Pedagogisk Forskning i Sverige årg 20 nr 1-2 2015 issn 1401-6788

Validitet och lärares bedömningar

Stefan Johansson
Institutionen för pedagogik och specialpedagogik,
Göteborgs universitet.

Rättvis bedömning är avgörande för skolans likvärdighet. En mycket viktig del av lärarkompetensen är därför förmågan att adekvat kunna bedöma elevernas kunskapsnivåer. Hög precision i kunskapsbedömningen är av stor betydelse för lärares möjlighet att kunna ge lämplig återkoppling, vilket i sig är av stor vikt för elevers möjlighet att lära. I denna studie undersöks validiteten, det vill säga giltigheten i lärares bedömningar av elevers läs- och skrivförmåga. För att analysera validiteten användes data från den storskaliga undersökningen Progress in International Reading Literacy Study 2001 (PIRLS), där elever i åk 3 har genomfört ett skriftligt läsprov som bedömts externt. Dessutom har elevernas lärare bedömt sina elevers läs- och skrivförmåga utifrån 12 olika aspekter. Analyserna visar att lärares bedömningar överensstämmer relativt väl med provresultat i PIRLS inom den egna klassen. Däremot tenderar bedömningsnivån att skilja sig mellan olika lärare, trots att skolklasser uppvisar jämbördig nivå på läsprovet. Lärares bedömningar påverkas också av elevernas kön och socioekonomiska bakgrund.

BAKGRUND

Att bedöma elevers kunskaper och färdigheter är ett av skolans viktigare uppdrag och därför riktas mycket uppmärksamhet mot lärares förmåga att bedöma elevers kunskapsnivåer. Många lärare i Sverige ett dubbelt uppdrag i och med att de betygsätter eller ger elever betygsliknande omdömen, samtidigt som de gör bedömningar för att stödja elevernas lärande. Lärare måste kunna bedöma sina elevers kunskaper och färdigheter på ett rättvist och rättssäkert sätt, i linje med de kunskapskrav som är preciserade i styrdokument. Rättssäkert sätt innebär dessutom strävan efter likvärdig bedömning mellan olika lärare så att tolkningen av uppdraget sker på ett samstämmigt sätt. Samtidigt

måste lärare kunna identifiera enskilda elevers styrkor och svagheter under lärandeprocessen, för att kunna ge lämplig återkoppling, en kompetens som är av extra vikt för elever i behov av särskilt stöd, då rätt insatser kan vara avgörande för deras fortsatta skolgång.

Den här artikeln fokuserar två huvudfrågor. En fråga handlar om hur väl lärare kan differentiera mellan elevernas kunskaper i den egna klassen. Det standardiserade läsprovet inom PIRLS används för att analysera validiteten i lärares bedömningar. Elevernas resultat på läsprovet ska dock inte betraktas som en standard, eller sant kriterium. Båda måtten på kunskaper har sina brister, men också styrkor. Jämförelsen mellan provresultaten och lärarnas bedömningar kan därför ses som en ömsesidig belysning av validiteten. Vid jämförelser på klassrumsnivå har dock PIRLS-provet fördelen att vara lika för alla, och inte påverkat av till exempel den enskilda lärarens bedömning, då proven rättas externt. I Sverige har vi för närvarande inte några mått på kunskaper, vid sidan av de internationella undersökningarna, som kan ge en god bild på systemnivå (se t.ex., Gustafsson, Cliffordson, & Erickson, 2014), men alla ämnen omfattas inte av dessa undersökningar. Skolverket (2009; 2012) har rapporterat om en viss likvärdighetsproblematik i betygsättning. Av dessa studier framgår det att lärares betygssättning skiljer sig åt, trots att eleverna får jämbördiga resultat på nationella prov. Den andra frågan av intresse i den här artikeln är därför att undersöka graden av överenstämmelse mellan olika lärares bedömningar med hjälp av detta externa bedömningsinstrument som PIRLS erbjuder. Således är det övergripande syftet med föreliggande studie att undersöka hur väl lärares bedömningar fungerar som mått på elevers kunskaper och färdigheter, såväl inom klassrum som mellan olika klassrum. Särskilt undersöks om och i så fall hur elevens kön och socioekonomiska bakgrund påverkar validiteten, det vill säga giltigheten, i lärarnas bedömningar.

HUR GILTIGA ÄR LÄRARES BEDÖMNINGAR?

Intresset för adekvat och likvärdig bedömning är stort hos elever, föräldrar, rektorer, myndigheter, politiker och massmedia. Under det senaste decenniet har diskussionen om bedömning i skolan intensifierats. Ett tecken på detta är det ökande antal fackböcker och avhandlingar som har producerats i allt snabbare takt inom bedömningsområdet (se t.ex., Andreasson, 2007; Balan, 2012; Hirsh, 2013; Johansson, 2013; Jönsson, 2008; Klapp Lekholm, 2008; Lundahl, 2006; Nyström, 2004; Westlund, 2013). Vad är det då som gör att intresset för bedömning i allmänhet, och lärares bedömning i synnerhet, är så stort? Det finns några historiska omständigheter i svensk skola under de senaste decennierna som kan anges som bidragande faktorer till denna utveckling.

I början av 1990-talet avreglerades skolväsendet. Från att ha varit centralstyrd blev skolan istället kommunalt styrd och privata aktörer etablerade sig på skolmarknaden. Under samma period började också lärare från en reformerad lärarutbildning att arbeta i skolorna. År 1994 sjösattes en ny läroplan och nya kursplaner med betygskriterier. De målrelaterade betygskriterierna innebar ett större tolkningsutrymme men även ett större ansvar för lärarna att upprätthålla likvärdig och rättvis bedömning och betygsättning. Det dröjde dock inte länge förrän kritiska röster höjdes om giltigheten i lärarnas bedömning, eller i varje fall om giltigheten i det målrelaterade bedömningssystemet (Selghed, 2004; Tholin, 2006).

Tholin (2006) visade att det var mycket svårt att få enskilda skolor att tolka nationella styrdokument och samtidigt skapa lokala betygskriterier på ett likvärdigt sätt. Några effektiva hjälpmedel för att kalibrera bedömningar lärare emellan som standardiserade centralprov fanns inte, och själva betygsystemet var inte uppbyggt för att kunna göra likvärdiga bedömningar. Kriterierna var något vagt framskrivna och skapade inte en gemensam referensram på nationell nivå bland lärarna. Lärarna hade visserligen tillgång till nationella prov i några ämnen, men dessa förmådde inte heller understödja likvärdig bedömning. Eftersom flertalet av de verksamma lärarna hade erfarenheter och utbildning från tiden då det normrelaterade betygsystemet gällde, var det knappast någon överraskning att det skulle ta tid att implementera det nya betyg- och bedömningssystemet. Selghed (2004) menade att lärarna efter dryga fem år av målrelaterat betygsystem fortfarande levde kvar i normrelaterade strategier, såsom klassens medelvärde eller liknande, då de gjorde sina bedömningar.

