Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning

Del 2 – Bilagor

Betänkande av Utredningen om nationella prov

Stockholm 2016

STATENS OFFENTLIGA UTREDNINGAR

SOU 2016:25

SOU och Ds kan köpas från Wolters Kluwers kundservice.
Beställningsadress: Wolters Kluwers kundservice, 106 47 Stockholm
Ordertelefon: 08-598 191 90
E-post: kundservice@wolterskluwer.se
Webbplats: wolterskluwer.se/offentligapublikationer

För remissutsändningar av SOU och Ds svarar Wolters Kluwer Sverige AB
på uppdrag av Regeringskansliets förvaltningsavdelning.
Svara på remiss – hur och varför
Statsrådsberedningen, SB PM 2003:2 (reviderad 2009-05-02).
En kort handledning för dem som ska svara på remiss.
Häftet är gratis och kan laddas ner som pdf från eller beställas på regeringen.se/remisser

Layout: Kommittéservice, Regeringskansliet
Omslag: Elanders Sverige AB
Tryck: Elanders Sverige AB, Stockholm 2016

ISBN 978-91-38-24427-2
ISSN 0375-250X

Inledning

Det här är del 2 till Utredningen om nationella provs betänkande *Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning* (SOU 2016:25). Del 2 innehåller bilagorna 2–5. Bilaga 1, som innehåller direktiven till utredningen, ingår i del 1 till betänkandet.

I bilaga 2 – *Provbetyg, lärarbetyg och deras inbördes relationer* – behandlas relationen mellan provbetyg och ämnes- eller kursbetyg.

Bilaga 3 – *Provsystem i förändring* – innehåller en historisk skildring av hur provsystemet har förändrats över tid från 1940-talet till i dag.

I bilaga 4 – *Provs mätfel* – ges en beskrivning av provs mätfel och av olika testteorier för att uppskatta mätfelen.

Bilaga 5 – *Provbetygens stabilitet och tillförlitlighet i gymnasieskolan* – handlar om de nationella provens stabilitet i gymnasieskolan.

Innehåll

Provbetyg, lärarbetyg och deras inbördes relationer

Sammanfattning

Examinerande prov övergavs i Sverige under 1950- och 1960-talen till förmån för prov som mer skulle fungera som stöd för lärarnas betygssättning. Övergången byggde på ökade insikter om att prov är osäkra instrument för att mäta enskilda individers samlade kunskaper. Däremot har de bättre mätsäkerhet när det gäller gruppers resultat. Detta ledde till en modell där provresultaten styrde den genomsnittliga betygsnivån i en klass även om läraren hade till uppgift att fördela betygen.

Modellen tillämpades, i varje fall på gymnasienivå, tämligen strikt under de relativa betygens tid och fram till 1990-talets reformer. Relationen mellan provbetyg och lärarbetyg var under denna tid förhållandevis klart reglerad. I och med införandet av det mål- och kunskapsrelaterade betygssystemet avskaffades den mer strikt reglerade relationen mellan provbetyg och lärarbetyg. Grunden för betygssättningen blev mer än tidigare en fråga om tolkning av verbalt formulerade betygskriterier än en fråga om procentuell betygsfördelning på prov. Proven tappade därmed sin normerande roll.

Samtidigt har dock betyg även efter 1994 stor betydelse när det gäller urval till gymnasieskolan och den högre utbildningen. Frågor om rättvis och likvärdig betygssättning har därför hög prioritet. Detta har lett till att betygsstödjande nationella prov har fortlevt även i det mål- och kunskapsrelaterade systemet, men med en annan och mer oreglerad relation till betygen.

I denna bilaga ges en samlad beskrivning av relationen mellan provbetyg och lärarbetyg ur lite olika perspektiv, bl.a. vad gäller betygens utveckling över tid, betygens fördelning och nivåer i olika

ämnen, hur provbetyg och lärarbetyg förhåller sig för olika grupper (främst skolenheter och klasser) samt förändringar över tid i olika avseenden.

Slumpens inflytande diskuteras sällan när det gäller prov och betyg. I denna bilaga redovisas resultat som visar att slumpen genom de mätfel den genererar har en avsevärd betydelse för i synnerhet små elevgrupper och för enstaka elever (även om resultat på elevnivå inte står i fokus i den här bilagan).

De viktigaste empiriska resultaten är att betygsnivån för olika ämnen varierar i hög grad. Genomgående är betygsnivån som högst i ämnet engelska för såväl provbetyg som lärarbetyg. Ämnet engelska har också minst systematisk skillnad mellan medelvärdena för lärarbetyg och provbetyg på nationell nivå. Ämnet svenska har lägre betygsnivåer och större skillnad mellan lärarbetyg och provbetyg. Klart lägsta betygsnivå och största skillnad mellan lärarbetyg och provbetyg har de olika proven i matematik. Dessa skillnader har funnits och varit relativt oförändrade på nationell nivå sedan de nationella proven började användas i slutet av 1990-talet och fram till i dag.

På skolenhetsnivå och klassnivå har det under hela perioden funnits stora variationer i avvikelse (skillnad mellan lärarbetyg och provbetyg) mellan olika skolenheter och olika klasser. Även de här mönstren har varit mycket beständiga över tid. Oberoende av ämne och betygsnivå för ämnet har betygsvariationen (standardavvikelsen) för såväl provbetyg som lärarbetyg, och skillnaden mellan dem, varit påtagligt konstant över tid. Ett tydligt mönster är dock att de elever som har fått betyget F (tidigare IG) på provet är den grupp som främst har fått ett högre betyg av läraren.

Ur ett rättvise- och likvärdighetsperspektiv är det inte försvarbart att avvikelserna är så olika för olika skolenheter och olika klasser. Trots insatser från Statens skolverk har någon nämnvärd förändring inte skett sedan de nationella proven infördes på 1990-talet. Utredning anser därför att någon form av modell behöver utvecklas för att tydliggöra relationen mellan provbetyg och lärarbetyg på gruppnivå. I bilagan presenteras och diskuteras en tänkbar sådan modell.

Inledning

En del i utredningens uppdrag handlar om att analysera om resultaten på nationella prov i förändrad grad ska vara styrande för betygssättningen eller om kopplingen mellan provresultat och betygssättning på annat sätt bör tydliggöras eller förändras. Det kan t.ex. göras genom att ange hur mycket medelvärdet av lärarens betyg (i denna bilaga kallat lärarbetyg) för en grupp får skilja sig från betygsmedelvärdet på provet, dvs. en modell liknande den som användes för de centrala proven på klassnivå i det grupprelaterade betygssystemet. Klassbegreppet är dock lite mer oklart i dag än det var för några decennier sedan då de centrala proven gavs, så kanske en jämförelse på skolenhets- eller huvudmannanivå kan vara mer lämplig i dag om modellen bedöms användbar. Men det är en senare fråga och för att kunna besvara den behövs först en mer ingående granskning av de resultat som finns tillgängliga om prov och betyg i det nuvarande mål- och kunskapsrelaterade systemet:

- Hur ser relationen mellan provbetyg och lärarbetyg ut?
- Är den likartad för olika ämnen och kurser?
- Kan en generell modell användas för alla prov och ämnen, eller måste varje kurs eller ämne ha sin speciella utformning?

Bakgrund

Resultat på de nationella proven kan presenteras på olika sätt för att betona olika perspektiv och aspekter. I den här bilagan är syftet att dels närmare granska och diskutera de sätt som för närvarande tillämpas, dels ge exempel på alternativa presentationsformer. Dessutom redovisas ett par alternativa sätt att analysera de befintliga resultaten avseende relationen mellan provresultat och betyg.

Med de många olika nationella prov som finns i grundskolan och gymnasieskolan är det inte möjligt att redovisa fullständiga resultat för alla olika prov utan vi har valt ut några som redovisas i denna bilaga medan andra läggs i en appendixdel till bilagan. Vissa prov redovisas inte alls – antingen för att de görs av så små och

varierande grupper[1] eller för att de är så nya att provkonstruktionen fortfarande mer eller mindre kan ses som en försöksverksamhet[2]. Det innebär att redovisningen kommer att fokusera på de sedan länge etablerade provämnena svenska[3], engelska och matematik.

För grundskolans del görs en fullständig redovisning för ämnet svenska i årskurs 9, medan motsvarande redogörelser för engelska och matematik är mer kortfattade. För gymnasieskolans del görs redovisningar för några olika kurser i matematik, svenska och engelska.

Resultat redovisas delvis på nationell nivå men i huvudsak på skolenhetsnivå och i vissa fall på gruppnivå, eftersom det är där betygssättningen görs. I vissa fall är dock grupperna så små att det är tveksamt om det är meningsfullt att bryta ner resultaten så långt som till gruppnivå. Med små grupper tenderar den statistiska osäkerheten att bli stor. Därför kan det vara svårt att dra slutsatser. Samma resonemang gäller i viss mån även för skolenheter som också kan vara små. Den statistik som används är insamlad av Statistiska centralbyrån (SCB) och är densamma som den Skolverket använder i sina redovisningar.

Skolverket redovisar prov- och betygsresultat antingen som frekvenser eller andelar elever med olika betyg[4], eller i form av så kallad genomsnittlig betygspoäng (GBP). GBP kan ses som ett viktat betygsmedelvärde där viktningen baseras på de föreskrivna betygspoäng som gäller vid beräkning av meritvärden (se tabell 1). Denna betygspoäng ligger också till grund för beräkningen av en elevs meritvärde.

Tabell 1 Gällande betyg och betygspoäng.

Betyg	F	E	D	C	B	A
Betygspoäng	0	10	12,5	15	17,5	20

[1] Gäller främst de prov för gymnasieskolans kurser som genomförs under höstterminen.
[2] Gäller prov i årskurs 6, eftersom betygssättningen där är tämligen ny. Proven i SO- och NO- ämnena har endast några år på nacken och har dessutom introducerats samtidigt som den nuvarande betygsskalan. Inte heller dessa ingår.
[3] Samma prov används för svenska som andraspråk, men inte heller detta ämne ingår i den här redovisningen.
[4] http://www.skolverket.se/statistik-och-utvardering/statistik-i-tabeller/grundskola/betyg-ak-9/betyg-och-provresultat-i-grundskolan-lasar-1997-98-1.26367
http://siris.skolverket.se/siris/f?p=SIRIS:7:0::NO:::

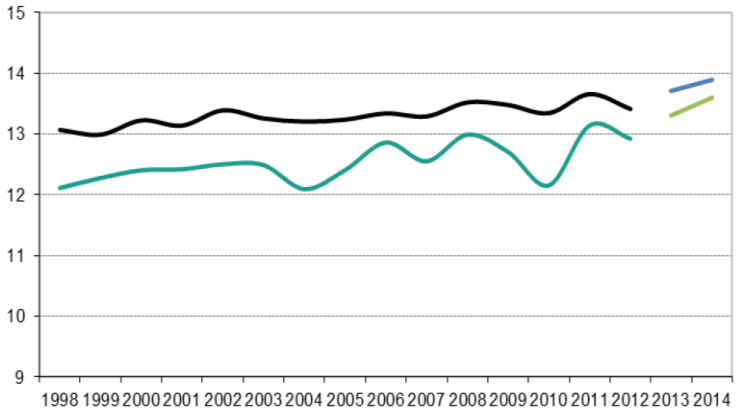
Det är värt att notera den sneda fördelningen i betygsskalan. Den har viss betydelse då den används i beräkningar, t.ex. av genomsnittlig betygspoäng, eftersom skalsteget mellan F och E ger fyra gånger så många betygspoäng som övriga skalsteg. Motivet för utformningen är att betyget E, som innebär godkänt resultat, ska markeras särskilt tydligt.

Grundskolans årskurs 9

Betygsskillnader varierar över tid och mellan ämnen

Provresultat och relationen mellan provbetyg och lärarbetyg redovisas på lite olika sätt av Skolverket. Ett vanligt sätt är diagram av det slag som visas i figur 1.[5] Diagrammen visar genomsnittlig betygspoäng (GBP) för provbetyg (Pbet) och lärarbetyg (Lbet) under den tid det funnits nationella prov inom ramen för ett mål- och kunskapsrelaterat betygssystem. Nedanstående figur visar utvecklingen för ämnet svenska.

Figur 1 **Genomsnittlig betygspoäng för provbetyg (turkos linje) och lärarbetyg (svart linje). Åren 1998–2012 gällde 1994 års fyrgradiga betygsskala och åren 2013–2014 gällde nuvarande sexgradiga skala (där ljusgrön linje anger provbetyg och blå lärarbetyg). Svenska.**

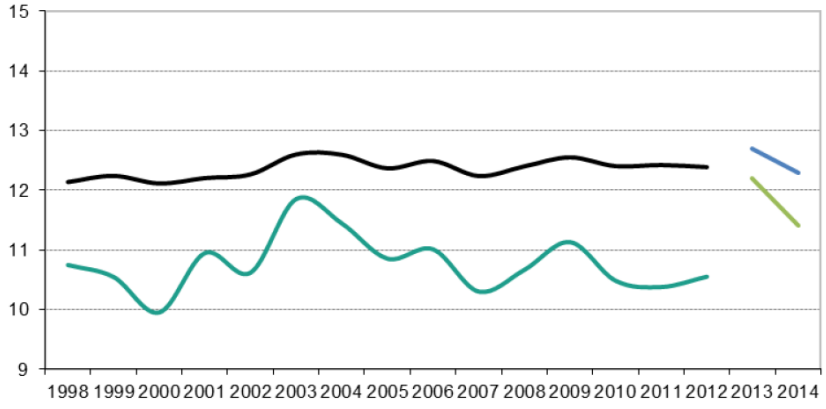


[5] Ur Skolverket (2015a).

Värdena på den vertikala axeln redovisas i genomsnittlig betygspoäng[6]. Värdena gäller endast de elever som har både provbetyg och lärarbetyg.[7]

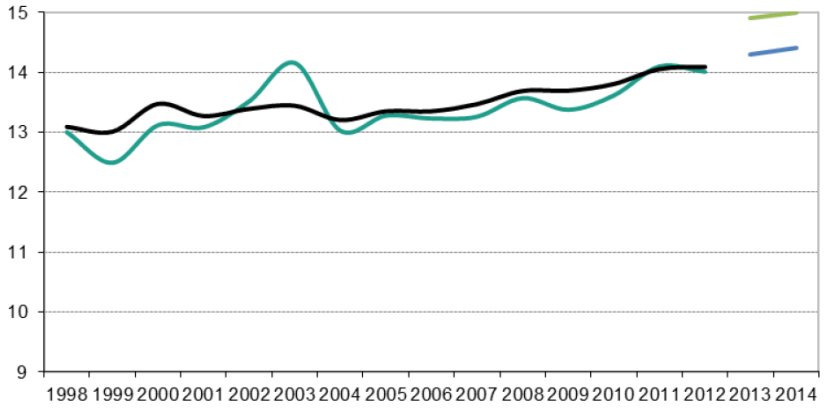
För matematik respektive engelska gäller nedanstående mönster (figur 2 och 3).

Figur 2 **Genomsnittlig betygspoäng för provbetyg (turkos linje) och lärarbetyg (svart linje). Åren 1998–2012 gällde 1994 års fyrgradiga betygsskala och åren 2013–2014 gällde nuvarande sexgradiga skala (där ljusgrön linje anger provbetyg och blå lärarbetyg). Matematik.**



[6] Den genomsnittliga betygspoängen beräknas enligt GBP=(0*P(F)+10*P(E)+12,5*P(D)+15*P(C)+17,5*P(B)+20*P(A))/100 där P(F) – P(A) är andelen elever med respektive betyg (i procent). GBP kan anta värden mellan 0 och 20.
[7] För cirka 10 procent av eleverna uppfylls vanligen inte detta krav.

Figur 3 **Genomsnittlig betygspoäng för provbetyg (turkos linje) och lärarbetyg (svart linje). Åren 1998–2012 gällde 1994 års fyrgradiga betygsskala och åren 2013–2014 gällde nuvarande sexgradiga skala (där ljusgrön linje anger provbetyg och blå lärarbetyg). Engelska.**



Man kan notera några saker:

- Den genomsnittliga betygspoängen varierar mer mellan olika år för provbetygen än för lärarbetygen.
- Den genomsnittliga betygspoängen för provet är klart lägst i matematik och högst i engelska.
- Skillnaden mellan den genomsnittliga betygspoängen för lärarbetyget och provbetyget är klart störst i matematik och minst i engelska.
- Den genomsnittliga betygspoängen har över tid ökat, mest i engelska och minst i matematik.

Prov våren 2014

I föregående figurer visades skillnaderna mellan lärarbetyg och provbetyg uttryckta i genomsnittlig betygspoäng. Ett annat sätt att visa dessa skillnader är att ange hur många betygssteg det skiljer mellan lärarbetyget och provbetyget. Detta visas i tabell 2 för ämnet svenska våren 2014.

Tabell 2 Antal och andel elever med olika avvikelser uttryckt i betygssteg (Lbet = lärarbetyg, Pbet = provbetyg). Svenska åk 9, vt 2014.[8]

Skillnad Lbet–Pbet	Antal elever	Procent total	Giltig procent
-5	1	0,0	0,0
-4	4	0,0	0,0
-3	29	0,0	0,0
-2	449	0,5	0,6
-1	8 582	9,5	10,6
0	53 087	59,0	**65,8**
1	16 584	18,4	20,6
2	1 699	1,9	2,1
3	184	0,2	0,2
4	10	0,0	0,0
Summa	**80 629**	**89,5**	**100**
Saknade	**9 412**	**10,5**	
Total	**90 041**	**100**	

Avvikelser[9] varierar mellan ämnen på nationell nivå

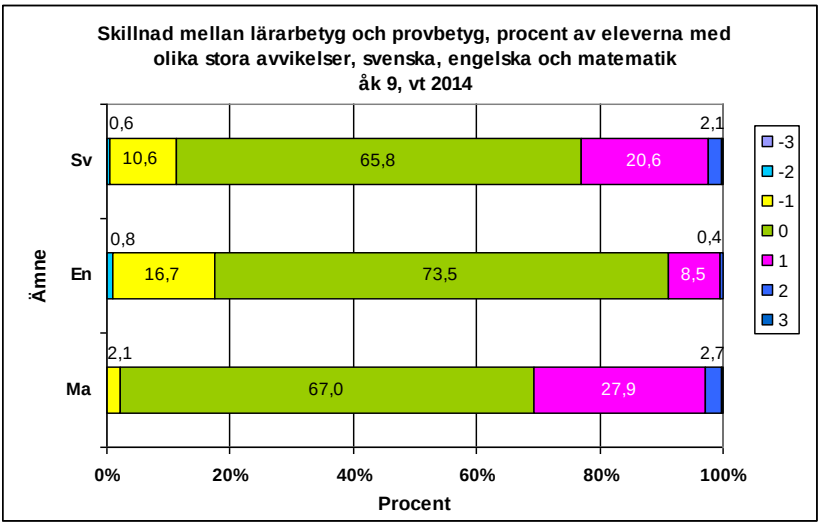
Av tabellen framgår att 65,8 procent av de elever som hade fullständigt provbetyg[10] fick samma betyg av läraren som de hade på provet medan 10,6 procent av eleverna fick ett lärarbetyg som låg ett steg under provbetyget och 0,6 procent fick ett lärarbetyg som låg två steg under provbetyget. Sammantaget fick cirka 11,2 procent av eleverna ett lägre betyg av läraren än de fick på provet.[11] Detta räknas som en *negativ avvikelse* på 11,2 procent. På motsvarande sätt blir den *positiva avvikelsen* cirka 22,9 procent. Med nettoavvikelse avses skillnaden mellan positiv och negativ avvikelse, dvs. positiv avvikelse - negativ avvikelse. Med total avvikelse menas positiv avvikelse + negativ avvikelse (eller alternativt 100 minus

[8] Ytterligare 7 981 elever gjorde provet och fick betyg i svenska som andraspråk.
[9] Man kan diskutera om avvikelse är ett bra begrepp, eftersom det ger intrycket att det finns något som är korrekt och något som inte är korrekt och därmed avviker. Det neutrala vore att tala om skillnaden mellan provbetyg och lärarbetyg. Dock är avvikelse etablerat som begrepp och används därför även här.
[10] Denna grupp utgjorde 59 procent av samtliga elever som var registrerade i ämnet. Bortfallet av elever motsvarade 10,5 procent av eleverna; dessa saknade resultat på ett eller flera delprov eller hade inte fått betyg av läraren.
[11] Något tiotal (29+4+1) elever har större avvikelse än -2, dvs. större än två betygssteg.

”ingen avvikelse” i procent, dvs. 100 - 65,8 = 34,2 procent i det aktuella fallet)[12].

Figur 4 illustrerar ovanstående avvikelser i svenska, engelska och matematik.

Figur 4 **Andel elever som fått samma provbetyg och lärarbetyg, respektive betyg som avviker med olika antal betygssteg. ”-1” betyder att lärarbetyget är ett steg lägre än provbetyget, etc.**



Av figur 4 framgår att engelska har minst total avvikelse, där 26,5 procent (100 – 73,5) av eleverna får ett annat betyg än vad de hade på provet. För svenska och matematik är andelen provtagare med samma provbetyg och lärarbetyg ungefär lika stor, 66–67 procent, vilket innebär att den totala avvikelsen är ungefär lika stor för dessa båda ämnen, 33–34 procent.

När det gäller nettoavvikelsen är skillnaden stor mellan främst engelska och matematik. För matematikens del blir nettoavvikelsen drygt 29 procent[13]. Så stor är alltså skillnaden mellan den andel av eleverna som får ett högre betyg av läraren än vad de hade på provet och den andel elever som får ett lägre lärarbetyg. För engelskans del blir nettoavvikelsen negativ, dvs. andelen elever som får lägre

[12] På grund av avrundningsfel kan olika sätt att beräkna ge olika värden på decimalen.
[13] 27,9 + 2,7 - 2,1 = 28,5 procent.

lärarbetyg än provbetyg är större än den andel som får högre lärarbetyg än provbetyg, och nettoavvikelsen blir cirka - 9 procent.[14] Cirka 65 till 75 procent av eleverna har alltså samma provbetyg och lärarbetyg.

Avvikelsen är olika för olika provbetyg

Om man delar upp provbetyg och lärarbetyg för ämnet svenska i en korstabell blir resultaten i enlighet med tabell 3. Summeras värdena i diagonalen blir summan 65,8, summeras värdena under diagonalen (Lbet > Pbet) blir summan 22,9 och summeras värdena ovanför diagonalen blir summan 11,2, dvs. samma värden som tidigare visats i tabell 2 och figur 4.

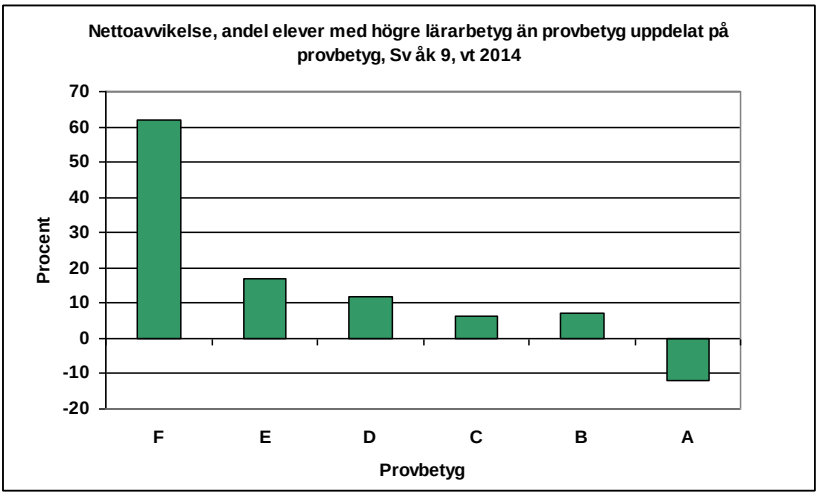
Tabell 3 **Samband mellan provbetyg och lärarbetyg i svenska, åk 9, vt 2014 (procent).**

Betyg	Provbetyg						Total
	F	E	D	C	B	A	
F	1,3	0,4	0,1				1,8
E	2,0	14,7	3,8	0,3			20,9
D	0,1	3,9	14,2	3,4	0,2		21,8
C		0,7	6,0	18,5	2,3		27,5
B		0,1	0,7	4,9	11,5	0,7	18,0
A			0,1	0,6	3,7	5,6	10,1
Total	3,4	19,8	24,9	27,7	17,7	6,4	100

Om man delar upp nettoavvikelsen per provbetyg, dvs. hur stor andel av eleverna med respektive provbetyg som får högre lärarbetyg än provbetyg får man det resultat som visas i figur 5.

[14] 8,5 + 0,4 - 16,7 - 0,8 = - 8,6 procent.

Figur 5 Nettoavvikelse för elever med olika provbetyg. Procent av antalet med respektive provbetyg.



Av figuren framgår att för drygt 60 procent av eleverna med provbetyg F har läraren satt ett betyg som är högre än provbetyget. Eftersom F är det lägsta provbetyget kan ett lägre betyg inte sättas av läraren. Nettoavvikelsen blir således densamma som den totala avvikelsen för elever med provbetyg F.

För provbetygen E till B ligger nettoavvikelserna på mellan cirka 5 och 15 procent av eleverna med respektive provbetyg. För betyg A gäller samma sak som för betyg F, men omvänt. Ingen lärare kan sätta ett högre betyg än provbetyg A och därför blir nettoavvikelsen lika med den totala avvikelsen som endast kan vara negativ. För cirka 12 procent av eleverna med provbetyg A i svenska gällde att läraren satte ett lägre betyg än provbetyg, vilket innebär att för övriga 88 procent sammanföll lärarens betyg A med provbetyg A.

Mönstret i figur 5 är karaktäristiskt för alla prov. Det betyg som har den största relativa avvikelsen är alltid betyget F.

Några begrepp i sammanfattning

Betygspoäng: Det värde respektive betyg tilldelas vid beräkningar (se tabell).
Total avvikelse: Andel elever med annat lärarbetyg än provbetyg.
Positiv avvikelse: Andelen elever med högre lärarbetyg än provbetyg.
Negativ avvikelse: Andelen elever med lägre lärarbetyg än provbetyg.
Nettoavvikelse: Positiv avvikelse - negativ avvikelse. Detta begrepp används av Skolverket.
Avvikelse i betygspoäng: Betygspoäng för lärarbetyg - betygspoäng för provbetyg.

Namn	Betyg och betygspoäng					
	F	E	D	C	B	A
Diff_6	1	2	3	4	5	6
Diff_20	0	10	12,5	15	17,5	20

Diff_1: Markerar att skillnaden (Diff) mellan lärarbetyg (Lbet) och provbetyg (Pbet) anges som nettoavvikelse baserad på tre möjliga värden: negativ (- 1), ingen (0) eller positiv (1), (dvs. Lbet<Pbet ger - 1, Lbet=Pbet ger 0, Lbet>Pbet ger + 1 oberoende av antal betygssteg). Diff_1 uttrycker avvikelsen som skillnad mellan andel positiv avvikelse och andel negativ avvikelse. Detta begrepp används av Skolverket.

Diff_6: Markerar att skillnaden mellan betyg görs enligt en skala med ett steg mellan varje betygssteg, 1–6 betygspoäng (jämför tabell 2). Diff_6 uttrycker avvikelsen som skillnad i betygssteg, Lbet–Pbet. Värdet blir positivt om Lbet>Pbet och negativt om Lbet<Pbet.

Diff_20: Markerar att aktuell beräkning baseras på officiell skala med olika viktning av betygssteg, 0–20 betygspoäng (jämför tabell 2). Diff_20 uttrycker skillnad i betygspoäng och används av Skolverket vid redovisning av GBP och meritvärden.

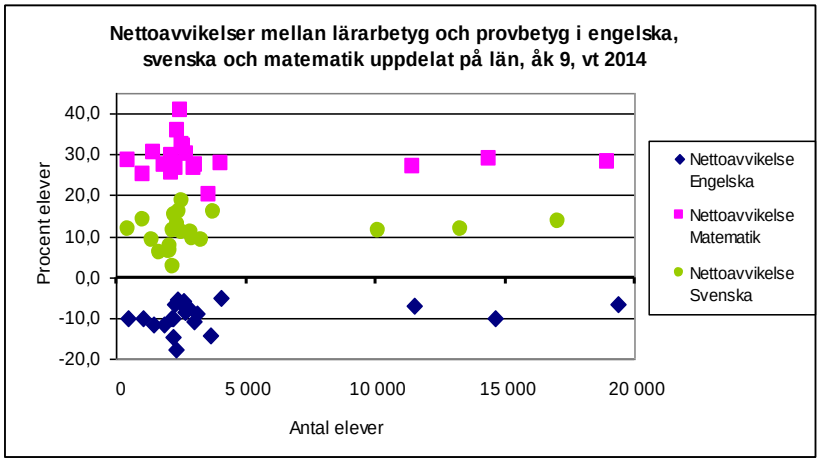
Avvikelser för olika grupper av elever

Provresultat kan förutom på nationell nivå även redovisas på gruppnivå – läns-, kommun- eller skolenhetsnivå[15]. I det här sammanhanget kommer provresultat i vissa fall också att redovisas på klassnivå, för om provbetygen i ökad grad ska styra lärarbetygen har det betydelse hur klassernas resultat ser ut och hur stora grupperna är. Samtidigt är många av de grupper som ingår i resultatinsamlingen små, vilket innebär betydande mätosäkerhet.

Länsnivå

Figur 6 visar nettoavvikelsen (Diff_1) för olika län för ämnena svenska, engelska och matematik.

Figur 6 **Genomsnittliga nettoavvikelser (Diff_1) på länsnivå för olika ämnen och efter antal elever i länet som gjort nationella provet i åk 9, vt 2014.**



Figuren visar, liksom tidigare figurer, att matematik har den största nettoavvikelsen och engelska den minsta. Den sistnämnda är till och med är negativ vårterminen 2014. Man kan också notera att

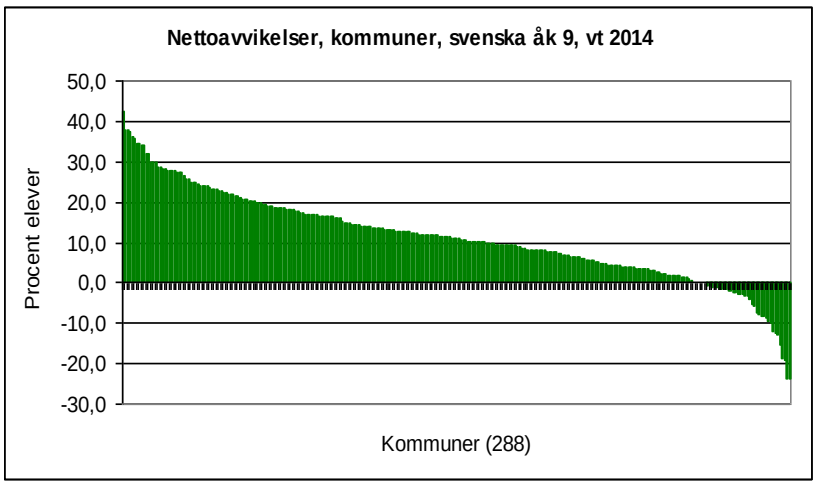
[15] De kan också gälla olika elevgrupper: flickor eller pojkar, elever med olika socioekonomisk bakgrund, utifrån etnicitet etc. men detta är inte aktuellt i det här sammanhanget.

variationen i genomsnittlig nettoavvikelse inom samma ämne är förhållandevis betydande (cirka 20 procentenheter) mellan olika län, framför allt för länen med färre än 5 000 elever. För de tre stora länen Stockholm, Västra Götaland och Skåne ligger de genomsnittliga nettoavvikelserna mer i linje med riksgenomsnittet för Sverige (29, 11 respektive - 10 procent för de tre ämnena).[16]

Kommunnivå

På kommunnivå är det vanligt att resultaten redovisas som i figurerna nedan. Figur 7 visar resultaten för ämnet svenska.[17] Figuren visar kommunerna rangordnade från den kommun som har störst genomsnittlig nettoavvikelse till den som har minst (i det här fallet negativ) avvikelse.

Figur 7 **Nettoavvikelser (Diff_1) för olika kommuner uppdelat efter avvikelsens storlek, svenska åk 9, vt 2014.**



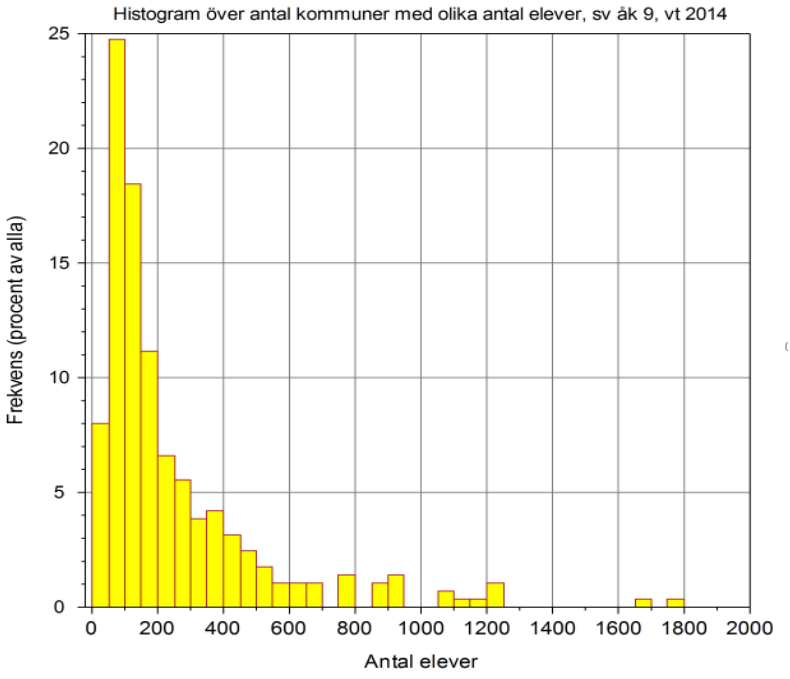
Av figuren framgår att nettoavvikelsen varierar påtagligt mellan olika kommuner. Medelvärdet är 11,0 procent och standardavvikelsen 11,5 procent.

[16] Data hämtade från Skolverkets databas Siris med nettoavvikelse räknad enligt Diff_1.
[17] Innefattar både kommunala och fristående skolor i respektive kommun.

Antalet elever har betydelse för avvikelsen

Figur 8 nedan visar ett histogram över hur antalet elever fördelar sig mellan olika kommuner. Vanligast är kommuner där 50–100 elever gjort provet i svenska (knappt 25 procent av kommunerna). Man ser också att var tolfte kommun (cirka 8 procent) har färre än 50 elever som genomför provet och får betyg.

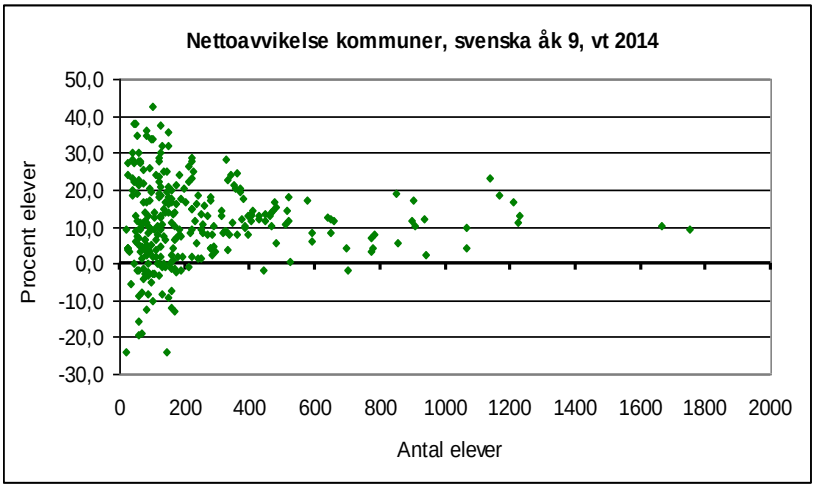
Figur 8 **Fördelning av kommuner efter antal elever, svenska åk 9, vt 2014.**



Om man visar samma nettoavvikelse som i figur 7 i relation till *antalet elever* i kommunen får man nedanstående bild.[18]

[18] De stora kommunerna Stockholm och Göteborg har så många elever att de ligger utanför skalan i figuren (avvikelse 16,2 respektive 13,2 procent).

Figur 9 Nettoavvikelse (Diff_1) för kommuner uppdelat efter antal elever, svenska åk 9, vt 2014.[19]



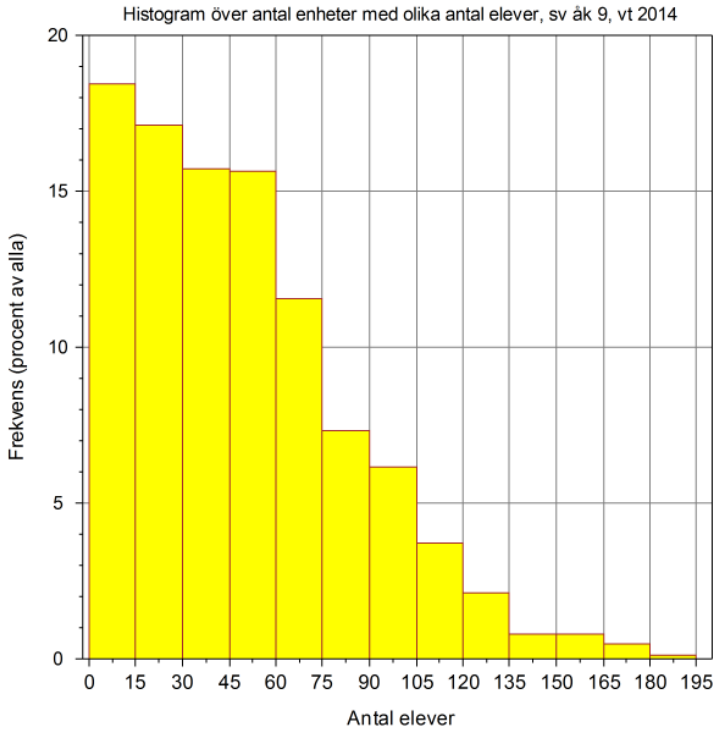
Av figur 9 framgår att variationen i nettoavvikelse ökar när antalet elever i kommunen minskar. Det är alltså framför allt de små kommunerna som står för såväl de stora som små nettoavvikelserna, vilket inte framgår av figur 7.

Skolenhetsnivå

Figur 10 visar ett histogram som anger skolenheter med olika antal elever. Den nedersta stapeln visar att drygt 260 skolenheter (drygt 18 procent av de 1 614 enheter som ingår i underlaget) har rapporterat färre än 15 elever som deltagit i provet och fått betyg i svenska.[20] Därefter följer skolorna i storleksordning. Det genomsnittliga antalet elever per skolenhet är 50.

[19] Stockholm med drygt 6 000 och Göteborg med cirka 3 500 elever ingår inte i figuren.
[20] De ingår dock inte i underlagen för figur 9 och 10.

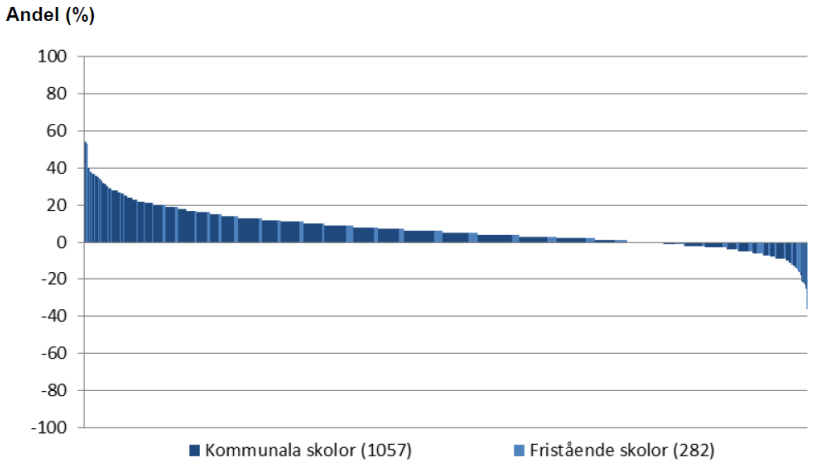
Figur 10 **Fördelning av skolenheter efter antal elever, svenska åk 9, vt 2014.**



Skolverket redovisar årligen nettoavvikelser för skolenheter med minst 15 elever som har både provbetyg och lärarbetyg. Figur 11 visar utfallet för proven i svenska våren 2014.[21]

[21] Figurerna hämtade från Skolverket (2015a).

Figur 11 Nettoavvikelse (Diff_1) uppdelat på skolenheter med minst 15 elever. Svenska, åk 9, vt 2014.



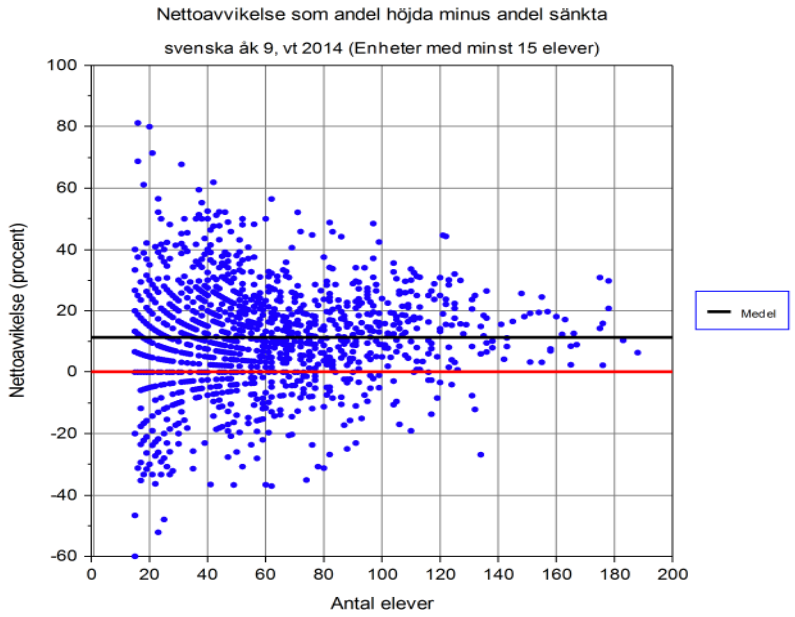
En jämförelse mellan figur 7 och figur 11 visar att bilderna är likartade. Skalorna är något olika men i båda fallen ligger de största värdena för nettoavvikelsen en bit över 40 procent och de lägsta ner mot - 40 procent.

Även på skolenhetsnivå har antalet elever betydelse för nettoavvikelsen. Figur 12 visar samma skolenheter som figur 11 men uppdelade efter antal elever.

Medelvärdet[22] är 11 procent och standardavvikelsen är 17 procent, vilket betyder att cirka två tredjedelar av skolenheterna har nettoavvikelser mellan 28 procent och - 6 procent.[23] En jämförelse med avvikelser på läns- och kommunnivå visar att den genomsnittliga avvikelsen ligger på samma nivå medan standardavvikelsen ökar när grupperna bli mer uppdelade, vilket man kan förvänta sig eftersom det handlar om en statistisk effekt.

[22] Beräknat som medelvärdet av skolenheternas medelvärden.
[23] Medelvärdet plus/minus en standardavvikelse.

Figur 12 **Nettoavvikelse (Diff_1) för skolenheter med minst 15 elever uppdelat efter antal elever. Svenska åk 9, vt 2014.**



Även på skolenhetsnivå är det tydligt att det är de små enheterna som varierar mest i nettoavvikelse.

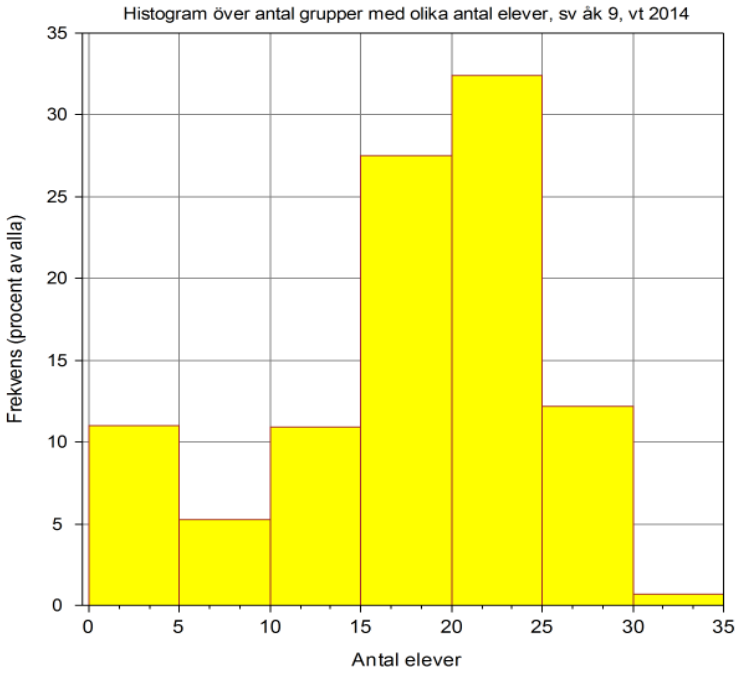
Det regelbundna mönstret till vänster i figuren uppstår på grund av att grupperna där är små. I en grupp på 15 elever motsvarar en elev cirka 7 procent (1/15) av eleverna i gruppen. Om 1 elev i en klass med 15 elever har ett högre lärarbetyg än provbetyg (och övriga elever samma provbetyg och lärarbetyg) blir således nettoavvikelsen 7 procent (1/15). Om 2 elever har högre betyg blir däremot avvikelsen 13 procent (2/15) osv. Om klassen i stället råkar ha 16 elever betyder nettoavvikelse för 1 elev att 6 procent av gruppen (1/16) avviker, för 2 elever 13 procent (2/16) osv.

Figur 12 blir också något missvisande för enheter med få elever, eftersom många enheter har samma värden och därför täcker varandra i diagrammet.

Gruppnivå

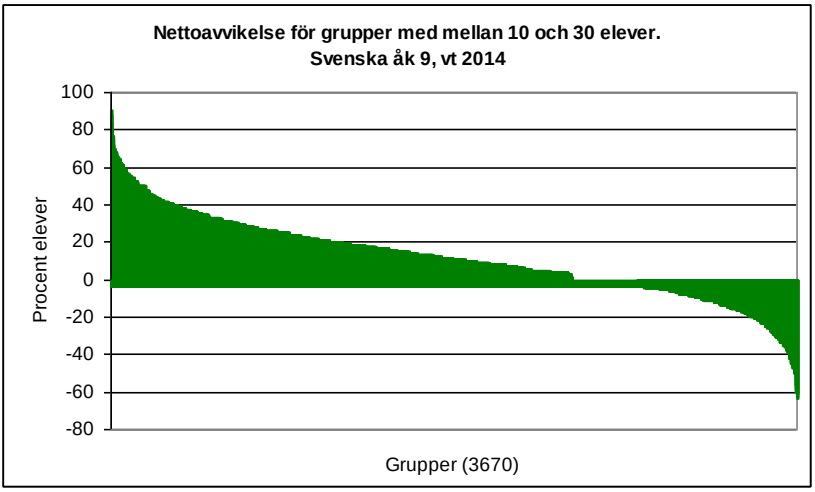
Rapporteringen av grupper varierar, och ibland är det svårt att avgöra av det statistiska underlaget vad som avses – om det är olika grupper inom ramen för en gemensam undervisningsgrupp eller om det är en sammanhållen undervisningsgrupp för olika ämnen. För grundskolans del verkar flertalet grupper i statistikunderlaget av beteckningarna att döma att ingå i en skolorganisation på samma sätt som traditionella klasser. Oavsett vilket gäller för svenskämnets del nedanstående fördelning (figur 13) av grupper utifrån det antal elever som ingår.

Figur 13 **Fördelning av grupper (klasser) efter antal elever. Svenska åk 9, vt 2014.**



Den vanligaste gruppstorleken är 20–25 elever (knappt 1 500 grupper av totalt 4 295). Genomsnittet för samtliga grupper är 19 elever (18,77) och standardavvikelsen är 9 (9,43).

Figur 14 Nettoavvikelser (Diff_1) på gruppnivånivå för grupper med minst 10 elever. Svenska åk 9, vt 2014.

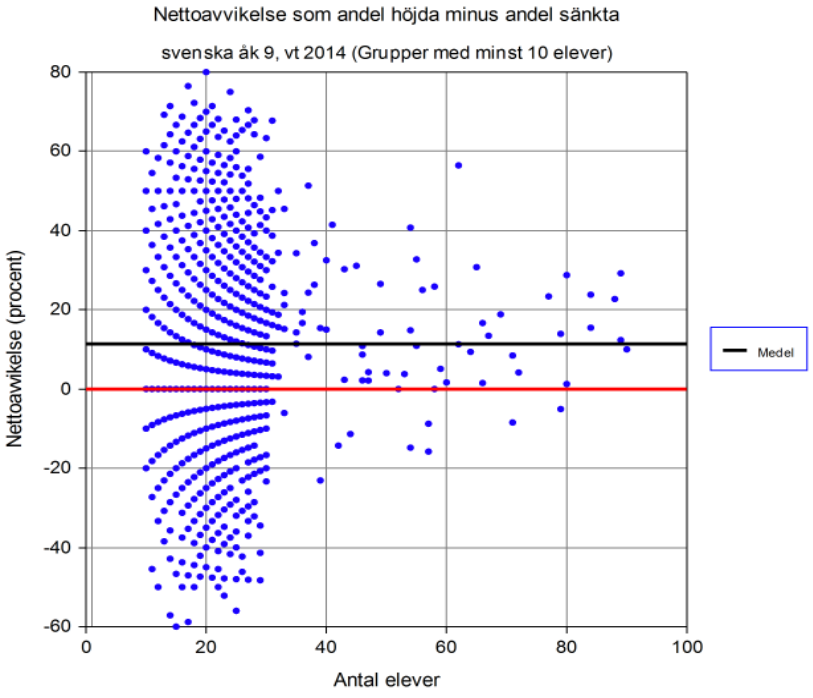


Medelvärdet i nettoavvikelse för grupper är 11,6 procent medan standardavvikelsen är 21,5 procent. För en enligt det statistiska underlaget genomsnittlig klass med 19 elever betyder det att de elever som fått högre lärarbetyg än provbetyg är två fler än de elever som fått lägre lärarbetyg än provbetyg.[24]

I figur 15 redovisas i stället fördelningen i relation till gruppernas storlek. Här framträder inte spridningens beroende av antalet elever lika tydligt som den gör för skolenheter, eftersom skillnaden mellan gruppstorlekarna är förhållandevis liten i relation till skillnaderna mellan skolenheter. Däremot framträder en del andra mönster för i synnerhet de små grupperna, där avvikelsen för 1 elev kan innebära en nettoavvikelse för gruppen på upp till 10 procent, för 2 elever 20 procent osv.

[24] 0,116 * 19 = 2,2 vilket avrundat blir 2 elever. Eftersom elever inte är delbara måste man räkna på flera klasser och elever om man ska få samma värde som för hela den grupp som gjort provet, dvs. 11,6 procent.

Figur 15 Nettoavvikelse (Diff_1) på gruppnivå uppdelat efter gruppstorlek. Svenska åk 9, vt 2014.



Kommentar

De tabeller och figurer som redovisas ovan utgör ett utsnitt av möjliga presentationer. Av tabellerna och figurerna framgår att resultat kan redovisas på många olika sätt och att olika presentationsformer ger olika intryck av resultaten. Ett exempel är de nettoavvikelser som visas i figurerna 7, 11 och 14. De ger intryck av att variationen är betydande (vilket den förvisso kan anses vara).

Om man i stället redovisar resultat som i figurerna 8, 12 och 15 ser man att avvikelserna också beror på enheternas storlek. De små enheternas resultat tenderar att variera betydligt mer än de stora enheternas. Det finns alltså statistiska och slumpberoende effekter med i bilden av nettoavvikelsen.

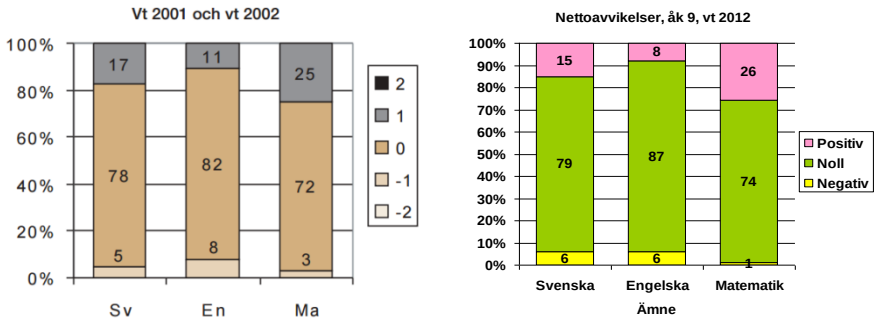
Liten förändring av avvikelse över tid

Figur 1–3 i förra avsnittet ger en bild av utvecklingen när det gäller den genomsnittliga betygspoängen för provbetyg och lärarbetyg under de år nationella prov har funnits inom ramen för det mål- och kunskapsrelaterade betygssystemet. Övriga figurer visar resultat från 2014 års prov i svenska i årskurs 9. Men hur ser bilden ut om man försöker granska betygsfördelningen och netto-avvikelserna över tid? Hur mycket var unikt för 2014 och hur mycket har sett likadant ut olika år?

Utveckling på nationell nivå

Figur 16 visar andelen elever (i procent) med positiv, ingen eller negativ avvikelse för 2001 och 2012 i de tre provämnen som var gemensamma de aktuella åren.

Figur 16 **Avvikelser[25] i svenska, engelska vt 2001 och matematik vt 2002 (den vänstra figuren) samt för samma tre ämnen vt 2012 (den högra figuren).**



Man kan notera att bilden är mycket likartad.[26] I matematik har 72–74 procent av eleverna samma provbetyg och lärarbetyg medan 25–26 procent har ett högre lärarbetyg och endast någon enstaka procent ett lägre. Motpolen är engelska där cirka 82 respektive 87 procent

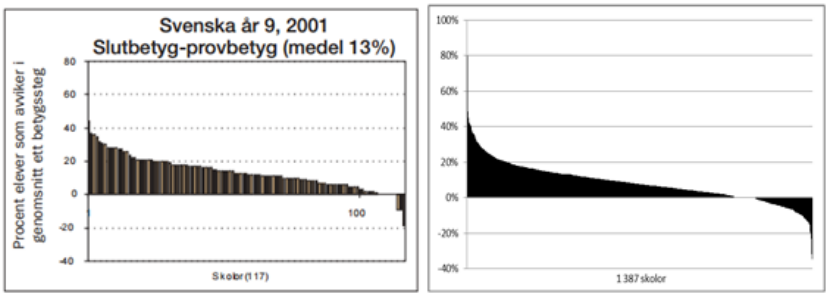
[25] Den vänstra figuren anger egentligen skillnad i betygssteg men andelen elever med avvikelse större än ett betygssteg är försumbar i sammanhanget.
[26] Se bilaga 1 för utförligare jämförelseunderlag.

har samma provbetyg och lärarbetyg medan andelen med högre respektive lägre lärarbetyg är ungefär lika stor för båda åren. Svenskämnet intar en mellanställning. Figur 16 visar att matematik på nationell nivå liksom tidigare har den största nettoavvikelsen och engelska den minsta. Man kan också konstatera att resultaten 2001 och 2012 är mycket likartade.

Förändring på skolnivå

Förändringen över tid på nationell nivå är i stort sett försumbar enligt figur 16. Frågan är då om relationen mellan provresultat och betyg förändrats på skolnivå. Figur 17 visar nettoavvikelsen för svenska i årskurs 9 på skolnivå 2001 och 2012.[27]

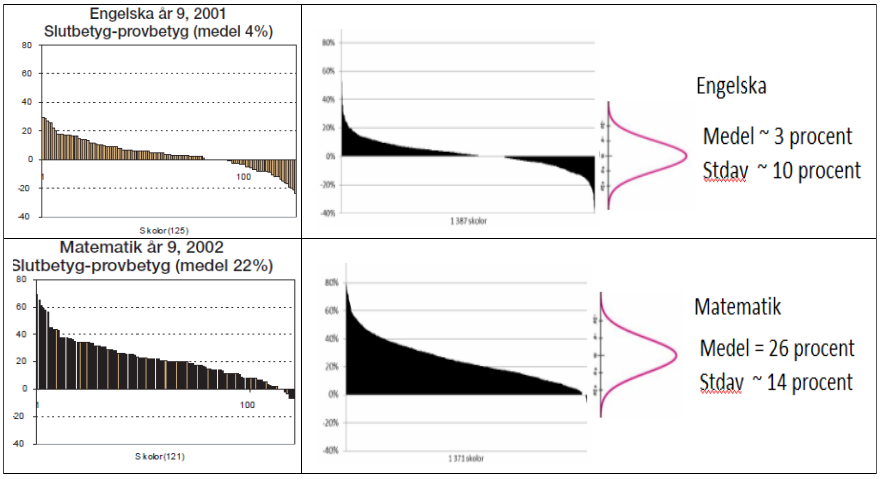
Figur 17 Nettoavvikelser för skolor vt 2001 och vt 2012, svenska åk 9.[28]



Av figur 18 framgår mönstret i engelska och matematik för vårterminen 2012 och vårterminen 2001/2002.

[27] Resultatinsamling gjordes från början på ett stickprov av skolor. Från och med läsåret 2002/03 samlas alla provresultat i årskurs 9 in.
[28] Ur Skolverket (2003) och Skolverket (2013a).

Figur 18 **Nettoavvikelser för skolor, engelska vt 2001 och 2012 samt matematik vt 2002 och 2012, åk 9.**[29]



Av figur 17 och 18 framgår att mönstren 2012 närmast är kopior av motsvarande mönster tio år tidigare, med skillnaden att 2001 och 2002 års resultat baseras på ett stickprov av skolor medan 2012 gäller samtliga skolor, vilket ger jämnare fördelningar.

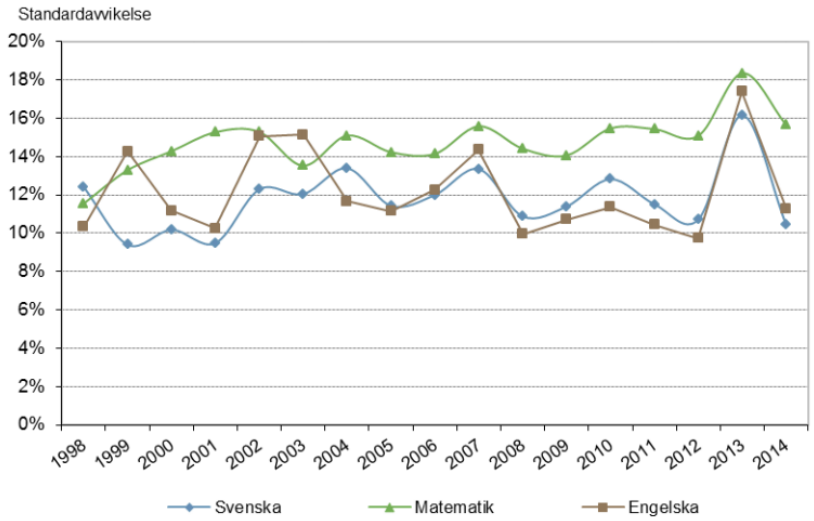
Även figur 19 nedan tyder på att förändringarna varit små. I varje fall ser man ingen tydlig trend när det gäller variationen. Bilden anger variationen i nettoavvikelse mellan skolor uttryckt som standardavvikelse.[30] Nettoavvikelserna har varierat mest i matematik och något mindre i svenska och engelska, dvs. samma mönster som framträtt tidigare.

Våren 2013 användes den nya sexgradiga betygsskalan för första gången i årskurs 9, vilket möjligen kan förklara den kraftiga uppgången det året. Våren 2014 ligger spridningen åter på ungefär samma nivå som under den fyrgradiga betygsskalans tid. Om detta är en tillfällighet kan inte bedömas efter så kort tid med den nya betygsskalan. Hur spridningen mellan skolor kommer

[29] Ur Skolverket (2003) och Skolverket (2013a).
[30] Ur Skolverket (2015a).

att se ut framöver kan inte heller bedömas efter endast två års resultat.[31]

Figur 19 **Spridning i skolors nettoavvikelse (Diff_1) för svenska, engelska och matematik vt 1998 till vt 2014, åk 9.**



Kommentar

Variationerna i nettoavvikelser uppmärksammades i början av 2000-talet, framför allt i samband med att Skolverket redovisade ett regeringsuppdrag om provsystemet.[32] Som en följd av redovisningen upprättade Skolverket en handlingsplan som ledde till ett antal åtgärder som sedan efterhand har utökats. Några exempel på åtgärder är att bedömningsanvisningarna till proven har förändrats, att sambedömning uppmuntras, att allmänna råd har utarbetats för bedömning och betygssättning, att rektorns ansvar för att följa upp provresultat och betyg betonas samt att Statens skolinspektion har granskat provanvändning och betygssättning. Trots insatserna upp-

[31] Eftersom Skolverket redovisar nettoavvikelse (Diff_1) får dock ökningen av antalet betygssteg troligen liten betydelse. Man kan därmed diskutera det rimliga i att endast räkna avvikelsens riktning (Lbet högre eller lägre än Pbet) utan ta hänsyn till hur många betygssteg som skiljer mellan lärarbetyget och provbetyget. Med flera betygssteg i skalan blir det vanligare att det skiljer två eller flera betygssteg.
[32] Skolverket (2003).

repas samma mönster när det gäller betygsnivåer, betygsfördelningar och skillnader mellan provbetyg och lärarbetyg. I vissa fall, t.ex. när det gäller genomsnittlig betygspoäng för provbetygen, kan de årliga variationerna vara ganska stora men den långsiktiga trenden och ordningen mellan ämnen är i huvudsak stabil. Framför allt är mönstren i relationerna mellan provbetyg och lärarbetyg stabila såväl på nationell nivå som på skolnivå.

Gymnasieskolan

Provresultaten i grundskolan och deras utveckling över tid väcker ett antal frågor. För gymnasieskolans del blir bilden än mer komplex, t.ex. för att:

- proven ges i olika versioner på höst- respektive vårterminen
- elevunderlagen i olika program kan variera mellan vår och höst och förändras över tid
- undervisningen i en och samma kurs kan utformas på olika sätt beroende på vilket program det gäller[33].

Dessutom ges många fler prov per år i gymnasieskolan. I grundskolans årskurs 9 ges årligen ett prov[34] i vardera svenska/svenska som andraspråk, engelska och matematik (vårterminen 1998–vårterminen 2014).[35] I gymnasieskolan gavs två prov i svenska/svenska som andraspråk, fyra prov i engelska och åtta prov i matematik[36] årligen (1995–2011) och därefter årligen fyra prov i svenska/svenska som andraspråk, fyra i engelska och arton i matematik (höstterminen 2011–tills vidare).

Den stora mängden prov i gymnasieskolan gör att endast resultat för vissa kurser redovisas här.[37] Resultat redovisas inte heller för

[33] Korp (2006).
[34] Provet består i allmänhet av olika delprov som kan ges vid olika tidpunkter.
[35] Efter 2010 har nationella prov tillkommit i SO- och NO-ämnen men de tas inte med i den här redovisningen.
[36] De första åren gavs inte prov i varje matematikkurs varje år. Se http://www.edusci.umu.se/np/np-b-d/tidigare-prov/
[37] Skolverket redovisar årligen resultaten på sin webbplats, se http://www.skolverket.se/om-skolverket/publikationer/sok?_xurl_=http%3A%2F%2Fwww5.skolverket.se%2Fwtpub%2Fws%2Fskolbok%2Fwpubext%2Ftrycksak%2FBlob%2Fpdf3487.pdf%3Fk%3D3487

höstterminens prov eftersom de flesta proven görs på våren. Vissa undantag finns dock – t.ex. gjordes proven i matematik A huvudsakligen på hösten i årskurs 1 av elever på Naturvetenskapsprogrammet medan man inom flertalet av de övriga programmen valde att göra vårens version av samma prov. Motsvarande förhållande gäller i dag, men eftersom provet i matematik A efter 2011 motsvaras av tre versioner av matematik 1 (1a, 1b och 1c) är det endast en mindre del av eleverna på Naturvetenskaps- och Teknikprogrammen som genomför vårens prov i matematik 1c. Däremot väljer man inom majoriteten av övriga program att genomföra proven 1a eller 1b på våren.

Sedan höstterminen 2011 samlas samtliga provresultat in från gymnasieskolan. Innan dess samlades resultat in från ett urval av skolor valda så att alla skolor skulle delta i insamlingen inom en sexårsperiod. Denna insamling gällde dessutom endast vårterminens prov.

Betygsskillnader varierar över tid och mellan ämnen

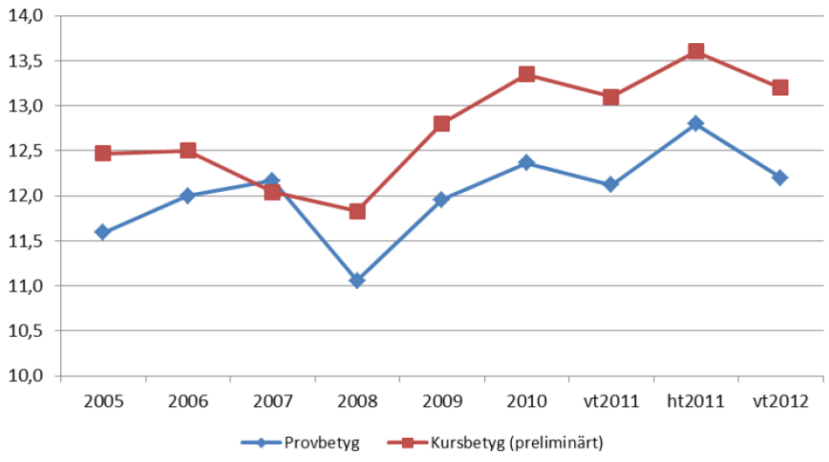
För de tidigare kärnämnena svenska B, engelska A och matematik A redovisar Skolverket nedanstående bilder. Röd kurva representerar kursbetyget (Lbet) och blå kurva provbetyget (Pbet).

Svenska

Åren 2002–2004 redovisades inga sammanfattande provbetyg i svenska och engelska.

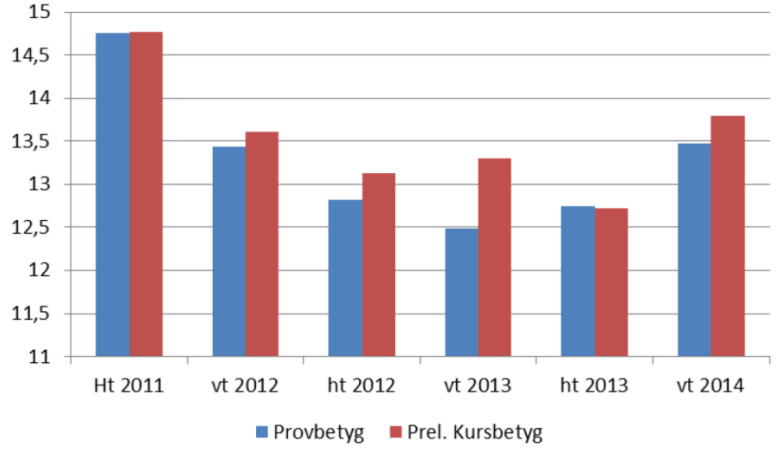
För höstterminen 2011 och vårterminen 2012 gäller att vissa elever genomför prov enlig Gy 2011 (se figur 20 för svenska B).

Figur 20 **Genomsnittlig betygspoäng (GBP) för provbetyg och kursbetyg. Svenska B.**[38]



Höstterminen 2011 började Gy 2011 och den nya sexgradiga betygsskalan att gälla, vilket gett följande resultat (figur 21) för svenska 1 sedan höstterminen 2011.[39]

Figur 21 **Genomsnittlig betygspoäng (GBP) för provbetyg och kursbetyg enligt Gy 2011. Svenska 1.**



[38] Ur Skolverket (2013b).
[39] Skolverket (2015b).

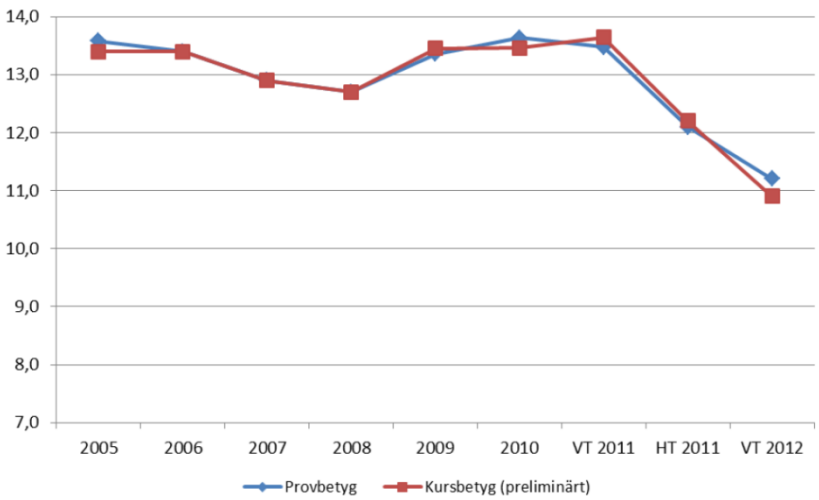
Engelska

För engelska, där engelska A motsvaras av engelska 5 i Gy 2011, redovisar Skolverket figur 22 för engelska A och figur 23 för engelska 5.

Eftersom en ny betygsskala och nya ämnesplaner började gälla fr.o.m. hösten 2011 får man räkna med osäkra resultat de närmast därpå följande terminerna. Detta avspeglar sig också i figurerna, vilka enligt erfarenhet från tidigare systemskiften bör tolkas med försiktighet tills resultaten börjar stabilisera sig.

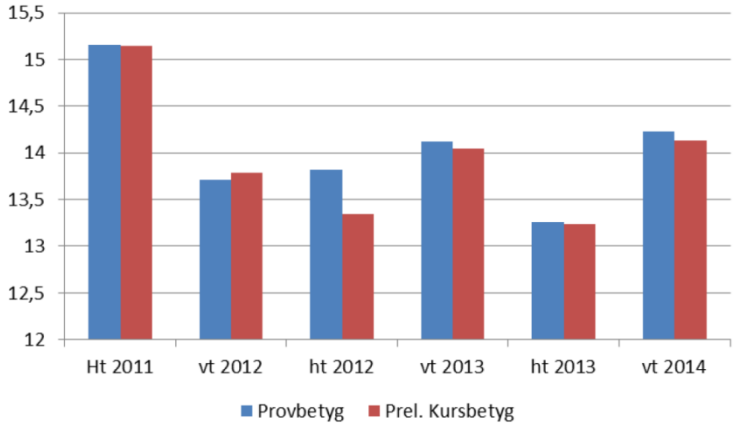
För engelska A kan man i figur 22 notera en tydlig nedgång i genomsnittlig betygspoäng när de två första proven enligt Gy 2011 gavs (höstterminen 2011 och vårterminen 2012). Detta kan antas bero på att de grupper som då gjorde proven i engelska A hade en annan sammansättning än tidigare års grupper, eftersom de elever som gjorde provet i engelska 5 inte ingick.

Figur 22 **Genomsnittlig betygspoäng (GBP) för provbetyg och kursbetyg. Engelska A.**



De elever som gjorde provet i engelska 5 ingår i stället i underlaget för figur 23, som visar genomsnittlig betygspoäng fr.o.m. höstterminen 2011 för provbetyg och lärarbetyg.

Figur 23 Genomsnittlig betygspoäng (GBP) för provbetyg och kursbetyg enligt Gy 2011. Engelska 5.



