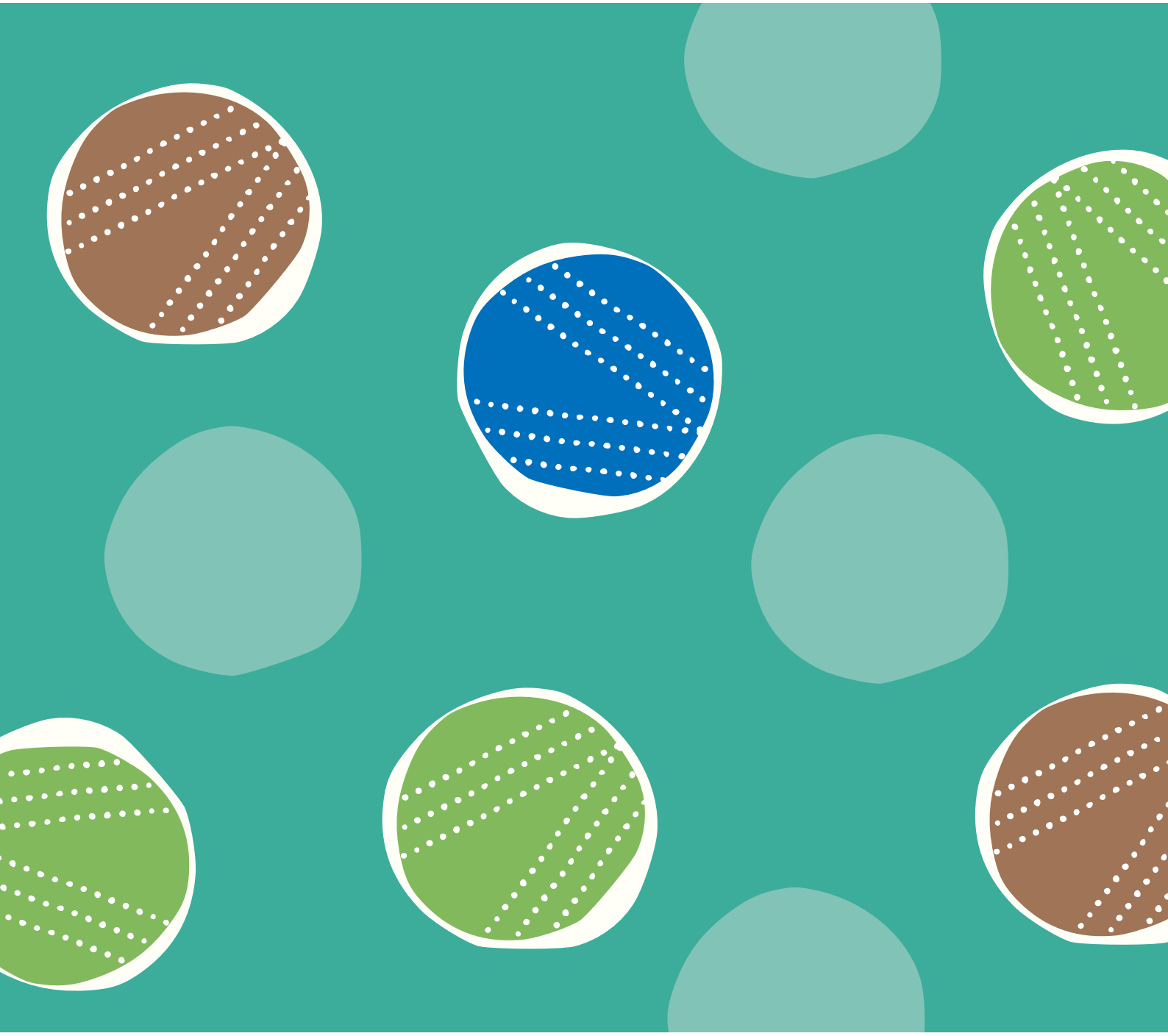
PROV OCH BEDÖMNING

Provpoängens tillförlitlighet

Om nationella prov



Skolverket

Provpoängens tillförlitlighet

Om nationella prov

Publikationen finns att ladda ner som
kostnadsfri PDF från Skolverkets webbplats:
skolverket.se/publikationer

ISSN: 1652-2508
ISBN: 978-91-7559-183-4

Grafisk produktion: AB Typoform
Skolverket, Stockholm 2015

Förord

Den här rapporten ingår i en serie rapporter som handlar om de nationella proven. Rapporterna kan ha lite olika inriktning, somliga analyserar och diskuterar mer innehållsliga aspekter eller redovisar resultat och resultatsammanställningar som inte ingår i den löpande redovisningen av provresultat. Andra rapporter är inriktade mot mer principiella och i vissa fall tekniska frågeställningar kring prov och provkonstruktion. Denna typ av rapporter vänder sig främst till läsare som har ett särskilt intresse för prov och provkonstruktion i stor skala. Den här aktuella rapporten är av det sistnämnda slaget.

De nationella provens betydelse har under senare år ökat påtagligt. Nya ämnen har tillkommit och proven har blivit obligatoriska i flera årskurser. Det finns också tendenser till att proven får allt större betydelse när det gäller betygssättning och att bedöma skolans och undervisningens kvalitet. Mycket lite finns dock skrivet om hur tillförlitliga provresultat egentligen är och hur säkra slutsatser man kan dra om elevers kunskaper utifrån deras prestationer på proven. Den här rapporten har ambitionen att på ett någorlunda lättillgängligt sätt försöka ge en beskrivning av tillförlitligheten i kunskapsbedömningar som baseras på prov uppbyggda av uppgifter som poängsätts eller där resultaten redovisas som antal erhållna belägg på olika kunskaper och förmågor. Däremot behandlas inte prov som i huvudsak bedöms helhetligt, till exempel uppsatser, muntliga prov, laborativa prov etc.

De analyser och resonemang som förs baseras på klassisk testteori. Denna kan se ut på olika sätt beroende på vilka antaganden och approximationer som görs, men det finns vissa gemensamma grundantaganden. Ambitionen här är som nämnts att försöka ge en beskrivning som inte kräver några djupare kunskaper i statistik eller matematik utan som mer visar på vilka styrkor och svagheter bedömningar av elevers kunskaper som baseras på provpoäng har. Hur sådana styrkor och svagheter ska hanteras är inte en teknisk fråga utan en fråga om beslut baserade på värden och värderingar, men för att sådana beslut ska vara välgrundade krävs, förutom gott omdöme, kunskap om provs möjligheter och begränsningar. Förhoppningsvis kan den här promemorian bidra till denna kunskap.

Rapporten innehåller en hel del formler från vilka den som så önskar kan bortse. De riktar sig främst till den som har ett mer tekniskt intresse av frågorna. Vissa avsnitt kan också ses som mer perifera, till exempel stora delar av kapitlet Några utvidgningar där dock avsnittet Slumpfelens betydelse för kravgränser och provbetyg kan vara av mer allmänt intresse.

Kristian Ramstedt har skrivit rapporten. Anders Auer, Lars-Åke Bäckman, Ingrid Jerkeman, Karin Nyqvist och Eva Wirén har lämnat värdefulla kommentarer.

Stockholm i april 2015

Karin Hector-Stahre
Enhetschef

Kristian Ramstedt
Expert

Innehåll

Ett grundläggande samband?

2002 års mottagare av ekonomipriset till Alfred Nobels minne, Daniel Kahneman, skriver i sin bok *Att tänka, snabbt eller långsamt*[1] att den viktigaste ekvation han känner till är

framgång = skicklighet + tur

där tur kan vara både positiv och negativ (otur).

