssystemet.
- Att tillräckligt med resurser har avsatts för att möjliggöra för elever att lära det som de ska bedömas på.

## Betyg som kunskaps- och urvalsmått i svensk skola

Frågan om lärarsatta betyg ställs på sin spets i de bedömnings- och utvärderingssystem där betyg har som syfte att ge information till styrande aktörer om kunskapsnivå och kunskapsutveckling samt där de används för urval till högre utbildning. I Sverige är det nuvarande betygssystemet utformat för att ge såväl information om kunskapsnivåer bland elever som att utgöra grund för intag till gymnasieprogram och högre utbildning. I detta avsnitt ska vi redogöra för forskning som undersökt hur väl olika svenska betygssystem uppfyllt funktioner som dessa på systemnivå och då framför allt utifrån den nationella styrnivån. En handfull internationella studier kommer refereras för att bredda perspektivet, men fokus ligger på studier av svenska betygssystem. Att det inte varit möjligt att hitta mer än en handfull relevanta studier i den internationella forskningen beror troligtvis på att lärarsatta betyg spelar en liten roll i den engelsktalande världen när det gäller syften som att mäta nationell kunskapsnivå och fungera som urval till högre utbildning. Det kan också bero på att länder där betyg fortsatt spelar en high-stakes roll i relation till dessa syften inte publicerar dessa studier på engelska.

Grovt sett kan vi säga att detta avsnitt redogör för svensk forskning om likvärdiga betyg ur ett systemperspektiv. Frågan om likvärdiga betyg har sedan införandet av det målrelaterade betygssystemet i mitten av 1990-talet varit en av de stora skolfrågorna och framstått som en av de stora utmaningarna för det svenska skolsystemet, vilket vi redan varit inne på i tidigare kapitel. Frågan har fått stor uppmärksamhet från såväl forskare, skolmyndigheter, politiker, elever och föräldrar som i media.

De svenska betygssystemen och till dessa kopplade frågor om likvärdighet har genom åren varit föremål för mängder av utredningar och studier. Exempelvis har frågan om vilket betygssystem – normrelaterat, målrelaterat eller individrelaterat – som är det bästa för svensk skola varit föremål för flera statliga utredningar (SOU 1942:11, SOU 1977:9, SOU 1992:86). Frågan om likvärdiga betyg har därtill utretts vid ett flertal tillfällen av ett flertal statliga styrorgan (se t.ex. Riksrevisionen 2004, 2011, Skolinspektionen 2014, Skolverket 2007, 2009). För nyligen genomförda forskningsstudier och översikter på området se Gustafsson (2013), Gustafsson, Cliffordson et al. (2014), Böhlmark och Holmlund (2011) och Holmlund, Häggblom et al. (2014).

Det faktum att man på nationell nivå redan tar frågan om likvärdig betygssättning på största allvar, att det redan finns god kompetens på systemnivå kring dessa frågor samt att det finns ett pågående arbete med att öka likvärdigheten gör att ambitionen med detta avsnitt inte så mycket blir att fördjupa och syntetisera kunskap som att tydliggöra forskningsläget och ytterligare betona några av de utmaningar som föreligger i relation till det nuvarande betygssystemet. Likvärdighet i det svenska betygssystemet kommer i detta avsnitt att belysas utifrån följande likvärdighetsdimensioner:

- Hur väl betygssystemet fungerar kompensatoriskt i syfte att jämna ut sociala skillnader i samhället.
- Hur väl betygssystemet är utformat för att sålla ut de som klarar sig bäst i högre studier.
- Hur väl betygssystemet ger information om elevers kunskapsnivå och kunskapsutveckling.

Situationen och utmaningarna för dagens målrelaterade betygssystem kommer att ges viss historisk belysning i syfte att visa på effekter av lärarsatta betyg i olika betygssystem i Sverige.

*Tabell 8. Sammanställning av analyserade studier under temat "Betyg som kunskaps- och urvalsmått i svensk skola".*

| **Effektstudier eller forskningsöversikter som baseras på effektstudier** | **Andra studier**<br>*** markerar att studien baseras på empiriskt underlag** |
|---|---|
| Bowers 2010; Bowers, Sprott et al. 2013*; Brookhart 2013b*; Cliffordson 2008; Gustafsson 2006; Gustafsson et al. 2014*; Neumann, Trautwein et al. 2011; Sawyer 2013; Thorsen 2014; Thorsen och Cliffordson 2012; Vlachos 2010; Wikström 2005b | Andersson 1991*; Hyltegren 2014*; Lundahl 2006*; Stobart 2005b; Tholin 2006*; Widén 2010*; Wyatt-Smith, Klenowski et al. 2010* |

## Betyg som kunskapsmått ur systemperspektiv

Diskussionen kring huruvida svensk skola ska ha ett norm- eller målrelaterat betygssystem har i princip pågått sedan 1930-talet (Andersson 1991, s. 53). Från att ha haft ett normrelaterat system, vilket började införas under 1950-talet, gick Sverige i början av 1990-talet över till ett målrelaterat system, utformat i början av 1990-talet och implementerat under andra halvan av 1990-talet. Gymnasieskolan blev samtidigt kursutformad, vilket innebar att betyg gavs för varje genomförd kurs och sammanräknades i slutbetyget. De svenska betygssystemens historia och förändring finns utförligt behandlat i flera avhandlingar (Andersson 1991, Hyltegren 2014, Lundahl 2006, Tholin 2006, Widén 2010).

Länge var det så att målrelaterade betyg ansågs för otillförlitliga för att kunna införas. De normrelaterade betygen infördes bland annat av det skäl att man i studier funnit stark evidens på att lärare var bra på att rangordna elever men sämre på att bestämma elevers kunskapsnivå, därför lät man kunskapsnivån regleras genom centralt anordnade prov i ett urval ämnen (Gustafsson 2006). Med tiden kom det normrelaterade betygssystemet att utsättas för massiv kritik såsom att det uppmuntrade till konkurrens, att lärare inte följde de anvisningar de gavs, att normeringen inte stöddes av centralprov i alla ämnen m.m. (Gustafsson 2006, Lundahl 2006, Tholin 2006, Widén 2010). Kritiken bidrog till beslutet att överge det normrelaterade för det målrelaterade betygssystemet. Andra bidragande faktorer till denna övergång var den kunskapsstandardrörelse som växte fram internationellt under 1980- och 1990-talen (Porter 1993). I relation till kunskapsfrågan var den starkaste kritiken mot de normrelaterade betygen att de sa för lite om elevers faktiska kunskaper (Gustafsson 2006).

I både det normrelaterade och det målrelaterade betygssystemet har betyget ytterst satts av läraren. Den starka kontrast som signaleras av de båda namnen motsvaras inte av hur de båda systemen rent faktiskt utformats. Det normrelaterade systemet styrde inte betygssättningen på individnivå, utan enbart på gruppnivå. Man brukar kalla denna typ av normrelaterade system för gruppnormerande. Det fanns alltså frihet för läraren att inom gruppen sätta de betyg han eller hon ansåg lämpliga, även om det fanns vissa direktiv gällande exempelvis fördelningen av betyg, vilka dock lättades upp efter hand (Gustafsson 2006, Lundahl 2006). I det målrelaterade systemet finns på samma sätt som det fanns i det normrelaterade stöd för lärare i att bedöma elevernas kunskapsnivå i form av centralt administrerade (nationella) prov, vilka alltså bland annat har till syfte att just normera betygen, även om det inte är huvudsyftet med proven och de har en mängd andra funktioner att fylla jämfört med de standardprov som användes i det normrelaterade systemet (Gustafsson, Cliffordson et al. 2014, jfr Lundahl 2009).

Både det normrelaterade och det målrelaterade betygssystemet använde sig alltså av normerande prov. Dessutom var de normrelaterade betygen relaterade till kunskaper hos eleverna, de avsåg spegla de mål och det innehåll som läroplanen specificerade. Skillnaden mellan de båda betygssystemen kan alltså verka större än den faktisk är om man bara ser till namnen. Trots detta finns viktiga och intressanta skillnader på systemnivå. När det gäller den senaste betygsreformen 2011 finns ännu inga effektstudier genomförda, men forskare pekar på att dess utformning är inriktad mot att ge elevers svagaste sidor stark vikt vid betygssättning (Gustafsson, Cliffordson et al. 2014, s. 22). Hur detta kommer påverka betygen ur rättvisesynpunkt återstår att undersöka.

Flera studier har visat att de målrelaterade betygen inte går att använda för att på nationell nivå mäta kunskapsnivå och kunskapsutveckling över tid (Gustafsson, Cliffordson et al. 2014, Vlachos 2010, Wikström 2005b). Det målrelaterade betygssystemet infördes bland annat med syfte att ge information på nationell nivå om kunskapsnivå och kunskapsutveckling. Den slutsats som kan dras baserad på de studier som finns är att det målrelaterade systemet inte kunnat leva upp till detta syfte och att Sverige därför inte haft ett fullgott system för uppföljning av kunskapsresultat sedan införandet av det målrelaterade betygssystemet. Det finns sammanfattningsvis evidens för att lärarsatta betyg, oavsett hur de utformas, inte kan uppfylla kunskapsfunktionen när det gäller den nationella nivån. Istället behövs separat utformade kunskapsbedömningssystem för detta syfte (Gustafsson, Cliffordson et al. 2014).

## Betyg och urval till högre utbildning

När betyg används som urval till nästa nivå i utbildningssystemet innebär det att det ställs höga krav på att betyg är likvärdiga (Stobart 2005b). Det finns starka indikationer på att det nuvarande målrelaterade betygssystemet inte uppfyller detta krav (Gustafsson, Cliffordson et al. 2014). Dels har studier visat att det föreligger betygsinflation över tid (Wikström 2005b), dels att konkurrens mellan skolor bidrar till att ytterligare spä på betygsinflationen, även om konkurrensdelen troligen utgör en mindre andel av den totala betygsinflationen (Vlachos 2010), och dels visar studier att likvärdigheten brister i betygssättning mellan skolor och lärare (Gustafsson, Cliffordson et al. 2014, jfr Skolverket 2007).

En annan aspekt av likvärdigheten är frågan om vad ett betyg är mått på. Det finns stark evidens i såväl svensk som internationell forskning på att betyg utöver att vara mått på kunskaper även innehåller aspekter som flit, motivation, temperament och kommunikativ förmåga (Klapp Lekholm och Cliffordson 2008). Dessa ”mjuka” delar av betyg ser olika ut beroende på betygssystem, med olika typer av konsekvenser för likvärdighet. Se även kapitel 2. För att ge några exempel på detta så visar Thorsen (2014) i sin studie av normrelaterade betyg att den så kallade allmänna betygsdimensionen favoriserade flickor och elever från studievana hem. Den allmänna betygsdimensionen kan populärt sägas vara den del av betyget som inte kan förklaras av ämneskunskaper utan som beror på andra faktorer, faktorer som egentligen inte ska vägas in i betyget men som en mängd studier visat att lärare ändå väger in. Förutom kön gäller det framför allt faktorer såsom motivation och attityd. Klapp Lekhom och Cliffordson (2008) menar att det är denna dimension som förklarar varför lärarsatta betyg är bättre än centralt administrerade prov som högskoleprovet på att förutsäga framgång i studier. För att nå framgång i studier krävs nämligen inte bara kognitiva förutsättningar utan även motivation, en vilja att orka studera.

Det mål- respektive normrelaterade betygssystemen har visat sig skilja sig när det gäller vilka grupper av elever som gynnas respektive missgynnas. Medan det normrelaterade systemet framför allt gynnade elever vars föräldrar hade en hög utbildningsnivå så gynnar det målrelaterade främst elever vars föräldrar har en låg utbildningsnivå. Den mest sannolika förklaringen till det senare tros vara kompensatorisk betygssättning för missgynnade elever som ligger mellan ett underkänt och ett godkänt betyg (Thorsen 2014, Thorsen och Cliffordson 2012, jfr kapitel 1).

Även om det kan finnas goda skäl till att sätta ett annat betyg än det som borde ges med avseende på vad betygen ska vara mått på enligt regelverket så är det problematiskt när dessa aspekter inte är synliggjorda i systemet, i synnerhet om inte alla gör på samma sätt. Detta gör att betygens värde för urval urholkas även om vi i nästa avsnitt ska se att studier ändå visar att de trots dessa brister verkar ha en hel del andra fördelar i jämförelse med andra urvalsinstrument.

Det finns en i sammanhanget intressant tysk studie av Neumann, Trautwein et al. (2011). Den lyfter fram samspelet mellan lärares bedömningar och nationella standardiserade tester. Neumann, Trautwein et al. (2011) frågade sig om lärare använder olika kriterier när de sätter betyg på kurser jämfört med när de betygssätter centrala examinationer. De valde det tyska betyget ”Abitur”, som i Tyskland utgör grunden för antagning till högre utbildning. Då kampen om utbildningsplatser kan vara hård är det av vikt att de *Abitur* elever får är jämförbara. *Abitur* baseras på både de senaste årens kursbetyg som eleven erhållit samt resultat på centralt administrerade examinationer. Forskarna gjorde statistiska jämförelser mellan betyg på kursdelar och betyg på centralt administrerade examinationer och frågade sig om det fanns skillnader mellan skolor. De använde

kvalitativa delar av de centralt administrerade examinationerna vilka rättas av samma lärare som sätter kursbetygen för dessa elever. De ämnen som undersöktes var engelska och matematik. Slutsatsen de kunde dra var att jämförbarheten mellan skolor när det gäller resultat på de centrala examinationerna var klart högre än när det gäller jämförbarhet i kursbetyg, vilket alltså talar för att *Abitur* borde baseras enbart på de centrala examinationerna. Trots detta så menar forskarna att det kan finnas goda skäl att inte avfärda kursbetyg då det kan leda till negativa effekter för elever som presterar under genomsnittet, såsom minskad motivation. Det verkar nämligen vara så att de skolor där elever kompenseras mest är skolor som har mindre resurser och sämre lärmiljöer jämfört med skolor där elever inte kompenseras på samma sätt Med andra ord kan det finnas goda skäl till kompensatorisk betygssättning. Författarna hävdar att *Abitur*, såsom det de facto utformas, kan anses utgöra en bra sammanvägning av olika typer av bedömningar som sammantaget leder till ett mer rättvist system än om enbart det ena eller det andra måttet använts. En invändning som kan resas mot detta sätt att resonera är att det inte lever upp till vad vi tidigare nämnt skulle gälla rättvisa system, nämligen öppenhet och medvetenhet. Utifrån det resonemanget hade det kanske varit rimligare att kompensera de skolor som har sämre förutsättningar med resurser istället för genom högre betyg. Det senare kan dock ses som en nödlösning i system som inte fungerar som de bör och verkar förekomma även i Sverige (se t.ex. Klapp Lekholm och Cliffordson 2008).

Om man ser till rättvisefrågan så kan man notera att det saknas kraftfulla analyser av de svenska betygs- och antagningssystemen utifrån etnicitet och funktionsnedsättning. Därtill är de studier som finns av kön och social klass inte så kraftfulla som de troligen hade varit om det funnits ett mer tillförlitligt nationellt kunskapsmätningssystem som tillåtit bättre jämförelser.

## Betyg som förutsägelse av studieavhopp respektive studieframgång

Om man nu ger betyg från tidig ålder och det därmed finns rik betygsdata för varje enskild elev, hur kan dessa data då användas på systemnivå? Ett flertal nyare studier har undersökt prediktionsförmågan hos lärarsatta betyg, dels i relation till studieframgång i högre utbildning, dels i relation till elever som riskerar att misslyckas med sina studier. Rent allmänt är det väl belagt att betyg är bättre än tester på att förutsäga skolframgång (Cliffordson 2008, Sawyer 2013, Thorsen 2014, Thorsen och Cliffordson 2012). När det gäller prediktionsförmåga i relation till olika högskoleprogram så gäller att högskoleprovet uppvisar en högre varians jämfört med både det mål- och det normrelaterade betygssystemet. Tittar man närmare på enbart Högskoleprovets varians gäller att det bättre förutsäger framgång inom humanistiska och samhällsvetenskapliga utbildningar än tekniska och naturvetenskapliga (Cliffordson 2008, s. 71). Minst varians i relation till typen av högskoleprogram har de normrelaterade betygen.

De målrelaterade betygen har visat sig bättre på att förutsäga framgång i högre utbildning jämfört med de normrelaterade (ibid.). Detta är delvis förvånande då det normrelaterade betygssystemet utformades för att fungera som just urvalsinstrument medan det målrelaterade haft en mängd andra syften att fylla. En förklaring till att det målrelaterade systemet fungerar bättre är att det baseras på en större mängd enskilda bedömningar därtill gjorda över tid, en följd av den kursutformade gymnasieskola som infördes samtidigt med de målrelaterade betygen (ibid.). En annan orsak kan vara det sätt som normering skett på inom det normrelaterade systemet, vilket kan ha gynnat respektive missgynnat grupper på sätt som minskat dessa betygs förmåga att förutsäga studieframgång (ibid.). Man skulle kunna tro att betygsinflationen som nämndes i förra avsnittet skulle göra att de målrelaterade betygens prediktionsförmåga skulle dras ner kraftigt. Att betygsinflationen inte påverkar de målrelaterade betygens prediktionsförmåga så mycket som man skulle kunna tro beror på att de som antas till högre utbildning tillhör relativt närliggande årskullar, varför effekterna av betygsinflation mellan dessa kullar är små (ibid.). För den enskilde individen kan denna ojämlikhet av förklarliga skäl ha stora konsekvenser (han eller hon kommer eller kommer inte in på ett program som konsekvens av att han eller hon fått ett högre eller lägre meritvärde som följd av betygsinflationen).

Samtidigt som de målrelaterade betygen verkar fungera väl som instrument för att förutsäga studieframgång inom högre utbildning så är de sämre än normrelaterade betyg på att förutsäga framgång i gymnasieskolan (Thorsen 2014, Thorsen och Cliffordson 2012). Denna skillnad beror troligtvis på den gräns mellan godkänt och icke godkänt betyg som finns i det målrelaterade systemet och, som vi redan berört, bidrar till

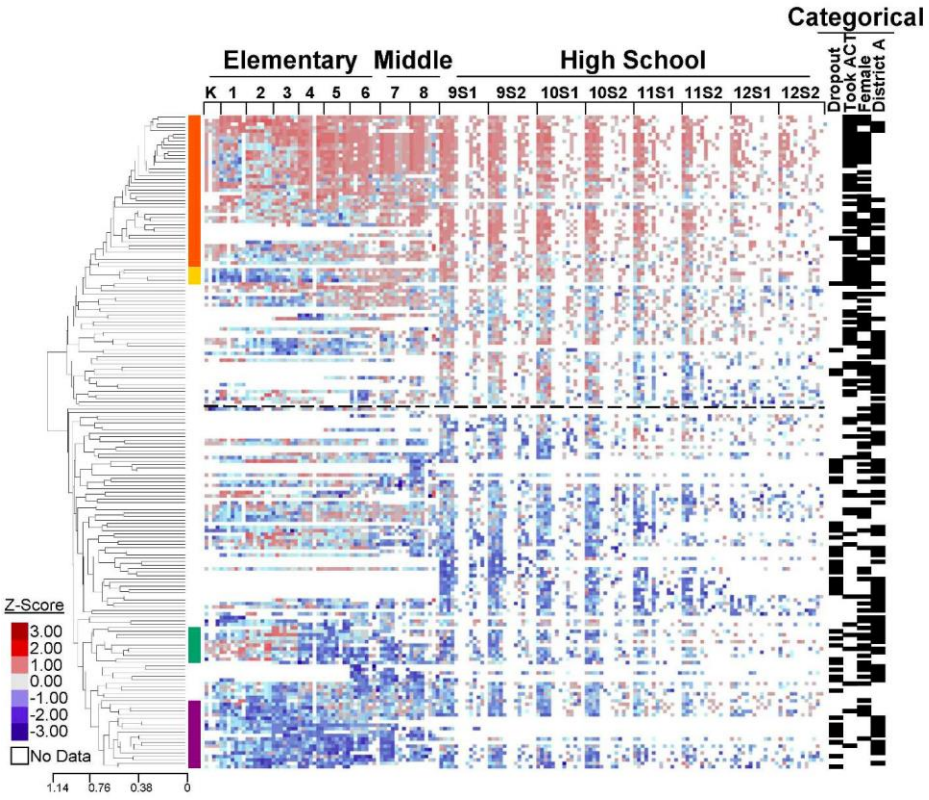
kompensatorisk betygssättning vilket alltså troligtvis är en aspekt som drar ner dessa betygs prediktiva förmåga när det gäller studieframgång på gymnasienivå. Man bör alltså i sammanhanget notera att de målrelaterade betygen inte har samma utformning i grundskola som gymnasieskola. Medan de är ämnesbaserade i grundskolan med ett slutbetyg i varje ämne så är de kursbaserade i gymnasieskolan, vilket innebär att betyget efter gymnasiet innehåller flera mätpunkter i varje ämne som därtill är gjorda över tid. Denna utformning av ett målrelaterat system verkar alltså leda till högre tillförlitlighet i relation till studieframgång, som vi var inne på i förra stycket (Cliffordson 2008). Målrelaterade system kan alltså utformas olika med konsekvenser för likvärdighet och prognosvärde.

Varje givet betygssystem gynnar respektive missgynnar grupper av elever på olika sätt varför prediktionsförmåga inte kan vara den enda frågan utan alltid måste ses i relation till andra aspekter av rättvisa i systemet. För att ge ett inspel från internationell forskning på området så fann Sawyer (2013) att betyg var bättre på att förutsäga framgång under första året på en universitetsutbildning jämfört med test i de fall där det inte var hög konkurrens om platserna och universitetet inte var högrankat. På de utbildningar där det var hög konkurrens och universitetet dessutom var högrankat så var test bättre på att förutsäga framgång. Allra bäst som mått på framgång var en sammanvägning av test och betyg (ibid.).

Givet varje system och de eventuella ojämlikheter som detta system bär på kommer det när det gäller antagning även in andra överväganden än enbart framgångsfaktorer, såsom att antagningssystem ska vara förutsägbara över tid och att de ska uppvisa en hög grad av uniformitet och objektivitet (Sawyer 2013). Att skapa system där olika regler gäller för olika skolor och där detta kan ändras över tid, beroende på exempelvis om en utbildning eller ett universitet räknas som högt rankat det året eller ej, skulle skapa problem med gränsfall och oförutsägbarhet och sådant är inte önskvärt. Ur detta perspektiv är det intressant med den tidigare refererade tyska studien av *Abitur*, som visar på medelvägar där lärares kursbetyg sammanvägs med betyg på nationella prov (Neumann, Trautwein et al. 2011). Men även sådana sammanvägningar kan ha brister i relation till vissa typer av utbildningar (Sawyer 2013).