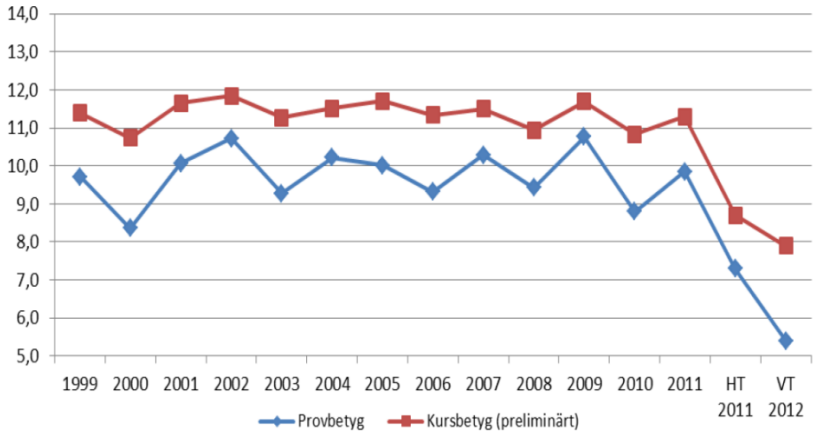
Även för engelska 5 är resultaten svårtolkade eftersom en ny ämnesplan och betygsskala gäller och genom att gruppsammansättningarna kan variera mellan provtillfällena.

Matematik

För matematik A gäller att det inte finns en motsvarande kurs i Gy 2011 utan matematik A har ersatts av tre kurser: matematik 1a, 1b och 1c.[40] Höstterminen 2011 och i synnerhet vårterminen 2012 läste flertalet elever kurser enligt Gy 2011, vilket resulterat i mycket låg genomsnittlig betygspoäng på provet (figur 24) för de relativt få elever som gjorde provet i matematik A vårterminen 2012 (tabell 4).

[40] Se Skolverket (2011a).

Figur 24 **Genomsnittlig betygspoäng (GBP) för provbetyg och kursbetyg. Matematik A.**



Tabell 4 nedan visar antalet elever som genomförde prov enligt den tidigare gymnasieskolan vårterminen 2012. Av tabellen framgår att endast knappt 2 883 elever på de yrkesförberedande programmen gjorde provet i matematik A; de flesta kan antas ha gjort provet i matematik 1a enligt den nya gymnasieskolan som infördes höstterminen 2011. Till skillnad mot matematik A genomfördes proven i svenska B liksom i de högre matematikkurserna (B–D) fortfarande av många elever vårterminen 2012.

Tabell 4 **Antal elever som genomförde prov enligt den tidigare gymnasieskolan vt 2012. Tabellen anger hur stor andel som hade samma provbetyg och kursbetyg (”Lika”) samt hur stor andel som hade avvikande betyg. Man kan notera att cirka 70–75 procent av eleverna har samma provbetyg och kursbetyg.**

Ipf 94		Antal betygssteg						
Kurs	Antal elever	-3	-2	-1	Lika	+1	+2	+3
Engelska A	5 096	0	0,5	5,9	82,6	10,9	0,1	.
Engelska B	50 421	0	0,8	4,9	76,6	17,6	0,2	0
Matematik A	2 883	.	.	1	70,4	28,3	0,3	.
Matematik B	23 221	.	0	0,7	69	29,6	0,7	0
Matematik C	12 195	.	0	1	73,5	24,7	0,7	0
Matematik D	6 662	.	0,1	1,2	76	21,6	1	.
Svenska B	42 630	0	0,4	6,1	70,7	21,7	1	0
Svenska som andraspr. B	1 128	0	0,4	4,3	68,9	25,4	0,9	0

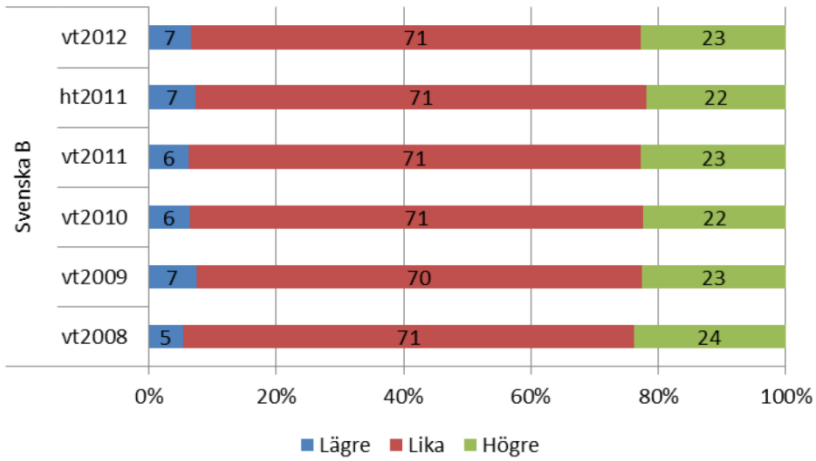
Tabell 4 visar också hur stor andel av eleverna som fått samma lärarbetyg som provbetyg (kolumnen ”Lika”) och hur stora andelar som har lärarbetyg som avviker med ett, två eller tre betygssteg från provbetyget. Den vanligaste avvikelsen, framför allt i matematik, är att lärarbetyget ligger ett betygssteg över provbetyget.

Man kan också notera att avvikelser som är större än ett betygssteg är mycket ovanliga (mindre än en procent). De inringade värdena anger att andelen elever med samma lärarbetyg och provbetyg ligger kring 70 procent +/- 5 procentenheter. Engelska A är det prov som främst bryter mönstret.

Små förändringar i avvikelser över tid[41]

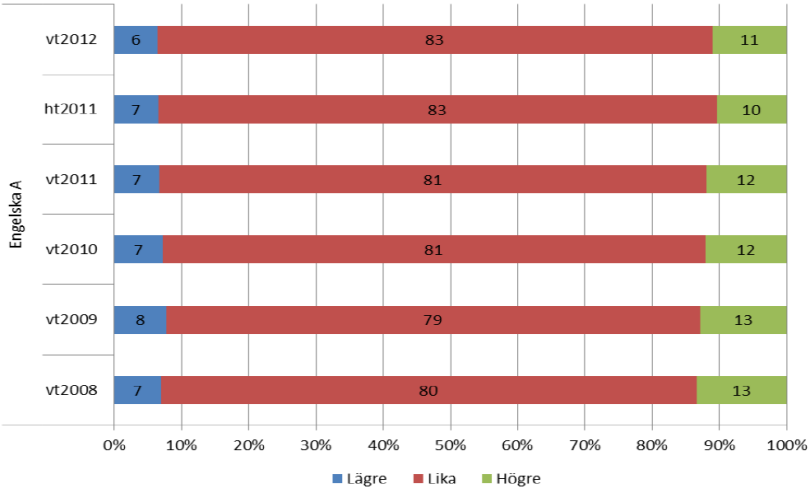
När det gäller förändringen av avvikelser över tid visar figur 25 mycket stabila värden för avvikelserna i svenska B 2008–2012. Även höstterminen 2011 faller in i mönstret, trots att det är den enda av de ingående grupperna som gjort provet på hösten.[42]

Figur 25 **Relationen mellan lärarbetyg och provbetyg för svenska B, vt 2008–vt 2012.**[43]



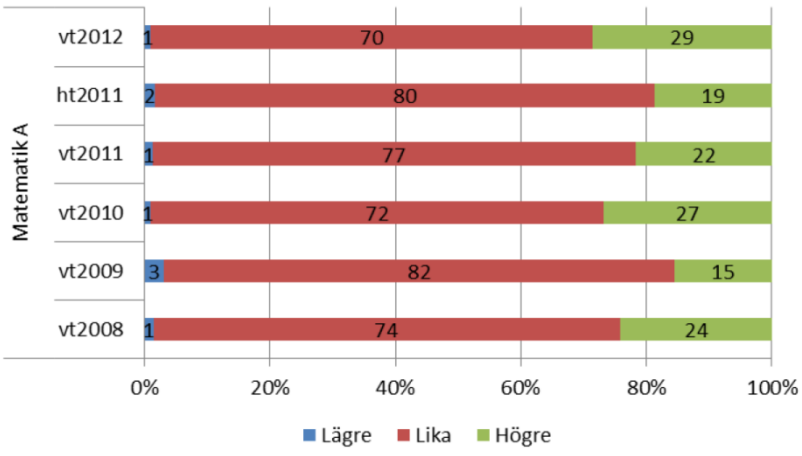
[41] Här visas ingen tidsserie för avvikelser i betyg enligt Gy 2011 eftersom de data som är tillgängliga ger för korta och osäkra tidsserier.
[42] Totalinsamling startade som nämnts 2011.
[43] Ur Skolverket (2013).

Figur 26 **Relationen mellan lärarbetyg och provbetyg för engelska A, vt 2008–vt 2012.**



Avvikelserna för kärnämnena engelska A och matematik A visas i figur 26 (ovan) och figur 27 (nedan). Värdena varierar något mer, framför allt i matematik A, medan engelska A visar en svagt stigande andel elever där lärarbetyg och provbetyg är lika.

Figur 27 **Relationen mellan lärarbetyg och provbetyg för matematik A, vt 2008–vt 2012.**



Av figur 24 framgår att den genomsnittliga betygspoängen sjönk markant höstterminen 2011 och vårterminen 2012 för matematik A, vilket dock inte påtagligt bryter mönstret när det gäller relationen mellan provbetyg och lärarbetyg (figur 27). Man kan notera en viss tendens till större andel höjda betyg (29 procent) vårterminen 2012 då provbetygen ligger på mycket låg nivå.

Liksom för grundskolan är det slående hur små de årliga förändringarna är på nationell nivå i främst svenska och engelska, medan matematik visar större årlig variation men liten förändring över längre tid.

Prov våren 2014

Några data för beräkning av avvikelser på länsnivå och kommunnivå har utredningen inte haft tillgång till för gymnasieskolans del. Redovisningen görs därför på nationell nivå, på skolenhetsnivå (minst 15 elever) och i något fall på gruppnivå (minst 10 elever).

Alla kursprov

Inledningsvis redovisas vissa sammanfattande data för samtliga kursprov. Vissa kursprov är obligatoriska för alla program medan övriga är obligatoriska för vissa program och frivilliga för andra. Tabell 5 redovisar antal elever från respektive program som erhållit både provbetyg och kursbetyg vårterminen 2014.

Tabell 5 **Antal elever med både provbetyg och lärarbetyg som genomfört kursprov vt 2014 uppdelat på program.**[44]

Program	Kurs															Totalt
	En 5	En 6	Ma 1a	Ma 1b	Ma 1c	Ma 2a	Ma 2b	Ma 2c	Ma 3b	Ma 3c	Ma4	Sva 1	Sva 3	Sv 1	Sv 3	
BA	3 228	1 088	3 261	2	1	184	7		2	5		159	15	3 122	776	11 850
BF	2 035	1 222	1 765	6	5	309	2					208	59	1 874	1 349	8 834
EE	3 399	1 826	3 508	3	9	429	14	3	18	10	1	196	55	3 036	1 042	13 549
EK	8 836	8 044	9	8 624	9	1	6 863	3	4 352	5	107	349	211	8 125	6 363	51 921
ES	6 013	6 143	9	5 491	3	9	2 197	34	311	5	6	122	93	5 360	5 711	31 507
FT	2 542	357	2 385		3	24				1		168	20	2 301	208	8 009
HA	1 860	1 103	1 765	8	2	261	4		10			234	69	1 755	752	7 823
HT	822	690	718		1	22	1		2			61	19	777	384	3 497
HU	522	657	1	473	0	0	269		50		1	18	12	471	643	3 117
HV	1 777	668	1 661	4		44	4		0	1		87	11	1 607	586	6 450
IB	465	182		45	165			98		1	1	74	17	410		1 458
IM	287	77	83	119	34	2	24	23	1	7	4	132	7	170	16	986
IMSPR	47	5	13	33	11		2	8		1	1	8				129
IN	1 186	489	858	4	54	176	2	30	1	113	61	56	8	1 134	244	4 416
IP											15		6		23	44
NA	10 126	9 781		3	2 816	1	13	5 871	2	4 447	4 844	851	645	9 219	8 501	57 120
NB	1 822	831	1 654	1		247		5	2	68	53	8	19	1 890	623	7 223
RL	1 420	550	1 427	3	2	32			1			50	6	1 549	316	5 356
RX	187	139	181			12	8			9		12	1	158	66	773
SA	14 495	14 423	18	13 457	17	0	13 098	19	2 102	5	40	617	405	13 243	12 759	84 698
TE	6 517	6 024	5	11	2 140	1	1	3 887	2	4 037	1 989	276	137	5 933	5 206	36 166
VF	809	195	739		2	42			1	4		43	2	811	141	2 789
VO	2 346	1 327	2 162	1	3	572	2		7			420	153	1 924	1 083	10 000
Totalt	**70 741**	**55 821**	**22 222**	**28 288**	**5 277**	**2 368**	**22 531**	**9 981**	**6 864**	**8 719**	**7 123**	**4 149**	**1 970**	**64 869**	**46 792**	**357 715**

Som framgår av tabellen[45] är det främst matematikkurserna som är öronmärkta för olika program medan engelska 5 och svenska 1 är obligatoriska för alla program.

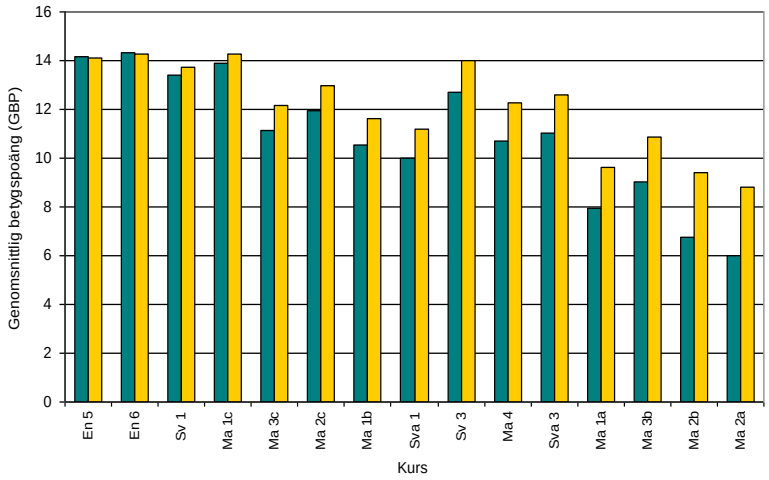
Olika betygsnivåer för olika kurser

Figur 28 visar betygsnivåerna för provbetyg och lärarbetyg uttryckta i genomsnittlig betygspoäng. Engelska 5 och 6, matematik 1c (som är en obligatorisk kurs för elever på Naturvetenskaps- och Teknikprogrammet) samt svenska 1 och 3 har de högsta genomsnittliga betygspoängen medan flertalet matematikkurser har betydligt lägre värden.

[44] Av appendix 3 till bilagan framgår vad de olika programförkortningarna står för.
[45] Detta framgår av att grupperna är stora, se också Skolverket (2011a). De gulmarkerade kurserna redovisas utförligare senare.

Av skillnaden mellan GBP för lärarbetyg respektive provbetyg kan man sluta sig till att kurserna i matematik har de största avvikelserna och kurserna i engelska de minsta. Dock är det svårt att av figuren få en uppfattning om avvikelsernas storlek i mer konkreta termer.

Figur 28 **Genomsnittlig betygspoäng för provbetyg (grön) och lärarbetyg (gul), samtliga kursprov vt 2014.**



Olika avvikelser för olika kurser på nationell nivå

Tabell 6 visar avvikelserna på nationell nivå för proven vårterminen 2014.[46] Vi går inte närmare in på detaljer i tabellen, men kan notera att avvikelserna varierar mellan olika prov. Det man också kan notera är att andelen elever med lika lärarbetyg och provbetyg är lägre i tabell 6 än i tabell 4, vilket kan ses som en logisk följd av att det finns fler betygssteg och att varje nytt betygssteg innebär en ny möjlighet till avvikelse. Värdena varierar också mer i tabell 6 – från 52 procent lika för svenska 3 till 84 procent lika för matematik 1c.

Man kan notera att avvikelser som är större än ett betygssteg är ovanliga och att höjning med ett betygssteg (+ 1) är klart vanligast (förutom Lika). Dock är andelen med höjning + 2 större än i den

[46] Ur Skolverket (2015b).

tidigare fyrgradiga skalan, vilket är en följd av att fler betygssteg ger kortare avstånd mellan betygsgränserna.

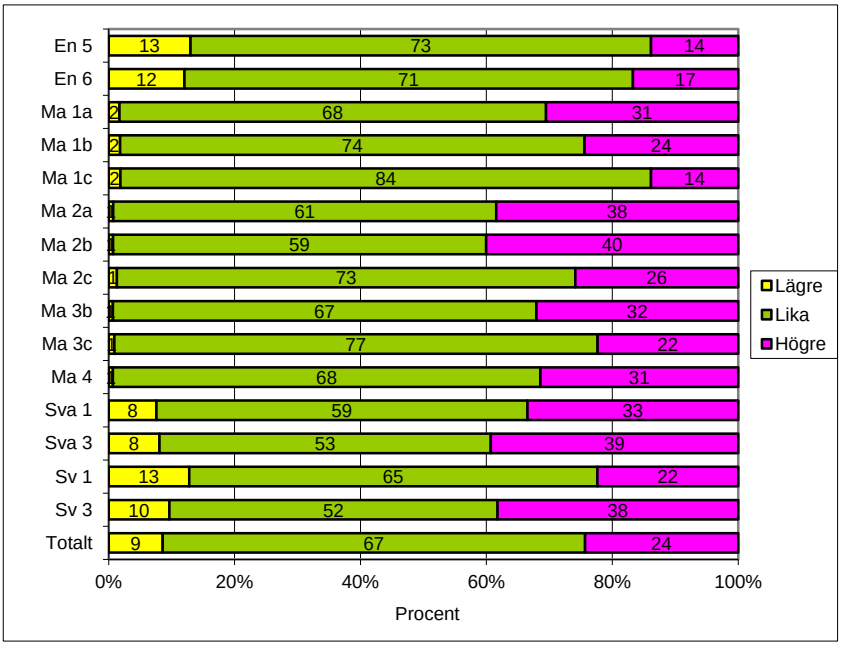
Tabell 6 **Antal elever som genomfört olika kursprov vt 2014 samt avvikelser mellan provbetyg och kursbetyg uttryckt i betygssteg.**

Gy 2011	Antal	Antal betygssteg										
Kurs	elever	-5	-4	-3	-2	-1	Lika	+1	+2	+3	+4	+5
Engelska 5	69 727	0,2	0,4	0,6	1,1	12,0	73,2	14,6	0,8	0,1	0,1	0,0
Engelska 6	55 160	0,4	0,3	0,8	1,3	10,6	71,3	17,4	1,3	0,2	0,1	.
Matematik 1A	21 909	.	0,4	0,3	0,3	2,2	67,6	27,8	2,8	0,3	0,0	.
Matematik 1B	28 045	.	0,1	0,0	0,2	2,1	73,7	23,5	1,9	0,1	0,0	.
Matematik 1C	4 821	0,1	.	0,1	0,2	1,8	84,2	15,4	0,8	0,1	.	.
Matematik 2A	2 298	.	.	.	0,2	1,1	61,3	34,0	3,9	0,6	.	.
Matematik 2B	22 322	.	.	0,1	0,1	1,1	59,2	34,1	5,4	0,8	0,2	0,1
Matematik 2C	9 516	.	0,1	0,0	0,1	1,5	72,7	25,7	2,4	0,4	0,1	0,1
Matematik 3B	6 657	.	.	.	0,1	0,9	67,2	29,4	3,1	0,5	0,1	0,1
Matematik 3C	8 285	.	.	0,1	0,2	0,9	76,9	22,1	1,9	0,3	0,2	.
Matematik 4	6 730	.	0,2	0,1	0,1	0,6	67,1	30,9	3,4	0,9	0,2	0,1
Svenska 1	63 911	0,2	0,3	0,4	1,4	12,0	64,9	21,2	2,8	0,5	0,1	.
Svenska som andraspr. 1	4 022	.	0,4	0,6	2,0	8,0	58,7	29,1	4,6	0,6	0,2	.
Svenska 3	45 623	0,1	0,2	0,3	1,2	9,4	52,1	33,3	8,4	1,8	0,4	0,1
Svenska som andraspr. 3	1 901	.	0,7	0,3	1,0	9,0	52,6	32,5	8,1	1,6	0,2	0,3

Tabellen är något svåröverskådlig; en tydligare bild ges i figur 29.[47]

[47] Här uttrycks avvikelserna på den tregradiga skalan (Diff_1).

Figur 29 **Andel elever med lägre, lika eller högre lärarbetyg än provbetyg för alla kurser vt 2014.**

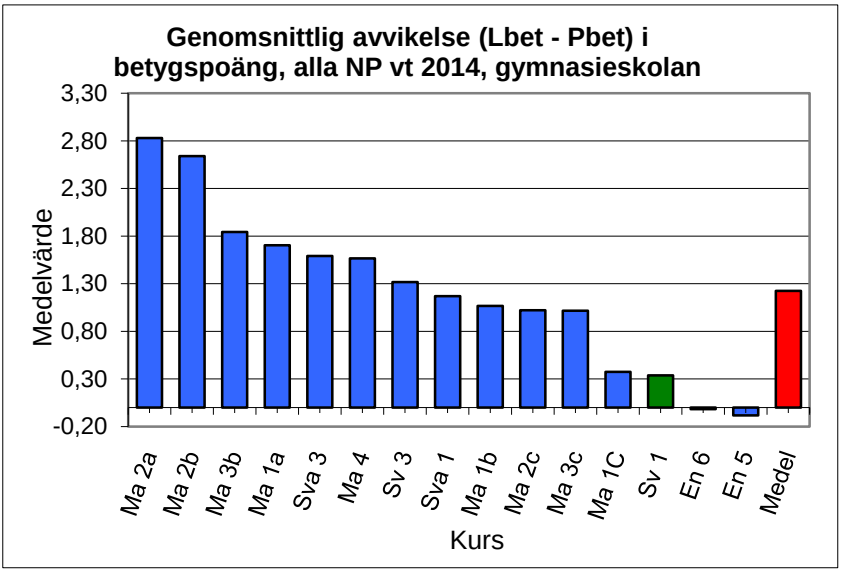


Av figur 29 framgår matematikämnets särställning tydligt, där en mycket liten andel elever får ett lägre lärarbetyg än provbetyg. Däremot är det omvända vanligt.

I figur 30 redovisas skillnaderna mellan genomsnittlig betygspoäng för kursbetyg respektive provbetyg (dvs. medelvärdesskillnaden Diff_20 med den terminologi som används här[48]) för samtliga kursprov. Alla provdeltagare med provbetyg och kursbetyg ingår i underlaget som således anger den genomsnittliga avvikelsen på nationell nivå för respektive program. Antalet elever i respektive prov anges i tabell 7.

[48] Diff_20 = genomsnittlig betygspoäng för lärarbetygen – genomsnittlig betygspoäng för provbetygen.

Figur 30 Genomsnittlig avvikelse (Diff_20) för samtliga elever som deltagit i kursprov och fått betyg vt 2014.



Av figur 30 framgår att avvikelsen mellan betygspoängens medelvärden i svenska 1 (grön stapel) är förhållandevis liten i relation till övriga kurser. Det framgår också att både engelska 5 och 6 har mycket liten total avvikelse, och liksom i grundskolan är avvikelsen svagt negativ. De största avvikelserna kan man notera för olika kurser i matematik.

Tabell 7 Antal elever som har både kursbetyg och provbetyg vt 2014.

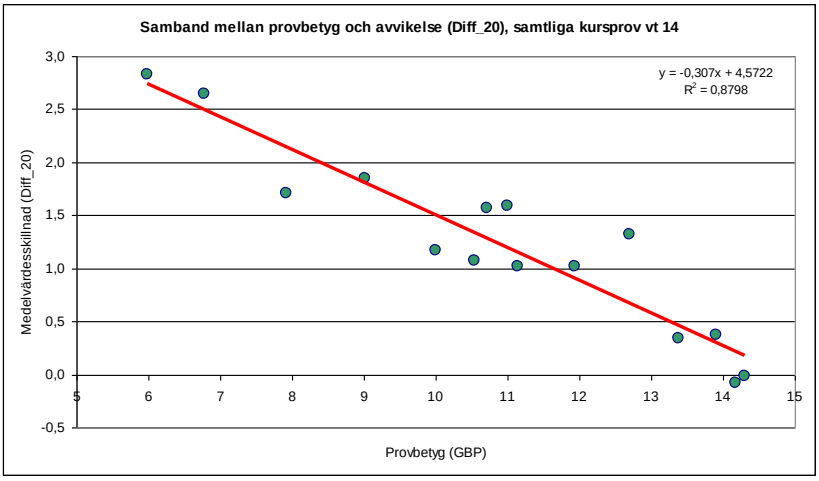
Kurs	Betyg
Ma 2a	2 368
Ma 2b	22 531
Ma 3b	6 864
Ma 1a	22 222
Sva 3	1 970
Ma 4	7 123
Sv 3	46 792
Sva 1	4 149
Ma 1b	28 288
Ma 2c	9 981
Ma 3c	8 719
Ma 1c	5 277
Sv 1	64 869
En 6	55 821
En 5	70 741

Den största avvikelsen har matematik 2a men den kursen har förhållandevis få elever vårterminen 2014 (se tabell 7). Därför redovisas i stället resultat från kurs 2b lite mer i detalj senare. Som kontrast redovisas också engelska 5 som har den lägsta avvikelsen. Skillnaden mellan resultaten för grupper och enheter är förhållandevis lika i övrigt, så därför nöjer vi oss med att undersöka resultaten på enhetsnivå för svenska 1, matematik 2b och engelska 5.

Olika avvikelser för olika betygsnivåer

Om man relaterar skillnaden (Diff_20) mellan kursbetyg och provbetyg till den genomsnittliga betygspoängen (GBP) för provbetygen kan man se nedanstående samband (figur 31). Som figuren visar finns det ett mycket tydligt samband mellan avvikelse och betygsnivå. Låga genomsnittliga provbetyg för en kurs medför således klart större genomsnittlig avvikelse.

Figur 31 Samband mellan provbetyg och avvikelse, samtliga kursprov vt 2014.



Sammanfattningsvis kan man konstatera att det finns ett tydligt samband mellan betygsnivå och avvikelse och att det främst är elever och elevgrupper i kurser med låga provbetyg som får högre lärarbetyg än provbetyg. Detta är viktigt att ha i åtanke vid bedömningen av vilka konsekvenser ändrade anvisningar för relationen mellan provbetyg och lärarbetyg kan få.

Avvikelser på programnivå, skolenhetsnivå och gruppnivå för några kurser

Det finns många kurser att välja på för att redovisa avvikelser för olika skolenheter och grupper. Av utrymmesskäl görs en mer ingående redovisning av resultaten endast för svenska 1. Därefter görs något mindre ingående redovisningar för engelska 5 och matematik 2b. Kursen svenska 1 är vald eftersom ämnet svenska används som modell i redovisningen av grundskolans resultat. Engelska 5 väljs för att samtliga elever deltar och för att avvikelserna är mycket små. Matematik 2b väljs för att det provet har den näst största avvikelsen och genomförs av en förhållandevis stor grupp elever på vårterminen. För de två sistnämnda ämnena redovisas dock resultat endast för skolenheter eftersom skillnaden i

resultat mellan skolenheter och grupper är tämligen marginella. Mönstren sammanfaller i huvudsak.

Svenska 1

Svenska 1 är en så kallad gymnasiegemensam kurs på 100 poäng. Den ingår i alla program, dvs. den genomförs av alla elever.[49] Alla elever gör också det nationella kursprovet i svenska 1. Vissa skolenheter eller program kan genomföra kursen på en termin medan andra kan välja att göra den på två terminer. Den samlade statistik över provresultat som redovisas i Siris anger endast betygens fördelning på program, inte vilken årskurs eleverna går. Därmed är det inte möjligt att koppla slutgiltigt kursbetyg till motsvarande provbetyg.[50]

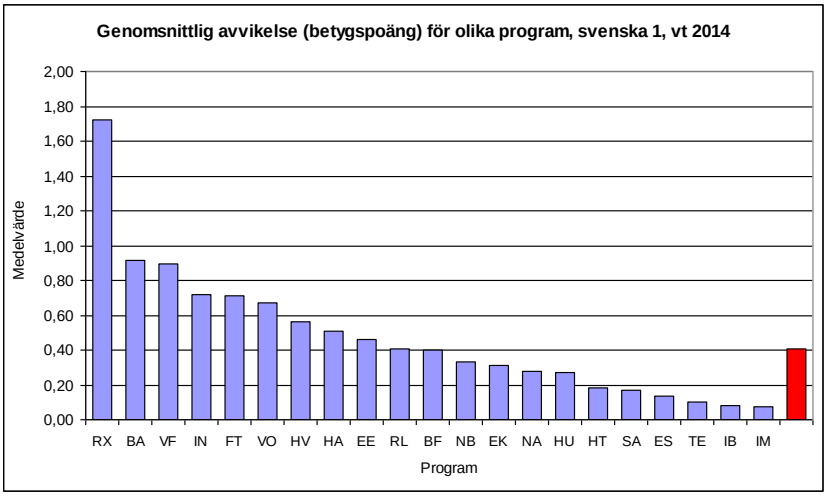
När det gäller de resultat som redovisas här baseras de på de provbetyg och lärarbetyg som rapporteras från skolorna efter provgenomförandet. Det innebär att alla resultat baseras på resultat från samma provversion.

Avvikelse på programnivå

Ytterligare ett sätt att jämföra avvikelser är att utgå från de olika programmen. Eftersom svenska 1 är en gemensam kurs för alla program finns det förhållandevis många elever i varje program som gör provet (se tabell 6 ovan). Figur 32 visar avvikelsen som skillnaden mellan genomsnittlig betygspoäng för lärarbetyg och provbetyg (Diff_20) för de olika program som genomfört prov och fått betyg i svenska 1.

[49] Jämför tabell 7.
[50] Dock redovisas nettoavvikelser i Siris, men inte heller de ger några detaljer om vilka terminer proven genomförts. Se http://siris.skolverket.se/siris/ris.kursprov.kursprov_slutbetyg för närmare information.

Figur 32 **Genomsnittlig avvikelse (Lbet – Pbet) i betygspoäng (Diff_20) för olika program, svenska 1, vt 2014.**

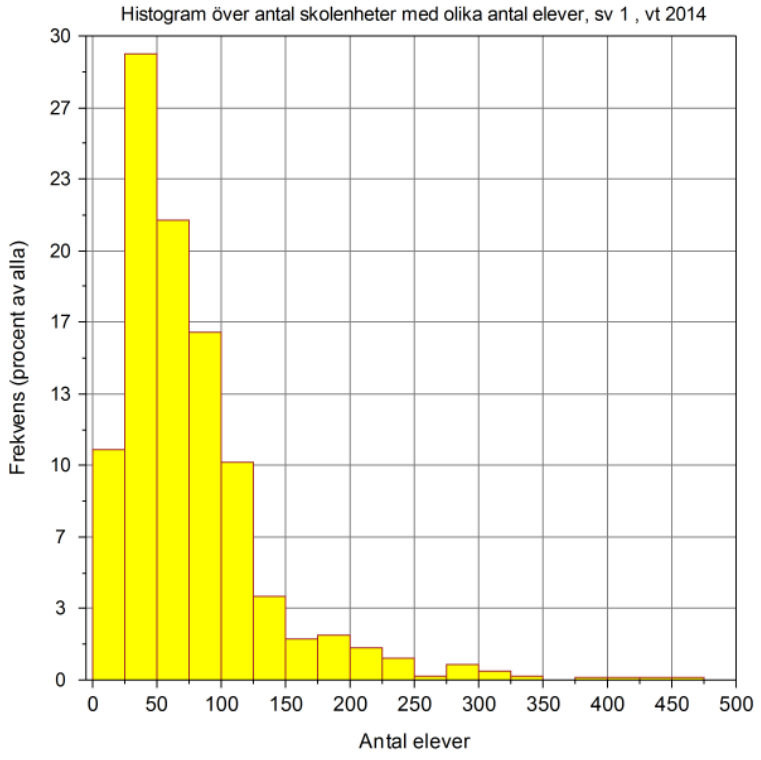


Medelvärdet för de olika programmen är 0,41 betygspoäng (röd stapel) och standardavvikelsen 0,38 betygspoäng. Man kan konstatera att skillnaden i genomsnittlig avvikelse mellan olika program är påtaglig, trots att det gäller samma prov och kurs. Samtidigt är avvikelsen i genomsnittlig betygspoäng svår att värdera.

Avvikelse på skolenhetsnivå

Figur 33 visar fördelningen av antal elever på skolenheter.

Figur 33 **Skolenheter (totalt 996 enheter med minst 15 elever) med olika antal elever som deltagit i provet i svenska 1 vt 2014.**

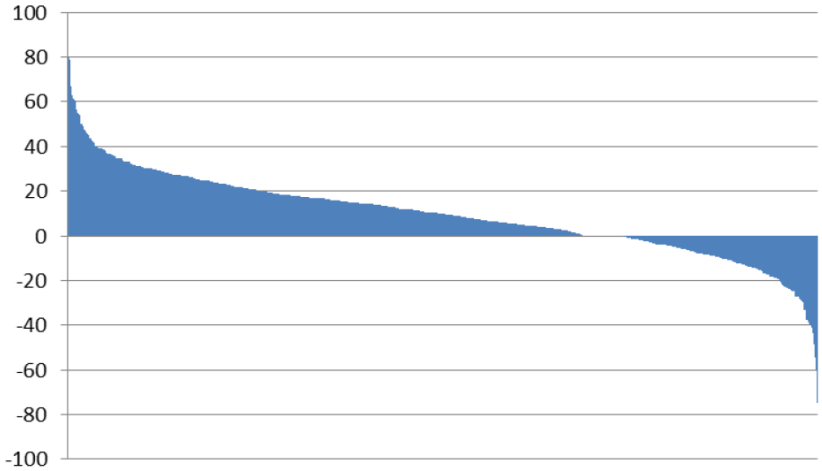


På knappt 30 procent av skolenheterna gör mellan 25–50 elever provet i svenska 1. Drygt 10 procent av de totalt 1 099 enheterna har färre än 25 elever medan närmare 70 procent har 25–100 elever. Medelvärdet är 77 elever (median 60) och standardavvikelsen är 60 elever.

Figur 34 återger nettoavvikelser (Diff_1) för svenska 1, vårterminen 2014 (gäller 1 009 av 1 099 skolenheter med minst 15 elever) så som den redovisas av Skolverket.[51]

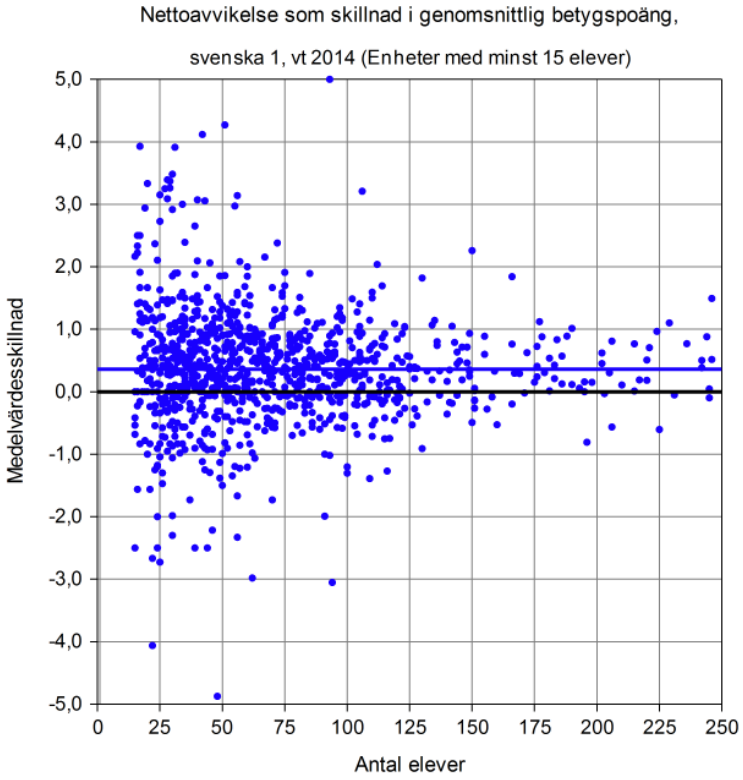
[51] Ur Skolverket (2015b).

Figur 34 Nettoavvikelse (Diff_1) för skolenheter med minst 15 elever, svenska 1, vt 2014.



Om figur 34 finns inte så mycket att säga. Mönstret ser ut som det har gjort tidigare. Figur 35 är däremot mer intressant. Där redovisas avvikelsen mellan lärarbetyg och provbetyg som skillnaden i skolenhetens genomsnittliga betygspoäng för lärarbetyg och provbetyg (Diff_20). Detta har visserligen ingen nämnvärd betydelse för mönstret i bilden jämfört med tidigare sätt att redovisa (Diff_1). Anledningen till att medelvärdesskillnaden baserad på 20-skalan används är att det är den skala som dels gäller enligt lagar och andra bestämmelser, dels används av Skolverket för såväl betygsredovisning som meritvärden.

Figur 35 **Nettoavvikelse (Diff_20) för skolenheter med minst 15 elever, svenska 1, vt 2014. Den blå linjen anger medelvärdet för skolenheterna.**



Avvikelsen räknas alltså här som skillnaden i enhetens genomsnittliga betygspoäng för lärarbetyg respektive provbetyg (medelvärdesskillnad). Den genomsnittliga avvikelsen för de skolenheter som visas i figuren är 0,36 betygspoäng och standardavvikelsen är 1,0 betygspoäng. Cirka 68 procent av enheterna har således avvikelser som ligger inom intervallet ett steg över respektive under medelvärdet för samtliga grupper (den blå linjen i figuren), dvs. de har genomsnittliga medelvärdesskillnader som ligger mellan - 0.64 och 1,36 betygspoäng.[52] Trots att skillnaden (avvikelsen) i genom-

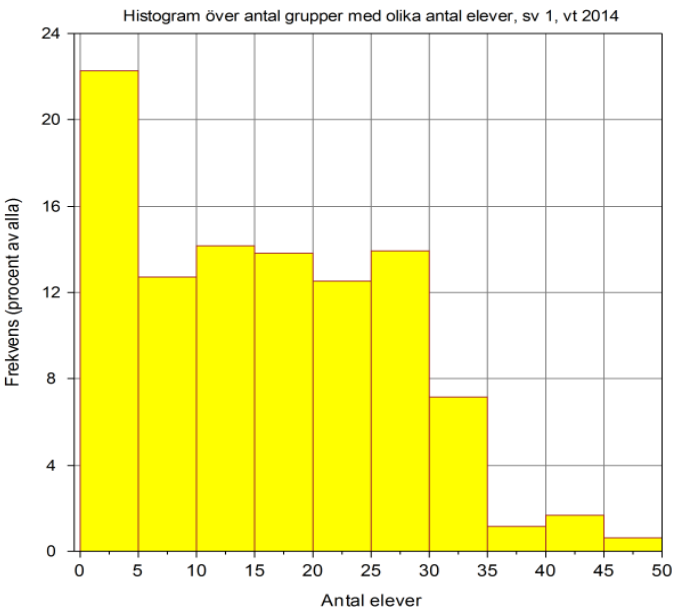
[52] I figur 36 syns inte det tydliga mönster som framträtt i en del av de tidigare figurerna. Det beror på att dessa har redovisat Diff_1 där avvikelser endast kan ha värdena + 1 eller - 1. I figur 36 räknas däremot avvikelse i betygspoäng, vilket innebär att en avvikelse på ett

snittlig betygspoäng är relativt liten enligt figur 30 ovan (0,36 betygspoäng på en skala från 0 till 20) så är den tämligen stora spridningen mellan olika skolenheters avvikelser problematisk. Det man också kan notera i figur 35 är det tydliga sambandet mellan avvikelse och gruppstorlek som vi tidigare konstaterat.

Avvikelse på gruppnivå

Gruppindelningen i svenska 1 (figur 36) har ett annat utseende än gruppindelningen i svenska i årskurs 9 (figur 13). Där var den vanligaste gruppstorleken 20–25 elever. För svenska 1 anges något överraskande 1–5 elever vara vanligast (cirka 23 procent av grupperna). I övrigt är fördelningen tämligen jämn mellan grupper av olika storlek.

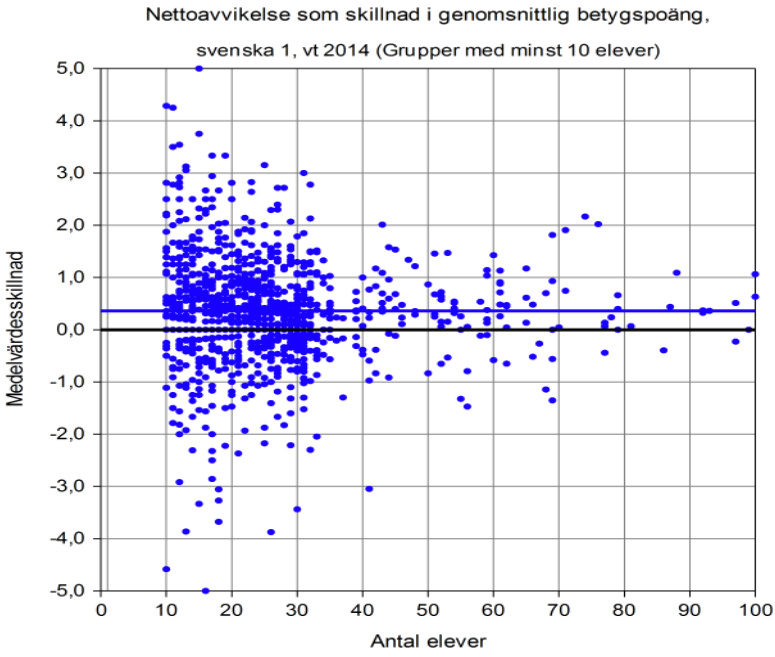
Figur 36 Fördelning av grupper efter antal elever, svenska 1, vt 2014.



betygssteg kan ge 10 betygspoäng (mellan betygen F och E) eller 2,5 betygspoäng (för övriga betygssteg). Detta gör att skillnader mellan betygsmedelvärden blir olika för samma andel avvikelser beroende på vilka betygssteg avvikelserna gäller. Därmed grupperar sig inte avvikelserna systematiskt i figur 36 på samma sätt som i tidigare figurer, även om det skulle gälla samma elever och samma prov.

Figur 37 visar avvikelsen för de 1 175 grupper som har 10–50 elever.

Figur 37 **(Diff_20) för grupper uppdelat efter gruppstorlek (minst 10 elever), svenska 1, vt 2014.**



Figur 37 visar samma sak som tidigare figurer på samma tema, dvs. att avvikelsen beror på gruppstorleken. För svenska 1 är medelvärdet av gruppernas avvikelse 0,35 betygspoäng och standardavvikelsen 1,1 betygspoäng, dvs. i stort sett samma värden som för skolenheterna.

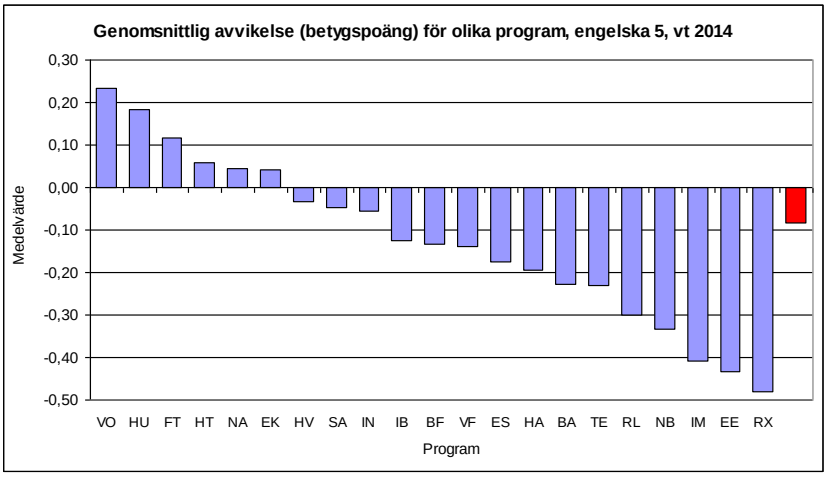
Nedan följer i lite mer sammanfattande form en genomgång av avvikelserna för några ytterligare kursprov i gymnasieskolan.

Engelska 5

Engelska 5 är en gymnasiegemensam kurs på 100 poäng för alla program. Det innebär att samtliga elever genomför provet, även om det kan göras under olika terminer eller årskurser.

Avvikelse på programnivå

Figur 38 Avvikelse (Lbet – Pbet) i betygspoäng (Diff_20) för olika program, engelska 5, vt 2014.

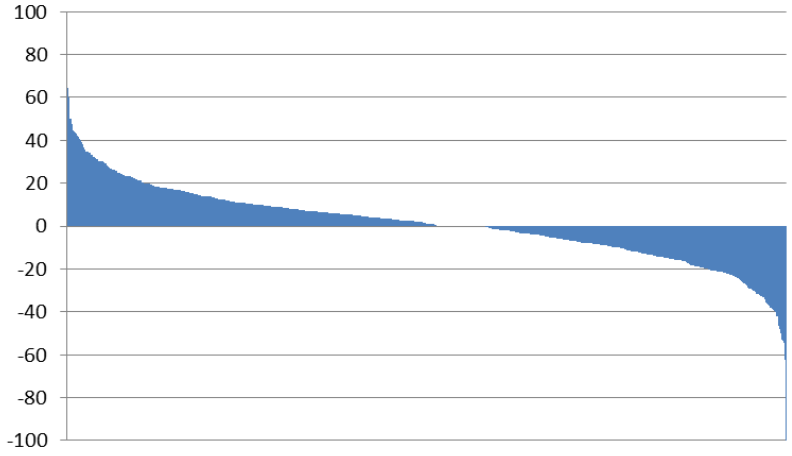


För engelska 5 är medelvärdet av programmens avvikelser - 0,13 betygspoäng och standardavvikelsen 0,18 betygspoäng. Spridningen i avvikelse mellan program är således betydligt lägre i engelska 5 än i svenska 1, vårterminen 2014 (0,18 mot 0,38 betygspoäng för svenska 1).

Avvikelse på skolenhetsnivå

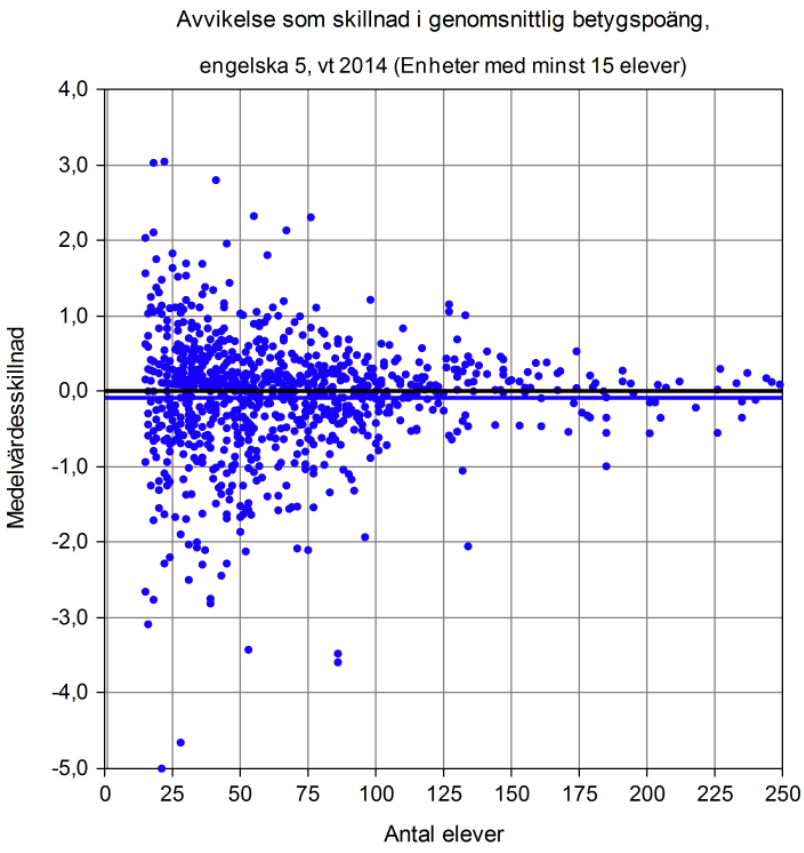
Skolverket redovisar nedanstående nettoavvikelser (Diff_1) för de 1 029 av 1 121 skolenheter som har minst 15 elever i engelska 5.

Figur 39 **Nettoavvikelse (Diff_1) för skolenheter med minst 15 elever, engelska 5, vt 2014.**



Figur 39 stämmer väl överens med den gängse bilden av avvikelse på skolenhetsnivå i engelska. Figur 40 visar hur avvikelserna beror på enheternas storlek.

Figur 40 Nettoavvikelse (Diff_20) för skolenheter med minst 15 elever, engelska 5, vt 2014.



Man kan notera att för engelska 5 ligger medelvärdet av skolenheternas skillnader i den genomsnittliga betygspoängen för provbetyg och lärarbetyg nära varandra. Däremot finns det i enlighet med det generella mönstret en inte obetydlig spridning mellan i synnerhet mindre enheter.

Matematik 2b

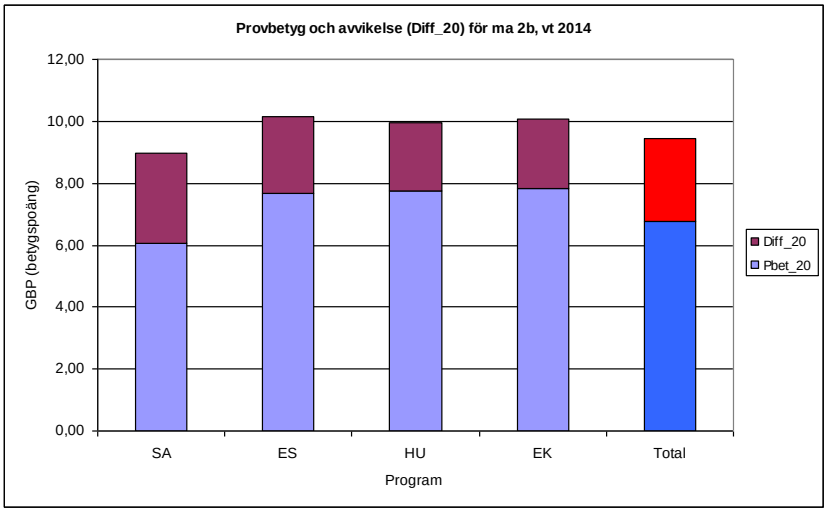
Matematik 2b är en gymnasiegemensam kurs på de högskoleförberedande programmen Ekonomiprogrammet (EK) och Samhällsvetenskapsprogrammet (SA), men kursen läses också av elever

från en del andra program, framför allt Estetiska programmet (ES) och delvis Humanistiska programmet (HU), se tabell 5 ovan.

Avvikelse på programnivå

Betygspoängen och avvikelsen för de få program som kursprovet i första hand gäller framgår av figur 41.

Figur 41 **Genomsnittlig betygspoäng för provbetyg (Pbet) och avvikelse (Diff_20) uppdelat på program, matematik 2b, vt 2014.**



Man kan notera att betygsnivån är låg. GBP ligger för provbetygen klart under 10, vilket är betygspoängen för betyget E.

Samhällsvetenskapsprogrammet, som har flest elever, har den lägsta genomsnittliga betygspoängen (6) och också den största avvikelsen. Övriga program har mer likartade genomsnittliga betygspoäng. Betygsnivån är dock generellt mycket låg, vilket innebär en stor genomsnittlig avvikelse på 2,6 betygspoäng. Som tidigare nämnts är dock betygspoängen svår att tolka.

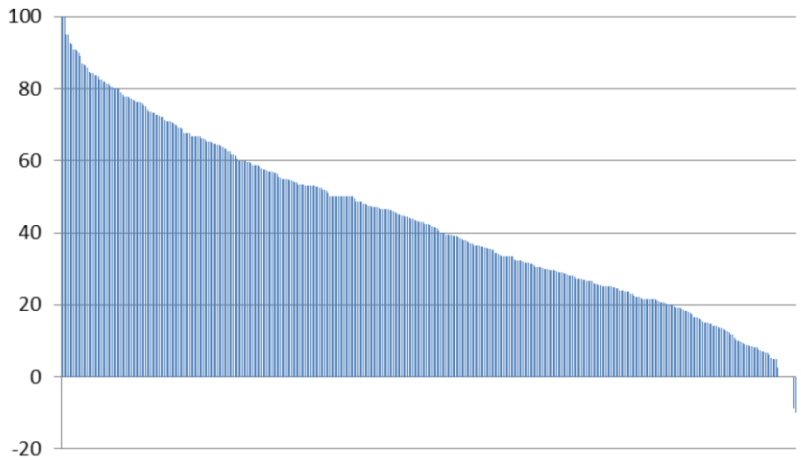
Uttryckt på annat sätt innebär det för det aktuella provet en nettoavvikelse (Diff_1) på 39 procent, dvs. 39 procent av eleverna

hade ett högre lärarbetyg än provbetyg,[53] och uttryckt i betygssteg (Diff_6) betyder det att i genomsnitt för varje elev är lärarbetyget nästan ett halvt betygssteg (0,46 steg) högre än provbetyget. Skillnaden mellan de olika program där provet görs är måttlig i den meningen att de olika programmen har ungefär samma genomsnittliga avvikelse på provet.

Avvikelse på skolenhetsnivå

Skolverket redovisar nedanstående avvikelser för skolenheter vårterminen 2014 i matematik 2b.[54]

Figur 42 **Nettoavvikelser (Diff_1) för skolenheter med minst 15 elever (418 av 515 enheter), matematik 2b, vt 2014.**

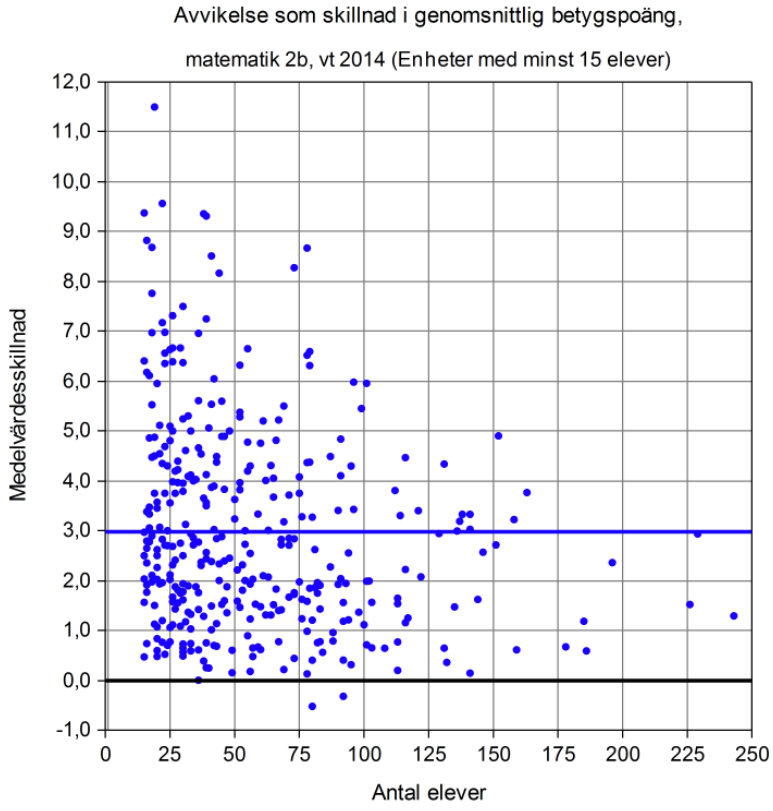


Figur 42 illustrerar den stora avvikelsen. Medelvärdet av nettoavvikelserna för skolenheterna ligger på 43 procent och standardavvikelsen är 24 procent. Här har en enhet i genomsnitt 47 elever som gjort provet i matematik 2b.

Om vi också redovisar resultatet så att avvikelserna relateras till enheternas storlek får vi fram nedanstående bild (figur 43).

[53] För betygsstegen E till A motsvarar 2,6 betygspoäng mer än ett betygssteg. Om det gäller betygssteget F till E räcker det bara till drygt en fjärdedels betygssteg.
[54] Skolverket (2015b).

Figur 43 **Nettoavvikelse (Diff_20) för skolenheter med minst 15 elever, matematik 2b, vt 2014.**



Här kan man konstatera en stor spridning i avvikelsen mellan olika skolenheter (standardavvikelse 2,1 betygspoäng) samt en stor skillnad mellan medelvärdet för enheternas genomsnittliga lärarbetyg och medelvärdena av den genomsnittliga betygspoängen för motsvarande provbetyg. Skillnaden anges av den blå medelvärdeslinjen, dvs. den är cirka 3 betygspoäng.[55]

[55] Observera att det här gäller medelvärdet av enheternas avvikelser, där *varje enhet* har samma vikt oberoende av hur många elever den har. När man beräknar avvikelsen på nationell nivå räknas varje individ lika. Det innebär att de stora enheterna väger tyngre än de små. Av figuren framgår att de stora enheterna i allmänhet har lägre värden än de små. Detta gör att den genomsnittliga avvikelsen blir mindre på nationell nivå (2,6 betygspoäng) än när den beräknas som medelvärdet av enheternas avvikelser (3,0 betygspoäng).

Kommentar

Det här avsnittet innehåller inga jämförelser över tid, vilket inte beror på att underlag saknas. Skolverket har på sin webbplats publicerat ett antal årliga rapporter där avvikelser redovisas. Skälet är snarast att bilden är densamma som den som har redovisats för grundskolan, dvs. att mönstret över tid är mycket likartat.

Relationen mellan lärarbetyg och provbetyg (avvikelsen) är mycket olika mellan olika kurser, där den är minst i engelska och störst i vissa matematikkurser. Betygsnivån är högst i engelska och lägst i matematik. Även detta mönster känns igen från grundskolan.

Ett ytterligare mönster som känns igen är att spridningen i avvikelse är avsevärd på alla nivåer: programnivå, skolenhetsnivå och gruppnivå. Ju färre elever som ingår i grupperna, desto större blir avvikelserna. Det finns alltså en tydlig statistisk effekt i mönstren.

På samma sätt som för grundskolan skulle man kunna sammanfatta analysen av relationen mellan provbetyg och lärarbetyg för gymnasieskolan med att det finns ett stabilt mönster av instabilitet. Detta är dessutom snarast förstärkt i relation till grundskolan.

I nästa avsnitt ska vi granska vissa av provresultaten och avvikelserna ur ett mer statistiskt perspektiv.

Slumpens roll för avvikelsen

Vi tänker oss ett scenario där en lärare har genomfört ett prov och enligt bedömningsanvisningarna fått en viss fördelning av provbetyg. Vidare har läraren utifrån sitt samlade betygsunderlag satt de betyg hon eller han anser att eleverna visat kunskaper för. Om nu läraren vill jämföra nivån på sina betyg med nivån på elevernas provbetyg är det förstås inte alldeles uppenbart hur detta kan eller bör göras. Som visats tidigare kan det ske via nettoavvikelser (Diff_1), via genomsnittlig avvikelse i betygssteg (Diff_6) eller via skillnad i genomsnittlig betygspoäng för klassens provbetyg och lärarbetyg (Diff_20).

För den här redovisningen väljer vi Diff_1 för grundskolan och Diff_6 eller Diff_20 för gymnasieskolan. Dock väljer vi att inte försöka undersöka skillnader på lärarnivå, eftersom det skulle innebära att granskningen görs på klassnivå (gruppnivå). Många grupper i framför allt gymnasieskolan är som vi sett små och av statistiken

framgår inte i vilken utsträckning samma lärare har elever från olika program i samma klass. Vi nöjer oss därför med att jämföra resultat på skolenhetsnivå, oberoende av om en eller flera lärare på enheten haft grupper som genomfört det aktuella provet.

Granskningen innebär att vi jämför provbetyg och lärarbetyg för att se om de skiljer sig åt på ett statistiskt signifikant sätt för de olika skolenheter som deltagit i provet och som har minst 15 elever.[56] Det vill säga kan man eller kan man inte med viss statistisk säkerhet säga att provbetyg och kursbetyg för en skolenhet skiljer sig så mycket åt att man kan säga att lärarens eller lärarnas betyg för eleverna i den aktuella gruppen avviker signifikant från motsvarande elevers provbetyg.

Som metod används s.k. parvisa t-test eftersom samma elever och elevgrupper har ett provbetyg och ett lärarbetyg. Signifikansnivån är satt till 95 procent.

Grundskolan

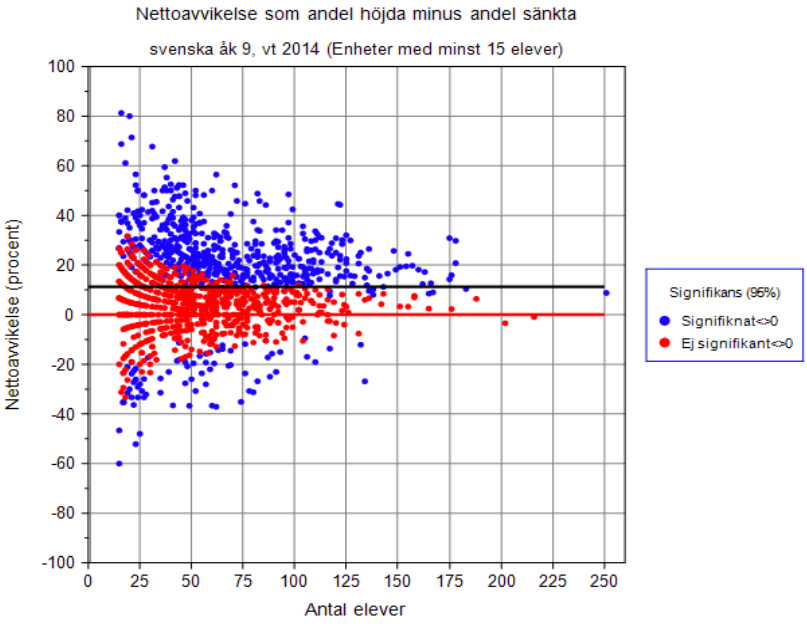
Svenska i årskurs 9

Figur 44 visar de skolenheter som genomfört provet i svenska i årskurs 9 vårterminen 2014. Observera att figuren visar genomsnittlig nettoavvikelse, dvs. den baseras på Diff_1.

Den röda linjen markerar nettoavvikelsen noll och den svarta den genomsnittliga nettoavvikelsen för samtliga skolenheter (medelvärdet för avvikelsen är 11 procent och standardavvikelsen 17 procent). För de rödmarkerade enheterna skiljer sig inte den genomsnittliga avvikelsen signifikant från noll (95 procents nivå). De blåmarkerade enheterna har däremot en nettoavvikelse som är signifikant skild från noll.

[56] Vi kunde ha valt mindre grupper, t.ex. minst tio elever. Spridningen skulle då ha blivit större eftersom ännu mindre grupper tillkommit. De värden på spridning som anges kan därför ses som underskattningar.

Figur 44 **Nettoavvikelse (Diff_1) för skolenheter med minst 15 elever. Blå enheter har nettoavvikelse signifikant skild från noll, svenska åk 9, vt 2014.**



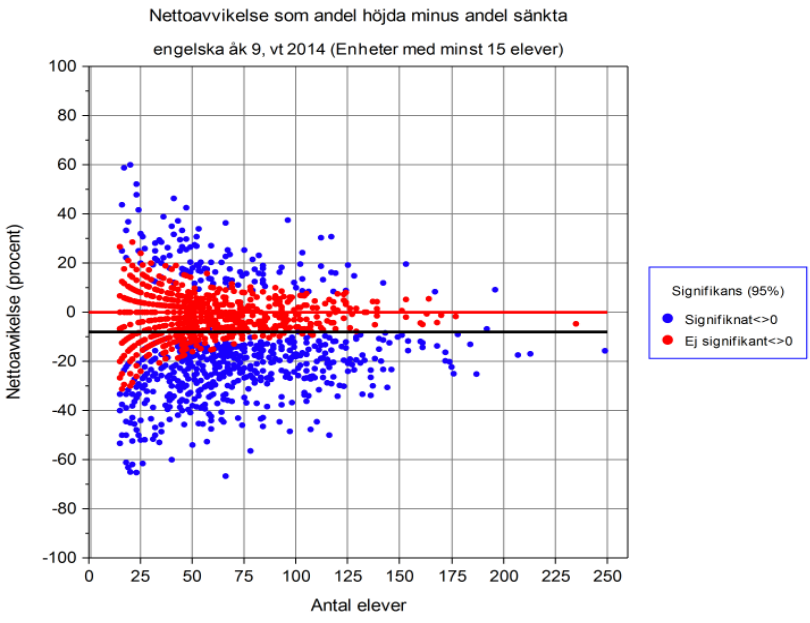
Figuren visar att det finns en systematisk avvikelse som markerar skillnaden mellan lärarkollektivets samlade betygsbedömning och provbetygets genomsnittliga nivå, dvs. provkonstruktörernas bedömning. Figuren visar också att för små grupper kan avvikelserna vara större utan att för den skull vara signifikant skilda från noll.[57]

Engelska i årskurs 9

Figur 45 visar utfallet för engelska i årskurs 9 vårterminen 2014.

[57] Man kan notera att gränsen mellan signifikant och icke signifikant avvikelse inte är skarp. Det beror på att t-testet utgår från varje stickprovs egna värden vid skattning av de parametrar som ingår i t-testet. Detta är den metod som får användas om populationsvärden inte finns tillgängliga, vilket skulle vara fallet vid den tidpunkt betyg ska sättas.

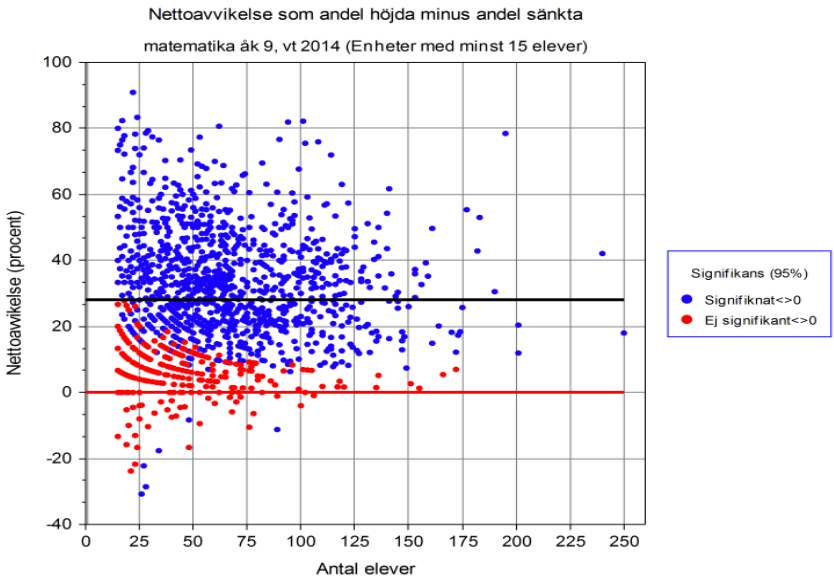
Figur 45 **Nettoavvikelse (Diff_1) för skolenheter med minst 15 elever. Blå enheter har nettoavvikelse signifikant skild från noll, engelska åk 9, vt 2014.**



För engelska är som tidigare noterats avvikelsen negativ, dvs. lärarbetygen (den svarta linjen) är i genomsnitt lägre än provbetygen.

Matematik i årskurs 9

Figur 46 **Nettoavvikelse (Diff_1) för skolenheter med minst 15 elever. Blå enheter har nettoavvikelse signifikant skild från noll, matematik åk 9 vt 2014.**

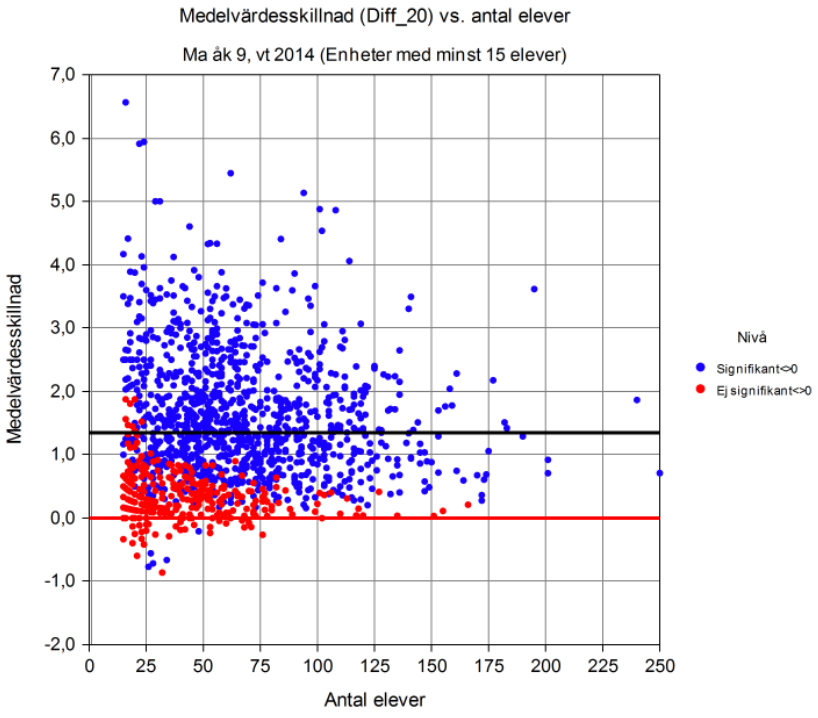


Figur 46 visar att den genomsnittliga nettoavvikelsen för de skolenheter som gjort proven är knappt 30 procent för matematik i årskurs 9 vårterminen 2014 (den svarta linjen). Variationen mellan enheter är också betydande (standardavvikelse 18 procent). Som synes finns det också några enstaka skolenheter som har signifikant lägre lärarbetyg än provbetyg.

I ovanstående figurer har avvikelsen visats som nettoavvikelse (Diff_1), dvs. skillnaden mellan den andel elever (i procent) som har högre lärarbetyg än provbetyg och den andel som har lägre lärarbetyg än provbetyg. Det tas dock ingen hänsyn till om avvikelsen är ett eller flera betygssteg. Ett mer precist och i andra sammanhang mer använt mått är den genomsnittliga betygspoängen, även om den som tidigare nämnts har sina nackdelar genom den ojämna viktningen av olika betyg.

Figur 47 visar avvikelsen för matematik i årskurs 9 vårterminen 2014, dvs. samma grupp som i föregående figur, men uttryckt som skillnad i genomsnittlig betygspoäng (Diff_20).

Figur 47 **Skillnad i genomsnittlig betygspoäng (Diff_20) för skolenheter med minst 15 elever. Blå enheter har medelvärdesskillnad som är signifikant skild från noll, matematik åk 9, vt 2014.**



Medelvärdet i betygsavvikelse för hela gruppen av enheter är 1,3 betygspoäng (enligt skalan 0–20) och standardavvikelsen 1,0 betygspoäng. Man kan notera att den övergripande bilden av fördelningen är tämligen likartad i de båda figurerna, men figur 46 uppvisar det karaktäristiska mönster som syns för små grupper när det gäller nettoavvikelse. När avvikelsen som i figur 47 visas i betygspoäng suddas sådana mönster ut i varierande grad, eftersom avvikelser på olika betygsnivåer ger olika skillnader i betygspoäng (ett steg mellan F och E ger 10 betygspoäng, medan ett steg mellan exempelvis E och D ger 2,5 poäng).