Vid sidan om likvärdighetsproblematiken, visade de internationella komparativa studierna (t.ex., PIRLS, TIMSS, PISA) en nedåtgående resultattrend i Sverige, samtidigt som de målrelaterade betygen visade något helt annat. Gustafsson och Yang-Hansen (2009) studerade betygsutvecklingen över tid och kunde påvisa att meritvärdet hade ökat under en längre period med störst ökning några år efter det nya betygssystemets införande 1994. Med största sannolikhet är dessa resultat tecken på betygsinflation, vilken har varit gällande i alla ämnen i varierande grad (Gustafsson & Yang-Hansen, 2009).

Det fördjupade intresset för lärares bedömningar har fått konsekvenser för bedömningspraktiken och för lärarprofessionen. En följd av detta är till exempel att nationella prov har implementerats i ökad omfattning för att om möjligt få en mer likvärdig bedömning. Sedan 2008 har även lärare i årskurs 3 fått tillgång till nationella prov med förhoppningar om att de tidigare åldrarna ska garanteras mer likvärdig utbildning. På det hela taget har utvecklingen gått mot en skärpt kontroll av skolorna och lärarnas bedömning av elevers kunskaper och färdigheter (Jarl & Rönnberg, 2010).

VALIDITET

Validitet eller giltighet är en av två övergripande kvalitetsaspekter som rör mätning och bedömning. En bedömning har god validitet om den mäter det som avsetts och om dess slutsatser och konsekvenser är riktiga och sunda. Den andra kvalitetsaspekten är reliabilitet eller tillförlitlighet och syftar på konsistens och precision i en bedömning, till exempel att olika bedömningar (av olika bedömare) av samma sak ger samma utfall. I en del forskning separeras begreppen validitet och reliabilitet men ofta ses reliabilitet som en del i det vidgade validitetsbegreppet, där reliabilitet betraktas som en nödvändig men inte tillräcklig, förutsättning för validitet (Messick, 1989). Att säga att något fungerar bra, eller som det är tänkt, är att säga att något har god validitet. I den här artikeln är fokus riktat mot hur bra lärares bedömningar fungerar, det vill säga hur valida de är. Validitetsbegreppet är nära sammankopplat med begreppet rättvisa – vilket återkommande fokuseras i diskussionen om bedömning och betygsättning. Kane (2010) har beskrivit validitet och rättvisa som att de refererar till samma grundläggande fråga – om den föreslagna slutsatsen och användningen av en bedömning är lämplig för en viss målgrupp. Den gängse definitionen av validitet som har vuxit fram över tid (Cronbach, 1971; Messick, 1989; Kane, 2006) är alltomfattande. I mitten av 1900-talet lades inte någon tyngd på bedömningens slutsatser och konsekvenser, utan på själva mätinstrumentet (Wolming, 1998). Ett högt samband mellan en bedömning och ett kriterium av intresse förenades med hög validitet (Guilford, 1946). I det alltomfattande validitetsbegreppet behövs mer än empirisk evidens för hög validitet; även de teoretiska utgångspunkterna ska överensstämma med begreppet som valideras. Ett ordprov skulle kunna fungera som ett mått på läsförmåga därför att korrelationen är hög – men eftersom ordprovet saknar ett teoretiskt ramverk som är i samklang med en bredare förståelse av begreppet läsning försämras validiteten. Vidare krävs att slutsatser och beslut fattade på grundval av en bedömning är adekvata.

När den alltomfattande definitionen av validitet används i relation till lärares bedömningar, kan det vara lämpligt att bryta ner själva valideringen av bedömningen i två centrala beståndsdelar. Den ena innebär att lärares bedömningar ska vara ett mått på adekvata kunskaper i relation till aktuella mål och kriterier. En sätt att bedöma relevansen och täckningen av lärarnas bedömda domän är att undersöka om det föreligger ett starkt samband med ett annat, liknande mått på kunskaper och färdigheter som till exempel ett nationellt prov eller andra relevanta bedömningar gjorda av någon annan. Bedömningen måste också ta sin utgångspunkt i ett teoretiskt ramverk, det vill säga de styrdokument som är aktuella och knutna till bedömt kunskapsområde. Den andra centrala aspekten av valideringen innebär att de slutsatser som dras, och de handlingar som utförs, ska bli korrekta. Dessa ska bygga

på giltigt underlag som kan motivera resultaten och som kan accepteras som mått på adekvata kunskaper. Vid användning av bedömningsresultaten kan dock validiteten försämras, trots att de inledande kraven är uppfyllda. Ett exempel på att validiteten kan bli sämre är då högskoleprovsresultat används vid antagning till högre studier. Cliffordson (2004) har visat att både norm- och målrelaterade betyg prognostiserar genomströmningen i civilingenjörsprogrammet klart bättre än högskoleprovsresultat. Konsekvensen av att anta studenter till högre utbildning på basis av högskoleprovsresultat blir att genomströmningen minskar. Därmed sjunker också validiteten i provet. Det bör i det här sammanhanget uppmärksammas att flera forskare menar att konsekvenser över huvud taget inte ska blandas in i validitetsbegreppet (se t.ex. Borsboom, Cramer, Kievit, Scholten, & Franic, 2009; Mehrens, 1997). Validitet definieras av dem som en egenskap tillhörande enbart mätinstrumentet, i det här fallet lärarnas bedömningar. Om bedömningen inte kunde mäta det som var avsett att mätas i första steget, finns det enligt Mehrens (1997) och Borsboom m.fl., (2009) inte något skäl att undersöka eller resonera i termer av konsekvensvaliditet. Använder man ett alltomfattande validitetsbegrepp är det viktigt att tillse att mätinstrumentet mäter det som det som var avsett, innan resonemang i termer av konsekvenser och användning kan komma ifråga. Även om konsekvensvaliditet är en viktig och intressant aspekt av bedömning, undersöks i föreliggande studie inte lärares bedömningar utifrån dess användning eller konsekvenser.