En nära relaterad aspekt till betyg som predikator för framgång i studier är betyg som tecken på elever som riskerar att misslyckas med sina studier. Bowers, Sprott et al. (2013) undersökte 110 olika typer av ”drop-out flags” som använts på studier av amerikanska elever, bland dem betyg, då inte i form av avgångsbetyg, utan sammanvägd betygshistorik så kallad *grade point average*, förkortat GPA. En ”drop-out flag” är en indikator på risken att en elev inte kommer att fullgöra sina studier. Indikatorn kan användas på exempelvis skol- eller kommunnivå och när den indikerar risk för avhopp så kan man vidta åtgärder för att förhindra att en elev inte avslutar sina studier i förtid. Bowers, Sprott et al. (2013) kunde konstatera att ingen av dessa indikatorer gav helt korrekta resultat, många byggde därtill på komplicerade konstruktioner och var av dessa skäl inte lika användbara för exempelvis skolor och beslutsfattare som andra och mer lättbestämda mått. De fann att sammanvägd betygshistorik, GPA, låg bra till som indikator (ibid.). Samtidigt varnade de för att alltför mekaniskt utforma ett system baserat på dylika riskfaktorer. Ett sådant system kommer exempelvis att identifiera elever som tillhörande riskgruppen trots att de inte gör det. Faktum är att elever riskerar att misslyckas bara genom det faktum att de identifierats som tillhörande riskgruppen.

Figuren nedan är hämtad från Bowers (2010, s.9) och visar en så kallad hierarkisk klusteranalys gjord på basis av elevers betygshistoria. I figuren ges hela elevens betygshistoria, från första klass (vitt markerar punkter där betygsdata saknas). Eleverna har därtill klustrats ihop i grupper av elever som har liknande betygsprofiler och betygsutveckling. Nyanserna i figuren indikerar olika höga betyg. Två stora kluster i figuren utgörs av dels elever med höga betyg som tar examen i tid (det röda), dels elever med låga betyg och där majoriteten av eleverna hoppar av sina studier i förtid (det blåa). Ett annat intressant kluster utgörs av elever som har relativt höga betyg i början av sina studier men där dessa sedan faller i takt med att de framskrider genom utbildningssystemet och där en hög andel av dem hoppar av (det gröna). Denna typ av data hade tillsammans med andra data kunnat bidra till en bättre grundad fördelning av resurser och planering av interventioner på lokal nivå menar Bowers (2010). Resultat från USA är dock inte direkt överförbara till svenska förhållanden men visar att betyg har potential att användas till annat av beslutsfattare än vad de används till idag.




***Figur 1. En så kallad hierarkisk klusteranalys gjord på basis av elevers betyg, hämtad från Bowers (2010, s. 9). I figuren ges hela elevens betygshistoria, från första klass (vitt markerar punkter där betygsdata saknas). Eleven har därtill klustrats ihop med elever som har liknande betygsprofiler och betygsutveckling. Nyanserna i figuren indikerar olika höga betyg där röda anger höga betyg och blå låga.***

## Diskussion och slutsatser

Att genomföra en systematisk litteraturundersökning av betyg ur ett systemperspektiv innebär att man låser sig vid att betyg ska finnas som en komponent i ett utbildningssystem. Detta är inte självklart. De funktioner betyg fyller kan även fyllas på andra sätt, ibland på ett bättre sätt än via betyg. Därtill har vi svårigheten att isolera betygens del i de systemeffekter man önskar undersöka.

Som framgått har betygens roll i många utbildningssystem reducerats de senaste decennierna. Tydligast är det amerikanska systemet där betyg i början av 1900-talet hade en bärande roll men där de idag främst används för att rapportera skolresultat till föräldrar (Brookhart 2013b).

Med detta sagt visar genomgången av betyg ur ett systemperspektiv att betyg är bättre som urvalsinstrument för högre utbildning jämfört med högskoleprov. Målrelaterade betyg så som de är utformade i gymnasieskolan är därtill bättre än normrelaterade betyg när det gäller att förutsäga studieframgång. Förklaringen till det målrelaterade gymnasiebetygets prediktiva förmåga tycks vara att det innehåller ett större antal kursbetyg snarare än sammanfattande ämnesbetyg. En annan viktig faktor tycks vara att kursbetygen inte erhållits vid ett tillfälle utan att de erhållits utspritt över tid. Dessa båda faktorer gör att de målrelaterade gymnasiebetygen innehåller rikare och därmed bättre information kring en elevs studieförmåga jämfört med när betyg sätts i ämnen och i form av slutbetyg. Det finns dock indikationer på att en kombination av test och lärarsatta betyg kan ge ännu bättre prediktionsförmåga (Sawyer 2013, jfr Neumann, Trautwein et al. 2011).

När det gäller det svenska målrelaterade betygssystemet så ger det ett inte tillfredsställande mått på elevers kunskaper på systemnivå då det inte har tillräcklig precision. Betyg skiljer sig dels mellan skolor, dels mellan lärare. Betygen är inte heller jämförbara från en årskull till en annan.

När det gäller grundskolan så gynnade det normrelaterade betygssystemet elever från studievana hem medan det målrelaterade systemet gynnar elever från studieovana hem. Det senare förklaras troligen av att lärare i valet mellan ett underkänt och ett godkänt betyg oftare väljer godkänt, det vill säga det finns en tendens att elever som ligger strax under gränsen för godkänt betyg får godkänt betyg trots att de egentligen inte borde haft detta betyg. Även om denna tendens att ge högre betyg än det eleven borde ha kan ses som en brist i tillämpningen av dagens betygssystem så kan det också förstås ur ett kompensatoriskt perspektiv, nämligen att många lärare inte upplever att dessa elever fått de resurser som de haft rätt till och därför kompenserar dessa elever genom att ge dem det betyg de anser att de borde haft om de fått den hjälp borde fått. Vi har även varit inne på andra orsaker till detta i kapitel 2, nämligen att det är tidsödande med elever som måste arbeta upp sitt betyg till godkäntnivån och att det kan hända att lärare ger dem godkänt för att slippa detta extraarbete.

Även om statliga myndigheter följt utvecklingen av och granskat det målrelaterade betygssystemet noggrant ur ett likvärdighetsperspektiv så saknas det större perspektivet av rättvisa och likvärdighet, något som även inkluderar att kunskapsfrågan belyses noggrant. Det saknas också en analys av brister i nuvarande system i relation till kön, klass, ålder, etnicitet och funktionsnedsättning. Ett problem i relation till bättre undersökningar av dessa fem aspekter är det faktum att Sverige inte har ett kvalitetssystem gjort för att kunna följa utvecklingen av elevers kunskapsnivå, ett sådant instrument hade gett bättre stöd för studier av social ojämlikhet i nuvarande betygssystem, tillförlitligare resultat och följaktligen starkare incitament till förändring i riktning mot ett mer rättvist och bättre fungerande betygssystem. Observera att ett sådant system också hade inneburit att instrumentets egna brister noggrant undersökts och varit del av analyserna. Det finns inget ”objektivt test”, som Gipps (1995), påminde oss om.

Slutligen, betyg kan aldrig förstås isolerat utan måste ses i relation till ett lands utbildningssystem, dess bedömnings- och utvärderingssystem och de kulturella och historiska omständigheter som omgärdar betyg. Att exempelvis jämföra i vilka årskurser olika länder börja sätta betyg innebär ofta att man jämför äpplen med päron. Vi kommer beröra jämförande aspekter utförligare i kapitel 4.

För att stärka förståelsen av det svenska betygssystemet så bör konkurrerande komponenter i bedömnings- och utvärderingssystem belysas med hjälp av systematiska litteraturstudier. En given sådan komponent är tester. Tester har börjat inta en allt mer central roll i många utbildningsystem och sannolikt kommer den utvecklingen bli tydligare även i Sverige framöver. Det finns för- och nackdelar med såväl lärarsatta betyg som de som ges av tester, men också en stor fara i att överger lärarsatta betyg helt – en risk att man så att säga slänger ut barnet med badvattnet då tester av allt att döma har brister på punkter där betyg fungerar bättre. Över huvud taget bör utvecklingen av instrument för att mäta och kontrollera skolutveckling och prestationer undersökas, såsom *value-added models* och modeller för skolutveckling, såsom NCLB och dess efterföljare. Vi har valt att belysa tester, *value-added models* och skolutvecklingsmodeller (*accountability* modeller) i nästa kapitel eftersom de ur ett systemperspektiv verkar på samma arena och därmed konkurrerar med betyg om vilken av dessa som uppfyller olika givna syften på bäst sätt.

# BETYGSSÄTTNING UR OLIKA KOMPARATIVA PERSPEKTIV.

De flesta av oss har personligen bara upplevt ett betygssystem, även om vi hört från våra föräldrar eller barn om andra sätt att få betyg. De elever som nu går i skolan och de som gick i skolan i mitten på 1990 talet har fått uppleva två olika betygssystem. De har möjlighet att göra jämförelser och uttala sig om skillnaderna mellan olika typer av betyg. Just möjligheten till jämförelser är något forskare försöker utnyttja. I vissa länder, som till exempel i USA och Tyskland, finns flera olika parallella betygssystem. Andra skillnader som kan utnyttjas för jämförelser av forskare är att betyg sätts av olika lärare, i olika ämnen och på olika elever. Betygsystem ser också olika ut mellan olika länder och betyg ändvänds till olika saker och får olika konsekvenser. I syfte att vidga bilden av vad betyg kan vara har vi i detta kapitel analyserat olika typer av komparativ forskning om betyg.

## Metodbeskrivning

I Sverige kom den senaste betygsreformen 2011 att motiveras med att flera länder inom EU anpassat sig till en gemensam norm för hur betyg bör se ut, nämligen den 6-gradiga ECTS-skalan (Prop 2008/09:66, Prop 2008/09:87, Ds 2008:13). Jämförelsen med andra länder blev alltså ett politiskt argument för att förändra betygsskalan till en sex-gradig skala, även om svenska lärares synpunkter om att betyget G i den tregradiga skalan blivit för omfattande också spelade roll (Ds 2008:13). Betygen flyttade samtidigt ned i åk 6. Detta motiveras också som en anpassning till när man satte betyg i andra länder, och 2014 kom en utredning som förslog att betyg skulle sättas redan i åk 4 (Promemoria 2014-08-20). Nu tillkom också argumentet att länder som presterade bättre än Sverige på PISA-proven som regel började med betyg tidigare. Med tanke på hur diskussionerna kring betyg sett ut i Sverige under 2000-talet är det rätt naturligt att se vad för slags jämförelser forskare gör mellan olika betygssystemen, och om det finns andra intressanta skillnader mellan länder eller olika betygssystem än i vilka årskurser betyg ges och hur många skalsteg som används.

Vi bestämde oss för att söka efter användbara perspektiv och resultat i databaserna ERIC och JSTORE. ERIC samlar de största pedagogiska tidskrifterna medan JSTORE också omfattar andra discipliner inom humaniora och samhällsvetenskap, t.ex. språk, ekonomi.

I den kartläggning av hur betyg sätts i andra länder har vi avgränsat oss till Europa. Vi har dock systematiskt gått igenom *Assessment in Education* och de artiklar som beskriver länder och världsdelar för att ge en lite bredare bild. Dessa artiklar som funnits med från starten av tidskriften 1991 benämns: *Profiles of educational assessment system worldwide*. De finns samlade i en särskild portal (http://www.tandf.co.uk/journals/pdf/AIEProfiles.pdf). Genomgången av de europeiska länderna utgår från Europeiska kommissionens databas EURYDICE. Denna bygger på de europeiska ländernas självrapportering om sina utbildningsystem utifrån vissa bestämda rubriker. Principen med självrapportering har den fördelen att landets utbildningssystem tolkas så att säga inifrån. Det är oftast nationella myndigheter och departement som tillhandahåller informationen. Vi får en bild av vad man i respektive land tycker är viktigt att lyfta fram om betyg och bedömning, och sannolikt en ganska bra sådan bild. Ett problem däremot är att detta ska uttryckas på engelska varpå vissa landspecifika uttryck går förlorade. Som utomstående kan man riskera tolka de använda termerna på ett annat sätt än den som skrivit. Ett större problem är att vissa fakta kanske inte ens finns med. Landsrapporterna i EURYDICE följer endast vissa övergripande gemensamma standards. Vi har försökt ta hänsyn till detta genom att titta på styrdokument där sådana har refererats i EURYDICE och framför allt genom att försöka knyta an till annan forskningslitteratur. En sammanfattning av betygen i Europa presenteras i Appendix. Den sammanfattningen är uteslutande baserad på EURYDICE. I de exempel vi lyfter fram har vi emellertid också använt andra källor, särskilt ovannämnda forskningsartiklar från Assessment in Education, för att fördjupa oss kring vissa aspekter. Vi har valt att presentera Appendix på engelska för att inte ytterligare lägga till en språklig uttolkning. Vi har slutligen också gått igenom de rapporter OECD tagit fram som har särskilt fokus på betygssättning och resultatstyrning.

### Sökresultat

När vi gör en enkel sökning på JSTORE på söksträngen ”comparison AND grading AND education AND country” får vi närmare 2000 träffar 1990-2014. Med olika justeringar av ordens grundform, synonymer och genom att inte tillåta ”higher education” går det att komma ner till drygt 780 träffar för åren 2000-2014. Detta var också väl många att gå igenom och vi bedömde att vi kunde avgränsa oss till artiklar där vårt huvudsökord, ”grading/grades”, skulle nämnas i abstract. Då kom vi ner till 61 artiklar. Dessa artiklar lästes i sin helhet och 29 artiklar ansågs relevanta kring olika teman av jämförelser av betyg. Genom att införa krav på ”grading” i abstract försvinner de studier där man jämförde elevresultat i andra syften än att förstå principer för betygssättning, dvs. studier snarare av *learning outcomes*. Exempelvis ”slipper” vi på detta sätt alla jämförelser länder emellan utifrån PISA-resultat, något som behandlas utförligt i en annan Skolforsk-rapport (Lindblad, Pettersson & Popkewitz 2015). Däremot kvarstår artiklar där direkta och indirekta effekter på ”grading” av internationella kunskapsmätningar diskuteras. Några intressanta studier har också valts bort då de ligger närmare medicinsk diagnostisering än betygssättning, även om gränsen kanske inte alltid är helt uppenbar (t.ex. Dowdy och Kamphaus 2007)

I ERIC har vi sökt med Proquest och utgick där initialt från söksträngen: ”comparison AND ’assessment system’ AND country”. Vi avgränsade oss till *peer reviwed journals* 2000 – 2014. Detta gav i 44 träffar där en stor andel handlade om jämförelser mellan länder utifrån PISA resultat. Genom att ta bort country och byta ut ”assessment system” mot ”summative assessment” fick vi 14 träffar. Samma sökning men med ”grading” istället för ”summative assessment” gav 8 träffar. Vi har gått igenom träffarna för alla dessa tre sökningar. Ett tiotal titlar överlappar varandra och i slutändan hade vi från ERIC 18 artiklar som bedömdes relevanta och som inte fanns med i träffbilden från JSTORE. Det betyder att vi i detta kapitel analyserat knappt 50 artiklar där olika typer av jämförelser av och med betyg stått i fokus. Vi kan konstatera att det inte finns särskilt mycket forskning vare sig i Sverige eller internationellt som på ett systematiskt sätt arbetat med jämförelser av betygs- och bedömningssystem. Däremot finns det väldigt många studier som jämför elevers studieresultat så som de kommer fram med test och med betyg. De huvudteman vi identifierat är:

- Jämförelser mellan länder
- Effekter av internationella jämförelser
- Effekter av (internationellt inspirerade) *accountability*-modeller
- Jämförelser av skolinterna bedömnings- och betygsmodeller
- Jämförelser av externa och interna bedömningsmodeller

## Jämförelser mellan länder – vad är det som oftast jämförs?

Vad är det framförallt som jämförs när forskare gör internationella utblickar kring betyg och summativ bedömning? I den ofta citerade artikeln Lessons from around the World: How Policies, Politics and Cultures Constrain and Afford Assessment Practices (2005) diskuterar Paul Black och Dylan Wiliam hur det kommer sig att det är så stora skillnader mellan bedömning och betygssättning i England, Frankrike, Tyskland och USA. De menar, givet alla andra sätt på vilka organiseringen av ett utbildningssystem kan variera på, att det heller inte är konstigt att det finns stora skillnader i synen på bedömning och hur man praktiskt ordnar den. Några saker som tycks avgörande är: typen av läroplan, ämnenas organisering, valfrihet, övergångar mellan skolformer, urval, synen på rättvisa, policy kring läromedel etc.

Det Black och Wiliam lyfter fram som centrala variabler för hur olika länder skiljer sig åt är när i åldrarna bedömningar görs, vem som gör dem (internt – externt), i hur många skalsteg betygen ges, om betyg är relaterade till externa tester, kvalificeringssystem (dvs. övergångarna till högre utbildningar). Varje land har en kombination av många olika formativa och summativa bedömningar för mer eller mindre tydliga syften. Vi kommer längre fram titta närmare på hur dessa variabler skiljer sig åt Europa. Det är i alla fall uppenbart att olika länder hittat sina vägar, vilket på gott och ont präglar deras system för betyg och bedömning (se även OECD 2005). Black och Wiliam beskriver som exempel det amerikanska betygssystemet:

> Beginning in the third or fourth grade (and continuing through to postgraduate level!), almost all formally assessed student work is assessed on the same literal grade scale: A, B, C, D, F (fail), typically corresponding to percentage scores of 90–100, 80 – 89, 70 – 79, 60 – 69 and 0 – 60 respectively. Grades are cumulated by converting them back to numbers (A=4, B=3, C=2, D=1, F=0) and calculating the 'grade- point average' over the year. However, unlike scores or grades given in most European countries, the grade is usually not a pure measure of attainment, but will include how much effort the student put into the assignment, attendance, and sometimes even behaviour in class. Paul Dressel's definition of a grade was 'an inadequate report of an inaccurate judgement by a biased and variable judge of the extent to which a student has attained an undefined level of mastery of an unknown proportion of an indefinite material' (Chickering, 1983), and while this may be a bit unfair, there can be little doubt that the meaning of a grade varies substantially from school to school, and even from teacher to teacher. (Black och Wiliam 2005, s. 257)

Traditioner och skilda uppfattningar kring skolan påverkar hur länder utformar sina bedömningssystem. Kring de olika variabler vi nämnde ovan där nationella skillnader finns, listar Black och Wiliam en rad faktorer som formar dessa variationer:

- beliefs about what constitutes learning;
- beliefs in the reliability and validity of the results of various tools;
- trust in the objectivity of formal testing;
- a preference for and trust in numerical data, with bias towards a single number;
- trust in the judgements and integrity of one's children's teachers;
- trust in the judgements and integrity of the teaching profession as a whole;
- belief in the value of competition between students;
- belief in the value of competition between schools – the market model of education;
- belief that test results are a meaningful indicator of school effectiveness;
- fear of national economic decline and belief that education is crucial to improvement;
- belief that the key to schools' effectiveness is strong top-down management. (Black och Wiliam 2005, s. 258f)

Det är med andra ord en hel del olika faktorer som bidrar till att länders bedömningssystem varierar. Trots artikelns namn, Lessons from around the world... förefaller den främsta lärdomen vara att vi inte kan lära så mycket av andra länders bedömningssystem, mer än att alla länder tycks slita med att få summativa och formativa bedömningar att passa ihop på ett funktionellt sätt:

> Thus not only is there no 'royal road' to an assessment system that effectively serves both formative and summative functions that each country could follow, but it seems likely that the idiosyncratic road that will need to be taken in each country will also be very hard going. (Black och Wiliam 2005, s. 260)

I en studie av skillnaderna i bedömning mellan England, Sverige och Tyskland fokuserar Florian Waldow (2014) särskilt på betydelsen av hur man ser på rättvisa procedurer för urval, *procedurial justice* (se även kapitel 3). Waldow pekar på att den viktigaste funktionen betyg historiskt sett fyllt är att beteckna en merit. Meriter är alltid jämförbara. Poängen med meriter är att någon har bättre och andra sämre. På basis av meriter går det således att göra ett rättvist urval – förutsatt att om de som bestämmer vad en merit är har legitimitet. Det finns i varje meritokrati "gatekeepers", grindvakt, men vad en gatekeeper tycker är meriter och rättvisa sätt att bedöma dem kan variera väldigt. Grindvaktens hela legitimitet bygger på att andra ansluter sig till dess uppfattning om rättvis bedömning. Waldow pekar på att det är den proceduriella rättvisan som är viktigast för elever, och att just deras uppfattning därför är väldigt viktig för systemets legitimitet. Waldow visar att medan Tyskland och Sverige bygger sina examinationers legitimitet (av gymnasiestudenter) på en idé om professionalism och att läraren känner sina elever, anses ett sådant system i England skapa orättvisa. Där anses snarare en bedömning vara rättvis när endast det man presterar vid examinationen spelar roll. Där är det självklart att examinatorn är helt extern.

Att det uppstår olika sätt att se på vad som är en rättvis examination har att göra med utbildningsväsendets struktur och dess behov av legitimitet avseende just examinationerna. I England finns en diskussion om respektive extern examinations svårighetsgrad (är de t.ex. lika svåra), i Tyskland diskuteras om det är lika svårt att få en examen i respektive *Länder* (Tyska delstater), i Sverige handlar diskussionen om likvärdig bedömning mellan enskilda skolor. En sak som förbryllar Waldow är att en central metod för att hantera dessa diskussioner om rättvisa i bedömningar finns i England och Tyskland men saknas i Sverige. Det gäller nämligen rätten att överklaga sitt betyg. Waldow menar att en anledning till att denna rätt inte finns i Sverige, och att regeringen 2006 – 2010 avfärdade förslagen från utredningen om rättvis bedömning till just en sådan rätt, kan spåras till att de svenska nationella proven alltjämt anses som en tillräcklig garant för en rättvis bedömning (Waldow 2014, s. 337). Precis som Black och Wiliam ovan, visar Waldow på den mängd sociala faktorer som gör att bedömningssystem varierar länder emellan, och därför är svåra att jämföra. Försök saknas dock inte.

Ett tema för de internationella jämförelserna är att forskare använder sitt eget ”hemlandsperspektiv” och utifrån det gör en utblick mot ett annat land. Exempelvis jämför den svenska forskaren Christina Wikström (2009) bedömning i Sverige med England (se även ”lektioner från Finland”, Hendrickson 2012, Dobbins och Martens 2012, från Shanghai och Singapore, Tan 2011). Kring högre utbildning och betyg finns ytterligare ett par artiklar men som faller utanför intresset för den här rapporten (t.ex. Billing 2004, Dahl, Lien, Lindberg-Sand 2009). Wikström beskriver hur England och Sverige från motsatta positioner gått mot en kriteriebaserad bedömning utifrån givna standards; England genom en utveckling som inneburit mer centralisering och Sverige genom mer decentralisering. Därmed ansluter sig båda länderna till det som Wikström beskriver som den dominerande internationella trenden, det som ibland kallas *standards based curriculum*.