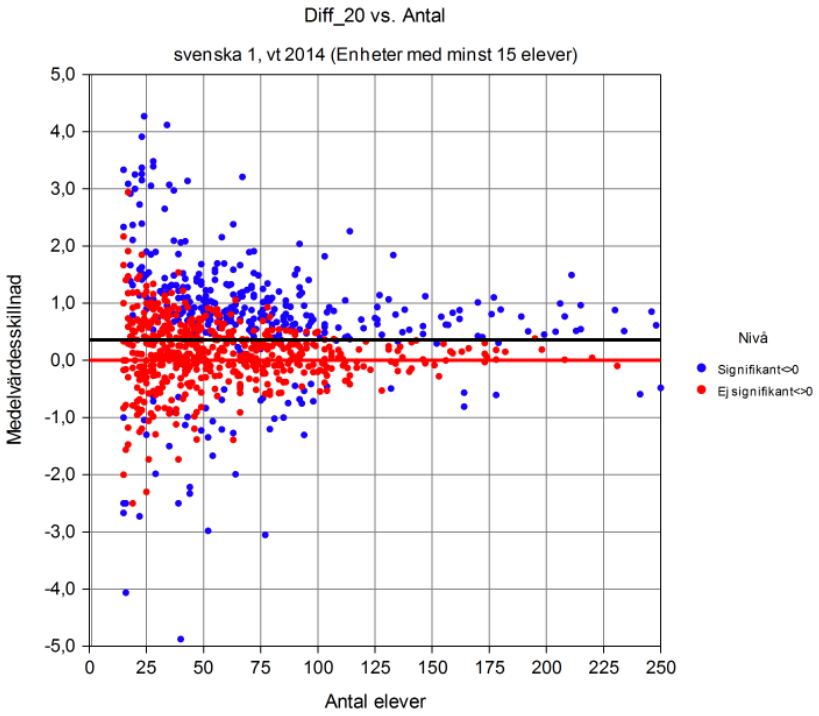
Gymnasieskolan

Även för gymnasieskolan visas skolenheters avvikelsemönster för några kurser. Här visas avvikelserna enligt skalan 0–20 betygspoäng.

Svenska 1

Avvikelsen för svenska 1 vårterminen 2014 visas i figur 48, Medelvärdet för enheterna är 0,35 betygspoäng och standardavvikelsen 0,97.

Figur 48 **Skillnad i genomsnittlig betygspoäng (Diff_20) för skolenheter med minst 15 elever. Blå enheter har medelvärdesskillnad som är signifikant skild från noll, svenska 1, vt 2014.**

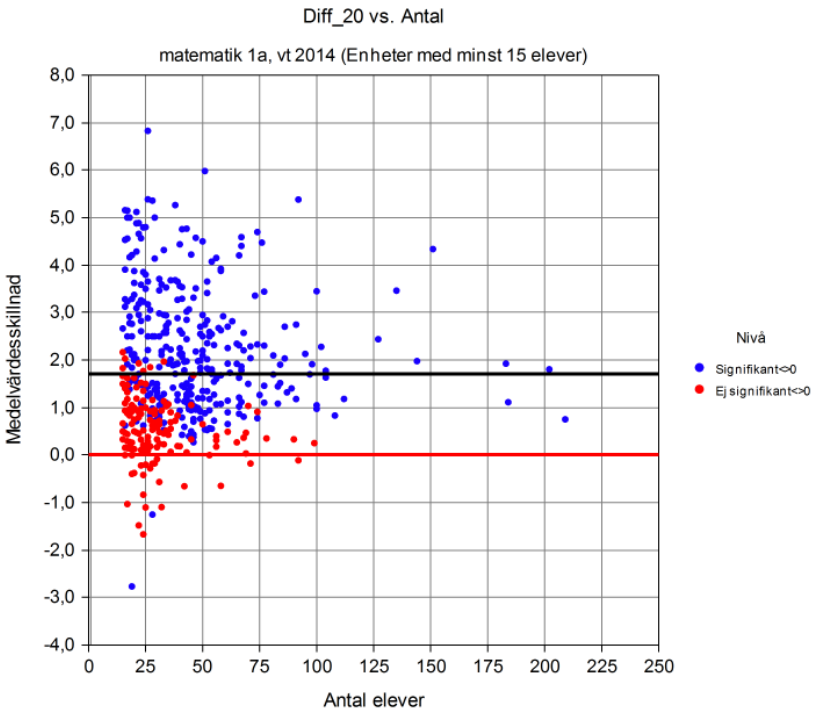


Spridningen är ganska stor och vissa skolenheter visar betydande avvikelser, men den genomsnittliga skillnaden i betygspoäng avviker måttligt från nollinjen.

Matematik 1a

Matematik 1a är en gymnasiegemensam kurs för yrkesprogrammen. Figur 49 visar mönstret för denna kurs. Den genomsnittliga avvikelsen för de ingående enheterna är 1,7 betygspoäng och standardavvikelsen 1,4.

Figur 49 **Skillnad i genomsnittlig betygspoäng (Diff_20) för skolenheter med minst 15 elever. Blå enheter har medelvärdesskillnad som är signifikant skild från noll, matematik 1a, vt 2014.**

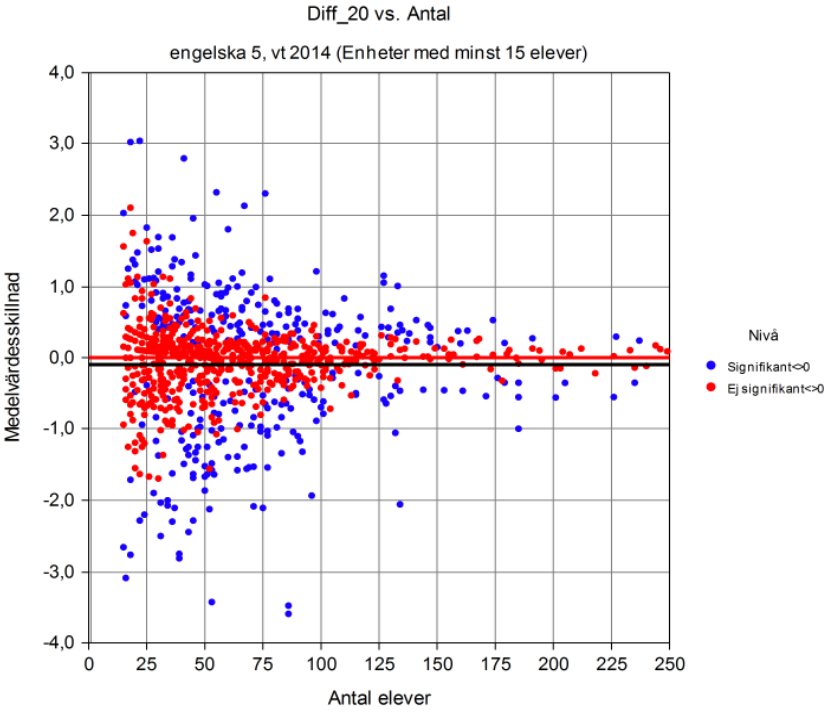


Antalet skolenheter med resultat i matematik 1a är dock mindre än antalet enheter med resultat i svenska 1 och engelska 5, vilket alla elever läser.

Engelska 5

För engelska 5 är den genomsnittliga avvikelsen liten och negativ (- 0,09 betygspoäng). Standardavvikelsen (0,79 betygspoäng) är också mindre än för svenska 1 och i synnerhet mindre än för matematik 1a. Trots detta finns det vissa enheter som sticker ut med stora medelvärdeskillnader mellan provbetyg och lärarbetyg.

Figur 50 **Skillnad i genomsnittlig betygspoäng (Diff_20) för skolenheter med minst 15 elever. Blå enheter har medelvärdesskillnad som är signifikant skild från noll, engelska 5, vt 2014.**

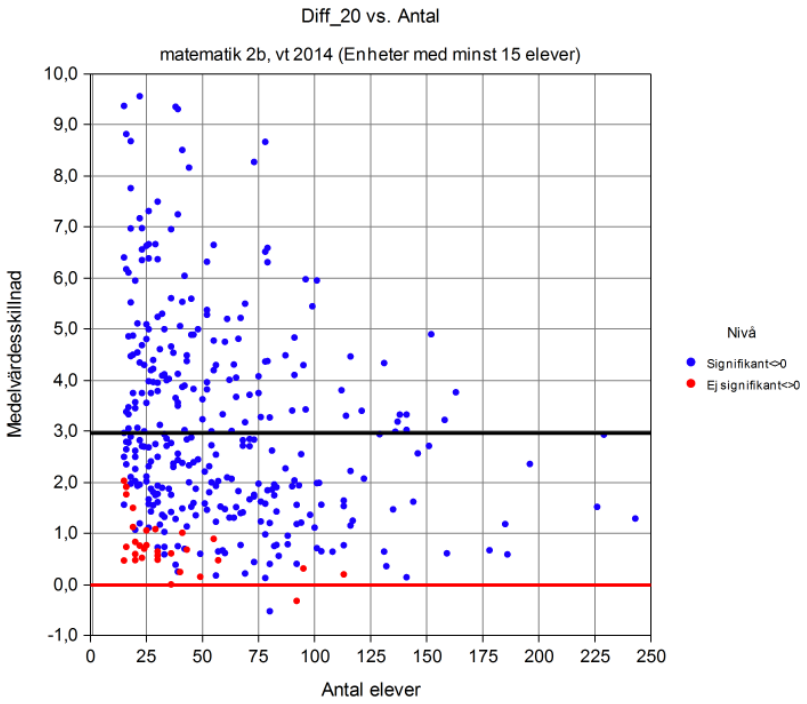


Matematikämnet har många kurser och kursprov och vi visar därför också ett par prov med särskilt stora avvikelser.

Matematik 2b

Medelvärdet av skolenheternas genomsnittliga betygsskillnad mellan lärarbetyg och provbetyg är 3,0 betygspoäng, och spridningen (standardavvikelsen) är 2,1 betygspoäng. Som framgår av figur 51 är fördelningen påtagligt sned. Man kan också notera att nästan alla skolenheter har avvikelser som är signifikant skilda från noll.

Figur 51 **Skillnad i genomsnittlig betygspoäng (Diff_20) för skolenheter med minst 15 elever. Blå enheter har medelvärdesskillnad som är signifikant skild från noll, matematik 2b, vt 2014.**

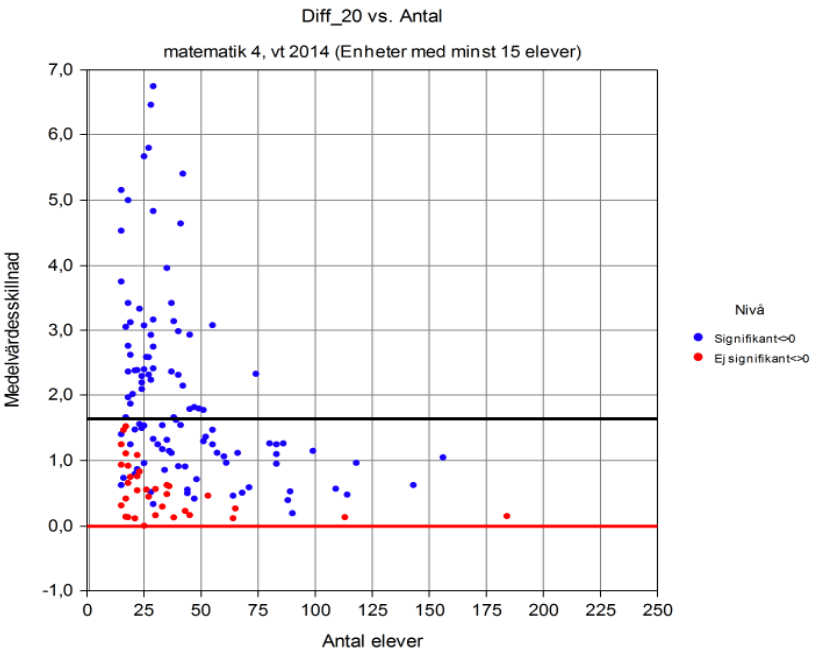


Matematik 4 (figur 52 nedan) är en kurs med förhållandevis få skolenheter – endast 147 stycken har minst 15 elever. Även här är avvikelsen stor; medelvärdet för enheterna är 1,7 betygspoäng och standardavvikelsen mellan enheterna 1,7 betygspoäng.

Trots att matematik 4 enbart gäller vissa elever på Naturvetenskaps- och Teknikprogrammet blir spridningen mellan olika enheter stor och fördelningen är påtagligt sned.

Matematik 4

Figur 52 **Skillnad i genomsnittlig betygspoäng (Diff_20) för skolenheter med minst 15 elever. Blå enheter har medelvärdesskillnad som är signifikant skild från noll, matematik 4, vt 2014.**



Kommentar

I det här avsnittet understryks ytterligare den stora variationen mellan olika ämnen, kurser och program när det gäller betygsnivåer och avvikelser. Det pekar också på att de statistiska förutsättningarna varierar och att variationer i avvikelser finns, oberoende av om stora elevgrupper som hela län eller små klasser undersöks. Den stora utmaningen med en så spretande problembild är att försöka finna en acceptabel modell för alla de olika nationella proven och hur dessa på ett tillförlitligt och likvärdigt sätt ska kunna stödja lärarnas betygssättning.

I nästa avsnitt diskuteras några förutsättningar som gäller för nuvarande prov och betyg och hur dessa förutsättningar harmonierar eller står i strid med några tänkbara modeller för provens betygsstödjande roll.

Modeller för betygsstöd

Det finns många tänkbara modeller för att använda prov som underlag för stöd eller styrning av betyg. För svensk del har vi valt att de nationella proven ska ha en betygsstödjande roll, och denna roll föreslås nu av utredningen att renodlas. Innan vi går in på några idéer om en svensk modell ska vi kort titta på hur man gör i några andra länder.

Modeller i några andra länder

Det finns många olika betygssystem och provsystem i världen. Många länder har examensprov som är avgörande för betygen medan andra ger mer ansvar till de undervisande lärarna. Generellt kan man säga att examensproven har gamla anor medan de examinationer och godkännanden som mer bygger på undervisande lärares bedömningar, ofta i kombination med resultat på vissa nationella prov, är en mer sentida företeelse.

Sverige var tidigt influerat av amerikanska strömningar och kan sägas ha varit en typisk representant för den senare kategorin tillsammans med främst vissa delstater i engelskspråkiga länder som USA, Kanada, Australien, Nya Zeeland och Skottland. Dock har dessa länder ofta också olika former av test eller prov vars resultat kombineras med lärarbetyg och andra meriter vid t.ex. urval till högre utbildning.

Mer traditionella examensländer är de nordiska länderna, förutom Sverige, det kontinentala Europa och flera ostasiatiska länder, t.ex. Kina, Korea och Singapore. Även England har olika former av examinerande prov men har under senare år liksom Sverige hållit ett högt reformtempo och befinner sig i en omställningsfas när det gäller prov och betygssystem på främst gymnasienivå.

Varje land har sin mer eller mindre specifika skolkultur. Därför är det svårt att implementera metoder och synsätt från ett land till ett annat. Åtgärder som varit framgångsrika i ett land behöver inte vara det i ett annat land med en annan skolkultur och andra traditioner. Detta gäller läroplaner, kursplaner, undervisningsmetoder, ledarskap m.m. likväl som prov, betygssättning och examinationssystem. När vi således tämligen översiktligt redovisar några drag från andra länders modeller för betygssättning och nationella prov

(betygsstödjande eller examinerande) ska de inte ses som konkreta förslag utan mer som översiktliga inspel av idéer som kanske i modifierad form kan leda till förbättring också i en svensk kontext.

Vi ska kortfattat beskriva prov- och betygssystemen i några länder och särskilt fokusera på relationen mellan provresultat och betyg. De länder som förefaller särskilt intressanta är Danmark, Finland, Norge och England.

Danmark

När det gäller betygssättning och nationella prov ligger det för svensk del nära till hands att vända sig mot Danmark. I de direktiv som låg till grund för den nuvarande sexgradiga svenska betygsskalan hänvisades till den betygsskala och det betygssystem som infördes i Danmark 2006. Till grund för det nya danska systemet låg en expertutredning.[58]

Den nya danska skalan skulle ha färre betygssteg (sju) än den tidigare så kallade 13-skalan med tio steg. Samma skala skulle användas i hela utbildningssystemet, alltså även i högskolan. Skalan skulle också vara användbar i internationella sammanhang. Därmed låg det nära till hands att koppla den till den ECTS-skala[59] som används inom högskolevärlden. Det som framför allt gör den danska skalan intressant är att den är relativ samtidigt som betygssystemet är mål- och kunskapsrelaterat (kriterierelaterat). Det innebär att på nationell nivå eftersträvas en bestämd procentuell fördelning av betygen medan provkonstruktörer och lärare ska utgå från betygskriterierna när de konstruerar prov och sätter betyg.

Betygsskalan har bestämda kriterier för nivån godkänd. De elever som inte når den gränsen blir inte godkända. Betygen för de elever som når godkändnivån ska sedan på nationell nivå som norm ha nedanstående betygsfördelning.[60] För de icke-godkända betygen F och Fx finns inga procentsatser utan de baseras enbart på kriterier.

[58] Undervisningsministeriet, Uddannelsestyrelsen. (2004). http://pub.uvm.dk/2004/karakterer/karakterer.pdf
[59] European Credit Transfer and Accumulation System.
[60] Denna har samma betygsfördelning som ECTS-skalan.

Figur 53 **Procentuell fördelning av godkända betyg i den danska sjugradiga betygsskalan samt motsvarande betygspoäng.**

Betyg	F	Fx	E	D	C	B	A
Andel	"-"	"-"	10%	25%	30%	25%	10%
Betygspoäng	-3	0	2	4	7	10	12

Till skalan hör också en kortfattad beskrivning av vad som krävs för de olika betygen.[61]

7-trins-skalaen

Karakter	Betegnelse	Beskrivelse	ECTS
12	Den fremragende præstation	Karakteren 12 gives for den fremragende præstation, der demonstrerer udtømmende opfyldelse af fagets mål med ingen eller få uvæsentlige mangler.	A
10	Den fortrinlige præstation	Karakteren 10 gives for den fortrinlige præstation, der demonstrerer omfattende opfyldelse af fagets mål med nogle mindre væsentlige mangler.	B
7	Den gode præstation	Karakteren 7 gives for den gode præstation, der demonstrerer opfyldelse af fagets mål med en del mangler.	C
4	Den jævne præstation	Karakteren 4 gives for den jævne præstation, der demonstrerer en mindre grad af opfyldelse af fagets mål med adskillige væsentlige mangler.	D
02	Den tilstrækkelige præstation	Karakteren 02 gives for den tilstrækkelige præstation, der demonstrerer den minimalt acceptable grad af opfyldelse af fagets mål.	E
00	Den utilstrækkelige præstation	Karakteren 00 gives for den utilstrækkelige præstation, der ikke demonstrerer en acceptabel grad af opfyldelse af fagets mål.	Fx
-3	Den ringe præstation	Karakteren -3 gives for den helt uacceptable præstation.	F

Kilde: Den nye karakterskala. Udgivet af Undervisningsministeriet i 2007.

7-trins-skalaen skal anvendes absolut i forhold til målene, så det vurderes, hvad den studerende har lært, og hvor godt vedkommende har lært det. Bedømmelsen er således uafhængig af andre studerendes præstation.

7-trins-skalaen skal anvendes absolut, men har dog en bestemt tilsigtet fordeling af de bestående karakterer, der lægger sig op ad ECTS-skalaens fordeling, jf. kapitel 2 nedenfor. ECTS-skalaens fordeling forventes ikke at være opfyldt på det enkelte hold eller den enkelte institution eller uddannelse i et bestemt år. Men ECTS-skalaen rummer en forestilling om, at karakterfordelingen i store populationer – fx på typer af uddannelser og over tid – skal svare til den forventede fordeling.

Ovanstående bild redovisar de allmänna kriterierna för de olika betygen. Dessa kompletteras med de mål och innehållsbeskrivningar

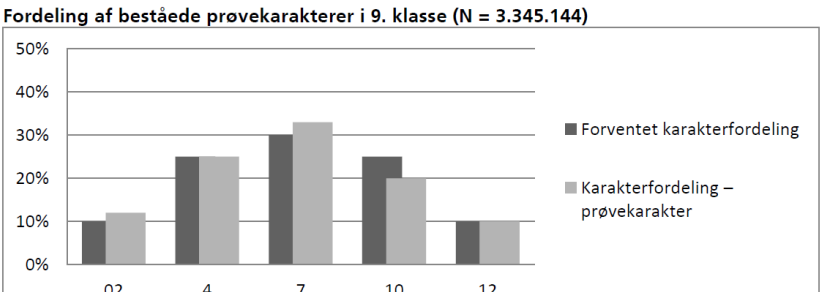
[61] Ur Danmarks Evalueringsinstitut (2013).

som finns i läroplanerna[62]. I jämförelse med svenska kunskapskrav är dock de danska motsvarigheterna betydligt mindre omfattande.

Det som gör den danska modellen intressant är att den har ambitionen att förena kriterierelaterade[63] och grupprelaterade betyg, där det förstnämnda gäller lärare och provkonstruktörer medan det senare gäller den nationella nivån. Tanken i den danska modellen är vidare att systemet ska utvärderas cirka vart femte år. Om det då visar sig att betygsfördelningen på nationell nivå avviker mer än vad som bedöms rimligt från den föreskrivna normfördelningen får betygskriterierna omformuleras så att betygsfördelningen blir mer i enlighet med den nationella normen.

Den första utvärderingen gjordes 2013 av Danmarks Evalueringsinstitut (EVA)[64]. Utvärderingen visar nedanstående fördelning för provbetygen och lärarbetygen i årskurs 9.[65] Proven är de examensprov som används i Danmark.

Figur 54 **Fördelning av godkända (bestående) provbetyg i relation till den nationella normen för betygsfördelning (förväntad betygsfördelning).**

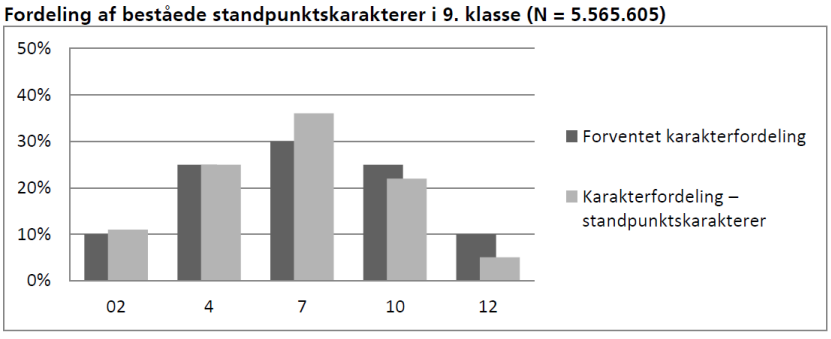


Av figuren framgår att den avvikelse som finns främst gäller fördelningen mellan betygen 7 och 10 samt att andelen med det lägsta godkända provbetyget 2 var ett par procentenheter högre än normen 10 procent.

[62] Se t.ex. http://www.uvm.dk/Uddannelser/Gymnasiale-uddannelser/Fag-og-laereplaner/Fag-paa-stx/Matematik-stx
[63] Motsvarar det svenska mål- och kunskapsrelaterade betygssystemet.
[64] Danmarks Evalueringsinstitut (2013).
[65] Figurer 54 och 55 är hämtade ur EVA:s rapport.

För lärarbetygen (*standpunktskarakterer*) gäller den fördelning som visas i figur 55.

Figur 55 **Fördelning av lärarbetyg (standpunktskarakterer) i relation till den nationella normen för betygsfördelning.**



Av figurerna framgår att överensstämmelsen får anses god. Andelen elever med medelbetyget 7 är något högre än normen i de två fallen. Däremot är andelen elever med de högsta betygen snarast lägre än normen. Det är också svårt att se någon tendens till att lärarbetygen skulle ligga högre än provbetygen.

Dessutom är det intressant att konstatera att de icke godkända betygen, vilka enbart bestäms enligt kriteriebaserade principer, endast utgör en liten andel av samtliga betyg. För såväl folkskolan som gymnasieskolan var cirka 5 procent av provbetygen icke godkända, medan cirka 3 procent av lärarnas slutbetyg var icke godkända.

EVA:s sammanfattande slutsatser framgår av nedanstående ruta.

Centrale konklusioner

- Generelt afviger karaktergivningen fra den forventede karakterfordeling med mindre end 10 procentpoint, og nogle karakterer rammer præcis den forventede fordeling.
- I langt de fleste tilfælde har karaktergivningen varieret med 0-7 procentpoint. Der er en generel tendens til, at karakteren 7 har nærmet sig den forventede fordeling.
- På uddannelsesområder, der har både prøvekarakterer og standpunkts-/årskarakterer, ligger karakteren 12 over den forventede fordeling, hvis man ser på prøvekaraktererne, mens den ligger under, hvis man ser på standpunkts-/årskaraktererne.
- De ikke-bestående karakterer udgør maksimalt 9 % af de karakterer, der er givet. Heraf udgør karakteren -3 som hovedregel ikke mere end 10 % af karaktererne. I den sammenhæng udgør universitetsuddannelserne og de kunstneriske uddannelser under Uddannelsesministeriet og Kulturministeriet dog væsentlige undtagelser.

Kommentar

Vid en första anblick kan den danska modellen förefalla märklig, men vid närmare eftertanke är det kanske inte så förvånande att modellen tycks fungera. Om man som i den svenska modellen enbart ska arbeta utifrån tolkning av texten i kunskapskraven är det svårt för den betygssättande läraren att veta om en tolkning är rimlig eller inte. De danska kunskapskraven är knappast formulerade på ett sådant sätt att de i sig är mer lättolkade och entydiga än svenska kunskapskrav. Det torde snarare vara det utrymme som finns för en relativ aspekt på tolkningen som, i varje fall med tiden, underlättar för danska lärare och provkonstruktörer att finna nivåer som är mer samstämmiga än vad fallet är i Sverige. Den nationella danska normen kan förstås av slumpmässiga skäl inte förväntas gälla annat än approximativt för enskilda klasser. Den ska således inte styra klassens betygssättning vare sig när det gäller genomsnitt eller spridning.

Några jämförelser av lärarbetyg, provbetyg och avvikelser på olika nivåer av det slag vi gör i Sverige redovisas inte av EVA. Det är därför svårt att dra några slutsatser om betygsavvikelser på skolenhets-, klass- och lärarnivå. Betygsfördelningar på nationell nivå är det som kan utläsas.

I Sverige förordas i dag inte relativ betygssättning även om en sådan kan ha fördelar i jämförelse med kriterierelaterad bedömning och betygssättning. Det kan vara lättare för en lärare att rangordna sina elever efter de kunskaper de visar än att relatera de visade kunskaperna till mer eller mindre abstrakta formuleringar i kunskapskrav och målbeskrivningar. En normfördelning kan i sådana fall tjäna som en återhållande kraft mot alltför yviga utsvävningar vid betygssättningen.

I praktiken finns det knappast några renodlade normrelaterade eller grupprelaterade betygssystem, i varje fall inte om praktiserande lärare är inblandade i betygssättningen. Har man ett norm- eller grupprelaterat system lär sig lärarna ganska snart vad elever på olika betygsnivåer kan, och därmed kan de också av elevernas visade kunskaper avgöra deras betygsnivå, dvs. de kan med tiden sätta rimligt tillförlitliga betyg på eleverna även utan direkt statistiskt stöd.[66] Men

[66] Oberoende av betygs- och provsystem kommer det dock alltid att råda osäkerhet vid betygsgränser.

samtidigt är verkligheten föränderlig och om bedömningen ska förbli tillförlitlig torde fortlöpande betygsstöd i statistisk eller annan form vara behövligt.

På liknande sätt kan ett normstöd av dansk modell tänkas bidra till att staga upp tolkningen av kunskapskraven så att de efterhand kan internaliseras i lärarnas kollektiva medvetande i form av konkreta föreställningar om vad elever på olika betygsnivåer presterar i form av olika utsagor. Det utesluter naturligtvis inte att tolkningar av innehållsbeskrivningar, kunskapskrav och lärandemål behöver göras och diskuteras – men kanske inte främst för bedömningens och betygssättningens skull utan i lika hög grad för att främja undervisning och lärande.

Norge

Norge har något de kallar *nasjonale prøver*. Dessa är dock inte obligatoriska och betygsstödjande som de svenska nationella proven, utan de är snarast en form av digitaliserade diagnostiska eller formativa prov. Däremot finns det obligatoriska examensprov.[67]

Examensproven konstrueras enligt traditionella metoder och betygssätts av särskilda censorer. Betygsskalan är sexgradig från 1 till 6, där 1 är icke godkänt och 6 högsta betyget. För censorerna gäller nedanstående anvisning[68].

[67] Se http://www.udir.no/Vurdering/Eksamen-grunnskole/ för närmare information.
[68] http://www.udir.no/Vurdering/Eksamen-grunnskole/#Karakterer

Grunnlag for karaktersetting

Når sensorene skal sette en karakter, må hun eller han forholde seg til kompetansemålene i læreplanen, forskrift til opplæringsloven, kjennetegn på måloppnåelse for faget og til det tolkningsfellesskapet sensorene er kommet fram til på sensorskoleringen.

Karakteren skal fastsettes etter en samlet vurdering av den kompetansen kandidaten viser i eksamensbesvarelsen. Deloppgaver vektes ikke.

Sensor skal ha en positiv holdning og se etter hva eleven får til. Mange oppgaver gir eleven mulighet til å velge ulike løsninger og innfallsvinkler for å svare på oppgaven.

Det hører med til sensors profesjonelle og faglige skjønn å være åpen for ulike faglige løsninger og synspunkter, så lenge eleven svarer på oppgavene som er gitt til eksamen.

Här finns inga angivelser om relativ betygsfördelning utan systemet beskrivs som helt mål- och kunskapsrelaterat.

Lärarens betyg (*standpunktskarakteren*) liknar den svenska lärarens och baseras på den samlade bedömningen av elevens kunskaper relaterade till vägledande beskrivningar av s.k. kännetecken för olika betyg.[69] Kännetecknen liknar de svenska kunskapskraven i det att det är några förmågor som rangordnas genom verbala beskrivningar.[70]

För Norge gäller att avgångsbetyget innehåller såväl betyg på examensproven som lärarens betyg. Denna modell tillämpas även i Danmark.[71]

[69] Se t.ex. http://www.udir.no/Vurdering/Standpunktvurdering-i-fag/
[70] Se t.ex. http://www.udir.no/globalassets/upload/vurdering/kjennetegn/matematikk_-kjennetegn_nn.pdf
[71] Se t.ex. http://www.uvm.dk/Uddannelser/Gymnasiale-uddannelser/Proever-og-eksamen/Eksamensbeviser-paa-de-gymnasiale-uddannelser

Finland

I Finland finns inga nationella prov av svensk typ i grundskolan. Inte heller finns examensprov av det slag som används i Danmark och Norge utan betygssättningen görs av läraren med stöd av följande anvisning.[72]

> Kriterierna beskriver kunskaperna som förutsätts för vitsordet åtta (8 = ”goda”). En elev får vitsordet åtta, om han eller hon i genomsnitt uppvisar kunskaper som motsvarar kriterierna. Svagare kunskaper inom något delområde kan kompenseras med kunskaper som överstiger kriterienivån inom andra delområden.

Betygssättningen är liksom i Danmark och Norge kompensatorisk, vilket den inte är i Sverige. Betygsskalan går från 4 till 10 där 4 är icke godkänt och 10 det högsta betyget. Det finns alltså bara kriterier för en sorts medelbetyg av de fem godkända betygen. De finländska lärarna i grundskolan ska alltså sätta betyg med stöd av kriterier för ett av de sju betygen och utan stöd av nationella prov. Läraren har tillsammans med den lokala skolan ansvar för betygssättningen av de egna eleverna och tydligen anses det fungera tillfredsställande. Möjligen kan Finlands goda resultat i PISA bidra.

För gymnasieskolans del används i Finland en relativ betygsskala med gamla latinska benämningar där L är högsta betyg.[73]

Laudatur	L	7	5%
Eximia cum laude approbatur	E	6	15%
Magna cum laude approbatur	M	5	20%
Cum laude approbatur	C	4	24%
Lubenter approbatur	B	3	20%
Approbatur	A	2	11%
Improbatur	I	0	5%

[72] http://www.edu.fi/planera/grundlaggande_utbildning/elevbedomningen/slutbedomningen_av_elever
[73] https://www.ylioppilastutkinto.fi/se/statistik/allmaent-om-examen

Som synes är skalan relativ och sjugradig.

Sammanfattningsvis gäller att för grundskolans del används i Finland en form av kriterierelaterad bedömning och betygssättning men inga nationella prov. För gymnasieskolan används i stället en traditionell studentexamen baserad på examensprov i varierande ämnen och med relativ betygssättning som grund.

England och Skottland

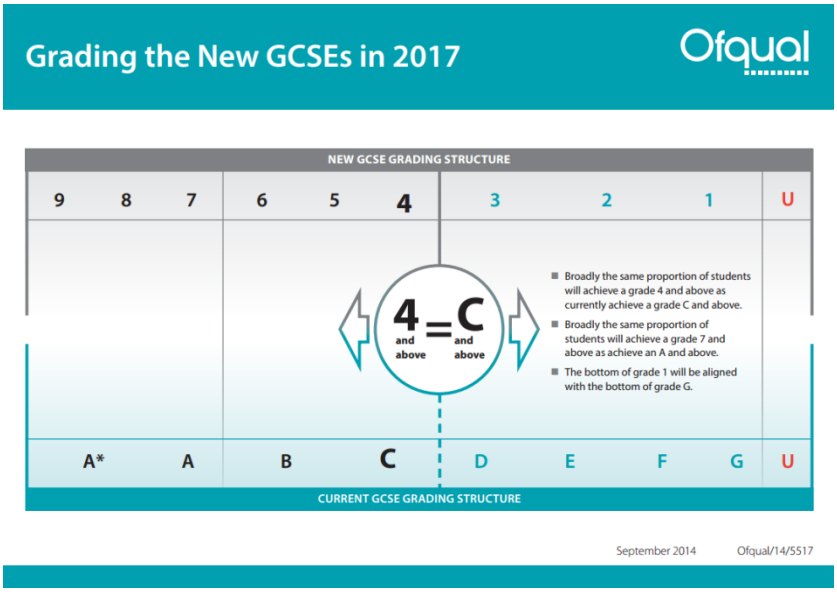
England och Skottland har under senare decennier genomfört många förändringar i sina prov- och betygssystem. Kingdon (2009) skriver:

> The Scottish system of school leaving qualifications (in common with other UK countries) appeared to have been the subject of far more changes in recent decades that most other systems. (s. 34)

Denna tendens har fortsatt även efter 2009. För Englands motsvarighet till gymnasieskolan pågår ett införande av en ny betygsskala inför 2017. Den beskrivs kortfattat på nedanstående sätt (figur 56) av Ofqual.[74]

[74] Office of Qualifications and Examinations Regulation. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/377768/2014-09-12-grading-the-new-gcses-in.pdf

Figur 56 Beskrivning av den nya sifferbaserade betygsskala som börjar gälla i England 2017.



Det nya systemet innebär en förstärkning av examensprovens roll. Ofqual[75] skriver bl.a.:

> Assessment will be mainly by exam, with other types of assessment used only where they are needed to test essential skills.

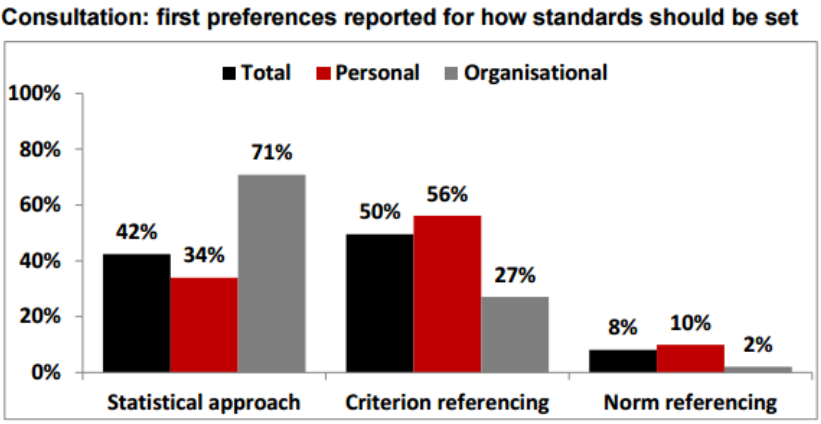
Nya kursplaner med innehållsbeskrivningar och bedömningsanvisningar infördes 2015. Kursplanerna är tämligen kortfattade jämfört med svenska kursplaner när det gäller syften och kunskapskrav. Däremot är innehållsbeskrivningarna mer detaljerade[76], antagligen som en följd av att lärarbedömningen har reducerats till förmån för examinerande prov och det av rättviseskäl är viktigt att alla elever ges samma möjligheter till lärande av för provet relevanta kunskapsområden.

[75] https://www.gov.uk/government/publications/get-the-facts-gcse-and-a-level-reform/get-the-facts-gcse-reform
[76] Se exempelvis https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254441/GCSE_mathematics_subject_content_and_assessment_objectives.pdf

Det sammanfattande intrycket av det nya engelska systemet på gymnasienivå är att betoningen av examensprov har ökat. När det gäller betygssättningen är strävan att försöka behålla ungefär samma betygsfördelning i den nya betygsskalan som i den tidigare, i varje fall under en övergångsperiod, med det tidigare betyget C som riktmärke. En sammanfattande beskrivning av huvuddragen i det nya systemet ges av Ofqual.[77]

Det nya systemet är inte renodlat relativt eller normrelaterat. I samband med införandet genomförde Ofqual en omfattande enkätundersökning, som pekade på att få användare förordade ett normrelaterat system (figur 57).[78] Däremot hade andra mindre styrande former av statistiskt stöd större uppslutning. Störst stöd hade dock kriterierelaterad betygssättning.

Figur 57 Enkätsvar om olika avnämares preferenser för att bestämma betyg.[79]



Den lösning som beslutats tycks gälla en statistisk ansats. Tendensen är att kursplaner och de bedömningskriterier som ingår tonas ned medan provresultaten ges större vikt.

[77] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/465873/your_qualification_our_regulation.pdf
[78] Detta gällde främst övergången mellan nya och gamla systemet.
[79] Ur https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/377771/2014-09-12-board-paper-for-new-gcses-in.pdf

Det engelska systemet skiljer sig från det svenska eftersom lärarna inte är inblandade i provverksamheten utan den sköts av ett antal särskilda organisationer. Dessa bestämmer också betygsgränser för proven, vilket innebär att de kursplaner som vänder sig till lärarna inte betonar bedömningskriterierna i någon högre grad. Däremot anges vikter för olika kunskapsområden inom olika ämnen, dvs. underlag för kvantitativ sammanvägning av olika provdelar, ungefär på samma sätt som för svenska nationella prov i engelska och svenska.

Även inom grundskolan (motsvarande) sker förändringar. Under senare år har antalet prov snarare minskat medan lärarbedömningar (*teacher assessment*[80]) ökat i betydelse. För högstadiet (*key stage 3*) togs de nationella proven bort 2010.

Det är svårt att se några direkta paralleller mellan det svenska och det engelska provsystemet under senare år. I England har som tidigare nämnts antalet nationella prov minskat i grundskolan medan lärarbedömningens betydelse har ökat. I Sverige har det snarast varit tvärtom. På gymnasial nivå ökar nu betoningen av examensproven i England medan lärarbedömningens betydelse minskar.

Sammanfattningsvis är det svårt att få grepp om det engelska systemet, framför allt för att det tycks vara under mer eller mindre kontinuerlig förändring. Där liksom på många andra ställen pendlar prov- och betygssystemet mellan grupprelaterade och kriterierelaterade ansatser. Monroe (2009) redovisar samma erfarenhet från Skottland.

> ... norm and criteria referencing represent opposing ends of a continuum; norms are often defined in criteria terms and vice versa. Current SQA grading procedures are located in the middle range of the continuum and more towards the criteria referenced end. Its grade-related criteria are general definitions, which are complemented by marking guidelines for specific exam papers. In addition, the percentage of candidates to be awarded a grade on the basis of their scores and the quality of their work is compared with the percentage of candidates usually being awarded the grade. (s. 16).

Motsvarande växlingar mellan relativa och kvalitativa bedömningar gör vi också i Sverige. Danmark har växlat mellan relativa och

[80] Se https://www.gov.uk/government/publications/key-stage-2-assessment-and-reporting-arrangements-ara/teacher-assessment

kriterierelaterade system. Finland har i sin studentexamen också en svårgenomskådlig mix av relativa och mer kvalitativa bedömningar.

Kommentar

Att den stora variationen mellan olika grupper inom det svenska systemet är problematisk ur ett likvärdighets- och rättviseperspektiv torde det inte råda någon tvekan om. Att förklara och förstå orsakerna är dock betydligt svårare. En del av förklaringen kan ligga i svårigheterna att göra samstämmiga tolkningar av verbala kunskapskrav och betygskriterier.

En av grundpelarna för ett provsystem består av de läroplaner, kursplaner, innehållsbeskrivningar och kunskapskrav som gäller för det utbildningssystem där proven ska användas. Ett provsystem är alltså oskiljaktigt förenat med det utbildningssystem det ingår i. För att provsystemet ska fungera måste därför också utbildningssystemet vara utformat så att båda dessa system är samstämmiga och kompatibla.[81]

En del i utredningens uppdrag gäller om kopplingen mellan provresultat och betygssättning bör tydliggöras eller förändras. Ett sätt att göra det är att ange en maximalt tillåten avvikelse mellan provresultat och betyg för en grupp, en skolenhet eller en huvudman, dvs. införa en normering på gruppnivå. En sådan normering kan förstås införas men svårigheten är att motivera valet av modell. Som framgått av redovisade resultat skulle en generell anvisning för tillåten avvikelse slå mycket olika mot olika ämnen och kurser. Å andra sidan behöver ett system med olika anvisningar för olika ämnen och kurser kunna motiveras tydligt för att få acceptans. Ambitionen bör därmed vara att finna en modell som är enkel och robust och som samtidigt uppfattas som rättvis och rimlig för alla ämnen och kurser. Strävan i de förslag utredningen lägger fram är att höja de nationella provens kvalitet, vilket på sikt bör bidra till att skillnaderna mellan ämnen minskar. Då bör det också bli lättare att hitta en modell som är användbar för alla ämnen och kurser.

Granskningen av några andra länders system pekar på vissa möjliga åtgärder. I flera fall används kombinationer av prov- och bedömnings-

[81] Se t.ex. Skolverket (2015c).

stöd baserade på både statistiska och kriterierelaterade grunder. I såväl Danmark som England är man inne på sådana tankegångar. Frågan för svensk del är i vilken utsträckning ett införande av vissa relativistiska eller statistiska tankegångar är gångbara och önskvärda. Den danska modellen är betydligt enklare i sin uppbyggnad än den svenska men den behöver granskas mer ingående innan en motsvarande svensk lösning eventuellt skulle kunna föreslås.

En annan intressant del med den danska modellen är att den i likhet med den norska innebär att såväl provbetyg som lärarbetyg redovisas i examensbevisen. Betygens viktigaste användning är som grund för urval till gymnasieskolan och högre utbildning. I det avseendet finns det mängder av forskning som visar att betyg har bättre prognostisk förmåga för framgång i studier än varje annan urvalsgrund. Dessutom förefaller träffsäkerheten bli bättre ju fler underlag som finns. Därför fungerar också meritvärden baserade på betyg bättre ju fler betyg som ingår.

Vilken prognostisk förmåga meritvärden och jämförelsetal baserade på både provbetyg och lärarbetyg skulle ha i en svensk kontext är dock oklart. Om båda typerna av betyg skulle redovisas i samma dokument (som i t.ex. Danmark) skulle skillnaderna (avvikelserna) mellan lärarbetyg och provbetyg bli tydliga på ett annat sätt än i dag när de bara framträder vid statistiska undersökningar eller Skolinspektionens granskningar. Nedan visas ett exempel på ett danskt bevis för studentexamen.[82]

[82] Från http://www.uvm.dk/Uddannelser/Gymnasiale-uddannelser/Proever-og-eksamen/Eksamensbeviser-paa-de-gymnasiale-uddannelser

Bevis for Studentereksamen (stx)
Aflagt i henhold til lovgivningen om de gymnasiale uddannelser

Navn: NN
Cpr.nr: ******-****
Eksamen er afsluttet juni 2014

	Årskarakterer			Prøvekarakterer			Særlige oplysninger		
Fag	Vægt	Karakter	ECTS	Vægt	Karakter	ECTS	Institution	Termin	Merit
Dansk A, mdt.	1	7	C	-	-	-			
Dansk A, skr.	1	7	C	1	4	D			
Engelsk A, mdt.	1	4	D	-	-	-			
Engelsk A, skr.	1	4	D	1	7	C			
Historie A	2	10	B	-	-	-			
Matematik A, mdt.	1	10	B	1	12	A			
Matematik A, skr.	1	10	B	1	10	B			
Fransk fortsættersprog B, mdt.	0,75	7	C	1,5	10	B			
Fransk fortsættersprog B, skr.	0,75	7	C	-	-	-			
Fysik B, mdt.	0,75	10	B	1,5	7	C			
Fysik B, skr.	0,75	7	C	-	-	-			
Samfundsfag B	1,5	7	C	-	-	-			
Biologi C	1	7	C	1	7	C	inst.nr.	t aa	stx
Filosofi C	1	4	D	-	-	-			
Idræt C	1	02	E	-	-	-			
Kemi C	1	7	C	-	-	-			
Musik C	1	7	C	-	-	-	inst.nr.	t aa	stx
Oldtidskundskab C	1	10	B	-	-	-			
Religion C	1	4	D	-	-	-			
Almen studieforberedelse	-	-	-	1,5	10	B			
Studieretningsprojekt	-	-	-	2	12	A			

Ikke medtællende fag	Karakter	ECTS	Særlige oplysninger		
			Institution	Termin	Merit
Almen sprogforståelse	4	D			
Naturvidenskabeligt grundforlob	10	B			

Studieretning: Matematik A, Fysik B, Filosofi C
Studieretningsprojekt: Fysik, Historie
Almen studieforberedelse: Fysik, Samfundsfag

Forelobigt eksamensresultat: **7,8**
Eksamensresultat: **7,8**

Bemærkninger:

Institutionens stempel/logo etc.
Dato og underskrift

Årskarakterer är de betyg läraren satt och *provekarakterer* är provbetygen. Man kan också notera att de olika ämnena ges olika vikt

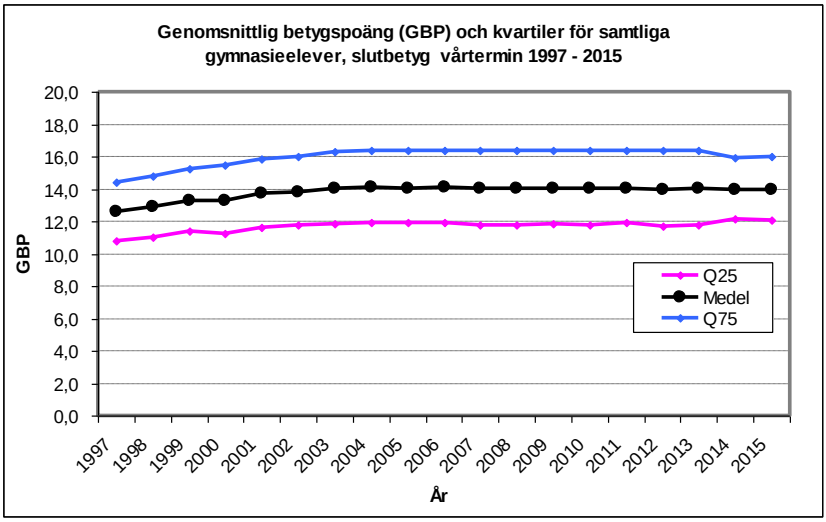
vid sammanräkning till meritvärden.[83] En danskinspirerad modell skulle kunna vara möjlig i Sverige, även om utredningen har tagit ställning för att inte föreslå en sådan modell eftersom det finns en risk att de nationella proven i så fall skulle styra betygen i för stor utsträckning på individnivå. Dessutom torde en modell liknande den danska kunna möta motstånd eftersom den knappast kan sägas harmoniera med rådande svensk syn på bedömning och betygssättning.

Innan vi lämnar jämförelsen med utländska modeller kan man undra hur en utländsk granskare skulle reagera inför det svenska systemet om de endast skulle utgå från vad som finns publicerat på olika webbplatser. Skulle de upptäcka de avvikelser och variationer som beskrivs i den här rapporten? Det är inte säkert. Kanske skulle de i stället få upp nedanstående bilder (se figur 58 och 59) över betygsmedelvärdenas utveckling i gymnasieskolan som helhet. Av figurerna framgår en liten förändring av kvartilvärdena i riktning mot medelvärdet 2014, vilket är det år då den nya sexgradiga betygsskalan får genomslag för första gången. Spridningen i genomsnittlig betygspoäng minskar alltså medan medelvärdet är i det närmaste konstant sedan 2003.[84] Figurerna avslöjar dock ingenting av de variationer och avvikelser som döljer sig under ytan.

[83] Används vid beräkning av de jämförelsetal som ligger till grund för antagning till universitet och högskolor. I Sverige viktas betygen efter hur många poäng kursen ger.

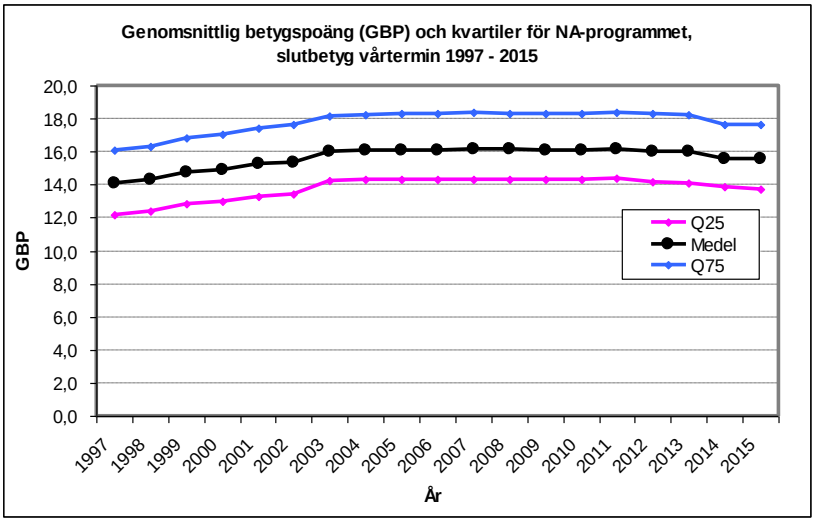
[84] Vi ser en förändring av kvartilerna mot medelvärdet 2014, det år då den nya sexgradiga betygsskalan får genomslag för första gången. Man kan notera att även under de första åren med den fyrgradiga betygsskalan var spridningen mindre än vad den sedan var. Det är en vanlig iakttagelse vid byte av betygssystem att betygsspridningen är liten i början, när lärarna är osäkra, och sedan ökar när de tycker sig bättre förstå de olika betygens innebörder. Man kan därför förvänta sig att betygsspridningen kommer att öka något de närmaste åren.

Figur 58 Den genomsnittliga betygspoängen (GBP) samt kvartilvärden för 25 och 75 procent (Q25 och Q75) för samtliga gymnasieelever med slutbetyg olika vårterminer.



För eleverna på Naturvetenskapsprogrammet är bilden likartad men betygsnivån högre.

Figur 59 Den genomsnittliga betygspoängen (GBP) samt kvartilvärden för 25 och 75 procent (Q25 och Q75) för elever på Naturvetenskapsprogrammet med slutbetyg olika vårterminer.



Ett förslag till svensk modell

I det föregående avsnittet diskuterades några olika förutsättningar och utgångspunkter för betygsstödjande prov och de konsekvenser som olika ansatser kan antas få för konstruktionen av en modell för relationen mellan provbetyg och lärarbetyg på gruppnivå. I det här avsnittet beskrivs en möjlig sådan modell.

Ett sätt att försöka hantera frågan om hur en modell kan utformas är att granska tidigare tillvägagångssätt. Vi inleder därför med en kort beskrivning av de två prov- och betygssystem som har gällt i Sverige sedan mitten av 1950-talet. De kan också i grova drag sägas vara representativa för vad som gäller och har gällt i andra länder, dock med olika nationella modifieringar. I ett första avsnitt granskas en modell som användes för prov i det relativa systemet.

Bakgrund

Det relativa systemet

Under de relativa betygens tid i Sverige (1962–1994)[85] fanns för gymnasieskolan en regel som angav att om betygsmedelvärdet för det nationella provet för en klass avvek mer än 0,2 betygssteg från medelvärdet av lärarens betyg var läraren skyldig att lämna en skriftlig motivering. Det är en modell av det slaget som brukar förespråkas av dem som önskar en motsvarande anvisning för det nutida prov- och betygssystemet. Det finns emellertid ett antal skillnader mellan det dåtida relativa betygssystemet och dagens mål- och kunskapsrelaterade system. I stora drag kan följande skillnader noteras.[86]

1. Det relativa systemet hade fem betygssteg där medelbetyget skulle vara 3 och skulle tilldelas 38 procent av eleverna, övriga betyg (1, 2, 4 och 5) skulle tilldelas bestämda andelar av den population som gjorde provet. Normalfördelningen valdes som modell för att fördela betygen och procentandelarna valdes så att det skiljde en standardavvikelse mellan varje betyg. Det gav en betygsfördelning på 7, 24, 38, 24 och 7 procent för betygen 1 till 5.

[85] Systemet började användas redan på 1940-talet men föreskrevs i Lgr 62 när den femgradiga betygsskalan med sifferbetygen (1–5) infördes.
[86] Detta innebär att vissa komplikationer och problem inte tas upp.

2. Alla ämnen skulle ha samma betygsfördelning och betygsnivå.[87]
3. Betygen skulle ha samma nivå och fördelning olika år.[88] Betygsglidning var i princip inte möjlig oberoende av om de underliggande kunskaperna steg eller sjönk.
4. Den föreskrivna betygsfördelningen gällde på nationell nivå för den population som läste ämnet och deltog i provet. Däremot behövde den inte gälla mindre grupper (skolor, klasser etc.) där medelvärdet och betygsfördelningen kunde vara en annan beroende på hur gruppens resultat var fördelat i relation till betygsgränserna på de nationella proven.
5. Betygsgränserna på proven fastställdes utifrån insamling av provresultat från ett representativt slumpmässigt urval av elever som genomfört provet. Poänggränserna för olika betyg bestämdes så att betygens fördelning för stickprovet så nära som möjligt överensstämde med den föreskrivna fördelningen. Sedan poänggränser för betygen bestämts skickades de ut till skolorna. Tillvägagångssättet krävde att proven gavs i så god tid att en preliminär bedömning och poängsättning kunde göras av lärarna, resultat skickas in till provkonstruktören och betygsgränser bestämmas för att sedan distribueras till skolorna. Därefter fastställde läraren de definitiva provbetygen och rapporterade dem till rektorn, varefter provbetyg och slutliga ämnesbetyg föredrogs för rektorn och lärarna i kollegiet.
6. Om skillnaden mellan medelvärdet av lärarens betyg avvek mer än 0,2 betygssteg uppåt eller neråt från medelvärdet av provbetygen kunde rektorn kräva en skriftlig motivering av läraren.
7. Systemet innebar att det fanns mål och föreskrifter som angav vad som ingick i kursen, men betygen baserades på att eleverna rangordnades efter sina visade kunskaper i relation till övriga elevers visade kunskaper och betygssattes utifrån sin position i denna rangordning. Systemet baserades vad avser nationella prov främst på kalibrerande mätning som angav klassens position i relation till den nationella betygsfördelningen.

[87] Detta gällde i princip. Efterhand infördes dock modifikationer för vissa ämnen.
[88] Så behöver det inte vara, men för det svenska systemet gällde detta. Ett sådant system kallas ibland kohortrelaterat eftersom varje årskull är sin egen norm.

Det mål- och kunskapsrelaterade systemet

Det mål- och kunskapsrelaterade systemet infördes 1994 och modifierades 2011. Betygskriterier blev kunskapskrav och den fyrgradiga betygsskalan ersattes av en sexgradig skala. Vidare infördes innehållsbeskrivningar i form av centralt innehåll. Någon mer grundläggande förändring gjordes dock inte vad avser nationella prov och betyg utan systemet är av samma typ som det som infördes 1994.

1. Systemet har i dag sex betygssteg (F–A). Betygen fastställs utifrån lärarens bedömning av på vilken nivå eleven uppfyller ett kunskapskrav i sin helhet (gäller lärarbetyg; för provbetyg gäller delvis andra bestämmelser).
2. Ingen nationell genomsnittlig betygsnivå eller nationell betygsfördelning är föreskriven. I princip kan alla elever få betyget A i ett ämne eller en kurs.
3. Olika ämnen kan ligga på olika betygsnivåer och ha olika betygsfördelningar.
4. Betygsnivåerna och betygsfördelningarna kan variera mellan olika år.
5. Betygsgränserna för provet fastställs före dess genomförande och tillhandahålls i anslutning till genomförandet, vilket innebär att provet kan läggas sent i kursen.
6. Det finns olika metoder och tillvägagångssätt för att fastställa betygsgränser eller kravgränser på prov. Dessa bygger främst på bedömning och tolkning av kunskapskraven, ibland med visst stöd av statistik från utprövningar. Förslag till betygsgränser tas fram i samverkan mellan de lärosäten som konstruerar proven och aktiva lärare. Skolverket fattar sedan beslut och ger trycklov.
7. Systemet innebär att elever i första hand betygssätts utifrån lärarens sammanfattande bedömning av i vilken grad kunskapskraven är uppfyllda. De nationella proven ska vara betygsstödjande men vad detta innebär och vilken relationen ska vara mellan nationella prov och betygssättning är inte reglerad.

De skillnader man kan utläsa ur ovanstående summariska beskrivning av de två systemen visar att förutsättningarna är olika i flera avseenden.

Varför var en avvikelse på 0,2 betygssteg lämplig för centrala prov i det relativa systemet?

Att provbetyg och lärarbetyg inte alltid överensstämmer är ingen ny företeelse. Även olika lärares betyg för samma prestationer varierar. I själva verket är denna variation i bedömning ett av huvudskälen till att införa betygsstyrande eller betygsstödjande prov. Provens syfte brukar anges vara att leda till en mer rättvis och likvärdig betygssättning. Samtidigt är prov inte några precisa mätinstrument. På individnivå har de betydande mätfel, men på gruppnivå är felen mindre – och ju större grupp, desto mindre mätfel för gruppens medelvärde, vilket vi visat tidigare.

Hur kommer man då fram till vilken skillnad mellan provbetyg och lärarbetyg som kan anses lämplig eller rimlig? Någon allmänt vedertagen eller vetenskapligt grundad anvisning finns inte utan den beslutas mer eller mindre godtyckligt av den som har auktorisation för detta. För de centrala prov som fanns för gymnasieskolan inom ramen för det relativa betygsystem som föregick det nuvarande mål- och kunskapsrelaterade systemet hade Skolöverstyrelsen bestämt att om klassens medelvärde för lärarens betyg avvek mer än 0,2 betygssteg från klassens medelvärde för provbetygen skulle läraren lämna en skriftlig motivering till rektorn. Rektorn och kollegiet avgjorde sedan om motiveringen var acceptabel och då godkändes lärarens betyg.[89]

Men varför sattes gränsen just vid 0,2 betygssteg? Någon dokumentation som beskriver motiven eller exakt när regeln infördes har utredningen inte kunnat finna, utan detta var mest troligt ett internt beslut av Skolöverstyrelsen. Man kan dock tänka sig att grunden ligger i det medelvärde och den standardavvikelse som var stipulerad för betygsfördelningen på nationell nivå, det vill säga en fördelning med medelbetyget 3,0 och standardavvikelsen 1,0. Beskrivningen är något teknisk och ges i nedanstående ruta.

[89] Se appendix 4 till bilagan som är Skolverkets version av Skolöverstyrelsens tidigare föreskrift.

En lite mer teknisk beskrivning

Om man i ett relativt system tänker sig att en genomsnittlig klass består av 25 elever betyder det att medelvärdet för slumpmässigt valda grupper med 25 elever fördelar sig med en standardavvikelse som är populationens standardavvikelse (1,0) dividerad med kvadratroten av antalet elever i stickprovet. Eller uttryckt som en formel

$$\frac{1}{\sqrt{25}} = \frac{1}{5} = 0{,}2$$

Detta är då medelvärdets standardfel som skattning av populationsmedelvärdet. Om man drar många stickprov kommer stickprovens medelvärden att fördela sig runt medelvärdet av stickprovens medelvärden (vilket blir den bästa skattningen av populationsmedelvärdet) med en standardavvikelse på 0,2. Om vi antar att ett stickprov på 25 elever har ett betygsmedelvärde på 3,25. Då kan vi vara till 95 procent säkra (konfidenta) att populationsmedelvärdet ligger i intervallet 3,25 - 1,96*0,2 till 3,25 + 1,96*0,2 eller uträknat i (konfidens)intervallet 2,86 till 3,64.

I vårt fall handlar dock frågan inte om att skatta populationsmedelvärdet ur ett stickprovsvärde utan om att jämföra två stickprov – ett ur populationen *provbetyg* och ett ur populationen *lärarbetyg* för att se om de kan antas vara dragna från två kongruenta populationer, dvs. populationer med samma medelvärden och fördelningar (vilket var stipulerat för provbetyg och lärarbetyg i det relativa systemet). Om vi då utgår från att normvärdet är provbetygens medelvärde skulle den stipulerade regeln innebära att om medelvärdet av lärarbetygen ligger inom intervallet medelvärdet för provbetyget $\pm$ 0,2 så är avvikelsen acceptabel. Detta är ett godtyckligt val som innebär att man kan vara till 68 procent (den andel som ligger inom $\pm$ 1 standardavvikelse) konfident att skillnaden mellan provbetyg och lärarbetyg inte skiljer från noll. En ganska låg grad av konfidens, vanligen brukar två standardavvikelser användas vilket ger konfidens till 95 procent.

Förutsättningar för användning inom ramen för dagens system

Om man skulle försöka använda motsvarande regel som gällde för det relativa systemet i det nuvarande systemet saknas två grundläggande förutsättningar. För det första finns inget stipulerat nationellt betygsmedelvärde och för det andra finns ingen föreskriven standardavvikelse för betygen på nationell nivå.

Avsaknaden av dessa förutsättningar skapar problem för både provkonstruktörer och betygssättande lärare. Provkonstruktörerna ska tolka kunskapskraven vid konstruktionen av prov och fastställande av betygsgränser. Lärarna ska göra detsamma vid sina sammanfattande bedömningar av elevernas kunskaper i relation till kunskapskraven. Det är då inte bara betygsmedelvärden som ska bestämmas utan också gränser för olika betyg utifrån vilka sedan ett betygsmedelvärde kan beräknas.

Förutsättning 1: Bestämda betygsmedelvärden

Det relativa betygssystemet hade ett bestämt betygsmedelvärde (3) och en bestämd betygsfördelning (standardavvikelsen 1) medan det mål- och kunskapsrelaterade systemet saknar bestämmelse om betygsmedelvärde och betygsfördelningen är implicit angiven i form av kunskapskrav som ska tolkas innan de kan generera betyg och därmed en betygsfördelning som ger underlag för att beräkna betygsmedelvärden på proven. Någon teoretisk grund för bestämning av betygsmedelvärden och betygsfördelning finns inte i systemet. Därmed återstår endast möjligheten att bestämma värdena empiriskt om de inte helt enkelt stipuleras, något som utredningen inte förordar.

Som vi sett råder det för vissa ämnen och kurser betydande diskrepans mellan betygsmedelvärden baserade på prov och lärarbedömning. Den grundläggande frågan blir då vilka betyg som ska ligga till grund för det nationella medelvärdet: Ska det vara lärarnas betyg? Eller provkonstruktörernas betyg? Eller någon kombination av de två?

Låt oss anta att den samlade lärarkårens betyg är den mest väl underbyggda tolkning som finns tillgänglig genom det stora antal lärare som medverkar i provbedömning och betygssättning och därmed också i tolkningen av kunskapskravens innebörder. Vi väljer

därmed lärarnas betygmedelvärde och betygsfördelning som norm. För att underlätta jämförelsen mellan provbetyg och lärarbetyg bör då provbetygen justeras till samma nivå som lärarbetygen. En enkel modell är att undersöka hur stor avvikelsen mellan lärarbetyg och provbetyg tenderar att vara (medelvärdet av skillnaden mellan lärarbetyg och provbetyg enligt vald skala för betygspoäng under ett bestämt antal år, t.ex. de fem senaste) och sedan addera denna genomsnittliga differens till det aktuella provbetygets medelvärde. Denna tanke utvecklas närmare senare i den här bilagan.

Därmed antas för den första förutsättningens del att det nationella (och därmed lokala) betygsmedelvärdet bestäms som provmedelvärdet plus medelvärdet av tidigare års avvikelser på nationell nivå.

Förutsättning 2: Bestämd betygsfördelning

I det relativa betygssystemet var betygsfördelningen föreskriven utifrån en normalfördelad modell med standardavvikelsen 1. Någon motsvarande regel är inte bestämd för det mål- och kunskapsrelaterade systemet. Betygen kommer givetvis att fördela sig ändå men utifrån betygssättarens tolkningar av kunskapskravens innebörder och elevens kunskaper så som de visar sig i olika utsagor, varav nationella prov är en.

Någon teoretisk modell som kan användas som underlag för betygens fördelning är således inte bestämd. Den enda utgångspunkten är kunskapskraven och de betyg som de genererar. Eftersom någon betygsfördelning inte är beslutad utgår vi från den betygsfördelningen som empiriskt visat sig vid tillämpningen av kunskapskraven.

Här dyker återigen frågan upp om vems tolkning som ska gälla som empirisk bedömning: Ska det vara det samlade lärarkollektivets betygsfördelning? Ska det vara provbetygens fördelning? Ska det vara någon form av sammanvägd fördelning av lärarbetyg och provbetyg?

Vi väljer att även här utgå från det samlade lärarkollektivets bedömning. För att få så tillförlitliga värden som möjligt vore det bästa att utgå från värden under flera år. Detta är dock inte möjligt i den här redovisningen. Det nuvarande sexgradiga betygssystemet är ganska nytt och de underlag som är tillgängliga för utredningen är data

från grundskolan för 2013 och 2014 och för gymnasieskolan från vårterminen 2014. Därför används de som underlag för de exempel som visas. För en modell i bruk är dock tanken att de värden som ingår i modellen ska baseras på betygsunderlag från ett antal år.

För den andra förutsättningens del antas att betygsfördelningen kan bestämmas empiriskt genom att man utgår från lärarbetygens standardavvikelse under ett antal år.

Ett exempel baserat på empiriska data

I det här avsnittet redovisas en tillämpning av samma modell som gällde för de tidigare centrala proven men med den skillnaden att vi nu använder empiriska värden för betygens medelvärden och spridning i stället för stipulerade. De värden som används i exemplet är inte medelvärden över tid, vilka är tänkta att gälla när modellen tas i praktiskt bruk. I stället används de data som utredningen har haft tillgång till, dvs. provresultat och betyg för grundskolans årskurs 9 vårterminerna 2013 och 2014 samt motsvarande resultat för gymnasieskolan vårterminen 2014.

I tabellerna nedan listas betygsmedelvärden och standardavvikelser för provbetyg och lärarbetyg för grundskolan vårterminerna 2013 och 2014 samt för gymnasieskolans kursprov vårterminen 2014. Här används betygspoäng från 1 (betyget F) till 6 (betyget A) för att få en betygsskala som liknar den tidigare femgradiga skalan men med sex steg i stället för fem. Lbet_6 och Pbet_6 markerar lärarbetyg respektive provbetyg. Vi börjar med att undersöka avvikelserna på gruppnivå och undersöker sedan avvikelserna på systemnivå (nationell nivå).

Avvikelser på gruppnivå

Tabellerna visar antal elever, hela elevgruppens betygsmedelvärden och skillnader mellan betygsmedelvärdena (Diff_6), standardavvikelse för lärarbetyg och provbetyg samt betygsmedelvärdenas standardfel för slumpmässigt valda grupper av 25 elever.[90]

[90] Det sistnämnda värdet är beräknat på samma sätt som det tidigare värdet 0,2, dvs. standardavvikelsen för den totala populationen dividerad med kvadratroten av antalet elever i stick-

Tabell 8 Medelbetyg och standardavvikelse för lärarbetyg och provbetyg för nationella prov i åk 9 vt 2013 och 2014 beräknat enligt skala 1 till 6 betygspoäng för betygen F till A, samt standardavvikelse för stickprov med 25 elever.

Vårtermin	Ämne	Antal	Medel	Diff_6	Stdav	Std/rot(25)
2013	Ma_Lbet_6	86704	3,37		1,44	**0,29**
	Ma_Pbet_6	86704	3,21	**0,17**	1,51	**0,30**
	Sv_Lbet_6	81181	3,60		1,29	**0,26**
	Sv_Pbet_6	81181	3,46	**0,14**	1,23	**0,25**
	En_Lbet_6	87663	3,94		1,39	**0,28**
	En_Pbet_6	87663	4,05	**-0,11**	1,35	**0,29**
2014	Ma_Lbet_6	87043	3,25		1,40	**0,28**
	Ma_Pbet_6	87043	2,93	**0,32**	1,41	**0,28**
	Sv_Lbet_6	80629	3,69		1,30	**0,26**
	Sv_Pbet_6	80629	3,56	**0,14**	1,26	**0,25**
	En_Lbet_6	88299	4,01		1,39	**0,28**
	En_Pbet_6	88299	4,10	**-0,09**	1,33	**0,27**
Medel					**1,36**	**0,27**

Av tabell 8 framgår att standardavvikelsen för betygsmedelvärdet[91] för slumpmässigt valda stickprovsgrupper om 25 elever är mycket likartad för såväl provbetyg som lärarbetyg och kan på nationell nivå avrundat till en decimal sägas vara 0,3 betygssteg för grundskolans del. Skillnaden mellan provbetyg och lärarbetyg kan ses som försumbar.

För gymnasieskolans del är bilden något mer varierande, vilket är naturligt med tanke på att de olika proven där genomförs av selekterade och därigenom i allmänhet mer homogena grupper. Tabell 9 visar resultaten.

provet (25). Skillnaden är att för 0,2 var det den föreskrivna standardavvikelsen 1 som gällde. Här är det den empiriskt funna standardavvikelsen för lärarbetygen som används. Man kan notera att skillnaderna i standardavvikelse för provbetyg och lärarbetyg är liten.

[91] Detta kallas också medelvärdets standardfel, men avser då stickprovets medelvärde som skattning av populationsmedelvärdet.

Tabell 9 Medelbetyg och standardavvikelse för lärarbetyg och provbetyg för nationella prov i gymnasieskolans olika kurser vt 2014 beräknat enligt skala 1 till 6 betygspoäng, samt standardavvikelse för medelvärdet för stickprov med 22 respektive 25 elever.

Kurs	Betyg	Antal	Medel	Diff_6	Stdav	Std/(rot(22)	Std/rot(25)
En05	Lbet_6	70741	3,77		1,37	0,29	0,27
	Pbet_6	70741	3,77	0,00	1,29	0,28	0,26
En06	Lbet_6	55821	3,83		1,34	0,29	0,27
	Pbet_6	55821	3,80	0,03	1,25	0,27	0,25
Ma1a	Lbet_6	22222	2,35		1,04	0,22	0,21
	Pbet_6	22222	2,03	0,32	0,93	0,20	0,19
Ma1b	Lbet_6	28288	2,95		1,31	0,28	0,26
	Pbet_6	28288	2,70	0,24	1,28	0,27	0,26
Ma1c	Lbet_6	5277	3,84		1,43	0,31	0,29
	Pbet_6	5277	3,72	0,12	1,44	0,31	0,29
Ma2a	Lbet_6	2368	2,27		1,14	0,24	0,23
	Pbet_6	2368	1,85	0,43	1,07	0,23	0,21
Ma2b	Lbet_6	22531	2,46		1,22	0,26	0,24
	Pbet_6	22531	1,99	0,46	1,10	0,23	0,22
Ma2c	Lbet_6	9981	3,47		1,48	0,31	0,30
	Pbet_6	9981	3,20	0,27	1,46	0,31	0,29
Ma3b	Lbet_6	6864	2,77		1,32	0,28	0,26
	Pbet_6	6864	2,42	0,35	1,27	0,27	0,25
Ma3c	Lbet_6	8719	3,24		1,54	0,33	0,31
	Pbet_6	8719	3,01	0,24	1,52	0,32	0,30
Ma4	Lbet_6	7123	3,28		1,56	0,33	0,31
	Pbet_6	7123	2,94	0,35	1,54	0,33	0,31
Sv1	Lbet_6	64869	3,60		1,32	0,28	0,26
	Pbet_6	64869	3,49	0,11	1,29	0,27	0,26
Sv3	Lbet_6	46792	3,71		1,40	0,30	0,28
	Pbet_6	46792	3,35	0,36	1,40	0,30	0,28
Medel					1,31	0,28	0,26

Tabell 9 visar att stickprovens standardavvikelse är något lägre än för grundskolans del, eller 0,26 om man räknar på samma klassstorlek (25 elever) som i grundskolan. Klasserna är dock i allmänhet något mindre i gymnasieskolan och om man utgår från en grupp på 22 elever blir standardavvikelsen 0,28, dvs. i stort sett samma som i grundskolan. Sammantaget kan man säga att det avrundade värdet blir 0,3 betygspoäng, vilket motsvarar 0,3 betygssteg i den skala som har samma avstånd mellan varje betygssteg.