Exemplet golf

Sambandet gäller i de flesta mänskliga sammanhang; aktiehandlare, idrottsmän, forskare, husköpare etc. Kahneman nämner som ett exempel golfspelare. Vissa dagar rullar puttarna i, andra dansar de runt koppkanten. Somliga dagar sitter svingen på plats, eller gör det inte, med påföljd att bollarna hamnar i ruffen. Låt oss anta att vi intresserar oss för en okänd spelare A som deltar i en viss tävling. Det vi så småningom får reda på är hur många slag spelare A hade på sin runda, låt oss säga att han hade 68 slag. Vad kan vi då säga om spelare As skicklighet? Det enda vi vet är spelare As *resultat*, men vi vet inte hur mycket *tur* spelare A hade under sin runda och därmed vet vi inte heller hur stor del av resultatet som bygger på *skicklighet*. Vi kan därmed endast säga att spelare A visat ett resultat som tyder på att As skicklighet motsvarar 68 slag, men osäkerheten ligger i att vi inte vet hur mycket tur (eller otur) som var med i spelet. Med skicklighet avser vi då en mer permanent förmåga som inte förändras från dag till dag utan är tämligen konstant över tid. Att dagsformen kan variera får hänföras till turfaktorn.

Antag nu att A ska gå en ny runda påföljande dag. Vad bör vi då gissa att hen får för resultat? Är det troligast att hen får ett bättre resultat, samma resultat eller ett sämre resultat? Det vi nu vet utöver As resultat är att det genomsnittliga resultatet för samtliga spelare dag ett var lika med banans par, 72 slag. 68 slag var således ett resultat som var bättre än genomsnittet.

Kahneman menar att ekvationen för den som haft *stor* framgång (större än genomsnittet) kan skrivas

stor framgång = något större skicklighet + mycket tur

Eftersom turen kan antas variera slumpmässig (annars är något annat än tur inblandat) blir konsekvensen att spelare A hade mycket tur den första dagen, men ändå antagligen har hyfsad skicklighet i jämförelse med de andra som deltar i spelet eftersom hen ligger klart under par (det genomsnittliga resultatet). Det är dock inte troligt att spelare A har mycket tur även dag två (eftersom det normala är att ha genomsnittlig tur). Skickligheten antas vara densamma dag ett och dag två och den bästa skattningen av spelare As resultat dag två blir därför att hen får ett något sämre resultat dag två än dag ett, men att det antagligen ligger något bättre än par (genomsnittet). Den bästa gissningen för dag två blir därmed att spelare A får ett något sämre resultat närmare genomsnittet dag två.

1 Kahneman, D (2012). *Att tänka, snabbt eller långsamt.*

Dag ett hade vi alltså inledningsvis endast tillgång till spelare As resultat (antar vi), då är den bästa gissningen av As skicklighet att den är lika med det uppnådda resultatet. Vi har inget underlag för att skatta hur stor roll turen spelade.

Inför dag två har vi fått mer information, nämligen att genomsnittet för samtliga deltagande spelarna första dagen var 72 slag. Detta indikerar att spelare A troligen hade mer tur än otur första dagen och att det troliga är att detta kommer att ändras dag två eftersom tur och otur i det långa loppet tenderar att jämna ut varandra. Då kommer sannolikt en *regression mot medelvärdet* att ske vilket får till följd att spelare A kommer att behöva fler slag dag två än dag ett. Hur många fler kan vi dock inte skatta utan det får bli en gissning, till exempel 70 slag.

Låt oss nu anta att vi får veta att spelare A heter Tiger Woods. Då får vi ytterligare information, till exempel att 68 slag inte längre framstår som lika tursamt. Vi vet att det handlar om en mycket skicklig spelare och då kanske turen inte alls har spelat samma roll. Om vi har många resultat för Tiger Woods kanske vi ser att han i genomsnitt har 66 slag på sina tidigare rundor. Då framstår det snarare som om Tiger Woods haft otur dag ett (om skickligheten antas konstant). En regression mot Tiger Woods medelvärde skulle då mer sannolikt innebära att Tiger Woods skulle behöva färre slag än 68 dag två, till exempel 67, vilket dock förstås fortfarande bara är en gissning.

I Tiger Woods fall antar vi att de många resultaten ger oss anledning tro att hans ”sanna” skicklighet ligger på 66 slag, dvs. medelvärdet av hans erhållna resultat i många tävlingar. Tiger Woods har förstås tur och otur ibland som alla andra, men i långa loppet jämnar de ut varandra och därför kan man säga att medelvärdet av hans tävlingsresultat är lika med hans ”sanna” skicklighet, 66 slag.

Slutsats

Det vi ser av ovanstående exempel är att vi inte kommer ifrån en okänd slumpfaktor vid varje enskild tävling. Vi kan med hjälp av viss statistisk information sluta oss till om det är mest troligt att ett visst resultat tyder på att tur eller otur varit med i spelet. Om vi som i det första fallet saknar statistisk information och endast vet den enskilda spelaren As resultat blir den bästa gissningen att As skicklighet är lika med resultatet. Vi har inget underlag för att bedöma graden av tur.

I det andra fallet visste vi mer genom att vi hade informationen om resultatet för hela den grupp som genomförde tävlingen. Då kunde vi göra en skattning av As skicklighet genom att jämföra As resultat med gruppens medelvärde. Om As resultat var bättre än medelvärdet är det mest troliga att As skicklighet är något mindre än det erhållna resultatet eftersom det positiva resultatet indikerar mycket tur. Vi kan med andra ord utgå från att det mest rimliga är att det föreligger en regression mot hela gruppens medelvärde. As ”sanna” resultat bör vara något lägre (närmare gruppens medelvärde) än As observerade resultat.

I det tredje fallet antogs vi känna till den ”sanna” skickligheten genom att vi för den spelaren (Tiger Woods) hade många resultat vilkas medelvärde kunde tjäna som skattning av den spelarens ”sanna” skicklighet. Här har vi mer information vilket gör att vi har två medelvärden och två regressioner mot medelvärdet att ta hänsyn till. I ena fallet kan vi anta att det sker en regression mot medelvärdet för hela gruppen. I det andra fallet kan vi anta att det sker en regression mot den enskilda individens ”sanna” värde (medelvärdet av många

tävlingar). Om vi antar att vi känner den sanna skickligheten betyder det att resultatvariationen vi olika tävlingar endast beror på graden av tur eller otur (skickligheten antas konstant). Men naturligtvis kan Tiger komma ur form. Då sjunker resultaten och med tiden kommer de sämre resultaten att påverka medelvärdet av alla resultat så att det sjunker, vilket alltså innebär att det sanna värdet på skickligheten sjunker. Den sanna skickligheten är således endast temporärt sann (och på sikt förgänglig som allt annat), men här har vi för enkelhets skull antagit att den är konstant.

Man kan sammanfattningsvis konstatera att i vissa fall verkar de två regressionerna i samma riktning, i andra i motsatt riktning.

En följd av ovanstående blir att den som lyckats mycket bra på en tävling sannolikt inte lyckas lika bra nästa gång och vice versa. För de tävlande som är ungefär lika skickliga blir det graden av tur eller otur som vid enskilda tävlingar avgör hur de placerar sig. Att det verkligen förhåller sig på det sättet är lätt att kontrollera genom att jämföra resultat från olika tävlingar.

Från golf till prov

Om vi istället för golf tänker oss att det handlar om provresultat skulle motsvarande resonemang innebära följande:

1. När vi endast har *ett* provresultat (och ingen annan information) för en individ är detta resultat den bästa skattning vi kan göra av provdeltagarens skicklighet, kunskaper, förmågor eller vad det råkar handla om. Vi vet att det troligen finns ett mätfel, men inte mer.

2. Om vi dessutom vet *hela provgruppens* resultat kan vi använda denna information genom ”lagen” om regression mot medelvärdet. Den provdeltagare som har resultat över medel kan antas ha haft mer tur än otur på provet och därmed ha en ”sann poäng” som sannolikt är lägre än den erhållna provpoängen. För den som har en provpoäng under medel är det tvärtom.[2]

3. Om vi därutöver har tidigare information om provdeltagaren som kan tjäna som underlag för att skatta dennes ”sanna poäng”, t.ex. många tidigare provresultat, finns ytterligare en regression mot medelvärdet att ta hänsyn till, nämligen regressionen mot det på tidigare resultat baserade sanna värdet. Regressionen i punkt 2 och regressionen i punkt 3 kan ge olika indikationer beroende på hur det uppnådda resultatet förhåller sig till medelvärdet för gruppen respektive det ”sanna” värdet för individen. Ibland verkar de i samma riktning ibland i motsatta riktningar.[3]

---

2 Det finns exempel på hur man kan skatta denna förskjutning som vi återkommer till (t.ex. Kelleys formel).

3 Ett typiskt exempel på denna motsättning kan ses i frågan om vem som är bäst lämpad att bedöma ett provresultat, en bedömare som inte vet någonting om provdeltagaren eller en bedömare som känner provdeltagrens tidigare prestationer. Provresultatets syfte måste ju vara att ge ett så ”sant” mått som möjligt på provdeltagarens skicklighet.