Wikström lägger särskild vikt vid hur nationella prov används för accountability och betygssättning. Wikström menar att de engelska proven (*Key stage test*) har ett väldigt tydligt syfte, att ställa skolor till svars för låga resultat. De engelska testen är verkligen *high stake* för skolan. Detta har skapat debatt, forskning och bidragit till att reliabiliteten i testen ökar, menar Wikström. Samtidigt har validiteten äventyrats i det att skolor systematiskt arbetar för att klara testen, men därigenom begränsar sitt arbete utifrån läroplanen som är vidare i sitt anspråk än vad som kan mätas på detta sätt. Wikström ser ändå en positiv skillnad jämfört med Sverige och det är att de engelska proven tvingat fram forskning och debatt kring bedömning på ett sätt som inte skett i Sverige där nationella prov från statens sida beskrivs som *low stake*.

Efter de senaste PISA-mätningarna 2009 och 2012 har intresset för Sydkorea, Singapore och Shanghai ökat medan det minskat något för Finland (Waldow, Takayama et al. 2014). I ett försök att förklara varför flera Asiatiska länder lyckas så bra i dessa test har Charlene Tan (2011) analyserat en del skillnader och likheter mellan Singapore och Shanghai. Hon fokuserar särskilt på kulturens betydelse för studieresultaten.

Både Singapore och Shanghai är stora, rika och globala städer med en klart högre genomsnittlig utbildningsnivå än andra städer i regionen. Utbildningspolicyn i både Singapore och Shanghai präglas av en ambition om att eleverna behöver rustas för en global kunskapsmarknad. I grunden finns också en utilitaristisk och teknokratisk utbildningsmoral, menar Tan, som i kombination skapar traditioner av att studieframgång är att prestera bra på prov och examinationer. Det finns höga förväntningar på eleverna och i t.ex. Shanghai läser mellan 50 – 60 procent av eleverna extra på kvällsskola i ämnen där de redan har höga betyg. I Singapore tillbringar en vanlig högstadieelev dagligen runt 3h av sin fritid med särskilda test och provböcker (*assessment books*) samt ytterligare 2h i kvällsskola (Tan 2011, s. 162).

Tan menar att det inte är konstigt att elever från dessa städer presterar bra på internationella kunskapsmätningar då de både kan mycket och är vana vid att tävla med kunskapsresultat genom papper och penna prov. Det finns dock en ytlighet i lärandet menar Tan och hänvisar till studier av arbetsgivare som klagar på hur skolan fostrar framtidens arbetskraft. Arbetsgivarna upplever att de anställda ofta saknar nyfikenhet, förmåga att ifrågasätta och kan främst lösa problem i en ”papper och penna”-kontext. Detta oroar även den politiska nivå skriver Tan, men så länge eleverna presterar bra på internationella kunskapsmätningar och kulturen hyllar resultat på enskilda examina, så är det svårt med mer progressiva reformer (2011, s. 164).

# Effekter av internationella jämförelser på nationella system

Att internationella jämförelser av elevers kunskaper spelar roll på nationell nivå är det flesta överens om, även om det ser olika ut i olika länder (Pettersson 2008, Pettersson och Wester 2010). Ett av de länder där PISA-resultat länge ignorerades var Frankrike. På nationell politisk nivå fanns inget intresse av internationella jämförelser och de låga resultat Frankrike fick i PISA 2000 och 2003 avfärdades ofta med att PISA inte mätte rätt saker. Men efter låga resultat även i PISA 2006 började den politiska nivån att reagera med reformförslag. I artikeln Towards an Education Approach a la "Finlandaise"? French Education Policy after PISA (2012), beskriver forskarna Dobbins och Martens hur franska politiker allt mer började snegla mot Finland:

> Against this background and in view of the clear deterioration of the French PISA results, the French government began to take an alarmist stance and a new education minister, Xavier Darcos, was mandated with the implementation of a broad range of secondary school reforms 'with a Finnish touch'. (Dobbins och Martins, 2012, s. 35)

Exempelvis infördes ett två-terminssystem, skolor fick större självständighet vad gällde implementering av nya reformer och de fick ett större och ekonomiskt inflytande. Författarna menar att mycket av detta var positivt för franskt vidkommande, men betonar också mycket av förändringarna diskuterats länge i Frankrike. Att hänvisa till Finland löste upp vissa politiska motsättningar mellan höger och vänsterblocken. Fransmännens, som författarna skriver ”reflexmässiga skepsis” till OECD och amerikanska styrmodeller kunde med hjälp av hänvisningar specifikt till Finland också kringgås.

I flera andra mer kritiska artiklar till hur internationella kunskapsmätningar används i nationell policy, beskriver forskarna hur politiker ofta är selektiva i vad de väljer att jämföra med. Att jämföra sig med ett högpresterande land kan exempelvis vara ett sätt att legitimera nationella reformer, vilket Paul Morris visar med hur England refererar till skolresultat från Hong Kong (2012). Även Israeliska forskare pekar på hur internationella kunskapsmätningar lett till nationella reformer som flyttat fokus från svårare frågor om jämlikhet till ”quick fix” genom fler nationella mätningar, vilket forskarna inte bedömer kommer att lösa problemen (Feniger, Livneh et al. 2012).

Paul Morris (2012) artikel lyfter också fram en del metodologiska svårigheter med jämförelser av kunskapsresultat och framförallt om hur man kan använda dem på nationell nivå. Detta tema finns i ytterligare några artiklar, och exempelvis blir det svårt att tolka resultaten när egna mätningar och elevers betyg pekar i en annan riktning. Vilken ska då bedömas som den mest rättvisande bilden? Detta dilemma har lyfts fram såväl i USA som i Europa (Wang 2001, Bautier, Crinon et al. 2006).

# Effekter av (internationellt inspirerade) accountability modeller

När vi söker på jämförelser av bedömning och bedömningssystem så blir det tydligt att betyg och bedömning har en extern och en intern dimension. Bedömningar av elevers kunskaper sker inom skolan och av aktörer utanför skolan. Resultaten kan oavsett vem som tagit initiativ till dem användas internt i skolan eller externt, t.ex. för utvärdering, kontroll och policy.

Några studier som är särskilt intressanta att lyfta fram är studier som jämför olika system för att arbeta med *Value added*-modeller. Poängen är helt enkelt att en verksamhet ska värderas efter vad den tillför elevernas lärande snarare än utifrån slutresultat då skolor i välbeställda miljöer ofta har lättare att nå höga slutresultat, oaktat vad de egentligen har tillfört. *Value added*-modeller diskuteras flitigt på den internationella pedagogiska scenen men kanske mer än någon annan stans i USA. Exempelvis har Wang, Aubrey et al. 2013 visat på en utveckling i USA av system som gör det lättare för lärare att från dag till dag använda sig av de skolresultat som löpande genereras. Istället för svagt integrerade statliga system (som AYP, *'adequate yearly progress*'), bör fokus, menar forskarna, ligga på lokala data och lokal användning. De statliga systemen visar på en stark korrelation mellan elevresultat och elevgruppernas sammansättning, men ger lite information om relationen resultat och organisation eller undervisningsmetoder, dvs. sådant som skolorna kan påverka (se även Tekwe, Carter et al. 2004 för ett europeiskt perspektiv). Samma typ av problem står svenska SIRIS och SALSA-

instrumenteten inför, vilka tillhandahålls av Skolverket. De kan inte användas för att se vilka lokala åtgärder det är som kan fungera för att ändra resultaten. Däremot kan de förklara hur stor del av resultaten skolorna har svårt att påverka.

Ett flertal artiklar handlar förstås om effekter av styrmodeller som sätter *learning outcomes*, t.ex. betygsresultat, i fokus, så kallade *accountability models*. Den här forskningen är mycket omfattande och det vi lyfter upp är endast när den har ett jämförande perspektiv och mer uttalat berör betyg. En vanlig uppfattning om *accountability*-system inom utbildning är att de riskerar begränsa arbetet utifrån läroplanen till det som är enkelt att mäta (Hout och Elliot 2011). Bilden kan dock utvecklas och nyanseras.

En studie baserad på internationella jämförelser av *accountability policy* och på en analys av engelska skolor baserat på låg- och högpresterande skolor visar att en allt för hårt driven resultatstyrning tenderar ha mest negativa effekter på lågpresterande skolor, där det också ofta saknades en bra samarbetskultur mellan lärarna och program för att hantera elever med särskilda behov. Ökade krav förutsätter menar studien också att det finns organisatoriska förutsättningar för att hantera dem (Rustique-Forrester 2005).

Det sannolikt största initiativet till en *accountability* reform är amerikanska *No Child Left Behind* där årliga *high-stake* tester genomfördes i syfte att identifiera skolor och elever som ”halkade efter” (se även kapitel 3). Sanktioner riktades mot skolor som inte uppfyllde kraven. I en stor utvärdering av effekterna av denna reform såg man att eleverna i åk 4 i genomsnitt signifikant förbättrat sina resultat i matematik (e=0,23, år 2007). Bland elever i åk 8 hade resultaten förbättrats särskilt bland svagpresterande grupper. Däremot, vilket är intressant, fann man inga effekter vad gäller läsning (Dee och Jacob 2011). Det tycks vara så att reformer inriktade på bedömning, eller där man använder test och betyg som *learning outcomes*, kan ha olika effekt på olika ämnen. Det är viktigt dock att komma ihåg att effekterna av NCLB är väldigt omdebatterade (Hout och Elliot 2011, Ravitch 2010). Se även kapitel 3 i denna översikt.

Nicole Schneeweis (2011) har också noterat att externa bedömningsinstrument kan ha positiva effekter för vissa ämnen men inte för andra. I en analys av faktorer som underlättar integrationen av elever med utländsk bakgrund, fann Schneeweis med utgångspunkt i PISA- och TIMSS-data att länder med externa tester inom NO-ämnen (science) lyckades bättre för dessa elever än länder där sådana kontrollinstrument saknas. En förklaring som förs fram är att externa tester minskar effekten av lärares subjektivitet (se även Wößmann 2005). Särskilt inom matematik och NO kan språksvaga grupper därmed gynnas.

Det är tydligt att bedömnings och *accountability*-system kan påverka elevers lärande och skolresultat. Frågan är hur mycket, och vad som är generella effekter oavsett ämne. OECD har kartlagt detta i relation till PISA-resultat. OECD skiljer mellan bedömningssystem som externa examinationer och som externa standardiserade test. Man skiljer också mellan *accountability*-system för benchmarking å ena sidan och en resultatinsamling med en mer aktiv betydelse för resurstilldelning och policy å andra sidan. OECD finner att i genomsnitt för de olika OECD länderna så förklarar skillnaden i bedömnings- och *accountablity*-system bara 2 procent av variationen i PISA-resultat (OECD 2010, s. 46.) Av de olika typer av bedömningssystem som tillämpas inom OECD är det enbart standardiserade externa examinationer som har en positiv effekt på lärandet (mätt i förhållande till resultat på läsförståelsetest, ibid, s. 47). *Accountability*-modeller har generellt ingen effekt vad gäller att höja ett lands PISA-resultat, men bidar något till en ökad likvärdighet (ibid, s. 78). Externa tester i kombination med *accountablity*-modeller bekräftar ojämlikheten inom skolor men minskar den mellan skolor (ibid., s. 47). Baserat på PISA 2012 fanns inga positiva samband mellan att offentliggöra skolors betyg och testresultat och skolornas resultat på PISA (OECD 2013, s. 57–59). OECD illustrerar skillnaden i hur olika länder använder resultatdata i figur 2.

■ Figure IV.3.6 ■

**How school systems use student assessments**

| | | Infrequent use of assessment or achievement data for benchmarking and information purposes | Frequent use of assessment or achievement data for benchmarking and information purposes |
|---|---|---|---|
| | | Provide comparative information to parents: 32% | Provide comparative information to parents: 64% |
| | | Compare the school with other schools: 38% | Compare the school with other schools: 73% |
| | | Monitor progress over time: 57% | Monitor progress over time: 89% |
| | | Post achievement data publicly: 20% | Post achievement data publicly: 47% |
| | | Have their progress tracked by administrative authorities: 46% | Have their progress tracked by administrative authorities: 79% |
| **Infrequent use of assessment or achievement data for decision making** | Make curricular decisions: 60%<br>Allocate resources: 21%<br>Monitor teacher practices: 50% | Austria, Belgium,[1] Finland,[2] Germany, Greece, Ireland, Luxembourg, Netherlands,[1] Switzerland,[1] Liechtenstein | Hungary, Norway,[2] Turkey, Montenegro, Tunisia, Slovenia |
| **Frequent use of assessment or achievement data for decision making** | Making curricular decisions: 88%<br>Allocating resources: 40%<br>Monitor teacher practices: 65% | Denmark, Italy, Japan,[2] Spain, Argentina, Macao-China, Chinese Taipei, Uruguay | Australia,[1] Canada,[2] Chile, Czech Republic, Estonia,[2] Iceland,[2] Israel, Korea,[2] Mexico, New Zealand,[1] Poland,[1] Portugal, Slovak Republic, Sweden, United Kingdom, United States, Albania, Azerbaijan, Brazil, Bulgaria, Colombia, Croatia, Dubai (UAE), Hong Kong-China,[2] Indonesia, Jordan, Kazakhstan, Kyrgyzstan, Latvia, Lithuania, Panama, Peru, Qatar, Romania, Russian Federation, Shanghai-China,[1] Singapore,[1] Thailand, Trinidad and Tobago, Serbia |

Note: The estimates in the grey cells indicate the average values of the variables used in latent profile analysis in each group. See Annex A5 for technical details.
1. Perform higher than the OECD average in reading.
2. Perform higher than the OECD average in reading and where the relationship between students' socio-economic background and reading performance is weaker than the OECD average.
Source: OECD, *PISA 2009 Database*.
StatLink http://dx.doi.org/10.1787/888932343399

***Figur 2. Hur olika skolsystem använder elevresultat (OECD 2010, s. 78).***

En anledning till varför *accountability*-modeller inte visar sig vara så effektiva på skolans område har att göra med komplexiteten i att leda lärare till att åstadkomma mer med sina elever. I en stor norsk studie jämfördes ett tydligt uttalat *accountablity*-system, Oslos gymnasieskolor, med dess motsats, norska folkhögskolor (Christophersen, Elstad et al. 2012). Studien visar att det som framför allt påverkar lärares arbete med eleverna, ur ett ledarskapsperspektiv, är relationerna mellan rektor och lärare, tilltro till lärares kompetens och social moral (i betydelsen en vilja att göra sitt yttersta för eleverna). Författarna menar att *accountability*-modeller inte nödvändigtvis omöjliggör detta men att dess logik snarare utgår från en misstro till professionen och en övertro på att lärares undervisningskvalitet kan påverkas med yttre incitament.

OECD (2012) pekar på vikten av att harmoniera yttre bedömningssystem med interna bedömningssystem. Ett tema i flera artiklar är just jämförelser mellan olika interna betygsmodeller. Efter att tittat närmare på dem återkommer vi med diskussion om relationen externa och interna bedömningsinstrument.

## Jämförelser av skolinterna bedömnings- och betygsmodeller

I skolan bedöms kunskap i många olika syften och ur flera olika perspektiv. Det kan handla om att ställa diagnos, få underlag till betyg eller examinera. Här finns inslag av bedömarsubjektivitet och ämnestraditioner. Allt detta skapar möjligheter till jämförelser mellan olika interna bedömningssystem och praktiker. Vi har särskilt valt att titta på de studier som finns om bedömning och betygssättning där forskarna jämfört olika ämnen.

Ett av betygssättningens mest tydliga tema är frågan om rättvisa (t.ex. Klapp Lekholm & Cliffordson 2008, Lundahl 2010). Det är emellertid inte en så stor forskningsfråga som man hade kunnat tro (Waldow 2014). En israelisk studie som vi också lyfte fram i kapitel 2 belyser dock hur centralt det är att systematiskt jämföra hur rättvisa betygen är, mellan ämnen och mellan lärare (Biberman-Shalev, Sabbagh, et al. 2011). En utgångspunkt för studien är att lärares disciplinära hemvist också leder till en viss betygssättningsstil (*grading style*). Tidigare

studier har exempelvis visat att språk och NO-lärare i Israel lade mer vikt vid elevens ansträngning än elevens uppvisade resultat än vad lärare i matematik gjorde. En anledning kan vara att matematikämnet är mer hierarkiskt strukturerat, där det ena så att säga bygger på det andra, medan språk och NO har en mer dynamisk karaktär (Resh 2009). De israeliska forskarna ville undersöka detta lite mer på djupet för att förstå om synen på ämnet som svagt eller hårt strukturerat påverkade hur man såg på elevernas prestationer, dvs. om synen på ämnet medierade ett specifikt pedagogiskt förhållningssätt. Studien som bygger på ett urval av 372 gymnasielärare kunde visa att synen på ämnen spelade stor roll, men man upptäckte att matematiklärare och de lärare som undervisade i grammatik hade mer gemensamt med varandra än vad matematik och NO lärare hade. Vilket forskarna förklarar med att grammatik liksom matematik är mer strukturerat än vad naturvetenskap är. Vad som är mest rättvist får dock inget svar i studien. Se även Korp 2006, Senk, Beckmann et al. (1997) för liknande resultat. Frågan om rättvis bedömning har också studerats i jämförande forskning i förhållande till en rad andra variabler: lärarnas etnicitet (Williams, Garza et al. 1999); lärares lön (Dolton & Marcenaro-Gutierrez 2011); elevernas kön, klass och etnicitet (t.ex. Mechtenberg 2009); elever självbild (Nowell & Alston 2007, Marsh, Trautwein et al. 2005, Boehnke 2005).

En annan faktor som jämförts är betydelsen av att bedöma och betygsätta examensarbeten via dator eller via papper och penna format. En studie (omfattande 80 lärare) från Hong Kong visar att lärare som inte hade tidigare erfarenhet av att bedöma icke-digitala arbeten bedömde de digitala arbetena på ett tillförlitligt sätt redan från början, medan lärare med lång icke-digital erfarenhet hade problem med att göra korrekta bedömningar vid första bedömningstillfället. Studien visare emellertid att lärare som enbart haft erfarenhet av analoga prov relativt snart kunde göra korrekta bedömningar i datorformatet. Forskarna bakom studien argumenterar för att elever kan bli rättvist bedömda relativt fort vid en övergång från analoga prov till digitala prov (Coniam 2009). Den här typen av studier kan ha stor betydelse inför reformer för digitalisering av nationella prov.

Ett intressant och okonventionellt sätt att sätta betyg undersöktes i en studie av Danielewicz och Elbow (2009). De bestämde att studenterna (i en universitetskurs i pedagogik) skulle få betyget B (näst högsta betyg) om de lämnade in sina uppgifter. På varje uppgift fick eleverna feedback. Studenterna visste att de fick minst B så länge de följde ”kontraktet” och lämnade in, men om de också tog till sig av återkopplingen och förbättrade sin portfölj kunde de få ett A. För B skulle studenterna följa detta kontrakt:

You are guaranteed a B if you:

1. attend class regularly – not missing more than a week’s worth of classes:
2. meet due dates and writing criteria for all major assignments;
   participate in all in-class exercises and activities;

3. complete all informal, low-stakes writing assignments (e.g. journal writing or discussion-board writing):
4. give thoughtful peer feedback during class workshops and work faithfully with your group on other collaborative tasks (e.g., sharing papers, commenting on drafts, peer editing online discussion boars);
5. sustain effort and investment on each draft of all papers;
6. make substantive revisions when the assignment is to revise – extending or changing the thinking or organization – not just editing or touching up;
7. copyedit all final revisions of main assignments until they conform to the conventions of edited, revised English;
8. attend conferences with the teacher to discuss drafts;
9. submit you midterm and final portfolio. (Danielewicz och Elbow 2009, s. 245f)

Forskarna upplevde att studenterna kunde fokusera bättre på skrivprocessen när de inte behövde oroa sig för betygen. Kontraktsbetyg av det här slaget sätter fokus på processen snarare än slutresultaten, vilket kan vara lämpligt för vissa mål eller ämnen. Studien är dock liten och svår att dra generella slutsatser från. För liknande studier av alternativa sätt ett sätta betyg se även Muñoz & Álvarez (2010) och Gijbels, Dochy et al. (2005).

# Jämförelser av externa och interna bedömningsmodeller

Att det finns skillnader mellan olika bedömare och mellan externa och interna bedömare har diskuterats i svenska styrdokument och utbildningspolitiska texter åtminstone sedan 1700-talet (Lundahl 2006). Ett sista tema som visar sig i den jämförande forskningen om betyg och bedömning handlar om jämförelser ifråga om validitet och reliabilitet mellan externa och interna bedömningsmodeller. I artikeln Grades and Test Scores (Willingham, Pollack et al. 2002) ställs den enkla frågan: varför skiljer det sig åt?

I en tabell listar författarna flera olika variabler där test och betyg sannolikt kan skilja sig åt:

***Tabell 9. Möjliga orsaker till varför betyg och testresultat skiljer sig åt (Willingham, Pollack et al. 2002, s. 4).***

A. Content Differences Between Grades and Test Scores

1. Domain of general knowledge and skill
   a. Subjects covered, such as science and history; broad divisions within subjects, such as physics or European history
   b. General cognitive skills, such as reasoning, writing, or performance
2. Specific knowledge and skills as reflected in
   a. Course-based content throughout the school district, state, or nation (especially relevant to an external test)
   b. Classroom-based content (especially relevant to a teacher's grade)
   c. Individualized content (especially relevant to personal interests, skills, and course of study)
3. Components other than subject knowledge and skills
   a. Social objectives of education (e.g., leadership, citizenship)
   b. Academic and personal development (e.g., attendance and participation, completing assignments, disruptive behavior, effort and coping skills, interpersonal competence)
   c. Assessment skills and debilities (pertinent to test-taking or class assignments, genetal or specific to particular assessments, construct relevant or irrelevant, confidenceor anxiety)

B. Individual Differences That Interact With Content Differences

1. Early development and relevant learning acquired outside of school
2. Student characteristics that can affect academic motivation
   a. Behavior in and out of school
   b. Attitudes about school and learning
   c. Family circumstances
3. Teacher judgment regarding the student's performance

C. Situational Differences

   1. Differences across contexts

   2. Differencesover time

D. Errors in Grades or Test Scores

   1. Systematic error-Noncomparability
      a. Variation in grading standards (across schools, courses, teachers, and sections)
      b. Variation in test score scales (across forms; across time)
      c. Cheating (by students or schools, on class assignments or tests)

   2. Unsystematic measurement error-Unreliability (in grades and in test scores)

Med utgångspunkt i tabellen tog forskarna fram fem faktorer som kunde tänkas förklara skillnaderna mellan test och betyg: *subject covered*; *grading variations*; *reliability*; *student characteristics*; *teacher ratings*. Genom att analysera test och betygsresultat samt registerdata för 10849 *high school*-studenter försökte forskarna bestämma hur dessa faktorer bidrog till skillnaderna. Det som hade störst betydelse för skillnaderna mellan betyg och testresultat var skillnader i betygspraktik mellan skolor, avvikelser från kursinnehåll samt elevernas engagemang och attityder till skolan. Resultatet visar att de sammantaget förklarar upp mot 80 procent av variationen mellan elevens betyg och testresultat. Det innebär likväl att det är svårt att använda det ena för att predestinera det andra på individnivå. Forskarna menar att det är oerhört viktigt att förstå att testresultat och betyg ömsesidigt bör validera varandra. Två viktiga skäl till det är att betyg rymmer så mycket komplexitet, så många observationer, samt att elevers prestationer faktiskt kan variera ganska mycket mellan dessa olika observationstillfällen.