Om analogin med det värde (0,2) som gällde för det relativa betygssystemet är tillämplig skulle motsvarande värde i det nuvarande

systemet utifrån de standardavvikelser som gäller för lärarbetygen på nationell nivå vara 0,3 betygspoäng för klassmedelvärdet utifrån den använda sexgradiga skalan. Detta leder till följande, för exemplet tentativa, anvisning.

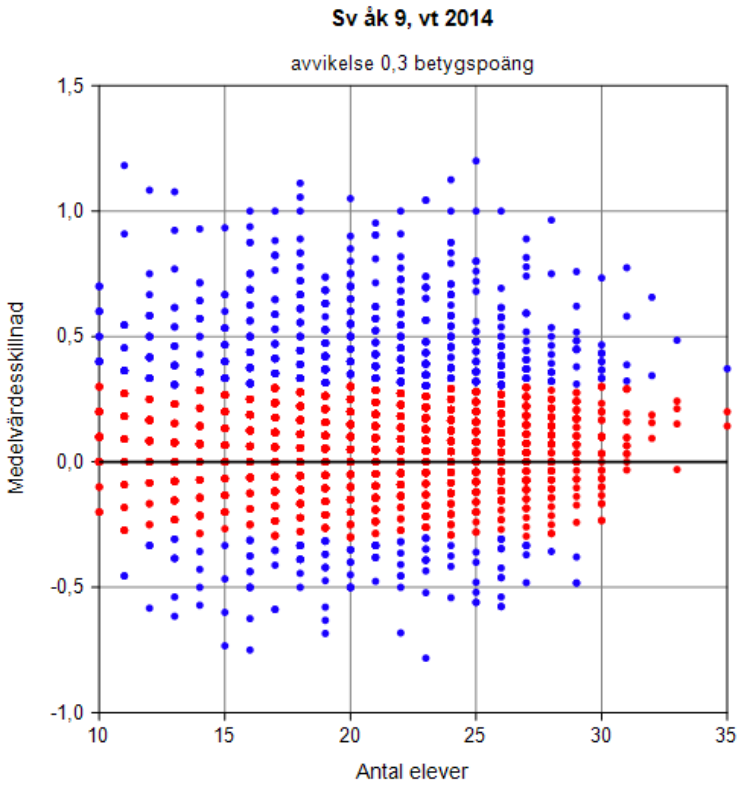
Accepterad avvikelse på gruppnivå: Skillnaden mellan medelvärdet för lärarbetygen och provbetygen ligger i intervallet - 0,3 till 0,3 betygspoäng räknat på en betygsskala med betygspoäng från 1 för F till 6 för A.

Hur skulle då bilden se ut om en avvikelse mellan betygsmedelvärden för provbetyg och lärarbetyg på 0,3 betygspoäng skulle användas som norm i det nuvarande systemet (med användning av en sexgradig skala)? Nedan anges några exempel. Dessa gäller dock endast avvikelsernas spridning, dvs. den del av modellen som baseras på förutsättning 2. I nästa avsnitt tillkommer den eventuella korrigering av avvikelsens nivå som baseras på förutsättning 1.

Grundskolan

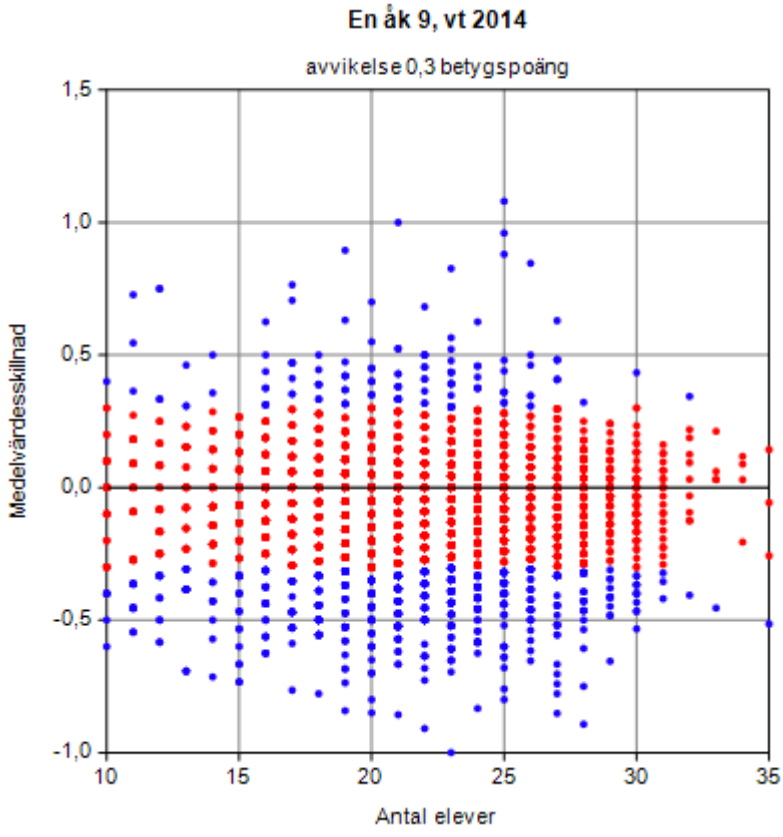
För grundskolan visas diagram för provämnena svenska, engelska och matematik i årskurs 9. Prickarna i diagrammet representerar grupper (klasser). De röda prickarna representerar grupper inom det accepterade intervallet.

Figur 60 Grupper med medelvärdesskillnader (Diff_6) mellan provbetyg och lärarbetyg mindre än 0,3 (röda prickar) och skillnader större än 0,3 (blå prickar). Svenska åk 9 vt 2014.



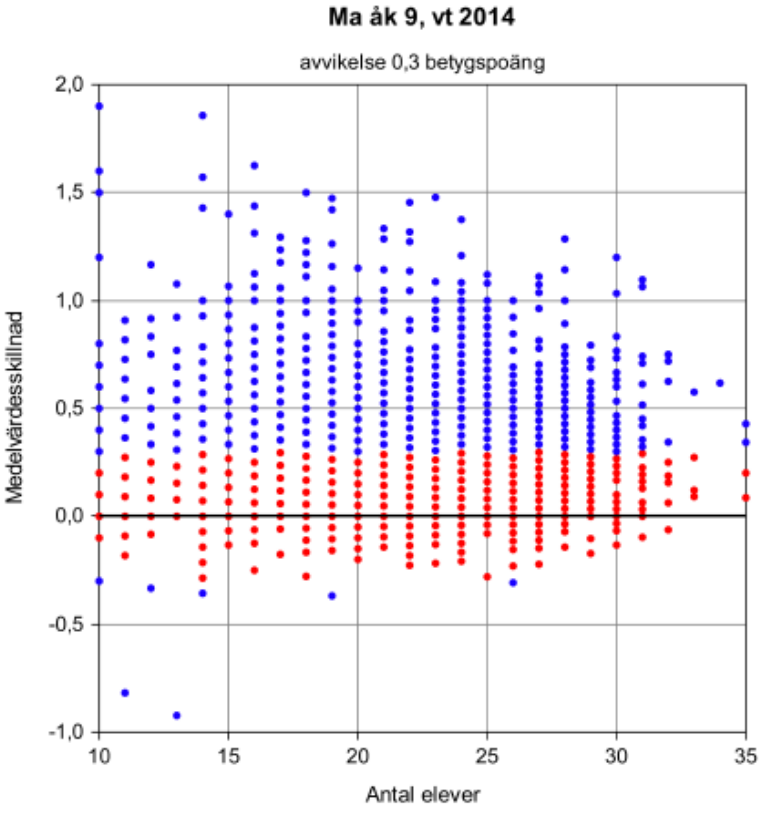
Av grupperna i svenska ligger 26 procent utanför gränsen 0,3 betygssteg. Flertalet avvikelser gäller som synes högre lärarbetyg än provbetyg.

Figur 61 Grupper med medelvärdesskillnader (Diff_6) mellan provbetyg och lärarbetyg mindre än 0,3 (röda prickar) och skillnader större än 0,3 (blå prickar). Engelska åk 9 vt 2014.



Avvikelse utanför gränsen 0,3 gäller för 20 procent av grupperna i engelska. Här har förhållandevis fler grupper lägre lärarbetyg än provbetyg. Effekten skulle för engelskans del därmed bli ett visst tryck mot högre lärarbetyg.

Figur 62 Grupper med medelvärdesskillnader mellan provbetyg och lärarbetyg mindre än 0,3 (röda prickar) och skillnader större än 0,3 (blå prickar). Matematik åk 9 vt 2014.

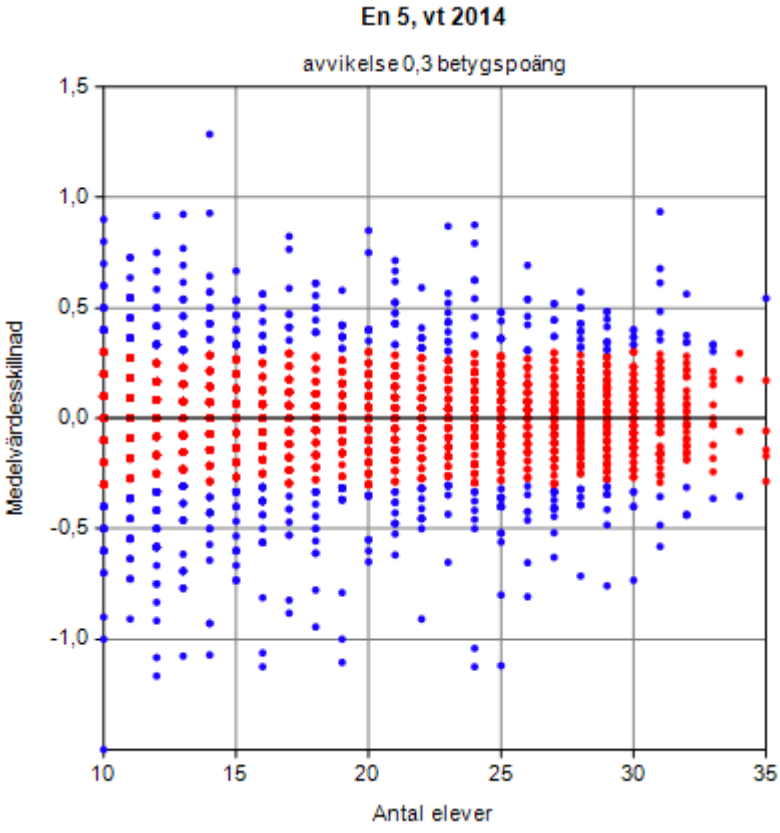


I matematik avviker 45 procent av grupperna med mer än 0,3 betygssteg. I stort sett alla avvikelser innebär högre lärarbetyg än provbetyg; endast ett fåtal grupper har negativa avvikelser.

Gymnasieskolan

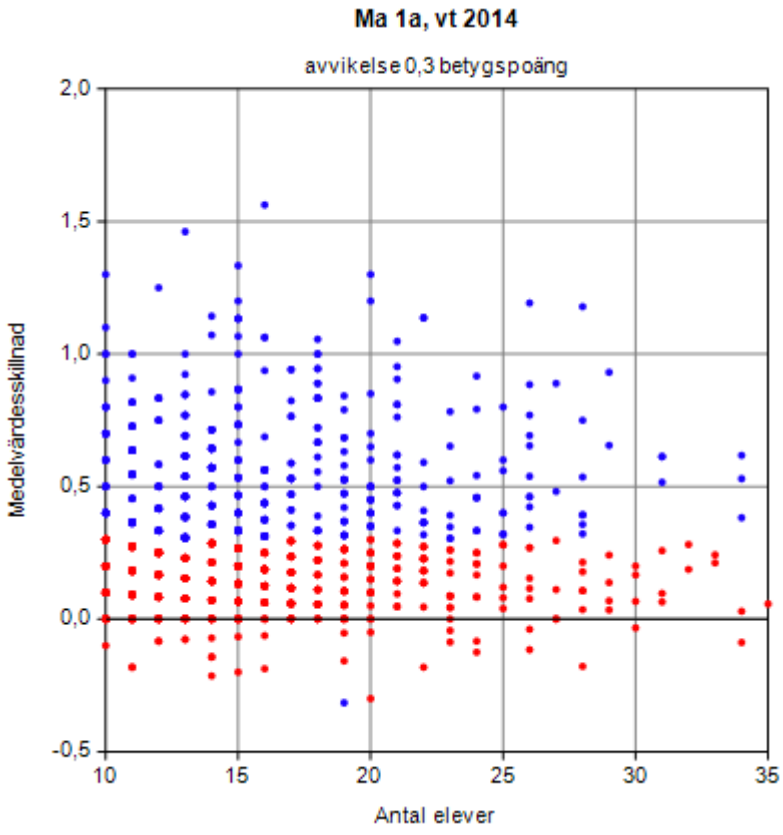
För gymnasieskolan visas också tre exempel: engelska 5, matematik 1a och svenska 3. Som framgått tidigare är resultaten och avvikelserna i gymnasieskolan i allmänhet betydligt mer varierande än i grundskolan.

Figur 63 Grupper med medelvärdesskillnader (Diff_6) mellan provbetyg och lärarbetyg mindre än 0,3 (röda prickar) och skillnader större än 0,3 (blå prickar). Engelska 5, vt 2014.



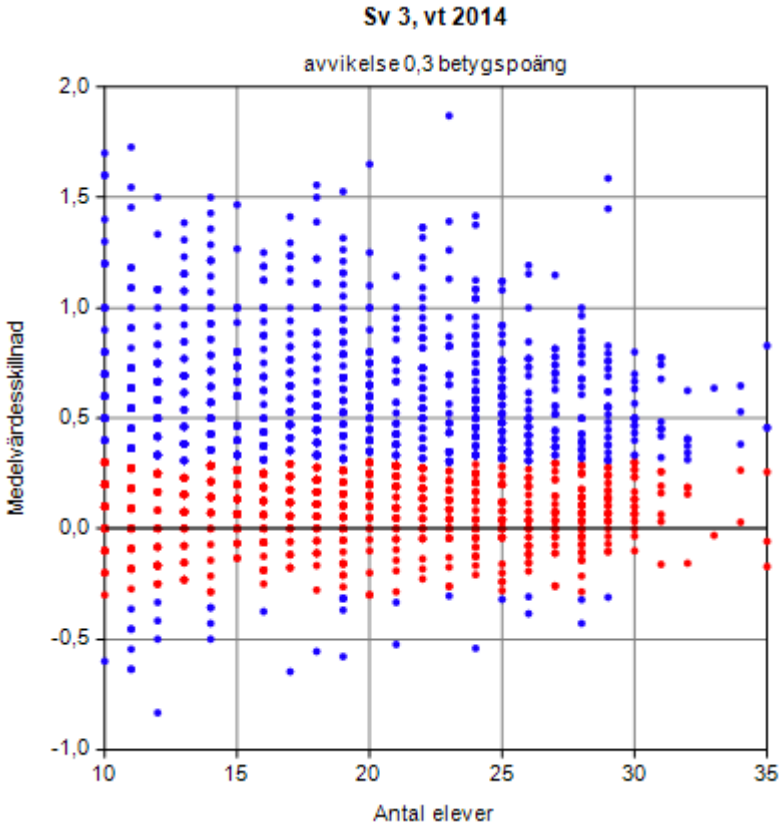
Engelska är som tidigare framgått det ämne som har den minsta genomsnittliga avvikelsen på nationell nivå och den jämnaste fördelningen mellan positiv och negativ avvikelse. För engelska 5 skulle en gräns på 0,3 betygssteg innebära att cirka 20 procent av grupperna ligger utanför den tillåtna avvikelsen (de blå grupperna).

Figur 64 Grupper med medelvärdesskillnader mellan provbetyg och lärarbetyg mindre än 0,3 (röda prickar) och skillnader större än 0,3 (blå prickar). Matematik 1a, vt 2014.



För grupperna med prov i matematik 1a har 43 procent en genomsnittlig avvikelse som är större än 0,3 betygssteg. Samtliga avvikelser (utom en) gäller att lärarbetygen för gruppen i genomsnitt ligger mer än 0,3 betygssteg över provbetygen. Det kan sägas vara den gängse bilden för matematikämnet.

Figur 65 Grupper med medelvärdesskillnader mellan provbetyg och lärarbetyg mindre än 0,3 (röda prickar) och skillnader större än 0,3 (blå prickar). Svenska 3, vt 2014.



För svenska 3 ligger hela 52 procent av grupperna utanför den tillåtna avvikelsen på 0,3 betygssteg. Även här innebär avvikelserna huvudsakligen att lärarbetyget är högre än provbetyget.

Kommentar

Modellen med tillåten avvikelse på 0,3 betygspoäng ger bilder som i huvudsak liknar de som visats tidigare i avsnittet *Slumpens roll för avvikelsen*. Att det finns likheter är inte förvånande, för i princip bygger de två modellerna på samma principer om statistiska

skillnader mellan olika stickprov.[92] Den modell som användes i de centrala proven är dock mer primitiv genom att den använder ett standardiserat värde baserat på en genomsnittlig klasstorlek (25 elever), medan de tidigare visade diagrammen anger avvikelsen med hänsyn tagen till gruppstorleken. Fördelen med den centralprovsmodell som använts ovan är att den är tydlig och lättanvänd för lärare och rektorer (förutsatt att den förenklade skalan 1–6 för betygspoäng används). Nackdelen är förstås att den är mindre precis. Små grupper har dessutom, som tidigare framgått, en större slumpvariation.

Figurerna i avsnittet visar att för engelskans del förefaller en modell som enbart gäller gruppmedelvärdenas avvikelser vara tillräcklig. Den systematiska skillnaden (medelvärdesskillnaden på nationell nivå) mellan provbetyg och lärarbetyg är i stort sett noll och behöver därmed inte korrigeras. Däremot kan spridningen mellan grupper anses vara för stor om 0,3 betygssteg bestäms som norm.

För matematikens del är bilden en annan. Där finns det, förutom skillnader mellan grupperna, också betydande skillnader mellan lärarbetygens och provbetygens nationella medelvärden, det vi kallar en systematisk skillnad. Den allmänna betygsnivån i matematik, i synnerhet när det gäller provbetygen, ligger dessutom på en så låg nivå att en betydande andel elever skulle bli underkända i många kurser om provbetygen skulle gälla.

Utredningen anser att den låga generella betygsnivån på de nationella proven och de stora avvikelser lärarna anser nödvändiga för att elevernas betyg ska bli i enlighet med lärarnas bedömningar av elevernas kunskaper är olycklig. Därmed kan det för matematikämnet vara aktuellt med en systematisk korrigering av provbetygen så att man får en modell som fungerar mer i linje med lärarnas bedömningar av rimlig betygsnivå. Ambitionen med de kvalitetshöjande insatser vi föreslår för de nationella proven är bl.a. att de på sikt ska bidra till att skillnaderna mellan olika ämnen och mellan ämnes- eller kursbetyg och provbetyg minskar. I så fall kan den

[92] Det gäller även om den tidigare använda modellen bygger på t-test där hänsyn tas till respektive grupps storlek och stickprovsvärden, medan den senare centralprovsmodellen bygger på populationsvärden (standardavvikelse) och antagande om normalfördelning.

systematiska korrigeringen för avvikelse på systemnivå efterhand minskas.

Avvikelse på systemnivå

Engelskan behöver utifrån befintliga resultat ingen korrigering av den genomsnittliga betygsnivån på proven, i varje fall inte uppåt. För engelskans del gäller snarare att de stigande betygen kan behöva bromsas för att inte skillnaderna mellan ämnen ska bli större än vad de redan är.

För svenskans del är underlaget alltför magert för att vi ska kunna bedöma hur stor korrigering som kan vara lämplig. I synnerhet svenska 3 hade en stor avvikelse vårterminen 2014 medan övriga tre svenskkurser låg mellan 0,11 och 0,14 i avvikelse (tabell 8 och 9). Med det lilla underlaget väljer vi att utelämna proven i svenska.

Mest akuta är proven i matematik. Om man ser på medelvärdesskillnaderna (Diff_6) enligt tabell 8 och 9 ser man att värdena är förhållandevis höga, med betydande skillnad mellan minsta värdet 0,12 betygspoäng och största 0,42 betygspoäng. Räknar man ut medelvärdet för samtliga matematikkurser får man värdet 0,30 betygspoäng.

Om man ska ha en anvisning för systematisk korrigering av provbetygen är det en fördel, främst av praktiska skäl, om den kan vara enhetlig för hela ämnet. Vi väljer därför att utgå från medelvärdet av Diff_6 för de olika matematikkurserna, vilket avrundat blir 0,3 betygspoäng. Det innebär en korrigering av provbetygets medelvärde så att det hamnar på samma nivå som genomsnittet av matematikbetygen från olika år (vilka ersätts med olika matematikkurser från utredningens underlag i exemplet) genom att addera korrektionen 0,3 betygspoäng till det observerade betygsmedelvärdet på provet. Detta är alltså en allmän korrektion som gäller alla som deltagit i provet. Syftet med korrektionen är att provbetygens medelvärde ska hamna på samma nivå som lärarbetygens medelvärde över tid.

Detta ger följande anvisning utifrån det i exemplet tillgängliga underlaget.

Accepterad avvikelse på systemnivå: Medelvärdet för provbetygen ökas för ämnet matematik med 0,3 betygspoäng räknat på en betygsskala med betygspoäng från 1 för F till 6 för A. För engelska och svenska anges däremot ingen avvikelse på systemnivå. För engelska tyder resultaten på att någon korrigerande faktor inte är aktuell. För svenska är underlaget för begränsat.

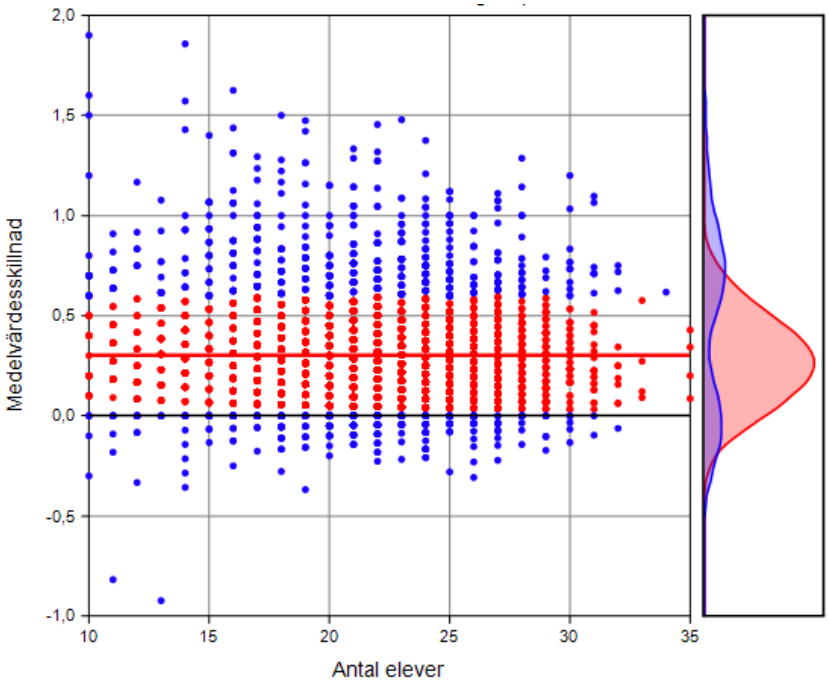
Det en ren tillfällighet att provbetygen i matematik ska höjas med just 0,3 betygspoäng (och att den systematiska avvikelsen därmed minskas med 0,3 betygspoäng), vilket är samma värde som intervallet för avvikelsen på gruppnivå. Om vi i stället valt ämnet svenska skulle korrektionen avrundad till en decimal ha blivit 0,1 betygspoäng (medelvärdet av Diff_6 för de olika kurserna i svenska i tabell 8 och 9). I engelska skulle den systematiska korrektionen på motsvarande sätt bli ett värde någonstans mellan 0 och -0,1 betygspoäng.

Nedan ges exempel på hur modellen skulle kunna tillämpas.

Tillämpning av modellförslaget

För att illustrera den modell utredningen anser möjlig väljs matematik i årskurs 9 vårterminen 2014 som exempel. I figur 66 redovisas fördelning av grupper inom och utanför accepterad avvikelse korrigerad på systemnivå, så att lärarbetyget i genomsnitt förväntas ligga 0,3 betygspoäng högre än provbetyget i enlighet med tidigare härlett genomsnitt för de matematikkurser som ingår i utredningens underlag. Det betyder att den anvisning som gäller för systematiskt okorrigerad avvikelse, dvs. (- 0,3 betygspoäng < avvikelsen < 0,3 betygspoäng) vid systematisk korrektion ersätts med intervallet (0.0 betygspoäng < avvikelsen < 0,6 betygspoäng), vilket innebär att avvikelsenivån flyttas upp 0,3 betygssteg. Figur 66 visar utfallet för matematik i årskurs 9, vårterminen 2014 med systematisk korrektion och accepterad avvikelse.

Figur 66 Fördelning av grupper inom (röda prickar) och utanför (blå prickar) accepterad avvikelse 0,3 betygspoäng (Diff_6). Korrigerad för systematisk avvikelse 0,3 betygspoäng. Matematik åk 9, vt 2014.



Av marginalfördelningarna framgår att den accepterade avvikelsen ligger på en nivå som är mer i överensstämmelse med lärarkårens genomsnittliga betygssättning. Dock är variationen mellan olika lärare (grupper) fortfarande betydligt större än de riktmärken för accepterad avvikelse (0,3 betygspoäng) som valts för exemplet. Liksom i tidigare figurer verkar det finnas betydande slumpeffekter inblandade.

Hur kan modellen användas?

För att modellen ska vara användbar och i praktiken stärka provens roll att fungera som betygsstöd måste den vara enkel att tillämpa för den som ska använda den, t.ex. lärare, rektorer, huvudmän och Skolinspektionen. Kravet på enkelhet vid tillämpningen gör att den officiella skalan med 0–20 betygspoäng inte är lämplig. Den sneda

fördelningen av betygspoäng gör beräkningar krångliga och svårtolkade. Utredningen förordar därför att skalan 1–6 används tillsammans med modellen.

Vi tänker oss i första hand att modellen bör användas som ett analysverktyg för lärare, rektorer, huvudmän och Skolinspektionen för att analysera provens betygsstödjande roll på skol-, huvudmanna- och nationell nivå.

På nationell nivå kan modellen användas av Skolinspektionen, eftersom den blir ett verktyg för myndigheten att granska skillnader mellan provbetyg och ämnes- eller kursbetyg. Systematiska skillnader på vissa skolor eller hos vissa huvudmän som avviker från vad som är accepterat enligt modellen kan göra att Skolinspektionen initierar en tillsyn.

På huvudmannanivå ges ett instrument för att följa upp kommunens eller den enskilda huvudmannens skolor. Modellen kan användas som en del i kvalitetsarbetet.

På skolnivå får rektorn ett analysverktyg och ges möjlighet att säkerställa kvaliteten i betygssättningen. Dessutom utgör modellen ett bra underlag i diskussioner med lärarna om deras bedömning och betygssättning.

Skolnivån innefattar även lärarna som kan använda modellen för att analysera sin bedömning och betygssättning i efterhand. En annan möjlighet som vi har undersökt är att en lärare vars provresultat och slutbetyg avviker mer än vad modellen medger ska motivera skillnaderna för rektorn. Detta förfarande skulle dock, enligt vår bedömning och i nuvarande läge, leda till att orimligt många lärare skulle behöva motivera sina avvikelser, vilket i sin tur skulle medföra en ökad administration för lärarna. Vi förordar därför att modellen åtminstone till att börja med används i efterhand för analyser av betygsnivåer och avvikelser på olika nivåer.

I stället för att använda modellen i efterhand för analys skulle den även kunna användas direkt och påverka elevernas betyg i likhet med hur modellen tillämpades i det relativa betygssystemet. Vi förordar dock inte en sådan användning i nuvarande läge. Det får bli en möjlighet som eventuellt kan undersökas vidare när det finns erfarenheter från den användning av modellen som utredningen förordar.

En liten fördjupning av modellen

Modellen som den beskrivits så här långt är förenklad genom att den utgår från en standardklass med 25 elever vid beräkningen av accepterad avvikelse (0,3 betygspoäng). Som framgått tidigare i bilagan är emellertid slumpeffekterna beroende av gruppens eller klassens storlek. I en lite mer genomarbetad modell kan man ta hänsyn till detta och därigenom få ett mer nyanserat resultat.

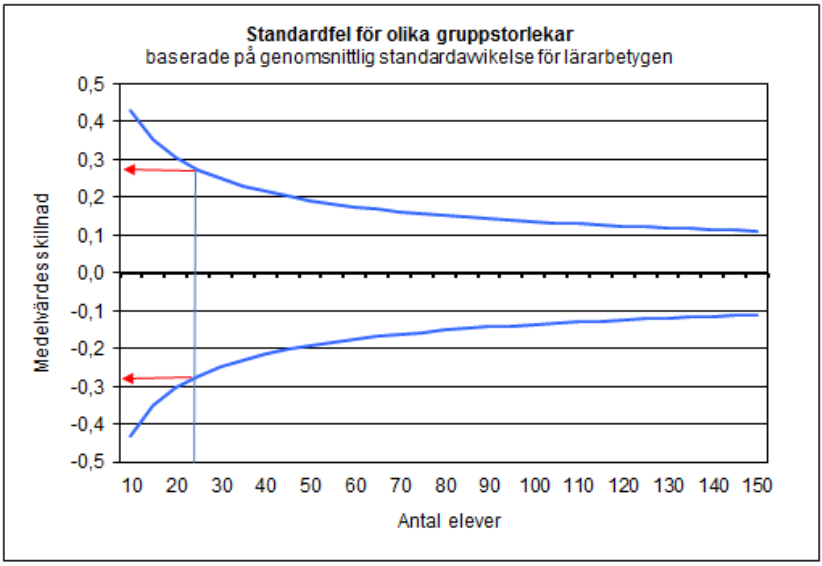
Modellen som ligger till grund för de bilder som visas beräknar avvikelsen i relation till gruppstorleken genom att utgå från den genomsnittliga standardavvikelsen för lärarbetygen för samtliga kurser[93] (jämför tabell 8 och 9). Denna standardavvikelse blir avrundat till en decimal 1,3 betygspoäng. Avvikelsen (SE)[94] kan då beräknas enligt följande:

$SE = 1{,}3/\sqrt{n}$, där n är antalet elever i gruppen eller skolan.

Ritar man upp grafen till denna beräknade avvikelse (positiv och negativ) får man nedanstående figur.

[93] Samtliga kurser kan ingå eftersom standardavvikelserna ligger nära varandra. Om skillnaderna var större skulle kanske endast standardavvikelser i samma ämnen vara lämpliga att använda.
[94] SE står för Standard Error, standardfelet eller stickprovsmedelvärdets standardavvikelse.

Figur 67 Avvikelser för grupper med olika antal elever. Baserad på medelvärdet av samtliga kursers standardavvikelse på lärarbetygen i det befintliga underlaget (tabell 8 och 9).



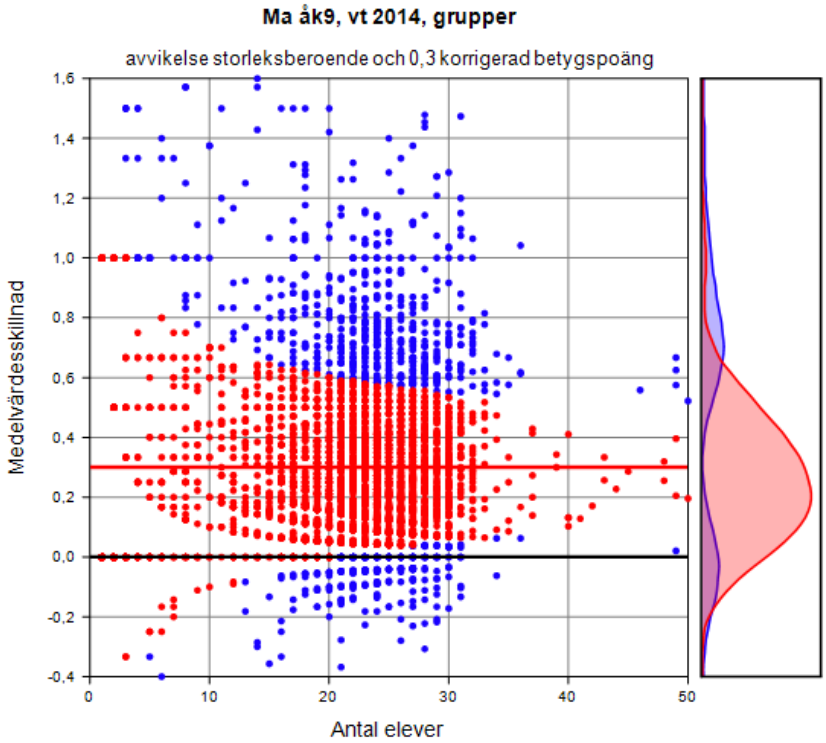
Den vertikala linjen i figuren markerar de tidigare använda värdena 0,3 betygspoäng baserat på en standardklass med cirka 25 elever.[95] Figuren visar också gränsen för accepterad avvikelse enligt den anvisning som valts (standardfelet för lärarbetygens medelvärde för stickprov av olika storlek, baserat på lärarbetygens totala standardavvikelse).

Av figur 67 framgår att för större grupper, exempelvis skolor med 150 elever, blir den accepterade medelvärdesskillnaden mellan lärarbetyg och provbetyg cirka 0,1 betygspoäng. Dessa värden kan användas för uppföljning av avvikelser på klassnivå, skolenhetsnivå, huvudmannanivå eller någon annan gruppnivå.

Figur 68 och 69 visar ovanstående variabla avvikelser tillämpade på matematik för årskurs 9. Först görs en uppdelning på grupper (klasser) och sedan på skolenheter.

[95] Värdena är avrundade och stämmer därför inte riktigt med diagrammet.

Figur 68 Fördelning av grupper (klasser) inom (röda prickar) och utanför (blå prickar) accepterad avvikelse baserad på gruppstorlek. Korrigerad för systematisk avvikelse 0,3 betygspoäng. Matematik åk 9, vt 2014.



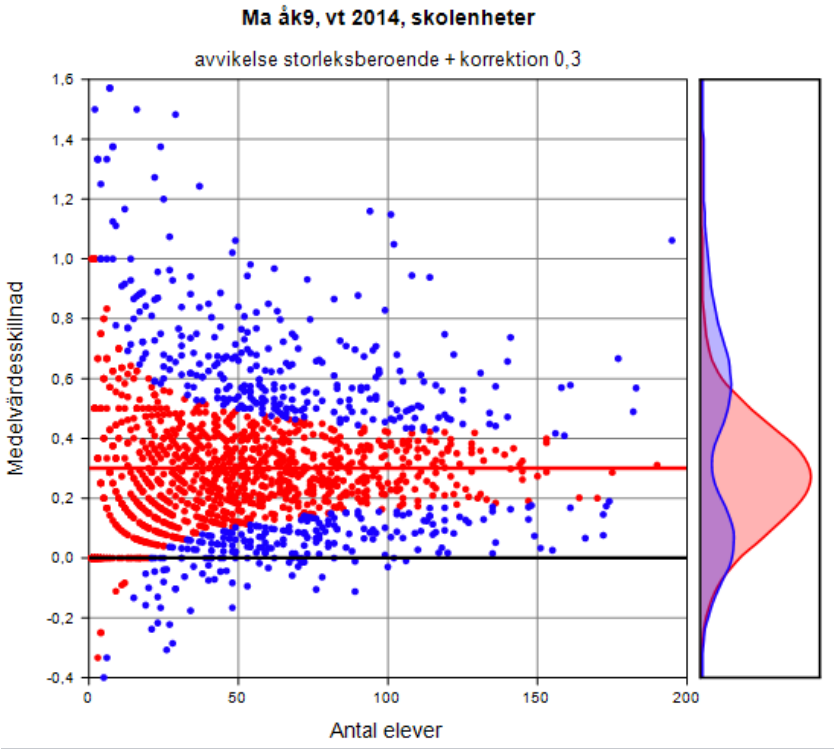
Här framgår att avvikelsen för grupper (klasser) inom intervallet 10–30 elever inte varierar särskilt mycket. Det är först när grupperna blir riktigt små som intervallet för accepterade avvikelser blir stora, dvs. slumpfelet får stor betydelse. Observera att figuren gäller matematik och därför visar värden som är systematiskt korrigerade med + 0,3 betygspoäng.

Om jämförelserna mellan medelvärden för lärarbetyg och provbetyg görs på skolenhetsnivå[96] blir resultatet som figur 69 nedan visar. Av figuren framgår att statistiken innehåller ett antal mycket

[96] Den tillgängliga statistiken använder uppdelning på skolenheter och grupper. Dessutom finns uppdelning på kommuner och huvudmän, men dessa grupperingar redovisas inte här.

små skolenheter (om statistikredovisningen kan antas vara korrekt). Figuren visar dock tydligt hur den accepterade avvikelsen minskar med ökad gruppstorlek. I figuren är avvikelserna korrigerade med 0,3 betygspoäng som gäller för matematik i underlaget.

Figur 69 **Fördelning av skolenheter inom (röda prickar) och utanför (blå prickar) accepterad avvikelse baserad på gruppstorlek och systematisk korrigering med 0,3 betygspoäng. Matematik åk 9, vt 2014.**



Den röda horisontella linjen ligger vid värdet 0,3 (den systematiska korrigeringen) och av bilden framgår att den accepterade avvikelsen för stora skolenheter med fler än 150 elever ligger mellan 0,2 och 0,4 (dvs. 0,3 plus/minus 0,1). Observera också att skolenheter med avvikelsen noll (samma poängmedelvärde för provbetyg och lärarbetyg) kommer att ligga utanför (under) intervallet för accepterad avvikelse och systematisk korrigering om de är större än 25 elever.

Det bör kanske påpekas att den valda gränsen för tillåten avvikelse kan ses som tämligen sträng. Den bygger på att gränsen är vald som ett standardfel för en slumpmässigt vald grupp av viss storlek. För den enkla modellen baserad på en ”normalklass” med 25 elever gav detta en accepterad avvikelse på 0,3 betygspoäng. Ett standardfel innebär emellertid att 68 procent av grupperna kan förväntas ligga inom det accepterade intervallet. Resterande 32 procent av grupperna kan därmed av slumpskäl förväntas ligga utanför intervallet. Det normala är att konfidensintervall väljs med två standardavvikelser, vilket ger 95 procent av grupperna inom det accepterade intervallet. Då blir intervallet dubbelt så brett, dvs. 0,6 betygspoäng i vårt exempel. Detta illustrerar dilemmat med statistiskt baserade gränser – man vet att av de grupper som ligger utanför ett accepterat intervall ligger vissa där av slumpskäl (de kan sägas vara falskt negativa), men man inte vilka de är. På motsvarande sätt kommer några grupper att av tur hamna inom intervallet (falskt positiva), men inte heller de är identifierbara ur statistiken.

I situationer när det nationella medelvärdet och standardavvikelsen för provbetyget är kända, vilket kan vara fallet när modellen används för uppföljning och analys i efterhand, skulle det faktiskt konstaterade medelvärdet för provbetyget kunna användas. Detta skulle kunna leda till att precisionen i modellens resultat ökar, eftersom det finns en viss variation i provbetygens medelvärden mellan åren.

Modellens styrkor och svagheter

Den modell som har skissats här kräver bearbetning och utökat statistiskt underlag innan beslut eventuellt fattas om ett införande. Som avslutning vill utredningen peka på några styrkor och svagheter som vi kan se i det här läget, vilka bör adresseras vid en fördjupad analys av modellen.

Styrkor

- Modellen baseras på lärarbetygens fördelning. Därigenom får den samlade lärarkårens tolkningsföreträde när det gäller kunskapskravens innebörder, elevernas kunskaper och relationen

dem emellan. Å andra sidan ligger lärarbetygens och provbetygens standardavvikelser i allmänhet nära varandra, så om underlaget baseras på det ena eller andra betygets varians är i praktiken egalt. Det ligger dock ett symboliskt värde i att utgå från lärarbetygen.

- Provbetygen får en ökad roll vid den slutliga betygssättningen genom att en avvikelse utanför ett visst fastslaget intervall innebär att betygssättningen kan komma att granskas på skol-, huvudmanna- eller nationell nivå. Det här kan förväntas skapa ett visst tryck på lärarna att vid betygssättningen inte gå utanför accepterade gränser.
- Modellen kan med tiden förväntas innebära att gränserna för acceptabel avvikelse blir snävare eftersom tillkommande lärarbetyg när modellen är i drift kan komma att få en mindre standardavvikelse, vilket minskar standardfelet och därmed intervallet för accepterad avvikelse. Om detta verkligen sker behöver dock prövas empiriskt.
- Modellen borde årligen kunna tillhandahållas av Skolverket med uppdaterade parametrar och sedan vara enkel att tillämpa för lärare, rektorer, huvudmän och Skolinspektionen för granskning på olika nivåer.
- Modellen borde inte innebära fördyring och extra insatser av de lärosäten som konstruerar proven.

Svagheter

- Modellen är oprövad. Den bygger på den statistiska parameter och på de data som är mest stabila vid analysen av tidigare prov- och betygsdata, nämligen lärarbetygens spridning (standardavvikelse). Metoden blir därmed i huvudsak empirisk och utan stabil teoretisk grund.
- Det är oklart hur modellen kommer att bemötas. Kommer den att anses tillämplig i ett mål- och kunskapsrelaterat system?
- Olika principer gäller för lärares betygssättning (icke-kompensatorisk) och provbetyg (mer eller mindre kompensatorisk, varierar med olika ämnen).

- Det är oklart hur förslaget om systematisk korrektion för vissa ämnen och kurser kommer att tas emot.
- Det saknas vedertagna teorier om varför olika ämnen ligger på olika betygsnivåer.
- Beräkningarna baseras på att betygsskalan poängsätts från 1 till 6 med ett steg mellan varje betyg. Det medför avsteg från nu gällande skala, och det kan framstå som opraktiskt med två parallella skalor för olika ändamål.
- Det kan finnas en risk att den övre gränsen för avvikelse i praktiken ses som normalvärde för avvikelsen, vilket kan leda till betygsglidning uppåt.

Referenser

Betänkandet Tydliga mål och kunskapskrav i grundskolan. Förslag till nytt mål- och uppföljningssystem (SOU 2007:28). Statens offentliga utredningar, Stockholm: Fritzes.

Danmarks Evalueringsinstitut (2013). 7-trins-skalaen Evaluering af anvendelsen af karakterskalaen. (Endast publicerad i elektronisk form.)
https://www.eva.dk/projekter/2012/evaluering-af-7-trins-skalaen/projektprodukter/7-trins-skalaen/view

Kingdon, M. (2009). Marks and grades. A review of underlying issues. SQA Research report 6, Scotland: Scottish Qualifications Authority.

Korp, H. (2006). Lika chanser i gymnasiet. En studie om betyg, nationella prov och social reproduktion. Malmö studies in educational sciences, No 24. Malmö högskola.

Skolverket (2003). Det nationella provsystemet – vad, varför och varthän? Dnr 2003:2038.

Skolverket (2004). Det nationella provsystemet i den målstyrda skolan. Stockholm: Fritzes

Skolverket (2011a). Läroplan, examensmål och gymnasie gemensamma ämnen för gymnasieskola 2011. Stockholm: Fritzes.

Skolverket (2011b). Skillnaden mellan betygsresultat på nationella prov och ämnesbetyg i årskurs 9, läsåret 2010/11. Dnr 2011:14.

Skolverket (2013a). Redovisning av uppdrag om avvikelser mellan provresultat och betyg i grundskolans årskurs 9. Dnr 2013:00164.

Skolverket (2013b). Redovisning av uppdrag om avvikelser mellan provresultat och kursbetyg i gymnasieskolan. Dnr 2013:00164.

Skolverket (2015a). Redovisning av uppdrag om relationen mellan provresultat och betyg i grundskolans årskurs 6 och årskurs 9. Dnr 2015:205.

Skolverket (2015b). Redovisning av uppdrag om avvikelser mellan provresultat och kursbetyg i gymnasieskolan. Dnr 2015:205.

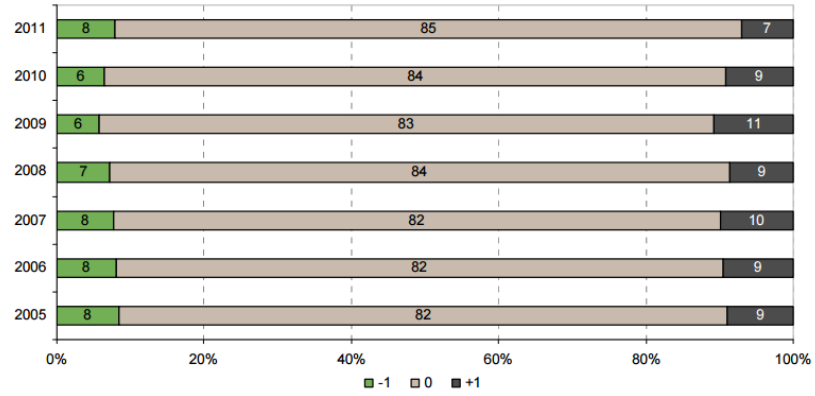
Skolverket (2015c). Skolreformer i praktiken. Hur reformerna landade i grundskolans vardag 2011–2014. Stockholm: Wolters Kluwers kundservice.

Undervisningsministeriet, Uddannelsesstyrelsen. (2004). Betænkning om indførelse af en ny karakterskala til erstatning af 13-skalaen afgivet af karakterkommissionen, november 2004 Köpenhamn: Undervisningsministeriets forlag.

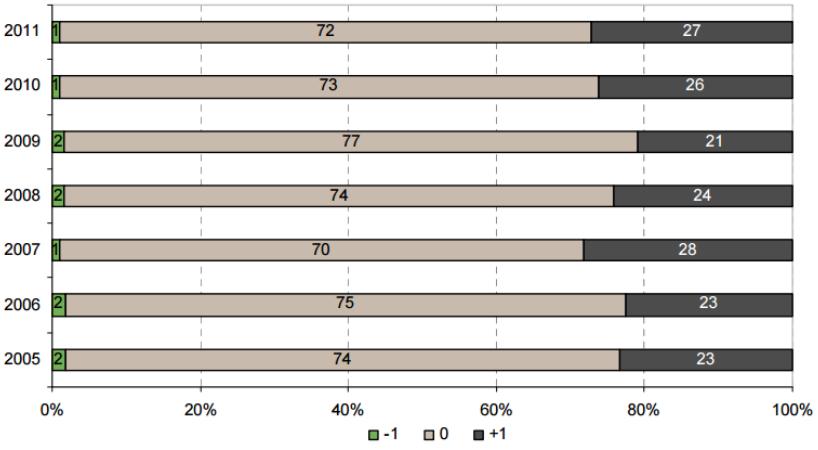
Appendix 1

Grundskolan, årskurs 9[97]

Figur 70 Nettoavvikelser 2005–2011. Engelska, åk 9.

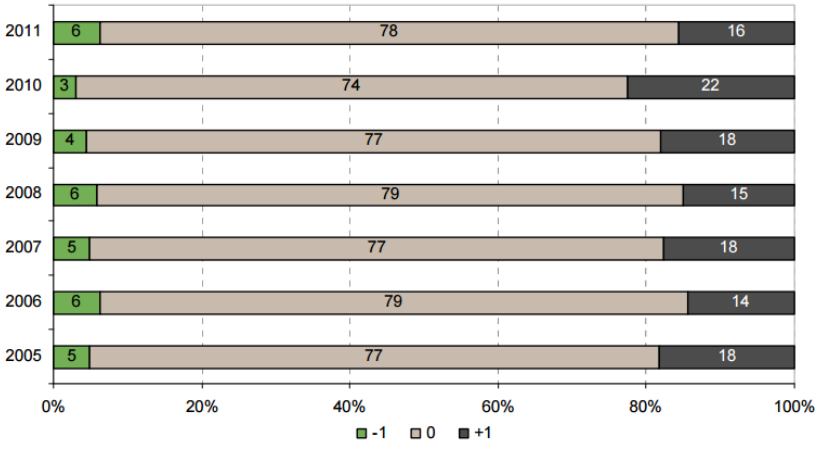


Figur 71 Nettoavvikelser 2005–2011. Matematik, åk 9.



[97] Skolverket (2011b).

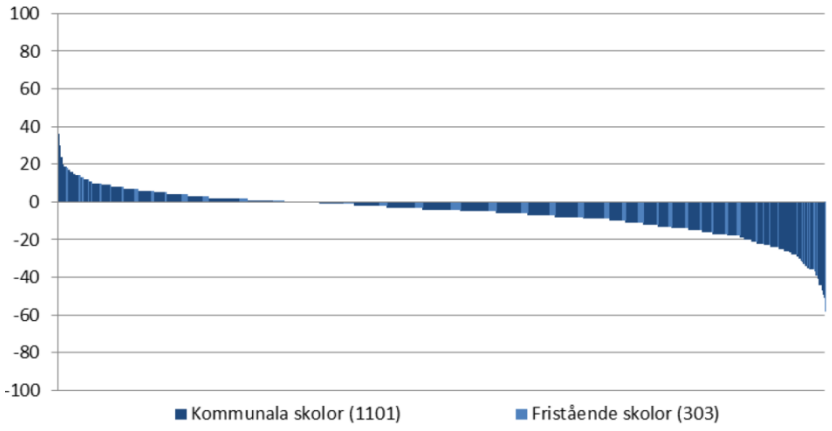
Figur 72 Nettoavvikelser 2005–2011. Svenska, åk 9.



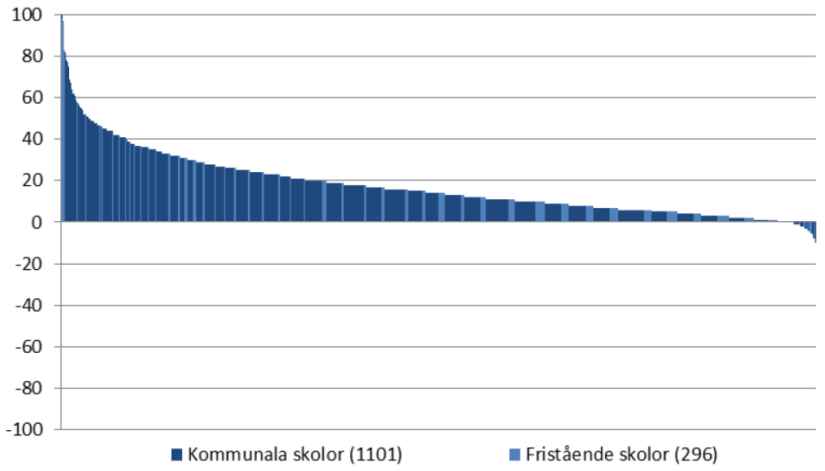
Appendix 2

Grundskolan, årskurs 9[98]

Figur 73 Nettoavvikelser, skolor med minst 15 elever. Engelska, vt 2014.



Figur 74 Nettoavvikelser, skolor med minst 15 elever. Matematik, vt 2014.



[98] Ur Skolverket (2015a).

Appendix 3

Tabell 10 Programförkortningar i gymnasieskolan.

Program	Programförkortning
Yrkesprogram	
Bygg och anläggning	BA
Barn och fritid	BF
El och energi	EE
Fordon och transport	FT
Handel och administration	HA
Hotell och turism	HT
Hantverk	HV
Industri	IN
Naturbruk	NB
Restaurang och livsmedel	RL
VVS och fastighet	VF
Vård och omsorg	VO
Högskoleförberedande program	
Ekonomi	EK
Estetiska	ES
Humanistiska	HU
Naturvetenskap	NA
Samhällsvetenskap	SA
Teknik	TE
Introduktionsprogram	*IM*
Preparandutbildning	IMPRE
Programinriktat individuellt val	IMPRO
Individuellt alternativ	IMIND
Språkintroduktion	IMSPR
Yrkesintroduktion	IMYRK
Övriga	
International Baccalaurate	IB
Riksrekryterande utbildningar	RX

Appendix 4

SKOLFS 1992:42

Utkom från trycket den 20 november 1992.

Skolverkets föreskrifter
om betygsättning på linjer och specialkurser i gymnasieskolan
1992-11-09

Med stöd av 4 kap. 1 § fjärde stycket förordningen (1992:396) med vissa bestämmelser för linjer och specialkurser i gymnasieskolan föreskriver Skolverket följande.

Inledande bestämmelser

1 § Dessa föreskrifter gäller betygsättning på linjer och specialkurser i gymnasieskolan.

Föreskrifterna i 3–8 §§ gäller inte för sådana studievägar och ämnen som avses i 9–11 §§.

Vad betyget avser

2 § Betyg som sätts innan ett ämne har avslutats, avser den del av årskursen som eleven har genomgått.

När ett ämne har avslutats, avser betyget hela kursen i ämnet.

Om en elev avgår från gymnasieskolan utan att ha slutfört kursen i ett ämne, avser betyget i ämnet den del av kursen som eleven har genomgått.

Indelning i referensgrupper

3 § Vid betygsättning enligt skalan 1–5 skall eleverna indelas i referensgrupper.

4 § De elever i riket som läser ett ämne enligt samma kursplan skall för varje årskurs utgöra en referensgrupp för ämnet i fråga.

5 § Utan hinder av 4 § skall elever som följer samma kursplan hänföras till olika referensgrupper, om de

1. tillhör de tvååriga linjerna respektive de tre- och fyraåriga linjerna,
2. läser ett visst språk som B- respektive C-språk,
3. läser allmän respektive särskild kurs i engelska,
4. läser ett ämne enligt olika timplaner och skillnaden i veckotimtal för en hel lärokurs i ämnet är minst tre timmar.

Betygsfördelning

6 § Vid betygsättningen skall, om inte något annat följer av 7 och 8 §§, beaktas att betygen för en referensgrupp i riket bör fördela sig på i huvudsak följande sätt:

Betyg	1	2	3	4	5
Procent	7	24	38	24	7

7 § Vid betygsättning i matematik, fysik, kemi och biologi på naturvetenskaplig och teknisk linje skall, om inte något annat följer av 8 §, beaktas att betygen för referensgruppen i riket bör fördela sig i huvudsak på följande sätt:

Betyg	1	2	3	4	5
Procent	5	18	34	28	15

8 § Vid betygsättning i högre årskurs skall förändringen av elevgruppen från föregående årskurs beaktas. Utgångspunkten skall vara att elever inte skall få sina betyg höjda eller sänkta endast därför att andra elever avslutat, avbrutit eller påbörjat studier i ämnet.

Treåriga yrkesinriktade studievägar och vissa små ämnen

9 § Betyg inom försöksverksamheten med treåriga yrkesinriktade studievägar skall sättas så att de motsvarar de betyg som ges på tvååriga linjer för samma prestationer.

10 § I ämnen som läses av mycket små grupper på enskilda studievägar skall riktvärdet för betygsnivån inom elevgruppen bestämmas med ledning av elevernas betyg i övriga ämnen, om inte något annat följer av 11 §.

11 § I ämnena specialidrott, bild – estetisk specialisering och musik – estetisk specialisering skall riktvärdet för betygsnivån inom elevgruppen bestämmas med ledning av elevernas betyg i idrott, bild respektive musik.

Betygsättning med ledning av centrala prov

12 § Betygen i ett ämne där ett centralt prov har getts under årskursen skall anpassas till fördelningen och medelvärdet på det centrala provet. Vid beräkningen av medelvärdet skall för elever som inte har deltagit i provet senast givna betyg ersätta provresultatet.

13 § Resultaten av de centrala proven bör, inom ramen för bestämmelserna 6–11 §§, beaktas även vid betygsättningen i ämnen där centrala prov inte ges.

Åtgärder vid betygsättningen

14 § Betygsättningen skall diskuteras mot slutet av varje termin i den ordning som rektor beslutar.

15 § Lärarna skall före det tillfälle som avses i 14 § sätta preliminära betyg för sina undervisningsgrupper och beräkna betygens medelvärde för varje grupp. Elever som inte har fått betyg i ett ämne skall inte medräknas då medelvärdet fastställs.

16 § Rektor skall ombesörja att översikter av betygsfördelning och betygsmedelvärde samt i förekommande fall resultat på centrala prov upprättas för varje referensgrupp eller motsvarande grupp enligt 9–11 §§ som finns vid skolan. Sådana översikter skall vara tillgängliga före betygsdiskussionen.

17 § I ett ämne med centralt prov får betygsmedelvärdet för en undervisningsgrupp avvika från provresultatet i gruppen med högst 0,20 betygsenheter, om det inte finns särskilda skäl för en större avvikelse. En sådan större avvikelse skall motiveras skriftligt.

18 § Om betygsnivån eller betygsfördelningen i en lärares undervisningsgrupp, i andra fall än som avses i 17 §, påfallande avviker från övriga lärares betygsättning eller från gruppens resultat på de centrala proven, skall läraren justera sin betygsättning eller skriftligen motivera den.

19 § Om en lärare vill ändra sin betygsättning efter det tillfälle som avses i 14 § så att medelvärdet för gruppen ändras, skall läraren anmäla detta till rektor som avgör om betygen skall diskuteras på nytt.

Dessa föreskrifter träder i kraft den 1 december 1992.

LENA LANDGREN

Åke Margell

Provsystem i förändring

För att kunna lämna välgrundade förslag till förändringar av ett provsystem behövs en historisk och teoretisk plattform att utgå ifrån. Det gör det möjligt att analysera några av de förutsättningar och principer som kan ha relevans för att förstå grunden för ett provsystems syfte och utformning. I den här bilagan försöker utredningen att lägga en sådan plattform. Det är dock svårt att avgränsa frågan och framställningen blir därför med nödvändighet schematisk och övergripande. Syftet med bilagan är inte heller att ge en heltäckande bild utan att försöka forma en viss referensram för att kunna bedöma vilka frågeställningar som behöver hanteras vid utformningen av ett provsystem.

Utgångspunkter

Paul Newton[1] diskuterar i en artikel frågor som rör olika former av prov och bedömning utifrån provens syften. Han tar upp några olika begrepp och deras funktion och roll för prov med olika syften. Newton skiljer på *judgement* (bedömning), *decison* (beslut) och *impact* (påverkan). I nedanstående text används några av Newtons begrepp som underlag för att dela upp de senaste 75 åren i tre perioder som i Sverige kan anses ha dominerats av mer eller mindre skilda typer av läroplaner. Dessa representerar delvis olika kunskapssyn och olika principer för kunskapsmätning, bedömningskulturer och beslutsfattande (betygssättning). Fokus ligger på Sverige, men en stor del av de förändringar som skett i Sverige har sina rötter utomlands, främst i USA.

[1] Newton (2007).

I det här sammanhanget är också den nivå som ligger under nivån *bedömning* intressant, dvs. den nivå som gäller sättet att generera empiriska underlag för bedömning och beslutsfattande. Denna underliggande nivå kallas här *mätning*, vilket i det här sammanhanget innefattar de åtgärder som vidtas för att ge elever möjligheter att prestera utsagor som kan bedömas och betygssättas, dvs. det vi kallar prov.[2] Själva bedömningen av utsagan kan vara av olika slag beroende på utsagans form. Den kan bestå av allt från okomplicerade uppgifter med fasta svarsalternativ som kan bedömas utan krav på tolkning och som ger underlag för statistiska analyser, till komplexa muntliga eller skriftliga framställningar där bedömningen blir starkt beroende av bedömarens tolkning.

I den här texten kommer betoningen att ligga på nivåerna mätning, bedömning och beslut. *Impact* (påverkan) som syftar på de avsedda eller icke avsedda effekter som proven och provsystemet får behandlas mer översiktligt. *Impact* kan exempelvis gälla att prov motiverar till ökat lärande och ökade ansträngningar (positiv avsedd effekt). Samtidigt kan det medföra att undervisningen blir mer styrd av vad som tas upp på proven än vad som anges i läroplanen (negativ icke avsedd effekt).

Olika kurs- och ämnesplaner ligger till grund för kunskapsmätningar, bedömningar och beslut om betyg. Det som betonas i kurs- och ämnesplanerna kan variera över tid och är ofta beroende av den rådande synen på vad kunskap är och hur människor lär. I ett visst skede kan mer faktainriktade kunskaper och specifika färdigheter dominera. I en annan tid kan betoningen ligga på mer generella färdigheter, kompetenser, förmågor eller vad man föredrar att kalla det undervisningen syftar till att utveckla hos eleven. Några klara och tydliga gränser mellan olika faser av kunskaps- och ämnessyn finns oftast inte utan de utgörs snarare av förlopp som successivt övergår i varandra för att vid en viss tidpunkt bli manifesterade i en formell läroplan. Det betyder att varje fas omges av gråzoner och i varje fas finns större eller mindre inslag av det tänkande som dominerar föregående och efterkommande faser.

[2] Mätning ska således ses mer som det som på engelska kallas assessment. Mätning uppfattas vanligen som tilldelning av kvantitativa mätenheter, medan assessment kan innefatta såväl detta som värdering i kvalitativa termer.

Kunskapsprov[3] baseras på och ska ha en tydlig koppling till motsvarande kurs- och ämnesplaner.[4] Detsamma gäller för undervisningen. Samtidigt har prov en återkopplande verkan på undervisningen vilket innebär att det finns en inbördes växelverkan mellan kurs- och ämnesplaner, undervisning och kunskapsprov. Genom denna koppling präglar således kurs- eller ämnesplanens kunskapssyn proven medan proven i sin tur tenderar att influera undervisningen och modifiera lärarens tolkning av kurs- eller ämnesplanen[5]. Tillsammans med bedömningsanvisningarna har proven även en normgivande roll när det gäller bedömning av de svar eleverna ger på uppgifterna, liksom för tolkningen av kunskapskraven och principerna för beslut om betyg. Detta beroende gör att kurs- eller ämnesplanens och läroplanens kunskapssyn via proven i hög grad avgör sättet att *mäta*, *bedöma* och *besluta* om utfallet av undervisningen. I det här sammanhanget innebär detta faserna: konstruktion av prov (mätande), bedömning och betygssättning av elevens provresultat (bedömning) och slutlig betygssättning baserad på elevens samlade resultat (beslut).

Tiden från 1940-talet – då grunderna för ett mer modernt sätt att konstruera och använda prov lades i Sverige – fram till i dag kan i huvudsak indelas i tre perioder. En första period med definitiv start i och med Lgr 62 som dominerades av grupprelaterade prov och betyg. En andra period med avstamp i 1994 års läroplaner då kriterierelaterade prov och betyg blev framskrivna. Och slutligen den period i vilken vi nu befinner oss där man snarast kan tala om standardsrelaterade eller standardsbaserade prov och betyg. De två första perioderna är tämligen klart åtskiljbara, men när det gäller kriterierelaterade prov och betyg i relation till de som gäller i den nuvarande tredje perioden blir skiljelinjerna mer oklara.[6]

Periodindelningen är således godtycklig och utgår från de svenska läroplaner som har sammanfallit med införandet av principiellt ändrade regler för betygssättning.

[3] Begreppet kunskap är inte entydigt. Det antas i det här sammanhanget innefatta det som anges som kunskap i den läroplan och kurs- eller ämnesplan som ligger till grund för provet.
[4] Det som på engelska kallas *alignment* och som saknar någon entydig svensk översättning. Även på engelska är begreppet något oklart. Se. t.ex. http://edglossary.org/alignment/
[5] Det som ibland kallas *implemented curriculum* på engelska. Se t.ex. (Unesco, 2013).
[6] Se t.ex. http://edglossary.org/standards-based/

Period 1 – normrelaterade prov och betyg

Den första perioden kan anses sträcka sig från mitten av 1940-talet till mitten av 1990-talet. På 1940-talet började utbildning få en allt större betydelse och frågan om införande av en sammanhållen nioårig grundutbildning för alla barn fick allt större uppmärksamhet. Andelen elever som ville gå vidare efter folkskolan hade efterhand ökat. Dåtidens betyg var mycket varierande och likvärdigheten mellan olika lärares och skolors betyg ansågs alltför bristfällig för att ligga till grund för urval till realskolor och läroverk. Regeringen gav därför matematiklektorn Frits Wigforss vid lärarseminariet Rostad i Kalmar i uppdrag att utreda förutsättningarna för att införa ett modernt och på rådande vetenskap baserat betygs- och provsystem.

Wigforss betänkande[7] kan sägas vara startpunkten för att införa ett prov- och betygssystem som var utformat utifrån en rationell teoretisk grund. Det skiljde sig mot tidigare system som till stor del hade baserats på examinerande bedömning under överinseende av erfarna och omdömesgilla personer med allmänt gott renommé (för gymnasieskolans och realskolans del benämnda censorer). Gymnasieskolans censorer utgjordes främst av representanter för universitet och högskolor. Ur ett mer modernt teoribaserat perspektiv kan de censorsbaserade examinationssystemen ses som kvarlevor från en tid då examinationer sågs som en form av initiationsriter till vuxenlivet. Studentexamen kallades för övrigt fram till 1905 mogenhetsexamen i Sverige. Först i mitten av 1960-talet gick systemet med censorer i graven.

Kunskapssyn och mätning

Läroplaner kan som tidigare nämnts betona olika aspekter av kunskaper och lärande. En inriktning kan vara att i läroplanerna lyfta och tydligt ange vilket *kunskapsstoff*[8] som ska behandlas. För svensk del kan detta sägas vara den syn som präglade de läroplaner

[7] SOU 1942:11.

[8] Med kunskapsstoff avses då de fakta och metoder (procedurer, algoritmer, regler etc.) som föreskrivs i kursplanen. Det som Betygsberedningen (SOU 1992:86) benämnde kunskaper i ämnet (till skillnad mot kunskaper om ämnet).

som tillkom efter 1950-talet, dvs. Lgr 62 och Lgr 69 för grundskolan samt Lgy 70 för gymnasieskolan.

Dessa läroplaner representerade en kunskapssyn som lämpade sig väl för mätning med det synsätt som då var dominerande inom den pedagogiska mätningsläran. Tydliga stoffangivelser innebar väl avgränsade och inramade kunskaper[9]. Detta medförde att prov med hög *innehållsvaliditet* (god stofftäckning[10]) och betoning på goda tekniska mätegenskaper (hög reliabilitet) kunde konstrueras. Hög reliabilitet främjas av prov med många uppgifter som mäter väl avgränsade kunskaper och genererar entydigt tolkningsbara svar, gärna i form av fasta svarsalternativ. Ett sådant system lämpar sig väl för statistisk rangordning av provdeltagarna och därmed även för ett grupprelaterat betygssystem vars främsta syfte var att ge god grund för urval till efterföljande utbildningar, dvs. prov som hade hög *prognostisk validitet*.

Bedömning och betygssättning av elevens provresultat

Bedömningen (judgement) av uppgifterna och konstruktionen av det samlande provresultatet blir i ett norm- eller grupprelaterat[11] system en tämligen enkel procedur eftersom de ingående uppgifterna i allmänhet är lättbedömda och inte ger särskilt stort tolkningsutrymme vid bedömningen.[12] Slutresultatet baseras därmed främst på en sammanräkning av respektive deltagares provpoäng.

Betygssättningen av provet kan även den ske genom tillämpning av fastställda rutiner. För svensk del skedde det genom att vid konstruktionen av betygsskalan utgå från en normalfördelning med medelvärdet 3, standardavvikelsen 1 och fem betygssteg. Varje betygssteg representerades av en standardavvikelse. Betyget 3 tilldelades de elever i normgruppen som låg i poängintervallets mitt +/- en halv standardavvikelse, betyget 2 till de elever som låg i intervallet -0,5 till -1,5 standardavvikelse under medelvärde, och betyget 1 till dem som låg ytterligare en standardavvikelse under

[9] Bernstein (1971).
[10] Hög *content validity* på engelska.
[11] Ibland också kallat relativt.
[12] Gäller förstås inte uppgifter av typen uppsatser och på 1950-talet fanns därför förslag om att sådana inte skulle ingå i dåtidens nationella prov (standardprov för grundskolan och centrala prov för gymnasieskolan). Förslaget avvisades dock av validitetskäl.

mittpoängen. På motsvarande sätt definierades betygen 4 och 5. Detta gav en betygsfördelning för normgruppen med procentsatserna (7, 24, 38, 24, 7) för betygen 1 till 5. Sedan kunde givetvis fördelningen i enskilda klasser se olika ut beroende på vilken sammansättning de råkade ha. Normen användes på nationell nivå för att fastställa poänggränser på nationella prov.

Betygssättning av eleven

Betygssättningen var styrd av de nationella proven och lärarnas slutgiltiga betyg var likaledes styrda av provresultaten, dock inte ovillkorligt. För gymnasieskolan fanns en föreskrift om att gruppens betygsmedelvärde inte fick avvika med mer än 0,2 betygssteg i relation till betygsmedelvärdet för provet. En högre avvikelse var inte förbjuden, men den skulle motiveras för rektor och kollegium. Beslutsrätten låg således hos läraren med stöd av provresultatet men med föreskrift om lokal övervakning av överensstämmelsen mellan provbetygens och de egna betygens medelvärden. För betygssättningen av enskilda elever hade dock läraren ensam beslutsrätt.

Kunskapsbedömningen baserade således i huvudsak på mätning med instrument där reliabiliteten var betonad. Proven konstruerades med utgångspunkt i den klassiska testteorin som i huvudsak brukar betraktas som färdigutvecklad i slutet av 1960-talet.[13] Besluten om betyg på provet grundades på en teoretisk modell (normalfördelningen). Modellen för fördelning var därmed teknisk och krävde ingen bedömning utöver den som eventuellt ingick i själva bedömningen av proven. Provpoängen angav rangordning av eleverna och elevens betyg på provet bestämdes av de poänggränser som angavs för olika betyg.[14]

För läraren gällde att beslutet om den enskilda elevens betyg skulle baseras på elevens samlade kunskaper vilket innebar att rangordningen på provet inte hade någon formellt styrande funktion i

[13] T.ex. Gulliksen (1950), Lord & Novick (1968).
[14] För vissa prov eller provdelar som inte baserades på poäng t.ex. uppsatser, krävdes dock mer helhetlig bedömning och betygssättning. Detta innebar lägre reliabilitet, vilket tidvis ledde till förslag om att uppsatser inte borde ingå i nationella prov. Av validitetsskäl avslogs dock sådana förslag.

det avseendet. Läraren var således fri att ändra på rangordningen mellan olika elever. Däremot gällde som tidigare nämnts för gymnasieskolan att om gruppens betygsmedelvärde avvek från betygsmedelvärdet på provet med mer än 0,2 betygssteg skulle läraren motivera sin betygssättning för rektor och kollegium. Det handlade alltså inte om ett ovillkorligt förbud mot större avvikelse.

Värdeord av typen godkänd, väl godkänd etc. fanns inte i det femgradiga relativa betygssystem som gällde från 1960-talet. Någon gräns för godkänt definierades heller aldrig. Sådana gränser kan i stället sägas ha blivit indirekt definierade genom efterföljande utbildningar där det t.ex. var vanligt att betyget 3 sattes som krav för behörighet. Därigenom kom betyget 3 ofta att uppfattas som en godkändgräns även om den inte var fastställd i någon förordning.

Kommentar

När grupprelaterade prov används som underlag för betygssättning blir det viktigt att avgöra vilken grupp som är normgrupp, dvs. för vilken grupp den föreskrivna procentuella fördelningen av betyg ska gälla. För de standardprov och centrala prov som introducerades i Sverige utgjorde normgruppen alla elever som ett visst år genomförde ett visst prov.