Hur väl skattningarna av skickligheten baserade på tävlingsresultat eller provresultat stämmer med den ”sanna” skickligheten” i de enskilda fallen kan vi inte veta. Statistiken ger oss endast vissa verktyg för att skatta sannolikheten för att de resultat vi uppnår i tävlingar eller på prov avspeglar vår sanna förmåga och hur mycket av resultatet som beror på tur eller otur.

I nästa avsnitt visas några grundläggande sätt att hantera provresultat. Där handlar det om prov som är baserade på olika typer av uppgifter som är poängsatta och där provresultaten utgörs av respektive elevs poängsumma för de ingående uppgifterna.

Poängbaserade prov

Kahneman fick sitt pris i ekonomi men är från början psykolog. Som sådan är han förstås välbekant med den klassiska testteorin och dess grundläggande ekvation:[4]

$$x_o = x_s + x_f$$

där x_o är den observerade provpoängen,

x_s den ”sanna” provpoängen och

x_f är den del av den observerade poängen som beror på turen, eller med andra ord representerar ett slumpfel (f).

Som synes är denna den klassiska testteorins grundekvation densamma som Kahnemans ekvation i golfexemplet.

Den ”sanna” provpoängen är inte möjlig att observera utan är ett abstrakt begrepp som antas finnas latent (ej observerbart) och som kan skattas mer eller mindre approximativt med hjälp av olika statistiska metoder utifrån de observerade poängen. Förenklat kan man säga att den sanna poängen för en elev definieras som den genomsnittliga poäng eleven skulle få om han eller hon skulle upprepa samma prov ett stor antal gånger. Detta är möjligt att tänka sig som teoretiskt antagande men inte att utföra i praktiken av olika skäl, inte minst att de elever som genomför samma prov gång på gång skulle drar lärdomar av tidigare provgenomföranden.

Det är således viktigt att vara medveten om att ovanstående ekvation *antas* gälla. Att den gäller är alltså ett *första grundläggande antagande* för den klassiska testteorin.[5] Med hjälp av detta antagande och ett par andra kan man härleda uttryck som kan ge bättre underlag för att skatta slumpens inverkan på ett provresultat.

Ett *andra antagande* är att för en *population* (t.ex. alla som spelar golf eller deltar i ett prov) gäller att felet (turen) är oberoende av den sanna poängen, dvs. felet varierar inte med den sanna poängen utan korrelationen mellan sann poäng och felpoäng är noll.

---

4 Den som vill ha en utförligare genomgång hänvisas till exempelvis Crocker & Algina(1986).

5 Det finns också en modern testteori, IRT (item response theory) som bygger på andra antaganden. Den tar vi inte upp här.

Ett *tredje antagande* är att felet på ett prov är oberoende av felet på ett annat prov,[6]

För en *population* som genomför ett prov kan medelvärdet för den observerade poängen anges som summan av medelvärdet för den sanna poängen och av feltermen

$$\bar{x}_o = \bar{x}_s + \bar{x}_f$$

Men för en population tenderar tur och otur att ta ut varandra och medelvärdet för ”turtermen” $\bar{x}_f$ går därför mot noll när populationen växer[7] eller mer formellt uttryckt

$$\bar{x}_f \approx 0$$

Vilket alltså betyder att medelvärdet av alla fel är lika med noll. För *populationen* gäller därmed

$$\bar{x}_o \approx \bar{x}_s$$

Det vill säga *medelvärdet för den observerade poängen för populationen är lika med medelvärdet för populationens sanna poäng*. Tur och otur antas ha tagit ut varandra. Medelvärdet för den ”sanna” poängen för en population är alltså lika med medelvärdet av den observerade poängen.

För *de enskilda individernas* resultat antas däremot felen (turen) fördela sig slumpmässigt (normalfördelat) runt deras respektive sanna värden, med samma sannolikhet för positiv som negativ inverkan. Den sanna poängen för en individ definieras enligt tidigare som medelvärdet av de poäng individen skulle erhålla om han eller hon genomförde samma prov ett stort antal gånger. Detta är som nämnts givetvis inte möjligt eftersom minnet från tidigare genomföranden förändrar provdeltagarens prestation; vi är lärande individer. *Den sanna poängen* på ett prov blir därmed på individnivå *en icke observerbar storhet*. Den kan dock skattas utifrån vissa antaganden. Vi återkommer till detta.

Den enskilda mätningens medelfel SEM

Eftersom den sanna poängen inte är observerbar försöker vi i stället skatta turens betydelse för relationen mellan den observerade poängen och den sanna poängen för en enskild individ. För en rimligt stor grupp tar enligt tidigare tur och otur ut varandra. Eftersom man där endast är intresserad av resultatet på gruppnivå behöver man således inte lägga någon större vikt vid turens betydelse för enskilda individer. Däremot får sådana slumpfel betydelse när det gäller skattningen av den sanna poängen för *de enskilda eleverna*. Dock är det oklart hur stor slumpens inverkan är för varje specifik elev.

---

6 Låter rätt självklart, men är ett antagande som används när man till exempel räknar ut ett mått på provets *reliabilitet* (tillförlitlighet) genom ”split half metoden” som innebär att uppgifterna i ett prov delas i två uppsättningar uppgifter eller två prov om man så vill. Korrelationen mellan resultaten på dessa prov (korrigerad till provets fulla längd med Spearman-Browns formel) ses sedan som ett mått på provets reliabilitet ”Coefficient alfa” som är det vanligaste måttet på reliabilitet kan ses ett medelvärde av alla för provet möjliga split half kombinationer. Det finns också andra mått på reliabilitet, men de är svåra att tillämpa i praktiken (se t.ex. Crocker & Algina, 1986).

7 Enligt ”de stora talens lag” http://sv.wikipedia.org/wiki/De_stora_talens_lag.

Utan att närmare gå in på härledningen[8] kan fördelningen av slumpfaktorn under de antaganden som angetts ovan härledas till följande uttryck

$$s_f = s_o\sqrt{1 - r_{xx'}}$$

där s_f är standardavvikelsen för ”turen” eller slumpfelet x_f och s_o är standardavvikelsen för den observerade poängen x_o på provet, dvs. standardavvikelsen för alla elevers poängsummor på provet. $r_{xx'}$ är ett mått på provets reliabilitet[9] Reliabiliteten kan sägas vara ett generellt mått på slumpens inflytande på provresultatet där en reliabilitet nära 1 anger mycket litet slumpinflytande och ett värde nära 0 ett mycket stort inflytande av slump. s_f betecknas ofta *SEM*, vilket står för Standard Error of Measurement, eller på svenska ”den enskilda mätningens standardfel” eller ”den enskilda mätningens medelfel”.

I klartext innebär formeln och de antaganden som gjorts att slumpfelet x_f för den enskilda provdeltagaren antas vara normalfördelat runt den observerade provpoängen för eleven, med standardavvikelsen s_f, eller med den vanligare beteckningen.[10] *Den enskilda provdeltagarens* sanna poäng x_s ligger då med 95 procents sannolikhet någonstans i intervallet

$$x_o - 1{,}96 * SEM \leq x_s \leq x_o + 1{,}96 * SEM$$

Det är alltså viktigt att notera att även om medelvärdet av den observerad poängen för en klass ger ett tämligen bra mått på *klassens* sanna genomsnittliga poäng[11] så gäller inte detta för den enskilda eleven, där kan den enskilda mätningens ”(slump)fel” vara betydande.