RELATIONEN MELLAN LÄRARES BEDÖMNINGAR OCH PROVRESULTAT

Forskare som uttryckt skepsis till lärares bedömningar menar att de kan vara föremål för påverkan av faktorer som relaterar till exempelvis elevens motivation, socioekonomiska bakgrund eller kön (Coladarci, 1986; Hoge & Coladarci, 1989). Dessa faktorer kan utifrån detta perspektiv på validitet betecknas som irrelevanta faktorer, eller *construct irrelevant variance* med Messicks (1989) terminologi, och därmed hota bedömningarnas giltighet. Det ska dock poängteras att exempelvis motivation är viktigt för studieframgång men att det inte enskilt ska utgöra grund för bedömning. Via ett stort urval av elever i årskurs 9 visade Klapp Lekholm (2008) i sin avhandling genom multivariata analyser av betygsvariansen att betygen återspeglar just faktorer som ligger utanför elevernas kognitiva kunskaper. På både individ- och skolnivå förklarades emellertid den största delen av variansen av den kognitiva förmågan, vilken i Klapp Lekholms studie mättes med de nationella provresultaten i svenska, matematik och engelska. För att vända blicken internationellt kan nämnas Brookhart's (2012) forskningsöversikt som visade att

tilliten till lärares bedömningar i USA inte är särskilt stor, vilket kan förklara användningen av och förtroendet för standardiserade prov som en kontrast till Sverige. Flera tidigare studier förstärker Brookharts resonemang: Cizek, Fitzgerald och Rachor (1996) studerade i sin studie lärarbedömning i låg-, mellan- och högstadiet i USA. Lärarna tillfrågades om vilka faktorer de lade vikt vid då eleverna skulle ges betyg. Generellt var elevens prestation i det aktuella ämnet det som lärarna fäste mest vikt vid när de bedömde eleverna, men även ansträngning, motivation och närvaro uppgavs som centralt för betygsättningen. I linje med dessa resultat är Llosas (2007) som visade att det finns konsistens mellan lärares bedömningar och elevers provresultat, men att det finns variation med avseende på vad lärare bedömer. Flera lärare uppgav att de baserade sina utlåtanden på elevkarakteristika, så som elevens personlighet, något som svensk forskning på individuella utvecklingsplaner och åtgärdsprogram bekräftar (se t.ex., Andreasson, 2007; 2012). Vidare har Martínez, Stecher och Borko (2009) i en serie multivariata analyser visat att flera elevfaktorer påverkar lärares bedömningar, så som elevens kön. Pojkarna i årkurs 3 och 5 i studien presterade bättre på det standardiserade matematikprovet, vilket inte reflekterades i lärarnas bedömning där flickors och pojkars prestationer bedömdes relativt lika. Samma mönster gällde även minoritetselever och svagpresterande elever. En förklaring som Martínez m.fl. (2009) lägger fram är att lärarna kan uppleva att dessa elever har olika svårigheter och därför kompenserar dessa elever genom att justera sina bedömningar uppåt, alternativt justerar sina förväntningar på måluppfyllelse nedåt.

Lärares bedömningar har dock ansetts vara giltiga, eftersom överensstämmelsen mellan dem och andra mått på prestation varit hög. Südkamp, Kaiser och Möller (2012) har gjort en meta-analys av studier som undersökt korrelationen mellan lärares bedömningar av elevers kunskaper och provresultat. Totalt 75 studier från olika länder i världen ingick i analysen. Medelkorrelationen mellan lärarnas bedömning och provresultaten var 0.63, vilket de tyska forskarna bedömde som relativt högt. Detta samband stämmer väl överens med resultaten från en tidigare forskningsöversikt gjord av Hoge och Coladarci (1989). I deras studie var medelkorrelationen 0.63 och mediankorrelationen 0.66. På grund av stor variation mellan sambanden i de ingående studierna (0.28 till 0.92) redovisades mediankorrelationen. Slutsatsen som drogs var dock att lärarbedömningarna bör ses som valida mått på elevers kunskaper och färdigheter samt att resultaten därmed undergräver de kritiska rösterna om lärares bedömningar. Ytterligare stöd för lärares bedömningar har framförts av Gipps (1994) som menade att lärares bedömningar kunde anses vara giltiga mått, då lärare generellt observerat sina elever under lång tid.

Som visats i tidigare forskning varierar tilltron till lärares bedömningar mellan olika studier men också mellan länder. De meta-analyser som finns

att tillgå visar dock att överensstämmelsen mellan lärares bedömning och andra mått på elevers kunskaper är relativt hög, mätt i korrelationskoefficienter. Däremot finns det lite stöd för att lärare förmår att göra likvärdiga bedömningar mellan olika klassrum. Få studier har kombinerat resultat där förhållandena reds ut mellan dels läraren och dennes egna elever, dels lärare emellan. Få studier redovisar också resultat för skolans tidigare år. En anledning till denna avsaknad av tillgängliga studier på området har att göra med svårigheterna att ta fram adekvata data. Lärarbedömningsdata måste finnas tillgänglig jämte ett annat mått på prestation, och hela klasser måste ha dragits i urvalet om det samtidigt ska gå att jämföra hur väl lärare kan bedöma sina egna elevers kunskaper, som hur samstämmigt lärarna bedömer elevers kunskaper generellt. I de tidigare skolåren är det ovanligt med betyg eller andra resultatmått, kanske speciellt i Sverige. Den föreliggande studien redovisar resultat från analyser av läsundersökningen PIRLS 2001, där elevers läsförmåga i årskurs 3 och 4 prövades. De tillgängliga data är väl lämpade för att studera validitet i lärares bedömningar.

METOD OCH DATA

The Progress in International Reading Literacy Study (PIRLS) är en av the International Association for the Evaluation of Educational Achievement (IEA) världsomspännande kunskapsmätningar för elever i årskurs 4. År 2001 ingick 35 länder i PIRLS-undersökningen och data från elever, föräldrar, lärare och skolledare finns tillgängliga. PIRLS design finns beskriven såväl i den internationella rapporten (Gonzalez & Kennedy, 2003) som i den svenska (Rosén, Myrberg, & Gustafsson, 2005). Detta år ingick även årskurs 3 i urvalet i Sverige och eftersom lärare i allmänhet haft sina elever under flera år i årskurs 3, men inte i årskurs 4, ansågs årskurs 3 lämpligare för denna studie. Totalt ingick 5271 elever och 351 lärare i studien.

Det material som är grundläggande för alla analyser är dels de aspektbedömningar, här omnämnda som lärares bedömningar, som lärare gjort av elevers kunskaper och färdigheter i läsning och skrivning utifrån ett antal aspekter, dels elevernas provresultat i PIRLS läsprov.

LÄRARES BEDÖMNING

I anslutning till PIRLS-undersökningen 2001 gjordes ett nationellt tillägg i den svenska designen som innebar att lärare också fick bedöma sina elevers läs- och skrivförmåga. Lärarbedömningarna bygger på iakttagelser och erfarenheter av elevernas läs- och skrivförmåga i det löpande skolarbete och de är inte explicit knutna till provuppgifterna i PIRLS-undersökningen. De

aspekter av läs- och skrivutveckling som läraren ombads att ta ställning till presenteras i Tabell 1. Lärarna som deltagit i undersökningen bedömde tolv olika aspekter av elevers läs- och skrivförmåga på en tio-gradig skala. Varje aspekt består av ett påstående som läraren skattar elevens förmåga i relation till. Till varje påstående gavs läraren också utrymme för att lämna kommentarer. När dessa 12 bedömningsaspekter senare studeras är det inte en och en utan tillsammans. Utgångspunkten är att variabeln *lärares bedömning* baseras på ett flertal kognitiva dimensioner, ungefär som ett betyg.

Tabell 1. Beskrivande statistik för de 12 bedömningsaspekterna.