Det innebär att varje ny årskurs blir sin egen normgrupp, varför metoden ibland kallas *kohortrelaterad*. Detta är knappast något problem om det handlar om stora grupper som inte ändras nämnvärt mellan olika år. Om det däremot gäller relativt små grupper kan skillnaden i deras prestationer mellan olika år bli så stor att det får betydelse för en elev om hon eller han har deltagit i provet det ena eller det andra året. Eftersom resultatet i ett grupprelaterat system beror på övriga provdeltagares prestationer är det inte enbart den egna prestationen som avgör provbetyget eller rangplatsen. Detta är en form av orättvisa som talar mot den typen av grupprelaterat system.

Ett annat sätt är att lägga fast en normgrupp, t.ex. de elever som gjorde provet år x och sedan rangordna efterföljande grupper i relation till den ursprungliga gruppen. Detta tillvägagångssätt har inte samma svagheter som den tidigare beskrivna kohortrelaterade metoden, men innebär å andra sidan tekniska svårigheter när det

gäller att ange förhållandet mellan provresultat och betygsgränser för olika årgångar så att det blir lika svårt eller lätt att få ett visst betyg olika år.

För svensk del kan man säga att den kohortrelaterade metod som användes i grundskolan inte hade de negativa effekterna eftersom elever i grundskolan inte behövde konkurrera om några platser med elever från andra årgångar. Dessutom genomfördes proven av hela årskullar. För gymnasieskolans del kunde det däremot få ganska stor betydelse eftersom många kurser var förhållandevis korta med små elevgrupper. Dessutom kunde elevgrupper ändras under utbildningens gång genom avhopp eller byte av linje, vilket ofta medförde att betygsfördelningen ändrades. Detta, plus själva utgångspunkten att för betygen var jämförelsen med andra elever viktigare än de faktiska kunskaperna, utgjorde grunden för den kritik som så småningom ledde till nedläggningen av det gruppreleterade prov- och betygssystemet.

Trots de fördelar som låg i de strikta tillvägagångssätten när det gällde konstruktionen, bedömningen och betygssättningen av proven, liksom för besluten om elevens betyg, så ledde kritiken mot det relativa systemet till slut till ett byte av prov- och betygssystem. En nackdel som visat sig med tiden är att betygsmedelvärdena i olika ämnen numera hamnar på skilda nivåer i betygsskalan när de inte normeras i relation till ett föreskrivet medelvärde.

Period 2 – kriterierelaterade prov och betyg

Det norm- eller grupprelaterade systemet hade sina rötter i de test som utvecklades i främst USA under det tidiga 1900-talet. Syftet där och då var till stor del urval av rekryter till den amerikanska militären. Eftersom syftet var urval var den prognostiska validiteten central. De grupprelaterade proven vann efterhand insteg även i utbildningsväsendet och den klassiska testteorin utvecklades snabbt under de följande decennierna.

För skolans del började dock efterhand kritik riktas mot de normrelaterade proven med utgångspunkt i att det inte bara var intressant att på ett reliabelt sätt rangordna och sortera eleverna sinsemellan utan också att veta vad de faktiskt kunde och borde kunna för att tilldelas olika betyg.

Kunskapssyn och mätning

Det nya sättet att se på prov krävde en delvis annan ansats vid konstruktionen. De kriterier för faktiska kunskaper som användes gällde inte enbart att uppnå en viss poängnivå. Nu gällde det också att eleven visade att hon eller han i tillräcklig utsträckning klarade av uppgifter som bedömdes ligga på den kvalitativa nivå som ett visst kriterium angav. Eleven skulle alltså ligga över en viss betygsgräns eller kravgräns (*cut off*-gräns)[15]. Därmed kunde man säga att en elev med det eller det betyget hade visat sig klara uppgifter på en viss kvalitativ nivå.

I USA började idéerna om kriterierelaterade prov få genomslag i slutet av 1960-talet. Efterhand blev uppslutningen allt större också i Sverige och i mitten av 1990-talet fick det kriterierelaterade tänkandet tydligt genomslag när de nya läroplanerna Lpo 94 och Lpf 94 beslutades. För grundskolans del hade det grupprelaterade systemet börjat upplösas redan med Lgr 80. Då övergavs den strikta betygsfördelningen efter givna procentsatser och ersattes av en mer svävande rekommendation om att betyget 3 skulle vara det vanligaste betyget och att andelen 2:or och 4:or skulle vara större än andelen 1:or och 5:or. För gymnasieskolans del gällde dock den på normalfördelningen baserade procentuella fördelningen i huvudsak oförändrad fram till 1994.[16]

Det nya betygssystem som introducerades 1994 fick efter vissa diskussioner benämningen *mål- och kunskapsrelaterat*. Det är en för Sverige unik beteckning och det är svårt att se på vilket sätt den preciserar vad det är som ska betygssättas eller hur det ska gå till. Det finns knappast något betygssystem som inte gäller mål och kunskaper. Någon tydlig hänvisning till den etablerade benämningen kriterierelaterade betyg finns inte. Måhända var det kursplanernas nydanande konstruktion med *mål att uppnå* (samtidigt kriterier för betyget G i grundskolan) och *mål att sträva mot* som den valda benämningen syftade på.

[15] I praktiken innebär det dock ofta att den kvalitativa nivån definieras med en bestämd poänggräns.
[16] Med vissa undantag som gjordes för att exempelvis öka andelen elever på vissa linjer och med vissa avvikelser i slutet av perioden då flera olika system användes parallellt.

Även kunskapsbegreppet fick en ny innebörd genom de resonemang som fördes i utredningen *Skola för bildning*[17] som låg till grund för 1994 års reform. Den tidigare betoningen av kunskapsstoff och färdigheter med rötter i främst behavioristiska[18] och konstruktivistiska[19] teorier tonades ner. I stället introducerades ett modernare kunskapsbegrepp som delvis baserades på konstruktivistiska idétraditioner, men främst på sociokulturella teorier om kunskapens natur och lärandets villkor.[20] Denna kunskapssyn kom att få benämningen ”de fyra f:en: fakta, förståelse, färdighet och förtrogenhet”. Dessa kunskapsformer stod inte i något hierarkiskt förhållande till varandra utan

> ... de kompletterar varandra och utgör varandras förutsättningar. Kortfattat kan de beskrivas som:
>
> Fakta är kunskap som information.
>
> Förståelse är kunskap som meningsskapande.
>
> Färdighet är kunskap som utförande.
>
> Förtrogenhet är kunskap som omdöme.(s. 80).

Bedömning

Det kriterierelaterade systemet, inte minst i sin svenska form, krävde delvis annorlunda prov och framför allt annorlunda bedömningsanvisningar för bedömning av proven och betygssättning. Det handlade inte längre om att relatera resultaten till en skala med förutbestämda egenskaper (medelvärde och standardavvikelse) och inte heller om att jämföra eleverna med varandra. I det nya systemet handlade det om att bedöma elevernas kunskaper i relation till i text framskrivna kriterier om vilka kunskaper som skulle visas för att berättiga till de olika betygen.

För svensk del var tilltron till det nya systemet stor på den politiska nivån. Där uppfattades de nya textbaserade betygskriterierna i sig tillräckliga som underlag för lärarnas betygssättning och därmed skulle några nationella betygsstödjande prov egentligen

[17] SOU 1992:94.
[18] Lgr 62.
[19] Lgr 80.
[20] Se kapitel 2 i SOU 1992:94.

inte längre behövas. Det visade sig dock att så inte var fallet. Bestämningen av kriteriegränser (betygsgränser) visade sig vara en betydligt besvärligare uppgift än den som gällt för de normrelaterade proven. Själva proven behöver inte se särskilt annorlunda ut. De ska liksom tidigare uppfylla krav på validitet och reliabilitet. Det som blev det stora problemet var (och är) fastställandet av kravgränser[21] för olika provbetyg.

Eftersom kriterierna är formulerade i text måste de tolkas och de flesta texter kan tolkas mer eller mindre olika. Detta fick redan den svenska nestorn Fritz Wigforss – som utredde förutsättningarna för ett svenskt provsystem på 1940-talet – att utesluta ett kriterierelaterat alternativ (även om han inte använde den termen). Wigforss menade att så precisa formuleringar kunde inte formuleras att de skulle bli entydigt tolkningsbara för olika användare.[22] Hans förslag blev därför ett provsystem baserat på relativa betyg, vilket också blev vad som infördes.

Även mer sentida forskare har haft dubier i frågan.

> It is linguistically naïve to believe that criteria … can ever be made sufficiently precise for their use not to involve subjective judgements of the type which they are intended to avoid … It is technically naïve to expect the use of complex aggregation rules to enable detailed descriptions of candidates' attainments to be inferred from summary measures like grades or that such rules, because they are explicit, necessarily operate in a way which is consistent with natural notions of fairness … It is philosophically naïve to assume that fair judgements can only be made if every candidate's script is judged by precisely the same set of criteria … It is psychologically naïve to assume … that performance is not profoundly affected by the context of the task being carried out. Cresswell (2000)[23]

Trots de problem som många påtalat, eller möjligen på grund av dem, har en uppsjö av metoder för *standard setting*, dvs. kravgränssättning, lanserats och publicerats i olika handböcker.[24] Alla intar inte en lika uppgiven hållning som Cresswell. Ziecki & Pieri skriver som följande:

> All procedures for setting cut scores require the application of judgment. For example, some types of cut score studies require judges

[21] *Cut off*-scores eller *benchmarks* på engelska. Själva processen brukar kallas *standard setting*.
[22] Då kan man betänka att Wigforss utredde prov för folkskolan.
[23] Cresswell (2000).
[24] Se t.ex. Zieky & Perie (2006).

> to estimate the probability that a hypothetical group of students would know the answer to a test question. Another type of study requires judges to examine a student's performance and to decide whether the performance is good enough for some particular purpose. No purely objective methods exist. There are no "true" cut scores that a group of perfectly selected, perfectly trained judges using a perfect method will find. The cut scores, rather, reflect the combined judgments of the people involved. (s. 7)

De drar vidare följande slutsats som torde ha bred uppslutning inom professionen på området.

> It is impossible to prove that a cut score is correct. Therefore, it is crucial to follow a process that is appropriate and defensible. Ultimately, cut scores are based on the opinions of a group of people. The best we can do is choose the people wisely, train them well in an appropriate method, give them relevant data, evaluate the results, and be willing to start over if the expected benefits of using the cut scores are outweighed by the negative consequences. (s. 23)

Det är uppenbart att bedömningen (judgement) har en mycket mer central roll i det kriterierelaterade systemet än i det normrelaterade. Detta innebär reliabilitetssvårigheter av ett annat slag än vad som gäller för normrelaterade prov vilka ofta konstrueras med hög reliabilitet som utgångspunkt. I ett kriterierelaterat system är oftast validitetsfrågor i centrum vilket innebär uppgifter med större tolkningsutrymme. Det gör att bedömarreliabiliteten på uppgiftsnivå blir lägre. I validitetshänseende kan man säga att medan innehållsvaliditet och prognostisk validitet var särskilt viktiga för normrelaterade prov, så är läroplansvaliditet[25] utmärkande för de kriterierelaterade proven.

[25] Detta är en fri översättning av begreppet *construct validity* som inte har någon väletablerad svensk översättning. Begreppet innefattar de kunskaper i vid mening (jämför SOU 1992:94) som uttrycks i kursplaner och läroplan, dvs. det som konstruerar eller konstituerar målet för undervisningen i ämnet enligt styrdokumenten. Hårdrar man detta innebär det t.ex. att intresse för ämnet kan ingå i *constuct validity* om undervisningen har som mål att skapa intresse för ämnet hos eleven.

Betygssättning

Införandet av 1994 års kursplaner innebar att de tidigare mer styrande läroplanerna Lgr 80 och Lgy 70 upphörde att gälla. Detta innebar tillsammans med övriga förändringar ett paradigmskifte. Den förhållandevis starka inramning och avgränsning av undervisningens innehåll som de tidigare läroplanerna angav fanns inte längre utan det blev de professionellas (lärarnas) uppgift att tillsammans med eleverna bestämma ämnets gränser och innehåll. En uppgift som de knappast var förberedda för. Detta bidrog till att försvåra lärarnas bedömning av elevernas kunskaper, i synnerhet om bedömningen skulle vara likvärdig med den som gjordes av andra lärare vid andra skolor. De nationella proven gav inte heller mycket vägledning inledningsvis eftersom även dessa var anpassade till de nya signalerna om lokalt val av stoff och val av tidpunkt då detta stoff skulle ingå i undervisningen.

Situationen medförde att inte bara bedömningen av provresultaten utan även de beslut (decision) om betygen som skulle fattas blev skakiga. Lärarna skulle tolka kriterier för olika betyg – varav kriterier för det högsta betyget mycket väl godkänt (MVG) saknades de första fem åren – på egen hand eller tillsammans med sina kollegor, och utifrån detta fatta beslut om vilket betyg respektive elev skulle få.

Den nya kunskapssynen, baserad på de fyra f:en, där de olika kunskapsformerna inte var rangordnade utan snarare växelverkade med och förutsatte varandra avvek påtagligt från tidigare sätt att betrakta kunskap och kognitiv utveckling. Dessa sätt var som nämnts mer hierarkiskt ordnade i form av taxonomier[26], vilka i sin tur kunde bygga på idéer om de kognitiva förmågornas succesiva och åldersstyrda utveckling[27]. Den mer strikt deterministiska syn som dominerade 1960-talets läroplaner mjukades efterhand upp för mer konstruktivistiska idéer om att kunskapsutveckling och kognitiv utveckling inte enbart berodde på biologisk mognad utan även på växelverkan med omgivningen. Kunskap uppstod inte enbart genom att läraren lärde ut utan även genom att eleven själv konstruerade sin kunskap i samverkan med sin omgivning. Denna

[26] Mest känd är Blooms taxonomi.
[27] Till exempel utifrån Piagets stadieindelning. Se t.ex. Piaget (2013), som skrevs 1968.

mer konstruktivistiska kunskapssyn kunde spåras i Lgr 80. Dock kvarstod idén om kunskapens hierarkiska ordning.

Betygssättning är en rangordnande verksamhet. Den leder till en värdeskala[28] som används för urval och får därmed rättsverkan i sammanhang som kan vara av stor betydelse för eleven. Det är även en myndighetsutövning av läraren och då bör den givetvis bygga på ett underlag som så långt möjligt underlättar rangordningen, dvs. den betygssättning som sker på beslutsnivån.

Den nya kunskapssynen kan vara tilltalande ur kunskapsteoretiska perspektiv, men den medförde problem för lärarnas bedömning och betygssättning. Betygskriterierna var även de svårtolkade och gav inte uttryck för något systematiskt sätt att rangordna elevers kunskaper.[29] Detta ledde till att kunskapsformerna i de fyra f:en i praktiken kom att rangordnas på olika sätt, ofta i form av en trappa där f:et ”fakta” ansågs representera betyget godkänt (G), varefter övriga trappsteg tilldelades övriga kunskapsformer.

De nationella proven brottades delvis med samma problem och kunde under de första åren inte ge lärarna särskilt tydlig vägledning.

Det är knappast förvånande att de betygssättande lärarna låg lågt de första åren i osäkerhet om vilka kunskaper de olika betygen representerade. Någon föreskrift om förhållandet mellan provresultat och betyg, liknande de som gällde för det relativa systemet, fanns inte heller. Efterhand ökade förstås lärarnas förtrogenhet med systemet och när kriterier år 2000 publicerades även för det högsta betyget MVG stabiliserades betygen i betydande grad på nationell nivå. På lokal nivå (skolor och klasser) har dock en avsevärd variation[30] i relationen mellan provbetyg och ämnes- eller kursbetyg funnits under hela den tid de kriterierelaterade nationella proven använts.

[28] Värdeskala i kvantitativ mening genom att den utgörs av en enligt vissa regler fastlagd sammanräkning av betygspoäng för ingående betyg, t.ex. meritvärden för grundskolan och jämförelsetal för gymnasieskolan. Respektive betygs betygspoäng är dessutom uttryck för en politiskt gjord värdering av betygets värde.

[29] Se t.ex. SOU 2007:28.

[30] Eftersom ingen annan föreskrift fanns än att de nationella proven skulle vara en del av lärarens betygsunderlag, var det svårt att värdera de skillnader som fanns mellan olika skolor eller klasser.

Kommentar

Sammanfattningsvis kan man säga att konstruktionen av prov inte skiljer sig särskilt mycket i ett normrelaterat och kriterierelaterat system. Möjligen gäller att det kriterierelaterade systemet kräver något fler komplexa uppgifter för att stämma med kriterier på mer avancerad nivå. Klassisk testteori kan användas, men även den moderna testteorin (IRT)[31] fungerar bra eftersom den ger möjligheter att välja uppgifter som differentierar väl vid betygsgränserna och eftersom den ger underlag för att bedöma vilka uppgifter som ligger på en specifik betygsnivå. Om uppgifterna publiceras ger de därmed god information om vilka kunskaper som krävs för olika betyg.[32]

Både bedömningen av och besluten om såväl provbetyg som betyg på enskilda elevers kunskaper bygger i ett kriterierelaterat system i hög grad på tolkning och bedömning. Den underliggande mätningen har inte samma starka roll som vid normrelaterade betyg där provstödet i form av poänggränser har stor betydelse. Däremot är kopplingen till kursplanernas skrivningar i allmänhet tydligare i ett kriterierelaterat prov och man kan säga att kravet på kursplanevaliditet är större för kriterierelaterade prov än för normrelaterade. Men till priset av en lägre reliabilitet.

Trots de uppenbara svårigheter som ansatsen ger när det gäller mätning och bedömning har de kriterierelaterade proven haft och har stor uppslutning. I Sverige innebar systemskiftet svårigheter och osäkerheten om tillämpningen har varit stor.

[31] Item Response Theory.

[32] Detta är egentligen mer en effekt av IRT än av att provet är kriterierelaterat. Till exempel används uppgifter för att illustrera olika nivåer i PISA trots att det är en form av grupprelaterat prov. Det senare beror på att i ett grupprelaterat prov rangordnas provdeltagarna först och därefter rangordnas uppgifterna baserat på den skala som gäller för provdeltagarna. I ett kriterierelaterat prov rangordnas uppgifterna först (utifrån de kriterier som är formulerade för olika nivåer) och därefter klassificeras eleverna efter hur väl deras resultat sammanfaller med uppgifternas placering. (Detta gäller i teorin. I praktiken är bilden mer komplicerad med ett otal metoder för *standard setting*.)

Period 3 – standardsbaserade prov och betyg

Någon klar avgränsning mellan period 2 och 3 är svår att definiera generellt. Det som är gemensamt är främst betoningen av innehållet i läro- och kursplaner vid bedömning och betygssättning. Synsättet började som beskrivits i förra avsnittet få genomslag i slutet av 1960-talet och har därefter vuxit allt mer i styrka. Innehållet i ämnen och kurser har förstås alltid varit av intresse, men det som har ökat är betoningen på att kunna bestämma vilken kunskapsnivå som ska uppnås för olika betyg och hur provdeltagaren kan eller ska visa att hon eller han har uppnått en viss nivå.

I engelskspråkig litteratur talar man ofta om utvecklingen som *the standards movement*. Den innebär i korthet att när det gäller innehållet definieras det i en uppsättning *content standards*, närmast motsvarande det som kallas centralt innehåll i nuvarande svenska kurs- och ämnesplaner.

När det gäller kunskapsnivåerna uttrycks dessa i *performance standards*, dvs. en uppsättning satser som anger vilka kunskaper som ska visas på olika betygsnivåer. I Sverige motsvaras *performance standards* av kunskapskrav för olika betyg. Detta synsätt är i dag klart dominerande i flertalet med Sverige jämförbara länder och präglar de flesta länders läro- och kursplaner.

De kunskapsnivåer som bestäms uttrycker någon vald form av kunskapsprogression i ämnet. Denna progression är mer eller mindre godtycklig och är beroende av ämnets karaktär. Vissa ämnen har i sig en tydlig hierarkisk ordning (kan t.ex. gälla vissa moment i matematik), medan andra ämnen har en progression som mer innebär en vidgning av kunskapsfältet (kan t.ex. gälla främmande språk). För samtliga ämnen gäller dock att progression är en kombination av fördjupning och breddning, men i olika proportioner. När de olika nivåerna bestämts är det nödvändigt att de uttrycks i ämnestermer och på ett så tydligt sätt som möjligt för att ge provkonstruktörer och betygsättare förutsättningar för en rättvis och likvärdig bedömning.

Begreppslig oklarhet

Skillnaderna mellan kriterierelaterade prov och de nutida prov som baseras på *standards* är oklara och flytande, inte bara för svensk del utan över huvud taget. Definitioner av kriterierelaterade prov och för de prov som vuxit fram under de senaste decennierna – standardsrelaterade, standardsbaserade, *outcome-relaterade* etc. – tenderar att flyta samman. Skillnaderna mellan dem är oklara och benämningarna ofta mer beroende av vilka länder det gäller än av specifika skillnader i vad som avses. I Australien och Nya Zeeland är exempelvis benämningen *outcome-based* vanlig. I USA talar man oftare om *standards-referenced* och *standards-based*. Ett försök att åtskilja olika provtyper inom ramen för standardsbaserade prov görs nedan.[33]

Kriterierelaterade prov är utformade för att mäta elevens provresultat mot en förutbestämd uppsättning kriterier (*standards*). Om eleven visar kunskaper över en fastlagd gräns, exempelvis genom att uppnå en poängsumma högre än gränsen, bedöms eleven ha visat kunskaper på den nivå som ligger över gränsen. Var den fastlagda gränsen (kravgränsen) ska ligga är en bedömning som görs av en eller flera bedömare. Ansatsen är kompensatorisk eftersom någon åtskillnad inte görs mellan olika delar av kunskapskravet. Om tillräckligt många poäng uppnåtts anses kravet uppfyllt. Det innebär att någon tydlig bild av vad eleven kan och kan göra inte framgår av provbetyget. Om en annan person eller grupp med annan tolkning av kunskapskraven hade fastställt kravgränserna kunde de ha fått en annan placering. Fastställandet av kravgränser handlar om bedömning.

Standardsrelaterade prov innebär att undervisning och prövning härleds från de mål som anges för kursen (*learning standards*).[34] I standardsrelaterade system vägleds undervisning och prov av standards som formuleras som syften, mål eller kunskapskrav. Innehållet i provet är helt standardsrelaterat och styrt av de kursmål som anges och de kunskapskrav som ställs. En viss andel av kunskapskraven för ett visst betyg ska vara uppfyllda för att ge det betyg kunskapskravet gäller. Dock behöver inte läraren veta vilka

[33] Bygger på Abbott (2015).
[34] Se t.ex. Abbot (2015). Kan jämföras med våra tidigare mål att sträva mot.

specifika delar av kunskapskravet eleven inte uppfyllt utan det räcker med att ”medelvärdet” uppfyller kraven. Betygssättningen är oftast poängbaserad och liknar i hög grad det traditionella sättet att betygssätta prov och elever. Såväl provbetygen som elevgruppens betyg sätts som medelvärden av betygen på de olika delproven eller de ingående proven (vilket förutsätter att betygen tilldelas kvantitativa värden). Likheten med kriterierelaterade prov är påtaglig. Det som skiljer är kanske den mer uttalade kopplingen till syften, mål och kunskapskrav. Denna typ av prov har främst ett summativt syfte.

Standardsbaserade prov (standards-based assessment[35]). Begreppet standardsbaserad refererar till en undervisning som strävar efter att säkerställa att eleven verkligen uppnår förväntade kunskaper, dvs. uppfyller de kunskapskrav som gäller.[36] Det finns därmed en stark formativ innebörd i begreppet. I ett standardsbaserat system blir elever godkända på ett prov eller i en kurs endast om de visar att de till fullo uppfyller de specifika kunskaper och färdigheter som gäller för kunskapskravet. Eleven kan därför behöva göra ett prov flera gånger eller behöva särskilt stöd för att visa att hon eller han har uppnått de kunskaper som krävs.[37] Det innebär också att proven inte bör omfatta alltför stora delar av ämnet eller kursen utan att det sker en fortlöpande växelverkan mellan undervisning och prövning.

I det standardsbaserade systemet kan provbetygen se annorlunda ut än i de andra systemen där ett sammanfattande provbetyg i allmänhet rapporteras. Det är i det standardsbaserade systemet vanligare att prov ges oftare och att de innefattar specifika delar av kunskapskraven. För svensk del skulle det kunna innebära att läraren använder olika prov för olika förmågor. I undervisningssammanhang (där den här modellen har störst tillämpning) innebär det också att betyg eller andra former av systematiserad bedömning ges för varje enskild förmåga eller varje enskilt innehåll. Rapporteringen av betyg sker således på förmågenivå (eller delprovsnivå) och ofta är rekommendationen att inte konstruera ett sammanfattande betyg eller provbetyg.

[35] Assessment är ett vidare begrepp än prov. Det innefattar prov, men även andra former av bedömning.
[36] Abbot (2015).
[37] Den här ansatsen har således mer relevans för lärarens egna prov eller för ett flexibelt nationellt bedömnings- och betygsstöd som läraren kan använda vid behov.

Prov utformade och använda i enlighet med den standardsbaserade modellen har ett utpräglat formativt syfte. Det är också ganska vanligt att underlag för succesiv betygssättning inte bara samlas in för traditionella kunskapsområden och färdigheter utan även för mer icke-kognitiva förmågor, metakognition, attityder och annat som är framskrivet i styrdokumenten. Sammanställningar av elevernas progression i olika avseenden rapporteras sedan som information till föräldrar och vårdnadshavare.

Kommentar

Sammanfattningsvis kan man konstatera att det i vissa avseenden, t.ex. när det gäller typ av uppgifter och deras koppling till kursplaner och andra styrdokument, är små skillnader mellan kriterierelaterade och standardsrelaterade prov och betyg. Däremot är det större skillnader i betoningen av summativ respektive formativ roll. De kriterierelaterade proven är i allmänhet uttalat summativa med syfte att stödja beslut om sammanfattande betyg.

De standardsbaserade proven (i sin renodlade form) är däremot mer uttalat formativa och begränsade till innehållet (men å andra sidan fler till antalet). De syftar huvudsakligen till att fatta beslut om undervisningsinsatser för att enskilda elever ska ges möjlighet att uppfylla alla standards. Därmed är det också mer tveksamt hur väl de lämpar sig som underlag för beslut om sammanfattande betyg. De förväntas ge bra underlag för successiv bedömning men samtidigt är det inte ovanligt att de länder som har standardsbaserade system också har särskilda summativa prov för betygssättning eller urval till efterföljande utbildningar.[38]

Lpo 94 och Lpf 94

Det svenska systemet har sedan 1994 gått under benämningen mål- och kunskapsrelaterat. Vid en jämförelse med de standardsbaserade provformerna kan man konstatera att 1994 års kursplaner inte hade några uttalade *content standards* eftersom urvalet av innehåll (stoff) i undervisningen var en uppgift för läraren och eleverna. Även om

[38] Gäller exempelvis delar av Australien och USA.

det främst blev läromedlen eller tidigare erfarenheter som styrde så fanns det ingen gemensam styrning genom en läroplan. Detta var en viktig skillnad jämfört med det tidigare norm- eller gruppreleterade systemet. Att inte ange något innehåll är inte heller i linje med någon av de beskrivna modellerna. När det gäller *performance standards* saknades också tydliga angivelser. Betygskriterierna uppfattades ofta som oklara och svårtolkade och som tidigare nämnts saknades kriterier för det högsta betyget fram till år 2000.

Någon tydlig modell för kunskapsprogression fanns inte heller att luta sig mot. De fyra f:en skulle inte tolkas i någon hierarkisk ordning och det blev därför provkonstruktörernas och lärarnas uppgift att fastställa en progression som kunde harmonieras med betygskriterierna. Eftersom gemensamma nationella riktlinjer saknades innebar det att olika provkonstruktörer kunde ha olika sätt att se på frågan. Det beror förstås delvis på att olika ämnen har olika karaktär, men också på att en gemensam syn på kunskapsprogression generellt saknades. Med tiden utvecklades en praxis där de nationella proven kan antas ha spelat en inte obetydlig roll. Samtidigt tenderar alla system att med tiden etablera mer eller mindre gemensamma praktiker om de har möjlighet att ta del av gemensam information och kommunicera på olika nivåer.

Om man ska sammanfatta den första fasen med det mål- och kunskapsrelaterade systemet (1994–2000) kan man säga att varken prov- eller betygssystem kan ses som ett tydligt exempel på något av de sätt att konstruera och bedöma prov samt fatta beslut om betyg som beskrivits. Det fanns en form av *learning standards* (mål att sträva mot) och ansatser till *performance standards* (betygskriterier), men det saknades *content standards*.

Vissa inslag i linje med den standardsbaserade modellen fanns några år (2002–2004) då några sammanfattande provbetyg inte gavs i engelska och svenska, utan endast betyg på de delprov som gällde olika förmågor (tala, höra, skriva och läsa). Sedan var det lärarens uppgift att sätta såväl provbetyg (i den mån det efterfrågades) som betyg på elevens samlade kunskaper. Modellen övergavs dock eftersom den väckte kritik för att betygssättningen inte skulle bli rättvis och likvärdig om det inte fanns gemensamma regler för hur de sammanfattande provbetygen skulle bestämmas.

Lgr 11 och Lgy 11

De kurs- och ämnesplaner som började gälla 2011 behöll de grundläggande delar från 1994 som gällde benämningen på bedömningssystemet, dvs. mål- och kunskapsrelaterat. Någon ändring av kunskapssynen gjordes inte heller utan de fyra f:en kvarstod. Den utredning *Tydliga mål och kunskapskrav*[39] som låg till grund för de nya läroplanerna behandlade kunskapsbegreppet översiktligt (jämfört med ett helt kapitel i *Skola för bildning*) och någon diskussion av progressionsbegreppet i kunskapstermer gjordes därför inte. Däremot uppmärksammades avsaknaden av innehållsangivelser och därför har 2011 års kurs- och ämnesplaner ett angivet centralt innehåll för olika ämnen och kurser. Dessa är allmänt formulerade om man jämför med gamla tiders stoffkataloger, men i linje med nutida sätt att formulera *content standards*.

Tydliga mål och kunskapskrav tog också fasta på den kritik som funnits mot att betygskriterierna var abstrakta och svårtolkade. Detta ledde till försök att finna en terminologi som kunde gälla för alla ämnen och kurser och som därmed skulle underlätta förståelsen av betygskriterierna, eller kunskapskraven som de kom att heta i de nya kurs- och ämnesplanerna. Eftersom någon progression baserad på kunskapssyn inte fanns konstruerades en progression baserad på så kallade värdeord. Kunskapskravens progression formulerades alltså som en form av grammatisk progression av typen ”goda kunskaper”, ”utvecklade kunskaper” och ”väl utvecklade kunskaper”. I kunskapskraven anges också ibland hur detta kan visas. Som exempel återges nedan progressionen för ett par av de förmågor som ingår i kunskapskraven för matematik i årskurs 9 (värdeorden i fetstil).[40]

Kunskapskrav för betyget E i slutet av årskurs 9

Eleven kan lösa olika problem i bekanta situationer på ett i **huvudsak** fungerande sätt genom att välja och använda strategier och metoder med viss anpassning till problemets karaktär samt **bidra till att formulera** enkla matematiska modeller som kan tillämpas i sammanhanget. Eleven för **enkla och till viss del** underbyggda resonemang om val av tillvägagångssätt och om resultatens rimlighet i förhållande till

[39] SOU 2007:28.
[40] Ur Skolverket (2011).

problemsituationen samt **kan bidra** till att ge **något** förslag på alternativt tillvägagångssätt.

Kunskapskrav för betyget C i slutet av årskurs 9

Eleven kan lösa olika problem i bekanta situationer på ett **relativt väl** fungerande sätt genom att välja och använda strategier och metoder med **förhållandevis god** anpassning till problemets karaktär samt **formulera** enkla matematiska modeller som **efter någon bearbetning** kan tillämpas i sammanhanget. Eleven för **utvecklade och relativt väl** underbyggda resonemang om tillvägagångssätt och om resultatens rimlighet i förhållande till problemsituationen samt kan ge **något** förslag på alternativt tillvägagångssätt.

Kunskapskrav för betyget A i slutet av årskurs 9

Eleven kan lösa olika problem i bekanta situationer på ett **väl** fungerande sätt genom att välja och använda strategier och metoder med **god** anpassning till problemets karaktär samt **formulera** enkla matematiska modeller som kan tillämpas i sammanhanget. Eleven för **välutvecklade** och **väl** underbyggda resonemang om tillvägagångssätt och om resultatens rimlighet i förhållande till problemsituationen samt kan ge **förslag** på alternativa tillvägagångssätt.

Om man jämför detta sätt att formulera *performance standards* med vad som är vanligt i andra länder så framstår de olika nivåerna som abstrakta och svårtolkade. Samtidigt är det inte meningen att kunskapskraven ska tolkas fristående utan tillsammans med beskrivningarna av ämnets syfte och det centrala innehållet.

För det nuvarande systemet gäller även att för att eleven ska få ett visst betyg (E, C eller A) måste respektive kunskapskrav vara uppfyllt i sin helhet. Det är en icke-kompenserande regel. Om en förmåga ligger på E-nivå har det ingen betydelse om övriga förmågor ligger på högre nivå. Det sammanfattande betyget blir E. Denna regel gäller när läraren sätter betyg.

För de nationella proven gäller inte den icke-kompensatoriska regeln lika strikt. Genom att ett prov begränsas av praktiska hänsyn (tid) kan inte alla förmågor prövas i tillräcklig omfattning för att resultaten på enskild förmågenivå ska bli rimligt reliabla. Det betyder att mätosäkerheten kan anses bli för stor för att medge icke-kompensatoriskt beslutsfattande vid bestämningen av provbetygen.

Den här skillnaden, att läraren ska sätta icke-kompensatoriska betyg på elevens kunskaper medan provbetygen i varierande grad är

kompensatoriska, skapar förvirring hos många lärare och leder till diverse resonemang om i vilka sammanhang betyg kan ges eller inte ges och vad betyg representerar. Några av de lärosäten som Statens skolverk anlitar för provkonstruktion argumenterar därför för att ta bort övergripande provbetyg och/eller delprovsbetyg. Andra gör försök att konstruera prov som fördelar uppgifter så att så många förmågor som möjligt prövas på så många nivåer som möjligt. Därigenom anses uppgifterna ge belägg för nivån på uppnådda kunskaper för olika förmågor. Detta leder till att matriser används för att kunna överblicka fördelningen av belägg på förmågor och nivåer. De endimensionella ”trappor” som var vanliga 1994–2010 har således i vissa fall ersatts av tvådimensionella matriser efter införandet av 2011 års läroplaner.

Problemen med bristen på samstämmighet mellan sätten att bestämma provbetyg och ämnes- eller kursbetyg leder också till att vissa röster höjs för att även om läraren sätter ett betyg på provet så bör eleven inte få veta detta betyg eftersom det är en annan sorts betyg än det slutliga.

I vilket fall kan man konstatera att det nuvarande systemet visar likheter med både det kriterierelaterade och det standardsbaserade systemet. Det kriterierelaterade systemet förväntas resultera i ett sammanfattande provbetyg. Det standardsbaserade systemet förutsätter att *learning standards* och *performance standards* är formulerade i förmågor och progression i dessa. Progressionen så som den uttrycks i kunskapskraven är visserligen abstrakt, men det finns stödjande material som med tiden borde kunna öka likvärdigheten i tolkningen av kunskapskravens innebörd. Man kan alltså se att progressionen i kunskapskraven följer ett antal spår (förmågor).

Ett annat förhållande som liknar det standardsbaserade systemet är att samtliga förmågor ska uppnås till en viss nivå för att berättiga till ett visst slutbetyg. Detta är en ansats som anses gälla för standardsbaserade prov. I Sverige föreskrivs den emellertid enbart för lärarens sammanfattande betygssättning och inte för nationella prov. Detta leder till viss förvirring genom att proven i Sverige i huvudsak är summativa medan de standardsbaserade proven i internationella sammanhang i första hand är formativa. Man kan alltså se de aktuella svårigheterna som ett uttryck för problemen med att förena kriterierelaterade ansatser med standardsbaserade i ett gemensamt system.

Sammanfattning och kommentar

Newtons artikel[41] och modell syftar till att ge referensramar för att bestämma sambandet mellan provs syften och deras utformning. Han formulerar en referensram som bygger på begreppen bedömning, beslut och effekt (*impact*). Till detta kan man lägga en underliggande nivå som gäller insamlingen av det som ska bedömas, i det här sammanhanget kallat mätning.

I takt med att synen på kunskap, lärande och undervisning har förändrats har även sättet och grunderna för att konstruera prov förändrats. I grova drag kan man säga att utvecklingen i Sverige under de senaste 75 åren har gått från ett påtagligt mätbaserat provsystem mot ett provsystem som i allt högre grad baseras på tolkning av verbalt formulerade mål och kriterier för bedömning, vanligen med en gemensam term kallade standards. En strömning har således varit att proven gått från att vara strikta mätinstrument baserade på tydliga föreskrifter om det innehåll som ska undervisas och de kunskaper som ska läras, till att bli underlag för att bedöma i vilken grad elever utvecklat de förmågor som skrivs fram i de standards som anger målen för och innehållet i undervisningen samt de standards som anger vad förmågor på olika nivåer innebär. För svensk del uttryckt som ämnets syfte, centralt innehåll och kunskapskrav. Denna förskjutning i synen på kunskap, lärande och bedömning har lett till att prov som utgår från standards i dag är dominerande i flertalet länder.

En annan strömning är att tyngdpunkten har förskjutits från summativa prov mot formativa prov. Under ett par decennier har främst engelska (Dylan Wiliam, Paul Black m.fl.), men också australiensiska och nya zeeländska forskare (John Hattie, Helen Timperley m.fl.) argumenterat starkt för det formativa förhållningssättet. Även i Sverige har rörelsen haft flera förespråkare (Anders Jönsson, Christian Lundahl m.fl.). De här strömningarna har förstås även haft inverkan på den svenska synen på prov. Många forskare och lärarutbildare har haft utbyte med utländska förespråkare för olika typer av formativa och standardsbaserade ansatser och i olika grad anammat dessa synsätt. Även lärare har förstås genom fortbildning och kompetensutveckling påverkats.

[41] Newton (2007).

När det gäller konstruktionen av de nationella proven har det inte funnits någon gemensam syn på hur de summativa och formativa ansatserna ska hanteras. Detta har fått till följd att sättet att se på prov och deras funktion i viss mån varierar. De uttryckta syftena för dagens prov är att de ska ”stödja en likvärdig och rättvis betygssättning” och att de ska ”ge underlag för en analys av i vilken utsträckning kunskapskraven uppfylls på skolnivå, på huvudmannanivå och på nationell nivå”. Dessutom anges att proven kan ha en roll i vissa andra avseenden, t.ex. att de ”kan användas formativt” även om de främst är summativa.[42]

Här har alltså den formativa ansatsen som är utmärkande för det standardsbaserade sättet att se på prov fått viss inverkan. Dock med en tolkning av begreppet som inte är i linje med den som vanligen anges för formativa prov (styrs av läraren, ges vid behov, omfattar mindre delar av kursinnehållet, prövar specifika förmågor, etc.). Man kan däremot säga att vissa av proven fångar upp en del idéer som gäller för de standardsbaserade proven, t.ex. när det gäller att de enskilda förmågorna ska prövas. Detta har lett till att uppgifterna i proven klassificeras efter vilken förmåga som prövas och på vilken nivå (uttryckt som E-poäng, C-poäng eller A-poäng, i vissa ämnen används beteckningen *belägg* som ännu tydligare markerar kopplingen till den enskilda förmågan och kunskapskravet). En följd av denna ansats blir också att resultathanteringen blir omfattande eftersom provet är summativt och omfattar en hel kurs. Det formativa synsättet leder då till att varje förmåga registreras för sig och för varje förmåga ska de olika beläggen för betygsnivån anges. För att kunna hantera detta krävs mer eller mindre omfattande matriser. Eftersom provet är summativt måste sedan matrisen på något sätt sammanfattas till ett provbetyg. Detta sker genom olika anvisningar för olika ämnen.

I vilket fall blir bedömningsproceduren omfattande genom att den hämtat inspiration från den formativa ansatsen (som innebär korta men återkommande prov på enskilda eller ett fåtal förmågor i taget), men används summativt (vilket innebär täckning av så mycket som möjligt av ämnet eller kursen på så många nivåer som möjligt).

[42] http://www.skolverket.se/bedomning/nationella-prov

Sverige har alltså vissa problem eftersom de nationella proven ska uppfylla flera syften och därtill ofta ges ytterligare funktioner, t.ex. den formativa. De andra nordiska länderna har snarare valt olika prov för olika syften. Norge och Danmarks digitala nationella prov i grundskolan har en mer uttalat formativ (och delvis utvärderande) funktion och är inte betygsstödjande. För den summativa uppgiften finns särskilda examensprov. Examensproven betygssätts av censorer och lärarna sätter betyg utifrån sin samlade bedömning. I examensbeviset ingår sedan både examensprovbetyget och lärarbetyget.

Newton betonar, liksom flertalet forskare på området, att prov måste konstrueras med specifika syften. Det är också provkonstruktörens uppgift att informera om vad prov *inte* lämpar sig för. Sedan är det en annan sak att så fort provresultat publiceras så har inte konstruktören och den ansvariga utgivaren (för svensk del Skolverket) längre kontroll över hur resultaten används av politiker, media, olika intressegrupper, föräldrar m.fl. Det viktiga i det här sammanhanget är att den som ska konstruera provet får ett tydligt och avgränsat syfte med provet angivet av den som har ansvaret för provverksamheten. Är det ett betygsstödjande summativt prov? Är det ett formativt prov? Är det ett utvärderande prov? Är det ett uppföljande prov? Är det ett examinerande prov? Är det ett validerande prov? Ska provet mäta förändring över tid?

Referenser

Abbott, S. (red.). (2015). The glossary of education reform. http://edglossary.org/

Bernstein, B. (1971). On the classification and framing of educational knowledge. In MFD Young (red.). Knowledge and Control: New directions for the sociology of education. London: Collier MacMillan, 47–69.

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

Betänkandet Ett nytt betygssystem (SOU 1992:86). Slutbetänkande av Betygsberedningen. Stockholm: Utbildningsdepartementet.

Betänkandet Skola för bildning (SOU 1992:94). Betänkande av läroplanskommittén. Stockholm: Utbildningsdepartementet.

Betänkandet Tydliga mål och kunskapskrav i grundskolan (SOU 2007:28). Betänkande av Utredningen om mål och uppföljning i grundskolan. Stockholm: Utbildningsdepartementet.

Cresswell, M. (red.). (2000). Research studies in public examining. Guildford: Associated Examining Board.

http://webarchive.nationalarchives.gov.uk/20141031163546/

http://www.ofqual.gov.uk/files/2007-comparability-exam-standards-d-chapter2.pdf

Haertel, E. H. & Lorie, W. A. (2004). Validating standards-based test score interpretations. Measurement: Interdisciplinary Research and Perspectives, 2, 61–104.

Gulliksen, H. (1950). Theory of mental tests. New York: John Wiley.

Lord, F. M. and Novick, M. R. (1968). Statistical theories of mental test scores. Reading Mass: Addison-Wesley.

Newton, P. E. (2007). Clarifying the purposes of educational assessment. Assessment in Education. Vol. 14, No. 2, July 2007, 149–170.

Piaget, J. (2013). Barnets själsliga utveckling. Lund: Studentlitteratur.

Skolverket (2011). Läroplan för grundskolan, förskoleklassen och fritidshemmet 2011. Stockholm: Skolverket.

Zieky, M. & Perie, M. (2006). A Primer on Setting Cut Scores on Tests of Educational Achievement. Princeton: ETS. https://www.ets.org/Media/Research/pdf/Cut_Scores_Primer.pdf

Unesco (2013). IBE Glossary of Curriculum Terminology. Genèeve: Unesco. http://www.ibe.unesco.org/fileadmin/user_upload/Publications/IBE_GlossaryCurriculumTerminology2013_eng.pdf

Provs mätfel

I den här bilagan redogör utredningen kortfattat för några olika typer av mätfel som prov är förenade med. Dessutom beskriver vi olika testteorier för att uppskatta mätfelen.

Provs mätfel

Nationella prov och andra prov är alltid behäftade med olika typer av fel. Det finns slumpmässiga mätfel på individnivå (*slumpfel*), *bedömningsfel* på grupp- eller lärarnivå och *systematiska fel* på nationell nivå.[1]

Slumpfel

Slumpfel handlar i stor utsträckning om vilka frågor som ställs på ett prov och hur dessa relaterar till den enskilda elevens kunskaper. Eftersom inget prov kan täcka in ett helt kunskapsområde på rimlig tid blir varje prov ett stickprov av alla frågor som kan ställas. Med lite tur i urvalet av uppgifter kan elevens resultat leda till en överskattning av förmågan, med lite otur till en underskattning. I prov som låter eleverna välja mellan olika givna svarsalternativ finns ofta utrymme för gissning, där slumpen har stort inflytande på utfallet. Ett annat slumpfel kan vara att en elev råkar ha en dålig dag.

[1] Avsnittet om provs mätfel bygger bl.a. på Skolverket (2015).

Bedömningsfel

Bedömningsfel handlar om att bedömningen av elevens svar görs på olika sätt av olika bedömare. Denna felkälla är ofta ett mindre problem för korta uppgifter. Ju mer komplexa och omfattande svaren är desto större tenderar dock variationen mellan olika bedömare att bli. Den här typen av fel uttrycks vanligen i termer av bedömaröverensstämmelse eller interbedömarreliabilitet[2]. Ofta finner man att olika bedömare är olika stränga i sina bedömningar, liksom att det finns skillnader mellan olika bedömares rangordning av svaren från en grupp elever. Eftersom det i Sverige oftast är elevens egen lärare som bedömer proven kan man säga att denna typ av fel gäller på gruppnivå, vilket innebär att lärarens grad av stränghet påverkar alla elever i klassen på samma sätt, förutsatt att läraren bedömer alla prov efter samma normer. Det kan dock inte förutsättas att läraren fullt ut bedömer alla elevers prov lika eftersom det är väl belagt att en bedömare som känner den elev som har gjort provet tenderar att påverkas av denna kännedom, s.k. *haloeffekter*.

En svårighet vid bedömningar av mer komplicerade uppgifter är att bedömningsanvisningarna inte kan ange specificerade rätta svar utan endast mer allmänna kriterier och exempel på olika typer av svar. Bedömaren behöver alltså tolka såväl elevens utsaga som bedömningsanvisningarna.

Statens skolinspektions ombedömningar av nationella prov kan ge ett visst underlag för att avgöra hur stor andel av lärarna som gör olika bedömningar. Andelen avvikelser varierar stort med typen av prov. Prov baserade på *item*[3] har färre bedömningsavvikelser medan prov med omfattande uppgifter av typen uppsatser kan ha stor andel avvikelser. Skolinspektionens ombedömningar tyder också på att osäkerheten i lärares bedömningar, mätt som procentuell överensstämmelse, ökar när antalet betygssteg ökar.[4]

[2] Man talar också om *intra*bedömarreliabilitet som gäller hur lika en och samma bedömare bedömer vid olika tillfällen. Sådana bedömningar kan skilja sig åt en hel del. Se t.ex. Baird, J-A m.fl. (2013).
[3] Små, korta uppgifter som vanligen bedöms rätt eller fel dvs. ger 0 eller 1 poäng eller ”belägg”. En uppgift kan bestå av flera item.
[4] Se också t.ex. Qualifications and Curriculum Authority (2009).

Systematiska fel

Systematiska fel kan handla om att kravgränserna för olika betyg ligger på ”fel” nivå. Den princip som med nuvarande bestämmelser gäller för kravgränssättning är att betygsgränser och betygskrav ska vara fastställda när proven genomförs. Det är därmed inte möjligt att avvakta provresultaten innan betygskraven fastställs. Ett sådant systematiskt fel i provresultatet eller provbetyget har inte någon större betydelse om resultatet endast ska gälla för jämförelser mellan elever och elevgrupper det aktuella året, eftersom alla då drabbas av samma fel. Om resultatet däremot ska användas för jämförelser av provresultat från år till år, t.ex. när elever från olika årskurser söker till samma högre utbildning, eller för utvärderingsändamål, får denna typ av systematiska fel betydelse.

Olika testteorier för att uppskatta mätfel

Slumpfel kan uppskattas med hjälp av olika testteorier. Man talar huvudsakligen om två testteorier: den klassiska och den moderna.[5]

Den klassiska testteorin

Den klassiska testteorin bygger på grundantagandet att en provdeltagares provpoäng kan ses som summan av en ”sann” poäng och ett mätfel. Med sann poäng menas det poängmedelvärde en provtagare skulle få om hon eller han kunde upprepa samma prov många gånger. Detta är förstås inte möjligt och därför finns olika statistiska metoder för att skatta den sanna poängen. Utifrån grundantagandet och flera andra antaganden, som t.ex. gäller att mätfelet är slumpmässigt med medelvärdet noll och att mätfelens storlek är oberoende av testpoängen, kan formler för beräkning av reliabilitet och mätfel bestämmas.

Den klassiska testteorin kan användas i många olika provsammanhang, men den har också nackdelar. Till exempel kan man inte avgöra i vilken utsträckning uppgifternas eller provens egen-

[5] Avsnittet om olika testteorier bygger bl.a. på Skolverket (2003) och Crocker, L. & Algina, J. (1986).

skaper beror på de som gör provet eller på det aktuella provet. Uppgifternas egenskaper, t.ex. deras svårighetsgrad, beror både på provet och på gruppen. Detsamma gäller provets egenskaper i form av reliabilitet och mätfel. För att eliminera denna effekt försöker provkonstruktörerna skapa prov som är så likartade (parallella) som möjligt. Det innebär att man försöker ha lika många uppgifter i proven, se till att uppgifterna är av samma typ och svårighetsgrad samt säkerställa att uppgifterna har så likartat innehåll som möjligt. Därigenom antas proven bli mer jämförbara och eventuella skillnader i lösningsproportioner eller poängfördelningar kan då hänföras till de prövade grupperna.

Den moderna testteorin

Den moderna testteorin, *item response theory* (IRT), är baserad på statistiska modeller. Dessa modeller uttrycker sannolikheten för ett korrekt svar på en uppgift som en funktion av individens latenta förmåga och av olika egenskaper hos uppgiften, t.ex. dess svårighetsgrad. Med latent förmåga avses en inte direkt iakttagbar förmåga som konstrueras med hjälp av den statistiska modellen och de ingående uppgifterna. Även uppgifternas egenskaper konstrueras i samma process. Dessa kan sedan antas vara konstanta och uppgifterna kan användas för att bestämma den latenta förmågan hos nya provtagare. När man har bestämt uppgifters egenskaper kan man överföra dem till en gemensam skala. Med hjälp av gemensamma, överlappande uppgifter är det möjligt att skapa en stor mängd olika prov med kända mätegenskaper. Alla dessa prov uttrycker sedan resultat på en och samma skala.

Med IRT är det alltså inte, som i klassisk testteori, nödvändigt att prov är parallella för att resultat ska bli jämförbara. IRT bygger bl.a. på antaganden om endimensionalitet[6] och lokalt oberoende[7], men det finns också flerdimensionella modeller. Även i de fall antagandena inte är fullt ut uppfyllda har IRT-ansatsen visat sig vara användbar för att lösa många praktiska mätproblem. Ett sådant är att skapa prov som ger resultat som är jämförbara över tid, vilket

[6] Betyder att uppgifterna i ett prov i huvudsak mäter samma förmåga.
[7] Betyder att resultatet på en uppgift inte bygger på resultat från en annan uppgift.

bl.a. utnyttjas i de internationella studierna, t.ex. PISA och TIMSS. IRT bygger inte heller på antagandet att mätfelens storlek är oberoende av prestationsnivån (vilket den klassiska testteorin gör)[8]. Det gör det möjligt att konstruera prov som har god mätsäkerhet vid vissa viktiga punkter på skalan, t.ex. vid en definierad gräns för godkänt.

En nackdel med den moderna testteorin är att den är statistiskt och tekniskt komplicerad. För att hantera den tekniska komplexiteten finns ett stort antal väl utprövade datorprogram att tillgå. Vid användning av IRT-modellerna presenteras en elevs resultat i form av poäng på en godtyckligt bestämd kontinuerlig skala.[9] Ett problem som kan uppstå är att rangordningen mellan elever på grundval av dessa poäng i vissa fall kan bli en annan än när man baserar rangordningen på observerade poäng på provet. En fördel är dock att olika IRT-baserade prov kan kalibreras så att de uttrycker poäng på en i förväg definierad skala, t.ex. med medelvärdet 500 och standardavvikelsen 100.

De svenska nationella proven har hittills främst grundat sig på klassisk testteori. Detta framstår som något paradoxalt mot bakgrund av att den klassiska mätläran i första hand är utvecklad för att lösa problem i anslutning till mätning av individuella differenser i prestationer med hjälp av normrelaterade (gruppprelaterade) prov. Den moderna mätläran är däremot mer lämpad att stödja konstruktion och användning av den typ av nivåindelade skalor som används i kriterierelaterade prov av det slag vi i dag använder i Sverige genom att den ger möjlighet att välja uppgifter så att mätsäkerheten blir optimal vid betygsgränserna.

[8] Det finns dock varianter av klassisk testteori som hanterar s.k. betingade mätfel (*conditional standard error of measurement*) som varierar med provpoängen.
[9] Till skillnad från en vanlig poängskala som används för prov och där skalan i allmänhet uttrycks i hela poäng.

Referenser

Baird, J.-A. m.fl. (2013). Marker effects and examination reliability. Ofqual.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: Holt, Rinehart and Winston.

Qualifications and Curriculum Authority (2009). Research into marking quality. Studies to inform future work on national curriculum assessments. QCA/09/4042.

Skolverket (2003). Det nationella provsystemet – vad, varför och varthän? Dnr 2003:2038.

Skolverket (2015). Provbetygens stabilitet. Om nationella prov – Åk 9 1998–2012.

Provbetygens stabilitet och tillförlitlighet i gymnasieskolan

Sammanfattning

De nationella proven har under senare år fått allt större betydelse. En ny läroplan och en ny betygsskala infördes 2011. I denna bilaga ingår dock de nationella prov som gavs i den tidigare gymnasieskolan innan 2011 för att få tidsserier som sträcker sig över en någorlunda lång tidsperiod. Innan 2011 gällde också den fyrgradiga betygsskalan med betygen icke godkänt (IG), godkänt (G), väl godkänt (VG) och mycket väl godkänt (MVG).

Varje prov är behäftat med olika typer av mätfel och syftet med denna bilaga är främst att granska tre typer av fel. För det första fel som beror på att poänggränser för olika provbetyg inte är stabila över tid (s.k. kravgränsfel). För det andra slumpbaserade fel som främst beror på om en elev har tur eller otur med de uppgifter som ingår i provet, men även andra slumpfaktorer som har betydelse vid själva provtillfället kan inverka (s.k. slumpfel). För det tredje fel som är kopplade till att olika bedömare gör olika bedömningar av elevlösningar (s.k. bedömningsfel).

Resultaten redovisas i form av hur stor andel av eleverna som kan bedömas få ett felaktigt provbetyg på grund av något av de tre felen. De kursprov som främst granskas är de tidigare kärnämneskurserna svenska B, engelska A och matematik A.[1] De olika felen varierar beroende på vilka kurser proven gäller. Detta har att göra med hur proven är konstruerade. Proven i svenska och engelska består av delprov där vissa förmågor prövas med s.k. itembaserade

[1] Benämningen kärnämne används inte i den gymnasieskola som infördes 2011 utan där används benämningen gymnasiegemensamma ämnen. Dessutom betecknas kursernas progression med siffror i stället för med bokstäver, t.ex. svenska 1, svenska 2 osv.

och poängsatta delprov och andra förmågor med längre skriftliga uppgifter. Matematikprovens resultat baseras i huvudsak på uppgifter som poängsätts. Det betyder att proven i svenska och engelska har vissa delprov (itembaserade) där slumpfel dominerar och vissa (längre skriftliga uppgifter) där bedömningsfel dominerar. För matematikens del är det slumpfelen som dominerar eftersom få uppgifter visar nämnvärd bedömningsavvikelse.

För alla prov visar resultaten att kravgränsfelet är det fel som berör minst andel elever. I jämförelse är detta fel något större för matematikproven än för de andra proven. Kravgränsfelet medför att 5–6 procent av eleverna får ett annat provbetyg än vad som kan förväntas utifrån den långsiktiga trenden. Detta kan jämföras med 2–3 procent av eleverna för proven i svenska och engelska.

När det gäller slumpfelets betydelse finns inte data för att bedöma detta för de delprov i svenska och engelska som är itembaserade. För matematikens del visar analysen att slumpfelet medför att cirka 20–30 procent av eleverna får för högt eller lågt provbetyg i förhållande till deras ”faktiska” kunskaper. Å andra sidan visar de skriftliga delproven i svenska och engelska bedömningsfel av motsvarande storleksordning 10–30 procent beroende på delprov (enligt Statens skolinspektions ombedömningar).

Sammanfattningsvis kan man säga att för de aktuella proven tyder resultaten på att provbetygen är tillförlitliga, dvs. avspeglar elevernas kunskaper på ett korrekt sätt, till cirka 70 procent. Av provresultaten kan man dock inte särskilja vilka elever som ingår i dessa 70 procent.

Hur och om de olika felen samverkar kan inte avgöras av befintliga data. Det krävs också mer sofistikerade metoder för att konstruera ett mått som integrerar de olika felen. De redovisade resultaten bygger, som tidigare nämnts, på den fyrgradiga betygsskalan. Med 2011 års sexgradiga betygsskala[2] beräknas mätfelen att öka med cirka 50 procent. Med fler betygssteg blir betygsskalan mer preciserad, men samtidigt mindre robust.

[2] I den sexgradiga betygsskalan ingår betygen F, E, D, C, B och A.

Inledning

I en tidigare rapport har de nationella provbetygens stabilitet i årskurs 9 redovisats för åren 1998 till 2012.[3] Den handlade om att försöka få en bild av i vilken utsträckning betyg på de nationella proven kan anses likvärdiga mellan olika år. I den här bilagan görs motsvarande redovisning för gymnasieskolans nationella kursprov i svenska, engelska och matematik. Bilagan kan alltså ses som en fortsättning på den tidigare grundskolerapporten.

De nationella proven har främst två syften nämligen 1) att tjäna som betygsstöd för läraren och 2) att ge underlag för resultatjämförelser på olika nivåer i skolsystemet. För att jämförelser över tid ska vara möjliga krävs att provbetygen mellan olika år kan antas likvärdiga, dvs. att samma kunskapsnivå ska gälla för olika provbetyg olika år. Eller annorlunda uttryckt att provbetygen är stabila över tid. Syftet med denna bilaga och den tidigare nämnda rapporten för grundskolan är att försöka undersöka i vilken utsträckning detta antagande kan anses uppfyllt.

För att få ett rimligt tillförlitligt underlag behövs tidsserier som sträcker sig över en någorlunda lång tidsperiod. Det gör att tidsserien för prov konstruerade utifrån 2011 års ämnesplaner är för kort för att ge tillförlitliga resultat. De här redovisade resultaten baseras därför på prov enligt 1994 års kursplaner[4]. Metoderna för att fastställa provbetyg har dock inte ändrats med 2011 års ämnesplaner[5] och därför kan de resultat som erhålls för provbetyg enligt 1994 års betygsskala ligga till grund för en skattning av stabiliteten även enligt den nya betygsskalan.

De kunskapskrav som gäller sedan 2011 har en annan utformning och delvis annan vokabulär, men de kräver i likhet med 1994 års betygskriterier uttolkning och operationalisering i form av konkreta exempel på vad olika kunskapskrav innebär i form av iakttagbara elevprestationer. En sådan utveckling sker med tiden, men hur skrivningarna i 2011 års kunskapskrav tolkas av provkonstruktörer och betygssättande lärare i relation till motsvarande tolkning av 1994 års betygskriterier vet ingen i dagsläget. Man kan alltså inte uttala sig om nivån i de nya provbetygen i relation till de

[3] Skolverket (2015a).
[4] Till vilka även räknas de revisioner som gjorts, t.ex. Kursplan 2000.
[5] Ny benämning på det som tidigare kallades kursplaner.

gamla, men eftersom de använda metoderna är desamma bör man kunna dra vissa slutsatser om den förväntade stabiliteten i de nya provbetygen på basis av stabiliteten i de gamla. Detta är det ena övergripande syftet med denna bilaga.

Det andra övergripande syftet gäller provens betygsstödjande roll och frågan om hur tillförlitligt provbetyget är som underlag för den enskilda elevens betyg. Här kommer frågor som rör tillförlitligheten i bedömningen av proven in, liksom frågor om slumpens betydelse när det gäller urvalet av uppgifter till provet, elevens tillstånd vid det specifika provtillfället och annat som kan bidra till att ett begränsat prov inom en begränsad tidsram vid en viss bestämd tidpunkt inte till fullo förmår ge en komplett bild av en elevs kunskaper och förmågor i ett visst ämne eller en viss kurs.

Bilagan försöker således undersöka två saker, dels provbetygens långsiktiga stabilitet på nationell nivå där slumpmässiga fel och lokala bedömningsskillnader tenderar att ta ut varandra, dels provbetygens tillförlitlighet i relation till den enskilda individen som berörs av de olika felen.

Bakgrund

> Two types of error are likely to occur when cut scores on tests are used to classify students. These errors of classification do not occur because someone made a mistake. The errors will occur because no test can be perfectly reliable or perfectly valid, and because no method of setting cut scores is perfect.[6]

Alla kunskapsprov ger mer eller mindre felaktiga resultat. Kunskap är en flyktig och svårfångad materia och ju mer komplex och omfattande den är desto svårare är den att fånga. Dessutom är kunskap inte konstant utan varierar över tid.

Begreppet fel

Det kan vara på sin plats att betona att ”fel” i det här sammanhanget inte betyder att någon gjort fel eller handlat felaktigt. Det handlar mer om antingen skattningsbara statistiska avvikelser eller

[6] Zieky & Perie (2006).

om olika tolkningar och bedömningar där någon ”sann” bedömning inte finns eller kan konstrueras statistiskt. För prov baserade på poäng finns det tekniker för att skatta fel i form av avvikelser från statistiska väntevärden (”sanna” värden) eller avvikelser från en trendlinje (som kan ses som ett glidande medelvärde). När det gäller t.ex. bedömning av uppsatser och tolkningen av bedömningsanvisningar och kunskapskrav finns normalt ingen ”sann” tolkning tillgänglig, varken i absolut eller statistisk mening. I statistisk mening skulle en ”sann” tolkning kunna konstrueras genom att ett stort antal bedömare tolkar och betygssätter samma dokument varefter ett medelvärde beräknas som ”sann” bedömning. Ett sådant förfarande är dock av praktiska och ekonomiska skäl inte möjligt för en reguljär provverksamhet.

Slumpbaserade fel

Att utifrån ett (vanligen) skriftligt prov som en elev genomför under några timmar en viss bestämd dag göra en rättvisande bedömning av den elevens kunskaper och förmågor är en uppgift förenad med betydande osäkerhet. Eleven kan ha tur eller otur med de uppgifter som kommer på provet. Eller eleven kan av olika skäl ha en dålig dag: pollenallergi, kärleksbekymmer, tandvärk eller någon form av konflikt. Eller kanske har något hänt som i stället gör att eleven kan överprestera i relation till sin normala förmåga. Det senare är lite svårare att exemplifiera, men ett grundantagande i den testteori som ligger till grund för konstruktionen av prov är att de slumpfaktorer som påverkar en elevs provresultat tenderar att jämna ut sig för en hel grupp av elever (eller för samma elev om hon eller han får göra många prov).

När det gäller en enskild elevs provresultat vet man dock inte i vilken utsträckning och på vilket sätt resultatet är påverkat av slump. Ändå finns det olika metoder för att skatta detta slumpfel och när man analyserar provresultat måste hänsyn tas till denna typ av fel. Felen brukar kallas *den enskilda mätningens standardfel* och förkortas SEM.[7] Detta är en typ av *slumpbaserade* fel som gäller

[7] Efter den engelska benämningen SEM = *Standard Error of Measurement*. Vid redovisning av provresultat bör enligt den standard som fastställts av de professionella organisationerna AERA, APA och NCME (2014) provets standardfel anges.

generellt för alla provresultat och alla provdeltagare. Det är också ett dolt fel. Det är inte möjligt att säga att en viss elevs provresultat har si eller så stort mätfel. Däremot kan man säga att sannolikheten för ett en elev ska ha ett visst resultat bestäms av storleken på SEM.

Kravgränsfel

Ett annat fel tillkommer om provresultaten ska ligga till grund för någon form av klassificering, t.ex. genom att betyg ska sättas på provet. På prov som baseras på poäng eller belägg (vilka kan ses som en form av poäng som är kopplade till kunskapskrav på olika nivåer) innebär det att ange hur många poäng som krävs för olika provbetyg. För prov som utgörs av en enda stor uppgift, t.ex. en uppsats, kan i stället en helhetsbedömning baserad på de bedömningsanvisningar som provkonstruktören tillhandahåller ligga till grund för betyget. Vissa prov kan vara sammansatta av delprov där vissa kan vara poängbaserade och andra helhetsbedömda. I samtliga fall ska dock bedömningen resultera i att varje elev tilldelas ett provbetyg.

Det är dessa provbetyg som blir det manifesta utfallet av provet och som ger det resultat som t.ex. publiceras på Statens skolverks webbplats. Att fastställa de kravgränser som ska gälla för olika provbetyg är en grannlaga uppgift. Därför är det inte förvånande att kravgränser ibland fastställs och sedan, när väl de samlade resultaten finns, visar sig avvika från vad man kan förvänta sig. Detta är en typ av fel som vi här kallar *kravgränsfel.*

Bedömningsfel

Om det handlar om poängbaserade prov ligger svårigheten främst i att bestämma vilka poäng som ska krävas för olika betyg. Om det gäller uppsatser ligger provkonstruktörernas svårigheter framför allt i att formulera tydliga bedömningsanvisningar och att hitta förebildliga exempel på elevarbeten som av konstruktörerna klassificeras till olika betyg. För bedömaren ligger svårigheten i att tolka och värdera de elevarbeten som ska bedömas och betygssättas i relation till bedömningsanvisningarna. Här görs bedömningen ofta direkt i betygstermer med betydande påverkan på provbetyget.

För poängbaserade prov är utrymmet för tolkning i allmänhet mindre och bedömningen resulterar i ett kvantifierbart mått, vanligen ett antal poäng. Här blir kravgränsfelet inte så beroende av en enstaka bedömning utan mer av vid vilka poängnivåer kravgränserna lagts.

För prov med större inslag av helhetsbedömning blir å andra sidan tolkningsutrymmet större och här kommer kravgränsfelet att mer bero på hur olika bedömare tolkar uppgifter av t.ex. uppsatstyp. Som den korrekta tolkningen och bedömningen skulle man kunna se medelvärdet av de provbetyg en grupp av kvalificerade inbördes oberoende bedömare var och en skulle ge uppsatsen. Det finns forskning som visar att även bland kvalificerade bedömare finns det en betydande variation i bedömningen, men en bedömaröverensstämmelse över 70 procent brukar anses som acceptabel. För nationella prov i Sverige finns i allmänhet endast en bedömare, även om det i många fall också sker olika former av sam- eller medbedömning. Men i vilket fall torde man kunna säga att en stor del av de kravgränsfel som finns i prov med större inslag av helhetsbedömning består av *bedömningsfel*. Detta fel är inte direkt observerbart, men de ombedömningar som Skolinspektionen gör kan ses som en indikator.

Syfte och disposition

Bilagans syften kan formuleras i fem punkter:

- Att sammanställa provbetygens variation över tid.
- Att bestämma kravgränsfelets storlek.
- Att jämföra kravgränsfelet med standardfelet (SEM) och de fel i klassificering till olika betyg som SEM leder till.[8]
- Att jämföra kravgränsfel med bedömningsfel och slumpfel.
- Att ge ett underlag för att bedöma vilka felnivåer som är acceptabla inom ramen för prov av god kvalitet.

[8] Det som kallas *classification accuracy* på engelska.

Avsnittet *Kravgränsfel* i bilagan behandlar punkterna 1 och 2, provbetygens variation över tid och bestämningen av kravgränsfelet för de olika nationella prov som används i gymnasieskolan.

I avsnittet *Slumpbaserade fel* diskuteras slumpfelet med hjälp av ett par exempel. Där redogörs kortfattat för storleken av slumpmässiga mätfel och klassificeringsfel[9].

Avsnittet *Bedömningsfel* tar kortfattat upp fel som gäller bedömning av mer omfattande uppgifter av typen uppsatser. I detta fall används underlag från Skolinspektionens ombedömningar som exempel.

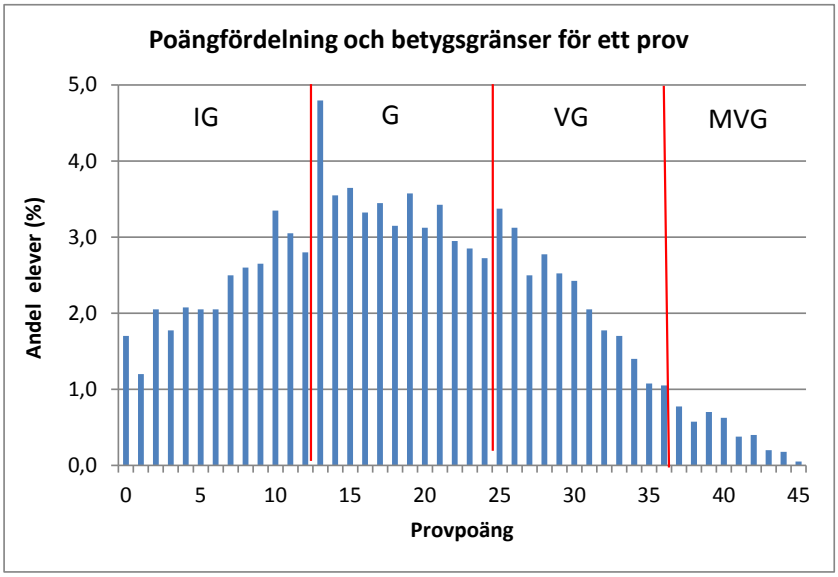
Slutligen sammanfattas och diskuteras de samlade resultaten i avsnittet *Sammanfattande diskussion*.

Kravgränsfel

För ett poängbaserat prov påverkar en förflyttning av en kravgräns ett poängsteg åt ena eller andra hållet många elevers provbetyg. Figur 1 visar ett exempel på poängfördelningen på ett prov med poänggränser för olika provbetyg.

[9] Utförligare redovisningar finns i Skolverket (2015b).

Figur 1 Exempel på poängfördelning på ett prov och poänggränser för olika betyg (1994 års betygsskala).



Av figuren framgår t.ex. att en förflyttning av den nedre poänggränsen för betyget G från 12 till 13 poäng innebär att andelen elever med betyget IG (EUM)[10] ökar med cirka 4,3 procentenheter (med motsvarande minskning för andelen elever med G). Om man tänker sig att det är ett nationellt prov som gäller en årskull på cirka 100 000 elever innebär höjningen således att cirka 4 300 fler elever får provbetyget IG. Det är alltså uppenbart att placeringen av kravgränserna har stor betydelse för provbetygens fördelning.

Data och metod

Skolverket samlar årligen in resultaten på genomförda nationella prov. Sedan 2003 samlas samtliga resultat in för grundskolans del. För gymnasieskolan samlades provresultat fram till 2011 in för ett stickprov av de elever som genomförde vårterminens version av respektive prov, men fr.o.m. höstterminen 2011 samlas samtliga

[10] EUM = ej uppnått målen. Exemplet är från grundskolan och där fanns inte betyget IG. Därför användes beteckningen EUM.

resultat, såväl höst- som vårtermin, in även på gymnasieskolans nationella prov.