Ovanstående kan måhända förefalla något abstrakt så låt oss se på ett konkret exempel.

Ett exempel på skattning av sann provpoäng

Som underlag för exemplet används data från det nationella provet i fysik för årskurs 9 våren 2011. I provet används inte begreppet poäng utan resultatet räknas i ”belägg” som är kopplade till olika delar av kursplanens kunskapskrav i ämnet. Vissa belägg karaktäriseras som belägg på G-nivå, andra på VG- respektive MVG-nivå. En och samma uppgift kan ge underlag för olika kombinationer av belägg vi bedömningen. En enkel uppgift kan t.ex. ge 1 G-belägg, en mer komplex uppgift kan t.ex. ge, 2 (1+1) VG-belägg och 1 MVG-belägg, osv. Dessutom klassificeras beläggen efter tre övergripande innehållsliga ”aspekter” som finns angivna i kursplanen för fysik och andra NO-ämnen. Varje uppgift bedöms således enligt en tvådimensionell 3x3 matris (G, VG, MVG x A1, A2, A3) där olika uppgifter prövar olika celler i matrisen. Man kan säga att varje uppgift består av ett antal ”item” som rätt besvarade ger belägg i en eller flera

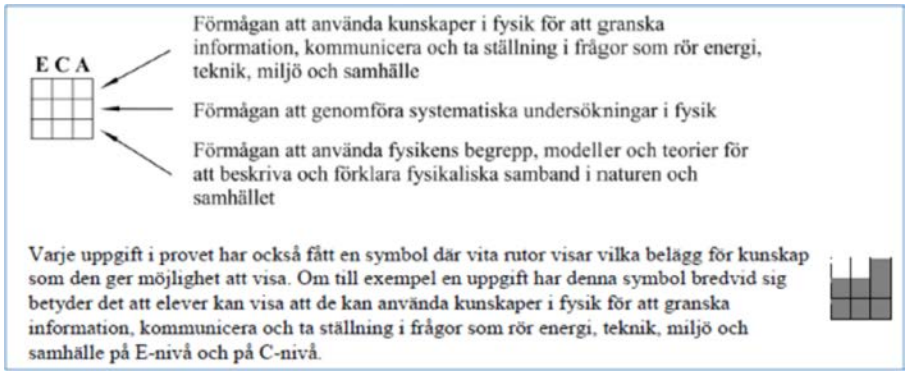
8 Se t.ex. Crocker & Algina (1987) för en härledning.

9 Vanligen används coefficient alfa som mått på reliabiliteten eftersom den ingår i t.ex. programvaran SPSS och därför är lätt att beräkna. Det lämpliga i att använda detta mått kan diskuteras men tas inte upp här.

10 Egentligen anger formeln hur den *observerade* poängen är fördelad runt den *sanna* poängen. Men som approximation kan man vända på ekvationen och se den som att den anger den sanna poängens fördelning runt den observerade poängen. (se t.ex. Harvill, 1991)

11 Förutsatt att inga systematiska fel förekommer.

celler i matrisen (jämför med bilden nedan där den nya betygsskalans beteckningar E, C och A används).[12]



Betygssättningen baseras sedan på betygsgränser som även de anges i olika kombinationer av belägg. För provbetyget G krävs för det aktuella provet minst 12 belägg (oberoende av nivå) och belägg inom varje aspekt (de tre pilarna i figuren). För VG gäller minst 22 belägg totalt samt minst 8 belägg på VG- eller MVG-nivå. För MVG slutligen gäller minst 30 belägg totalt samt minst 3 belägg på MVG-nivå.

Sättet att bedöma och betygssätta provresultatet kan diskuteras, men i det här sammanhanget är endast provet och provpoängen (beläggen) intressanta som underlag för att undersöka relationen mellan observerad poäng och sann poäng samt slumpfaktorns (s_f eller *SEM)* storlek. Det betyder att för det aktuella provet betraktas alla belägg som poäng på samma nivå oberoende av om de kallas G-, VG-, eller MVG-belägg. Någon åtskillnad för belägg inom olika aspekter görs inte heller i de inledande redovisningarna. Vi återkommer senare till frågan om vad det i bedömningsanvisningarna angivna sättet att hantera belägg har för inverkan på slumpfaktorns storlek.

Data för det aktuella provet

Nedanstående tabell anger de beskrivande data och parametrar för provet som behövs för bestämningen av *SEM*. De är framtagna med hjälp av SPSS.

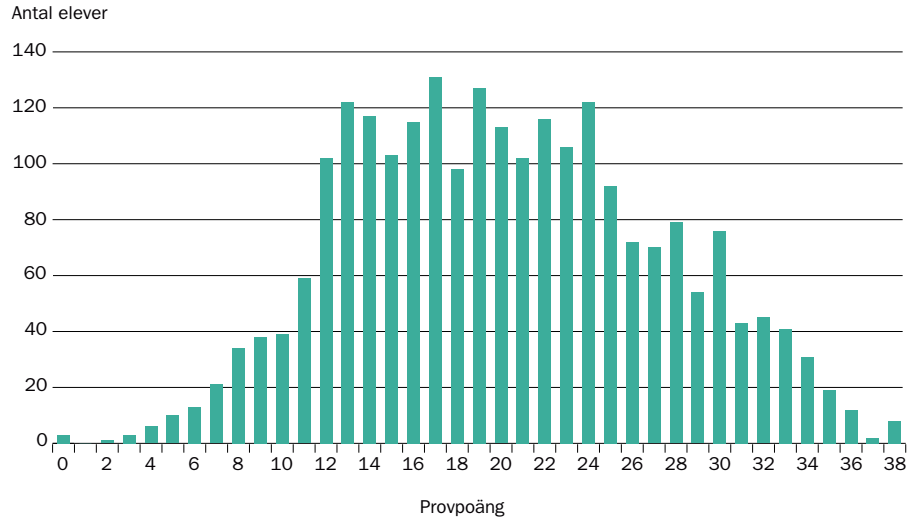
Tabell 1. Data och statistiska parametrar för provet.

Antal prov	Medelpoäng	Std	Min	Max	Reliabilitet (α)
2 345	20,2	7,1	0	38	0,88

För att få en uppfattning om hur de provdata man arbetar med ser ut är det alltid bra att visa poängfördelningen i ett diagram.

12 http://www5.edusci.umu.se/np/AP-info/vt13/Bedomningsanvisningar_Fysik.pdf

Figur 1. Poängfördelning för prov Fy åk9



Provet består således av 38 item[13] som vart och ett ger ett belägg eller en poäng som det kallas här. De 38 beläggen är fördelade på 18 uppgifter.[14]

Medelvärdet 20,2 verkar av figuren att döma rimligt och standardavvikelsen 7,1 som indikerar att cirka två tredjedelar av provpoängen ligger inom 20±7, dvs. mellan 13 och 27 poäng, verkar också stämma.

Stinas provresultat

Låt oss nu anta att Stina har fått 23 poäng på detta prov. Med den information vi har tillgång till, hur bör vi bedöma Stinas poäng? Hennes observerade poäng (x_0^{Stina}) är 23, men som vi vet kan den antas bestå av en sann poäng (x_s^{Stina}) som vi inte känner och en ”slumppoäng (x_f^{Stina})” som vi inte heller känner Tack vare att vi har data för hela gruppen kan vi dock skatta hur stort slumppoängen eller det genomsnittliga mätfelet *SEM* är.