Variabel	Påstående			
	Eleven…	N	Medel	Std.Av
01	…bygger meningar korrekt	5208	7.67	2.16
02	…känner igen ofta återkommande ord i texten	5213	8.35	1.93
03	…kan knyta en berättelse till egna erfarenheter	5162	8.26	1.85
04	…tar hjälp av textens sammanhang vid läsning	5207	8.05	2.05
05	…skriver sammanhängande (berättelser, egna upplevelser, fantasiberättelser)	5209	7.84	2.18
06	…förstår textinnehåll efter egen läsning	5124	8.30	2.00
07	…känner igen alla bokstäver och kopplar bokstavstecken till ljud	5136	9.48	1.27
08	…kan läsa obekanta ord	5133	8.11	2.03
09	…kan reflektera kring en berättelse	5083	8.09	1.90
10	…läser flytande	5135	8.32	2.10
11	…kan förbättra sin egen text	5072	7.11	2.24
12	…har ett rimligt stort ordförråd, dvs. som räcker i olika situationer	5132	8.30	1.89

Påståendena som lärarna utgick från i sina bedömningar var utarbetade efter diagnosinstrumentet ”Språket lyfter” som Skolverket gav ut 2002. Skolverket (2002) utformade det diagnostiska materialet dels eftersom de tidiga skolåren tidigare saknat ett material för systematisk uppföljning och utvärdering, dels för att det inte fanns något fast regelverk i form av kursplaner med mål och kriterier för dessa skolår. Eftersom de tidigare inte regelmässigt följt upp elevers kunskaper och färdigheter i de lägre årskurserna på nationell nivå, ansågs det vara av största vikt att pröva Skolverkets instrument (Rosén,

m.fl., 2005). Inför PIRLS-undersökningen 2001 gavs möjligheten att pröva innehållet i instrumentet mot elevernas prestationer i PIRLS provet för att ömsesidigt belysa validiteten. Diagnosinstrumentet utarbetades för de tidiga skolåren (årskurs 2-6) och fokuserar på elevens språkutveckling. Inför PIRLS 2001 anpassades instrumentet för att passa i en storskalig studie. Det innebar att lärarnas i PIRLS gjorde en bedömning på en 10-gradig skala istället för att formulera sina observationer i text.

De teoretiska utgångspunkterna i PIRLS 2001 och den svenska läroplanen går att sätta i relation till varandra, och för den aktuella studien kan det vara relevant att belysa överensstämmelsen dem emellan. Det diagnostiska material som Skolverket (2002) har utformat för uppföljning och bedömning av elevers läsförmåga till stöd för lärarnas läsundervisning, överensstämmer väl med de aspekter av läsförmåga som PIRLS prövat. I den svenska rapporten kommer forskarna fram till att syftena med PIRLS-undersökningen står väl i samklang med den svenska läroplanen (Rosén, m.fl., 2005). Vidare har Skolverket (2006) undersökt kursplanen i svenska och dess mål för elever i årskurs fem och jämfört dessa med PIRLS intentioner med sina uppgifter. Kursplanens mål är bland annat att elever i femte årskursen ska ”*kunna läsa med flyt både högt och tyst och uppfatta skeenden och budskap i böcker och saklitteratur skrivna för barn och ungdom, kunna samtala om läsningens upplevelser samt reflektera över texter*” (Skolverket, 2000, sid. 101). Slutsatsen som dras av Skolverket (2006) är att de kunskaper som prövas i PIRLS stämmer väl överens med kursplanens innehåll.

ÖVRIGA VARIABLER

Vidare användes som tidigare nämnts elevernas resultat på det standardiserade PIRLS-provet i analyserna. Uppgifterna i PIRLS-undersökningen är rigoröst utprövade av IEA och ett flertal åtgärder har vidtagits för att säkerställa datas kvalitet. Elevsvaren är dessutom rättade av externa bedömare och inte av lärarna i de deltagande klasserna. I den svenska rapporten från PIRLS 2001 (Rosén m.fl., 2005) står de väsentliga och övergripande delarna av kvalitetssäkringen att läsa samt i den motsvarande internationella tekniska rapporten (Martin, Mullis & Kennedy, 2003) finns en mer detaljerad beskrivning av studiens kvalitetsaspekter.

För att öka validiteten i ett prov kan det göras mer omfattande, genom att fler frågor ställs inom varje område. Bedömningen av en elevs läsförmåga blir då mer heltäckande. Problemet med ett sådant förfarande är att provet kommer att bli väldigt långt och att det är osäkert om det finns någon elev som skulle orka göra hela provet. Därför skulle resultaten riskera att inte bli

giltiga. I PIRLS och andra storskaliga mätningar har man löst detta med så kallad matrissampling. Detta innebär att eleverna tillsammans gör alla uppgifter men att varje enskild elev inte gör alla själv. PIRLS 2001 bestod av 10 häften med uppgifter och varje elev gjorde ett häfte. Några uppgifter är unika för varje häfte, medan några uppgifter är identiska med uppgifter i andra häften. Med tillräckligt stort urval fås på det här viset en giltig bedömning av läsförmågan samt god information om hur elever presterar relativt varandra. Alla elever får sedan en poäng på en gemensam skala, vilken räknas ut med hjälp av Item Response Theory (IRT) teknik. IRT tar både frågornas svårighetsgrad och elevens färdigheter i beaktande när en totalpoäng beräknas, som eleven skulle ha fått om hon gjort alla uppgifter i PIRLS. I och med den här proceduren förbättras tillförlitligheten i provet avsevärt.

För att undersöka faktorer som kan påverka lärarnas bedömningar, har information om elevernas socioekonomiska bakgrund (SES) och kön inhämtats. Information om variablerna presenteras i Tablå 1 på nästa sida.

När det gäller begreppet SES operationaliseras det i den här studien med hjälp av fem olika indikatorer på kulturellt och ekonomiskt kapital. Dessa var: antal böcker hemma, hushållets ekonomiska situation, hushållets årsinkomst, högsta utbildning i hemmet, samt typ av arbete. Liknande indikatorer har tidigare används med framgång i effektstudier (Sirin, 2005).