Att det blir så olika utfall av olika mätningar kan förklaras med hjälp av klassisk testteori. Det finns mätfel även i de allra mest ambitiöst konstruerade proven. Ska man göra vettiga tolkningar av ett prov måste man därför ha en aning om mätfelets storlek. Black och Wiliam (2012) har ett resonemang om hur man kan tänka om det genom att ha en hypotes om elevernas rätta resultat. Det finns inget meningsfullt prov där elever skulle få samma resultat varje gång. Elever gör olika fel vid varje mättillfälle och bedömare gör olika rättningar vid olika tillfällen. Men om man lade ihop en elevs resultat på fem till sex liknande prov under en begränsad tid skulle man få fram ett genomsnittligt resultat som kallas *true score* – det rätta resultatet. Ett sätt att åstadkomma detta i praktiken är att arbeta med *split half* metoden som innebär att man gör ett prov som kan delas i två delar. Sedan jämför man utfallet på de två delarna. Är det hög överensstämmelse har uppgifterna en hög *inre konsistens* avseende vad de mäter. Överensstämmelsen är dock också beroende av hur man delar upp testet och därför måste man korrelera alla tänkbara rimliga sätt att dela testet på med varandra. Då får man ett värde som kallas Cronbachs alpha och som uttrycks mellan 0 och 1, där 0 betyder att proven ger slumpmässiga utfall och 1 att provet är helt reliabelt – varje gång vi gör det får vi samma resultat. En vanlig uppfattning är att Cronbach alpha bör ligga på 0.7 och uppåt om testet ska vara användbart, men det beror givetvis på vad det faktiskt är man mäter. För att förstå vilken effekt olika grader av reliabilitet faktiskt kan få t.ex. för vilket provbetyg en elev får behöver vi kombinera Cronbach alpha med ett mått på elevens sanna resultat (*the true score*).

För att undersöka hur ett provs inre konsistens påverkar resultatet för en elev kan Cronbach alpha sättas i relation till standardavvikelsen, dvs. den genomsnittliga avvikelsen från medelvärdet. En bra illustration till hur man kan räkna finns i Black och Wiliam (2012). På en normalfördelningskurva faller 68 procent av resultaten inom en standardavvikelse och 96 procent inom två standardavvikelser. Genom att kombinera dessa mått går det att få fram ett förväntat *standardfel*, SEM. Standardfelet anger för varje reliabilitetsnivå den förväntade spridningen av felprocent inom en och samma faktisk kunskapsmängd.

Formeln för SEM är $X \sqrt{(1-r)}$

Om r är reliabilitet så betyder detta att SEM på ett prov med en reliabilitet på 0.85, där man kan få 50 poäng och där standardavvikelsen (X) är 7,5 poäng blir 2,9 poäng (SEM=7,5$\sqrt{}$(1-0,85)=2,9). Det innebär att den ”sanna poängen” för en elev med 35 provpoäng till 68 procents sannolikhet ligger mellan 32 och 38 poäng. Vill man ha 95 procents säkerhet kan man säga att den ligger mellan 29 och 41. Detta är i själva verket en approximation, men används allmänt i professionell provanalys. Detta innebär hursomhelst att i en klass på 30 elever så är det minst en elev, vi vet aldrig vem, som avviker mer än 12 procent i positiv eller negativ riktning från sitt riktiga resultat. Minst tio elever avviker 6 procent från sitt sanna resultat relaterat till provens bristande inre konsistens (som i det här exemplet trots allt inte var så farligt hög). Effekten för den enskilda individen kan bli enorm. Black och Wiliam skriver:

> even the best tests can be widly inaccurate for a few individual students /…/ This is why testing experts invariably say that high-stakes decisions should never be based solely on the results of a single test. (Black och Wiliam 2012, s. 252)

Det är också av detta och likande skäl som nationella prov inte ska styra elevernas betyg. Om de gör det kommer vissa elever ändå att få fel betyg. Se också Brennan, Kim et al. (2001) för en liknande diskussion. En möjlig slutsats av den här typen av studier är att en kombination av externa testresultat och betyg vore mer rättvisande än enbart betyg eller enbart externa prov (se även kapitel 3). Vi ska strax visa hur olika länder i Europa har löst detta där några länder som t.ex. Danmark och Finland har just denna typ av kombination.

Innan vi kommer in på våra egna jämförelser av betygssystemen i Europa ska vi peka på några slutsatser av vår genomgång av forskning om betyg ur komparativa perspektiv.

## Diskussion och slutsatser

När vi söker på bedömning och internationella jämförelser ser vi att betyg inte får en särskilt framträdande plats i artiklarna. I huvudsak är det tre områden forskarna fokuserar vid dessa jämförelser: det är för det första system för *accountability*, för det andra kulturella förklaringar till varför bedömnings- och betygssystem ser olika ut i olika länder och för det tredje variationer mellan olika lärares bedömningar i olika ämnen eller av olika elevgrupper. De länder som dominerar i artiklar kring internationella jämförelser är USA och England men även Tyskland, Israel, Sverige, Frankrike, Kina och Japan förekommer i fler än en artikel om internationella jämförelser kring *assessment*/*grading*.

Några viktiga iakttagelser i vår genomgång är att det länge, vilket torde vara välkänt för de flesta, funnits en internationell trend mot att upprätta olika system för ökad ansvarsskyldighet (*accountability*) för skolans resultat. Dessa resultat mäts främst i elevprestationer på test eller i betyg. Denna trend är delvis en effekt av internationella kunskapsmätningar och OECDs ökande inflytande på den internationella policydiskursen, men det hänger också ihop med en restrukturering av nationella välfärdssystem (se vidare *Assessment in education – country profiles* http://www.tandf.co.uk/journals/pdf/AIEProfiles.pdf). Precis som flera kritiska studier har även OECD noterat, att förhoppningen om att jämförelser av skolors resultat ska leda till resultatförbättringar varit väl optimistiska. Studier av den amerikanska NCLB-reformen visar på blandade resultat och att effekten kan variera mellan skolor och mellan ämnen. En del menar att *accountability*-modeller borde avvecklas medan andra forskare pekar på att de kan utvecklas. Uppenbarligen behöver i alla fall externa modeller bättre kalibreras med skolans behov och organisatoriska villkor. Aktörer på skolnivå behöver verktyg för att analysera och förstå resultat som de kan påverka.

Vi finner inga indikatorer i komparativ forskning på att något betygssystem skulle vara bättre än något annat. OECD (2010) har pekat på att länders system för kunskapsbedömning förklarar i det närmaste ingenting av variationen i PISA-resultat (två procent). Andra studier pekar på att externa bedömningar kan bidra till ökad rättvisa i det att de på olika sätt reducerar subjektiviteten i lärares bedömningar. Detta har visat sig exempelvis gagna språksvaga elever i ämnen där språket inte är det som huvudsakligen ska bedömas. OECD 2013 har pekat på att vissa typer av betygsbeteckningar leder till större likvärdighet, men det återkommer vi till nedan.

De skillnader som finns mellan länders olika betygs- och bedömningssystem har ofta djupa kulturella rötter och är starkt kopplade till varje lands specifika sätt att organisera sitt skolväsende på. Vi ska därför titta närmare på betygen i Europa, hur och varför de skiljer sig åt.

# Betygen i Europa

I ett uttalande om vikten av att börja med betyg tidigare sa tidigare utbildningsminister Jan Björklund: ”I stort sett hela världen börjar med betyg tidigare än vad Sverige gör. De allra flesta ger betyg från första klass. Vårt grannland Finland ger från trean-fyran. Länder som ligger högt i Pisa börjar väldigt tidigt” (SVT 20/8 2014). Hur han vet detta är en fråga i sig. Det finns en lista i en snabbutredning om betyg i tidigare åldrar (Utbildningsdepartementet PM 2014-08-20, s. 37) där referensen går till ”(bl.a. OECD 2013)”. I de publikationer från 2013 som OECD släppt finns dock inte dessa uppgifter, däremot i en av dem vilka betygsskalor som används i olika OECD länder (OECD 2013). På en direkt fråga till utredaren om var dessa uppgifter kommer ifrån hänvisar han till att det fanns en lista på utbildningsdepartementet. När vi frågar ansvarig tjänsteman på utbildningsdepartementet får vi veta att:

> Det verkar tyvärr som att det har blivit ett fel vid referenshänvisningen när det gäller tabellen över läsåret och elevens ålder för första betyg i skolan (s 37). Det är Utbildningsdepartementet som har tagit fram ett eget underlag med dessa uppgifter. Informationen är dels hämtad från Nätverket Eurydike, som tillhandahåller information om och analyser av utbildningssystem och utbildningspolitik i Europa, dels genom kontakter med utbildningsdepartement eller motsvarande i övriga länder. (Mejl U2014/5534/S Skolenheten, Utbildningsdepartementet 7 oktober 2014)

Den lista som användes i utredningen om betyg från årskurs 4 kan mycket väl vara riktig men går det att kontrollera informationen? Frågan i det här avsnittet är, hur ser betygssystem ut i andra länder och vilka jämförelser är alls möjliga att göra och vilka är kloka att göra? I vilken utsträckning menar vi samma sak när vi i olika länder i Europa talar om betyg?

Grunden för den här översikten och sammanställningen av betygssystem i Europa i Appendix är främst den information som EU-kommissionens organ *Education, Audiovisual, and Culture Executive Agency* har publicerat i databasen Eurydice. Onlineplattform Eurypedia har information från medlemsländer i EU, samt Bosnien och Herzegovina, Island, Liechtenstein, Montenegro, Makedonien, Norge, Serbien och Turkiet (Eurydice 2014a). Information om ländernas bedömningssystem samlade vi in under hösten 2014. Dessutom har vi använt två av Eurydice schematiska diagram som visar den obligatoriska utbildningen i Europa (2014b) och strukturen för de europeiska utbildningssystemen 2014/15, som publicerades i november 2014 (Eurydice 2014c). Vi har med andra ord försökt få fram så uppdaterade uppgifter som möjligt.

Vidare bygger exemplen på ”Country profiles” som har publicerats för europeiska länder i tidskriften *Assessment in Education: Principles, Policy & Practice* 1997-2014. Vi har också använt några av OECD:s underlagsrapporter: *Review on Evaluation and Assessment Frameworks for Improving School Outcomes*. Dessa tidskriftsartiklar och rapporter har dock inte använts för att komplettera information som saknas i Eurydice. Snarare har de använts för att upparbeta de kategorier vi skapat i Appendix, och för att ge en något fylligare beskrivning.

Att skapa kategorierna visade sig vara ett ganska tidskrävande tolkningsarbete. Det som i Eurydice förefaller okomplicerat visar sig ofta ha flera nyanser. I några fall är engelskan som används rent ut sagt dålig. Detaljnivån på informationen varierar mycket, och informationen är producerad av väldigt olika typer av författare. Den metodologiska utmaningen att använda Eurydice till att sammanställa information om betygsystemen i Europa är en poäng i sig själv som vi gärna vill framhålla. Det framgår klart av tabellen i Appendix att för många teman är information för ojämn för att användas som jämförelseunderlag. Det är alltså med största försiktighet vi ska dra slutsatser om jämförelser med andra länders betygssystem utifrån enbart den här typen av data. Det gäller såväl i forskning som i policy.

## Elevernas ålder vid betygssättning i Europa

Om vi studerar de beskrivningar som finns om europeiska länders betygssystem i Eurydice (enligt den information som fanns där hösten 2014) kan vi för det första konstatera att 12 länder inte alls nämner när de börjar sätta betyg.

Detta betyder att det är svårt att kontrollera för generella påstående om hur många länder som sätter betyg i vilka åldrar, då den informationen ibland bara finns på originalspråk via ländernas statliga myndigheter. Dessutom är det intressant att så många länder inte tycker detta är viktig information. I de beskrivningar som finns av när och hur betyg ges i olika europeiska länder är det tydligt att betygen i många fall inte är ett särskilt väl reflekterat pedagogisk fenomen.

Helt klart är det i alla fall att de flesta länder börjar mycket tidigare med betyg än vad vi gör i Sverige – men av olika anledningar och på mycket olika sätt.

***Tabell 10: Tidiga betyg***

| | |
|---|---|
| Ålder 5 | **1 land:** Nordirland |
| Ålder 6 | **9 länder:** Cypern, England, Frankrike, Ungern, Italien, Polen, Rumänien, Wales, Österrike |
| Ålder 7 | **4 länder:** Belgien (franska), Luxemburg, Turkiet, Tyskland (i flertalet delstater) |
| Ålder 8 | **3 länder:** Grekland, Malta, Slovenien |
| Ålder 9 | **2 länder:** Portugal, Slovakien, |
| Ålder 10 | |
| Ålder 11 | **2 länder:** Lichtenstein, Litauen |
| Ålder 12 | **1 land:** Sverige |
| Ålder 13 | **1 land:** Norge |
| Ålder 14 | **2 länder:** Danmark, Finland* |

Ej redovisat (12): Belgien (flamländska), Belgien (tyska), Bulgarien, Estland, Irland, Island, Kroatien, Lettland, Nederländerna, Skottland, Spanien, Tjeckien.

* I Finland varierar tidpunkten för första betygen lokalt, se beskrivning nedan.

Betyg vid sex års ålder (som oftast är åk 1) är alltså, utifrån den information som finns i Eurydice, den tidpunkt som samlar flest länder i Europa, även om det totalt sett är vanligare att vänta till åk 2, 3 eller 4. Betyg i åk 6 och senare ges i ett färre antal länder, givet den information som finns via Eurydice. Det finns dock också några länder som inte direkt reglerar när betyg ska ges. Hur betygen ges varierar dock mellan länderna även när de ger dem vid samma tidpunkt.

Om vi bara uppehåller oss vid länder som ger betyg i åk 1 kan den stora variationen av system och principer för hur betyg blir till, enkelt framträda (se även Lundahl 2014). Bland länder som ger betyg i åk 1 har vi Österrike som ger ett samlande betyg i slutet av åk 1, ibland tillsammans med muntliga tillägg. Först från åk 2 får barnen betyg i alla ämnen. Även Lettland tillämpar ett system där eleverna får ett mer kvalitativt övergripande omdöme i åk 1, för att sedan få betyg på skalan 1-10 från åk 2 i modersmål och matematik. För de andra ämnena används mer kvalitativa omdömen fram till åk 4.

I Cypern får eleverna varje år från åk 1 ett framstegscertifikat som visar om eleven klarat av årskursen. Detta är en förutsättning för att få påbörja nästa årskurs. Vid slutet av åk 6 får eleven ett avgångsbetyg från *primary school*. Också i Litauen är godkända resultat från åk 1 en förutsättning för att flyttas upp. Eleverna kan flyttas upp innan de avslutat klass 1-3 om de anses klara med det läroplanen täcker för dessa år.

I Finland har eleverna rätt till en avrapportering från åk 1 om hur de fortskrider i studierna. Det bestäms i den lokala arbetsplanen hur detta ska se ut, om det ska utformas verbalt eller numeriskt, eller som en kombination av dem upp till åk 7. Därefter är betygen numeriska. Det är därför i bästa fall oprecist att tala om att man i Finland sätter betyg från trean eller fyran, som i citatet som inleder detta avsnitt.

Frankrike tillämpar ett system där man följer elevernas studier i bestämda cykler och i slutet av en sådan cykel ges ett prov. Resultaten från dessa prov sammanfattas i en bok (*livret scolaire*). Denna rapportbok används i kommunikationen mellan lärare och föräldrar och inför bytet mellan olika stadier, i syfte att ge en kontinuerlig bild av barnens utveckling.

I Turkiet gör eleverna ett färdighetstest i slutet på åk 1 för att avgöra vilken undervisning de ska få från åk 2. I slutet av varje årskurs från åk 2 till åk 8 gör eleverna sedan ett prov på vilket de måste få betyget *Fair* (3) för att få avsluta årskursen.

Ungern har ett system där lärarna ska ge eleverna regelbundna ”marks” under året och sammanfatta dessa i slutet med ett betyg. I Italien får eleverna i slutet av varje läsperiod olika kommentarer som sammanfattas i bedömningsdokument i slutet av året. Eleverna får också omdömen i uppförande. Polen och Schweiz, slutligen, där får eleverna betyg på skala 1 – 6 från åk 1 (se vidare Appendix).

Det finns länder där elevernas kunskapsnivå bestäms i tidiga år, även om det inte nödvändigtvis är så att betyg används för bedömningen. Det är bedömningen som är reglerad, inte formen för den som till exempel på Cypern. Det är också så att betyg i vissa länder, vad gäller de lägre åldrarna är ett resultat på ett färdighetsprov, och inte ett kursplanestyrt betyg så som vi konceptualiserar betyg i Sverige.

Flertalet landsbeskrivningar i Eurydice ger inga indikationer på att betygssystemen reformertas de senaste decennierna. Några få länder har reformerat sina betygssystem mot senare betygssättning. Liechtenstein avskaffades betyg helt i primärskolan 1999/2000. I Slovakien har man sedan några år tillbaka enbart krav på muntlig bedömning upp till åk 4. I Litauen har det skett en övergång mot framför allt beskrivande och kvalitativa omdömen i primärskolan.

Spelar det då någon roll för resultaten på PISA-mätningarna när i ålder eleverna får sina första betyg? Om vi samkör den information vi har med vilka resultat eleverna i Europa fick på PISA-mätningen 2012 ser vi inga signifikanta effekter av att börja tidigt eller sent (tabell 11).

Korrelationer (Pearsons *r*) mellan Pisa-resultat i matematik, läsning och naturkunskap och ålder för betygsättning i olika länder är icke signifikanta ($r$ = 0, 12 till 0, 20; $p$ = 0, 271 - 0, 524). Detta resultat innebär att det inte finns några samband mellan ålder för betygsättning och Pisa-resultat i matematik, läsning och naturvetenskap. Vidare är resultaten för regressioner på ålder för betygsättning för Pisa-resultat för matematik, läsning och naturkunskap icke signifikanta och visar på att det inte finns några orsakssamband mellan när länder börjar sätta betyg och PISA-resultat. Detta resultat, som har länder som analysnivå (aggregerad nivå) kan skilja sig mot resultat där analyserna är på individnivån (eleven). Vi har dock inte haft tillgång till individdata.

*Tabell 11. Sambandet mellan betygsstart i Eurydice och PISA-resultat*

**Correlations**

| | | pisa_ma | pisa_read | pisa_scie | Grading |
|---|---|---|---|---|---|
| pisa_ma | Pearson Correlation | 1 | ,870** | ,903** | ,201 |
| | Sig. (2-tailed) | | ,000 | ,000 | ,271 |
| | N | 36 | 36 | 36 | 32 |
| pisa_read | Pearson Correlation | ,870** | 1 | ,914** | ,117 |
| | Sig. (2-tailed) | ,000 | | ,000 | ,524 |
| | N | 36 | 36 | 36 | 32 |
| pisa_scie | Pearson Correlation | ,903** | ,914** | 1 | ,133 |
| | Sig. (2-tailed) | ,000 | ,000 | | ,467 |
| | N | 36 | 36 | 36 | 32 |
| Grading | Pearson Correlation | ,201 | ,117 | ,133 | 1 |
| | Sig. (2-tailed) | ,271 | ,524 | ,467 | |
| | N | 32 | 32 | 32 | 33 |

**.Correlation is significant at the 0.01 level (2-tailed). PISA-data från kommer från OECD PISA web site (revised edition February 2014), Excel snapshot file from chapter 1: http://dx.doi.org/10.1787/888932937035. Dessa kompletterades för UK och Belgien från OECD:s landsrapporter: http://www.oecd.org/unitedkingdom/PISA-2012-results-UK.pdf och: http://www.oecd.org/education/PISA-2012-results-belgium.pdf. (nedladdat 2015-01-19)

Samma resultatbild får vi om vi utgår ifrån den skala Utbildningsdepartementet har tagit fram (tabell 12). I utredningen En bättre skolstart (Promemoria 2014-08-20) finns en lista över när betyg sätts i 36 länder inkluderande även USA och några länder i Asien, och några länder i Sydamerika. Vi har inte kunnat verifiera hur mycket på denna lista det är som stämmer, men konstaterar att för några länder sätts betyg egentligen något senare än vad utbildningsdepartementet anger (s. 37). Trots det finner vi inte heller här att det finns något samband, mellan betygsstart och resultat på PISA. Det finns med andra ord inga belägg för antydda samband av slaget: ”Länder som ligger högt i Pisa börjar väldigt tidigt med betyg”.

***Tabell 12. Sambandet mellan betygsstart enligt Utbildningsdepartementets förteckning och PISA-resultat***

**Correlations**

| | | pisa_ma | pisa_read | pisa_scie | Utb. dep. |
|---|---|---|---|---|---|
| pisa_ma | Pearson Correlation | 1 | ,926** | ,940** | ,125 |
| | Sig. (2-tailed) | | ,000 | ,000 | ,448 |
| | N | 46 | 46 | 46 | 39 |
| pisa_read | Pearson Correlation | ,926** | 1 | ,950** | ,012 |
| | Sig. (2-tailed) | ,000 | | ,000 | ,942 |
| | N | 46 | 46 | 46 | 39 |
| pisa_scie | Pearson Correlation | ,940** | ,950** | 1 | ,019 |
| | Sig. (2-tailed) | ,000 | ,000 | | ,908 |
| | N | 46 | 46 | 46 | 39 |
| Utb. dep. | Pearson Correlation | ,125 | ,012 | ,019 | 1 |
| | Sig. (2-tailed) | ,448 | ,942 | ,908 | |
| | N | 39 | 39 | 39 | 40 |

**. Correlation is significant at the 0.01 level (2-tailed). PISA-data från kommer från OECD PISA web site (revised edition February 2014), Excel snapshot file from chapter 1: http://dx.doi.org/10.1787/888932937035. (nedladdat 2015-01-19)

## Betygsskalor i Europa

Betygen i Europa varierar utöver när de sätts också utifrån hur många skalsteg betygen ges i och hur dessa uttrycks (tabell 13 och 14).

***Tabell 13 antal skalsteg:***

| | |
|---|---|
| >10 nivåer | **1 land:** Frankrike (20 nivåer) |
| 9 nivåer | **5 länder:** England, Nordirland, Skottland, Turkiet, Wales |
| 8 nivåer | |
| 7 nivåer | **2 länder:** Finland, Lettland |
| 6 nivåer | **6 länder:** Bulgarien, Danmark, Norge, Polen, Sverige, Tyskland |
| 5 nivåer | **9 länder:** Cypern, Estland, Grekland, Island, Kroatien, Slovakien, Slovenien, Ungern, Österrike |
| 4 nivåer | **2 länder:** Portugal, Rumänien |
| Ej redovisat | **12 länder:** Luxemburg, Italien, Malta, Tjeckien, Belgien (flamländska, franska, tyska), Irland, Nederländerna, Spanien, Lichtenstein, Litauen |

Bland de länder som redovisar hur många steg bedömningen sker i är fem skalsteg vanligast (9 länder) medan sex länder tillämpar en 6-gradig skala. Det är emellertid viktigt att vara försiktig vid klassificeringen av dessa data, eftersom det kan se ut som att några länder rapporterar hur många skalsteg de använder utan att precisera vilka gradbeteckningar som används (t.ex. Bulgarien och Estland).