Provbetyg

Varje nationellt prov ska åtföljas av anvisningar om vilka krav som ställs för olika betyg på det aktuella provet. Om det är ett prov som bedöms med poäng (numera ibland kallade belägg), anges hur många poäng som krävs för respektive betyg. Ibland anges också vilka kombinationer av poäng graderade som G-poäng, VG-poäng och MVG-poäng som krävs, vilket främst gäller prov i matematik.

Om det å andra sidan är prov med mer omfattande uppgifter av typen uppsatser kan uppgiften betygssättas direkt. Det gäller främst för vissa delprov i svenska och engelska. För att öka likvärdigheten i bedömningen tillhandahåller proven bedömningsanvisningar med exempel på poäng- eller betygssatta elevarbeten och argument för de gjorda valen. Kraven för olika provbetyg ska således vara fastställda i förväg.

Att bestämma dessa krav är en grannlaga uppgift. Det finns olika metoder, delvis beroende på ämnets och provens karaktär och utformning. Flertalet av metoderna bygger på bedömning av en grupp kvalificerade bedömare med goda insikter i styrdokumenten och god kännedom om de elevgrupper det handlar om.[11] Alla provuppgifter utprövas i förväg och då samlas vanligen statistik in över elevernas resultat. Även denna statistik kan sedan vara ett underlag vid bestämningen av kravgränserna för olika betyg.

Grundskolan

Ett rimligt antagande är att kunskapsförändringar för stora grupper som läser ämnen under många år är tröga processer. Då är det inte troligt att olika gruppers kunskaper i ett ämne varierar särskilt mycket mellan olika år, i synnerhet torde det gälla grupper som gör prov närliggande år. Plötsliga kunskapsfall (eller ökningar) vissa år kan då antas vara ett tecken på att mätningen inte är stabil utan påverkas av någon form av systematiskt mätfel. Ett sådant antagande förefaller åtminstone rimligt för grundskolan där hela

[11] Cizek & Bunch (2007).

årskullen (cirka 100 000 elever) läser samma ämne och gör samma prov, och där sedan 2003 samtliga provresultat samlas in. I det fallet kan med fog årliga variationer i huvudsak antas vara uttryck för svårigheten att bestämma betygsgränser eller betygskrav som är stabila över tid.[12]

Gymnasieskolan

För gymnasieskolan är bilden mer komplicerad. I de tidigare kärnämnena svenska B (sv B), engelska A (eng A) och matematik A (ma A) var proven obligatoriska, vilket innebär att alla elever berördes. Proven i högre kurser – engelska B (eng B), matematik B (ma B), matematik C (ma C) och matematik D (ma D) – var obligatoriska för de program där kurserna var avslutande. Även för mellanliggande kurser rekommenderades lärarna att använda proven. För elever på exempelvis Naturvetenskapsprogrammet var således för matematikens del endast proven i ma A och ma D obligatoriska, men rekommendationen var att genomföra även proven för ma B och ma C. I praktiken innebar det att de allra flesta eleverna gjorde även dessa prov.

Några faktorer som försvårar möjligheterna att beräkna provbetygens stabilitet för gymnasieskolans del är följande:

- Varje prov i gymnasieskolan ges både höst- och vårtermin. Provresultat som finns tillgängliga har samlats in endast för vårterminerna.
- Elevsammansättningen med avseende på program varierade mellan vår- och hösttermin. Elever på yrkesförberedande program genomförde t.ex. oftast provet i ma A under den andra terminen, dvs. vårterminen, medan elever på t.ex. Naturvetenskapsprogrammet i allmänhet gjorde det första terminen, dvs. höstterminen. Elevsammansättningen vår- och hösttermin torde därmed ha varit olika. Men eftersom resultat endast har samlats in för vårterminens prov förefaller det rimligt att anta att den insamlade elevgruppens sammansättning inte har varierat alltför mycket mellan olika år. Det är då möjligt att skatta provbetygens avvikelser för vårterminens kursprov. Fastläggningen

[12] Se Skolverket (2015a) för närmare beskrivning.

av kravgränser sker på samma sätt för vårens och höstens kursprov. Därför är det rimligt att anta att de genomsnittliga avvikelser som kan skattas för vårterminens provbetyg i huvudsak även gäller för höstterminens kursprov trots att andelen elever med höga provbetyg kan antas vara högre på hösten för t.ex. ma A[13]. Det är inte den genomsnittliga betygsnivån i sig (uttryckt i hur många procent av eleverna som har respektive provbetyg) som indikerar avvikelse utan hur mycket de årliga resultaten avviker från den långsiktiga nivån (trenden) över hela den ingående tidsperioden.

- Provresultat har endast insamlats från ett stickprov av skolor (cirka en sjättedel varje år så att alla skolor har deltagit i insamlingen inom sex år). Hur representativt varje stickprov är för hela gruppen är oklart. Antalet rapporterade provresultat varierar mellan år och kurser, vilket gör det tveksamt i vilken utsträckning det är rimligt att bortse från urvalsfelen. Det här gäller i synnerhet för de högre matematikproven (ma C och ma D) och är den mest osäkra punkten.[14] I bilagan görs dock försök att bedöma urvalfelets betydelse.
- För proven i sv B, eng A och eng B redovisades fram till vårterminen 2005 endast delprovsbetyg, men inget samlat provbetyg. Delprovsbetygen var därtill redovisade för olika antal elever, vilket gör det svårt att i efterhand beräkna ett samlat provbetyg. Från och med vårterminen 2005 redovisas ett samlat provbetyg av Skolverket men först vårterminen 2008 anges att det samlade provbetyget endast baseras på resultat för elever som deltagit i och har resultat på samtliga delprov. Vad som gällde i det avseendet 2005–2007 är oklart, men om det visar sig att resultaten på dessa prov inte avviker på något anmärkningsvärt sätt i förhållande till resultaten på efterföljande prov kommer de att ingå i underlaget för studien.

[13] Eftersom flertalet elever på Naturvetenskapsprogrammet genomförde provet på hösten.
[14] Kan möjligen skattas om man undersöker hur eleverna fördelar sig på olika program de olika åren. För många program blir dock elevgrupperna små så det är tveksamt hur tillförlitligt resultatet blir. En annan tänkbar möjlighet är att välja endast det största programmet (Samhällsvetenskapsprogrammet för flertalet kursprov), men då blir elevantalet i stickprovet ändå tämligen begränsat (jämför figur 1).

Data

De data som används har hämtats från Skolverkets webbplats.[15] Nedanstående bild ger ett exempel på hur data kan se ut. De resultat som används i analyserna gäller ”Totalt program”. Som framgår av bilden rapporterades den aktuella terminen (vårterminen 2005) totalt 5 875 provresultat.

Figur 2 **Exempel på Skolverkets redovisning av provresultat i Siris.**

SiRiS

Gymnasieskolan - Resultat på kursprov i Svenska B

Elever som började hösten 2010 eller tidigare

Vald termin: VT05 Vald organisation: Riket

ANALYSSTÖD EXCEL NYTT URVAL

Nationella program samt utbildning vid fristående skolor med anknytning till nationella program		Antal elever	Provbetyg Betygsfördelning (%) IG	G	VG	MVG	Genm. prov-betp.
BF	Barn- och fritidsprogrammet	407	14	59	26	..	10,0
BP	Byggprogrammet	270	13	76	10	..	9,3
EC	Elprogrammet	361	14	68	18	..	9,6
EN	Energiprogrammet	51	..	71	22	.	10,3
ES	Estetiska programmet	513	10	44	38	7	11,7
FP	Fordonsprogrammet	298	11	72	16	..	9,7
HP	Handels- och administrationsprogrammet	402	15	62	22	..	9,6
HR	Hotell- och restaurangprogrammet	232	17	58	23	2	9,7
HV	Hantverksprogrammet	102	5	54	40	..	11,6
IP	Industriprogrammet	97	19	70	10	..	8,8
LP	Livsmedelsprogrammet	14	..	50	..	..	9,6
MP	Medieprogrammet	420	11	51	35	3	10,9
NP	Naturbruksprogrammet	198	4	44	43	10	12,8
NV	Naturvetenskapsprogrammet	684	3	28	51	18	14,1
OP	Omvårdnadsprogrammet	125	9	57	31	..	11,0
SP	Samhällsvetenskapsprogrammet	1 350	7	34	44	15	13,0
TE	Teknikprogrammet	351	9	49	36	6	11,5
	Totalt program	5 875	10	49	34	7	11,5
IB	International baccalaureate	-	..	..	..	..	.
IV	Individuellt program	..	..	..	..	..	..
OV	Övriga	83	..	66	29	..	11,4
SM	Specialutformat program	799	8	49	38	5	11,6

Rapportbeskrivning och definitioner etc.

Resultaten redovisas utifrån den utbildning som angivits vid inrapporteringen.

Från och med vårterminen 2008 redovisas endast resultaten för elever som har genomfört ett ordinarie nationellt prov i respektive kurs samt har ett korrekt provbetyg inrapporterat. T ex krävs att man utfört alla delprov i det ordinarie provet för att kunna få ett provbetyg.

För övriga kursprov är tillgängliga data av liknande slag som ovan. Med tanke på de begränsningar och osäkerheter som tidigare nämnts får slutsatserna av analyserna tolkas med försiktighet. Men om resultaten blir i paritet med motsvarande analyser för grundskolan kan det rimligen ses som en indikation på att stabiliteten i

[15] http://siris.skolverket.se/siris/f?p=SIRIS:114:0::NO

provbetyg i grundskolan och gymnasieskolan är av samma storleksordning, vilket är vad man kan förvänta sig eftersom tillvägagångssättet vid bestämning av kraven för olika provbetyg bygger på samma metoder.

Metod

Den metod som används är densamma som redovisas i motsvarande grundskolerapport.[16] Det innebär att trendlinjer skattas för de tidsserier av data som används, främst andelen elever i procent som fått olika provbetyg.[17] Därefter jämförs det värde trendlinjen ger med motsvarande datavärde ur Siris. Skillnaden mellan dessa värden (residualen) beräknas och kan sedan ses som ett mått på hur stor avvikelsen är för de olika provbetygen det aktuella året uttryckt i *procentenheter*.[18]

Andel avvikelser och andel elever med avvikande provbetyg

I vissa tabeller kommer summan av absolutvärdena (alla värden räknas som positiva) av residualerna att användas som mått på totala avvikelserna (se t.ex. tabell 3). Detta ger totala andelen avvikelser från trendlinjen oberoende av om de är positiva eller negativa. Samtidigt är summan av andelarna som har respektive betyg hela tiden 100. Om det är för många elever i en betygsgrupp måste det vara för få i någon annan för att summan ska vara 100. En avvikelse uppåt i en grupp måste således kompenseras av en avvikelse neråt i någon (eller några andra grupper). Detta gör att andelen avvikande elever i sådana fall där summan ska vara 100 procent är *hälften så stor* som andelen avvikelser.

Resultat baserade på hur stor andel av eleverna som fått respektive provbetyg olika år redovisas i första hand, och i andra

[16] En närmare beskrivning finns i Skolverket (2015a).
[17] Till mindre del även för genomsnittlig betygspoäng GBP.
[18] Det innebär alltså att resultatet uttrycks i andel av samtliga elever (100 procent). Ett annat sätt att uttrycka avvikelsen är i procent, dvs. hur stor avvikelsen är i relation till den förväntade andelen för det aktuella betyget. (Jämför t.ex. valresultat. Partiet X minskar från 10 till 8 procent, dvs. 2 procentenheter. I procent däremot blir minskningen 2/10 = 0,20, dvs. 20 procent). Både procentenheter och procent kommer att användas i resultatredovisningen.

hand för den genomsnittliga betygspoängen (GBP)[19]. Ett grundantagande är att en *liten årlig variation* (residual, avvikelse från trendlinjen) i andelen elever som får respektive provbetyg är *ett mått på hög stabilitet*.

Genomsnittlig betygspoäng

GBP ger ett grovt mått som inte kan översättas till andel eller antal elever, utan endast tjänar som en indikation. Detta beror på den asymmetriska poängsättningen av betygsskalan.[20] Beräkningar baserade på andel elever med olika provbetyg ger ett mer precist mått.

Resultat

Resultaten visas först för svenska B, därefter för engelska A och slutligen för matematik A. Dessa prov har genomförts av alla elever. Resultaten för svenska B redovisas tämligen utförligt för att tydliggöra den använda metoden. För övriga kurser ges en mindre ingående beskrivning eftersom samma metodik används för redovisningen av varje kursprov. För att inte tynga framställningen med alltför många diagram och tabeller läggs dessutom resultaten på proven i engelska B och matematik B–D i appendix 1.

De avvikelser som redovisas här antas gälla avvikelser som beror på att betygsgränser eller betygskrav varit för stränga eller för milda. Om alla betygsgränserna varit korrekta skulle avvikelserna vara nära noll (förutsatt att den antagna linjära modellen är korrekt). Det finns andra typer av fel som beror på reliabilitetsbrister av olika slag.[21] Från sådana fel bortses i den här resultatredovisningen. Däremot tas de upp i senare avsnitt.

[19] GBP = (0*IG+10*G+15*VG+20*MVG)/100, där IG, G, osv. är andelen elever i procent med respektive betyg.
[20] Den är främst politiskt motiverad för att användas vid beräkning av meritvärden med särskild vikt lagd vid att uppnå betyget G.
[21] Den enskilda mätningens standardfel eller på engelska *Standard error of measurement* (SEM) för främst poängbaserade prov samt bedömningsfel beroende på olika bedömares olika syn på komplexa uppgifter av typen uppsatser, t.ex. avvikelse mellan lärarens bedömning och bedömningen av Skolinspektionens ombedömare.

Svenska B

Observerade data

Tabell 1 visar resultat på provet i svenska B: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna[22]. Före 2005 rapporteras som nämnts inga sammanfattande provbetyg i svenska och engelska.

Tabell 1 **I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, sv B.**

Sv B							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2005	5875	10	49	34	7	11,5	11,4
2006	6089	9	45	37	10	11,9	12,1
2007	7515	6	48	36	10	12,1	12,2
2008	7119	11	47	33	9	11,4	11,5
2009	5570	8	47	36	10	12,0	12,1
2010	7890	7	44	39	11	12,3	12,5
2011	8099	8	45	35	12	12,1	12,2
Medel	6880	8,4	46,4	35,7	9,9	11,9	12,0
Std	1026	1,7	1,8	2,0	1,6	0,3	0,4

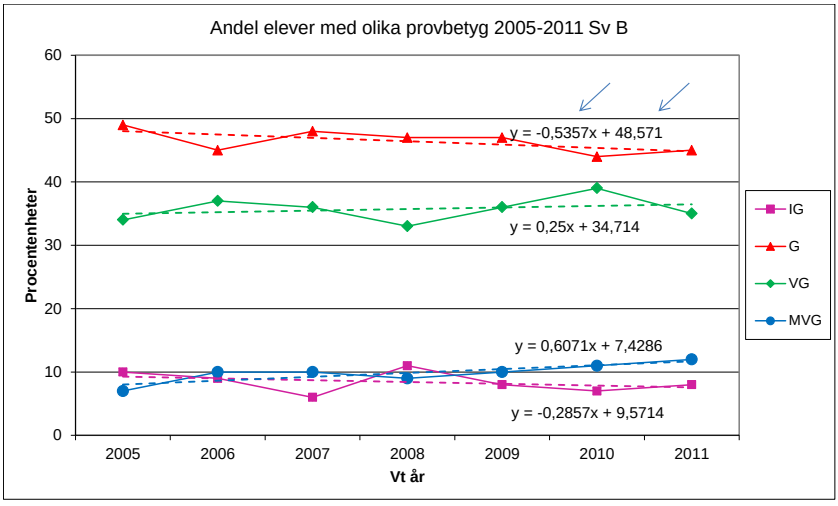
Tabellen visar exempelvis att antalet elever i stickproven varierar mellan 5 570 och 8 099 elever och G och VG är de vanligaste provbetygen (cirka 80 procent sammantaget). Man kan t.ex. notera att en förhållandevis stor andel elever fick IG vårterminerna 2005 och 2008, samtidigt som en förhållandevis liten andel fick MVG. Detta ledde till låga värden på GBP. Vissa data sticker alltså ut, men det är svårt att mer i detalj värdera om avvikelsen eller stabiliteten i provbetygen för svenska B kan anses som stor eller liten utan en mer systematisk granskning.

[22] Om skillnaderna mellan de olika GBP-måtten indikerar att det finns vissa avrundningsfel kan inte avgöras utifrån befintliga data i Siris.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 3). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer.[23] De senare används sedan för att beräkna modellvärden för de olika procentandelarna. I figur 3 visas värdena från Siris samt anpassade trendlinjer (de streckade linjerna) med tillhörande ekvationer.

Figur 3 **Andel elever med olika provbetyg i svenska B vt 2005–2011 samt trendlinjer med ekvationer.**



Av figuren kan man t.ex. notera att andelen elever med G på provet i genomsnitt avtar med drygt 0,5 procentenheter per år (anges av negativt värde på koefficienten i ekvationen för G, k=-0,5357) och att andelen elever med provbetyget MVG ökar med i genomsnitt 0,6 procentenheter per år (k=0,6071). Detta kan också utläsas ur den högra delen av tabell 2 där modellvärdena för varje år är beräknade utifrån de värden på k och m som anges i tabellhuvudet.

[23] Görs med hjälp av kalkylprogrammet Excel. Ekvationen för en rät linje har formen y = k*x + m där y är värdet på y-axeln (procentenheter i det aktuella fallet), k är koefficienten framför x (anger lutningen på linjen och hur många procentenheter värdet på y ändras per år) och m slutligen anger var linjen skär y-axeln.

Tabell 2 **Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). "k" och "m" är parametrar för respektive betygs trendlinje, sv B.**

Sv B														
Vt	Antal	Betyg (%)				GBP		k=	-0,286	-0,536	0,25	0,607		
								m=	9,57	48,57	34,71	7,43	GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad	Löpnr	IG	G	VG	MVG	modell	
2005	5875	10	49	34	7	11,5	11,4	1	9,3	48,0	35,0	8,0	11,7	
2006	6089	9	45	37	10	11,9	12,1	2	9,0	47,5	35,2	8,6	11,8	
2007	7515	6	48	36	10	12,1	12,2	3	8,7	47,0	35,5	9,3	11,9	
2008	7119	11	47	33	9	11,4	11,5	4	8,4	46,4	35,7	9,9	12,0	
2009	5570	8	47	36	10	12,0	12,1	5	8,1	45,9	36,0	10,5	12,1	
2010	7890	7	44	39	11	12,3	12,5	6	7,9	45,4	36,2	11,1	12,2	
2011	8099	8	45	35	12	12,1	12,2	7	7,6	44,8	36,5	11,7	12,3	
Medel	6880	8,4	46,4	35,7	9,9	11,9	12,0	Medel	8,4	46,4	35,7	9,9	12,0	
Std	1026	1,7	1,8	2,0	1,6	0,3	0,4	Std	0,6	1,2	0,5	1,3	0,2	

Sammanfattande resultat

Om man antar att trendlinjen representerar en stabil flerårig utveckling kan *skillnaden mellan den observerade andelen elever med ett visst betyg och motsvarande andel enligt modellen (trendlinjen)* ses som ett mått på hur stor andel elever som tilldelas ett provbetyg som avviker från det betyg de kan förväntas ha enligt modellen. Avvikelsen, eller felet om vi antar att modellen är korrekt, blir då andelen betyg i tabellens vänstra del minus motsvarande andel betyg i den högra delen av tabell 2, (O-M)[24]. För 2005 blir således, för betyget IG, avvikelsen lika med 10-9,3 = 0,7 procentenheter, dvs. andelen elever med IG på provet var 0,7 procentenheter högre än förväntat utifrån den långsiktiga trenden (räknat på samtliga elever, dvs. 100 procent).

Om man i stället undrar hur stor del av eleverna med provbetyget IG som enligt trendvärdet inte borde ha IG (positiv skillnad) eller borde ha fått IG (negativ skillnad) får man i stället beräkna hur stor andel 0,7 är av 9,3 (modellvärdet för IG), avrundat till heltal ger det 8 procent. De elever som fick IG på provet 2005 var alltså 8 procent för många enligt modellen eftersom värdet är positivt. Dessa elever borde ha haft betyget G (något annat alternativ finns ju inte

[24] O(bserverat värde)-M(odellvärde).

för den som har fått IG). Tabell 3 visar motsvarande resultat för samtliga betyg.[25]

Tabell 3 **Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. "Sum(ABS)" anger hur stor andel elever (procent av alla) totalt[26] som har från trendlinjen avvikande betyg (för högt eller för lågt). "Medel(ABS)" anger genomsnittlig årlig avvikelse i procent av den andel elever som enligt modellen förväntas ha respektive betyg, sv B.**

	Avvikelse(% av totalt)				
År	IG	G	VG	MVG	Sum(ABS)
2005	0,7	1,0	-1,0	-1,0	4
2006	0,0	-2,5	1,8	1,4	6
2007	-2,7	1,0	0,5	0,7	5
2008	2,6	0,6	-2,7	-0,9	7
2009	-0,1	1,1	0,0	-0,5	2
2010	-0,9	-1,4	2,8	-0,1	5
2011	0,4	0,2	-1,5	0,3	2
Medel	0,0	0,0	0,0	0,0	4
Std	1,6	1,4	1,9	0,9	2

	Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG
2005	8	2	-3	-13
2006	0	-5	5	16
2007	-31	2	2	8
2008	31	1	-8	-9
2009	-2	2	0	-4
2010	-11	-3	8	-1
2011	6	0	-4	3
Medel(ABS)	13	2	4	8
Std	19	3	5	10

"Sum(ABS)" i den vänstra delen av tabellen anger totala andelen avvikelser (procentenheter) från modellens värden oberoende av om andelen varit för hög eller för låg. Tabell 3 visar att avvikelserna är mellan cirka 2 (2009) och 7 (2008) procent under perioden vårterminen 2005 till vårterminen 2012, med ett medelvärde för andelen avvikande eller felaktiga betyg på 4 procent.

Medel(ABS) i den högra delen av tabellen anger hur stor andel av eleverna (procent) som i genomsnitt fått ett annat betyg än det förväntade i relation till det enligt den långsiktiga trenden förväntade betyget. Man kan konstatera att för sv B är det främst betyget IG som visar stora avvikelser. År 2007 var gruppen elever med IG 31 procent mindre än förväntat och 2008 var den 31 procent större än förväntat. I genomsnitt avviker gruppen som fått IG med 13 procent (uppåt eller neråt) i relation till vad den enligt modellen

[25] Siris visar andelen elever i hela procent, vilket gör att avrundningsfelen kan bli ganska stora. De beräknade värdena räknas i tiondelar av procent, vilket får hållas i minnet då resultaten värderas.
[26] Dvs. procentenheter.

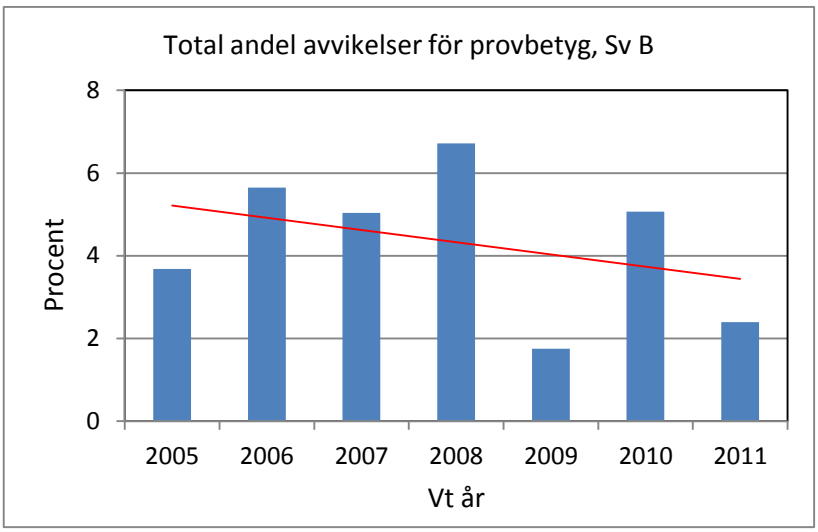
förväntas vara. Variationen mellan olika år är dessutom stor, vilket framgår av tabellen och standardavvikelsen som är 19 procent. Även MVG visar ganska stor avvikelse medan den är mindre för G och VG.

Skillnaderna mellan tabellens vänstra och högra del beror på att avvikelsen jämförs med 100 procent i vänstra delen (procentenheter) och med procentandelen för det aktuella betyget i högra delen (procent). Båda värdena är korrekta, men *värderingen* kan bli olika beroende på vilket man väljer att lyfta fram.

Figur 4 nedan illustrerar den genomsnittliga avvikelsen i klassifikation räknat på samtliga avvikelser (Sum(ABS)). Man kan notera att trenden är avtagande, dvs. variationen i provbetygens avvikelse tycks minska, eller omvänt – provbetygens stabilitet tycks visa en tendens att öka. *En avtagande trend för avvikelsen innebär således en över tid ökande stabilitet i provbetygen.* ”Fel” betyder den andel av eleverna som klassificerats med ett annat provbetyg än vad som kan förväntas utifrån den långsiktiga betygsutvecklingen, dvs. trenden för den aktuella tidsperioden 2005 till 2011.

Observera att om man vill uttrycka avvikelsen som andel elever som fått ”fel” betyg får man *halvera värdena* i figur 4.

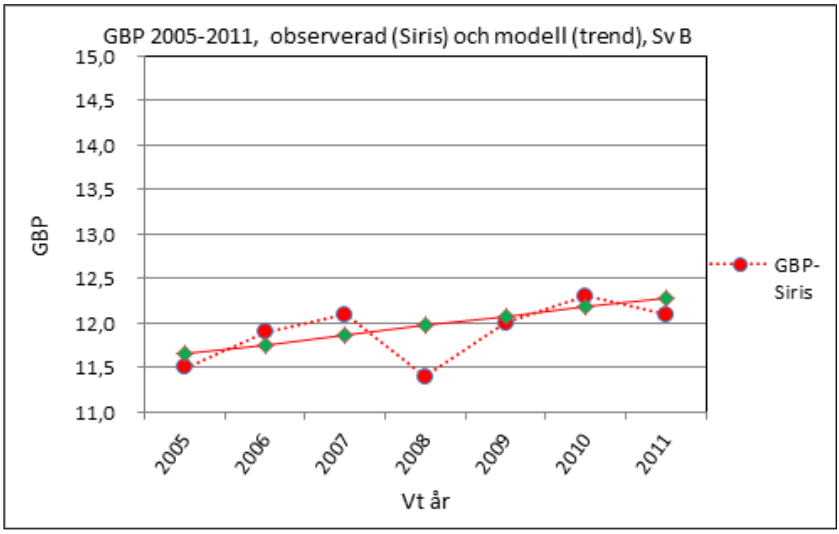
Figur 4 **Total andel avvikelser för svenska B. Andelen elever med avvikande betyg=hälften.**



Genomsnittlig betygspoäng

Figur 5 visar den genomsnittliga betygspoängen (GBP) för provbetygen i sv B under samma period. GBP ökar under perioden, vilket stämmer med att såväl andelen VG som MVG ökar medan IG och G minskar (figur 3 och tabell 2). I synnerhet ändringar mellan IG och G får stor betydelse på grund av den skala som används för betygspoäng och som ligger till grund för beräkningen av GBP (steget IG–G ger 10 poäng mot övriga betygssteg 5 poäng).

Figur 5 **Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (de gröna punkterna på trendlinjen), svenska B.**



Man kan notera en kraftig avvikelse 2008, vilket stämmer med tabell 3 där det framgår att ovanligt många elever fick betyget IG. I övrigt är det dock svårt att dra några mer precisa slutsatser av den genomsnittliga betygspoängen. Det är t.ex. inte möjligt att på något enkelt sätt översätta GBP till hur stor andel elever som fått från trenden avvikande provbetyg olika år. GBP får därmed mer ses som en långsiktig trendindikator än som ett instrument för att bestämma storleken på olika avvikelser.

Engelska A

Observerade data

Tabell 4 visar resultat på provet i engelska A: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna.

Tabell 4 **I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, eng A.**

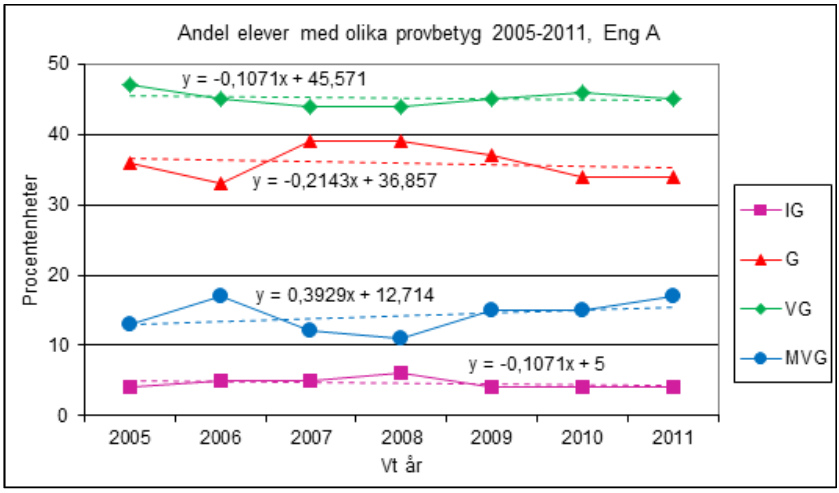
Eng A							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2005	8094	4	36	47	13	13,2	13,3
2006	11474	5	33	45	17	13,5	13,5
2007	13505	5	39	44	12	12,9	12,9
2008	10703	6	39	44	11	12,7	12,7
2009	8373	4	37	45	15	13,3	13,5
2010	12820	4	34	46	15	13,5	13,3
2011	12537	4	34	45	17	13,5	13,6
Medel	11072	4,6	36,0	45,1	14,3	13,2	13,2
Std	2144	0,8	2,4	1,1	2,4	0,3	0,3

Inget särskilt kan noteras i tabell 4 utom möjligen att antalet elever i stickproven varierar en hel del.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 6). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer. De senare används sedan för att beräkna modellvärden för de olika procentandelarna.

Figur 6 **Andel elever med olika provbetyg i engelska A vt 2005–2011 samt trendlinjer med ekvationer.**



Man kan notera att andelen IG, G och VG avtar medan andelen MVG ökar över tid.

Tabell 5 **Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). "k" och "m" är parametrar för respektive betygs trendlinje, eng A.**

Eng A							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2005	8094	4	36	47	13	13,2	13,3
2006	11474	5	33	45	17	13,5	13,5
2007	13505	5	39	44	12	12,9	12,9
2008	10703	6	39	44	11	12,7	12,7
2009	8373	4	37	45	15	13,3	13,5
2010	12820	4	34	46	15	13,5	13,3
2011	12537	4	34	45	17	13,5	13,6
Medel	11072	4,6	36,0	45,1	14,3	13,2	13,2
Std	2144	0,8	2,4	1,1	2,4	0,3	0,3

k=	-0,107	-0,214	-0,107	0,393	
m=	5	36,86	45,57	12,71	GBP
Löpnr	IG	G	VG	MVG	modell
1	4,9	36,6	45,5	13,1	13,1
2	4,8	36,4	45,4	13,5	13,1
3	4,7	36,2	45,2	13,9	13,2
4	4,6	36,0	45,1	14,3	13,2
5	4,5	35,8	45,0	14,7	13,3
6	4,4	35,6	44,9	15,1	13,3
7	4,3	35,4	44,8	15,5	13,4
Medel	4,6	36,0	45,1	14,3	13,2
Std	0,2	0,5	0,2	0,8	0,1

Inte heller tabell 5 visar några särskilda avvikelser.

Sammanfattande resultat

Resultattabellen (tabell 6) redovisar, liksom tidigare för sv B, avvikelsen i procentenheter i den vänstra delen och avvikelsen som procent av förväntad andel elever med det aktuella betyget i den högra delen (negativa värden anger för få givna betyg, positiva värden för många).

Tabell 6 **Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. "Sum(ABS)" anger hur stor andel elever (procent av alla) totalt som fått från trendlinjen avvikande betyg (för högt eller för lågt). "Medel(ABS)" anger genomsnittlig årlig avvikelse i procent av den andel elever som enligt modellen förväntas ha respektive betyg, eng A.**

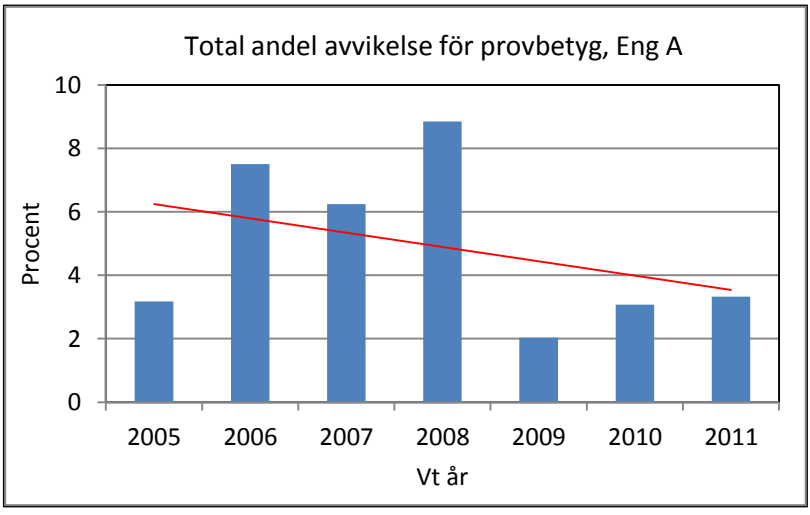
	Avvikelse(% av totalt)				
År	IG	G	VG	MVG	Sum(ABS)
2005	-0,9	-0,6	1,5	-0,1	3
2006	0,2	-3,4	-0,4	3,5	8
2007	0,3	2,8	-1,2	-1,9	6
2008	1,4	3,0	-1,1	-3,3	9
2009	-0,5	1,2	0,0	0,3	2
2010	-0,4	-1,6	1,1	-0,1	3
2011	-0,3	-1,4	0,2	1,5	3
Medel	0,0	0,0	0,0	0,0	5
Std	0,8	2,4	1,0	2,2	3

	Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG
2005	-18	-2	3	-1
2006	4	-9	-1	26
2007	7	8	-3	-14
2008	31	8	-3	-23
2009	-10	3	0	2
2010	-8	-4	2	0
2011	-6	-4	0	10
Medel(ABS)	12	6	2	11
Std	16	7	2	16

För engelska A är den genomsnittliga totala avvikelsen under perioden 5 procent (medelvärdet för Sum(ABS)). Åren 2006 och 2008 utmärker sig med de största avvikelserna. Under de senare åren är dock avvikelsen betydligt lägre och trenden pekar på ökad stabilitet i provbetygen. Detta illustreras av figur 7.

Man kan också notera i den högra delen av tabellen att det framför allt är gruppen elever med förväntade betyg IG eller MVG som visar stor inbördes variation. Där kan gruppen med förväntat betyg IG variera mellan att vara 31 procent för stor till 18 procent för liten. För MVG gäller liknande siffror. Det som leder till de höga procentsatserna är att antalet elever i kategorin är förhållandevis litet, i synnerhet för IG (jämför tabell 5).

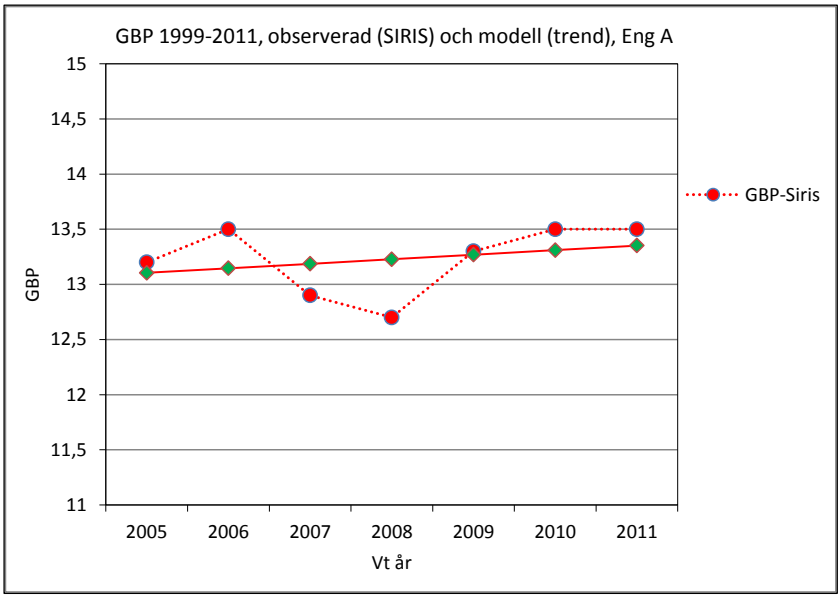
Figur 7 **Total andel avvikelser i engelska A. Andelen elever med avvikande betyg=hälften.**



Genomsnittlig betygspoäng

Diagrammet över GBP pekar på större avvikelser för 2006 och 2008 och ger därmed en likartad bild av utvecklingen.

Figur 8 Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (trendlinjen), engelska A.



Matematik A

Observerade data

Tabell 7 visar resultat på provet i matematik A: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna. För kursproven i matematik finns provbetyg rapporterade fr.o.m. vårterminen 1999 eller 2000 och resultaten blir därmed mer tillförlitliga än för svenska och engelska.

Tabell 7 I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, ma A.

Ma A							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
1999	6526	26	41	26	8	9,5	9,6
2000	8007	36	40	17	6	7,9	7,8
2001	11995	18	53	20	8	10	9,9
2002	8356	18	45	26	11	10,6	10,6
2003	8316	25	49	21	6	9,2	9,3
2004	8909	22	40	29	8	10	10,0
2005	10417	24	38	27	10	10	9,9
2006	9987	28	42	22	8	9,1	9,1
2007	13183	18	48	27	7	10,2	10,3
2008	10067	24	48	22	6	9,3	9,3
2009	9034	18	43	29	10	10,7	10,7
2010	12433	31	39	24	6	8,8	8,7
2011	11775	22	47	23	8	9,8	9,8
Medel	9923	23,8	44,1	24,1	7,8	9,6	9,6
Std	1979	5,5	4,6	3,6	1,7	0,8	0,8

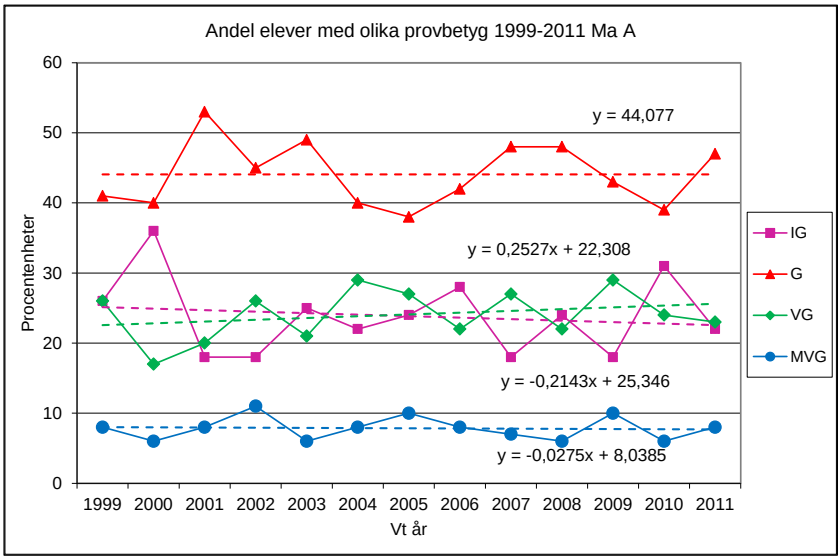
I jämförelse med tidigare redovisade provbetyg för svenska och engelska kan man för matematik A konstatera en betydligt större andel IG samt betydligt lägre värden på GBP.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 9). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer. De senare används sedan för att beräkna modellvärden för de olika procentandelarna.

I figur 9 kan man utan svårighet se att den årliga variationen i provbetyg, eller de årliga avvikelserna för olika provbetyg om man jämför med trendlinjen, är betydande i relation till vad vi sett för svenska och engelska.

Figur 9 **Andel elever med olika provbetyg i matematik A vt 1999–2011 samt trendlinjer med ekvationer.**



Av figur 9 framgår att för betyget G är trenden att ingen förändring alls skett över tid. Däremot är den årliga variationen betydande. Detta gäller också IG och VG, men i mindre grad MVG. Tabell 8 visar resultaten i siffror.

Tabell 8 **Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). "k" och "m" är parametrar för respektive betygs trendlinje, ma A.**

Ma A														
Vt	Antal	Betyg (%)				GBP		k=	-0,214	0	0,253	-0,028		
								m=	25,35	44,08	22,31	8,04	GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad	Löpnr	IG	G	VG	MVG	modell	
1999	6526	26	41	26	8	9,5	9,6	1	25,1	44,1	22,6	8,0	9,4	
2000	8007	36	40	17	6	7,9	7,8	2	24,9	44,1	22,8	8,0	9,4	
2001	11995	18	53	20	8	10	9,9	3	24,7	44,1	23,1	8,0	9,5	
2002	8356	18	45	26	11	10,6	10,6	4	24,5	44,1	23,3	7,9	9,5	
2003	8316	25	49	21	6	9,2	9,3	5	24,3	44,1	23,6	7,9	9,5	
2004	8909	22	40	29	8	10	10,0	6	24,1	44,1	23,8	7,9	9,6	
2005	10417	24	38	27	10	10	9,9	7	23,9	44,1	24,1	7,8	9,6	
2006	9987	28	42	22	8	9,1	9,1	8	23,6	44,1	24,3	7,8	9,6	
2007	13183	18	48	27	7	10,2	10,3	9	23,4	44,1	24,6	7,8	9,7	
2008	10067	24	48	22	6	9,3	9,3	10	23,2	44,1	24,8	7,8	9,7	
2009	9034	18	43	29	10	10,7	10,7	11	23,0	44,1	25,1	7,7	9,7	
2010	12433	31	39	24	6	8,8	8,7	12	22,8	44,1	25,3	7,7	9,8	
2011	11775	22	47	23	8	9,8	9,8	13	22,6	44,1	25,6	7,7	9,8	
Medel	9923	23,8	44,1	24,1	7,8	9,6	9,6	Medel	23,9	44,1	24,1	7,8	9,6	
Std	1979	5,5	4,6	3,6	1,7	0,8	0,8	Std	0,8	0,0	1,0	0,1	0,1	

Standardavvikelsen i den vänstra tabelldelen visar störst variation för IG och minst för MVG (vilket också framgår av figur 9). I den högra halvan kan man notera att enligt modellen (trendlinjen) är andelen elever med provbetyget G (44,1 procent) konstant över perioden, dvs. samma sak som vi kunde se i diagrammet.

Sammanfattande resultat

Resultattabellen (tabell 9) redovisar liksom tidigare i den vänstra delen avvikelsen i procentenheter och i den högra delen som procent av förväntad andel med det aktuella betyget.

Tabell 9 **Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. "Sum(ABS)" anger hur stor andel elever (procent av alla) totalt som fått från trendlinjen avvikande betyg (för högt eller för lågt). "Medel(abs)" anger genomsnittlig årlig avvikelse i procent av den andel elever som enligt modellen förväntas ha respektive betyg, ma A.**

	Avvikelse(% av totalt)				
År	IG	G	VG	MVG	Sum(ABS)
1999	0,9	-3,1	3,4	0,0	7
2000	11,1	-4,1	-5,8	-2,0	23
2001	-6,7	8,9	-3,1	0,0	19
2002	-6,5	0,9	2,7	3,1	13
2003	0,7	4,9	-2,6	-1,9	10
2004	-2,1	-4,1	5,2	0,1	11
2005	0,1	-6,1	2,9	2,2	11
2006	4,4	-2,1	-2,3	0,2	9
2007	-5,4	3,9	2,4	-0,8	13
2008	0,8	3,9	-2,8	-1,8	9
2009	-5,0	-1,1	3,9	2,3	12
2010	8,2	-5,1	-1,3	-1,7	16
2011	-0,6	2,9	-2,6	0,3	6
Medel	0,0	0,0	0,0	0,0	12
Std	5,5	4,6	3,5	1,7	5

	Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG
1999	3	-7	15	0
2000	44	-9	-25	-25
2001	-27	20	-13	1
2002	-27	2	11	39
2003	3	11	-11	-24
2004	-9	-9	22	2
2005	1	-14	12	27
2006	18	-5	-10	2
2007	-23	9	10	-10
2008	3	9	-11	-23
2009	-22	-2	16	29
2010	36	-12	-5	-22
2011	-3	7	-10	4
Medel(ABS)	17	9	13	16
Std	23	10	15	21

Av den vänstra tabelldelen framgår att den totala avvikelsen i genomsnitt är 12 procent. Uttryckt i *andel elever* betyder det att i genomsnitt per år får *6 procent* av eleverna ett felaktigt provbetyg på grund av instabila betygsgränser. För tidigare redovisade kursprov i svenska och engelska var de procentuella felen störst för IG och MVG. Så är det också för provbetygen i ma A, men samtidigt är felen i andelen med förväntat provbetyg G respektive VG betydligt större i matematik A än i svenska B och engelska A.

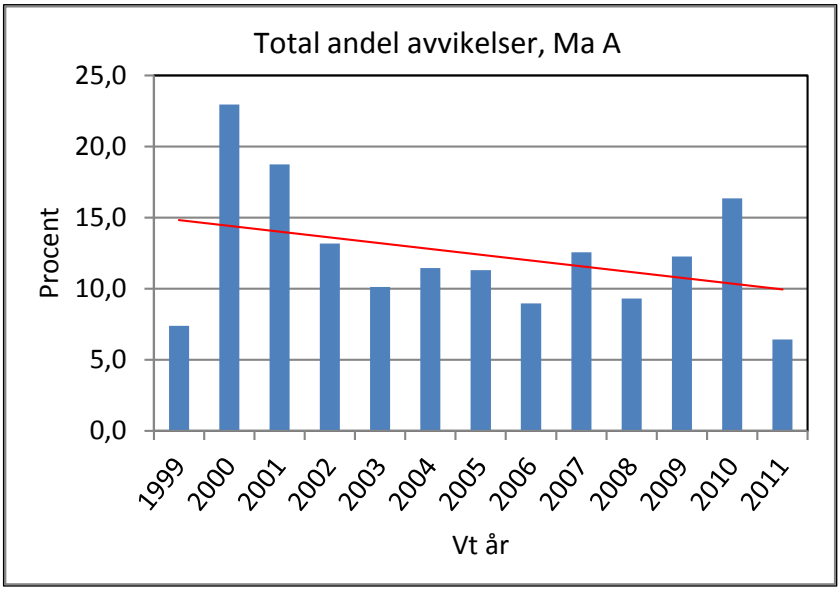
Av tabellens högra del framgår att i genomsnitt har den grupp som förväntas få provbetyget IG varit 17 procent för stor eller liten med en variation mellan att vara 44 procent för stor och 27 procent för liten. Bilden för MVG är nästan densamma och VG ser inte särskilt mycket bättre ut.

Några uppenbara förklaringar till de förhållandevis stora avvikelserna i relation till de andra kärnämnena är svåra att se. Stickproven är stora, i genomsnitt nästan 10 000 elevresultat (tabell 7) så det kan inte gärna handla om urvalsfel. Det handlar också om ett kärnämnesprov så det är i princip samma elever som deltagit i övriga

kärnämnesprov som deltagit i provet för ma A. Vi får återkomma till frågan i slutdiskussionen.

Trenden när det gäller avvikelsen för provet ma A är dock avtagande (figur 10). Det betyder alltså att stabiliteten är (var) ökande, men det sista året, 2011, var den totala avvikelsen ändå 10 procent (dvs. 5 procent av provdeltagarna fick ett annat provbetyg än det av modellen förväntade).

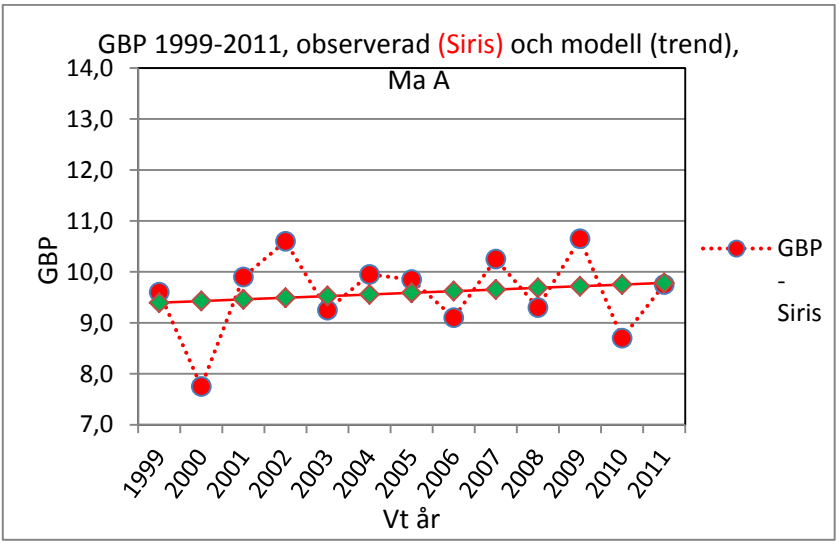
Figur 10 **Total andel avvikelser för matematik A. Andelen elever med avvikande betyg=hälften.**



Genomsnittlig betygspoäng

Bilden av den genomsnittliga betygspoängen under perioden 1999–2011 är i linje med tidigare redovisade resultat (figur 11).

Figur 11 Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (trendlinjen), matematik A.



Det är dock tveksamt om den svagt stigande trendlinjen ska tolkas som en resultatförbättring. Av tabell 8, den högra delen, framgår att förbättringen främst beror på att andelen IG minskat med cirka 2,5 procentenheter. Eftersom steget IG till G ger dubbel betygspoäng får detta också dubbel verkan på GBP.

Sammanfattning och kommentarer

De redovisade resultaten visar att andelen elever som får ett visst provbetyg varierar mellan olika år. Med det mål- och kunskapsrelaterade betygssystem som gäller är inte andelen elever som ska ha ett visst provbetyg bestämd i förväg. Ökade kunskaper leder till att en större andel av eleverna får högre betyg, vilket förstås i sin tur leder till en minskande andel elever med lägre betyg. Om kunskaperna å andra sidan sjunker händer det omvända. Sådana kunskapsförändringar sker långsamt när det gäller stora populationer av det slag som deltar i nationella prov, i synnerhet gäller detta för de nationella prov som genomförts av alla elever, dvs. sv B, en A och

ma A. I den aktuella studien fångas sådana långsiktiga förändringar upp av de använda trendlinjerna.

De mer kortsiktiga förändringar som kan noteras i provbetygen mellan närliggande år är därmed inte främst en följd av den långsiktiga utvecklingen utan antas bero på att provet är ett mätinstrument som i likhet med andra mätinstrument är behäftat med olika fel. I de nu redovisade resultaten manifesteras dessa i form av årliga svängningar i andelen elever som får ett visst provbetyg. Dessa svängningar kan ses som uttryck för ett mätfel hos det använda instrumentet. Det är detta systematiska mätfel som vi kallar *kravgränsfel*. Felet är systematiskt eftersom det är en egenskap hos provet (till vilket föreskrifterna om hur provbetyget fastställs räknas) och inte hos provdeltagarna.

Ett grundläggande antagande är således att kunskaperna hos den elevgrupp som genomför proven inte ändras nämnvärt mellan närliggande år utöver den systematiska och långsiktiga förändring som fångas upp av trendlinjen. De årliga fluktuationerna i relation till tilltrendlinjen ses därmed som ett uttryck för provets kravgränsfel. En väsentlig orsak till denna typ av ”klassificeringsfel” är att betygsgränserna eller betygskraven ligger för högt (eller för lågt) i relation till det läge de borde haft för att andelen skulle sammanfalla med trendlinjens värde.[27] För ett i huvudsak poängbaserat prov betyder det att poänggränsen för ett betyg ligger för högt eller för lågt.

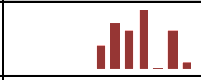
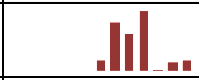
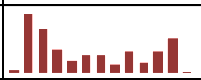
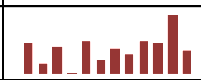
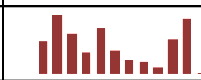
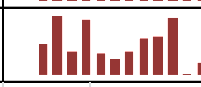
För ett prov som mer baseras på essäliknande uppgifter med holistisk bedömning kan bedömningsunderlagen vara mer svårtolkade och i högre grad bygga på den bedömande lärarens egen tolkning medan de poäng- och itembaserade[28] proven i allmänhet har mer precisa bedömningsanvisningar med mindre tolkningsutrymme för den som bedömer provet. Denna skillnad kan ha betydelse för de olika resultatbilder vi ser mellan ämnena i denna studie.

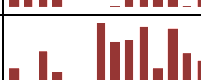
[27] Vi kan alltså inte bestämma något fel i relation till kunskapskravens formuleringar eftersom dessa inte är konstruerade för sådana jämförelser.
[28] Med *item* avses mindre provuppgifter kopplade till någon specifik information eller något visst tema. De bedöms med poäng, oftast 0 eller 1. Det kan t.ex. gälla ett hörförståelseprov där frågor ska besvaras om innehållet i det hörda eller läsförståelseprov där frågor baserade på den lästa texten ska besvaras.

Tabell 10 visar andelen elever med från trenden avvikande betyg för de aktuella populationerna (stickproven).[29] Resultat från årskurs 9, som fr.o.m. 2003 bygger på hela provpopulationen, finns med som jämförelse. Man kan notera att för samtliga prov i svenska och engelska samt för matematik i årskurs 9 varierar andelen med från trenden avvikande provbetyg med mellan cirka 2 och 3 procent. För proven i årskurs 9 innebär det över den aktuella perioden att i genomsnitt mellan 2 000 och 3 000 elever per år får fel provbetyg. För gymnasieskolans del är det svårare att översätta procentandelarna till antal eftersom det inte finns statistik över hur många elever totalt som gjort respektive prov.

[29] Även här är andelen uttryckt i konkreta elever hälften av andelen total avvikelse.

Tabell 10 Genomsnittlig andel elever med avvikande provbetyg för olika prov samt avvikelsens årliga variation och olika statistiska parametrar (max, min etc.). Trend anger om avvikelsen ökat eller minskat under perioden. Minustecken anger att den minskat, dvs. att stabiliteten i provbetygen ökat. Positiva värden anger omvänt att stabiliteten minskat.

Prov	Andel elever med fel provbetyg (%)					Trend
	Diagram 1999-2013	Max	Min	Medel	Stdav	%/år
Sv B 2005-2011		3,4	0,9	2,2	0,9	-0,15
Eng A 2005-2011		4,4	1,0	2,4	1,3	-0,23
Eng B 2005-2012		4,3	1,0	2,2	1,0	-0,20
Ma A 1999-2011		11,5	3,2	6,2	2,3	-0,20
Ma B 2000-2011		8,7	1,7	4,7	1,7	0,21
Ma C 2001-2012		8,2	2,0	5,0	2,0	-0,23
Ma D 2001-2013		10,5	1,1	6,1	2,9	-0,23

Prov Åk 9	Andel elever med fel provbetyg (%)					Trend
	Diagram 1998-2012	Max	Min	Medel	Stdav	%/år
Svenska		7,4	1,3	2,9	1,6	0,13
Engelska		3,4	0,8	1,9	0,9	-0,02
Matematik		5,6	0,6	3,2	1,6	0,11

Av tabellen framgår att det främst är matematikproven i gymnasieskolan som har stora totala avvikelser i relation till övriga prov, mellan knappt 5 till drygt 6 procent. För proven i svenska och engelska är de totala avvikelserna ungefär hälften så stora.

Kolumnen längst till höger i tabellen, ”Trend”, anger avvikelsernas utveckling över tid. Ett positiv värde t.ex. 0,11 (röd) anger att den absoluta avvikelsen i genomsnitt ökat med 0,11 procentenheter per år, dvs. provbetygens stabilitet har *minskat* i denna utsträckning. Ett negativt värde (grönt) uttrycker omvänt att provbetygens stabilitet ökat. Av tabell 10 framgår således att stabiliteten i provbetyg har ökat över tid för sex av de sju gymnasieproven, det vill säga andelen elever som får fel provbetyg har i genomsnitt minskat med cirka 0,2 procent per år. Variationen mellan olika år är dock betydande som diagrammen visar.

Hur tillförlitliga är skattningarna av kravgränsfelen?

Som tidigare nämnts har Skolverket (och Statistiska centralbyrån) fram till 2011 endast samlat in ett stickprov av provresultat från gymnasieskolan. Insamlingarna har gjorts för vårterminens prov och urvalet har gjorts så att alla skolor har blivit valda en gång under en sexårsperiod. Det betyder att urvalet inte är obundet slumpmässigt utan att det kan finnas vissa klustereffekter. Från dessa bortses i det här sammanhanget.

Eftersom studien bygger på stickprov finns urvalsfel i underlaget. Dessa kan tämligen enkelt skattas eftersom resultaten i Siris är givna som proportioner.[30]

En annan begränsning ligger i att Siris endast redovisar resultaten i hela procent vilket gör att det finns ett avrundningsfel på maximalt ±0,5 procentenheter. Det felet kan också tämligen enkelt skattas.[31]

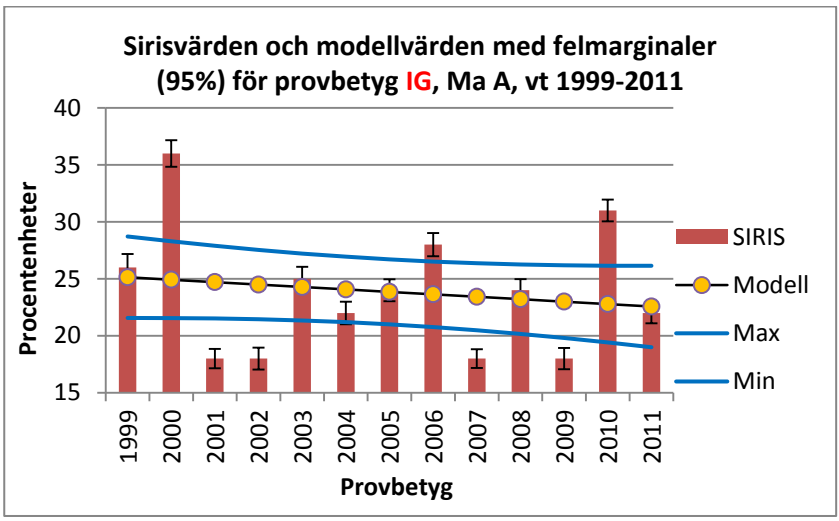
Några exempel på urvalsfelens (beroende av stickprovets storlek) och avrundningsfelens betydelse redovisas nedan.

I diagrammen i figur 12 (nedan) anger de blå linjerna konfidensintervallet (95 procent) för trendlinjen. Staplarna representerar de värden Siris anger med felmarginaler angivna (95 procent). Cumming (2012) anger att man som tumregel kan säga att om stapeln (med felmarginaler) överlappar de blå linjerna till mindre

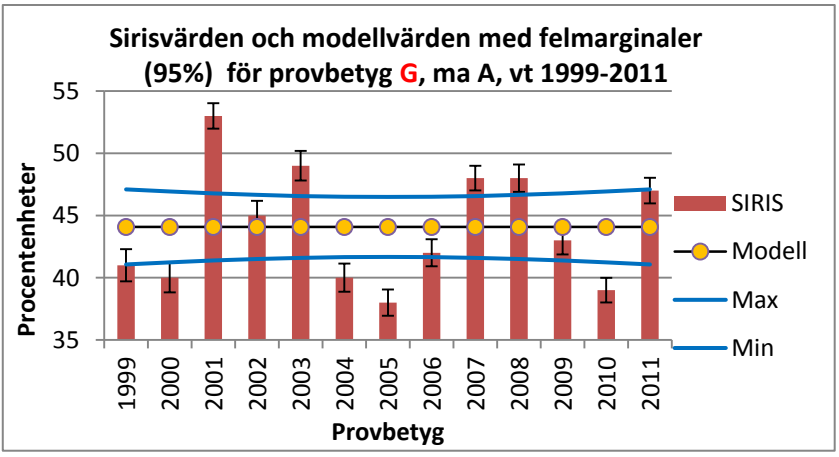
[30] Det ger standardfelet $s_e = \sqrt{p * (1-p)/n}$ där p är proportionen (procentandelen) som har betyget och n är stickprovets storlek (anges i tabell 1 och motsvarande för övriga prov).
[31] Standardfelet för en uniform fördelning är $s_e = \sqrt{((b-a)\hat{}2/12)}$ där a är felets undre gräns (-0,5 i det här fallet) och b dess övre (0,5) vilket ger standardfelet $s_e = 0{,}29$ procentenheter vid avrundning till hela procent.

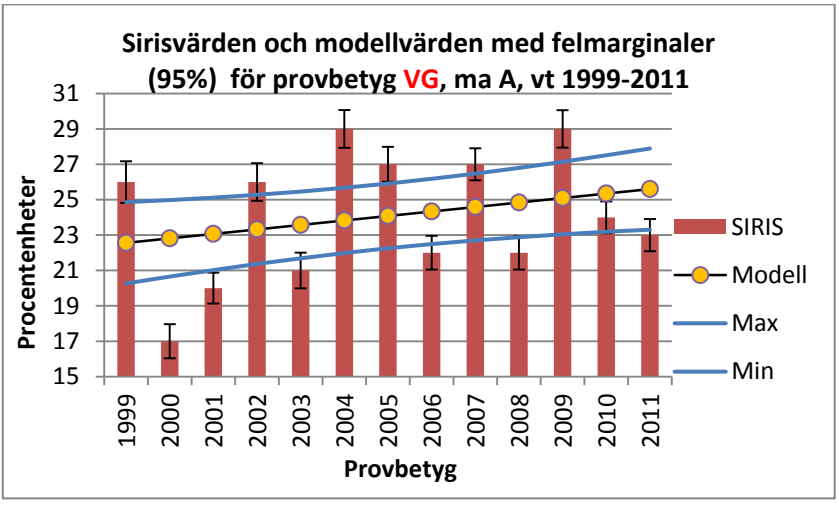
än 50 procent kan avvikelsen betraktas som statistisk signifikant (95 procent).[32] Därmed kan avvikelserna för provbetyget IG anses statistisk signifikanta vårterminerna 2000, 2001, 2002, 2006, 2007, 2009 och 2010 medan avvikelserna återstående vårterminer är mer osäkra. Detta resultat stämmer väl med bilden av andelen elever med provbetyget IG i matematik A i figur 9. På motsvarande sätt kan bilderna för övriga provbetyg bedömas. Den sammanfattande bedömningen blir att de aktuella slumpfelen inte på något avgörande sätt ändrar den övergripande bilden.

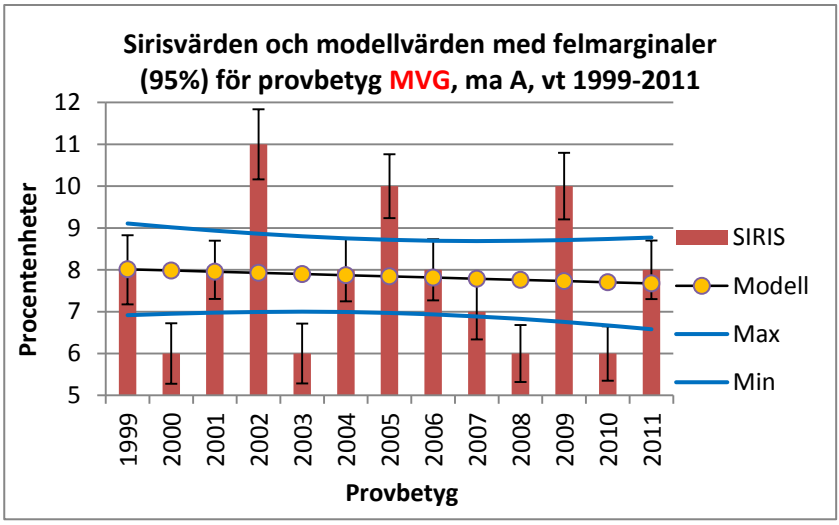
Figur 12 **Andel elever med olika redovisade provbetyg olika år för ma A med anpassade trendlinjer samt felmarginaler (95 procent). Om den redovisade andelen (Sirisvärdet) ligger utanför de blå linjerna kan avvikelsen mellan förväntad andel och redovisad andel anses signifikant.**



[32] Förutsatt att felen är oberoende. Detta gäller inte fullt ut här eftersom konfidenslinjen för trendlinjen innefattar varje års värde (stapel) på andelen med betyget. Det är dock inte mödan värt att konstruera en särskild trendlinje (utan aktuellt stapelvärde), inklusive konfidenslinjer, för varje stapelvärde och därefter jämföra överlappning. Tumregeln betraktas som approximativt giltig.


Sirisvärden och modellvärden med felmarginaler (95%) för provbetyg G, ma A, vt 1999-2011
Procentenheter
55
50
45
40
35
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Provbetyg
SIRIS
Modell
Max
Min


Sirisvärden och modellvärden med felmarginaler (95%) för provbetyg VG, ma A, vt 1999-2011
Procentenheter
31
29
27
25
23
21
19
17
15
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
Provbetyg
SIRIS
Modell
Max
Min



Av de fyra diagrammen framgår att för olika provbetyg i ma A är avvikelserna signifikanta för mellan cirka hälften och tre fjärdedelar av åren. Det innebär rimligen att de avvikelser som redovisats i tidigare tabeller och som sammanfattas i tabell 10 kan ses som goda indikatorer på de avvikelser som beror på det vi kallat kravgränsfel.

För svenska B och engelska A som redovisas i appendix 2 är signifikansnivån lägre i den meningen att avvikelsen mellan trendlinjen och Siris-värdet är signifikant (95 procent) för färre år än vad som var fallet för ma A (mellan 0 och 3 år för de sju år som ingår i serien). Den lägre andelen signifikanta skillnader beror främst på att färre år ingår i tidsserien för sv B och eng A (samt för eng B). Även om signifikansnivån är lägre är dock den bästa skattning som kan göras den som utgår från medelvärdena och som ligger till grund för de redovisade resultaten. Indikationerna på avvikelse blir med andra ord något mindre tillförlitliga för dessa prov men är de bästa som ges av befintliga data.

Slutsatser och diskussion

Syftet med detta avsnitt var att få ett mått på kravgränsfelets storlek uttryckt i andelen felklassificerade elever utifrån den långsiktiga trenden. Vill man veta hur stor andel av eleverna som skulle

behöva ändra provbetyg för att avvikelsen skulle närma sig noll får man ta hälften av antalet felklassificerade betyg.

Här har endast det sammanfattande provbetyget och dess variation över tid studerats. Någon hänsyn har inte tagits till hur det sammanfattande provbetyget konstruerats. Detta kan dock antas ha betydelse för hur mycket det slutliga provbetyget varierar över tid. Man kan notera att provbetygen i matematik har de största avvikelserna medan proven i engelska och svenska ligger på en väsentligt lägre nivå. Detta kan förefalla något överraskande och det är svårt att finna någon uppenbar förklaring. Man kan tänka sig åtminstone två faktorer som kan vara av betydelse. För det första provens utformning och sättet att konstruera provbetyg. För det andra provens bedömning.

Provens utformning

De olika proven konstrueras enligt olika principer och de sammanfattande provbetygen bestäms likaledes på olika sätt. Prov i svenska B består av delprov som prövar dels läsförståelse, dels skriftlig förmåga. Även muntlig förmåga prövas. Detsamma gäller proven i engelska, vilka till skillnad från proven i svenska även prövar förmågan att lyssna. Proven i läsförståelse och hörförståelse poängsätts medan proven i skriftlig förmåga (uppsatser) bedöms holistiskt, dvs. betygssätts i enlighet med angivna kriterier och exempel. För båda ämnena gäller att de olika delproven betygssätts var för sig varefter de olika provbetygen sammanvägs till ett sammanfattande provbetyg enligt någon för ämnet mer eller mindre specifik metod.

I korthet kan man säga att matematikprov i huvudsak baseras på uppgifter som kan ge en eller flera poäng. Dessutom används ofta olika typer av poäng (G-poäng, VG-poäng respektive någon form av MVG-poäng eller MVG-kvaliteter). Beroende på om de uppgifter en elev besvarat korrekt bedömts som G-, VG-, respektive MVG-uppgift har eleven tilldelats motsvarande typ av poäng. Vissa uppgifter kan bestå av olika deluppgifter som ger olika typer av poäng. De olika poängen summeras och betyg sätts utifrån olika villkor för respektive betyg. Ofta ska en viss totalpoäng ha uppnåtts för varje betyg. För ett högre provbetyg kan ytterligare gälla att en viss minsta summa av VG- och/eller MVG-poäng ska ha

erhållits och/eller att poängen måste erhållas på uppgifter som täcker in olika kunskapsområden.

Provens bedömning

I Sverige bedöms de nationella proven av undervisande lärare. Det betyder antingen att läraren helt själv bestämt provbetyget eller att det skett i större eller mindre samverkan med andra kollegor.

Kan provens utformning och bedömning förklara den större variationen för provbetygen i matematik?

Kan de två ovanstående faktorerna – provens utformning och provens bedömning – förklara den större variationen för provbetygen i matematik? Proven i matematik är nästan uteslutande baserade på uppgifter som poängsätts. Uppgifterna är i allmänhet avgränsade och kan ge ett begränsat antal poäng. De korrekta svaren är också i allmänhet tämligen avgränsade och det finns sällan något större utrymme för tolkningar. Har en elev gjort en felaktighet så är det oftast förhållandevis tydligt och om inte felet är trivialt[33] så innebär det i allmänhet någon form av poängavdrag även om läraren med stor säkerhet ”vet” att eleven har de kunskaper som behövs för att nå rätt svar. Matematikens krav på stringens skulle alltså kunna innebära att kravgränserna blir skarpa och att det därmed inte ges utrymme för en lärare som känner eleven att tolka in tidigare erfarenheter av elevens kunnande i sin bedömning. Samtidigt är dock inte tolkningsmöjligheter uteslutna. Om man granskar poängfördelningen för poängbaserade prov kan man ofta notera en viss förhöjd andel elever med poäng precis ovanför provets betygsgränser. Figur 1 och figur 15 kan tjäna som illustrationer.

För proven i engelska och svenska gäller motsvarande förhållande för de poäng- och itembaserade[34] delproven. Framför allt i engelska är de poängbaserade delarna omfattande. Dessutom finns i engelska och svenska ett delprov som består av en längre skriv-

[33] Till exempel ger 2+2=5 i ett gymnasieprov knappast något poängavdrag.
[34] Uppgifterna kan t.ex. utgöras av en text som ska läsas varefter att antal frågor (*items*) kopplade till texten ska besvaras.

uppgift (en uppsats). Detta delprov bedöms inte med poäng utan som tidigare nämnts holistiskt. Prov där en helhetsbedömning ska göras innefattar ofta ett större inslag av tolkning av den som bedömer provet än om bedömningen görs med poäng. Då kan det vara svårt för bedömaren att frigöra sig från tidigare uppfattningar om eleven. Det kan handla om den s.k. haloeffekten[35] eller om egna erfarenheter av elevens tidigare prestationer i ämnet. Denna tendens att tolka positivt (eller negativt) behöver inte vara en medveten process. Det förefaller dock som om det är vanligare att denna tendens ger utslag i positiv än negativ riktning. Detta skulle kunna vara en möjlig förklaring till att skillnaden mellan provbetyget och slutbetyget i genomsnitt är mindre i svenska och engelska än i matematik.

Urvalsfel och avrundningsfel

Eftersom analysen av provbetygen för gymnasieskolans del baseras på stickprov finns mer eller mindre stora urvalsfel beroende på stickprovens storlek och den slumpmässighet med vilken de valts. Dessa påverkar tillförlitligheten i analyserna. Data är dessutom angivna i hela procentsatser, vilket också innebär en begränsning i noggrannheten. De försök till skattningar av de olika felens storlek tillsammans med jämförelser med motsvarande resultat för grundskolan (vilka baseras på hela populationer) tyder dock på att analyserna av kravgränsfelens storlek ger rimliga resultat, framför allt för matematikämnet som har långa tidsserier jämfört med tidsserierna för svenska och engelska.