Tablå 1. Beskrivning av de kovariater som använts i analyserna

Variabel	Information/fråga	Källa
Läsprovsresultat	Elevernas resultat på PIRLS 2001 läsprov.	Elev
Kön	Elevens kön (Pojke=0, Flicka=1)	Elev
Antal böcker hemma	Ungefär hur många böcker finns hemma hos er? Alternativ (1-5): 0-10,11-25, 26-50, 51-100, fler än 100	Hem
Ekomomisk situation	Om man jämför er familj med andra, hur god är er familjs ekonomiska situation? Alternativ (1-5): Inte alls god, Inte så god, Medel, Ganska god, Väldigt god.	Hem
Årsinkomst	Inom vilken spännvidd ligger ert hushålls sammanlagda årsinkomst? Alternativ (1-6): Mindre än 180 000kr, 180 000-269 999kr, 270 000-359 999kr, 360 000-449 999kr, 450 000-539 999kr, mer än 540 000kr	Hem
Högsta utbildning	Högsta utbildningsnivå I hemmet. Alternativ (1-8): Ej fullgjord grundskola, Genomförd grundskolal, Två-årigt gymnasium, Tre-årigt gymnasium, KY-utbildning eller liknande, Minst två år universitetsstudier, Kandidatexamen, Masterexamen.	Hem
Typ av arbete	Kategorisering av hemmets huvudsakliga arbetssituation. Alternativ (1-3): Arbetare, Tjänsteman, Akademiker	Hem

Analysmetoder

Data i samhällsvetenskaplig forskning och speciellt inom utbildning är ofta av hierarkisk natur (Gustafsson, 2009; Hox, 2002). Det innebär att individer är klustrade inom ett klassrum (klassrumsnivå), att klassrummen är klustrade inom en skola (skolnivå) och så vidare. Individerna inom ett kluster tenderar att vara mer lika varandra än individer i andra kluster. Elever på en skola delar liknande erfarenheter, i det aktuella fallet också lärare och kamrater. Dessa omständigheter, som beror på hur urvalet är gjort, måste hanteras statistiskt för att tolkningar och slutsatser baserade på analyserna ska bli korrekta. Många statistiska test tillämpar ett antagande om oberoende mellan de observationer som analyserna baseras på, och bryts detta antagande kommer standardfelen att bli för små, vilket kan leda till signifikanta T-värden (Hox, 2002). Regressionskoefficienterna kan då tolkas felaktigt och orimliga slutsatser kan dras.

Flernivåmodellering hanterar problemet med beroenden inom nivåer. Genom att dela upp variansen i komponenter inom och mellan grupper går det att separera ut den variation som går att härleda till individuella respektive gruppskillnader. Den så kallade intraklass-korrelationen ger signaler om flernivåmodellering bör användas eller inte; den ger information om den andel av variationen som kan härledas till skillnader mellan grupper. Exempelvis varierar ofta kunskaper både mellan eleverna i en klass, och mellan olika klassrum eller skolor. I ett likvärdigt skolsystem, är klasser/skolors genomsnittliga prestationer relativt lika medan prestationer inom ett klassrum kan variera i högre grad. Om klassernas resultat varierar kraftigt kommer intraklasskorrelationen att vara hög, vilket indikerar heterogena prestationer bland klassrummen.

Strukturell ekvationsmodellering (SEM) var den huvudsakliga analysmetoden i studien, då den har flera viktiga fördelar i jämförelse med exempelvis multipel regressionsanalys. En sådan fördel är möjligheten att använda latenta variabler. Dessa variabler är inte direkt observerbara, vilket är fallet med manifesta variabler eller indikatorer. En latent variabel bygger på samvariationen mellan flera olika indikatorer och gör det möjligt att bättre operationalisera ett begrepp än vad som vore möjligt med endast enstaka manifesta variabler. Exempel på begrepp som bäst mäts med en latent variabel kan vara lycka, motivation och självförtroende. En annan fördel med latenta variabler är att de kan sägas vara fria från mätfel. Indikatorer är alltid behäftade med ett visst mått av mätfel, dvs. felvarians, men genom att konstruera en latent variabel sorteras mätfelen ut till en så kallad residual. Det som är kvar i den latenta variabeln är då ”sann” varians. Analyserna gjordes med hjälp av programmet Mplus 7.11 (Muthén & Muthén, 2007-2013) som användes i programmet STREAMS analysmiljö (Gustafsson & Stahl, 2005).

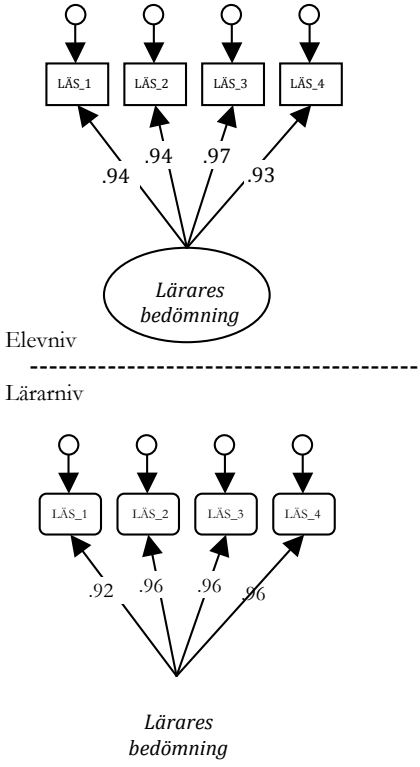
RESULTAT

Som ett första steg prövades validiteten i lärarnas bedömningar genom att undersöka relationen mellan lärarnas bedömningar och elevers provresultat i årskurs 3. Validiteten belystes dels genom att undersöka om lärare rankade elevernas kunskaper inom den egna klassen på liknande sätt som PIRLS-provet gjorde, dels genom att studera om lärare var samstämmiga i sina bedömningar. För att kunna studera dessa relationer användes tvånivåmodellering, där resultat för respektive nivå (individ- och klassrumsnivå) erhålls.

Eftersom indikatorerna som bildade den latenta variabeln *"lärares bedömning"* bestod av både bedömningar av läs- och skrivförmåga, gjordes först en tvånivåmodell med två faktorer: en för läsbedömning och en för skrivbedömning. Emellertid passade denna modell data dåligt och indikerade att *läs- och skrivfaktorerna* var högt korrelerade på både individ- och gruppnivå (r= .96). Således verkade det svårt att separera bedömningen av läsning och skrivning rent empiriskt, eftersom läraren tycktes göra en generell bedömning av den språkliga förmågan hos individuella elever. Däremot visade sig en modell med en generell bedömningsfaktor "*Språk*" ge en bättre modellanpassning, dock inte tillräckligt bra för att modellen skulle kunna accepteras. Därför gjordes fyra paketsummor av de tolv bedömningsaspekterna, vilket innebar fyra nya indikatorer, så kallade *Läs* 1-4 (se Figur 1). De fyra paketen består vart och ett av tre aspekter. Istället för en latent variabel konstruerad av tolv indikatorer med en skala från 1-10, består det slutgiltiga latenta variabeln av fyra indikatorer med skalan 1-30. Eftersom det inte förelåg någon multidimensionalitet i de tolv aspekterna tilldelades varje paket tre av dem helt slumpmässigt. Genom detta förfarande ökar indikatorernas varians och mätegenskaperna förbättras generellt (Little, Cunningham, Shahar, & Widaman, 2002). En utförligare diskussion gällande paketeringsförfarandet av dessa bedömningsaspekter finns publicerat i Johansson, Myrberg och Rosén (2012). Den slutgiltiga modellen för lärarbedömningarna presenteras nedan.