Frankrike skiljer sig från de övriga länderna i Europa med hela 20 betygsnivåer. Som det framgår av *country profile* artikeln är det dock inte nödvändigtvis så att alla nivåer används i praktiken:

> Marking, which is generally used at all levels, is also a personal affair even though the tendency in France, where the marking scheme is from 0 to 20, is to award marks in the medium range rather than to use extremes. A very good piece of work will rarely be rated 18 or 20; more likely 14 or 15. (Bonnet 1997, s. 296)

I tabell 13 och i Appendix är länderna klassificerade utifrån antal skalsteg från brytpunkten till underkänt, men det blir lite missvisande då några länder har flera nivåer för underkänt. I Danmark ligger gränsen är underkänt betyg 00, men eleverna kan också få ned till -3 vilket alltså är ännu mera underkänt. Lettland är ett annat exempel där man har en betygsskala som ser helt annorlunda ut. Enligt Eurydice har de det *fyra* nivåer för underkänt: 4 (Almost satisfactory), 3 (Weak), 2 (Very weak), 1 (Very, very weak).

Tjänsten STUDYinEUROPE (http://www.studyineurope.eu/grades) erbjuder elever och förälder ett verktyg för att möjliggöra jämförelser av betygsystem i Europa, så man kan se hur ens betyg kommer att ”översättas” om man börjar i skolan i ett annat land. Det är oklart för oss hur den har tagits fram och exempelvis tar den inte hänsyn till att olika länder sätter godkäntgränsen på olika sätt., STUDYinEUROPE lider av samma översättningsproblematik som Eurydice. Frågan om vem det är som har auktoritet och legitimitet för att göra dessa översättningar blir central. Vår samanställning visar hur problematiskt det kan vara att göra jämförelser och det är betänklig att flera länder, som Sverige, ändrar sina betygsskalor till ECTS (6-punkter) och därigenom skapar intryck av att betygen är jämförbara. Att betygsskalorna är jämförbara betyder givetvis inte att kunskapsstandarden för respektive nivå är det. Det är lite som att jämföra bilar utifrån karossen och inte bry sig om motorn i dem.

***Tabell 14: Typ av skala***

| | |
|---|---|
| Verbal och numerisk | **11 länder:** Österrike, Bulgarien, Island, Ungern, Slovakien, Estland, Norge, Polen, Tyskland, Finland, Lettland. |
| Numerisk | **5 länder:** Danmark, England, Nord Irland, Slovenien, Wales |
| Verbal | **2 länder:** Kroatien, Rumänien |
| Verbal och bokstäver | **2 länder:** Cypern, Grekland |
| Bokstäver | **1 land:** Sverige |
| Poängskala | **2 länder:** Luxemburg, Malta |
| Ej redovisat | **14 länder:** Frankrike, Italien, Skottland, Turkiet, Tjeckien, Belgien (franska, flamländska, tyska), Irland, Nederländerna, Lichtenstein, Portugal, Litauen, Spanien |

Bland de länder som redovisat vilken typ av skala som används dominerar betygssystem som använder en kombination av verbala och numeriska betyg. Det kan till exempel se ut som i Finland där man vanligen börjar med numeriska betyg först högre upp i åldrarna. En faktor som spelar roll för hur skalstegen har utformats är om landet haft en tradition av relativ bedömning eller kriterierelaterad. Detta spelar också viss roll för om betygen sätts med fokus på vad eleverna inte kan eller vad de kan. Betygssättning utifrån en normalfördelningskurva ledde som tidigare i Norge till att betygsskalan var indelad i relativa nivåer fram till 2007, långt efter det att det relativa betygssystemet faktiskt avskaffats. Det kunde bidra till att lärare återkopplade till eleverna utifrån betygskriterier som mer hade fokus på vilka kunskaper eleverna saknade (t.ex. din text har ingen tydlig struktur, din text är ologisk, ditt språk är oprecist och du har många stavfel) snarare än vad de behövde arbeta med (se vidare Tveit 2014). I kapitel 1 visar vi att den typ av feedback som ger bäst

resultat är den motsatta, det vill säga den som i god tid ger information om vad eleverna konkret behöver göra för att nå längre. Det är med andra ord viktigt att förstå hur olika skalor kan föra med sig olika fokus i bedömningen av elevers kunskaper. I sammanhanget kan vi också nämna att ECTS skalan i många länder tillämpas som en relativ skala och inte som i Sverige som en kriterierelaterad skala (se Ds 2008:13). Det är också viktigt att komma ihåg att man i vissa länder har olika typer av betyg för olika ämnen. I exempelvis Nederländerna kan man i gymnastik, konst och specialarbete bara få betyget Tillfredställande eller God.

Värt att notera här är att OECD (2012) konstaterar att likvärdigheten i bedömning är bättre i utbildningssystem som använder en verbal kvalitativ skala av slaget: Very good, Good, Satisfactory, Sufficient, Insufficient. De menar att dessa kvalitativa uttryck är allmänna och lätta att förstå och relatera till oavsett det gäller kvalitet på mat, kläder eller kunskaper.

Vi också försökt få fram vilka länder de är som använder betyg i ordning och uppförande. I Europa är det enligt Eurydice: Norge, Italien, Rumänien, Polen, Slovakien, Ungern, samt ungefär hälften av delstaterna i Tyskland.

## Betygsystem och skolorganisation

En bakomliggande förklaring till variationen i betygsättning är skolsystemens skilda strukturer (se Appendix). En markerad gränsdragning mellan primär- och sekundärskola har haft en viss betydelse för ett lands betygssystem då det krävts någon form av reglering vid övergången. Norden tillsammans med Bulgarien, Estland, Lettland, Kroatien, Portugal, Slovakien, Slovenien, Spanien, Turkiet och Ungern saknar en sådan uppdelning. För vissa länder med uppdelning mellan de lägre och högre åren i grundskolan finns dessutom en organisatorisk differentiering mot studier eller yrkesarbeten, t.ex. Tyskland, Österrike och Nederländerna (se tabell 15, se även Lundahl, Román & Riis 2010).

***Table 15: Typ av utbildningsstruktur***

**Single structure** – sammanhållen obligatorisk grundskola;
**Primary secondary** – grundskolans högre årskurser tillhör sekundärskolan tillsammans med gymnasiet. Vissa selektionsinslag kan finnas redan vid övergången till ”högstadiet”, eftersom eleverna i flera av dessa utbildningssystem måste nå en särskild kunskapsnivå för att komma in på ”högstadiet”.
**Tracked secondary** – elever väljer inriktning till sekundärskolan.

| | |
|---|---|
| **Single structure** | **15 länder:** Bulgarien, Danmark, Estland, Finland, Island, Kroatien, Norge, Portugal, Slovakien, Slovenien, Spanien, Sverige, Turkiet, Ungern |
| **Primary Secondary** | **12 länder:** England, Nordirland, Skottland, Wales, Cypern, Rumänien, Frankrike, Polen, Grekland, Malta, Tjeckien, Litauen |
| **Tracked secondary** | **9 länder:** Belgien (franska, flamländska och tyska), Irland, Lichtenstein, Luxemburg, Nederländerna, Tyskland, Österrike |
| **Inte konsistent beskrivit** | **2 länder:** Italien, Lettland |

Det är viktigt att komma ihåg att de flesta länder har en brytpunkt mellan primär och sekundärskolan efter det som motsvarar åk 6 i Sverige (se Appendix). Notera också att *Länderna* i Tyskland har olika sätt att organisera övergången från primärskola till sekundärskola. I vissa fall väljer man studie- eller yrkesinriktning på sekundärskolan redan i åk 4 andra Länder i åk 5 eller åk 6. Övergången går också till på olika sätt, i vissa *Länder* genom betyg, inträdesexamination och i andra *Länder* genom att man läser ett halvt år på försök.

Hur betyg sätts i de europeiska länderna har ofta landspecifika orsaker kopplade till landets utbildningshistoria. Ett exempel som kan illustrera hur det politiska och kulturella sammanhanget har betydelsefulla konsekvenser för bedömningssystemets utformning är Portugal. Den obligatorisk utbildning i Portugal brukade vara fyra år lång fram till slutet av Francesco-regimens fall 1974. Strax efter att demokrati

införts förlängdes den obligatoriska utbildningen till sex år, och ytterligare en förlängning till nio år genomfördes 1986. Detta förklarar landets struktur för den obligatoriska utbildningen i tre cykler (År 1-4, År 5-6 och år 7-9), där eleverna får sitt första betyg, som noteras i Appendix, vid slutförandet av andra cykeln (se vidare Fernandes 2009).

Som beskrivs av (Isaacs, 2010) har England en helt annan historia. England är ett av de länder som gjorde allmän utbildning obligatorisk tidigast. Redan 1947 införde de obligatorisk utbildning upp till 15 års ålder, men den var lokalt organiserad. Det är nästan 40 år tidigare än i t.ex. Portugal (1986) och drygt 20 år tidigare än Sverige. När i historien en förlängning av den obligatoriska skolan sker har betydelse för bedömningssystemets organisering. Medan Portugal gjorde stora investeringar för att utöka den obligatoriska utbildningen till 9 år i mitten av åttiotalet fanns det bara plats för små förändringar vad gäller studielängd att göra i England (till 16 år 1972 och 2015 till 18 år).

Hopmann (2003) skiljer mellan två typer av läroplanstraditioner som han kopplar till processtyrning och produktkontroll. Den förra är karakteristisk för kontinentala Europa, där läraryrket säkrade inflytande över, och genom nationell lärarutbildning, tog ansvar för undervisning i den statliga skolan. I länder där den statliga skolan av tradition varit mer lokal organiserad har staten inte haft samma ansvar för innehållet. Det har gjort att staten mer fått fokusera på produktkontroll. När Portugal gör sin stora utbyggnad av skolsystemet sker det utifrån en tradition av att lärare har stort inflytande över arbetssätt och bedömning i skolan, medan de mindre förändringarna i England kännetecknas av ytterligare mer produktkontroll (jfr Hopmann 2003). Sveriges utbildningshistoria skiljer sig på ytterligare ett vis, där produktkontroll under flera decennier var statens sätt att få legitimitet för att expandera utbildningsväsendet (exempelvis genom begreppet begåvningsreserv) vilket senare övergick, genom samma centrala instrument för bedömning, till resultatstyrning av en decentraliserad skolorganisation (t.ex. Lundahl & Tveit 2014). Tidpunkten för betyg och centralt utarbetade prov flyttades emellertid succesivt mot de åldrar där övergång till nästa skolform började bli aktuellt (Lundahl 2009), för att sedan 2011 få en lösare organisatorisk koppling i och med att betyg började ges igen i årskurs 6.

Det finns stora kulturella skillnader och skillnader mellan hur skolsystem är organiserade och vilka ämnen som ingår i läroplanen. Att ett betygssystem fungerar bra i ett land behöver inte betyda att samma betygssystem passar bra för andra länder. Grundprinciperna för betygssättning i Europa är att betygen är avgångcertifikat och urvalsinstrument. Betyg sätts traditionellt i de åldrar där övergångar och urval sker. Men ofta lyfter länderna även fram föräldrars behov av information och att betygen kan avgöra om elever behöver gå om eller få särskilt stöd. I de texter som ligger på Eurydice lyfts sällan betygens funktion som motivation för lärande eller som en del i målutvärdering fram. Två argument för betyg som ofta hörts i den svenska betygsdebatten.

Som vi visade ovan utifrån OECD:s statistik förklarar betygs- och examenssystem inte så mycket av PISA-resultaten på en aggregerad nivå. Det betyder inte att det inte spelar någon roll vilket betygssystem ett land har. Tvärtom, varje land bär på sin specifika bedömningshistoria och därifrån präglade bedömningskultur.

# SLUTDISKUSSION

Inför färdigställandet av den här forskningsöversikten om betyg har vi läst över 6000 abstracts ca 500 artiklar och ett 40 tal avhandlingar. De artiklar vi gått igenom är vetenskapligt granskade och publicerade i vetenskapliga tidskrifter. Våra sökningar och urval har varit systematiska. Ett övergripande resultat är att lärarsatta omdömen, som de svenska betygen, har en minskande betydelse internationellt sett. De fungerar som information till föräldrarna om barnets utveckling medan externa tester används vid selektion till andra utbildningar och som utvärderingsunderlag på olika nivåer inom systemet. Detta präglar det internationella forskningsfältet som oftare handlar om det vidare begreppet summativ bedömning än om betyg. Betyg må vara en stor fråga i svensk skoldebatt men det är en liten fråga i internationell forskning.

I den första delstudien har vi undersökt forskning om hur summativa bedömningar påverkar elevernas lärande, motivation för lärande och prestationer och vilka resultat den genererat. Några övergripande slutsatser vi drar är att resultaten från studierna till viss del är samstämmiga. Vuxna högpresterande studenter verkar påverkas positivt i sitt lärande och prestationer av feedback som innehåller mycket information som kommer i direkt anslutning till uppgiften och information bör vara positiv. Samtidigt framkommer det att vuxna studenter inte påverkas negativt om feedback kommer i form av betyg. Detta förklaras av att vuxna studenter på universitetsnivå ”kan” systemet och har lång erfarenhet av summativa bedömningar och har utvecklat strategier för att hantera detta system samt att de är högpresterande. Däremot verkar det vara annorlunda för yngre elever och när representativa urval undersöks. En slutsats som kan dras av resultaten från de inkluderade studierna är att betyg generellt differentierar och påverkar äldre och yngre elever och låg- och högpresterande elever på olika sätt. Lågpresterande och yngre elever verkar påverkas mer negativt av betygsättning jämfört med äldre och högpresterande elever. Ålder och erfarenheter av bedömning tycks spela en stor roll för hur elevers lärande, motivation för lärande och prestationer påverkas av betygsättning.

Den andra delstudien handlar om hur och vad lärare betygsätter och hur betyg påverkar undervisning. Vi har studerat internationell respektive svensk forskning för att beskriva skillnader dem emellan. Gemensamt är att validitetsfrågan är central men häri ligger också skillnaden. I svensk forskning är det relationen mellan lärarens betygsättning och *styrdokumenten* som dominerar perspektivet. Utanför Sverige är det framförallt frågan om *vad* läraren bedömer som dominerar, t.ex. elevens kunskaper eller personliga egenskaper.

Att lärares dagliga verksamhet påverkas av betygens inflytande är mer framträdande i den svenska forskning vi funnit. Här är det framförallt godkäntgränsen som problematiseras men även hur betyg tar tid från lärarens pedagogiska arbete. Betygens inverkan på lärarens undervisning är däremot inte centralt i forskningen utanför Sverige. Där dominerar istället kritiken mot ett ökat inflytande av *high-stakes* tester och hur lärare upplever dessa som meningslösa i sin undervisning. Standardisering av betygsättningen och *high-stakes* tester ifrågasätts utifrån att de kan riskera lärares möjlighet att verka som professionella bedömare. Över huvud taget framkommer i de studier som tar upp betygens dilemman en spänning mellan styrning och kontroll och pedagogiska aspekter av lärarens bedömning.

I den tredje och fjärde delstudien har vi gått mer explorativt tillväga, då det inte funnits internationell forskning som primärt fokuserat betyg ur styrperspektiv. I delstudie tre fann vi tre centrala teman om betyg ur styrperspektiv: 1) Betyg ur rättvise- och jämlikhetsperspektiv, 2) Betyg som kunskaps- och urvalsmått, 3) Betyg som *high-stakes* i bedömnings- och utvärderingssystem. Det tredje temat gjordes till en inramning för de andra två. Den forskning som berörde första temat poängterade bland annat att betygssystem måste sättas in i ett större perspektiv av ett rättvist bedömnings- och utvärderingssystem, med instrument för att följa upp rättviseaspekter i relation till olika elevgrupper m.m. Studierna poängterade vikten av transparens i systemen, så att grunder för bedömning och utvärdering liksom existerande orättvisor blir tydliga för systemets aktörer. Kunskapsfrågan lyftes också fram som central, det är lätt att anta att det som står i läroplanen – den kunskap som bedöms – är neutralt, men kunskapen har alltid konsekvenser och olika konsekvenser för olika grupper av elever. När det gällde tema två var ett tydligt resultat att betygens roll i många utbildningssystem reducerats de senaste decennierna. Samtidigt visar genomgången av betyg ur ett systemperspektiv att betyg är bättre som urvalsinstrument för högre utbildning jämfört med högskoleprov och andra liknande tester. I synnerhet kursbetyg på gymnasienivå som ges med stor bredd och i hög frekvens har en god predikativ förmåga. Detta

visar att betyg kan fylla viktiga funktioner i ett utbildningssystem och det på ett bättre sätt än andra instrument, och att den utveckling som man sett internationellt mot allt mer centralt administrerade examens- och antagningsprov inte bör anammas okritiskt. Ytterligare ett resultat som lyfts fram i det andra temat är att dagens målrelaterade betyg inte ger en tillräckligt bra information om elevers kunskapsnivåer och kunskapsutveckling på nationell nivå och att det är en svaghet i det nuvarande svenska utvärderings- och bedömningssystemet att man på nationell nivå inte har tillförlitlig information om kunskapsnivåer och kunskapsutveckling.

Den fjärde delstudien fokuserar betygen ur olika komparativa perspektiv. Det vi fokuserat på är vad betyg i sig jämför för något samt hur olika betygssystem jämförs med varandra på nationell och internationell nivå. När vi söker på bedömning och internationella jämförelser ser vi att det i huvudsak är tre områden som utgör fokus för jämförelser kring betyg: system för *accountability*; kulturella förklaringar till varför bedömnings- och betygssystem ser olika ut i olika länder; variationer mellan olika lärares bedömningar i olika ämnen eller av olika elevgrupper.

Några viktiga iakttagelser i vår genomgång är att det länge funnits en internationell trend mot att upprätta olika system för ökad ansvarsskyldighet (*accountability*) för skolans resultat. Dessa resultat mäts främst i elevprestationer på test eller i betyg. Såväl kritiska forskare som OECD har dock på senare tid noterat, att förhoppningarna om att jämförelser av skolors resultat ska leda till resultatförbättringar har varit överdrivna. De system olika länder har för bedömning och *accountability* förklarar i princip ingenting av variationen i PISA resultat. Det är snarare vad lärarna gör i klassrummet som har betydelse och lärare ha svårt att dra slutsatser om vad de bör göra utifrån de resultat som tillgängliggörs via *accountability*-modeller. Modellerna har sällan rätt informationsnivå för didaktiska slutsatser.

I kapitel 4 gör vi också en egen jämförelse av betygssystem i Europa i barn- och ungdomsskolan. Det första vi kan konstatera är att informationsläget är väldigt komplicerat. Det finns inte standardiserade data på hur betygssystem ser ut i olika länder varför alla jämförelser behöver bygga på komplicerat klassificeringsförfarande, där det ibland uppstår tolkningsproblem. Detta är inte bara ett problem för oss utan det finns i alla de jämförelser och hänvisningar till hur det ser ut i andra länder som också görs i den offentliga debatten om betyg. Enkla listor över när betyg ges i ålder eller i hur många skalsteg som används är ganska meningslös information utanför sitt kulturella och strukturella sammanhang.

En systematisk litteraturöversikt förutsätter att det dels finns tillräcklig volym av forskning av empirisk karaktär inom de områden man vill ha svar, dels att olika studier på en och samma fråga är jämförbara. När det gäller vår översikt av betyg så visar det sig att flertalet av de frågor och teman vi identifierat inte samlar särskilt många studier. I några fall är studierna därtill utförda i olika kontexter (länder), varför jämförbarheten minskar kraftigt. Följaktligen blir de grunder på vilka slutsatserna kan dras i flertalet fall ganska svaga och får därtill en mer allmän och övergripande karaktär.

Pedagogiska problemområden kan dessutom närmas från flera håll. I denna studie har vi utgått från betyg, men man kunde lika gärna utgått från exempelvis de funktioner som kopplas till betyg. Exempelvis, om det gäller urval till högre utbildning, hur löses det bäst? Här hade betyg i olika former varit en bland många lösningar.

För att sammanfatta våra erfarenheter av att använda den systematiska ansatsen så kan vi konstatera att det inte varit möjligt att hitta lösningar på samma sätt som inom exempelvis medicin. Däremot tvingar denna typ av undersökningar, som i första hand fokuserar empiriska studier med tydligt kvantifierbara resultat, fram en noggrann analys av empiriska belägg för eller emot vissa typer av lösningar. På så sätt kan metoden bidra till att hitta ”fasta punkter” inom utbildningsvetenskaplig forskning som kan bilda en viss bas för såväl policybeslut som initiativ till ny forskning.

Vad gäller vår kanske mest centrala frågeställning, vad betygen har för effekter på elevers lärande om motivation framträder vissa tydliga svagheter i de studier som finns om detta. Endast i tre studier var urvalet av elever nationellt representativt och det var i de tre svenska studierna. Att urvalet av deltagare är representativt är av stor vikt för att kunna dra generella slutsatser. I flera av studierna används elever och studenter som går på privata och selektiva skolor vilket innebär att resultaten inte går att generalisera till andra grupper av elever. En stor brist i dessa studier är avsaknad av diskussion om urvalet och de möjliga konsekvenser urvalet för med sig. Dock drar flera av studiernas författare långtgående slutsatser av resultaten vilket kan få konsekvenser för policyutveckling och reformarbete.

Bristande metodisk kvalitet, avsaknad av vetenskapliga bevis och bristen på detaljerade redovisningar av forskningsdesign riskerar att påverka policy och allmänhetens utvärderingar av vad som påverkar elevers lärande och prestationer. Trots att forskare inser begränsningarna med sina egna studier och varnar för att dra generaliserande slutsatser används resultaten i media och vid policyförändringar (Raymond & Hanushek 2003). Att genomföra randomiserade studier med elever ställer dock höga krav på etiskt förhållningssätt. Detta framförs av vissa författare som problematiskt och att det begränsar möjligheten att designa studier inom fältet som kan ge mer användbara resultat. Olika studier använder olika typer av utfallsvariabler vilka betyder lite olika saker. Inom det utbildningsekonomiska fältet används ofta utbildningslängd och inkomst som utfallsvariabler. Inom andra discipliner används mått på lärande (betyg eller resultat på prov) eller mått på motivation som utfall. Ett exempel kan vara att de kausala sambanden mellan att få betyg i de tidiga skolåren och lön i vuxen ålder är svåra att bevisa. Inom utbildningsvetenskap finns en mängd fenomen som kan orsaka att elever lär sig och presterar i skolan och detta kan i sin tur påverka utfall i vuxen ålder på en mängd olika sätt. Att kontrollera för alla tänkbara orsaker till låg eller hög lön i vuxen ålder är alltså problematiskt. När resultaten därför ska användas för policyändamål bör vissa aspekter tas i beaktande: inom vilken disciplin författaren skriver (till exempel pedagogik/psykologi eller ekonomi); vilka teoretiska utgångspunkter som används; vilka variabler som används; samt den metodiska kvaliteten.