Det är alltså standardavvikelsen för mätfelet vi kan beräkna med hjälp av formeln

$$SEM = s_o\sqrt{1 - r_{xx'}}$$

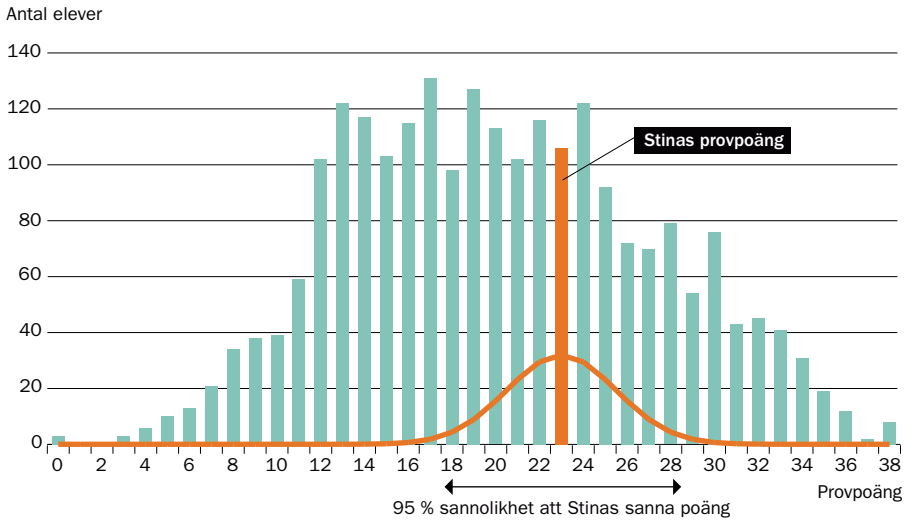
Sätter vi in värden från tabell 1 får vi $SEM = 7{,}1\sqrt{1 - 0{,}88} = 2{,}46$

Det genomsnittliga standardfelet *SEM* är således 2,46 poäng. Det innebär att sannolikheten för att Stinas sanna poäng ligger i intervallet 23±2,46 eller någonstans mellan 21 och 25 poäng (om vi avrundar till heltal) är cirka 68 procent.[15] Det betyder i sin tur att en tredjedel av dem som har 23 poäng har en sann poäng under 21 poäng eller över 25 poäng. Men med tur eller otur hamnar de på 23 poäng på det aktuella provet.

13 Ett item kan vara en uppgift som ger en poäng eller ett belägg, eller en del av en större uppgift som består av flera item som vart och ett ger en poäng (belägg).
14 I datamaterialet finns dels 18 uppgiftsvariabler med respektive uppgifts totalpoäng, dels 38 itemvariabler som ger 0 eller 1 poäng.
15 Arean mellan standardavvikelsen -1 och 1 i normalfördelning är 68 procent av hela arean.

Vill man vara till 95 procent säker på att täcka in en elevs sanna poäng, t.ex. Stinas, får man ta ett inter vall 23±2,46*1,96 = 23±4,82. För Stinas del innebär alltså provresultat med 95 procents säkerhet att hon kan antas ha en sann poäng som ligger någonstans i intervallet 18 till 28 poäng. Figur 2 illustrerar förhållandet.

Figur 2. Poängfördelning Fy åk9 sant Stinas provpoäng och konfidensintervall (95 %) för Stinas sanna poäng.



Om man kopplar tillbaka ovanstående figur till Kahnemans ekvation

framgång = skicklighet + tur

skulle Stinas ”framgång”, 23 poäng, utgöra summan av hennes skicklighet och hennes tur, där det enda vi vet är att turen (enligt de antaganden modellen bygger på) är normalfördelad runt den sanna poängen med en standardavvikelse på 2,46 poäng.[16] Något särskilt bra mått på skickligheten har vi dock inte. Det mest sannolika är att turen är noll, dvs. utifrån diagramet i figur 2 att Stinas sanna poäng är lika med hennes observerade poäng., men den sannolikheten är ju inte så stor eftersom vi med 95 procent sannolikhet endast kan säga att hennes sanna poäng ligger mellan 18 och 28 poäng.

Frågan är om vi blivit så mycket klokare av denna genomgång. Slutsatsen blir snarast att de 23 poäng Stina fått på provet inte känns särskilt tillförlitliga som mått på vad kon verkligen kan, dvs. hennes ”sanna” kunskaper eller skicklighet. Det mesta talar för att turen (eller oturen) haft ganska stor betydelse. Vi har alltså anledning att tvivla på hur bra mått på Stinas förmåga (skicklighet) de 23 poängen (framgången) egentligen är. Men det är kanske en klok insikt.

16 Detta är i själva verket en approximation eftersom vi inte känner den sanna poängen, men den observerade poängen kan ses som en approximation av den sanna poängen och *SEM* ger en uppfattning om slumpfelets storlek och fördelning. Mer om detta i nästa avsnitt.

Skattning av ett värde för den sanna poängen baserad på gruppens resultat

Kan man då inte få fram en bättre skattning av Stinas sanna poäng? I det här fallet har vi tillgång till resultat från hela populationen (antar vi, i själva verket handlar det om ett stickprov på 2345 elever, men ett så stort stickprov kan i det här sammanhanget ses som en population). Det betyder att vi kan utnyttja den regression mot medelvärdet som vi tidigare talat om.

Under vissa antaganden (som vi inte går in på här), men som kan antas giltiga i vårt fall, kan man beräkna en skattning av den sanna poängen för olika observerad poäng med följande formel[17]

$$\hat{x}_S = r_{xx'}(x_O - \bar{x}) + \bar{x}$$

Där $\hat{x}_S$ är den estimerade sanna poängen (att värdet är estimerat anges av ”hatten”), $r_{xx'}$ är reliabiliteten, x_O den observerade poängen och $\bar{x}$ medelvärdet av den observerade poängen för alla i gruppen som gjort provet.

Tillämpas denna formel på Stinas resultat skulle hennes sanna poäng bli (se tabell 1 för värdena i formeln)

$$\hat{x}_S = 0{,}88(23 - 20{,}2) + 20{,}2 = 22{,}7 \approx 23$$

I Stinas fall blir således regressionen mot medelvärdet svag och resulterar inte i något behov av att justera den observerad poängen (i varje fall inte om vi håller oss till heltalspoäng). Det beror på att reliabiliteten är rätt hög och att Stinas poäng ligger ganska nära medelvärdet. Bilden i figur 2 kvarstår således oförändrad även om man för Stinas del tar hänsyn till regressionen mot medelvärdet. Regressionen mot medelvärdet tillför således i det här fallet försumbar information för Stinas del.[18]

Figur 3 illustrerar regressionen mot medelvärdet för det aktuella provet och att den har liten inverkan nära medelvärdet där Stina ligger. Endast för de elever som har höga eller låga provpoäng avviker den observerade poängen från den sanna tillräckligt för att det ska ge utslag i hela poäng.[19]

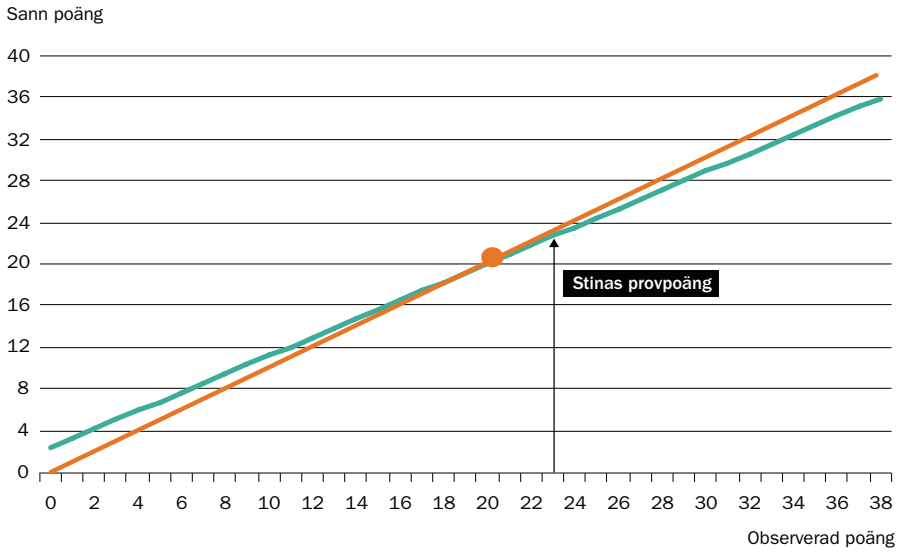
---

17 Kelleys formel, se t.ex. Crocker & Algina (1986) eller Lord & Novick (1967).

18 I provet finns endast hela poäng och den sanna poängen 22,7 avrundas således till 23, vilket är lika med Stinas observerade poäng.