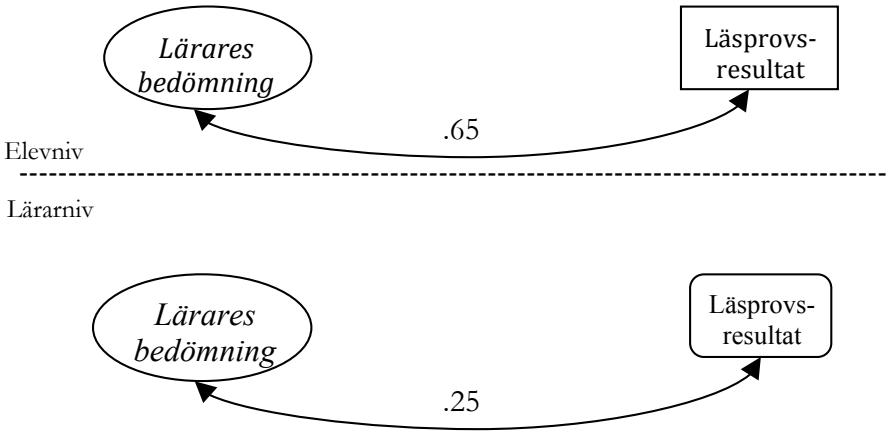
Modellen presenterad i Figur 1 passade data väl, med jämna och höga faktorladdningar. Ett första steg var att analysera variationen i lärarbedömningarna, inom och mellan klassrum. Ungefär 70 % av variationen i bedömningarna kunde härledas till variation inom klasser medan ungefär 30 % utgjordes av variation mellan klasser. Med andra ord varierade elevernas resultat mer än vad medelresultaten för klassrum gjorde. Genom att pröva sambandet mellan lärarnas bedömning och elevernas provresultat i PIRLS belystes validiteten i båda dessa mått på kunskaper och färdigheter. Som visas i Figur 2, var korrelationen inom klasser 0.65, medan den var endast 0.25 mellan klassrum. Dessa resultat innebär i praktiken att lärares bedömningar följer elevernas prestationer på PIRLS-testet relativt väl inom klassrum men

att olika lärares bedömningar varierar kraftigt eftersom korrelationen med klassernas medelresultat var låg.



Figur 1. Mätmodeller för lärares bedömning. Faktorladdningar för paketsummorna Läs1-4 visas.

I undersökningens andra del relaterades elevens socioekonomiska bakgrund (SES) och kön till lärares bedömningar och elevers resultat, främst för att undersöka om dessa elevkaraktäristika var något som lärare tog med i sin bedömning av elevens läs- och skrivkunskaper. Vad som kan noteras på individnivå är att både kön och SES har en liten effekt på lärares bedömning. Effekten består i att flickor och elever från högre socioekonomisk bakgrund får en något högre bedömning av lärarna, givet att deras prestationsnivå på provet är under statistisk kontroll, det vill säga samma som för pojkar och elever från lägre socioekonomisk bakgrund. Dessutom framgår det av analyserna att både SES och kön är positivt relaterade till läsprovsresultaten i PIRLS, då flickor och elever med högre SES har högre resultat på provet.



Figur 2. Sambandet mellan lärares bedömningar och elever PIRLS resultat, inom och mellan klasser.

När det gäller samband på klassnivå visar det sig att SES har en mycket stark påverkan på klassens resultatnivå. Klassrum med ett högt genomsnittligt SES visar betydligt högre prestationer. Klassers SES har också en relativt stark påverkan på lärares bedömningar, samtidigt som man kan notera att sambandet mellan klassens provresultat och lärares bedömningar inte längre är signifikant när SES introduceras i modellen. Det ser sålunda ut som att relationen mellan provresultat och lärares bedömningar helt förklaras av SES. Vid tidpunkten då lärarna skattade elevernas kunskaper och färdigheter kände de inte till den egna klassens, eller andra klassers testresultat, medan de hade viss information om elevernas SES och denna variabel är högt korrelerad med testresultat. Klassrummens SES ska ses som ett uttryck för den demografiska kontext som skolan befinner sig i, det vill säga karaktäristika som delvis bestäms av skolans upptagningsområde. Analyserna visar att stor variation i skolors resultat kan hänföras till SES.

De viktigaste resultaten av den här studien kan sammanfattas i följande punkter:

- Lärarna bedömer elevernas kunskaper i den egna klassen i samklang med de externa bedömningarna på PIRLS-provet
- Lärare har problem att samstämmigt bedöma elevers läsförmåga; referensramarna skiljer sig åt
- Lärarbedömningarna innehåller mer än bara faktiska kunskaper – elevkaraktäristika associerade med socioekonomisk bakgrund och kön bidrar till variation i bedömningen

DISKUSSION OCH SLUTSATSER

Det huvudsakliga syftet i den här studien har varit att undersöka validiteten i lärares bedömningar av elevers läs- och skrivkunskaper i årskurs 3. När förhållandet mellan lärares bedömningar respektive elevernas provresultat i PIRLS 2001 studerades, visade resultaten att båda instrumenten kan vara lämpliga mått på läsförmåga inom klassrum. Lärarnas bedömningar följde elevernas provresultat i PIRLS relativt väl. Resultaten inom klasser är således i linje med de resultat som Südkamp, Kaiser och Möller (2012) lägger fram i sin metaanalys av lärarbedömningars relation till andra bedömningsformer. Människor tenderar att förankra sina bedömningar i ett initialt värde, eller referensvärde (se t.ex., Tversky & Kahneman, 1974). Genom att förankra sin bedömning i någon elevs prestation, till exempel den högst presterande eleven, distribueras de övriga elevernas prestationer utifrån dennes. Däremot visade analyserna att elevens kön och socioekonomiska bakgrund kan påverka lärarens bedömning. Med hänsyn till rättviseaspekter kan inte dessa elevkaraktäristika förbises. Trots att provresultaten i PIRLS hölls lika för pojkar och flickor, blev flickor bedömda högre av lärarna. Detsamma gällde för elever från högre socioekonomisk bakgrund. Kön och socioekonomisk bakgrund kan vara kopplade till icke-kognitiva faktorer, som ansträngning och motivation, vilket lärare skulle kunna premiera. Även om dessa faktorer är viktiga för lärandet, utgör de inga relevanta faktorer för kunskapsbedömning, så som styrdokumenten är framskrivna i svensk skolkontext. Karami (2013) bland andra menar att bedömningar betraktas som rättvisa om de är fria från ”bias”, det vill säga att irrelevant varians, med Messicks (1989) termer (*construct irrelevant variance*), inte förs in i bedömningen. Begreppet rättvisa i bedömningar är tätt sammankopplade med validitet (se t.ex., Kane, 2010; Karami, 2013). Även om somliga resultat i den aktuella studien kan tolkas som att lärare orättvist fördelar sina bedömningar, kan det finnas rimliga förklaringar. Exempelvis kan flickor ha gjort bättre ifrån sig på uppgifter som inte omfattas i PIRLS läsprov, som t.ex. muntliga, och mer omfattande skriftliga uppgifter. Deras högre bedömda läsförmåga skulle därför kunna tillskrivas kunskaper som inte mäts i provet men som lärare tar med i sin bedömning.