Men även om forskningsläget är komplicerat och det alltid kommer vara svårt att arrangera experiment med betyg eller importera system som visat sig fungera i andra länder, kan vi dra vissa lärdomar från en forskningsgenomgång som den här inför kommande betygsreformer i Sverige.

Vi kan i ljuset av de reservationer som görs i den litteratur vi gått igenom konstatera att vi har ett betygssystem i Sverige som inte bygger på någon tydlig vetenskaplig grund för hur det bäst ska tjäna sina syften. Och förmodligen har betygssystemet därtill givits för många syften vilka ibland, sett till forskningslitteraturen, kommer i konflikt med varandra. Vi har t.ex. sett att bredd och hög frekvens kan vara bra ur ett predikativt perspektiv och att kursbetyg är att föredra framför ämnesbetyg i urvalssammanhang. De svenska betygen har dock fått kritik för att vara för oprecisa för att användas som underlag för utvärdering och det finns tendenser till betygsinflation mellan skolor och över tid (kapitel 3). Samtidigt pekar våra resultat mot att för hög standardisering av betygssättning hotar lärares professionalitet. För hög frekvens av externa summativa bedömningar påverkar också lärares arbete negativt (kapitel 2). Yngre elever och lågpresterande elever verkar inte heller gagnas av för mycket summativa bedömningar (kapitel 1).

Den här typen av motsättningar har man i många länder försökt att lösa genom att ha flera parallella system både för elevutvärdering och utvärdering av skolors kvalitet (kapitel 4). Sverige har kanske lagt allt för stora förhoppningar vid att nationella prov ska ge betygen en sådan kvalitet att de kan användas både för en rättvis bedömning av enskilda elever och för målutvärdering och ansvarsutkrävande.

Ur såväl ett rättssäkerhetsperspektiv som kvalitetsperspektiv förefaller det viktigt att utveckla fler modeller för bedömning av elevers kunskaper och för utvärdering av skolan. Det är i så fall viktigt att förstå hur det kan ske så att interna och externa behov kan tillgodoses genom samma instrument, eller när man behöver hålla dem åtskilda (Benveniste 2002). Vissa skolresultat ska kanske inte alltid offentliggöras. Andra skolresultat behöver anpassas så att lärare har nytta av dem. En ytterligare viktig aspekt är att förstå att interna bedömningar som betyg och externa bedömningar som nationella prov, behöver kalibrera och validera varandra (kapitel 3 och 4). Det blir dock lätt så att lärarna uppfattas vara de som gör fel bedömning (kapitel 2).

Baserat på vad vi har fått fram i den här översikten vill vi avluta med några rekommendationer. Det finns tydliga resultat som åtminstone bör mana till försiktighet om att vidare sänka åldern för betyg. Frågan är också på vilket sätt utblickar mot andra länders betygsstart kan hjälpa oss att ta kloka beslut om när vi ska börja med betyg. Snarare verkar det som att många länder i Europa inte reformerat sina betygssystem. Tidiga betyg har inte införts för att öka kraven på eleverna eller främja deras motivation. Tidig betygssättning har helt enkelt inte tagits bort. Men i många länder representerar inte ett tidigt betyg det vi menar med betyg i Sverige. Om man däremot går på djupet med de kvalitativa skillnaderna som finns i de olika ländernas betygssystem är det möjligt att extrahera vissa principer som kanske är överförbara till Sverige. Flera länder ger exempelvis lärare och skolor stor autonomi över hur bedömningarna i tidiga åldrar ska tillämpas, vilket kan tänkas ha positiva konsekvenser för lärares professionalitet i frågan.

Det är också viktigt att det svenska nuvarande betygssystemet bättre utvärderas på ett nyanserat sätt i förhållande till olika lärare, ämnen och elevgrupper. Betyg fungerar inte lika för alla. Det är också viktigt att fundera över hur vi utvärderar elevers resultat och om det finns möjlighet att kombinera fler modeller med varandra, så att vi bättre kan få data av ”value added”-karaktär samt för att följa kunskapsutvecklingen över tid. Studien visar också på flera olika plan vilka svårigheter det finns med översättning av forskningsresultat och information om utbildningssystem mellan olika länder och kontexter.

Så länge den svenska debatten om betyg är så livlig som den varit det senaste decenniet är det viktigt att denna fråga följs i forskningen, men då bör den knytas till den större frågan om alla slags summativa bedömningars effekter på skolan. Det är även centralt att vi får ett skolaktörsperspektiv på denna fråga, då många av de studier vi granskat visat att huvudproblemet kring summativ bedömning är att det används fel när skolans aktörer inte själva definierar vad de ska ha bedömning, betyg och utvärdering till i sin egen vardag. Det är också viktigt att förstå betygens relation till läroplan och kursplaner och att fortbildning och implementeringsinsatser kring bedömning även kan vara en väldigt bra väg för implementering av läroplanernas innehåll.

Vi vill dock betona att det är skillnad på direkta och indirekta effekter. Duktiga pedagoger kan säkert hitta sett att ge betyg utan att det behöver påverka elevernas självbild, men frågan hur det ska gå till och när elever har tillräcklig mognad för att hantera sådan information finns det inte mycket skrivet om i forskningen om betyg. Det finns i dagsläget inga belägg i forskningen för att vetskapen om låga betyg leder till ett kompensatoriskt stöd som minskar den negativa effekt ett lågt betyg kan ha på självkänsla och motivation.

Vår studie pekar på att lärarnas autonomi över bedömningssystemen, oavsett hur de ser ut, är det som kanske har störst betydelse. Att lärarna har verktyg som de kan använda i bedömning av elevernas kunskaper och i kommunikationen kring dessa kunskaper som lärarna själva upplever är meningsfulla och som gagnar den pedagogiska processen. Därför är det också av stor vikt att lärare ges möjlighet till fortbildning kring betyg och bedömning och att det kanske blir ett ännu mer markerat inslag i lärarutbildningen.

# REFERENSLISTA


Abu-Hamour, B., & Mattar, J. (2013). The applicability of curriculum-based-measurement in math computation in Jordan. *International Journal of Special Education, 28*(1), 111-119.

Allal, L. (2013). Teachers' professional judgement in assessment: a cognitive act and a socially situated practice. *Assessment in Education: Principles, Policy & Practice*, *20*(1), 20-34.

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education policy analysis archives*, *10*(18), 1-74.

Andersson, H. (1991). *Relativa betyg: några empiriska studier och en teoretisk genomgång i ett historiskt perspektiv* (Doktorsavhandling, Umeå universitet).

Annerstedt, C., & Larsson, S. (2010). "I Have My Own Picture of What the Demands Are...": Grading in Swedish PEH - Problems of Validity, Comparability and Fairness. *European Physical Education Review 16*(2), 97-115.

Artes, J., & Rahona, M. (2013). Experimental evidence on the effect of grading incentives on student learning in Spain. *Journal of Economic Education, 44*(1), 32-46.

Au, W. (2007). High-Stakes Testing and Curricular Control: A Qualitative Metasynthesis. *Educational Researcher*, *36*(5), 258-267.

Au, W. (2011). Teaching under the new Taylorism: high- stakes testing and the standardization of the 21st century curriculum. *Journal of Curriculum Studies*, *43*(1), 25-45.

Azmat, G., & Iriberri, N. (2009). *The importance of relative performance feedback information: Evidence from a natural experiment using high school students*. CEP Discussion. Paper No. 915, Centre for Economic Performance, London School of Economics and Political Science.

Azmat, G., & Iriberri, N. (2010). The importance of relative performance feedback information: Evidence from a natural experiment using high school students. *Journal of Public Economics*, *94*(7-8), 435-452.

Bagley, S. S. (2008). High school students' perceptions of narrative evaluations as summative assessment. *American Secondary Education, 36*(3), 15-32.

Baird, J.-A., Greatorex, J., & Bell, J. F. 2010. What makes marking reliable? Experiments with UK examinations. *Assessment in Education: Principles, Policy & Practice*, *11*(3), 331-348.

Baker, E. L., & O'Neil Jr, H. F. (1994). Performance Assessment and Equity: a view from the USA. *Assessment in Education: Principles, Policy & Practice*, *1*(1), 11-26.

Bandiera, O., Barankay, I., & Rasul, I. (2009). Team incentives: evidence from a field experiment. Mimeo.

Bandiera, O., Larcinese, V., & Rasul, I. (2008). Blissful ignorance? Evidence from a natural experiment on the effect on individual feedback on performance. Mimeo.

Bautier, É., Crinon, J., Rayou, P., & Rochex, J. Y. (2006). Performances en littéracie, modes de faire et univers mobilisés par les élèves: analyses secondaires de l'enquête PISA 2000. *Revue française de pédagogie*, *157*, 85-101.

Becker, W. E., & Rosen, S. (1992). The learning effect on assessment and evaluation in High School. *Economics of Education Review*, *11*(2), 107-118.

Benveniste, L. (2002). The Political Structuration of Assessment: Negotiating State Power and Legitimacy. *Comparative Education Review*, *46*(1), 89-118.

Bergman, L. (2007). *Gymnasieskolans svenskämnen: En studie av svenskundervisningen i fyra gymnasieklasser* (Doktorsavhandling, Malmö högskola).

Betts, J. R., & Grogger, J. (2003). The impact of grading standards on student achievement, educational attainment, and entry-level earnings (hög relevans). *Economics of Education Review*, *22*(4), 343-52.

Biberman-Shalev, L., Sabbagh, C., Resh, N., & Kramarski, B. (2011). Grading styles and disciplinary expertise: The mediating role of the teacher's perception of the subject matter. *Teaching and Teacher Education*, *27*(5), 831-840.

Bies-Hernandez, N. (2012). The effects of framing grades on student learning and preferences. *Teaching of Psychology, 39*(3), 176-180.

Billing, D. (2004). International Comparisons and Trends in External Quality Assurance of Higher Education: Commonality or Diversity? *Higher Education*, *47*(1), 113-137.

Black , P., Harrison, C., Hodgen J., Marshall B., & Serret N. (2010). Validity in teachers' summative assessments. *Assessment in Education: Principles, Policy & Practice*, *17*(2), 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*(4), 451-469.

Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education*, *5*(1), 7-74.

Black, P., & Wiliam, D. (2005). Lessons from around the world: how policies, politics and cultures constrain and afford assessment practices, *The Curriculum Journal, 16*(2), 249-261

Black, P., & Wiliam, D. (2007). Large-scale Assessment Systems Design principles drawn from international comparisons. *Measurement*, *5*(1), 1-53.

Black, P., & Wiliam, D. (2012). The reliability of assessments. In J. Gardner (Ed.), *Assessment and Learning* (pp. 243–263). Los Angeles and London: Sage Publications.

Blok, H. Otter, M. E., & Roeleveld, J. (2002). Coping with Conflicting Demands: Student Assessment in Dutch Primary Schools. *Studies in Educational Evaluation, 28*(2), 177-188.

Boehnke, K. (2005). Value orientations in relation to mathematical self-esteem: An exploratory study of their role in mathematical achievement among German, Israeli, and Canadian 14-year-olds. *European journal of psychology of education*, *20*(3), 227-241.

Bonesrønning, H. (2004). Do the teachers' grading practices affect student achievement? *Education Economics, 12*(2), 151-167.

Bonnet, G. (1997). Country Profile from France. *Assessment in Education: Principles, Policy & Practice*, *4*(2), 295–306.

Bowers, A.J. (2010). Analyzing the Longitudinal K-12 Grading Histories of Entire Cohorts of Students: Grades, Data Driven Decision Making, Dropping Out and Hierarchical Cluster Analysis. *Practical Assessment, Research & Evaluation (PARE), 15*(7), 1-18.

Bowers, A.J., Sprott, R., & Taff, S.A. (2013). Do we Know Who Will Dropout? A Review of the Predictors of Dropping out of High School: Precision, Sensitivity and Specificity. *The High School Journal, 96*(2), 77-100.

Braun, H. (2004). Reconsidering the Impact of High-stakes Testing. *Education Policy Analysis Archives, 12*(1), 1-43.

Brennan, R. T., Kim, J., Wenz-Gross, M., & Siperstein, G. N. (2001). The relative equitability of high-stakes testing versus teacher-assigned grades: An analysis of the Massachusetts Comprehensive Assessment System (MCAS). *Harvard educational review*, *71*(2), 173-217.

Brookhart, S. M. (2011). Starting the Conversation about Grading. *Educational Leadership, 69*(3), 10-14.

Brookhart, S. M. (2013a). Educational Assessment Knowledge and Skills for Teachers. *Educational Measurement: Issues and Practice, 30*(1), 3-12.

Brookhart, S. M. (2013b). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice, 20*(1), 69-90.

Butler, R. (1988). Enhancing and undermining intrinsic motivation: The effects of task-involving and ego-involving evaluation on interest and performance. *British Journal of Educational Psychology*, *58*(1), 1-14.

Böhlmark, A., & Holmlund, H. (2011). *20 år med förändringar i skolan: Vad har hänt med likvärdigheten?* Stockholm: SNS förlag.

Cameron, J. (2001). Negative effects of reward on intrinsic motivation—a limited phenomenon: Comment on Deci, Koestner, and Ryan (2001). *Review of Educational Research, 71*(1), 29-42.

Cameron, J., Banko, M., & Pierce, D. (2001). Pervasive negative effects of rewards on intrinsic motivation: the myth continues. *The Behavior Analyst, 24*(1), 1-44.

Carrillo-de-la-Peña, M. T., Baillès, E., Caseras, X., Martínez, A., Ortet, G., Pérez, J. (2009). Formative assessment and academic achievement in pre-graduate students of health sciences. *Adv Health Sci Educ Theory Pract., 14*(1), 61-67.

Chilisa, B. (2000). Towards Equity in Assessment: Crafting gender-fair assessment. *Assessment in Education: Principles, Policy & Practice, 7*(1), 61-81.

Christophersen, K-A., Elstad, E., & Turmo, A. (2012). Antecedents of Teachers Fostering Effort within Two Different Management Regimes: An Assessment-Based Accountability Regime and Regime without External Pressure on Results. *International Journal of Education Policy & Leadership*, *7*(6).

Cilliers, F. J., Schuwirth, L. W., Adendorff, H. J., Herman, N., & van der Vleuten, C. P. (2010). The mechanism of impact of summative assessment on medical students' learning. *Advances in Health Sciences Education, 15*(5), 695-715.

Cliffordson, C. (2004). Betygsinflation i de målrelaterade gymnasiebetygen. *Pedagogisk forskning i Sverige*, *9*(1), 1-14.

Cliffordson, C. (2008). Differential prediction of study success across academic programs in the Swedish context: The validity of grades and tests as selection instruments for higher education. *Educational Assessment*, *13*(1), 56-75.

Clymer, J. B., & Wiliam, D. (2007). Improving the way we grade science. *Educational Leadership, 64*(4), 36-42.

Congon, P. J. & McQueen, J. (2002). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement, 37*(2), 163-178.

Coniam, D. (2009). A comparison of onscreen and paper-based marking in the Hong Kong public examination system. *Educational Research and Evaluation*, *15*(3), 243-263.

Cooksey, R. W. Freebody P., & Wyatt-Smith, C. (2007). Assessment as Judgment-in-Context: Analysing how teachers evaluate students' writing, *Educational Research and Evaluation: An International Journal on Theory and Practice, 13*(5), 401-434.

Covington, M. V. (2000). Goal theory, motivation, and school achievement: An integrative review. *Annual Review of Psychology*, *51*, 171-200.

Cox, K. B. (2011). Putting Classroom Grading on the Table: A Reform in Progress *American Secondary Education, 40*(1), 67-87.

Crooks, T. (1988). The impact of classroom evaluation practices on students. *Review of Educational Research*, *58*(4), 438-481.

Cumming, J. J. (2008). Legal and educational perspectives of equity in assessment. *Assessment in Education: Principles, Policy & Practice*, *15*(2), 123-135.

Dahl, B, Lien, E., & Lindberg-Sand, Å. (2009). Conformity or confusion? Changing higher education grading scales as a part of the Bologna Process: the cases of Denmark, Norway and Sweden. *Learning and Teaching: The International Journal of Higher Education in the Social Sciences*, *2*(1), 39-79.

Danielewicz, J., & Elbow, P. (2009). A unilateral grading contract to improve learning and teaching. *College Composition and Communication, 61*(2), 244-268.

Daugherty, R. (2008). Reviewing national curriculum assessment in Wales: how can evidence inform the development of policy? *Cambridge Journal of Education*, *38*(1), 73-87.

De Luca, C. (1994). *The impact of examination systems on curriculum development: an international study*. Paris: UNESCO.

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological bulletin*, *125*(6), 627-668.

Deci, E. L., Koestner, R., & Ryan, R. M. (2001). Extrinsic rewards and intrinsic motivation in education: Reconsidered once again. *Review of Educational Research*, *71*(1), 1-27.

Deci, E. L., & Ryan, R. M. (1985). *Intrinsic motivation and self-determination in human behaviour*. New York: Plenum.

Dee, T. S. & Jacob, B. (2011). The impact of no Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*(3), 418-446.

Deutsch, M. (1979). Education and Distributive Justice: Some Reflections on Grading Systems. *American Psychologist,* 34(5), 391-401.

Dlaska, A., & Krekeler, C. (2013). Does grading undermine feedback? The influence of grades on the effectiveness of corrective feedback on L2 writing. *The Language Learning Journal*. DOI: 10.1080/09571736.2013.848226.

Dobbins, M., & Martens, K. (2012). Towards an Education Approach a la "Finlandaise"? French Education Policy after PISA. *Journal of Education Policy*, *27*(1), 23-43.

Docan, T. N. (2006). Positive and negative incentives in the classroom: An analysis of grading systems and student motivation. *Journal of Scholarship of Teaching and Learning*, *6*(2), 21-40.

Dolton, P., & Marcenaro-Gutierrez, O. D. (2011). If you pay peanuts do you get monkeys? A cross-country analysis of teacher pay and pupil performance. *Economic Policy*, *26*(65), 5-55.

Dowdy, E., & Kamphaus, R. W. (2007). A Comparison of Classification Methods for Use in Predicting School-Based Outcomes. *The California School Psychologist*, *12*(1), 121-132.

Dragemark Oscarson, A. (2008). *Self-Assessment of Writing in Learning English as a Foreign Language. A Study at the Upper Secondary School Level* (Doktorsavhandling, Göteborgs universitet).

Ds 2008:13. *En ny betygsskala*. Stockholm: Utbildningsdepartementet.

Duncan R. C., & Noonan, B. (2007). Factors Affecting Teachers' Grading and Assessment Practices. *Alberta Journal of Educational Research*, *53*(1), 1-21.

Duckworth, A. L., & Seligman, E. P. (2005). Self-Discipline outdoes IQ in predicting academic performance of adolescents. *Psychological Science*, *16*(12), 939-944.

Durlak, J. A. (2009). How to select, calculate, and interpret effect sizes. *Journal of Pediatric Psychology*, *34*(5), 917-928.

Dweck, C. S. (1992). The study of goals in psychology. *Psychological Science*, *3*(3), 165-167.

Ecclestone, K. (2004). Learning in a comfort zone: cultural and social capital inside an outcome based assessment regime. *Assessment in Education: Principles, Policy & Practice*, *11*(1), 29-47.

Elliott, A. J., Shell, M. M., Henry, K. B., & Maier, M. A. (2005). Achievement goals, performance contingencies, and performance attainment: An experimental test. *Journal of Educational Psychology*, *97*(4), 630-640.

Englund, T., Forsberg, E., & Sundberg, D. (red.) (2012). *Vad räknas som kunskap?: läroplansteoretiska utsikter och inblickar i lärarutbildning och skola*. Stockholm: Liber.

Eurydice (2014a). About Eurydice. Downloaded on December 20th 2014 from: http://eacea.ec.europa.eu/education/eurydice/about_eurydice_en.php

Eurydice (2014b). Compulsory Education in Europe 2014/2015, Facts and Figures, November 2014. Downloaded on December 20th 2014 from: http://eacea.ec.europa.eu/education/eurydice/documents/facts_and_figures/compulsory_education_EN.pdf

Eurydice (2014c). The Structure of the European Education Systems 2014/2015, Schematic Diagrams, November 2014. Downloaded on December 20th 2014 from: http://eacea.ec.europa.eu/education/eurydice/documents/facts_and_figures/EN_2014_15_diagrams_version_finale_pngs.pdf

Feniger, Y., Livneh, I., & Yogev, A. (2012). Globalisation and the politics of international tests: the case of Israel. *Comparative Education*, *48*(3), 323-335.

Fernandes, D. (2009). Educational assessment in Portugal. *Assessment in Education: Principles, Policy & Practice*, *16*(2), 227-247.

Forsberg, E. (2008). Framtidsvägen – en huvudled eller en skiljeväg? *Utbildning & Demokrati*, *17*(1), 75-98.

Forsberg, E., & Lundahl, C. (2006). Kunskapsbedömningar som styrmedia. *Utbildning & Demokrati*, *15*(3), 7-29.

Figlio, D. N., & Lucas, M. E. (2004). Do high grading standards affect student performance? *Journal of Public Economics*, *88*(9-10), 1815-1834.

Friedler, S. A.,Tan, Y. L., Peer, N. J., & Shneiderman, B. (2008). Enabling Teachers to Explore Grade Patterns to Identify Individual Needs and Promote Fairer Student. *Assessment Computers & Education, 51*(4 ), 1467-1485.

Frydenberg, E. (2008). *Adolescent coping: Advances in theory, research and practice*. London, New York, NY: Routledge & Francis Group.

Gewirtz, S. (1998). Conceptualizing social justice in education: mapping the territory. *Journal of Education Policy*, *13*(4), 469-484.

Gijbels, D., Dochy, F., Van den Bossche, P., & Segers, M. (2005). Effects of problem-based learning: A meta-analysis from the angle of assessment. *Review of educational research*, *75*(1), 27-61.

Gipps, C. (1995). What Do We Mean by Equity in Relation to Assessment? *Assessment in Education: Principles, Policy & Practice*, *2*(3), 271-281.

Gipps, C. (1999). Sociocultural aspects of assessment. *Review of Research in Education*, *24*(1), 355-392.

Gustafsson, J. E. (2006). *Barns utbildningssituation. Bidrag till ett kommunalt barnindex*. Stockholm: Rädda Barnen.

Gustafsson, J. E. (2013). Förändringar i kunskapsbedömningar på individ- och systemnivå i den svenska skolan under 25 år. I I. Wernersson & I. Gerrbo (red.), *Differentieringens janusansikte: En antologi från Institutionen för pedagogik och specialpedagogik vid Göteborgs universitet* (s. 45-75). Göteborg: Göteborgs universitet.