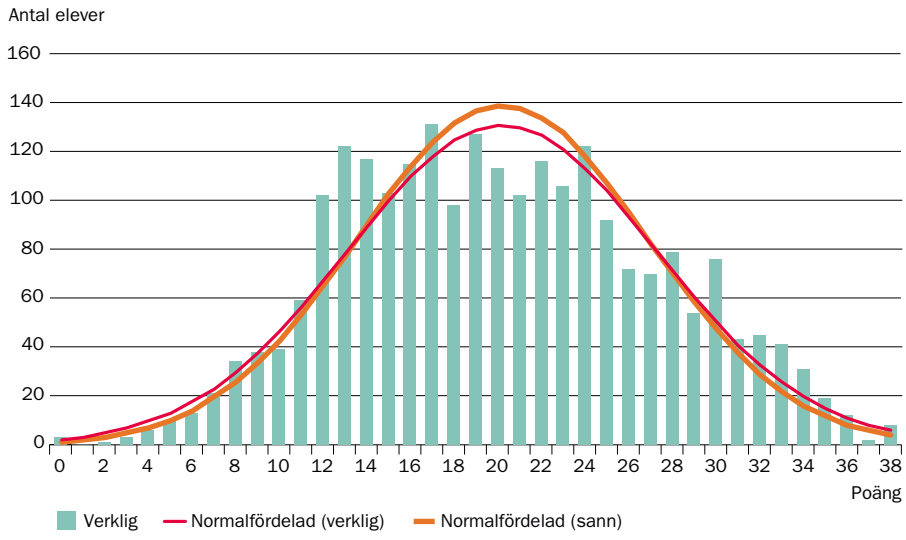
19 För en elev med observerad poäng 5 blir t.ex. resultatet att den sanna poängen är $\hat{x}_S = 0{,}88(5 - 20{,}2) + 20{,}2 = 7{,}42 \approx 7$, alltså en tydlig ”regression mot medelvärdet” (20,2). Den observerade poängen 5 estimeras innebära en *högre* sann poäng 7.
För en annan elev med t.ex. observerad poäng 35 blir resultatet att den sanna poängen är $\hat{x}_S = 0{,}88(35 - 20{,}2) + 20{,}2 = 33{,}2 \approx 32$, alltså också en tydlig ”regression mot medelvärdet” (20,2) som innebär att den sanna poängen estimeras till ett *lägre* värde än den observerade. Den observerade poängen 35 skattas representera en sann poäng på 32. (Jämför figur 3)

Figur 3. Observerad poäng och sann poäng för Fy åk 9. Den turkosa linjen illustrerar den sanna poängen som regression mot medelvärdet (ringen vid 20,2). Den orangea linjen är diagonalen.



Ett annat sätt att illustrera regressionen mot medelvärdet och den sanna poängens fördelning visas i figur 4.

Figur 4. Observerad fördelning och motsvarande normalfördelning samt fördelning av sann poäng (regression mot medelvärdet).



I figur 4 ser man att provpoängen har en rätt tydlig normalfördelad form och att regressionen mot medelvärde innebär att den sanna poängen har mindre spridning än den observerade, det vill säga ligger närmare medelvärdet 20,7 poäng.

Vi nämnde tidigare att den formel vi använder för att beräkna *SEM* är härledd för att beräkna den observerade poängens fördelning runt den sanna poängen. Som approximation används den dock vanligen tvärtom dvs. som

ett mått på den sanna poängens fördelning runt den observerade poängen. En mer korrekt formel för att bestämma den sanna poängens fördelning anges av Harvill (1991) i formeln

$$x_S = \hat{x}_S \mp s_O\sqrt{1 - r_{xx'}} * \sqrt{r_{xx'}}$$

där variablerna har samma innebörd som tidigare och $\hat{x}_S$ bestäms enligt den tidigare formeln

$$\hat{x}_S = 0{,}88(23 - 20{,}2) + 20{,}2 = 22{,}7$$

Sätter man in värdena för Stina fås som resultat att hennes sanna poäng x_s ligger i intervallet

$$x_S = 22{,}7 \mp 7{,}1\sqrt{1 - 0{,}88} * \sqrt{0{,}88} = 22{,}7 \mp 2{,}3$$

Sannolikheten blir då 68 procent att Stinas sanna poäng ligger i intervallet

$22{,}7 - 2{,}3 \leq x_S \leq 22{,}7 + 2{,}3$, eller uträknat $20{,}4 \leq x_S \leq 25{,}0$. Avrundar vi ser vi att den sanna poängen skattas till 23 och intervallgränserna till 20 och 25. Avrundning gör att intervallet blir osymmetriskt.

Vill vi ha 95 procentig säkerhet får vi multiplicera slumpfelet med 1,96 och får då intervallet mellan 18,2 och 27,2 eller avrundat mellan 18 och 27 poäng. Enligt en mer korrekt metod att beräkna Stinas sanna poäng skulle således det mest sannolika värdet vara 23 poäng men med 95 procentig sannolikhet ligga i intervallet 18 till 27 poäng.

Om vi jämför med de värden vi fick med den mer approximativa metod där vi utgick från Stinas observerade poäng och *SEM* = 2,5 ser vi att vi får samma skattning av Stinas sanna poäng (23) medan slumpfelet blir något större 2,47 eller avrundat 2,5. Intervallet för 95 procentig säkerhet för den sanna poängen blev med den tidigare approximationen 18 till 28 poäng. Beroende på avrundningar kan man få någon enstaka poängs skillnad, men på det hela taget kan man säga att den första approximationen förefaller ge tillräckligt goda värden på skattningen av felmarginalen för det sanna värdet och det finns åtminstone för Stinas del knappast anledning att använda några utökade formler.

Item eller uppgifter?

Man kan i sammanhanget undra lite över vilken betydelse det har att resultatet är uppdelade i enskilda belägg eller poäng i stället för i poäng på uppgifter. Reliabiliteten 0,88 är inget att anmärka på för ett prov med 18 uppgifter. Här är dock varje uppgift bedömd efter särskild anvisning per item och räknar man antal distinkta item eller belägg blir det totalt 38 belägg av olika slag. Man kan alltså säga att resonemangen så här långt gäller ett prov med 38 dikotoma item, dvs. uppgifter som ger 1 eller 0 poäng (belägg)

Om man i stället betraktar resultatet som ett prov med 18 uppgifter med varierande antal poäng per uppgift (genom att ta summan av de item som tillhör uppgiften) blir reliabiliteteten 0,80. Medelpoäng och standardavvikelse blir förstås desamma som tidigare 20,2 respektive 7,1 poäng. Sätter man in dessa värden i formeln för *SEM* får man $SEM = 7{,}1\sqrt{1 - 0{,}80} = 3{,}18$

Det betyder alltså att reliabiliteten blir lägre och slumpinflytandet större när uppgifterna ges en samlad poäng. En uppdelning av uppgifterna på item (belägg) medför således att reliabiliteten ökar och att *SEM* minskar från 3,2 till 2,5. En mer preciserad bedömning tycks således ge ett mer reliabelt prov med mindre mätfel. Detta förefaller inte förvånande eftersom en mer preciserad bedömning innebär att bedömningen innehåller mer information. Slutsatsen blir således att det ur ett mätperspektiv finns ett värde i att bedömningsanvisningarna är nedbrutna till item-nivå så att resultaten kan preciseras i enskilda belägg eller poäng.

Den enskilda individens sanna poäng eller sanna förmåga

Så här långt har resonemangen gällt vilka slutsatser man kan dra av resultaten på *ett* prov. Baserat på vissa antaganden och statistiska beräkningar kan man då skatta sanna värden och felmarginaler samt utifrån dessa dra slutsatser om hur tillförlitliga, eller snarare otillförlitliga, provresultat är när det gäller att bedöma enskilda elevers ”skicklighet”, kunskaper, förmåga eller vad man kallar den underliggande förmåga provet förutsätts mäta.