Trots att lärare relativt väl kan skilja på elevernas kunskaper i den egna klassen, har många lärare olika referensramar vid bedömning. Detta innebär att elever med ungefär samma kunskaper kan få olika bedömning, beroende på vilken lärare de har. Även om lärare väl kan identifiera elevernas prestationer i den egna klassen, finns svårigheter att bedöma den genomsnittliga kunskapsnivån relativt andra lärares elever. Detta mönster har visat sig tidigare, exempelvis med de normrelaterade betygen. Både anekdotisk och vetenskaplig evidens har visat att medelvärdet för de normrelaterade betygen kunde vara detsamma i två klassrum trots att den faktiska kunskapsnivån

skiljt sig markant mellan olika klassrum (Jarl & Rönnberg, 2010). Precis som då, har lärarna i den här studien utgått från kunskaperna i sin egen klass.

Socioekonomisk bakgrund visade sig förklara en hög andel varians i PIRLS-provet (R2=0.64) på klassrumsnivå. Detta innebär att de stora resultatskillnaderna mellan klassrum nästan helt kan förklaras av elevernas socioekonomiska bakgrund. Eftersom data i studien är inhämtade år 2001, finns det anledning att tro att det har skett vissa förändringar i samhället, både vad det gäller elevers kunskaper och färdigheter samt skolsegregation. Läskunskaperna bland 9-10-åringar har sjunkit sedan 1970 i Sverige, och framförallt har andelen mycket goda läsare sjunkit under de senaste decennierna (Rosén, 2012). Mycket tyder också på att segregationen med avseende på elevers socioekonomiska bakgrund har accentuerats det senaste decenniet. Vidare visade resultaten att det andra måttet på elevers läskunskaper – lärares bedömningar – inte hade lika högt samband med socioekonomisk bakgrund. Anledningen till att lärarbedömningarna korrelerade lägre med SES än vad provresultaten i PIRLS gjorde kan bero på takeffekter i variabeln som mätte lärares bedömning. Takeffekter kan bidra till att resultatskillnader som faktiskt existerar mellan klassrum med olika socioekonomisk bakgrund inte går att uppfatta. Till exempel kan Lärare A och B bedöma sina respektive klasser genomsnittligen med nio på den tio-gradiga skalan trots att medelresultaten på PIRLS-provet skiljer sig avsevärt mellan klassrummen. Klasserna blir då bedömda av lärarna som att de har likvärdiga kunskaper, eftersom inget ytterligare utrymme på skalan existerar. I det här fallet skulle den bristande samvariationen med SES bero på takeffekter i lärarnas bedömning.

Resultaten i den här studien visade att lärares bedömningar är problematiska vid summativa bedömningar som utgör basen för individuella utlåtanden och betyg. Eftersom bedömningen ligger till grund för urval till exempelvis högre utbildning, är det viktigt att den sker på ett likvärdigt sätt mellan klassrum och skolor i landet. Summativa bedömningar kan emellertid också ligga till grund för insatser som görs för att främja elevers lärande. Om formativ bedömning varierar stort i landet innebär detta att villkoren för lärande inte blir likvärdiga för eleverna. Undervisningen behöver inte ske på samma sätt i olika skolor men Utbildningsdepartementet (Skollagen, 2010:800) föreskriver att utbildningen i grundskolan ska vara likvärdig och att hänsyn därmed ska tas till elevernas olika förutsättningar och behov. Om bedömningen inte är likvärdig kan det få konsekvensen att en del skolor ger stöd och feedback till elever som behöver det, medan andra skolor inte gör det.

Huvudslutsatsen i denna artikel är att bedömningens likvärdighetsproblematik bygger på ett systemfel och att lärarna måste få adekvata verktyg

för att kunna kalibrera sina bedömningar mellan såväl olika klassrum i samma skola som mellan skolor. Ett sådant verktyg kan vara ett mer standardiserat kunskapsprov, som kompletterar de redan existerande nationella proven, i linje med de resonemang som Gustafsson, Cliffordson och Erickson (2014) för i en nyligen publicerad rapport. Ett sådant kunskapsprov skulle också underlätta för nationell uppföljning och utvärdering eftersom betygen eller de nuvarande nationella proven inte ger en god bild över kunskapstrenden i Sverige. Sambedömning, där lärare från olika skolor diskuterar bedömning för att nå konsensus kan också vara en framkomlig väg för ökad likvärdighet (Klenowski & Wyatt-Smith, 2010), något som Skolverket också poängterar som en styrka. Dock visar forskning om sambedömning att lärare snarare når samsyn kring mål och kriterier och får en ökad bedömarkompetens, än att lärare ger samma poäng eller betyg för samma elevprestation (Thornberg, 2014). Slutligen kan det konstateras att möjligheten att använda PIRLS som en form av valideringsinstrument har varit gynnsamt. Vidare forskning kring möjligheter att arbeta med externa prov i lärarprofessionens vardag har potential att verka för mer likvärdiga lärarbedömningar.

REFERENSER

Andreasson, I. (2007). *Elevplanen som text - om identitet, genus, makt och styrning i skolans elevdokumentation*. (Doktorsavhandling). Göteborg: Göteborgs Universitet.

Andreasson, I. (2012). Språk och elevidentiteter i skolans elevdokumentation. *Forskning om undervisning & lärande.* Stiftelsen SAF.

Balan, A. (2012). *Assessment for learning: a case study in mathematics education*. (Doktorsavhandling). Malmö: Malmö Högskola.

Borsboom, D., Cramer, A.O.J., Kievit, R.A., Zand Scholten, A. &Franić, S. (2009). The end of construct validity. In R. Lissnitz (Ed.). *The Concept of Validity. Revisions, New Directions and Applications*. (pp. 135-170). Charlotte, NC: IAP.

Brookhart, S. M. (2012). The use of teacher judgement for summative assessment in the USA. *Assessment in Education: Principles, Policy & Practice*, *20*, 69-90. doi:10.1080/0969594x.2012.703170

Cizek, G. J., Fitzgerald, S. M., & Rachor, R. E. (1996). Teachers' Assessment Practices: Preparation, Isolation, and the Kitchen Sink. *Educational Assessment, 3*(2), 159-179.

Gipps, C. (1994). *Beyond Testing. Towards a theory of educational assessment*. London: The Falmer Press.

Cliffordson, C. (2004). De målrelaterade gymnasiebetygens prognosförmåga. *Pedagogisk Forskning i Sverige; 9*(2), 129-140.

Coladarci, T. (1986). Accuracy of Teacher Judgments of Student Responses to Standardized Test Items. *Journal of Educational Psychology, 78*(2), 141-146.