Gustafsson, J. E., Cliffordson, C., & Erickson, G. (2014). *Likvärdig kunskapsbedömning i och av den svenska skolan: Problem och möjligheter*. Stockholm: SNS förlag.

Gustafsson, J. E., & Erickson, G. (2013). To Trust or Not to Trust?–Teacher Marking versus External Marking of National Tests. *Educational Assessment, Evaluation and Accountability*, *25*(1), 69-87.

Hacking, I. (1995). The Looping Effects of Human Kinds. I: D. Sperber, D. Premack, & A. J. Premack (red.). *Causal Cognition: A Multi-Disciplinary Approach*. Oxford: Clarendon Press.

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

Hanna, R. N., & Linden, L. L. (2012). Discrimination in Grading. *American Economic Journal: Economic Policy, 4*(4), 146-168.

Hanushek, E. A. (1986). The economics of schooling. *Journal of Economic Literature, 24*(3), 1141–1177.

Harlen, W. (2004). 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, University of London.

Harlen, W., & Deakin Crick, R. (2002). A systematic review of the impact of summative assessment and tests on students´ motivation for learning (EPPI-Centre Review, version 1.1*). In: *Research evidence in educational library*. Issue 1. London: EPPI-Centre, Social Science Research Unit, Institute of Education.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. London: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112.

Hendrickson, K. A. (2012). Assessment in Finland: A Scholarly Reflection on One Country's Use of Formative, Summative, and Evaluative Practices. *Mid-Western Educational Researcher, 25*(1-2), 33-43.

Hobfoll, S. E. (1989). Conservation of resources: A new attempt at conceptualizing stress. *American Psychologist, 44*(3), 513-524.

Holmlund, H., Häggblom, J., Lindahl, E., Martinson, S., Sjögren, A., Vikman, U. & Öckert, B. (2014). *Decentralisering, skolval och fristående skolor: resultat och likvärdighet i svensk skola*. Rapport 2014:25. Uppsala: IFAU – Institutet för arbetsmarknad och utbildningspolitisk utvärdering

Hopmann, S. T. (2003). On the evaluation of curriculum reforms. *Journal of Curriculum Studies, 35*(4), 459-478.

Hopmann, T. S. (2013). The end of schooling as we know it? *Journal of Curriculum Studies*, *45*(1), 1-3.

Hout, M., & Elliot S. W. (2011). *Incentives and test-based accountability in education*. Washington: The National Academies Press.

Hultqvist, E. (2011). Om lärarnas förändrade yrkesvillkor. *Pedagogisk Forskning i Sverige, 16*(3), 202-213.

Hyltegren, G. (2014). *Vaghet och vanmakt: 20 år med kunskapskrav i den svenska skolan* (Doktorsavhandling, Göteborgs universitet).

Isaacs, T. (2010). Educational assessment in England. *Assessment in Education: Principles, Policy & Practice*, *17*(3), 315-334.

Jacob, B. A. (2005). Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics*, *89*(5-6), 761-796.

Jansson, T. (2011). *Vad kommer på provet? Gymnasielärares provpraxis i samhällskunskap* (Licentiatavhandling, Karlstad universitet).

Karlsson, A. (2011). *Samhällsguide, individualist och moderator: samhällskunskapslärares professionella förhållningssätt i betygsättningsrelaterat arbete* (Licentiatavhandling, Karlstad universitet).

Klapp, A. (2014). Does grading affect educational attainment? A longitudinal study. *Assessment in Education: Principles, Policy and Practice*, DOI: 10.1080/0969594X.2014.988121.

Klapp, A., Cliffordson, C., & Gustafsson, J-E. (2014). The effect of being graded on later achievement: evidence from 13-year olds in Swedish compulsory school. *Eductional Psychology: An international Journal of Experimental Educational Psychology*. DOI: 10.1080/01443410.2014.933176.

Klapp Lekholm, A. (2010). *Vad mäter betygen*. I C. Lundahl & M. Folke-Fichtelius (red.), *Bedömning i och av skolan – praktik, principer, politik*. Lund: Studentlitteratur.

Klapp Lekholm, A., & Cliffordson, C. (2008). Discrepancies between school grades and test scores at individual and school level: effects of gender and family background. *Educational Research and Evaluation*, *14*(2), 181-199.

Klein, J. (2002). The failure of a decision support system: inconsistency in test grading by teachers. *Teaching and Teacher Education, 18*(8), 1023-1033.

Klenowski, V., & Wyatt-Smith, C. (2012). The impact of high stakes testing: the Australian story. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 65-79.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, *119*(2), 254-284.

Korp, H. (2003). *Kunskapsbedömning: hur, vad och varför*. Stockholm: Myndigheten för skolutveckling.

Korp, H. (2006). *Lika chanser i gymnasiet? En studie om betyg, nationella prov och social reproduktion*. (Doktorsavhandling, Malmö högskola).

Kroksmark, T. (2002). En tankes fall i praktiken – då den målrationella styrningen möter skolan. I: *Att bedöma eller döma: tio artiklar om bedömning och betygssättning*. Stockholm: Skolverket.

Kurth, J., Gross, M., Lovinger, S., & Catalano, T. (2012). Grading Students with Significant Disabilities in Inclusive Settings: Teacher Perspectives. *Journal of the International Association of Special Education 13*(1), 41-57.

Lindberg, V. (2002). Införandet av godkändgränsen - konsekvenser för lärare och elever. I: *Att bedöma eller döma: tio artiklar om bedömning och betygssättning*. Stockholm: Skolverket.

Lindberg, V. (2005). Svensk forskning om bedömning och betyg 1990–2005. *Studies in Educational Policy and Educational Philosophy*. E-tidskrift 2005:1.

Lindberg, V., & Forsberg, E. (2010). *Svensk forskning om bedömning: en kartläggning*. Stockholm: Vetenskapsrådet.

Lindberg, V. & Löfgren R. (2011). Vad krävs för godkänt i kemi? I: Eriksson, I. (red.). *Kemiundervisning, text och textbruk i finlandssvenska och svenska skolor: en komparativ tvärvetenskaplig studie*. Stockholm: Stockholms universitets förlag.

Lindblad, S., Pettersson, D. & Popkewitz, T. S. (2015) *International Comparisons of School Results: A Systematic Review of Research on Large Scale Assessments in Education*. Delrapport från skolforsk-projektet. Vetenskapsrådet. Stockholm: Vetenskapsrådet.

Lindström, L., Lindberg V., & Pettersson, A. (red.) (2011). *Pedagogisk bedömning: att dokumentera, bedöma och utveckla kunskap*. Stockholm: Stockholms universitets förlag.

Lundahl, C. (2006): *Viljan att veta vad andra vet. Kunskapsbedömning i tidigmodern, modern och senmodern skola* (Doktorsavhandling, Uppsala universitet).

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

Lundahl, C. (2010). Nationella prov – ett redskap med tvetydiga syften. I C. Lundahl & M. Folke-Fichtelius (red.), *Bedömning i och av skolan – praktik, principer, politik*. Lund: Studentlitteratur.

Lundahl, C. (2014). Kunskap in/om pedagogik – Produktion, visualisering och effekter av skolresultat. *Utbildning & Demokrati*, *23*(3), 7-31.

Lundahl, C. & Folke-Fichtelius, M. (red.) (2010). *Bedömning i och av skolan – praktik, principer, politik*. Lund: Studentlitteratur.

Lundahl, C., Román, H., & Riis, U. (2010). Tidigt ute med sena betyg – sent ute med tidiga! Svensk betygspolitik i ljuset av internationell betygsforskning och betygsättningen i Europa. *Pedagogisk forskning i Uppsala nr 157*. Uppsala: Uppsala universitet, Pedagogiska institutionen.

Lundahl, C., & Tveit, S. (2014). Att legitimera nationella prov i Sverige och Norge – en fråga om profession och tradition. *Pedagogisk forskning i Sverige*, *19*(4-5), 297-323.

MacLure, M. (2005). 'Clarity bordering on stupidity': where's the quality in systematic review? *Journal of Education Policy, 20*(4), 393-416.

Marsh, H. W., Trautwein, U., Lüdtke, O., Köller, O., & Baumert, J. (2005). Academic self-concept, interest, grades, and standardized test scores: Reciprocal effects models of causal ordering. *Child development*, *76*(2), 397-416.

Martinez, J.F., Stacher, B., & Borko, H. (2009). Classroom Assessment Practices, Teacher Judgments, and Student Achievement in Mathematics: Evidence from the ECLS. *Educational Assessment, 14*(2), 78-102.

Mastergeorge, A. M. & Martinez, J. F. (2010). Rating Performance Assessments of Students with Disabilities: A Study of Reliability and Bias. *Journal of Psychoeducational Assessment*, *28*(6), 536-550.

McDonnell, L. M. (1994). Assessment Policy as Persuasion and Regulation. *American Journal of Education*, *102*(4), 394-420.

McMillan, J. H. (2001). Secondary Teachers' Classroom Assessment and Grading Practices. *Educational Measurement: Issues and Practice*, *20*(1), 20-32.

McMillan, J. H, Myran, S., & Workman, D. (2002). Elementary teachers' classroom assessment and grading practices. *The Journal of Educational Research, 95*(4 ), 203-213.

Meadmore, D. (1995). Linking Goals of Governmentality with Policies of Assessment. *Assessment in Education: Principles, Policy & Practice*, *2*(1), 9-22.

Mechtenberg, L. (2009). Cheap talk in the classroom: How biased grading at school explains gender differences in achievements, career choices and wages. *The Review of Economic Studies*, *76*(4), 1431-1459.

Mehrens, W. A. (1998). Consequences of Assessment: What is the Evidence. *Education Policy Analysis Archives, 13*(6), 1-30.

Meyer, B. J. F., Wijekumar, K., Middlemiss, W., Higley, K., Lei, P., Meier, C., & Spielvogel, J. (2010). Web-based tutoring of the structure strategy with or without elaborated feedback or choice for fifth- and seventh-grade readers. *Reading Research Quarterly*, *45*(1), 62-92.

Mickwitz, L. (2011). *Rätt betyg för vem? Betygsättning som institutionaliserad praktik* (Licentiatavhandling, Stockholms universitet).

Mickwitz, L. (2015). *En reformerad lärare. Konstruktionen av en professionell och betygssättande lärare i skolpolitik och skolpraktik*. (Doktorisavhandling, Stockholms universitet).

Molden, D. C., & Dweck, C. S. (2006). Finding "meaning" in psychology: A lay theories approach to self-regulation, social perception, and social development. *American Psychologist, 61*(3), 192-203.

Morris, P. (2012). Pick'n'mix, select and project; policy borrowing and the quest for 'world class' schooling: an analysis of the 2010 schools White Paper. *Journal of Education Policy, 27*(1), 89-107.

Muñoz, A. P., & Álvarez, M. E. (2010). Washback of an oral assessment system in the EFL classroom. *Language testing*, *27*(1), 33-49.

Neumann, M., Trautwein, U., & Nagy, G. (2011). Do central examinations lead to Greater Grading comparability? A study of frame-of-reference effects on the university entrance Qualification in Germany. *Studies in Educational Evaluation, 37*(4), 206-217.

Nichols, S. L., Glass, G. V., & Berliner, D. C. (2006). High-stakes testing and student achievement: does accountability pressure increase student learning? *Education Policy Analysis Archives, 14*(1), 1-175.

Nieuwenhuis, J. (2010). Social justice in education revisited. *Education Inquiry, 1*(4), 269-287.

Nowell, C., & Alston, R. M. (2007). I thought I got an A! Overconfidence across the economics curriculum. *The Journal of Economic Education*, *38*(2), 131-142.

Nyström, P. (2014). *Rätt mätt på prov: om validering av bedömningar i skolan* (Doktorsavhandling, Umeå universitet, Pedagogiska institutionen).

Odenstad, C. (2010). *Prov och bedömning i samhällskunskap: en analys av gymnasielärares skriftliga prov* (Licentiatavhandling, Karlstad universitet).

OECD (2005). *Formative assessment: improving learning in secondary classrooms*. Facsimile ed. (2005). Paris: OECD.

OECD (2010). PISA 2009 Results: What Makes a School Successful? Resources, Policies and Practices (Volume IV). http://www.oecd.org/pisa/pisaproducts/48852721.pdf Nedladdad 2014-12-28.

OECD (2012). Grade expectations: how marks and education policies shape students' ambitions. Paris: Organisation for Economic Co-operation and Development (OECD).

OECD (2013). PISA 2012 Results: What Makes a School Successful? Resources, Policies and Practices. http://www.oecd.org/pisa/keyfindings/pisa-2012-results-volume-IV.pdf Nedladdad 2014-12-28.

Oscarsson, M., & Apelgren B.-M. (2011). Mapping Language Teachers' Conceptions of Student Assessment Procedures in Relation to Grading: A Two-Stage Empirical Inquiry. *System: An International Journal of Educational Technology and Applied Linguistics, 39*(1), 2-16.

Prendergast, C. (1999). The provision of incentives in firms. Journal of Economic Literature, *37*(1), 7-63.

Pettersson, D. (2008). *Internationell kunskapsbedömning som inslag i nationell styrning av skolan* (Doktorsavhandling, Uppsala universitet).

Pettersson, D., & Wester, A. (2010). Skolan i världen – internationella kunskapsmätningar. I C. Liberg, U. P. Lundgren & R. Säljö (red.), *Lärande skolan bildning*. Natur & Kultur: Stockholm.

Plank, S. B., & Falk Condliffe, B. (2013). Pressures of the Season: An Examination of Classroom Quality and High-Stakes Accountability. *American Educational Research Journal*, *50*(5), 1152-1182.

Pope, N., Green, S. K, Johnson, R. L, & Mitchell, M. (2009). Examining teacher ethical dilemmas in classroom assessment *Teaching and Teacher Education, 25*(5), 778-782.

Porter, A.C. (1993). School delivery standards. *Educational Researcher, 22*(5), 24-30.

Promemoria 2014-08-20. U2014/4873/S *En bättre skolstart för alla: bedömning och betyg för progression i lärandet*. Utbildningsdepartementet.

Pulfrey, C., Buchs, C., & Butera, F. (2011). Why grades engender performance-avoidance goals: The mediating role of autonomous motivation. *Journal of Educational Psychology, 103*(3), 683-700.

Randall, J. & Engelhard, G. (2009). Examining teacher grades using Rasch measurement theory. *Journal of Educational Measurement, 46*(1), 1-18.

Randall, J., & Engelhard, G (2010). Examining the grading practices of teachers *Teaching and Teacher Education* 26(7), 1372-1380

Ravitch, D. (2010). *The death and life of the great American school system: how testing and choice are undermining education*. New York, NY: Basic Books.

Raymond, M. E., & Hanushek, E. A. (2003). High-stakes research. *Education Next*, *3*(3), 48-55.

Redelius, K., Fagrell B., & Larsson H. (2009). Symbolic capital in physical education and health: to be, to do or to know? That is the gendered question. *Sport, Education and Society*, *14*(2), 245-260.

Regeringens proposition 2008/09:66. *En ny betygsskala*. Stockholm: Utbildningsdepartementet.

Regeringens proposition 2008/09:87. *Tydligare mål och kunskapskrav – nya läroplaner för skolan*. Stockholm: Utbildningsdepartementet.

Resh, N. (2009). Justice in grades allocation: teachers' perspective. *Social Psychology of Education*, *12*(3), 315-325.

Riksrevisionen (2004). *Betyg med lika värde? En granskning av statens insatser*. RiR 2004:23. Stockholm: Riksrevisionen.

Riksrevisionen (2011). *Lika betyg, lika kunskap? En uppföljning av statens styrning mot en likvärdig betygssättning i grundskolan*. RiR 2011:23. Stockholm: Riksrevisionen.

Rinne I. (2015). *Pedagogisk takt i betygssamtal. En fenomenologisk hermeneutisk studie av gymnasielärares och elevers förståelse av betyg* (Doktorsavhandling, Göteborgs universitet, Institutionen för didaktik och pedagogisk profession)

Ross, S. J. (2005). The impact of assessment method on foreign language proficiency growth. *Applied Linguistics*, *26*(3), 317-342.

Russel, J. A., & Austin, J. R. (2010). Assessment Practices of Secondary Music Teachers. *Journal of Research in Music Education*, *58*(1), 37-54.

Rustique-Forrester, E. (2005). Accountability and the pressures to exclude: A cautionary tale from England. *Education Policy Analysis Archives*, *13*(26).

Räihä, H. (2008). *Lärares dilemman* (Doktorsavhandling, Örebro universitet).

Sawyer, R. (2013). Beyond Correlations: Usefulness of High School GPA and Test Scores in Making College Admissions Decisions. *Applied Measurement in Education*, *26*(2), 89-112.

Schneeweis, N. (2011). Educational institutions and the integration of migrants. *Journal of Population Economics*, *24*(4), 1281-1308.

Scott, S., Webber, C. F., Lupart, J. L., Aitken, N., & Scott, D. E. (2014). Fair and equitable assessment practices for all students. *Assessment in Education: Principles, Policy & Practice*, *21*(1), 52-70.

Seger, I. (2014). *Betygsättningsprocess i ämnet idrott och hälsa: en studie om betygsättningsdilemman på högstadiet* (Licentiatavhandling, Örebro universitet).

Selghed, B. (2004). *Ännu icke godkänt. Lärares sätt att erfara betygssystemet och dess tillämpning i yrkesutövningen* (Doktorsavhandling, Malmö Högskola).

Selghed, B. (2010). Ett omöjligt uppdrag. Om lärares bedömningar och betygssättning. I: Brante, G. & Hjort, K. (red). *Dilemman i skolan - aktuella utmaningar och professionella omställningar*. Kristianstad: Kristianstad University Press.

Senk, S. L., Beckmann, C. E., & Thompson, D. R. (1997). Assessment and grading in high school mathematics classrooms. *Journal for research in Mathematics Education*, *28*(2), 187-215.

Shute, V. J. (2007). *Focus on formative feedback*. Princeton, NJ: ETS.

Sikes, P., & Vincent, C. (1998). Social justice and education policy: an introduction. *Journal of Education Policy, 13*(4), 463-467.

Silva, M., Munk, D. D., & Bursuck, W. D. (2005). Grading Adaptations for Students with Disabilities. *Intervention in School & Clinic, 41*(2), 87-98.

Simon, M. Tierney, R. D., Forgette-Giroux, R. Charland, J., Noonan, B., & Duncan, R. (2010). A Secondary School Teacher's description of the process of determining report card grades. *McGill Journal of Education*, *45*(3), 535-554.

Sjögren, A. (2010). *Graded children – evidence of longrun consequences of school grades from a nationwide reform*. Working paper 2010:7. Uppsala: IFAU – Institutet för arbetsmarknad och utbildningspolitisk utvärdering.

Skolinspektionen (2014). *Uppenbar risk för felaktiga betyg. En kortrapport om likvärdighet och kvalitet i skolors betygssättning*. Skolverkets kvalitetsgranskning, rapport 2014:08.

Skolverket (2000). *Skolverkets nationella kvalitetsgranskningar. Betygsättningen*. Stockholm: Skolverket.

Skolverket (2002). *Att bedöma eller döma: tio artiklar om bedömning och betygssättning*. Stockholm: Skolverket.

Skolverket (2007). *Provbetyg-Slutbetyg-Likvärdig bedömning? En statistisk analys av sambandet mellan nationella prov och slutbetyg i grundskolan*. Rapport 300. Stockholm: Skolverket.

Skolverket (2009). *Likvärdig betygssättning i gymnasieskolan? En analys av sambandet mellan nationella prov och kursbetyg*. (Rapport 338). Stockholm: Skolverket.

SOU 1942:11. *Betänkande med utredning och förslag angående betygssättningen i folkskolan*. Stockholm: Ecklesiastikdepartementet.

SOU 1977:9. *Betygen i skolan. Betänkande av 1973 års betygsutredning*. Stockholm: Utbildningsdepartementet.

SOU 1992:86. *Ett nytt betygssystem. Slutbetänkande av Betygsberedningen*. Stockholm: Utbildningsdepartementet.

Spillane, J. P. (1999). External reform initiatives and teachers' efforts to reconstruct their practice: The mediating role of teachers' zones of enactment. *Journal of Curriculum Studies, 31*(2), 143-175.

Sprietsma, M. (2013). Discrimination in grading: experimental evidence from primary school teachers. *Empirical Economics*, 45(1), 523-538.

Stobart, G. (2005a). Fairness in multicultural assessment systems. *Assessment in Education: Principles, Policy & Practice*, *12*(3), 275-287.

Stobart, G. (2005b). What does a grade mean? *Assessment in Education: Principles, Policy & Practice, 12*(2), 101-103.

Stobart, G. (2008). Removing obstacles to fairness. *Assessment in Education: Principles, Policy & Practice*, *15*(2), 121-122.

Stobart, G. (2009). Determining validity in national curriculum assessments. *Educational Research, 51*(2), 161-179.

Stobart, G., & Eggen, T. (2012). High-stakes testing – value, fairness and consequences. *Assessment in Education: Principles, Policy & Practice*, *19*(1), 1-6.

Sun, Y., & Cheng, L. (2014). Teachers' grading practices: meaning and values assigned, *Assessment in Education: Principles, Policy & Practice*, *21*(3), 326-343.

Suurtamm C., & Koch M. J. (2014). Navigating dilemmas in transforming assessment. practices: experiences of mathematics teachers in Ontario, Canada. *Educational Assessment, Evaluation & Accountability, 26*(3), 263-287.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology*, *104*(3), 743.

Svennberg, L., Meckbach J., & Redelius, K. (2014). Exploring PE teachers' 'gut feelings': An attempt to verbalise and discuss teachers' internalised grading criteria. *European Physical Education Review, 20*(2), 199-214.

Tan, C. (2011). Framing Educational Success: A Comparative Study of Shanghai and Singapore. *Education, Knowledge and Economy*, *5*(3), 155-166.

Tekwe, C. D., Carter, R. L., Ma, C-X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., & Resnick, M. B. (2004). An Empirical Comparison of Statistical Models for Value-Added Assessment of School Performance. *Journal of Educational and Behavioral Statistics, 29*(1), 11-35.

Tholin, J. (2006). *Att kunna klara sig i ökänd natur: En studie av betyg och betygskriterier – historiska betingelser och implementering av ett nytt system* (Doktorsavhandling, Högskolan i Borås, Institutionen för pedagogik och didaktik).

Thorsen, C. (2014). Dimensions of Norm-Referenced Compulsory School Grades and their Relative Importance for the Prediction of Upper Secondary School Grades. *Scandinavian Journal of Educational Research, 58*(2), 127-146.

Thorsen, C., & Cliffordson, C. (2012). Teachers' Grade Assignment and the Predictive Validity of Criterion-Referenced Grades. *Educational Research and Evaluation, 18*(2), 153-172.

Tierney, R D., Simon M., & Charland, J. (2011). Being Fair: Teachers' Interpretations of Principles for Standards-Based Grading, *The Educational Forum, 75*(3), 210-227.