Utöver ovanstående fel kan man diskutera den linjära trendlinjens giltighet över långa tidsintervall. Betyg med stigande trender skulle stiga mot 100 procent och betyg med fallande trender skulle gå mot noll procent. En sådan utveckling förefaller förstås inte trolig. Den linjära modellen förefaller dock fungera väl för den aktuella perioden. Men om betygsskalan skulle bli bestående under en längre tid och börja visa dålig anpassning till en linjär trend kan man tänka sig att låta trendlinjen vara ”glidande” genom att t.ex. endast innefatta de senaste 10–15 åren.

[35] Vetskap om att eleven är duktig (eller inte duktig) i andra sammanhang.

Sammanfattande resultat för kravgränsfelet

De erhållna resultaten (se tabell 10) tyder på att kravgränsfelens storlek innebär att i genomsnitt utgör andelen felklassificeringar cirka 4 till 5 procent för den aktuella perioden. Det betyder att 2 till 2,5 procent av eleverna är i fel betygsgrupp enligt den långsiktiga trenden (och skulle behöva flyttas från betygsgrupper med överskott av elever till betygsgrupper med motsvarande underskott av elever för att anpassning till trendlinjen skulle ske). Andelen elever som fått ett från trendlinjen avvikande betyg i matematik är ungefär dubbelt så stor, cirka 4,5 till 6 procent. Som tabell 10 också visar är variationen betydande. En viss minskning av kravgränsfelens storlek över tid (cirka 0,5 procentenheter per år i genomsnitt) kan dock noteras för provbetygen i flertalet ämnen. Om avvikelser av den rådande storleksordningen är rimliga och acceptabla är en fråga som behöver diskuteras.

Kravgränserna i form av poänggränser eller andra anvisningar för olika provbetyg bestäms av de lärosäten som konstruerar proven i samråd med Skolverket. De avvikelser som beror på kravgränsernas läge och övriga bestämningar är således i huvudsak ett resultat av provkonstruktörernas bedömningar.

Det sammanfattande empiriska resultatet är det som visas i tabell 10 när det gäller stabiliteten i provbetygen uttryckt i den andel elever som vid vårterminens nationella prov fått provbetyg som avviker från det värde som anges av den långsiktiga trenden. Detta är de avvikelser som framträder statistiskt i tabeller och grafiskt i diagram, en sorts synliga nettoavvikelser. Bakom dessa avvikelser finns emellertid andra fel som inte framträder vid analyser av det slag som gjorts hittills. Dessa fel finns på individnivå och tenderar att ta ut varandra om man jämför provresultat för olika grupper. Det beror på att det förutom systematiska avvikelser kopplade till kravgränser för olika provbetyg också finns slumpmässiga faktorer som påverkar enskilda elevers provresultat och därmed även deras provbetyg. Sådana osynliga slumpbaserade fel tas upp i nästa avsnitt.

Slumpbaserade fel

Inledning

Ett nationellt prov baseras oftast på många olika uppgifter. När ett sådant prov konstrueras måste ett urval av uppgifter göras. Till varje lite mer omfattande kunskapsområde kan mängder av olika uppgifter konstrueras. De kan konstrueras så att de täcker olika delar av kunskapsområdet och de kan vara av olika svårighetsgrad. Det betyder att de uppgifter som väljs att ingå i provet utgör ett stickprov av många för provet tänkbara uppgifter.

Uppgifterna väljs efter innehåll för att så väl som möjligt täcka det kunskapsområde som ämnesplanen föreskriver. Detta urval sker i samverkan med ämnesexperter och aktiva lärare, vilket är nödvändigt för att säkra provets relevans eller validitet. Uppgifterna väljs också efter svårighetsgrad och även där medverkar ämnesexperter och lärare, men även personer med kunskaper i pedagogisk mätningslära. Svårighetgrad kan förstås bedömas erfarenhetsmässigt, men för att få mer precisa mått behövs även empiriska underlag, vilka erhålls via olika utprövningar. [36]

Att varje prov är ett stickprov av uppgifter i relation till alla uppgifter som kan konstrueras för kunskapsområdet innebär att det finns ett urvalsfel. Om en annan uppsättning av uppgifter från samma kunskapsområden och av samma svårighetgrad hade valts visar erfarenheten att resultaten på de två proven kommer att variera i större eller mindre grad för enskilda elever. Detta beror på slumpen. Endast den som inte kan lösa någon uppgift eller alla uppgifter kommer att få samma resultat. För övriga kommer antalet uppgifter man kan lösa på respektive prov i allmänhet att variera. Man kan en viss (okänd) andel av alla tänkbara uppgifter. Och det är den okända andelen som provresultatet ska försöka fånga. Med lite tur råkar provet innehålla en större andel uppgifter som provdeltagaren klarar. Med lite otur en mindre andel. Utmaningen för den som ska värdera provresultatet för en elev är att bedöma vilken betydelse slumpen har. Att göra denna bedömning på ett systematiskt sätt är dock inte helt enkelt.

[36] Se t.ex. www.skolverket.se/bedomning/nationella-prov/hur-konstrueras-de-nationella-proven www.su.se/primgruppen/matematik

Graden av slumpinflytande på provresultatet är det som brukar anges som provets reliabilitet. Ambitionen är givetvis att få så hög reliabilitet som möjligt. Den mäts i allmänhet som ett tal mellan 0 och 1 där 0 betyder att resultaten är helt slumpmässiga och 1 att de helt saknar slumpinflytande. Inga prov når upp till 1 men värden kring 0,90 brukar anses bra för prov av typen nationella prov. Reliabiliteten är dock ett mått på ett visst provs tillförlitlighet för en viss grupp och är således inte en bestämd egenskap hos provet. Reliabilitetsmåttet kan också vara svårt att tolka i relation till en enskild elevs resultat och i nästa avsnitt kommer den frågan att tas upp. Det skulle föra för långt i det här sammanhanget att ytterligare fördjupa sig i resonemang om reliabilitetens och validitetens roll i samband med pedagogiska mätningar.[37] Vi utgår i stället från ett empiriskt exempel.

Ett empiriskt exempel

Det här avsnittet baseras på de metoder som finns närmare beskrivna i Skolverket (2015b). Metoderna upprepas inte här utan endast resultaten av deras tillämpning återges. Syftet är att ge en bild av vilken betydelse slumpfelet har för bedömningen av hur tillförlitlig den enskilda elevens provpoäng är och därmed också det provbetyg eleven får. I det här sammanhanget har inte kravgränsernas placering någon betydelse eftersom ingen jämförelse görs mellan olika år, utan här handlar det enbart om ett enskilt prov och de slumpfel som gäller för ett sådant.

Dataunderlag

De metoder som tillämpas här för att bestämma de parametrar eller storheter som behövs för att skatta slumpens betydelse för enskilda elevers resultat gäller främst för uppgiftsbaserade prov. Ett sådant kommer därför att användas som exempel. Även för prov med holistiska bedömningar kan reliabilitet (i form av bedömaröverens-

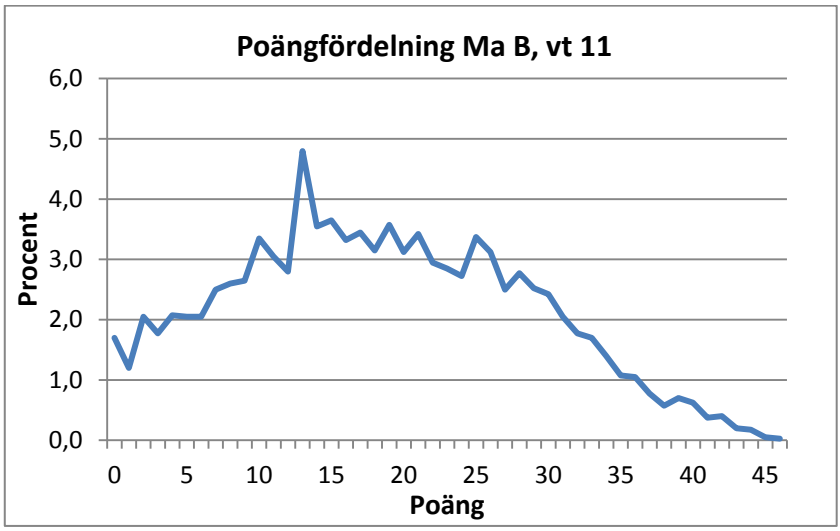
[37] Detta beskrivs i grundläggande litteratur om pedagogiska mätningar t.ex. Crocker & Algina (1986).

stämmelse) skattas och ligga till grund för bedömningar av mätfel, men denna typ av fel får anstå till nästa avsnitt.

Underlaget här utgörs av data för provet ma B vårterminen 2011. De har insamlats av Umeå universitet som även ansvarat för konstruktionen av provet.[38] Syftet med analysen är som nämnts att utifrån detta underlag ge en representativ bild av slumpfelens storlek och konsekvenser för poängbaserade prov. Tillvägagångssättet är detsamma som för motsvarande rapport för grundskolan.[39]

Poängfördelningen för det aktuella provet visas nedan i figur 13.

Figur 13 **Poängfördelning för stickprovet som genomfört provet i matematik B vt 2011.**



Uppgifterna bedöms med två typer av poäng som representerar kunskaper i enlighet med olika betygskriterier – G-poäng och VG-poäng, samt för uppgifter markerade med ”¤”, med ett antal MVG-kvaliteter.[40] Kravgränserna för olika provbetyg definieras i enlighet med nedanstående anvisning.

[38] Se www5.edusci.umu.se/np/np-prov/B-kursprov-vt11.pdf och www.edusci.umu.se/np/np-2-4/resultat/
[39] Skolverket (2015b).
[40] För närmare beskrivning se www5.edusci.umu.se/np/np-prov/B-kursprov-vt11.pdf

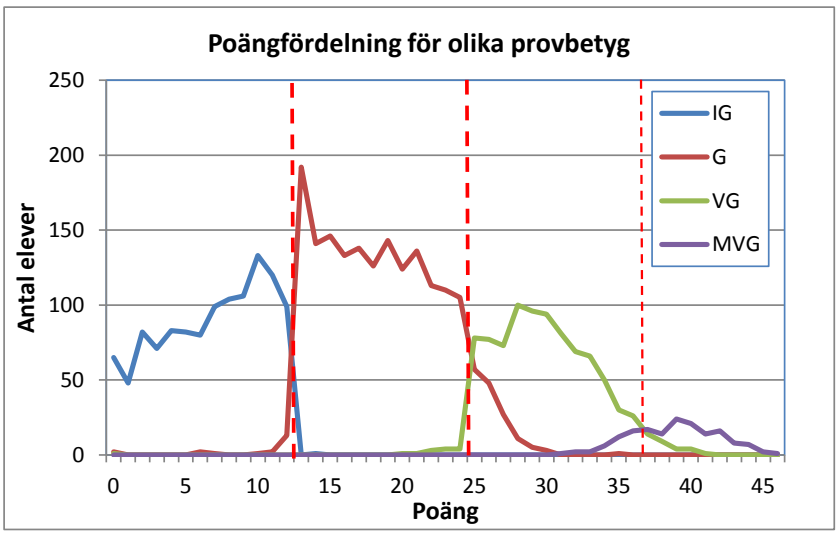
Kravgränser

Detta prov kan ge maximalt 45 poäng, varav 25 g-poäng.

Undre gräns för provbetyget	
Godkänt:	13 poäng.
Väl godkänt:	25 poäng varav minst 6 vg-poäng.
Mycket väl godkänt:	25 poäng varav minst 13 vg-poäng. Eleven ska dessutom ha visat prov på minst tre *olika* MVG-kvaliteter av de fyra MVG-kvaliteter som är möjliga att visa i detta prov.

Om man sammanställer den totala poängfördelningen för elever som fått olika provbetyg fås bilden i figur 14. Dessa poänggränser används i den fortsatta analysen.

Figur 14 **Poängfördelningar för elever i stickprovet matematik B vt 2011 med olika provbetyg.**



Man kan notera att för de lägre betygen IG och G stämmer poängfördelningarna väl med anvisningarna för kravgränserna (förhållandevis liten överlappning vid betygsgränsen). För betyget VG och i synnerhet MVG är det svårare att bedöma överensstämmelsen utan mer ingående analyser (större överlappning vid betygsgräns-

erna). Som framgår av figuren finns det en betydande överlappning och villkoren om VG-poäng och MVG-kvaliteter skapar uppenbarligen vissa omkastningar i provbetygen. För de analyser som görs här väljer vi dock för enkelhets skull att ange gränsen för MVG vid skärningen mellan kurvorna för VG och MVG.[41]

Den enskilda mätningens standardfel (SEM)[42]

Några grundläggande parametrar utifrån urvalet visas i tabell 11.[43]

Tabell 11 Värden för olika parameterskattningar, ma B vt 2011.

Parameter	Värde
Antal elever	4 004
Medelvärde	17,13
Standardavvikelse	10,74
Reliabilitet (alfa)	0,91
SEM (Standard error of measurement)	3,24

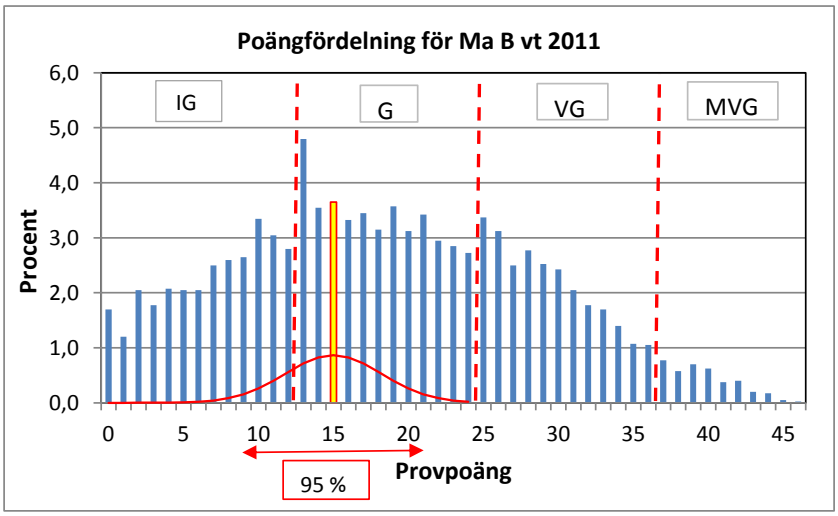
Medelvärde och standardavvikelse bestäms med hjälp av statistikprogrammet SPSS. Det gäller även reliabilitetsmåttet *coefficient alfa* som är ett av flera tänkbara mått, men det som oftast används. SEM är det genomsnittsvärde som oftast används och som enkelt beräknas.[44]

Figur 15 visar poängfördelningen och aktuella betygsgränser. Betygsgränserna är inlagda i figuren. Vidare är poängen för en elev med 15 poäng markerad (gul stapel). Den observerade poängen är alltid behäftad med ett (okänt) slumpfel. Det är dock möjligt att utifrån vissa antaganden skatta ett intervall inom vilket elevens ”sanna” poäng[45] kan antas ligga med viss sannolikhet. Denna sannolikhet fördelar sig som kurvan runt foten av den gula stapeln i

[41] Detta är i enlighet med en av de metoder som används för att bestämma kravgränser för olika betyg (*contrasting groups*-metoden).
[42] SEM = *standard error of measurement*.
[43] En närmare beskrivning av data för det aktuella provet ges i appendix 3.
[44] Se Skolverket (2015b).
[45] Med ”sann poäng” menas i det här sammanhanget den medelpoäng en elev skulle få om hon eller han kunde göra ett stort antal likvärdiga prov. Detta är av olika skäl inte möjligt, men det finns metoder för att skatta ett värde på den sanna poängen. Se t.ex. Skolverket (2015b) eller Crocker & Algina (1986).

det aktuella fallet. Standardavvikelsen för denna normalfördelade skattning kallas *det enskilda felets standardavvikelse* (*standard error of measurement* förkortat SEM på engelska)[46]. Enligt tabell 11 är SEM för det aktuella provet (och gruppen) 3,24 poäng. Ett 95 procentigt konfidensintervall för den sanna poängen blir då 15±1,96*3,24 poäng, dvs. den bör ligga mellan 9 och 21 poäng.

Figur 15 **Poängfördelning och betygsgränser för provet i ma B, vt 2011. Exempel på poäng (15) med fördelning och konfidensintervall för 95 procentigt konfidensintervall för den sanna poängens läge.**



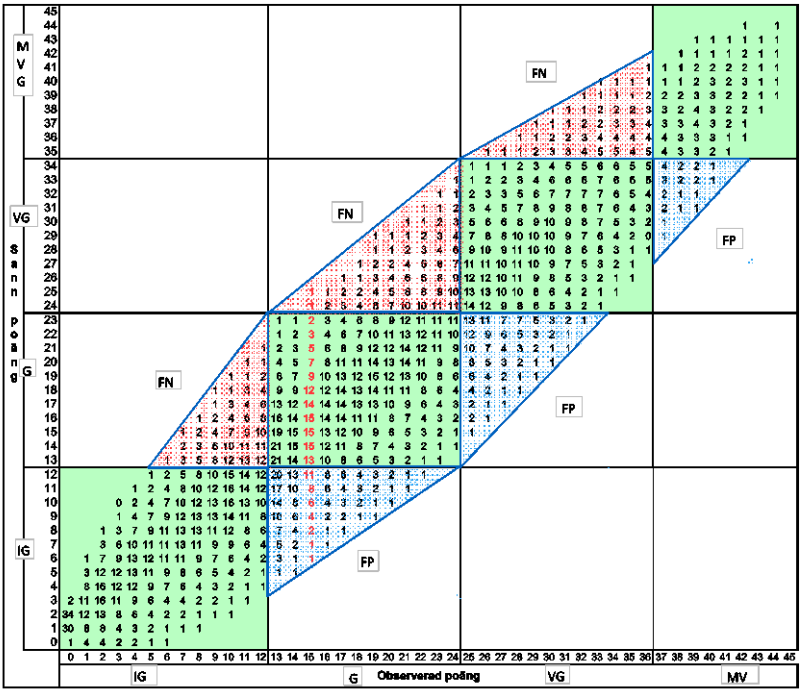
Standardfelet i det aktuella fallet visar således att det finns en icke oväsentlig sannolikhet att en elev med 15 poäng på provet och betyget G, har en sann poäng under gränsen för provbetyget G. Utifrån givna data kan man beräkna hur stor andel av eleverna med 15 poäng som kan antas ha en poäng lägre än gränsen för G. Men man kan inte utifrån provresultatet peka ut *vilka* dessa elever är. Man vet således att de finns men inte vilka de är.[47] Motsvarande resonemang gäller för alla observerade provpoäng.

[46] Se Skolverket (2015b) för en närmare beskrivning.
[47] Utifrån de antaganden som den klassiska testteorin bygger på i det här fallet. Man kan också använda modern testteori, vilken dock ger likartade skattningar.

Classification accuracy[48]

Av figur 15 och ovanstående resonemang torde det framgå att för elever med observerade poäng i närheten av en betygsgräns finns det betydande sannolikhet att eleven antingen haft tur och hamnat över den aktuella betygsgränsen, eller otur och hamnat under den. Låt oss ytterligare illustrera detta med aktuella data från provet i matematik B, vårterminen 2011. Utan att närmare gå in på tillvägagångssättet blir resultatet av en sammanställning det som visas i figur 16.[49]

Figur 16 **Skattad fördelning av antal elever med viss sann poäng (vertikal skala) för olika observerad poäng (horisontell skala). Baserad på den aktuella fördelningen av observerad poäng i stickprovet för ma B, vt 2011. FN = falskt negativa provbetyg, FP = falskt positiva provbetyg. Grön ruta = korrekt provbetyg.**



[48] Det finns ingen vedertagen svensk benämning på uttrycket. En möjlig benämning kunde vara klassifikationsriktighet. Observera dock att det är en annan form av klassificering än den som gällde i avsnittet *Kravgränsfel*.
[49] Jämför Skolverket (2015b).

Figuren visar på den vågräta axeln observerad provpoäng och på den lodräta motsvarande sann poäng.[50] De vertikala och horisontella linjerna markerar gränserna för respektive betyg på de två skalorna.[51]

De elever som ligger i de gröna diagonala rutorna har samma betyg på den observerade och den sanna skalan. De kan därmed antas ha fått korrekta provbetyg. Vilka eleverna är vet vi dock inte enligt tidigare resonemang, endast hur många de är i den aktuella gruppen.

De elever som ligger i intervallet för IG på skalan för observerad poäng, men i intervallet för G enligt den sanna skalan representerar de elever som haft otur på provet. De har en sann poäng som räcker till G, men genom otur med valet av uppgifter i provet, dålig dagsform eller vad det kan vara, har de inte nått sin sanna nivå. De råkar tillhöra det som brukar kallas gruppen ”falskt negativa” (FN, se figur 16).

På motsvarande sätt finns det en grupp elever som haft tur med valet av uppgifter och som därmed lyckas prestera över sin sanna nivå på det aktuella provet.[52] Denna grupp har fått högre betyg än förväntat och kallas därför ”falskt positiva” (FP, se figur 16). De har t.ex. en sann poäng som motsvarar IG, men har exempelvis genom tur med valet av uppgifter i provet fått en observerad poäng som räcker till G.

För varje observerat betygssteg finns en eller två grupper provdeltagare med falskt positiva eller falskt negativa betyg. Dessa provdeltagare kan alltså antas ha fått oriktiga provbetyg.

[50] Den sanna poängen är konstruerad utifrån regression mot medelvärdet med användning av Kelleys formel (se Skolverket (2015b) för närmare beskrivning). Fördelningen av elever med viss observerad poäng i y-led (sann poäng) baseras på CSEM (C står för *conditional*, vilket innebär att SEM:s värde varierar med den observerade poängen, se Skolverket (2015b)).

[51] Notera att för det tidigare exemplet med 15 observerade poäng är i figur 16 fördelningen av den sanna poängen vertikal (röda siffror). Man kan notera att den elev som har 15 poäng på provet lika gärna kan ha en sann poäng på 14 eller 16 poäng. Ett par elever med 15 poäng (och provbetyg G) har sannolikt 24 eller 25 i sann poäng och har alltså för lågt provbetyg. De borde ha VG och är alltså falskt negativa (FN). Betydligt fler elever med 15 poäng har sann poäng i IG-intervallet samt har för högt provbetyg och är falskt positiva (FP).

[52] Det aktuella provet ses alltså bara som ett (stick)prov av alla tänkbara prov som skulle kunna konstrueras utifrån det aktuella kunskapsområdet och de aktuella förmågorna. Den sanna poängen är en skattning av vad eleven kan prestera för hela kunskapsområdet och alla förmågor.

Räknar man samman antalet elever med korrekta betyg, falskt negativa och falskt positiva betyg och uttrycker dem i procent fås följande tabeller:

Tabell 12 **Andel elever med korrekta (diagonalen), falskt positiva (ovanför diagonalen) och falskt negativa (under diagonalen) betyg.**

		Observerad poäng (procent)			
	Betyg	IG	G	VG	MVG
Sann poäng	IG	26	5	0	0
	G	4	32	2	0
	VG	0	7	18	0
	MVG	0	0	3	3

Tabell 13 **Sammanfattning av andelen elever med korrekta respektive felaktiga (falskt negativa eller falskt positiva) provbetyg på provet ma B, vt 2011.**

Kategori	Procent
Korrekta betyg	79
FN	13
FP	7
Totalt	**100**

På grund av avrundningsfel blir summan mindre än 100.

Tabell 13 visar att för det aktuella provet i matematik har 79 procent av eleverna fått ett korrekt provbetyg medan 21 procent fått ett för högt eller för lågt provbetyg.[53]

Den viktiga frågan i sammanhanget är hur siffrorna i tabell 13 ska tolkas. Det bör observeras att ovanstående fel (falskt positiva och falskt negativa) till ganska liten del är beroende av var betygsgränserna har lagts. Om man flyttar en betygsgräns sker en ökning för exempelvis gruppen falskt positiva, men samtidigt minskar gruppen falskt negativa ungefär lika mycket (se figur 16). Det felet framför allt är ett uttryck för är provets mätegenskaper baserade på

[53] I engelskspråkig litteratur innebär det att provet har en *classification accuracy* (eller klassifikationsriktighet) på 79 procent.

de parametrar som anges i tabell 11 och då främst på provets standardavvikelse och reliabilitet.[54] De tidigare beskrivna kravgränsfelen och ovanstående typ av slumpfel är alltså i huvudsak två av varandra oberoende fel.

Jämförelse med ett engelskt exempel

Om man tycker det låter mycket att 20 procent av eleverna får fel provbetyg kan man jämföra med resultat på andra liknande prov. I en engelsk rapport[55] redovisas följande värden för *accuracy*, *false positive*, *false negative* m.m. i ett antal ämnen från den engelska motsvarigheten till gymnasieskolans nationella prov.[56]

[54] Dessa är för övrigt ingen generell egenskap hos provet utan en egenskap för provet givet till den aktuella gruppen. Om den aktuella gruppen kan anses representativ för hela den population som gjort det aktuella provet bör dock de erhållna parametrarna vara goda approximationer för de värden som gäller för hela populationen.
[55] Wheadon & Stockford (2010).
[56] GCSE och A-level examinations.

Tabell 14 ***Accuracy*** **(andel korrekt betygssatta elever) samt andel falskt positivt och falskt negativt betygssatta elever för ett antal engelska examensprov (A-level). IRT- baserade[57] resultat i den mittersta kolumnen, resultat baserade på Livingston & Lewis klassiska metod i den högra.[58]**

Specification	Unit	Accuracy	False Positive	False Negative	Accuracy	False Positive	False Negative
ACCOUNTING	ACCN1	0.61	0.17	0.22	0.53	0.22	0.25
ACCOUNTING	ACCN2	0.62	0.17	0.21	0.56	0.22	0.22
BIOLOGY	BIOL1	0.60	0.20	0.20	0.58	0.21	0.21
BIOLOGY	BIOL2	0.63	0.19	0.18	0.61	0.18	0.21
CHEMISTRY	CHEM1	0.67	0.16	0.17	0.64	0.19	0.17
CHEMISTRY	CHEM2	0.73	0.14	0.14	0.69	0.16	0.15
COMPUTING	COMP2	0.60	0.20	0.20	0.54	0.23	0.22
ELECTRONICS	ELEC1	0.67	0.17	0.16	0.62	0.21	0.17
ELECTRONICS	ELEC2	0.70	0.15	0.15	0.65	0.19	0.16
ENVIORONMENTAL STUDIES	ENVS1	0.57	0.20	0.23	0.54	0.22	0.25
ENVIORNMENTAL STUDIES	ENVS2	0.64	0.17	0.19	0.61	0.20	0.19
HUMAN BIOLOGY	HBIO1	0.67	0.16	0.18	0.65	0.16	0.19
HUMAN BIOLOGY	HBIO2	0.59	0.19	0.22	0.58	0.19	0.23
PHYSICS	PHYA1	0.67	0.17	0.16	0.64	0.19	0.17
PHYSICS A	PHYA2	0.68	0.16	0.16	0.66	0.18	0.16
PHYSICS B	PHYB2	0.64	0.18	0.18	0.62	0.20	0.17
PSYCHOLOGY A	PSYA1	0.60	0.19	0.21	0.56	0.23	0.21
PSYCHOLOGY A	PSYA2	0.60	0.19	0.20	0.57	0.23	0.20
PSYCHOLOGY B	PSYB1	0.55	0.22	0.23	0.54	0.23	0.23
SCIENCE IN THE SOCIETY	SCIS1	0.58	0.21	0.22	0.56	0.22	0.22

Som framgår vid en jämförelse med tabell 14 förefaller det aktuella svenska exemplet ha förhållandevis hög *klassifikationsriktighet*. Jämförelser med andra utländska exempel ger ungefär samma bild så detta är den osäkerhet man får leva med när det gäller examensprov av gängse typ. Dock måste man vara försiktig vid sådana jämförelser eftersom andelen falskt positiva och falskt negativa ökar i

[57] Baserade på användning av modern testteori, Item Response Theory (IRT).
[58] Ur Wheadon & Stockford 2010, s. 15.

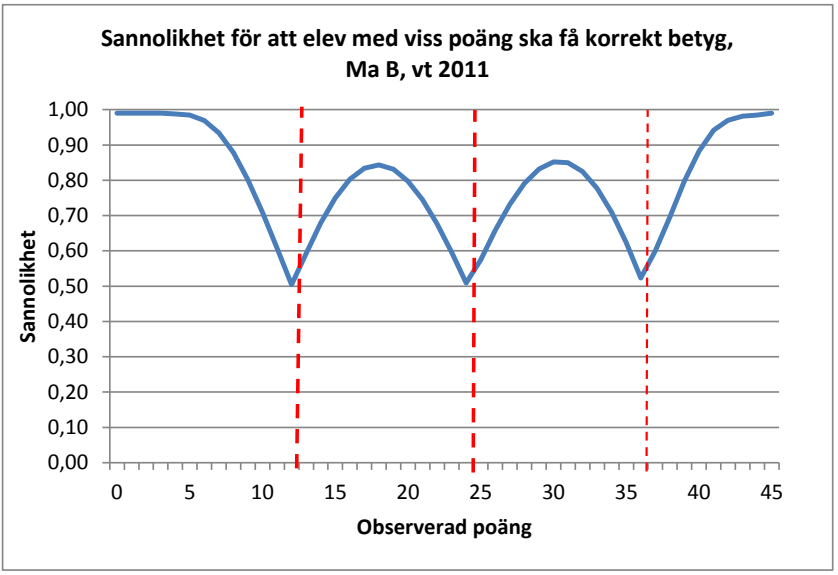
takt med antalet klasser. De svenska resultat som används i det här redovisade exemplet gäller fyra betygsklasser, medan de engelska värdena baseras på sex klasser (A–E samt *fail*). Det vill säga motsvarande den nu gällande svenska betygsindelningen. Om det aktuella provet gällt i dag skulle således ytterligare två betygsgränser med vidhängande falska klassificeringar (FN och FP) ha tillkommit vilket skulle lagt de svenska resultaten mer i nivå med de engelska.

Sannolikhet för korrekt betyg i relation till erhållen provpoäng

Som nämnts ett antal gånger är det inte möjligt att utifrån provresultaten avgöra vilka elever som har falskt positiva eller falsk negativa resultat. Däremot kan man skatta hur stor sannolikheten är för att en elev med en viss poäng fått ett korrekt provbetyg. Om man för enkelhets skull redovisar dessa sannolikheter i ett diagram får man bilden i figur 17.

Figuren visar att för elever med provpoäng i närheten av en betygsgräns är sannolikheten cirka 0,50 att eleven ska få ett korrekt betyg. För en grupp av elever i närheten av en kravgräns innebär det att cirka 50 procent får det ena betyget (korrekt) och cirka 50 procent ett annat (felaktigt) provbetyg. De elever som har poäng i mitten av ett betygsintervall har för det aktuella provet cirka 0,85 (85 procents) sannolikhet att få korrekt provbetyg, men också cirka 15 procent av dem kan anses ha fått ett felaktigt provbetyg (skulle haft ett högre eller lägre betyg). Ungefär var sjunde elev av dem som ligger mitt emellan två betygsgränser får således ett för högt eller lågt provbetyg på det aktuella provet. De elever som har riktigt höga eller riktigt låga provpoäng har största sannolikheten att få ett korrekt betyg IG eller MVG.

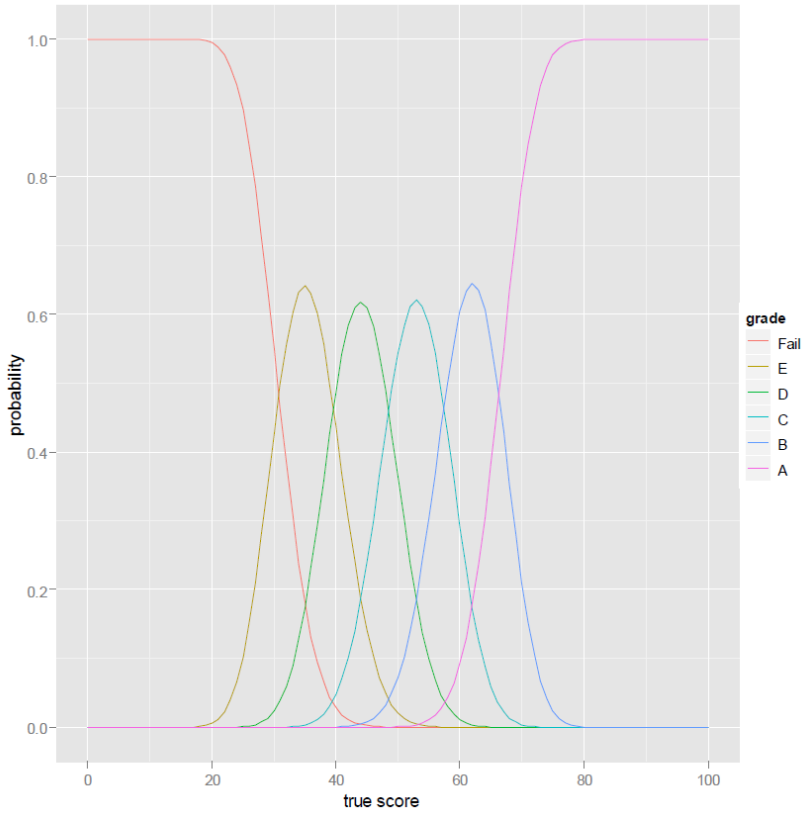
Figur 17 Skattning av sannolikheten för att en elev med en viss observerad poäng har fått korrekt provbetyg.



Även när det gäller ovanstående skattning kan man jämföra med den engelska motsvarigheten. [59]

[59] Ur Wheadon & Stockford 2010, s. 9. Se också Bramley & Dhawan 2010.

Figur 18 Sannolikheten för korrekt klassificering för elever med olika sann poäng för ett engelskt A-level prov.



Som vi ser är likheten mellan det svenska och det engelska diagrammet slående. I det engelska har man ritat ut hela fördelningen medan figuren för ma B endast innehåller de delar som har sannolikhet över 0,50.[60]

Vi ser i figuren att sannolikheten för korrekt betygsklassificering av elever i mitten av intervallen är lägre för det engelska

[60] Detta för att förenkla figuren. Det är dock lätt att rita om den på samma sätt som den engelska. Den engelska figuren visar förutom sannolikheten för korrekt betyg (de delar av graferna som ligger över skärningspunkterna) även sannolikheten för falskt negativt och falskt positivt provbetyg (de delar av graferna som ligger under skärningspunkterna). Observera också att figur 18 anger sann poäng på x-axeln medan figur 17 anger observerad poäng. Denna skillnad har endast marginell betydelse och görs för att förenkla förståelsen av figur 17.

provet (0,60–0,65), vilket inte beror på att det är mindre reliabelt utan på att det har fler betygssteg än det svenska exemplet. Vi kan förvänta oss att en analys av ett svenskt prov enligt 2011 års sexgradiga betygsskala uppvisar en bild liknande den engelska. Om man i figur 16 (och 17) tänker sig att man lägger in ytterligare två betygsgränser tillkommer ytterligare två områden med falskt positiva och falskt negativa resultat. Vi går då från fyra till sex betygsgrupper och om vi antar mätfel av ungefär samma storlek blir det totala felet cirka 50 procent större, dvs. det går från cirka 20 procent av eleverna till cirka 30 procent. Därmed skulle andelen korrekt klassificerade bli mellan 65 och 75 procent. En jämförelse med tabell 14 visar att dessa värden korresponderar tämligen väl med motsvarande engelska resultat.

Slutsatser

Som framgått av avsnittet är slumpfelen på individnivå betydande. Det aktuella värdet på SEM enligt stickprovet (3,24 poäng) är beräknat på gängse sätt. Det medför att värdet är ett genomsnitt som anses gälla lika för hela poängskalan. Noggrannare analyser visar emellertid att standardfelet i själva verket varierar så att det är större i mitten av skalan och minst ute vid ändarna.[61] I gängse litteratur på området[62] anses dock det genomsnittliga värdet tillräckligt för praktiskt bruk, men kan alltså betraktas som en approximation.

I figur 16 och vid bestämningen av klassifikationsfelet (*classification accuracy*) har för det första en viss approximativ metod använts för bestämning av ett betingat[63] SEM. För den fördelning inom vilken den sanna poängen antas ligga har för det andra en normalfördelning valts. För såväl betingat SEM som fördelning kan olika metoder väljas och de som valts här kan ses som ganska enkla och osofistikerade.[64] Syftet med analysen har dock inte varit att försöka hitta det mest korrekta värdet på klassificeringsriktigheten och på andelen (antalet) elever som hamnat i gruppen falskt

[61] Brukar kallas *conditional SEM* eller betingat SEM.
[62] Se t.ex. Crocker & Algina (1986).
[63] *Conditional*.
[64] Metoderna beskrivs närmare i Skolverket (2015b).

positiva respektive falskt negativa, utan mer att peka på storleksordningen. Jämförelserna med de engelska resultaten, som erhållits med modernare metoder, visar att den här använda metoden ger resultat i samma storleksordning. Det finns därför ingen anledning att tro att de använda approximationerna skulle ha någon avgörande betydelse för utfallet.

Den avgörande slutsatsen blir att trots det aktuella provets höga reliabilitet (0,91) är det rimligt att anta att cirka 20 procent av eleverna av rena slumpskäl får ett för lågt eller högt provbetyg på det aktuella provet. För den nya betygsskalan som gäller från 2011 kan slumpfelet antas öka till mellan 25 och 35 procent.[65]

Bedömningsfel

Den tredje typen av fel handlar om bedömningen av elevers provsvar och prestationer i övrigt. Framför allt handlar det om tolkning av texter, men även olika typer av problemlösningar, laborationer, praktiska övningar och inte minst muntliga prestationer ingår. Läraren gör också fortlöpande bedömningar som utgör viktiga underlag vid den slutliga betygssättningen.

När det gäller prov, och i synnerhet de nationella prov som är i fokus för den här bilagan, handlar det mer om enstaka punktmätningar. Proven är i allmänhet standardiserade i den meningen att de ges under bestämda villkor vad avser exempelvis tid, plats och tillåtna hjälpmedel, och med strikta regler för genomförandet. Syftet med detta är att elevernas prestationer ska utföras under så likartade förhållanden som möjligt. Detta kan ses som en grundläggande rättvisefråga.

Provet kan betraktas som en form av stimuli på vilket provtagaren ska generera utsagor som kan ligga till grund för en analys av i vilken utsträckning hon eller han behärskar det kunskapsområde som provet avser att pröva. Beroende på kunskapsområdets art är proven och utsagorna av olika karaktär och bedöms och tolkas på olika sätt och enligt olika principer.

[65] Detta kan förstås prövas empiriskt. De lärosäten som utvecklar de nationella proven anser dock att proven behöver ha varit i användning några år för att ha stabiliserats och att resultat på de tidiga proven inte bör ligga till grund för mer långtgående slutsatser.

Vissa prov baseras på ett antal förhållandevis korta uppgifter som bedöms var för sig, vanligen med poäng, belägg eller liknande. Resultaten anges då ofta som en poängsumma. För denna typ av prov finns utvecklade teorier[66] med hjälp av vilka det är möjligt att beräkna och uppskatta hur väl provresultaten kan anses ge en representativ bild av provtagarens kunskaper och med vilken säkerhet man kan uttala sig. Givetvis måste provet på ett så allsidigt sätt som möjligt spegla det aktuella kunskapsområdet (ha god validitet) och på olika sätt försöka minimera olika slumpfaktorers inverkan (ha hög reliabilitet). Att sådana krav uppfylls är en uppgift för provkonstruktörerna. I provkonstruktionen ligger också att konstruera uppgifter som ger så litet slumpinflytande som möjligt vid bedömningen. Flervalsfrågor kan ses som ideala ur det perspektivet. Även kortsvarsuppgifter, matchningsuppgifter och uppgifter där ord ska fyllas i ger hög överensstämmelse vid bedömningen. Samtidigt måste provkonstruktören avgöra om denna typ av uppgifter ger tillräcklig validitet. Reliabiliteten i form av bedömaröverensstämmelsen är hög för denna typ av provuppgifter, vilket styrks av många studier.[67]

Om proven å andra sidan utgörs av uppgifter där provtagarna ska producera mer omfattande texter blir bedömaruppgiften svårare. Denna typ av uppgifter ges ofta av validitetsskäl. Till exempel att en viktig kunskap är att kunna föra ett resonemang om någon fråga eller företeelse, eller att kunna producera en text av god kvalitet i olika genrer. Vid sådana prov flyttas kraven på validitet och reliabilitet till stor del över på bedömarna. Uppgifterna eller provet bedöms i sådana fall ofta helhetligt i betygstermer och då måste bedömaren till stor del själv tolka provtexten i relation till de krav eller kriterier som gäller för olika betyg (vilka också kräver tolkning). Även om provkonstruktören tillhandahåller förebildliga exempel krävs alltså betydande tolkning av bedömaren. Någon entydigt sann bedömning finns knappast när det gäller exempelvis uppsatser. Den säkraste bedömningen vore att låta många kompetenta bedömare bedöma varje uppsats och sedan beräkna ett medelvärde av deras bedömningar. Ett sådant tillvägagångssätt är dock inte möjligt av ekonomiska och praktiska skäl och därför lär bedömningen även framöver i huvudsak vila på en eller ett par

[66] Olika former av klassisk och modern testteori, se t.ex. Crocker & Algina (1986).
[67] Se t.ex. Massey & Raikes (2006).

bedömares uppfattningar.[68] Detta får till följd att skillnaden i bedömning är betydligt större mellan olika bedömare för denna typ av prov än för prov baserade på många avgränsade uppgifter med mer distinkta svar. Frågor om bedömarreliabilitet är komplicerade och det finns en hel del forskning på området.[69] Utrymmet här medger inte någon utförligare diskussion i frågan.

Skolinspektionens ombedömningar

Skolinspektionen har sedan några år ett uppdrag från regeringen att genomföra ombedömningar av ett urval av delprov som ingår i de nationella proven.[70] Frågan om relationen mellan prov och betyg är komplex och av litteraturen på området följer inga entydiga vetenskapligt fastlagda riktlinjer för vilken överensstämmelse som kan eller bör rekommenderas.[71] Dock kan överenstämmelsen mellan olika bedömare höjas genom träning och utbildning. Empiriska resultat visar att bedömaröverensstämmelsen i allmänhet rör sig på en glidande skala från höga värden för prov med många uppgifter och korta svar, mot betydligt lägre värden för prov med långa skriftliga svar eller uppsatser. Vissa prov kan innehålla både lättbedömda och svårbedömda uppgifter.

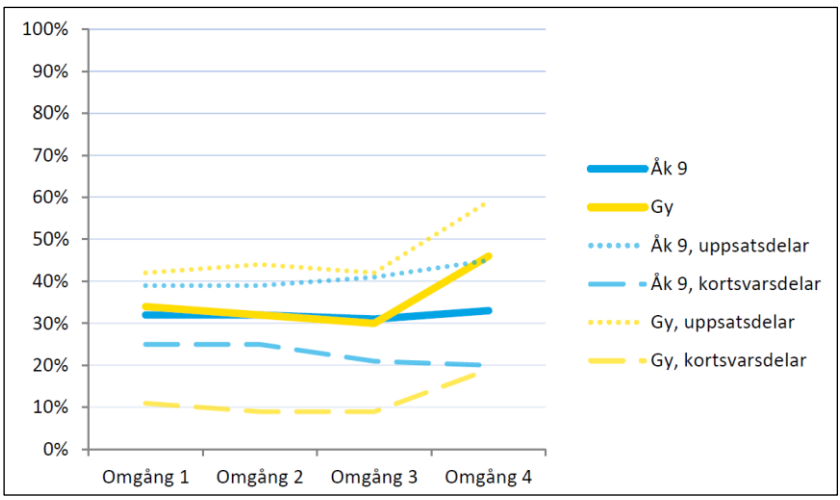
För en myndighet som ska bedriva tillsyn måste någon form av norm fastställas. Skolinspektionen har i sin rapport 2013 valt att för uppsatser sätta gränsen för acceptabel överensstämmelse mellan ursprungsbedömare och ombedömare till 0,70. Det betyder att när ett urval uppsatser från en skola har bedömts och betygssatts av ombedömarna ska minst 70 procent av deras betyg överensstämma med den bedömande lärarens betyg på samma uppsatser för att resultatet ska anses acceptabelt. Detta är ett vanligt värde i olika tekniska rapporter av provresultat. I sådana fall handlar det om överensstämmelse mellan två oberoende bedömare. Den svenska modellen med lärare som bedömer obligatoriska nationella prov är

[68] Det är dock möjligt att effektivisera bedömningen genom att t.ex. utveckla digitala metoder för automaträttning med dator. Även försök med extern bedömning och medbedömning som utredningens föreslår kan effektivisera bedömningen och göra den mer likvärdig.
[69] Se t.ex. Tisi m.fl. (2013) och Ofqual (2014) för allmänna kunskapsöversikter om forskning på området bedömarreliabilitet.
[70] Skolinspektionen (2010), Skolinspektionen (2011), Skolinspektionen (2012), Skolinspektionen (2013), Skolinspektionen (2015).
[71] T.ex. Tisi m.fl. (2013) och Ofqual (2014).

internationellt mycket ovanlig och ombedömningen i Sverige har mer rollen av kontrollfunktion, medan studierna i andra länder i huvudsak handlar om att genom träning finna en rimlig norm för vad som kan betraktas som tillräckligt god bedömarreliabilitet mellan oberoende bedömare.

Den svenska forskningen på området är inte särskilt omfattande[72] och för syftet i den här bilagan nöjer vi oss med några resultat från Skolinspektionens ombedömning. Figur 19 visar avvikelser för några delprov i ämnena svenska och engelska för åren 2009 (omgång 1) till 2012 (omgång 4). Av figuren framgår tydligt att avvikelserna mellan bedömarna är störst för de delprov som utgörs av uppsatser medan delprov som endast innefattar uppgifter med korta svar visar mindre skillnader.

Figur 19 **Avvikelse mellan ursprungsbedömare och ombedömare. (Bedömaröverensstämmelsen är 100 minus avvikelsen.) Ur Skolinspektionen (2013), s. 15.**



Man kan notera att avvikelserna ligger på ganska jämn nivå de tre första åren för att sedan öka betydligt det fjärde året, främst för gymnasieskolan. Förklaringen till detta är att 2012 var det första året då den nya sexgradiga betygsskalan tillämpades på de aktuella

[72] Se t.ex. Skolverket (2009).

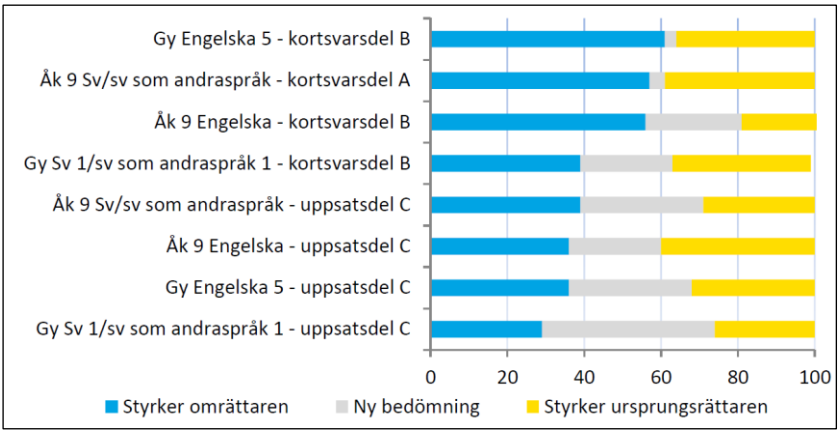
gymnasieproven. Eftersom den innebar att antalet betygssteg ökade från fyra till sex (med 50 procent) så finns det ytterligare två betygsgränser där avvikelser kan förekomma. Om alla betygsgränser är ungefär lika svårbedömda bör ökningen från fyra till sex betygssteg innebära att andelen avvikelser ökar i samma utsträckning, dvs. med 50 procent.

Av figur 19 framgår tydligt att för gymnasieskolan stämmer ökningen i avvikelse för uppsatsdelen mycket väl med ökningen av antal betygssteg, från drygt 40 procent avvikelse till knappt 60 procent, dvs. en ökning med knappt 50 procent (20 procentenheter). Det som är något svårförståeligt är att avvikelsen för kortsvarsdelar nästan fördubblas från 10 till 20 procents avvikelse. Denna typ av delprov brukar inte erbjuda särskilt stort tolkningsutrymme. Det är också svårt att förstå varför tillkomsten av ytterligare två betygssteg skulle få så stor inverkan på bedömningen av sådana delprov.

För gymnasieskolan som helhet anger Skolinspektionens rapport att avvikelsen under de tre år av ombedömning som gäller 1994 års betygsskala ligger på cirka 30 procent. Om man skulle utgå från att ombedömarna bedömer korrekt skulle således ”felet” vara cirka 30 procent. Nu kan man emellertid inte utgå från att ombedömarna är ”felfria” (vilket Skolinspektionen inte heller gör), vilket gör att felet rimligen bör kunna vara mindre än 30 procent.

I vissa fall där ursprungsbedömare och ombedömare varit oeniga har en ytterligare ombedömare fått ge sitt omdöme. Figur 20 visar resultaten. Man kan se en viss tendens till att den nya ombedömaren i högre grad stöder den första ombedömaren än ursprungsbedömaren. Det tycks framför allt gälla kortsvarsuppgifterna. För uppsatsdelarna är tendensen ännu svagare och en rimlig slutsats blir att ”felet” för dessa delprov är i det närmaste försumbart, i varje fall för de uppgifter som gått vidare till en andra ombedömning.

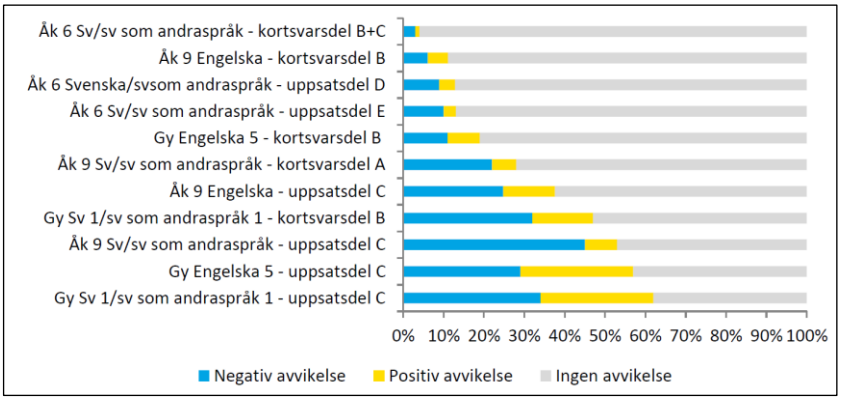
Figur 20 Prov med avvikelser som ombedömts en andra gång, för respektive delprov, omgång 4. Ur Skolinspektionen (2013), s. 18.



Dock kan man anta att de fall som ombedömts en andra gång i stor utsträckning rör bedömningar som ligger i närheten av betygsgränser. Dessa fall har valts för att det råder oenighet mellan ursprungsbedömare och första omgångens ombedömare samt eftersom andra omgångens bedömare i så stor utsträckning stöder såväl ursprungsbedömare som första omgångens ombedömare. För de elevarbeten som är typiska för respektive betyg torde bedömaröverensstämmelsen vara tämligen hög. Rimligen är det gränsfallen som skapar oenighet och då är det inte heller förvånande att ombedömarnas omdömen är tämligen jämt fördelade.

Av den litteratur och de rapporter som gäller bedömaröverensstämmelse kan man säga att en överenstämmelse på mer än 70 procent förefaller allmänt accepterad som ett mått på godtagbar överensstämmelse. Hur man ska tolka Skolinspektionens resultat i det avseendet är svårt att avgöra. Men om man utgår från att den första ombedömningen är korrekt förefaller den genomsnittliga överenstämmelsen för det aktuella urvalet av skolor ligga på gränsen för det acceptabla med en avvikelse på cirka 30 procent (för en fyrgradig betygsskala).

Figur 21 **Fördelning av andel avvikelser: positiva, negativa och ingen avvikelse mellan ursprungsbedömarens bedömning och Skolinspektionens bedömning för de olika delproven, omgång 4. Ur Skolinspektionen (2013), s. 14.**



Den sammanfattande bedömningen när det gäller såväl positiv som negativ avvikelse i Skolinspektionens ombedömning visas i figur 21. Om man utgår från uppsatsproven och ser både positiv och negativ avvikelse som ett mått på bedömningsfel ligger felet i intervallet cirka 35 till 60 procents avvikelse (sexgradig betygsskala). I så stor andel av fallen är alltså ursprungsbedömare och ombedömare oeniga.

Om man använder samma terminologi som för mätfelens effekter på poängbaserade delprov skulle ”negativ avvikelse” motsvara det som kallades FP (falskt positiva)[73]. Om vi utgår från årskurs 9 med fyrgradig betygsskala ser vi att för uppsatsdelen i engelska (delprov C) är andelen FP cirka 25 procent och FN cirka 13 procent. Andelen korrekt bedömda är drygt 60 procent. För svenska i årskurs 9 (delprov C) är motsvarande värden FP cirka 45 procent, FN cirka 8 procent, vilket ger en andel korrekta på knappt 50 procent. Vid en jämförelse med andelen elever med korrekta betyg respektive falskt positiva eller falskt negativa betyg

[73] Avvikelse är skillnaden mellan ombedömare och ursprungsbedömare (O-U). Negativ avvikelse innebär då att ursprungsbedömaren hade ett högre betyg än ombedömaren. Om vi antar att ombedömaren gör den ”sanna” bedömningen betyder det att ursprungsbedömaren satt för högt betyg, dvs. eleven är falskt positiv vad gäller provbetyget. På motsvarande sätt motsvarar positiv avvikelse andelen falskt negativa elever.

för matematik (tabell 13), framgår att andelen ”korrekta” betyg är lägre för engelska och betydligt lägre för svenska.

Så vilka slutsatser kan man dra om bedömningsfelets storlek? Skattningen ovan bygger på antagandet att ombedömarens bedömning alltid är korrekt. Detta är knappast realistiskt. Låt oss därför som skattning av genomsnittlig avvikelse för uppsatsprov ange intervallet 25–35 procent (för den fyrgradiga skala som undersöks i den här bilagan) med tolkningen att om en annan bedömare gjort bedömningen är detta sannolikheten för att betyget blivit ett annat. Man kan också notera att andelen falskt positiva är betydligt större än andelen falskt negativa, för matematikexemplet i tabell 13 var det tvärtom men där baserades resultaten på skattningen av sann poäng och provets betingade standardfel CSEM.

Sammanfattande diskussion

Inledningsvis i bilagan angavs nedanstående syften:

- Att sammanställa provbetygens variation över tid.
- Att bestämma kravgränsfelets storlek.
- Att jämföra kravgränsfelet med standardfelet (SEM) och de fel i klassificering till olika betyg som SEM leder till.[74]
- Att jämföra kravgränsfel med bedömningsfel och slumpfel.
- Att ge ett underlag för att bedöma vilka felnivåer som är acceptabla inom ramen för prov av god kvalitet.

I det här avslutande avsnittet görs ett försök att sammanställa och diskutera resultaten.

[74] Det som kallas *classification accuracy* på engelska.

Provbetygens variation över tid och bestämning av kravgränsfelets storlek

Kravgränsfelens storlek redovisas i de sammanställningar och analyser som finns i den första delen och i appendix 1. För grundskolans del, där samtliga provbetyg samlas in och hela populationen genomför samma prov, är detta en förhållandevis enkel uppgift[75]. För gymnasieskolans del är uppgiften svårare och resultaten osäkrare eftersom endast ett stickprov av provbetyg samlas in. Dessutom kan de grupper som genomför proven variera mellan olika år när det gäller programtillhörighet. Båda dessa förhållanden kan påverka resultaten. För svenskans och engelskans del (sv A och eng A) finns dessutom sammanfattande provbetyg endast för sju år (2005–2011) medan det för matematiken (ma A) finns redovisade provbetyg för tolv år (1999–2011). Detta gör att resultaten i svenska och engelska är mindre tillförlitliga än resultaten i matematik.

Den metod som använts för att skapa ett mått på den årliga avvikelsen i respektive betygskategori går ut på att en trendlinje anpassas till den andel elever (i procent) som tilldelats respektive provbetyg olika år. Därefter beräknas skillnaden (residualen) mellan den observerade andelen och det av trendlinjen angivna värdet. På detta sätt fås ett mått på hur stor andel elever (i procentenheter) i respektive betygskategori som fått för högt eller för lågt provbetyg i relation till trendlinjens värde. Absolutvärdet av samtliga avvikelser ett visst år summeras sedan och ger den totala avvikelsen i klassificeringen[76].

Ett observandum

Om man emellertid tänker sig att man för ett poängbaserat prov i efterhand skulle korrigera en betygsgräns så att andelen elever i en viss betygkategori (t.ex. IG) bättre skulle stämma med trendvärdet blir tolkningen av resultatet (avvikelsen) annorlunda. Det beror på att det finns ett randvillkor att summan av procentandelarna för de

[75] Redovisad i Skolverket (2015a).
[76] Observera att detta är den klassificering som baseras på andel elever som fått respektive provbetyg. Det ska inte sammanblandas med den klassificering (classification accuracy) som baseras på mätfelet eller den klassificering som bygger på jämförelser mellan ursprungsbedömare och ombedömare.

fyra betygskategorierna varje år är 100 procent. Om då en betygskategori ökar sin andel måste någon annan minska lika mycket för att summan av de fyra betygsandelarna fortfarande ska vara 100 (en sorts nollsummespel). Det betyder att om kategorin IG minskar med ett visst antal elever, exempelvis 200, genom att en betygsgräns flyttas en poäng, ökar kategorin G lika mycket. Det vill säga de 200 elever som legat en poäng under G får nu G i stället för IG. Ändringen av betyget för 200 elever minskar således andelen i kategorin IG med 200 och ökar den i kategorin G med 200, en ändring med 200 elever i var och en av två kategorier (en total ändring på 400 uttryckt i avvikelser). Men endast 200 fysiska elever har i realiteten fått ändrat betyg, de 200 som fick G i stället för IG.

Det betyder alltså att antalet felklassificeringar (avvikelser) anger dubbelt så stora värden som det faktiska antal elever som skulle beröras av en korrigering. Detta kan vara viktigt att ha i åtanke när man värderar kravgränsfelet.

Kravgränsfel och bedömningsfel

Prov i svenska och engelska består ofta av olika delprov där vissa är poängbaserade och andra är av uppsatstyp och helhetligt bedömda.

För de poängbaserade proven gäller i princip samma resonemang som ovan för matematikproven där en förskjutning av poänggränsen för något provbetyg direkt ändar andelen elever i de två av poänggränsen berörda grupperna. För uppsatsproven finns däremot ingen skarp kravgräns och inga möjligheter att justera någon poänggräns. Här är det den bedömande läraren som – med stöd av de anvisningar och exempel som tillhandahålls i provmaterialet – utifrån sina tolkningar och bedömningar avgör delprovsbetyget. Någon central instruktion från Skolverket kan därmed inte på samma sätt som en flyttad poänggräns ändra en given fördelning, eftersom den beror av varje enskild bedömande lärares helhetliga bedömning.

Betygsfördelningen på poängbaserade prov beror således i huvudsak på den bedömning den kravgränssättande gruppen hos provkonstruktören gjort om placeringen av poänggränser för olika provbetyg. Det är en bedömning gjord av cirka 10–15 särskilt utvalda personer. Provbetygens fördelning beror därmed i huvudsak på dessa personers bedömning och beslut. Sådana beslut fattas

varje år och det är knappast förvånande att det förekommer variationer mellan olika år när en så pass begränsad grupp gör sina bedömningar. I varje fall är det rimligt i relation till den stora mängd bedömare som är inblandade när provbetygen sätts för de uppsatsbaserade proven. Där är provbetygen grundade på hela lärarpopulationens bedömning[77] (de som fungerar som bedömare) medan provbetygen på poängbaserade prov baseras på ett i sammanhanget litet stickprov bedömares (kravgränssättares) bedömningar.

Ovanstående skulle kunna vara en förklaring till varför kravgränsfelet på aggregerad nivå är större för poängbaserade prov (matematik) än för prov som kombinerar holistiskt bedömda prov och poängbaserade prov (svenska och engelska).[78] Det ”stickprov” 10–15 personer[79] som sätter poänggränser för poängbaserade prov kan antas ha svårare att bibehålla en viss standard över tid än de många tusen bedömare som sätter provbetyg på uppsatser.[80] Det skulle innebära att de prov som innefattar delprov av olika slag (med poäng- respektive holistisk bedömning) kan förväntas få ett lägre kravgränsfel än ett renodlat poängprov.

Den poängbaserade betygsskalan innebär oftast att varje poängsteg motsvarar flera procentenheters ändring i andelen provbetyg på intilliggande nivåer. Detta gäller inte för de helhetligt bedömda proven där varje bedömare själv avgör provbetyget för varje enskilt prov. Där finns inga definierade gränser att överskrida, men självklart är osäkerheten störst vid betygsgränser även för holistiskt bedömda uppgifter och prov, men det finns inga statistiska underlag som möjliggör skattningar av andelen falskt positiva och falskt negativa provbetyg för den typen av prov.

Ett memento i sammanhanget är som vi tidigare konstaterat att de lärare som bedömer poängbaserade prov vet vid vilka poäng betygsgränserna ligger och att detta tenderar att leda till att en oproportionerligt stor andel elever har provpoäng just över gränsen för ett högre betyg. Figur 1 och 15 illustrerar företeelsen. Dock är denna avvikelse, som främst påverkar kravgränsfelet, i det närmaste

[77] Där varje lärare bedömer ett mindre stickprov av elever (vanligen en eller ett par klasser).
[78] Jämför tabell 10.
[79] Oberoende av om det är samma eller olika individer olika år.
[80] En sorts tillämpning av ”stora talens lag” eller den ”massans visdom” som Surowieccki (2004) talar om.

försumbar i relation till den stora felkällan för poängbaserade prov, nämligen slumpfelet SEM och de falskt positiva och falskt negativa provbetyg det genererar.

Slumpfelet SEM

Slumpfel finns på alla nivåer. Den grupp som sätter kravgränser kan ses som ett stickprov ur den totala grupp ämnesdidaktiker, lärarutbildare, lärare etc. som kan ingå i denna grupp. Kravgränsfelet kan ses som ett mått på detta fel om vi ser den långsiktiga trendlinjens värde som ett övergripande populationsmått. Det handlar då om slumpens inverkan på nationell nivå.

Olika bedömare gör i större eller mindre utsträckning olika bedömningar av samma uppsatser som Skolinspektionens granskning visar. Det betyder att för varje grupp eller klass som bedöms av samma bedömare finns det en slumpfaktor när det gäller om bedömaren är genomsnittlig, sträng eller mild. Här handlar det om slump på gruppnivå.

Det tredje slumpfelet är kopplat till den enskilda provdeltagaren. Det kan gälla dagsform, tur eller otur med valet av uppgifter till provet eller andra faktorer av individuell art som påverkar den enskilda elevens prestation. För poängbaserade prov kan storleken av sådana slumpfel skattas med hjälp av klassisk eller modern testeori. Detta ger ett mått, den enskilda mätningens standardfel (SEM), med vars hjälp tillförlitligheten i en provpoäng kan anges. I synnerhet för eleverna i närheten av poänggränserna för ett provbetyg blir osäkerheten stor och många elever tilldelas för höga eller för låga provbetyg och kan klassas som falskt negativa eller falskt positiva beroende på om de fått för lågt eller för högt betyg.

För bestämning av mätfel på individnivå finns många metoder. De bygger på delvis olika antaganden och kan därmed ge något olika resultat. Skillnaderna är dock i allmänhet marginella. I den här studien har tämligen elementära metoder använts. Jämförelsen med de engelska proven visar emellertid att de använda metoderna ger resultat av samma storleksordning som de engelska studierna, vilka baseras på mer sofistikerade metoder. Det viktiga i det här sammanhanget har inte varit att presentera så precisa resultat som

möjligt utan att peka på de olika feltypernas betydelse vid bedömningen av ett provbetygs tillförlitlighet och stabilitet.

De olika felens storlek och relation till varandra

Vi har i de olika avsnitten försökt skatta de olika felen uttryckt i hur stor andel av eleverna som kan antas få ett felaktigt provbetyg till följd av respektive felkategori.

Kravgränsfelen

Kravgränsfelen finns sammanfattade i tabell 10. Värdet i kolumnen ”Medel” för proven i svenska och engelska anger i genomsnitt kravgränsfel i storleksordningen *2 till 3* procent av eleverna. Så stor andel av eleverna kan uppskattas ha fått ett felaktigt provbetyg på grund av kravgränsfel. Observera dock att dessa värden är tämligen osäkra och får tolkas försiktigt.[81] För matematik är motsvarande andel *5 till 6* procent (i gymnasieskolan). Dessa värden är mer tillförlitliga eftersom de bygger på en längre tidsserie.

Den enskilda mätningens standardfel SEM

De data som utgjort underlag för analysen av SEM i den här bilagan gäller provet i matematik B vårterminen 2011. För denna typ av analyser krävs data för varje elev och varje uppgift. Sådana finns inte tillgängliga annat än som de stickprov som insamlas av de lärosäten som konstruerar de nationella proven. Umeå universitet har tillhandahållit de data som används här.

Tabell 13 sammanfattar resultaten på det aktuella provet. Där framgår att andelen elever som kan antas ha fått fel provbetyg uppgår till cirka 20 procent. Felet är beroende av såväl det aktuella provet som av den grupp som gjort provet och kan variera något.[82] Vi antar här att det aktuella stickprovet är representativt för den population som gjort provet. Om vi vidare antar att mätfelet med-

[81] Se appendix 2.
[82] Genom att reliabilitet och standardavvikelse som ingår i beräkningen av SEM är beroende av den aktuella gruppens resultat.

för att *15 till 25* procent av eleverna har fått ett ”falskt” provbetyg har vi knappast överdrivit. Denna skattning kan anses gälla allmänt för nationella prov i matematik (och andra ämnen) som är baserade på poängsatta uppgifter och en fyrgradig betygsskala. I en sexgradig skala ökar andelen elever med falska provbetyg med cirka 50 procent, till andelar på mellan 25 och 40 procent av provdeltagarna.

Bedömningsfelet

För andra typer av prov än de poängbaserade, främst de med holistisk bedömning, kan inte standardfel beräknas lika enkelt. Det skulle kräva att flera oberoende bedömare bedömde varje uppsats, varefter samstämmigheten i betygssättningen beräknades. Några data för att göra skattningar av bedömningsfelens storlek för de nationella proven finns inte utöver de som Skolinspektionen samlat in och granskat.[83] Deras resultat kan dock sägas vara i linje med vad utländska källor anger som rimligt så låt oss därför anta att samstämmigheten mellan olika bedömare ligger på cirka 70–90 procent, beroende på bedömningens komplexitet. Det skulle innebära att bedömningsfelet skulle kunna antas ligga i intervallet *10 till 30* procent. Så stor andel av eleverna skulle alltså få ett avvikande provbetyg (i relation till vad en annan bedömare kan antas ha ansett) på grund av bedömningsfel. Den lägre gränsen (10 procent) kan då t.ex. antas gälla något mer komplexa uppgifter i matematik eller frågor som kräver viss skriftlig redovisning. Den övre gränsen (30 procent) gäller för mer fullskaliga texter av typen uppsatser. För flervalsfrågor och andra enkla uppgiftstyper kan bedömningsfelet anses försumbart.

Det totala felet

Hur stort blir då det totala felet? Det är betydlig svårare att skatta. Som framgått ovan är de olika proven konstruerade på olika sätt. Proven i matematik är i huvudsak poängbaserade medan proven i

[83] Några mer ingående svenska studier på området finns inte. Skolverket publicerade 2009 en mindre studie, Skolverket (2009).

svenska och engelska innefattar vissa delprov som bedöms med poäng och andra som bedöms helhetligt.

Om vi utifrån tidigare erhållna resultat gör ett försök att grovt skatta storleken på de olika typerna av fel för proven i olika ämnen kan det se ut som nedan.

Tabell 15 Skattad storlek på olika typer av fel uttryckt i andel elever som får fel provbetyg utifrån en fyrgradig betygsskala.

Ämne	Andel elever (procentenheter)		
	Kravgränsfel totalt	Slumpfel (SEM) poängbaserat (del)prov	Bedömningsfel
Svenska	2–3	Okänt	10–30
Engelska	2–3	Okänt	10–30
Matematik	4–6	15–25	0–5

De olika felen påverkar och överlappar varandra i större eller mindre grad och det är därför svårt att utan mer ingående analyser skatta vad det totala felet blir. Det man kan notera är att det fel som vid en ytlig granskning ofta framstår som svårast – att bestämma kravgränser för olika betyg, antingen det handlar om att fastställa poänggränser för olika provbetyg eller att enas om exempel på uppsatser som ligger just ovanför eller under ett visst provbetyg – tycks vara det fel som leder till klart minsta antalet elever med felaktigt provbetyg. Såväl slumpfelen som bedömningsfelen tycks ha avsevärt större betydelse i det avseendet. Detta gäller på individnivå. På aggregerad nivå blir däremot slumpfel och bedömningsfel betydligt mindre än på individnivå, t.ex. genom att antalet falskt negativa och falskt positiva provbetyg tar ut varandra.

Betydelsen av antal betygssteg

För samtliga feltyper gäller att andelen elever som tilldelas fel provbetyg ökar med antalet betygssteg. Det man vinner i skalprecision genom fler betygssteg förlorar man i träffsäkerhet (*accuracy*). Det underliggande mätfelet på poängskalan blir inte mindre för att betygsskalan får fler steg, och varje betygssteg påverkas av mätfelet.

Ett par ytterligare kommentarer

Genomsnittlig betygspoäng

Den genomsnittliga betygspoängen (GBP) är ett mått som kan fungera som en grov indikator, men inte medger mer ingående analyser. GBP har sin funktion på aggregerad nivå för att beräkna genomsnittlig betygspoäng för klasser, skolor eller andra större grupper. Dessutom används GBP för att beräkna meritpoäng.

Det som gör den genomsnittliga betygspoängen så svårhanterlig och svårtolkad är den asymmetriska poängsättningen. För 1994 års betygsskala gäller:

Betyg	IG	G	VG	MVG
Betygspoäng	0	10	15	20

För 2011 års betygsskala gäller:

Betyg	F	E	D	C	B	A
Betygspoäng	0	10	12,5	15	17,5	20

För en elev som går från G till IG sjunker meritvärdet (eller GBP) med 10 poäng. För att kompensera detta krävs att två elever med betyget G eller högre höjer sitt betyg ett steg för att meritvärdet ska vara oförändrat, eller alternativt att en elev med G eller VG ökar ett betyg två steg. Betygspoängens sneda fördelning efter 1994 beror på att betyget G av politiska skäl markerats särskilt. Ett extra tryck skulle läggas på att nå betyget G.

För 2011 års betygsskala är asymmetrin ännu starkare. Här krävs att *fyra* elever med E höjer sitt betyg ett steg för att kompensera en nedgång för *en* elev från E till F.

Information om provresultat och felkällor

Bilagan har undersökt och diskuterat de olika typer av fel som förekommer i samband med nationella prov. Vid Skolverkets redovisning av provresultat finns ingenting nämnt om dessa olika fel. Inte heller de lärosäten som utvecklar de nationella proven redovisar mätfel för olika prov och delprov. Detta leder till en övertro på provresultat som mått på elevers kunskaper, i synnerhet när det gäller enskilda elever. Proven är konstruerade i enlighet med de principer som gäller för konstruktionen av nationella prov av stor betydelse. Men även om proven är välkonstruerade är provpoäng och provbetyg inte så stabila och tillförlitliga som man skulle önska, men de är de bästa hjälpmedel som står till buds. Det är då viktigt att de som ska använda proven och provresultaten – elever, lärare, rektorer, beslutsfattare, politiker, skoldebattörer m.fl. – är välinformerade om provs styrkor och svagheter.

Referenser

AERA, APA & NCME[84]. (2014). Standards for Educational and psychological testing.

Bramley, T. & Dhawan, V. (2010). Estimates and reliability of qualifications. Coventry: Office of Qualifications and Examinations Regulation.