Cronbach, L. J. (1971). Test Validation. In Robert L. Thorndike (Ed.), *Educational Measurement* (Second edition, pp. 443-507). Washington, D.C: American Council on Education.

Gonzalez, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 User Guide for the International Database.* Chestnut hill, MA: Boston College.

Guilford, J. (1946). New standards for test evaluation. *Educational and Psychological Measurement*, *6*(5), 427-438.

Gustafsson, J-E. (2009). Strukturell Ekvationsmodellering. I G. Djurfeldt & M. Barmark (Red.), *Statistisk verktygslåda 2 – Multivariat analys* (s. 269-321). Lund: Studentlitteratur.

Gustafsson, J-E., & Stahl, P. A. (2005). *STREAMS User's Guide, Version 3.0 for Windows 95/98/NT*. Mölndal: MultivariateWare.

Gustafsson, J-E., & Yang-Hansen, K. (2009). Resultatförändringar i Svensk grundskola. I Skolverket (Red.), (sid. 40-83). Vad påverkar resultaten i svensk grundskola? Kunskapsöversikt om betydelsen av olika faktorer. Stockholm: Skolverket.

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

Harlen, W. (2005). Trusting Teachers' Judgement: Research Evidence of the Reliability and Validity of Teachers' Assessment Used for Summative Purposes. *Research Papers in Education, 20*, 245-270.

Harlen, W. (2011). On the Relationship between Assessment for Formative and Summative Purposes. I J. Gardner (Ed.), *Assessment and learning*. (Second edition; pp. 61-80). Gateshead, UK: Sage.

Hirsh, Å. (2013). *The individual development plan as a tool and practice in Swedish compulsory school*. (Doktorsavhandling). Jönköping: Högskolan i Jönköping.

Hoge, R. D., & Coladarci, T. (1989). Teacher-Based Judgments of Academic Achievement: A Review of Literature. *Review of Educational Research, 59*(3), 297-313.

Hox, J. (2002). *Multilevel Analysis - Techniques and Applications*. Mawah, NJ: Lawrence.

Jarl, M., & Rönnberg, L. (2010). *Skolpolitik. Från riksdagshus till klassrum.* Malmö: Liber.

Johansson, S., (2013). *On the Validity of Reading Assessments: Relationships Between Teacher Judgements, External Tests and Pupil Self-assessments*. (Doktorsavhandling). Göteborg: Göteborgs Universitet.

Johansson, S., Myrberg, E., & Rosén, M. (2012). Teachers and tests: assessing pupils' reading achievement in primary schools. *Educational Research and Evaluation, 18*(8), 693-711. doi: 10.1080/13803611.2012.718491

Jönsson, A. (2008). *Educative assessment for/of teacher competency: a study of assessment and learning in the "interactive examination" for student teachers*. (Doktorsavhandling). Malmö: Malmö Högskola.

Kane, M. (2006). Validation. In R. Brennan (Ed.), *Educational measurement* (4th ed.). Washington, D.C: American Council on Education and National Council on Measurement in Education.

Kane, M. (2010). Validity and fairness. *Language Testing*, 27: 177–182, doi:10.1177/0265532209349467

Karami, H. (2013). The quest for fairness in language testing. *Educational Research and Evaluation: An International Journal on Theory and Practice, 19*(2-3), 158-169, doi:10.1080/13803611.2013.767618

Klapp Lekholm, A. (2008). *Grades and grade assignment: effects of student and school characteristics*. (Doktorsavhandling). Göteborg: Göteborgs Universitet. Hämtad från http://hdl.handle.net/2077/18673

Klenowski, V., & Wyatt-Smith, C. (2010). Standards-Driven Reform Years 1-10: Moderation an Optional Extra? *Australian Educational Researcher, 37*(2), 21-39.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To Parcel or Not To Parcel: Exploring the Question, Weighing the Merits. *Structural Equation Modeling, 9*, 151-173.

Llosa, L. (2007). Validating a Standards-Based Classroom Assessment of English Proficiency: A Multitrait-Multimethod Approach. *Language Testing, 24*, 489-515.

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

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2003). *PIRLS 2001 Technical Report*. Chestnut Hill, MA: Boston College.

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

Mehrens, W A. (1997). The Consequences of Consequential Validity. *Educational Measurement: Issues and Practice 16*, 16-18.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (Third edition, pp. 13-103). New York: American Council on Education/Macmillian.

Muthén, L. K., & Muthén, B. O. (2007-2012). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Nyström, P. (2004). *Rätt mätt på prov.* (Doktorsavhandling). Umeå: Umeå Universitet.

Rosén, M., Myrberg, E., & Gustafsson, J.-E. (2005). *Läskompetens i skolår 3 och 4. Nationell rapport från PIRLS 2001 i Sverige*. Göteborg: Göteborgs universitet.

Rosén, M. (2012). Förändringar i läsvanor och läsförmåga bland 9- till 10-åringar. Resultat från internationella studier. I SOU, 2012, rapport nr (Ed.), *Läsarnas marknad, marknadens läsare – en forskningsantologi.* Stockholm: Utbildningsdepartementet.

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

Sirin, S. R. (2005). Socioeconomic Status and Academic Achievement: A Meta-Analytic Review of Research. *Review of Educational Research, 75*, 417-453.

SFS 2010:800. 2010 års skollag. Stockholm: Utbildningsdepartementet.

Skolverket (2000). *Kursplaner för grundskolan*. Stockholm: Skolverket.

Skolverket (2002). *Språket lyfter! Diagnosmaterial i svenska och svenska som andra språk för åren före skolår 6.* Stockholm, Sverige: Skolverket.

Skolverket (2006). *Med fokus på läsförståelse. En analys av skillnader och likheter mellan internationella jämförande studier och nationella kursplaner.* Stockholm: Skolverket.

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

Skolverket (2012). Redovisning av uppdrag om avvikelser mellan provresultat och betyg i grundskolan årskurs 9. Stockholm: Skolverket.

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*, 743-762.

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). Göteborg: Göteborgs Universitet. Hämtad från http://hdl.handle.net/2077/16892

Thornberg, P. (2014). *Kan sambedömning leda till en mer likvärdig bedömning och betygsättning*? (Mastersuppsats). Kristianstad: Högskolan Kristianstad. Hämtad från http://www.diva-portal.org/smash/get/diva2:704983/FULLTEXT01.pdf

Tversky, A., & Kahneman, D. (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *4157*, 1124-1131.

Westlund, B. (2013). *Att bedöma elevers läsförståelse: En jämförelse mellan svenska och kanadensiska bedömningsdiskurser i grundskolans mellanår.* (Doktorsavhandling) Stockholm: Stockholms Universitet. Hämtad från: http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-95403.

Wolming, S. (1998). Validitet - Ett traditionellt begrepp i modern tillämpning. *Pedagogisk Forskning i Sverige, 3*(2), 81-103.