När det gäller den regression mot *individens* medelvärde man kunde tänka sig vid värderingen av Tiger Woods golfresultat så är den av lätt förståeliga skäl inte tillämplig på nationella prov. Eleven har inte någon serie av nationella provresultat i bagaget. Det förefaller dock rimligt att tänka sig att en motsvarande regression mot en enskild elevs individuella medelvärde kan ha betydelse när det visar sig att lärare tenderar att bedöma annorlunda än vad t.ex. Skolinspektionens bedömare[20] gör. Läraren har i allmänhet extra information om sina elever som kan innebära att provresultatet tolkas på ett annat sätt än vad en neutral granskare gör. Den senare har ingen tilläggsinformation som kan leda till regression mot det ”medelvärde” elevens samlade prestationer ger i lärarens föreställningsvärld.

Så länge läraren har frihet att sätta betyg spelar det i princip ingen roll om läraren väger in sin bedömning av turens eller oturens roll för provresultatet vid provrättningen eller senare vid betygssättningen. Men i takt med att trycket ökar på att bedömningen ska vara neutral i den meningen att ingen information utöver den som finns i elevens provresultat ska vägas in torde lärarens rättning tendera att bli mer formell. Är detta enbart av godo?

Man kan tänka sig att läraren i högre utsträckning än en oberoende bedömare bedömer elevens ”skicklighet” vid rättningen medan den senare mer bedömer den ”framgång” provet uttrycker (för att använda Kahnemans begrepp). Det som inte stämmer med den bilden är att tur och otur inte tenderar att ta ut varandra. Skolinspektionens oberoende lärare tenderar nämligen att systematiskt ge lägre provbetyg än de egna lärarna. Om det enbart handlade om slumpfel borde lärare sätta lägre betyg lika ofta som högre.[21] Detta förutsätter dock att de inte ser tidigare provresultat och andra resultat som kompensatoriska, dvs. att de vid bedömningen av det aktuella provresultatet läser in kunskaper eleven visat vid andra tillfällen och därmed tolkar mer välvilligt än en oberoende bedömare som enbart bedömer den skriftliga utsagan. Vilken typ av bedömning som ger

20 Skolinspektionen (2013)
21 Vi antar då för resonemangets skull att Skolinspektionens bedömningar är korrekta.

ett provresultat med en mer rättvisande bild av den enskilda elevens sanna kunskaper och förmågor är en fråga som kan diskuteras?[22]

Givetvis är det kontroversiellt att argumentera för att en subjektiv provbedömning skulle kunna vara mer rättvisande än en objektiv, och så länge det handlar om resultat på övergripande och aggregerad nivå finns det som framgått ingen anledning att inte sträva efter objektiv bedömning. När det handlar om examensprov eller andra prov av stor betydelse för individen bör ambitionen naturligtvis vara att bedömningen av eleven blir så riktig som möjligt, dvs. att olika slumpfaktorers roll minimeras; otur (eller tur) med val av provuppgifter, provångest, tillfällig värk eller vad det kan vara. Men i ett sådant fall är det också viktigt att tolkningen av provresultatet så väl som möjligt ger en korrekt bild av elevens förmåga.[23] Det handlar sist och slutligen om att i mesta möjliga mån fånga elevens skicklighet för att tala med Kahneman.

Sammanfattande kommentar

Vad har man då för nytta av den här sortens övningar och diskussioner? En förhoppning är att kunskapen ska vara av värde för den som ska förhålla sig till Skolverkets provuppdrag och de prov som blir följden av detta. De nationella proven har för närvarande två uttalade syften, nämligen:

- stödja en likvärdig och rättvis bedömning och betygssättning
- ge underlag för en analys av i vilken utsträckning kunskapskraven uppfylls på skolnivå, på huvudmannanivå och på nationell nivå.

Om vi börjar med den senare punkten kan vi alltså säga att i det avseendet är provresultaten tillförlitliga när det gäller det slumpmässiga mätfelet, eftersom det handlar om grupper och aggregerade värden. Då antas tur och otur ta ut varandra och de erhållna medelvärdena kan ses som i huvudsak sanna mått för gruppen vad *slumpfel* anbelangar.

Däremot kan förstås fortfarande *systematiska* fel finnas, t.ex. att en lärare bedömer proven på ett avvikande sätt (systematiskt fel på klassnivå), att en skola systematiskt bedömer proven på något särskilt sätt som inte är i enlighet med bedömningsanvisningarna (systematiskt fel på skolnivå), eller att kravgränserna för olika betyg ligger på ”fel” nivå (systematiskt fel på nationell nivå). Där det sistnämnda kan innebära att resultat mellan olika år inte blir jämförbara. I det här sammanhanget intresserar vi oss emellertid endast för sådana fel som kan hänföras till slumpmässiga variationer på individnivå och går därför inte in på de systematiska fel som kan förekomma på andra nivåer.

När det gäller den första punkten, att fungera som stöd för en likvärdig och rättvis bedömning (på individnivå), är det betydligt svårare att ur ett mätperspektiv uppfylla syftet. Om de nationella proven används som underlag för att bedöma skolans eller klassens nivå är medelvärdet av provpoängen som vi konstaterat ett tämligen tillförlitligt mått på klassens eller skolans nivå, och mer tillförlitligt ju större gruppen är. Men när proven tenderar att användas som examensprov för enskilda elever blir ”den enskilda mätningens standardfel” som vi sett betydande och då måste provresultat tolkas med stor försiktighet. Och

22 Här kommer vi in på validitetsfrågor som dock inte tas upp närmare i det här sammanhanget.
23 Se t.ex. Kane (2013) för att få en översikt över ett modernt sätt att se på validitet.

detta är kanske den främsta lärdomen av denna korta genomgång av den klassiska testteorins grundläggande teorier. *Provresultat som mått på enskilda elevers kunskaper är otillförlitliga, men de är de minst dåliga hjälpmedel vi känner till.* Provens betygsstödjande funktion måste därför också verka med andra medel, som till exempel stöd och information om klassers och skolor resultat och vilka slutsatser som kan dras av sådana när det gäller enskilda elevers resultat, det vill säga ökad förståelse av sådant som den här promemorian handlar om, tydliggörande av kunskapskravens innebörder genom förklarande analyser och kommenterade exempel etc. Det vill säga olika typer av stöd för de lärare som har till uppgift att gör sammanfattande bedömningar av elevernas skicklighet i de avseenden kursplanerna föreskriver och manifestera de visade förmågorna i rättvisa och likvärdiga betyg.

Det är också viktigt att komma ihåg att *SEM* bygger på ett antal antaganden och att det är ett mått som mer ska ses som en indikation än som ett precist mått. *SEM* är ett medelvärde av standardfelet för varje poängnivå och i själva verket varierar standardfelet med poängen på provet. Vi återkommer kort till detta i nästa avsnitt. Eftersom standardfelet varierar med poängnivån kan det vara intressant att få en indikation på hur stort standardfelet är i närheten av provets olika betygsgränser. Även frågan om hur man kan se på mätfelen då betygsgränserna på ett prov baseras på olika kombinationer av provpoäng (G-belägg, VG-belägg etc.) eller andra villkor kan behöva diskuteras. Några sådana frågor tas upp i nästa avsnitt.

Några utvidgningar

I det här kapitlet är de två första punkterna rätt tekniska och specifika till sin natur och kan överhoppas av den som inte finner innehållet intressant. Den tredje punkten däremot är av mer allmänt intresse.

I det här avsnittet ska vi göra några utvidgningar som innebär att vi undersöker

- Villkorade standardfel med och utan korrektion.
- Standardfel vid olika poängvillkor för provbetyg.
- Slumpfelens betydelse vid olika antal betygssteg.