Cizek, J.G. & Bunch, M.B. (2007). Standard setting. London: Sage publications.

Crocker, L. & Algina, J. (1986). Introduction to classical and modern test theory. New York: CBS College Publishing.

Cumming, G. (2012). Understanding the new statistics. Effect sizes, confidence intervals, and meta-analysis. New York: Taylor & Francis Group.

Massey, A.J. & Raikes, N. (2006). Item level examiner agreement. Paper presented at the 2006 Annual Conference of the British Educational Research Association, 6–9 September 2006, University of Warwick, UK.

Ofqual. (2014). Review of Quality of Marking in Exams in A Levels, GCSEs and Other Academic Qualifications Final Report. Coventry: Office of Qualifications and Examinations Regulation.

Raikes, N., Scorey, S. & Shiell, H. (2008). Grading examinations using expert judgements from a diverse pool of judges. A paper presented to the 34th annual conference of the International Association for Educational Assessment, Cambridge, UK.

Skolinspektionen (2010). Kontrollrättning av nationella prov i grundskolan och gymnasieskolan. Dnr 2009:2796.

Skolinspektionen (2011). Lika eller olika? Omrättning av nationella prov i grundskolan och gymnasieskolan. Dnr 2010:2643.

Skolinspektionen (2012). Lika för alla? Omrättning av nationella prov i grundskolan och gymnasieskolan under tre år. Dnr 2010:2643.

[84] American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

Skolinspektionen (2013). Olikheterna är för stora. Omrättning av nationella prov i grundskolan och gymnasieskolan, 2013.

Skolinspektionen (2015). Ombedömning av nationella prov 2014: ”Processerna spelar roll”.

Skolverket. (2009). Bedömaröverensstämmelse vid bedömning av nationella prov. Dnr 2008:286.

Skolverket. (2015a). Provbetygens stabilitet, årskurs 9. Dnr 2015:4.

Skolverket. (2015b). Provpoängens tillförlitlighet. Dnr 2015:5.

Surowieccki, J. (2004). Massans vishet. Stockholm: Santérus förlag.

Suto, W.M.I. & Greatorex, J. (2008). What goes through an examiner's mind? Using verbal protocols to gain insights into the GCSE marking process. British Educational Research Journal, 34(2), 213–233.

Tisi, J., Whitehouse, G., Maughan, S. & Burdett, N. (2013). A Review of Literature on Marking Reliability Research (report for Ofqual). Slough, National Foundation for Educational Research.

www.nfer.ac.uk/publications/MARK01/MARK01_home.cfm?publicationID=948&title=A%20%20review%20of%20literature%20on%20marking%20reliability%20research

Wheadon, C. & Stockford, I. (2010). Classification accuracy and consistency in GCSE and A Level examinations offered by the Assessment and Qualifications Alliance (AQA) November 2008 to June 2009, report for Ofqual by AQA Centre for Education Research and Policy.

Ziecky, M. & Perie, M. (2006). A primer on setting cut scores on tests of educational achievement. Princeton: Educational testing service.

Appendix

Appendix 1

Här redovisas analyser för kurserna engelska B och matematik B, C och D.

Engelska B

Observerade data

Tabell 16 visar resultat på provet i engelska B: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna.

Tabell 16 I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, eng B.

Eng B							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2005	8094	4	36	47	13	13,2	13,3
2006	7453	5	39	43	13	12,9	13,0
2007	8738	4	37	44	16	13,4	13,5
2008	6195	5	41	42	11	12,7	12,6
2009	4992	4	39	46	11	13,1	13,0
2010	9371	3	37	47	12	13,3	13,2
2011	10213	3	37	46	13	13,3	13,2
2012	50905	4	38	45	13	13,2	13,2
Medel	13245	4,0	38,0	45,0	12,8	13,1	13,1
Std	15310	0,8	1,6	1,9	1,6	0,2	0,3

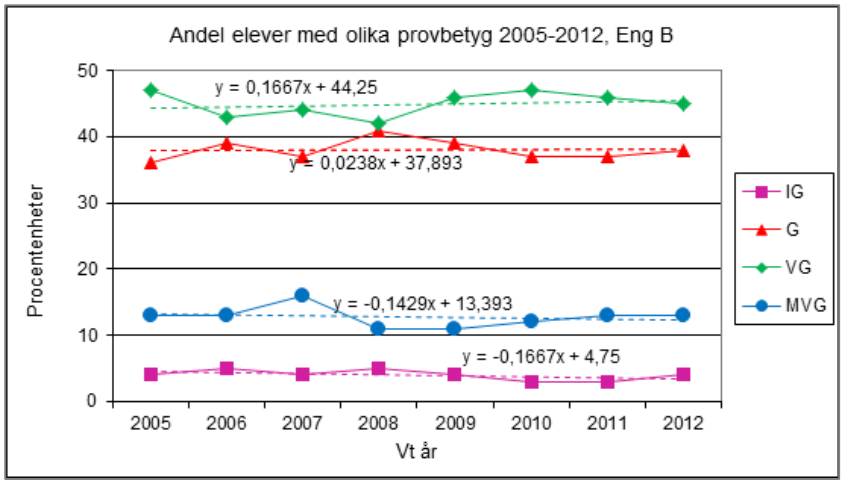
Inte heller för engelska B visar tabellen några speciella förhållanden. Dock kan man notera det stora antalet elever 2012. Det beror på att totalinsamling av provdata på gymnasienivån då hade trätt i kraft. Den stora samstämmigheten mellan dessa totaldata och stickprovs-

data från tidigare år styrker att sticksprovsdata av storleksordningen några tusen elever ger god säkerhet.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 22). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer. De senare används sedan för att beräkna *modellvärden* för de olika procentandelarna.

Figur 22 **Andel elever med olika provbetyg i engelska B vt 2005–2012 samt trendlinjer med ekvationer.**



Andelen VG ökar något medan andelen MVG och IG minskar svagt. Andelen G är i stort sett oförändrad. Vissa oregelbundenheter kan noteras de tidiga åren. Det är dock intressant att notera hur väl 2012 års resultat stämmer med trendlinjen.[85]

[85] Någon viktning efter gruppstorlek av de procentsatser som ligger till grund för skattningen av trendlinjen har inte gjorts. Alla år har samma vikt, t.ex. åren 2009 och 2012.

Tabell 17 **Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). ”k” och ”m” är parametrar för respektive betygs trendlinje, eng B.**

Eng B									k=	-0,167	0,024	0,167	-0,143	
Vt	Antal	Betyg (%)				GBP			m=	4,75	37,89	44,25	13,39	GBP
år	elever	IG	G	VG	MVG	Siris	Beräknad	Löpnr		IG	G	VG	MVG	modell
2005	8094	4	36	47	13	13,2	13,3	1		4,6	37,9	44,4	13,2	13,1
2006	7453	5	39	43	13	12,9	13,0	2		4,4	37,9	44,6	13,1	13,1
2007	8738	4	37	44	16	13,4	13,5	3		4,2	38,0	44,8	13,0	13,1
2008	6195	5	41	42	11	12,7	12,6	4		4,1	38,0	44,9	12,8	13,1
2009	4992	4	39	46	11	13,1	13,0	5		3,9	38,0	45,1	12,7	13,1
2010	9371	3	37	47	12	13,3	13,2	6		3,7	38,0	45,3	12,5	13,1
2011	10213	3	37	46	13	13,3	13,2	7		3,6	38,1	45,4	12,4	13,1
2012	50905	4	38	45	13	13,2	13,2	8		3,4	38,1	45,6	12,2	13,1
Medel	13245	4,0	38,0	45,0	12,8	13,1	13,1	Medel		4,0	38,0	45,0	12,7	13,1
Std	15310	0,8	1,6	1,9	1,6	0,2	0,3	Std		0,4	0,1	0,4	0,4	0,0

Tabell 17 visar i den högra delen vilka små förändringar som skett över tid när det gäller betygsfördelningen, vilket också framgår av den låga standardavvikelsen. Däremot sker viss variation runt trendlinjen vilket framgår av tabellens vänstra del och den högre standardavvikelsen.

Sammanfattande resultat

Resultattabellen (tabell 18) redovisar, liksom tidigare, i den vänstra delen avvikelsen i procentenheter och i den högra delen som procent av förväntad andel med det aktuella betyget.

Tabell 18 Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. ”Sum(ABS)” anger avvikelserna totalt från trendlinjen. Andelen elever med fel betyg är hälften så stor. ”Medel(ABS)” anger genomsnittlig årlig avvikelse i procent, i relation till trendlinjens värde, eng B.

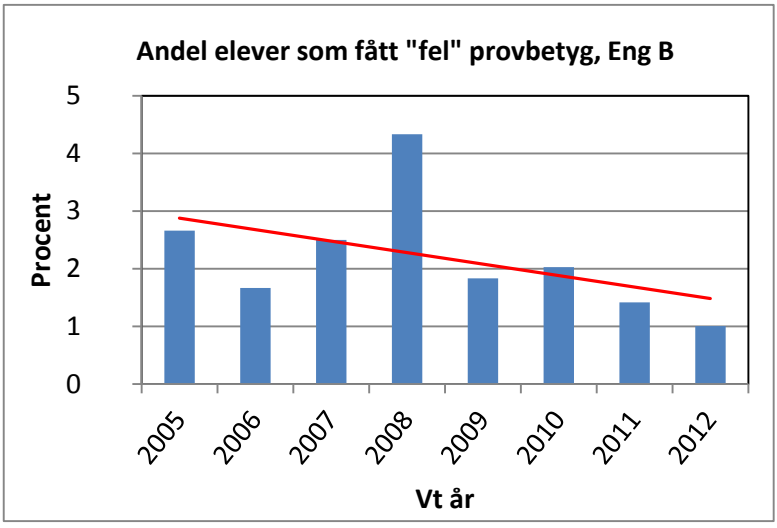
	Avvikelse(% av totalt)				
År	IG	G	VG	MVG	Sum(ABS)
2005	-0,6	-1,9	2,6	-0,2	5
2006	0,6	1,1	-1,6	-0,1	3
2007	-0,2	-1,0	-0,8	3,0	5
2008	0,9	3,0	-2,9	-1,8	9
2009	0,1	1,0	0,9	-1,7	4
2010	-0,7	-1,0	1,7	-0,5	4
2011	-0,6	-1,1	0,6	0,6	3
2012	0,6	-0,1	-0,6	0,8	2
Medel	0,0	0,0	0,0	0,0	4
Std	0,6	1,6	1,8	1,5	2

	Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG
2005	-13	-5	6	-2
2006	13	3	-4	-1
2007	-6	-3	-2	23
2008	22	8	-6	-14
2009	2	3	2	-13
2010	-20	-3	4	-4
2011	-16	-3	1	5
2012	17	0	-1	6
Medel(ABS)	14	3	3	9
Std	16	4	4	12

Här kan man notera en låg total avvikelse, 4 procentenheter (dvs. cirka 2 procent av eleverna), medan avvikelsen i procent av andelen enskilda betyg blir betydande främst för gruppen med enligt trenden förväntat betyg IG. I genomsnitt per år är gruppen 14 procent för stor eller för liten och varierar mellan att vara 20 procent för liten och 22 procent för stor i relation till trendvärdet. Även MVG visar förhållandevis stor procentuell avvikelse medan de stora grupperna G och VG visar avvikelse nästan i paritet med totalgruppen i vänstra tabelldelen.

Figur 23 illustrerar trenden som är avtagande avvikelse, dvs. en minskande andel elever får ett annat betyg än det förväntade. Eller med andra ord – stabiliteten ökar.

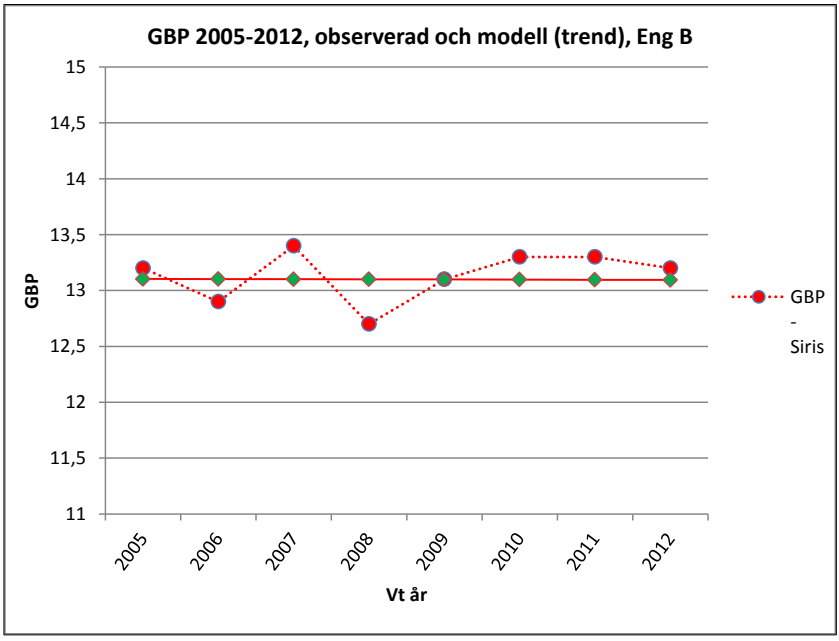
Figur 23 Andel elever som fått från trenden avvikande provbetyg i engelska B.



Genomsnittlig betygspoäng

Diagrammet över GBP styrker bilden av stabil betygsnivå under perioden och pekar på vissa variationer, främst 2007 och 2008. Aven samstämmigheten mellan 2012 års resultat och trendlinjen illustreras (figur 24).

Figur 24 Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (trendlinjen), engelska B.



Matematik B

Observerade data

Tabell 19 visar resultat på provet i matematik B: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna.

Tabell 19 I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, ma B.

Ma B							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2000	2654	27	45	21	7	9,1	9,1
2001	5019	26	38	24	12	9,8	9,8
2002	4072	25	38	21	16	10,2	10,2
2003	4315	28	38	22	12	9,6	9,5
2004	5572	33	41	21	6	8,4	8,5
2005	5638	28	37	23	12	9,5	9,6
2006	5088	29	41	18	13	9,3	9,4
2007	6142	27	38	24	11	9,6	9,6
2008	5014	33	42	18	7	8,3	8,3
2009	4681	36	37	21	6	8,1	8,1
2010	7254	26	36	29	10	9,9	10,0
2011	6972	32	43	21	4	8,3	8,3
Medel	5202	29,2	39,5	21,9	9,7	9,2	9,2
Std	1260	3,5	2,8	2,9	3,6	0,7	0,7

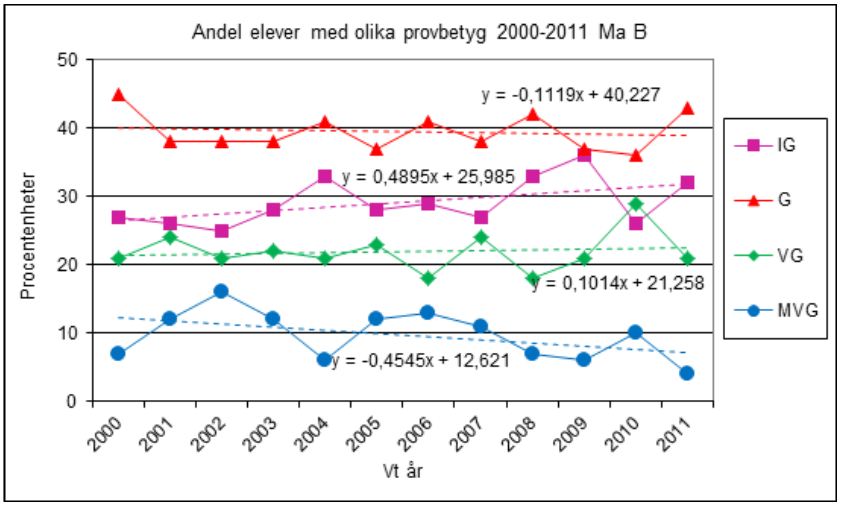
Man kan lägga märke till att stickproven för ma B i genomsnitt är ungefär hälften så stora som för ma A. Vidare ser vi att andelen elever med provbetyget IG är stort och i genomsnitt större än för ma A.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 25). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer. De senare används sedan för att beräkna modellvärden för de olika procentandelarna.

I figur 25 kan man se att variationen mellan olika år när det gäller hur stor andel av eleverna som har respektive betyg är betydande.

Figur 25 Andel elever med olika provbetyg i matematik B vt 2000–2011 samt trendlinjer med ekvationer.



IG är det näst vanligaste betyget och andelen är ökande under perioden. Tabell 20 visar siffrorna.

Tabell 20 Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). "k" och "m" är parametrar för respektive betygs trendlinje, ma B.

Ma B														
Vt	Antal	Betyg (%)				GBP			k=	0,49	-0,112	0,101	-0,455	
									m=	25,99	40,23	21,26	12,62	GBP
år	elever	IG	G	VG	MVG	Siris	Beräknad		Löpnr	IG	G	VG	MVG	modell
2000	2654	27	45	21	7	9,1	9,1		1	26,5	40,1	21,4	12,2	9,6
2001	5019	26	38	24	12	9,8	9,8		2	27,0	40,0	21,5	11,7	9,6
2002	4072	25	38	21	16	10,2	10,2		3	27,5	39,9	21,6	11,3	9,5
2003	4315	28	38	22	12	9,6	9,5		4	28,0	39,8	21,7	10,8	9,4
2004	5572	33	41	21	6	8,4	8,5		5	28,4	39,7	21,8	10,3	9,3
2005	5638	28	37	23	12	9,5	9,6		6	28,9	39,6	21,9	9,9	9,2
2006	5088	29	41	18	13	9,3	9,4		7	29,4	39,4	22,0	9,4	9,1
2007	6142	27	38	24	11	9,6	9,6		8	29,9	39,3	22,1	9,0	9,0
2008	5014	33	42	18	7	8,3	8,3		9	30,4	39,2	22,2	8,5	9,0
2009	4681	36	37	21	6	8,1	8,1		10	30,9	39,1	22,3	8,1	8,9
2010	7254	26	36	29	10	9,9	10,0		11	31,4	39,0	22,4	7,6	8,8
2011	6972	32	43	21	4	8,3	8,3		12	31,9	38,9	22,5	7,2	8,7
Medel	5202	29,2	39,5	21,9	9,7	9,2	9,2		Medel	29,2	39,5	21,9	9,7	9,2
Std	1260	3,5	2,8	2,9	3,6	0,7	0,7		Std	1,8	0,4	0,4	1,6	0,3

Andelen elever med G och VG ändras lite under perioden. Däremot ökar andelen med IG och minskar andelen med MVG med knappt 5 procentenheter (tabell 20, högra delen).

Sammanfattande resultat

Resultattabellen (tabell 21) redovisar, liksom tidigare, i den vänstra delen avvikelsen i procentenheter och i den högra delen som procent av förväntad andel med det aktuella betyget.

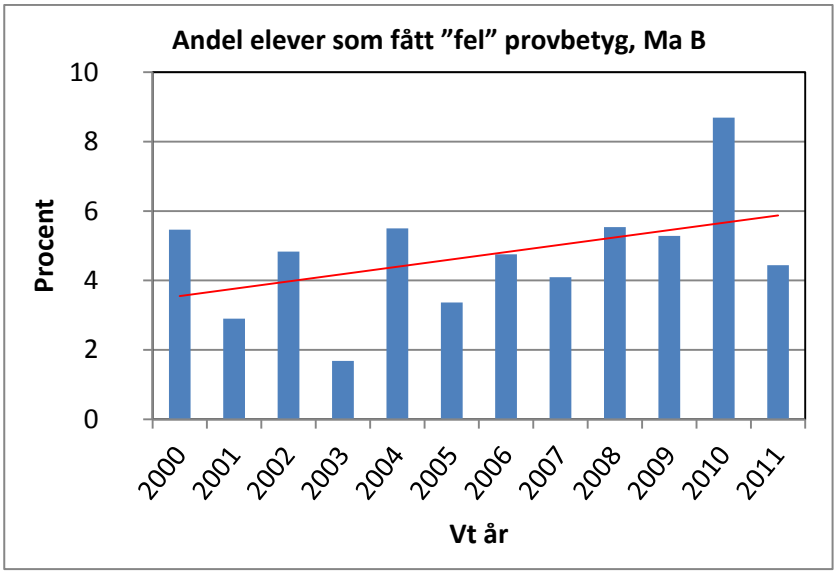
Tabell 21 **Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. "Sum(ABS)" anger avvikelserna totalt från trendlinjen. Andelen elever med fel betyg är hälften så stor. "Medel(ABS)" anger genomsnittlig årlig avvikelse i procent, i relation till trendlinjens värde, ma B.**

	Avvikelse(% av totalt)						Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG	Sum(ABS)	År	IG	IG	IG	IG
2000	0,5	4,9	-0,4	-5,2	11	2000	2	12	-2	-42
2001	-1,0	-2,0	2,5	0,3	6	2001	-4	-5	12	2
2002	-2,5	-1,9	-0,6	4,7	10	2002	-9	-5	-3	42
2003	0,1	-1,8	0,3	1,2	3	2003	0	-4	2	11
2004	4,6	1,3	-0,8	-4,3	11	2004	16	3	-4	-42
2005	-0,9	-2,6	1,1	2,1	7	2005	-3	-6	5	21
2006	-0,4	1,6	-4,0	3,6	10	2006	-1	4	-18	38
2007	-2,9	-1,3	1,9	2,0	8	2007	-10	-3	9	22
2008	2,6	2,8	-4,2	-1,5	11	2008	9	7	-19	-18
2009	5,1	-2,1	-1,3	-2,1	11	2009	17	-5	-6	-26
2010	-5,4	-3,0	6,6	2,4	17	2010	-17	-8	30	31
2011	0,1	4,1	-1,5	-3,2	9	2011	0	11	-7	-44
Medel	0,0	0,0	0,0	0,0	9	Medel(ABS)	7	6	9	28
Std	3,0	2,8	2,9	3,2	3	Std	10	7	13	33

Även för ma B kan man konstatera att den genomsnittliga avvikelsen är betydande. Av samtliga provbetyg är 9 procent andra än de förväntade (cirka 4,5 procent av eleverna får således fel provbetyg). Den betygsgrupp som framför allt avviker (i relativ mening) är MVG. Där är gruppen elever med provbetyget MVG i genomsnitt 28 procent för stor eller 28 procent för liten i relation till trendvärdet.

För matematik B är trenden ökande, dvs. avvikelsen har ökat över tid, vilket alltså betyder att stabiliteten har minskat. Figur 26 ger bilden.

Figur 26 **Andel elever som fått från trenden awikande provbetyg i matematik B.**

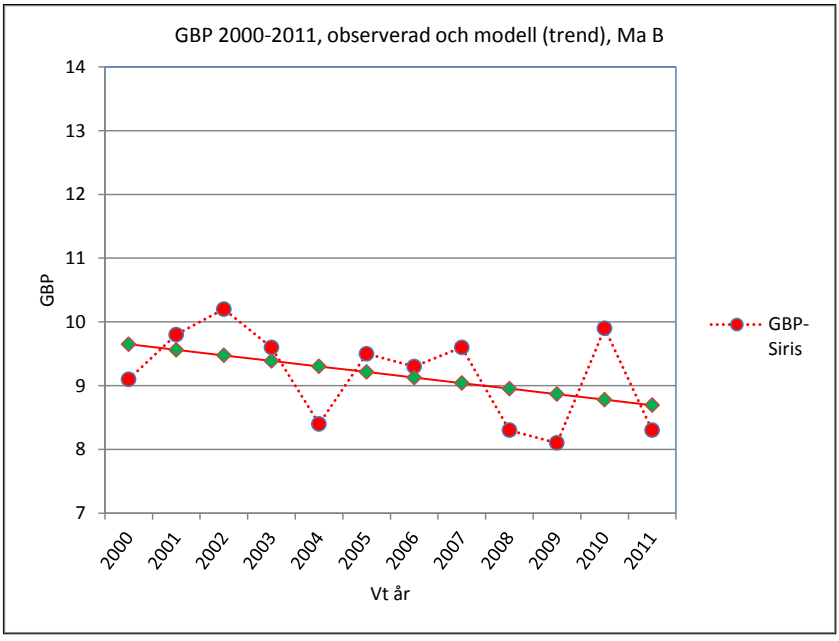


Det är, som framgår av figuren, främst resultatet 2010 som ger bilden av ökande avvikelse. I övrigt verkar förändringen över tid liten.

Genomsnittlig betygspoäng

Den genomsnittliga betygspoängen styrker den tidigare bilden.

Figur 27 Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (trendlinjen) matematik B.



Matematik C

Observerade data

Tabell 22 visar resultat på provet i matematik C: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna.

Tabell 22 I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, ma C.

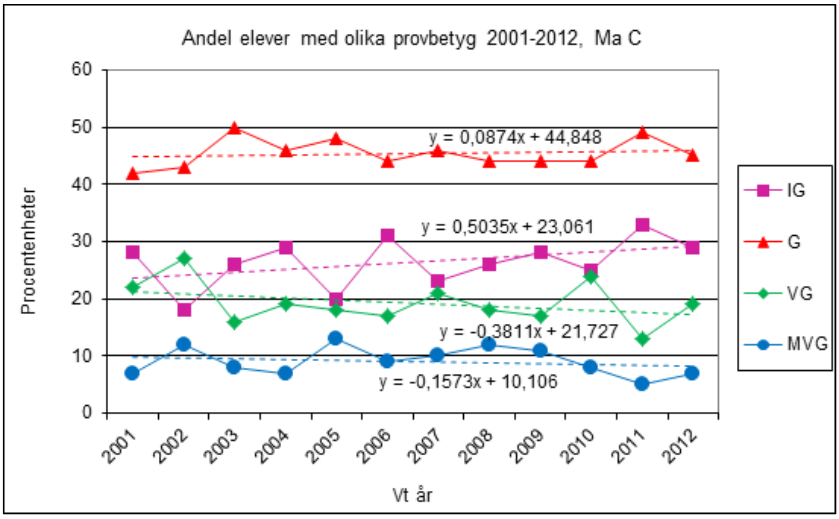
Ma C							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2001	2538	28	42	22	7	9	8,9
2002	1148	18	43	27	12	10,8	10,8
2003	1009	26	50	16	8	9,1	9,0
2004	1459	29	46	19	7	8,8	8,9
2005	1644	20	48	18	13	10,2	10,1
2006	1986	31	44	17	9	8,6	8,8
2007	2139	23	46	21	10	9,7	9,8
2008	1385	26	44	18	12	9,6	9,5
2009	1497	28	44	17	11	9,2	9,2
2010	2801	25	44	24	8	9,5	9,6
2011	2300	33	49	13	5	7,8	7,9
2012	12264	29	45	19	7	8,7	8,8
Medel	2681	26,3	45,4	19,3	9,1	9,3	9,2
Std	3069	4,4	2,5	3,8	2,5	0,8	0,8

Här framgår att stickprovsstorleken krymper betydligt jämfört med ma B och särskilt ma A. Medelantalet är knappt 2 700 elever, men då drar det höga antalet resultat 2012 upp medelvärdet (och standardavvikelsen). Utan 2012 är genomsnittet cirka 1 800 elever. Det man i övrigt kan notera är att även för ma C är andelen elever med provbetyget IG högt.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 28). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer. De senare används sedan för att beräkna modellvärden för de olika procentandelarna.

Figur 28 **Andel elever med olika provbetyg i matematik C vt 2001–2012 samt trendlinjer med ekvationer.**



Den mest påtagliga förändringen är ökningen av IG med cirka 0,5 procentenheter per år (ungefär samma som för ma B). Andelen VG avtar liksom i mindre grad andelen MVG. Andelen G är i stort sett oförändrad över tid. Variationen runt trendlinjerna är dock som framgår betydande.

Tabell 23 **Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). "k" och "m" är parametrar för respektive betygs trendlinje, ma C.**

Ma C							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2001	2538	28	42	22	7	9	8,9
2002	1148	18	43	27	12	10,8	10,8
2003	1009	26	50	16	8	9,1	9,0
2004	1459	29	46	19	7	8,8	8,9
2005	1644	20	48	18	13	10,2	10,1
2006	1986	31	44	17	9	8,6	8,8
2007	2139	23	46	21	10	9,7	9,8
2008	1385	26	44	18	12	9,6	9,5
2009	1497	28	44	17	11	9,2	9,2
2010	2801	25	44	24	8	9,5	9,6
2011	2300	33	49	13	5	7,8	7,9
2012	12264	29	45	19	7	8,7	8,8
Medel	2681	26,3	45,4	19,3	9,1	9,3	9,2
Std	3069	4,4	2,5	3,8	2,5	0,8	0,8

k=	0,504	0,087	-0,381	-0,157	
m=	23,06	44,85	21,73	10,11	GBP
Löpnr	IG	G	VG	MVG	modell
1	23,6	44,9	21,3	10,0	9,7
2	24,1	45,0	21,0	9,8	9,6
3	24,6	45,1	20,6	9,6	9,5
4	25,1	45,2	20,2	9,5	9,4
5	25,6	45,3	19,8	9,3	9,4
6	26,1	45,4	19,4	9,2	9,3
7	26,6	45,5	19,1	9,0	9,2
8	27,1	45,5	18,7	8,9	9,1
9	27,6	45,6	18,3	8,7	9,0
10	28,1	45,7	17,9	8,5	9,0
11	28,6	45,8	17,5	8,4	8,9
12	29,1	45,9	17,2	8,2	8,8
Medel	26,3	45,4	19,3	9,1	9,2
Std	1,8	0,3	1,4	0,6	0,3

Sammanfattande resultat

Resultattabellen (tabell 24) redovisar, liksom tidigare, i den vänstra delen avvikelsen i procentenheter och i den högra delen som procent av förväntad andel med det aktuella betyget.

Tabell 24 **Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. ”Sum(ABS)” anger avvikelserna totalt från trendlinjen. Andelen elever med fel betyg är hälften så stor. ”Medel(ABS)” anger genomsnittlig årlig avvikelse i procent, i relation till trendlinjens värde, ma C.**

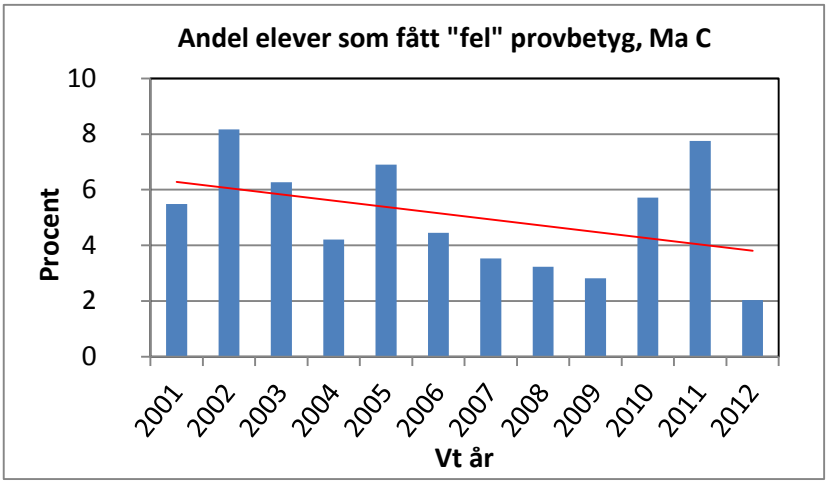
	Avvikelse(% av totalt)				
År	IG	G	VG	MVG	Sum(ABS)
2001	4,4	-2,9	0,7	-3,0	11
2002	-6,1	-2,0	6,0	2,2	16
2003	1,4	4,9	-4,6	-1,6	13
2004	3,9	0,8	-1,2	-2,5	8
2005	-5,6	2,7	-1,8	3,7	14
2006	4,9	-1,4	-2,4	-0,2	9
2007	-3,6	0,5	1,9	1,0	7
2008	-1,1	-1,5	-0,7	3,1	6
2009	0,4	-1,6	-1,3	2,3	6
2010	-3,1	-1,7	6,1	-0,5	11
2011	4,4	3,2	-4,5	-3,4	16
2012	-0,1	-0,9	1,8	-1,2	4
Medel	0,0	0,0	0,0	0,0	10
Std	4,0	2,4	3,5	2,4	4

	Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG
2001	19	-7	3	-30
2002	-25	-4	29	22
2003	6	11	-22	-17
2004	16	2	-6	-26
2005	-22	6	-9	39
2006	19	-3	-13	-2
2007	-13	1	10	11
2008	-4	-3	-4	36
2009	1	-4	-7	26
2010	-11	-4	34	-6
2011	15	7	-26	-40
2012	0	-2	11	-15
Medel(ABS)	13	4	14	23
Std	16	5	18	27

Även för ma C är den genomsnittliga avvikelsen stor, cirka 10 procent, dvs. ungefär 5 procent av eleverna får ett annat provbetyg än det förväntade. För enskilda betyg är den relativa variationen avsevärd för samtliga betyg utom G.

Trenden i avvikelsen är avtagande (figur 29), men till stor del beror det på hög avvikelse de första åren. Även under senare år (2010 och 2011) har avvikelsen varit betydande.

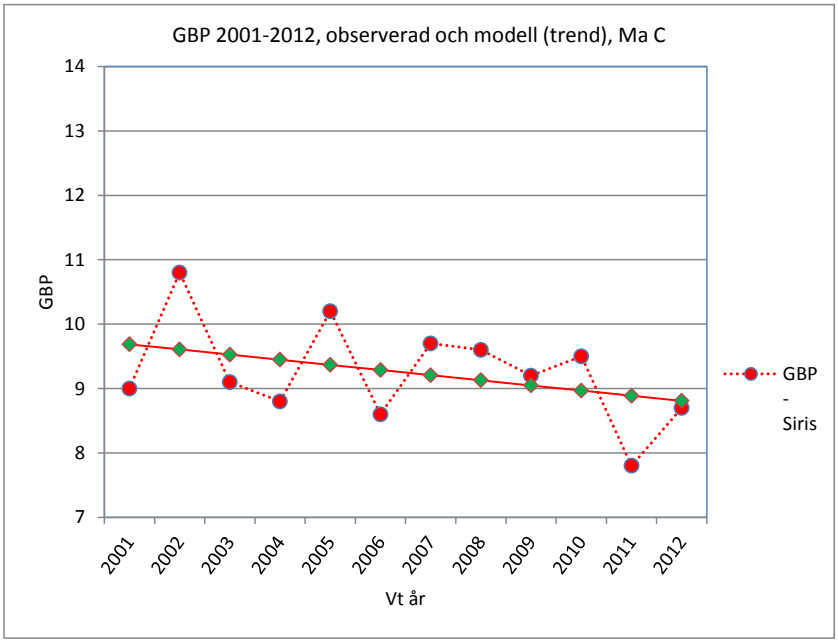
Figur 29 Andel elever som fått från trenden avvikande provbetyg i matematik C.



Genomsnittlig betygspoäng

Den genomsnittliga betygspoängen stärker den redan givna bilden.

Figur 30 Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (trendlinjen) matematik C.



Matematik D

Observerade data

Tabell 25 visar resultat på provet i matematik D: antal elevresultat som redovisats, andelen elever med respektive betyg uttryckt i procent samt GBP enligt Siris och beräknad utifrån de angivna procentsatserna.

Tabell 25 I Siris rapporterat antal elever, andel elever med olika betyg och GBP samt utifrån angivna betygsandelar beräknad GBP, ma D.

Ma D							
Vt	Antal	Betyg (%)				GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad
2001	2632	15	39	36	11	11,4	11,5
2002	1233	10	52	29	10	11,4	11,6
2003	1128	15	41	24	20	11,7	11,7
2004	1265	24	36	23	17	10,5	10,5
2005	1058	18	39	25	18	11,3	11,3
2006	1139	18	39	28	16	11,2	11,3
2007	1175	14	40	25	21	12	12,0
2008	943	18	48	25	10	10,4	10,6
2009	712	11	49	22	18	11,8	11,8
2010	1617	15	35	27	23	12,2	12,2
2011	1304	16	44	22	18	11,3	11,3
2012	6861	20	42	20	18	10,8	10,8
2013	1493	19	49	21	12	10,3	10,5
Medel	1735	16,4	42,5	25,2	16,3	11,3	11,3
Std	1607	3,7	5,4	4,2	4,3	0,6	0,6

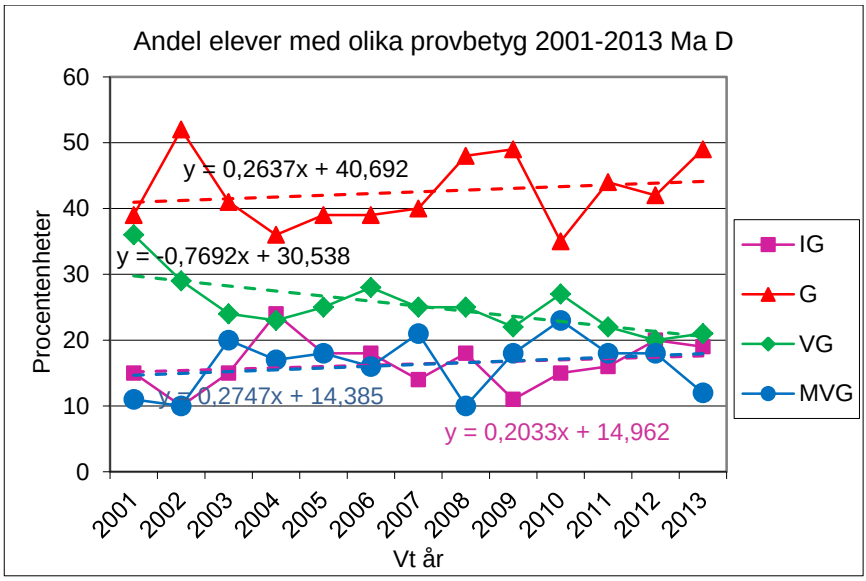
Här har stickprovet minskat ytterligare i storlek. I genomsnitt är det drygt 1 700 elever, men då drar totalinsamlingen 2012 upp värdet. De flesta stickproven ligger på lite drygt tusen elever. Man kan också lägga märke till att andelen elever med IG har sjunkit i relation till resultaten på övriga kursprov i matematik. Samtidigt ligger andelen fortfarande högt över motsvarande andel för kursproven i svenska och engelska.

Data i diagramform

En tydligare bild fås om tabellen visas i diagramform (figur 31). Diagrammet används också för att generera trendlinjer och motsvarande ekvationer. De senare används sedan för att beräkna modellvärden för de olika procentandelarna.

Diagrammet för matematik D påminner om diagrammen för övriga kursprov i matematik: stora variationer och för vissa betyg tämligen kraftig förändring över tid.

Figur 31 **Andel elever med olika provbetyg i matematik D vt 2001–2013 samt trendlinjer med ekvationer.**



Andelen elever med VG sjunker med cirka 0,8 procentenheter per år och variationen mellan olika år är betydande för i synnerhet G, men stor för samtliga betygssteg.

Tabell 26 **Antal elever i stickprovet och andel elever med olika provbetyg (vänstra delen), samt andel elever med olika provbetyg enligt trendlinjen (den högra delen av tabellen). ”k” och ”m” är parametrar för respektive betygs trendlinje, ma D.**

Ma D														
Vt	Antal	Betyg (%)				GBP		k=	0,203	0,264	-0,769	0,275		
								m=	14,96	40,69	30,54	14,39	GBP	
år	elever	IG	G	VG	MVG	Siris	Beräknad	Löpnr	IG	G	VG	MVG	modell	
2001	2632	15	39	36	11	11,4	11,5	1	15,2	41,0	29,8	14,7	11,5	
2002	1233	10	52	29	10	11,4	11,6	2	15,4	41,2	29,0	14,9	11,5	
2003	1128	15	41	24	20	11,7	11,7	3	15,6	41,5	28,2	15,2	11,4	
2004	1265	24	36	23	17	10,5	10,5	4	15,8	41,7	27,5	15,5	11,4	
2005	1058	18	39	25	18	11,3	11,3	5	16,0	42,0	26,7	15,8	11,4	
2006	1139	18	39	28	16	11,2	11,3	6	16,2	42,3	25,9	16,0	11,3	
2007	1175	14	40	25	21	12	12,0	7	16,4	42,5	25,2	16,3	11,3	
2008	943	18	48	25	10	10,4	10,6	8	16,6	42,8	24,4	16,6	11,3	
2009	712	11	49	22	18	11,8	11,8	9	16,8	43,1	23,6	16,9	11,2	
2010	1617	15	35	27	23	12,2	12,2	10	17,0	43,3	22,9	17,1	11,2	
2011	1304	16	44	22	18	11,3	11,3	11	17,2	43,6	22,1	17,4	11,2	
2012	6861	20	42	20	18	10,8	10,8	12	17,4	43,9	21,3	17,7	11,1	
2013	1493	19	49	21	12	10,3	10,5	13	17,6	44,1	20,5	18,0	11,1	
Medel	1735	16,4	42,5	25,2	16,3	11,3	11,3	Medel	16,4	42,5	25,2	16,3	11,3	
Std	1607	3,7	5,4	4,2	4,3	0,6	0,6	Std	0,8	1,0	3,0	1,1	0,1	

Sammanfattande resultat

Resultattabellen (tabell 27) redovisar, liksom tidigare, i den vänstra delen avvikelsen i procentenheter och i den högra delen som procent av förväntad andel med det aktuella betyget.

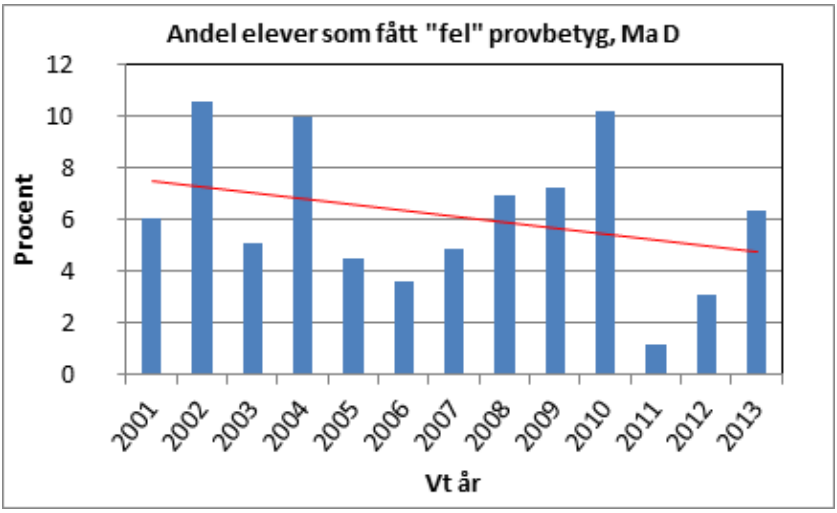
Tabell 27 Skillnad i andel elever med observerat provbetyg och provbetyg enligt modellen. Den vänstra delen anger andel av samtliga, den högra delen andel av de som enligt modellen förväntas ha respektive betyg. ”Sum(ABS)” anger avvikelserna totalt från trendlinjen. Andelen elever med fel betyg är hälften så stor. ”Medel(ABS)” anger genomsnittlig årlig avvikelse i procent, i relation till trendlinjens värde, ma D.

	Betygsskillnad(%)						Avvikelse(% av respektive betyg)			
År	IG	G	VG	MVG	Sum(ABS)	År	IG	G	VG	MVG
2001	-0,2	-2,0	6,2	-3,7	12	2001	-1	-5	21	-25
2002	-5,4	10,8	0,0	-4,9	21	2002	-35	26	0	-33
2003	-0,6	-0,5	-4,2	4,8	10	2003	-4	-1	-15	31
2004	8,2	-5,7	-4,5	1,5	20	2004	52	-14	-16	10
2005	2,0	-3,0	-1,7	2,2	9	2005	13	-7	-6	14
2006	1,8	-3,3	2,1	0,0	7	2006	11	-8	8	0
2007	-2,4	-2,5	-0,2	4,7	10	2007	-15	-6	-1	29
2008	1,4	5,2	0,6	-6,6	14	2008	9	12	3	-40
2009	-5,8	5,9	-1,6	1,1	14	2009	-34	14	-7	7
2010	-2,0	-8,3	4,2	5,9	20	2010	-12	-19	18	34
2011	-1,2	0,4	-0,1	0,6	2	2011	-7	1	0	3
2012	2,6	-1,9	-1,3	0,3	6	2012	15	-4	-6	2
2013	1,4	4,9	0,5	-6,0	13	2013	8	11	2	-33
Medel	0,0	0,0	0,0	0,0	12	Medel(ABS)	17	10	8	20
Std	3,6	5,3	3,0	4,1	6	Std	23	13	11	25

Återigen är avvikelsen betydande, 12 procent total avvikelse i genomsnitt innebär att 6 procent av samtliga elever tilldelas fel provbetyg utifrån angivna betygsgränser. För de enskilda betygsstegen är också avvikelserna stora. För gruppen MVG är i genomsnitt andelen som fått betyget 20 procent för stor eller 20 procent för liten. För övriga betygssteg är den mindre men ändå råder betydande variation mellan olika år, särskilt gäller detta IG och MVG.

Avvikelsen är något avtagande (figur 32), men bilden är spretig.

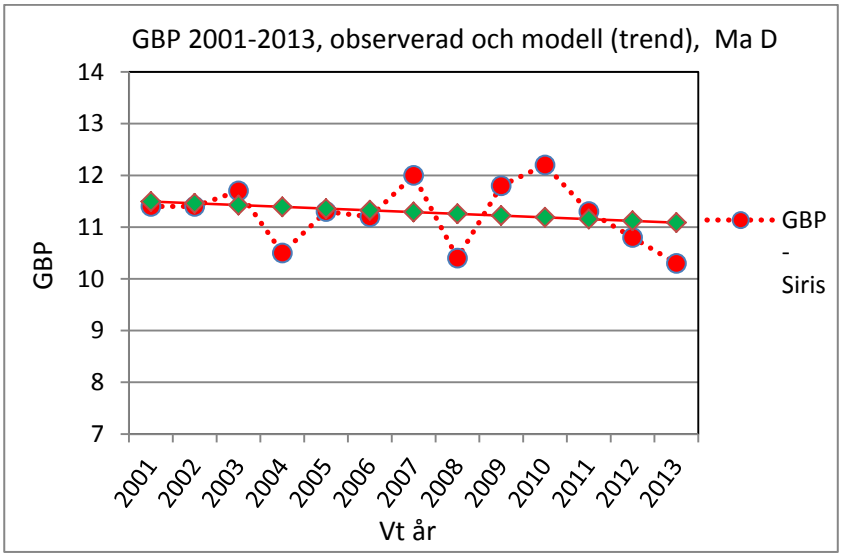
Figur 32 Andel elever som fått från trenden avvikande provbetyg i matematik D.



Genomsnittlig betygspoäng

Den genomsnittliga betygspoängen indikerar inte att provbetygen skulle ha blivit mer stabila.

Figur 33 Observerad genomsnittlig betygspoäng och genomsnittlig betygspoäng enligt den modellanpassade betygsfördelningen (trendlinjen) matematik D.



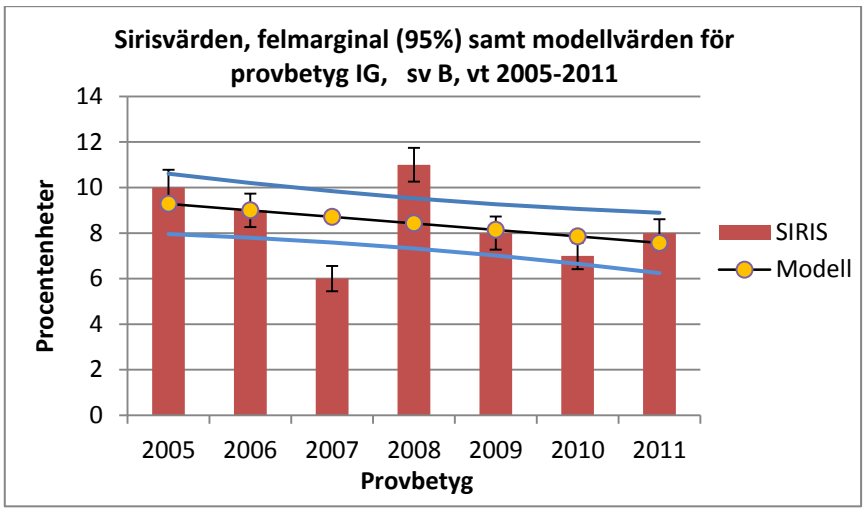
Appendix 2

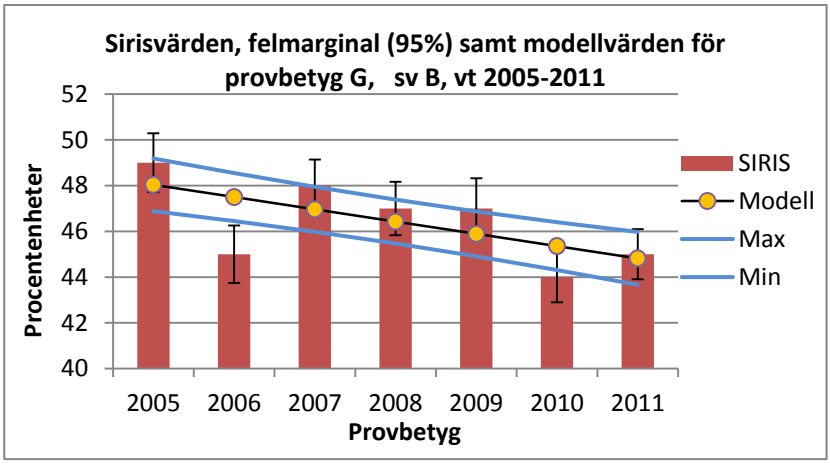
Standardfel för sv B, eng A och ma D

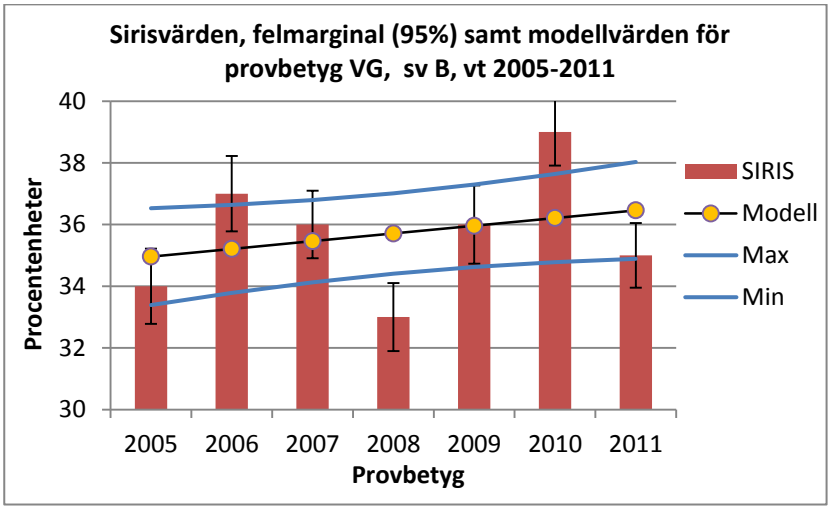
Här redovisas som jämförelse med tidigare redovisade standardfel för ma A motsvarande standardfel för proven i sv B, eng A och eng B.

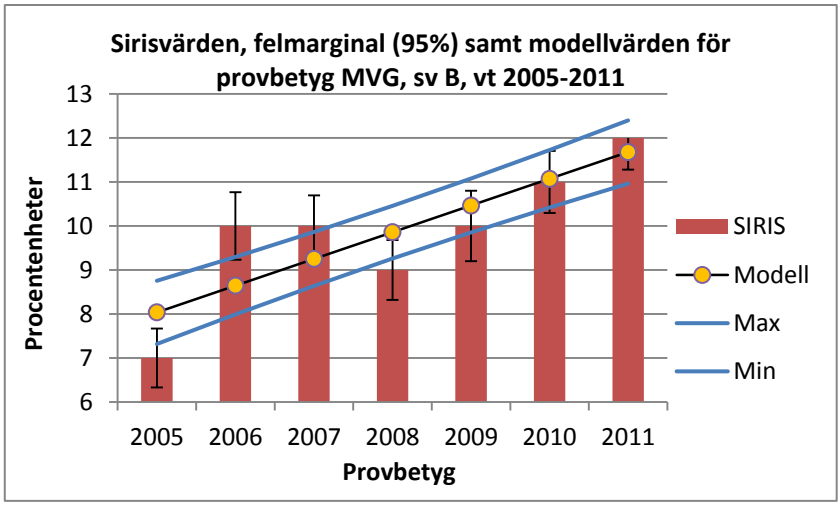
Svenska B

Figur 34 Andel elever med olika redovisade provbetyg olika år för sv B med anpassade trendlinjer samt felmarginaler (95 procent). Om den redovisade andelen (Sirisvärdet) ligger utanför de blå linjerna kan avvikelsen mellan förväntad andel och redovisad andel anses signifikant. (Obs. olika skalor på y-axeln.)




Sirisvärden, felmarginal (95%) samt modellvärden för provbetyg G, sv B, vt 2005-2011
Procentenheter
52
50
48
46
44
42
40
2005
2006
2007
2008
2009
2010
2011
Provbetyg
SIRIS
Modell
Max
Min

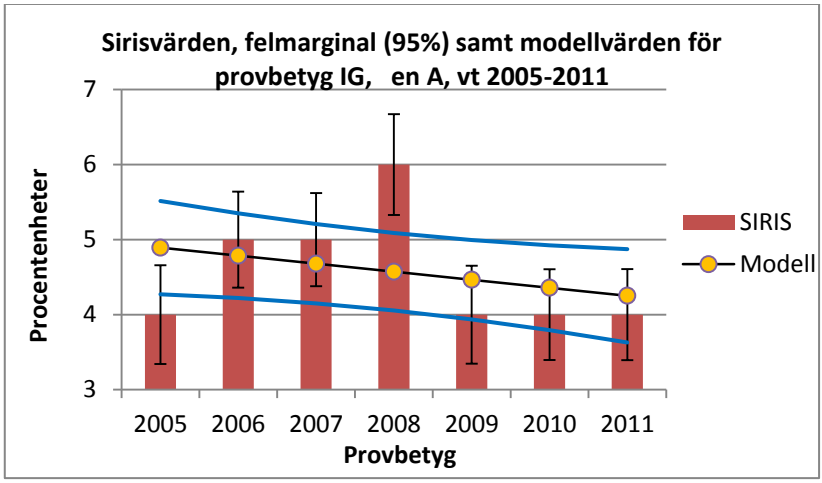

Sirisvärden, felmarginal (95%) samt modellvärden för provbetyg VG, sv B, vt 2005-2011
Procentenheter
40
38
36
34
32
30
2005
2006
2007
2008
2009
2010
2011
Provbetyg
SIRIS
Modell
Max
Min

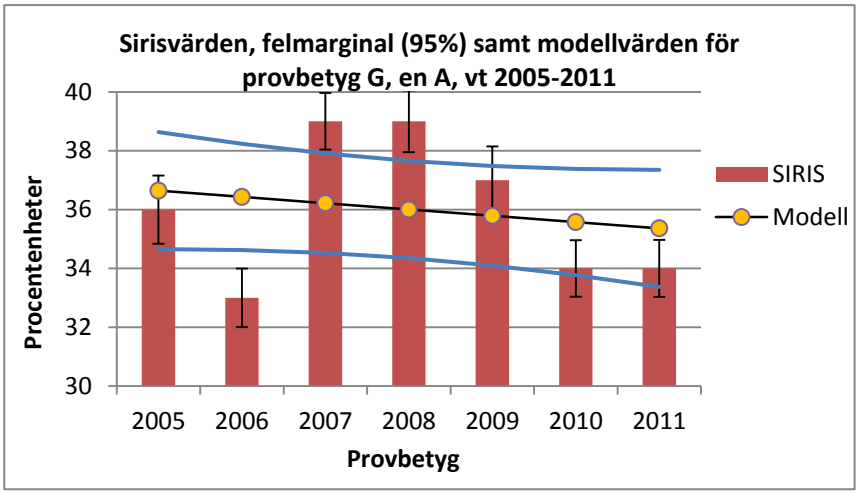


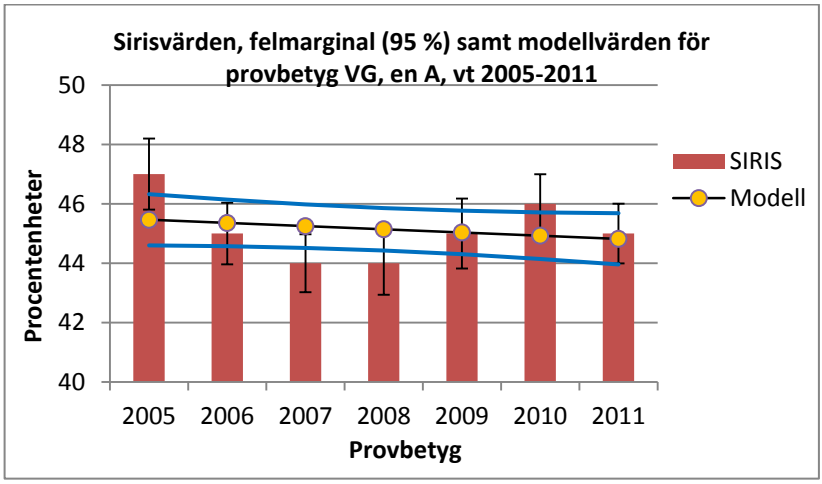
Bilderna visar att det är tveksamt om avvikelserna för sv B kan ses som statistisk signifikanta. Resultaten får tolkas med försiktighet.

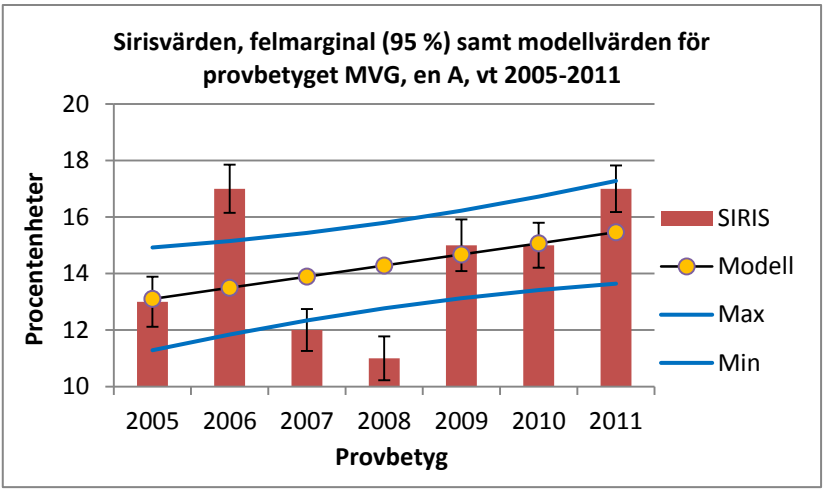
Engelska A

Figur 35 Andel elever med olika redovisade provbetyg olika år för eng A med anpassade trendlinjer samt felmarginaler (95 procent). Om den redovisade andelen (Sirisvärdet) ligger utanför de blå linjerna kan avvikelsen mellan förväntad andel och redovisad andel anses signifikant. (Obs. olika skalor på y-axeln.)




Sirisvärden, felmarginal (95%) samt modellvärden för provbetyg G, en A, vt 2005-2011
Procentenheter
40
38
36
34
32
30
2005
2006
2007
2008
2009
2010
2011
Provbetyg
SIRIS
Modell

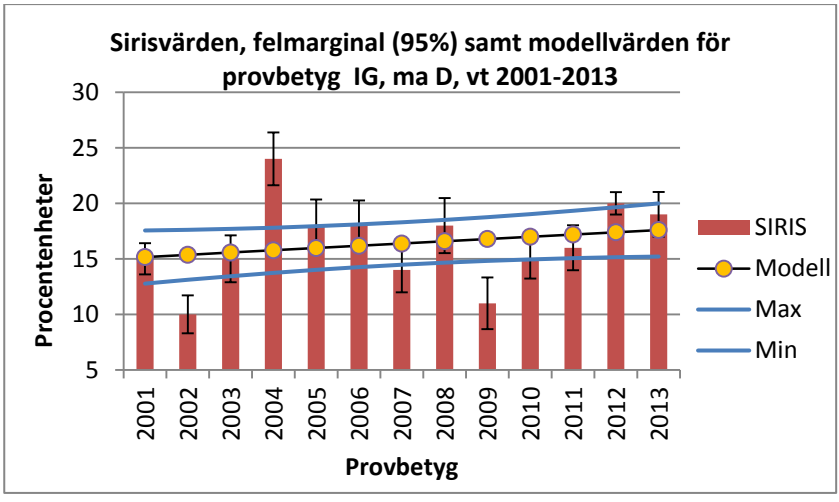

Sirisvärden, felmarginal (95 %) samt modellvärden för provbetyg VG, en A, vt 2005-2011
Procentenheter
50
48
46
44
42
40
2005
2006
2007
2008
2009
2010
2011
Provbetyg
SIRIS
Modell

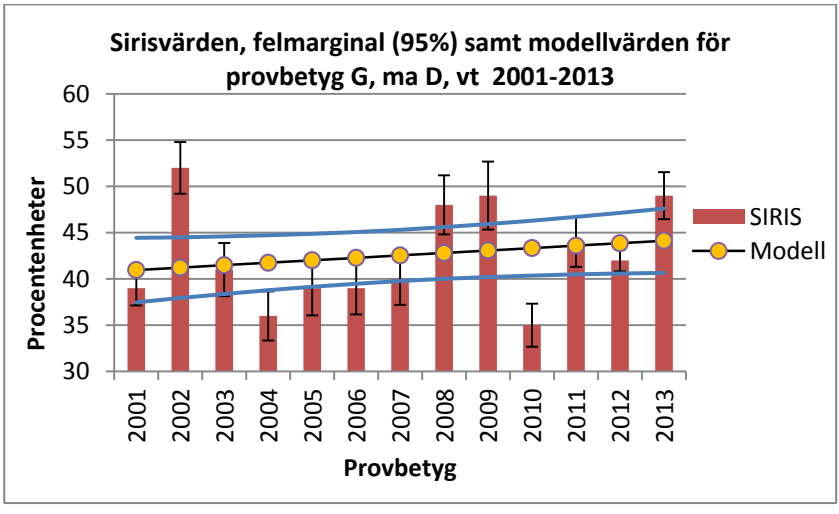


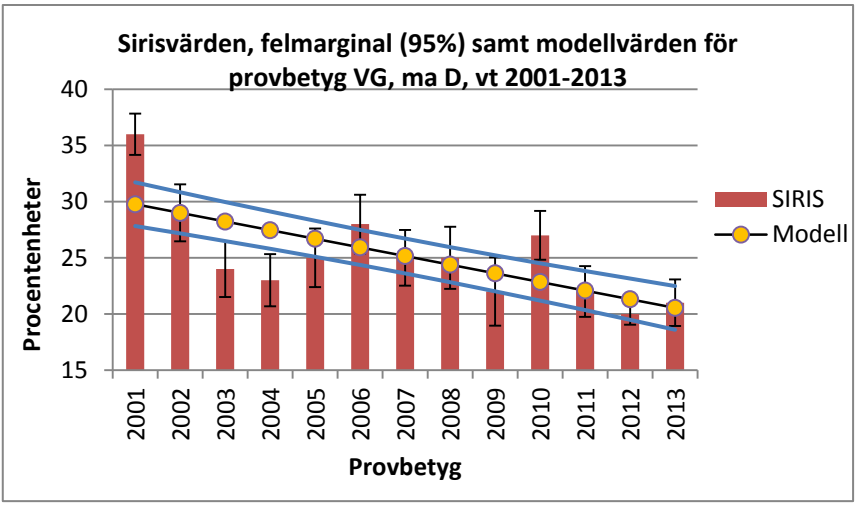
För såväl proven i svenska B som i engelska A gäller att det är tveksamt om avvikelserna kan betraktas som statistiskt signifikanta. Betydlig färre Sirisvärden än för matematik A faller utanför konfidenslinjerna för trendlinjen för respektive betyg. Resultaten för såväl svenska B som engelska A bör således tolkas varsamt.

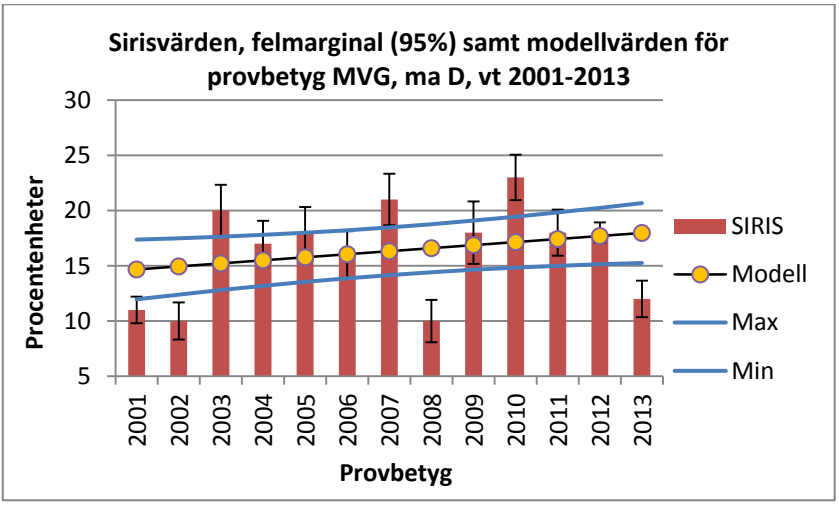
Matematik D

Figur 36 Andel elever med olika redovisade provbetyg olika år för ma D med anpassade trendlinjer samt felmarginaler (95 procent). Om den redovisade andelen (Sirisvärdet) ligger utanför de blå linjerna kan avvikelsen mellan förväntad andel och redovisad andel anses signifikant. (Obs. olika skalor på y-axeln.)









För såväl proven i svenska B som i engelska A gäller att det är tveksamt om avvikelserna kan betraktas som statistiskt signifikanta. För matematik D är avvikelserna mer tillförlitliga än för svenska B och engelska A.

Appendix 3

Provbetyg för ma B, vt 2011.[86]

Kravgränser	
Detta prov kan ge maximalt 45 poäng, varav 25 g-poäng.	
Undre gräns för provbetyget	
Godkänt:	13 poäng.
Väl godkänt:	25 poäng varav minst 6 vg-poäng.
Mycket väl godkänt:	25 poäng varav minst 13 vg-poäng. Eleven ska dessutom ha visat prov på minst tre *olika* MVG-kvaliteter av de fyra MVG-kvaliteter som är möjliga att visa i detta prov.

I tabellen nedan listas antalet elever efter hur många totalpoäng de har på provet (vertikalt) och efter vilket provbetyg de tilldelats av bedömande lärare (horisontellt). De röda siffrorna indikerar troliga felregistreringar. Tabellen är underlag för figur 14.

[86] www5.edusci.umu.se/np/np-prov/B-kursprov-vt11.pdf

Belägg (poäng)	Provbet (antal elever)				Total	Procent
	IG	G	VG	MVG		
0	65	2	1		68	1,7
1	48				48	1,2
2	82				82	2,0
3	71				71	1,8
4	83				83	2,1
IG 5	82				82	2,0
6	80	2			82	2,0
7	99	1			100	2,5
8	104				104	2,6
9	106				106	2,6
10	133	1			134	3,3
11	120	2			122	3,0
12	99	13			112	2,8
13		192			192	4,8
14	1	141			142	3,5
15		146			146	3,6
16		133			133	3,3
17		138			138	3,4
G 18		126			126	3,1
19		143			143	3,6
20		124	1		125	3,1
21		136	1		137	3,4
22		113	3		118	2,9
23		110	4		114	2,8
24		105	4		109	2,7
25		57	78		135	3,4
26		48	77		125	3,1
27		27	73		100	2,5
28		11	100		111	2,8
29		5	96		101	2,5
VG 30		3	94		97	2,4
31			81	1	82	2,0
32			69	2	71	1,8
33			66	2	68	1,7
34			50	6	56	1,4
35		1	30	12	43	1,1
36			26	16	42	1,0
37			14	17	31	0,8
38			9	14	23	0,6
39			4	24	28	0,7
40			4	21	25	0,6
(MVG) 41			1	14	15	0,4
42				16	16	0,4
43				8	8	0,2
44				7	7	0,2
45				2	2	0,0
46				1	1	0,0
Totalt	1175	1780	886	163	4004	100

Statens offentliga utredningar 2016

Kronologisk förteckning

1. Statens bredbandsinfrastruktur som resurs. N.
2. Effektiv vård. S.
3. Höghastighetsjärnvägens finansiering och kommersiella förutsättningar. N.
4. Politisk information i skolan – ett led i demokratiuppdraget. U.
5. Låt fler forma framtiden! Del A + B. Ku.
6. Framtid sökes – Slutredovisning från den nationella samordnaren för utsatta EU-medborgare. S.
7. Integritet och straffskydd. Ju.
8. Ytterligare åtgärder mot penningtvätt och finansiering av terrorism. Fjärde penningstvättsdirektivet – samordning – ny penningtvättslag – m.m. Del 1 + 2. Fi.
9. Plats för nyanlända i fler skolor. U.
10. EU på hemmaplan. Ku.
11. Olika vägar till föräldraskap. Ju.
12. Ökade möjligheter till modersmålsundervisning och studiehandledning på modersmål. U.
13. Palett för ett stärkt civilsamhälle. Ku.
14. En översyn av tobakslagen. Nya steg mot ett minskat tobaksbruk.S.
15. Arbetsklausuler och sociala hänsyn i offentlig upphandling – ILO:s konvention nr 94 samt en internationell jämförelse. Fi.
16. Kunskapsläget på kärnavfallsområdet 2016. Risker, osäkerheter och framtidsutmaningar. M.
17. EU:s reviderade insolvensförordning m.m. Ju.
18. En ny strafftidslag. Ju.
19. Barnkonventionen blir svensk lag. S.
20. Föräldraledighet för statsråd? Fi.
21. Ett klimatpolitiskt ramverk för Sverige. M.
22. Möjlighet att begränsa eller förbjuda odling av genetiskt modifierade växter i Sverige. M.
23. Beskattning av incitamentsprogram. Fi.
24. En ändamålsenlig kommunal redovisning. Fi.
25. Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning. Del 1 + 2. U.

Statens offentliga utredningar 2016

Systematisk förteckning

Finansdepartementet

Ytterligare åtgärder mot penningtvätt och finansiering av terrorism. Fjärde penningtvättsdirektivet – samordning – ny penningtvättslag – m.m. Del 1 + 2. [8]

Arbetsklausuler och sociala hänsyn i offentlig upphandling – ILO:s konvention nr 94 samt en internationell jämförelse. [15]

Föräldraledighet för statsråd? [20]

Beskattning av incitamentsprogram. [23]

En ändamålsenlig kommunal redovisning. [24]

Justitiedepartementet

Integritet och straffskydd. [7]

Olika vägar till föräldraskap. [11]

EU:s reviderade insolvensförordning m.m. [17]

En ny strafftidslag. [18]

Kulturdepartementet

Låt fler forma framtiden! Del A + B. [5]

EU på hemmaplan. [10]

Palett för ett stärkt civilsamhälle. [13]

Miljö- och energidepartementet

Kunskapsläget på kärnavfallsområdet 2016. Risker, osäkerheter och framtidsutmaningar. [16]

Ett klimatpolitiskt ramverk för Sverige. [21]

Möjlighet att begränsa eller förbjuda odling av genetiskt modifierade växter i Sverige. [22]

Näringsdepartementet

Statens bredbandsinfrastruktur som resurs. [1]

Höghastighetsjärnvägens finansiering och kommersiella förutsättningar. [3]

Socialdepartementet

Effektiv vård. [2]

Framtid sökes – Slutredovisning från den nationella samordnaren för utsatta EU-medborgare. [6]

En översyn av tobakslagen. Nya steg mot ett minskat tobaksbruk. [14]

Barnkonventionen blir svensk lag. [19]

Utbildningsdepartementet

Politisk information i skolan – ett led i demokratiuppdraget. [4]

Plats för nyanlända i fler skolor. [9]

Ökade möjligheter till modersmålsundervisning och studiehandledning på modersmål. [12]

Likvärdigt, rättssäkert och effektivt – ett nytt nationellt system för kunskapsbedömning. Del 1 + 2. [25]