Trotter, E. (2006). Student perceptions of continuous summative assessment (medel relevans). *Assessment & Evaluation in Higher Education, 31*(5), 505-521.

Tsagalidis, H. (2008). *Därför fick jag bara Godkänt… : Bedömning i karaktärsämnen på HR-programmet* (Doktorsavhandling, Stockholms universitet, Pedagogiska institutionen).

Tveit, S. (2014). Educational assessment in Norway, *Assessment in Education: Principles, Policy & Practice*, *21*(2), 221-237.

Vaden-Goad, R. (2009). Leveraging summative assessment for formative purposes. *College Teaching, 57*(3), 153-155.

van der Kleij, K., Eggen, T. J. H. M., Timmers, C. F., & Veldkamp, B. P. (2012). Effects of feedback in a computer-based assessment for learning. *Computers & Education, 58*(1), 263-272.

van Ewijk, R. (2011). Same Work, Lower Grade? Student Ethnicity and Teachers' Subjective Assessments *Economics of Education Review*, *30*(5), 1045-1058.

Vlachos, J. (2010). *Betygets värde – en analys av hur konkurrens påverkar betygssättningen vid svenska skolor*. Konkurrensverket uppdragsforskningsrapport 2010:6.

Waldow, F. (2014). Conceptions of Justice in the Examination Systems of England, Germany, and Sweden: A Look at Safeguards of Fair Procedure and Possibilities of Appeal. *Comparative Education Review*, *58*(2), 322-343.

Waldow, F., Takayama, K., & Youl-Kwan, S. (2014). Rethinking the pattern of external policy referencing: media discourses over the 'Asian Tigers'' PISA success in Australia, Germany and South Korea. *Comparative Education*, *50*(3), 302-321.

Wang, A. H., Walters, A. M., & Thum, Y. M. (2013). Identifying highly effective urban schools: comparing two measures of school success. *International Journal of Educational Management*, *27*(5), 517-540.

Wang, J. (2001). TIMSS Primary and Middle School Data: Some Technical Concerns. *Educational Researcher*, *30*(6), 17-21.

Wedin, A-S. (2007). *Lärares arbete och kunskapsbildning Utmaningar och inviter i den vardagliga praktiken* (Doktorsavhandling, Linköpings universitet, Institutionen för beteendevetenskap och lärande).

Welsh, M. E., D'Agostino, J. V., & Kaniskan, B. 2013. Grading as a reform effort: Do standards-based grades converge with test scores? *Educational Measurement: Issues and Practice*, *32*(2), 26-36.

Widén, P. (2010). *Bedömningsmakten - Berättelser om stat, lärare och elev 1960-1995* (Doktorsavhandling, Linköpings universitet).

Wikström, C. (2005a). *Criterion-referenced measurement for educational evaluation and selection* (Doktorsavhandling, Umeå universitet, Institutionen för beteendevetenskapliga mätningar).

Wikström, C. (2005b). Grade stability in a criterion- referenced grading system: the Swedish example. *Assessment in Education: Principles, Policy & Practice*, *12*(2), 125-144.

Wikström, C. (2009). National Curriculum Assessment in England – A Swedish Perspective. *Educational Research*, *51*(2), 255-258.

Williams, J. E., Garza, L., Hodge, A. A., & Breaux, A. (1999). The color of teachers, the color of students: The multicultural classroom experience. *Teaching sociology*, *27*(3), 233-251.

Willingham, W. W., Pollack J. M., & Lewis, C. (2002). Grades and Test Scores: Accounting for Observed Differences. *Journal of Educational Measurement*, *39*(1), 1-37.

Wyatt-Smith, C., Klenowski, V., & Gunn, S. (2010). The centrality of teachers' judgment practice in assessment: a study of standards in moderation. *Assessment in Education: Principles, Policy & Practice*, *17*(1), 59-75.

Wyse, D., & Torrance, H. (2009). The development and consequences of national curriculum assessment for primary education in England. *Educational Research*, *51*(2), 213-228.

Wößmann, L. (2005). The effect heterogeneity of central examinations: evidence from TIMSS, TIMSS-Repeat and PISA. *Education Economics*, *13*(2), 143-169.

Zoeckler, L. G. (2007). Moral aspects of grading: a study of high school English teachers' perceptions. *American secondary education*, *35*(2), 83-102.

# APPENDIX: BETYGEN I EUROPA

Developed by PhD-student Sverre Tveit in cooperation with professor Christian Lundahl

The tables were developed based on available data from the Eurydice network in the fall 2014. Some errors or lost information may occur. The table overview two main areas: Compulsory education structure including required formal grading, and grading scales. The information in the columns is explained below. Endnotes and stars (*) indicate amendments made to facilitate consistent classification. These are explained on the last page.

## STRUCTURE OF COMPULSORY EDUCATION SYSTEMS

**Starting age.** Students' age when commencing compulsory education. Note that the classification of Years/levels in Eurydice may not be fully comparable. Further different interpretations across the countries in this report may cause variations. Thus all indication of age for starting school, starting formal marking, differentiation etc. may be somewhat imprecise. If the Eurydice information provided is correct it is however unlikely that the descriptions are off by more than a year.

**Compulsory Years.** Number of years of compulsory education (for compulsory pre-school, see endnotes).

**School structure.** Lists the type of compulsory structure classified in three types: ***Single structure*** education systems of which there are no parallel education during compulsory years; ***Primary Secondary*** divided education system of which there may be merit based selection when moving from primary to secondary e.g. due to requirements for progressing to next level; and ***Tracked secondary*** education systems where students are differentiated according to merit/potential and/or interests.

**Differentiation.** Details the students' age and Year when the education system is differentiated. Counted with age (Year) when differentiated schools/programs starts. For single structure systems and Primary Secondary divided education system differentiation usually occur at the conclusion of compulsory education (usually the conclusion of lower secondary education), while it in tracked secondary education systems occurs at the beginning of secondary school.

**First certificate.** List the Year and age students get the first official certificate, to the extents this is described. Classified based on Year and students' typical age when completing that Year. As it is classified based on completion of the Year instead of when commencing, students' age is up by one compared to the classifications for "Starting school" and "Grading required". In some countries there are immatriculation certificates in the forthcoming year; these countries are still classified according to the conclusive Year before immatriculation.

**Grading required:** Lists the age and Year when teachers' grading is required. Classified with age when starting the relevant Year. This classification should be viewed as indicative, as many countries do not address the issue of formal grading explicitly in their reports.

# GRADING SCALES

**First grading scale:** List the first grading scale students meet in school. Much information missing. List the grading scales that have been identified in the generated data. In many countries (such as in the UK) students meet grading scales earlier than reported here, however in many cases Eurydice does not list scales used for classroom assessment. Some countries have different grading scales in primary and secondary education or other ways of combining multiple grading scales. Only the reported grading scales students meet first is listed here. This should not been taken as a exhaustive reporting of grading scales, as many countries have either not reported it properly and in several cases there was not even made note of it in Eurydice.

**Scale levels.** Number of grading scale levels, considered including 1 failure level (some countries have additional failing levels)

**Scale type:** Classifies the type of grading scale, to the extent it has been possible to establish, along the following attributes: Verbal, Verbal and letters, Verbal and numerical, Numerical, Letters, Points scale.

**Grading behavior**: Addresses whether the countries report to have a legal framework for grading students' behaviour with distinct grades. There are various ways and terms for grading behavior (order, conduct, diligence). This should not be regarded as an exhaustive classification; it appears only countries that employ such practices describe it, albeit it is not possible to know whether that implies the others do not.

| STRUCTURE OF COMPULSORY EDUCATION SYSTEMS | | | | | | |
|---|---|---|---|---|---|---|
| **COUNTRY** | **Starting age** | **Compulsory years** | **School structure** | **Differen-tiation** | **First certificate** | **Grading required** |
| **AUSTRIA** | Age 6 | 9 | Tracked secondary | Age 10 (Year 5) | Year 4 (Age 10) | Age 6 (Year 1) |
| **BELGIUM (Flemish)** | Age 6 | 9 | Tracked secondary | Age 12 (Year 7) | Year 6 (Age 12) | Not described |
| **BELGIUM (French)** | Age 6 | 9 | Tracked secondary | Age 12 (Year 7) | Year 6 (Age 12) | Age 7 (Year 2) |
| **BELIGUM (German)** | Age 6 | 9 | Tracked secondary | Age 12 (Year 7) | Year 6 (Age 12) | Not described |
| **BULGARIA** | Age 7 | 9 | Single structure | Age 15 (Year 9) | Year 8 (Age 15) | Not described |
| **CROATIA** | Age 6[i] | 9 | Single structure | Age 14 (Year 9) | Not described | Not described |
| **CYPRUS** | Age 6[ii] | 10 | Primary Secondary | Age 12 (Year 7) | Year 1 (Age 7) | Age 6 (Year 1) |
| **CZECH REPUBLIC** | Age 6 | 9 | Primary Secondary | Age 15 (Year 10) | Year 5 (Age 11) | Not described |
| **DENMARK** | Age 7[iii] | 9 | Single structure | Age 16 (Year 10) | Year 9 (Age 16) | Age 14 (Year 8) |
| **ESTONIA** | Age[iv] 7 | 9 | Single structure | Age 16 (Year 10) | Not described | Not described |
| **FINLAND** | Age 7 | 9 | Single structure | Age 16 (Year 10) | Year 9 (Age 16) | Age 14 (Year 8) |
| **FRANCE** | Age 6 | 10 | Primary Secondary | Age 11 (Year 6)* | Year 9 (Age 15) | Age 6 (Year 1) |
| **GERMANY** | Age 6 | 9 | Tracked secondary | Age 10** (Year 5) | Year 9 (Age 15) | Age 7 (Year 2) |
| **GREECE** | Age 6[v] | 10 | Primary Secondary | Age 12 (Year 7) | Year 9 (Age 15) | Age 8 (Year 3) |
| **HUNGARY** | Age 6[vi] | 9 | Single structure | Unclear | Year 1 (Age 7) | Age 6 (Year 1) |
| **ICELAND** | Age 6[vii] | 10 | Single structure | Age 16 (Year 11) | Year 10 (Age 16) | Not described |
| **IRELAND** | Age 6[viii] | 10 | Tracked secondary | Age 12 (Year 7) | Year 7 (Age 13) | Not described |
| **ITALY** | Age 6[ix] | 10 | Inconsitancy | Unclear | Year 10 (Age 16) | Age 6 (Year 1) |

| GRADING SCALES | | | | |
|---|---|---|---|---|
| **COUNTRY** | **First grading scale** | **Scale levels** | **Scale type** | **Grading behavior** |
| **AUSTRIA** | Very good (1), Good (2), Satisfactory (3), Sufficient (4), Insufficient (5). | 5 | Verbal and numerical | Not described |
| **BELGIUM (Flemish)** | Not described | Not described | Not described | Not described |
| **BELGIUM (French)** | Not described | Not described | Not described | Not described |
| **BELIGUM (German)** | Not described | Not described | Not described | Not described |
| **BULGARIA** | Excellent (6), Very good (5), Good (4), Fair (3) and Poor (2). | 6 * | Verbal and numerical | Not described |
| **CROATIA** | Excellent, Very good, Good, Sufficient, Insufficient. | 5 | Verbal | Not described |
| **CYPRUS** | General lower secondary: A=Excellent, B=Very Good, C=Good, D=Almost Good, E=Fail. | 5 | Verbal and letters | Not described |
| **CZECH REPUBLIC** | Not described | Not described | Not described | Not described |
| **DENMARK** | 12, 10, 7, 4, 02, 00, -3. | 6 ** | Numerical | Not described |
| **ESTONIA** | Very good (5), Good (4), Satisfactory (3), Poor (2), Weak (1). | 5 *** | Verbal and numerical | Not described |
| **FINLAND** | Excellent (10), Very good (9), Good (8), Satisfactory (7), Moderate (6), Adequate (5), Failure (4). | 7 | Verbal and numerical | Not described |
| **FRANCE** | 20-1. | 20 | Not described | Not described |
| **GERMANY** | Very good (1), Good (2), Satisfactory (3), Adequate (4), Poor (5), Very poor (6). | 6 | Verbal and numerical | 50 % of the länder |
| **GREECE** | Year 3 and 4: Excellent (A), Very Good (B), Good (C), Fairly Good (D). | 4 **** | Verbal and letters | Not described |
| **HUNGARY** | Very good (5), Good (4), Satisfactory (3), Pass (2), Fail (1). | 5 | Verbal and numerical | Conduct, diligence |
| **ICELAND** | Honors (9.00 - 10.00), First Class (7.25 - 8.99), Second Class (6.00 - 7.24), Third Class (5.00 - 5.99), Fail (0.00 - 4.99). | 5 | Verbal and numerical | Not described |
| **IRELAND** | Not described | Not described | Not described | Not described |
| **ITALY** | 10-1 (6 is pass) | Not described | Not described | Behavior |

| STRUCTURE OF COMPULSORY EDUCATION SYSTEMS | | | | | | |
|---|---|---|---|---|---|---|
| **COUNTRY** | **Starting age** | **Compulsory years** | **School structure** | **Differen-tiation** | **First certificate** | **Grading required** |
| **LATVIA** | Age 7[x] | 9 | Inconsitancy | Unclear | Year 9 (Age 16) | Not described |
| **LICHTENSTEIN** | Age 6 | 9 | Tracked secondary | Age 11 (Year 6) | Not described | Age 11 (Year 6) |
| **LITHUANIA** | Age 7 | 9 | Primary Secondary | Unclear | Year 4 (Age 11) | Age 11 (Year 5) |
| **LUXEMBOURG** | Age 6 | 12 | Tracked secondary | Unclear | Year 6 (Age 12) | Age 7 (Year 2) |
| **MALTA** | Age 5 | 11 | Primary Secondary | Unclear | Year 11 (Age 16) | Age 8 (Year 4) |
| **NETHERLANDS** | Age 6[xi] | 13 | Tracked secondary | Age 13 (Year 8) | Age 13 (Year 7) | Not described |
| **NORWAY** | Age 6 | 10 | Single structure | Age 16 (Year 11) | Year 10 (Age 16) | Age 13 (Year 8) |
| **POLAND** | Age 6[xii] | 10 | Primary Secondary | Unclear | Year 1 (Age 7) | Age 6 (Year 1) |
| **PORTUGAL** | Age 6 | 12 | Single structure | Age 15 (Year 10) | Year 6 (Age 12) | Age 9 (Year 4) |
| **ROMANIA** | Age 6 | 11 | Primary Secondary | Age 15 (Year 10) | Year 9 (Age 15) | Age 6 (Year 1) |
| **SLOVAKIA** | Age 6 | 10 | Single structure | Unclear | Year 1 (Age 7) | Age 9 (Year 4) |
| **SLOVENIA** | Age 6 | 9 | Single structure | Age 15 (Year 10) | Year 1 (Age 7) | Age 8 (Year 3) |
| **SPAIN** | Age 6 | 10 | Single structure | Unclear | Year 10 (Age 16) | Not described |
| **SWEDEN** | Age 7 | 9 | Single structure | Age 16 (Year 10) | Year 9 (Age 16) | Age 12 (Year 6) |
| **TURKEY** | Age 7[xiii] | 12 | Single structure | Age 15 (Year 9) | Year 4 (Age 11) | Age 7 (Year 1) |
| **UK-ENGLAND** | Age 5 | 11 | Primary Secondary | Age 16 (Year 12) | Year 11 (Age 16) | Age 6 (Year 2) |
| **UK:NORTHERN IRELAND** | Age 4 | 12 | Primary Secondary | Age 16 (Year 12) | Year 11 (Age 16) | Age 5 (Year 2) |
| **UK-SCOTLAND** | Age 5 | 11 | Primary Secondary | Age 16 (Year 12) | Year 11 (Age 16) | Not described |
| **UK-WALES** | Age 5 | 11 | Primary Secondary | Age 16 (Year 12) | Year 11 (Age 16) | Age 6 (Year 2) |

| GRADING SCALES | | | | |
|---|---|---|---|---|
| **COUNTRY** | **First grading scale** | **Scale levels** | **Scale type** | **Grading behavior** |
| **LATVIA** | 10 (With distinction), 9 (Excellent), 8 (Very good), 7 (Good), 6 (Almost good), 5 (Satisfactory), 4 (Almost satisfactory), 3 (Weak), 2 (Very weak), 1 (Very, very weak). | 7<br>***** | Verbal and numerical | Not described |
| **LICHTENSTEIN** | Not described | Not described | Not described | Not described |
| **LITHUANIA** | Not described | Not detailed | Not detailed | Not described |
| **LUXEMBOURG** | Upper secondary: Excellent (> 52 points), Assez bien (36 to 39 points), Bien (40 to 47 points), Très bien (48 to 51 points). | Not described | Points scale | Not described |
| **MALTA** | 100 point scale | Not described | Points scale | Not described |
| **NETHERLANDS** | Not described | Not described | Not described | Not described |
| **NORWAY** | 6 (exceptional), 5 (very high), 4 (high), 3 (fair), 2 (low), 1 (very low). | 6 | Verbal and numerical | Order and conduct |
| **POLAND** | Excellent (6), Very good (5), Good (4), Satisfactory (3), Acceptable (2), Unsatisfactory (1). | 6 | Verbal and numerical | Behavior (conduct) |
| **PORTUGAL** | 5, 4, 3, 2, 1 (3 minimum to pass). | 4 | Not detailed | Not described |
| **ROMANIA** | Primary: Very good, Good, Sufficient, Insufficient. | 4 | Verbal | Behavior |
| **SLOVAKIA** | Scale 1: Excellent (1), Laudable (2), Good (3), Satisfactory (2), Fail (5). | 5<br>4 | Verbal and numerical | Behavior |
| **SLOVENIA** | 5, 4, 3, 2, 1. | 5 | Numerical | Not described |
| **SPAIN** | Not described | Not described | Not described | Not described |
| **SWEDEN** | A, B, C, D, E, F. | 6 | Letters | Not described |
| **TURKEY** | Not described | 9 | Not described | Not described |
| **UK-ENGLAND** | Lower secondary GCSE: 9-8-7-6-5-4-3-2-1. | 9 | Numerical | Not described |
| **UK:NORTHERN IRELAND** | Lower secondary GCSE: 9-8-7-6-5-4-3-2-1 | 9 | Numerical | Not described |
| **UK-SCOTLAND** | Not described | 9 | Not described | Not described |
| **UK-WALES** | Lower secondary GCSE: 9-8-7-6-5-4-3-2-1. | 9 | Numerical | Not described |

## REMARKS TO APPENDIX TABLE:

**STARTING AGE:**

[i] **Croatia**: There is a compulsory pre-school year, however not counted as first school Year.
[ii] **Cyprus**: School starts Age 5 + 8mth. There is a compulsory pre-school year, however not counted as first school Year.
[iii] **Denmark**: There is one compulsory pre-school year, however not counted as first school Year. The optional 10th year before upper secondary is not counted, thus first secondary year is listed as Year 10.
[iv] **Estonia**: There is a compulsory pre-school year, however not counted as first school Year.
[v] **Greece**: There is a compulsory pre-school year, however not counted as first school Year
[vi] **Hungary**: There is a compulsory pre-school year, however not counted as first school Year
[vii] **Iceland**: There is a compulsory pre-school year, however not counted as first school Year
[viii] **Ireland**: There is a compulsory pre-school year, however not counted as first school Year
[ix] **Italy**: There is a compulsory pre-school year, however not counted as first school Year
[x] **Latvia**: There are two compulsory pre-school years, however not counted as first school Year. The optional 10th year before upper secondary is not counted, thus first secondary year is listed as Year 10.
[xi] **Netherlands**: There is a compulsory pre-school year, however not counted as first school Year
[xii] **Poland**: There is a compulsory pre-school year, however not counted as first school Year
[xiii] **Turkey**: School starts Age 6+6mth. There is a compulsory pre-school year, however not counted as first school Year

**COMPULSORY YEARS:**

* **Germany:** In 12 länder there are 9 compulsory years; in 5 länder there are 10 compulrosy years.

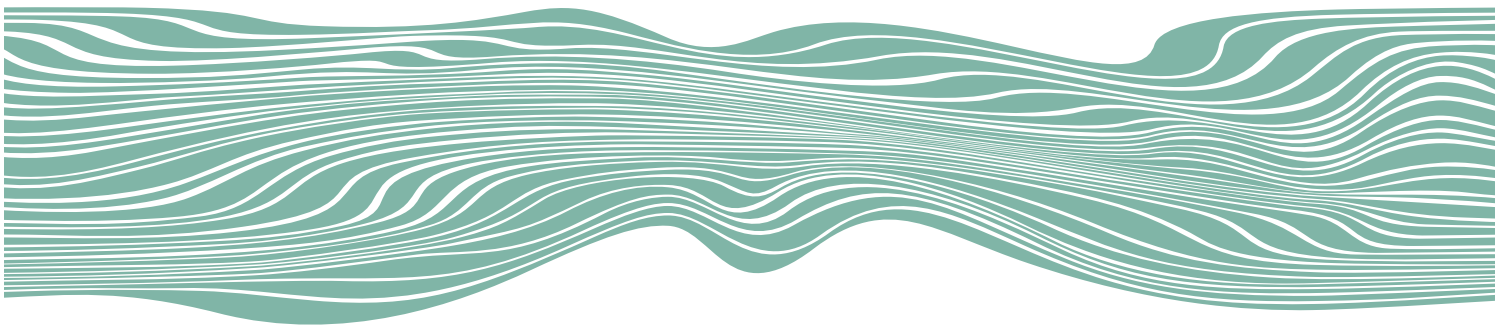
**DIFFERENTIATION:**

* **France:** There are many school types and thus highly complex to describe how students are differentiated based on merit.
** **Germany:** In Berlin and Brandenburg secondary levels start Age 12 (Year 7).

**GRADING LEVELS:**

* Bulgaria: Failing level not described
** Denmark: One additional failure level
*** Estonia: Failure level not addressed
**** Greece: Year 1-6 have 4 levels; Year 7-9 have 5 levels.
***** Latvia: Additional 3 failing levels.

Vetenskapsrådet genomförde under 2014 ett projekt, SKOLFORSK, för att kartlägga befintlig utbildningsvetenskaplig forskning. Arbetet skedde på uppdrag av regeringen för att resultera i kartläggningar av svenska och internationella forskningsresultat med relevans för skolväsendet. Syftet var att skapa en plattform av kunskapsunderlag till det nybildade Skolforskningsinstitutet. Slutsatserna i denna delrapport är författarnas egna. Vetenskapsrådets sammanfattande rapport, Forskning och skola i samverkan, med en beskrivning av projektet och med de frågeställningar, resultat och rekommendationer som redovisats inom delprojekten kan liksom de övriga delrapporterna laddas ner från Vetenskapsrådets webbplats.



Västra Järnvägsgatan 3 | Box 1035 | 101 38 Stockholm | Tel 08-546 44 000 | vetenskapsradet@vr.se | www.vr.se