Villkorade standardfel med och utan korrektion

Som oftast i statistiska sammanhang finns det flera olika tillvägagångssätt och metoder. När det gäller att estimera standardfel för provpoäng på olika nivåer (så kallade ”conditional SEM” eller på svenska ”villkorat standardfel”) anger Felt m.fl.[24] fem olika metoder som de testat. Samtliga ger värden av ungefär samma storleksordning och här väljer vi den metod som förefaller enklast att använda. Den formel som används ser ut på följande sätt

$$SEM_{cond} = \sqrt{\frac{x(k-x)}{k-1}}$$

där x är poängnivån och k är antalet uppgifter (item) i provet. Denna formel anses dock ge ett något för högt värde och därför används ibland något som kallas Keats korrektion, vilket innebär att provets reliabilitet beräknad på två olika sätt används som korrektionsfaktor. Detta ger nedanstående uttryck[25]

$$SEM_{cond} = \sqrt{\frac{x(k-x)}{k-1} * \frac{1-r_{xx'}}{1-r_{21}}}$$

där x är poängnivån, k är som tidigare antalet uppgifter, $r_{xx'}$ är reliabiliteten mätt som coefficient alfa och är reliabiliteten mätt som KR21[26].

Illustrerar vi hur standardfelet varierar med provpoäng enligt de två angivna formlerna fås nedanstående figur.

---

24 Feldt, Steffen & Gupta N. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement Vol. 9, pp.* (351–361). http://conservancy.umn.edu/bitstream/102190/1/v09n4p351.pdf

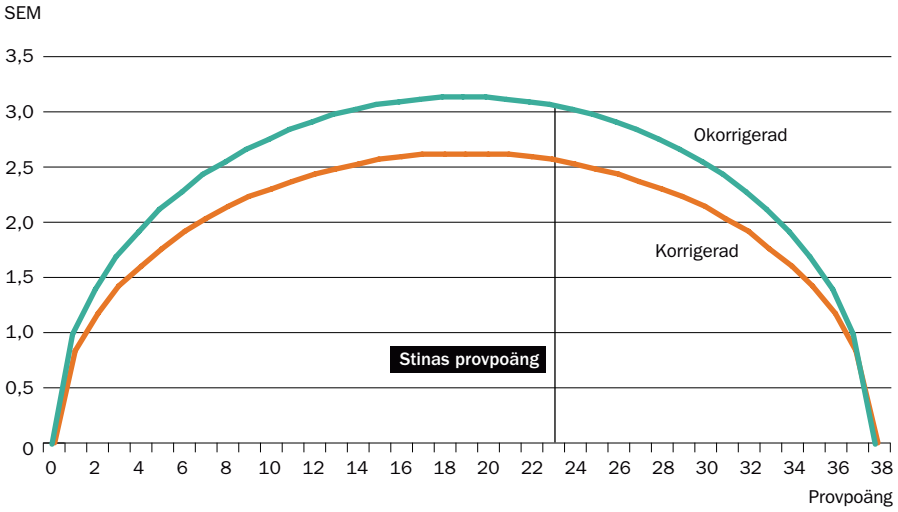
25 Den kallas ”Lord’s binominal approach: Keats’ modification”

26 Kuder Richardsons formel 21 som gäller för dikotoma uppgifter, dvs. enpoängsuppgifter.

$$KR_{21} = \frac{n}{n-1}\left(1 - \frac{M - \frac{(M)^2}{n}}{SD^2}\right)$$

n = number of items on the test
M = mean score on the test
SD^2 = Variance of scores (the standard deviation squared)

Figur 5. Villkorat standardfel *SEM*-(cond) (dvs. anpassat efter provpoäng) i okorrigerad och korrigerad form. Den vertikala linjen anger Stinas poäng.



Av figur 3 framgår att den korrigerade formen genomgående ger ett något lägre värde på *SEM*. Vi bör också komma ihåg att det *SEM* (=2,462,5) som vi använde tidigare var ett genomsnittsvärde för hela poängskalan.[27] Om vi jämför värdena i figur 5 ser vi att medelvärdet av de okorrigerade *SEM*-värdena förefaller ligga närmast det tidigare genomsnittliga *SEM*-värdet medan de korrigerade värdena ligger lägre än 2,5 utom i poängintervallet 14-25. Korrigeringen innebär således att det skattade standardfelet blir något lägre.

Stinas villkorade standardfel

För Stinas del ser vi att användningen av det genomsnittliga värdet 2,5 stämmer rätt bra med det värde hon skulle fått om en poänganpassad skattning använts (ger värdet 2,6 enligt ovanstående formel med korrigering). För Stinas del blev alltså skillnaden mellan att använda genomsnittligt standardfel och villkorat standardfel försumbar. Hade Stina däremot haft en påtagligt hög eller låg poäng skulle (det villkorade) standardfelet ha blivit mindre som framgår av figur 5.

Ovanstående innebär att säkerheten vid tilldelningen av provbetyg utifrån provpoäng blir olika beroende på vid vilken poäng en betygsgräns ligger. Provpoängens felmarginal är större för elever i mitten av poängskalan än för dem som har låga eller höga poäng.[28]

Låt oss då tillämpa ovan angivna formel på samma prov som tidigare, Fy åk9, med de betygsvillkor som angetts.

27 Dvs. ett medelvärde för den blå okorrigerade grafen viktad med antalet elever med respektive poäng.

28 Det handlar alltså här inte om i vilken utsträckning själva betygsgränsen är korrekt eller inte utan enbart om provpoängens mätfel.

Standardfel vid olika villkor för provbetyg

Tabell 2 anger villkoren för olika provbetyg på provet Fy åk9.

Tabell 2. Villkor för olika provbetyg

Betyg	Villkor 1	Villkor 2
G	minst 12 poäng totalt	inget
VG	minst 22 poäng totalt	minst 8 VG och/eller MVG-poäng
MVG	minst 30 poäng totalt	minst 3 MVG poäng

Båda villkoren ska vara uppfyllda för att ett visst provbetyg ska erhållas. I det här sammanhanget nöjer vi oss för tillfället med att endast utgå från villkor 1, totalpoängen[29]. Det gängse sättet att bedöma och betygssätta provresultat utgår från villkor 1 och det är på en sådan bedömning de mätfel som redovisats baseras. Vilken betydelse för standardfelets storlek har det då att det finns ett villkor 2?

Låt oss anta att vi har tre elever, A med 11 poäng totalt, B med 21 poäng totalt och C med 29 poäng sammantaget. Alla tre ligger således en poäng under respektive betygsgräns. Hur stor kan man då bedöma att sannolikheten är för att de i själva verket (deras sanna poäng) kan ha uppnått respektive betygsgräns.

Tabell 3 nedan visar en sammanställning av sannolikhetsfördelningen av olika sanna poäng för de tre eleverna.[30]

Elev A har 11 poäng på provet. På den nivån är korrigerad och villkorad *SEM* = 2,4 poäng (jämför figur 5). Elev As observerade poäng är 11 och sannolikheten för att detta ska vara hans sanna poäng är 17 procent[31] (läs av vid 11 poäng under A i tabell 3). Elev As observerade poäng når alltså inte över gränsen för betyget G. Om vi räknar samman de röda talen i tabell 3 får vi att sannolikheten är 58 procent att det är ett korrekt beslut att ge elev A betyget IG. Men å andra sidan betyder det att sannolikheten är 42 procent (summan av de gröna sannolikheterna) att elev A har en sann poäng som är 12 poäng eller högre och alltså borde ha betyget G. Det vi kan säga om elev A är således att det för As del är något större sannolikhet för att provbetyget bör vara IG än för att det bör vara G.

Om det råkar finnas ett någorlunda stort antal elever som har 11 poäng på provet kan man således anta att knappt 60 procent av dem får det korrekta provbetyget IG medan drygt 40 procent får IG fast de kan antas vara värda betyget G (de får det falskt negativa betyget IG), men vilka individer av dem med 11 poäng som får rätt respektive fel betyg kan vi inte avgöra utifrån provresultaten.

29 Vi återkommer till villkor 2 senare.

30 Beräkningarna är gjorda i Excel som normalfördelning med vald poäng som medelvärde och *SEM*-(cond) vid den aktuella poängen som standardavvikelse (se tabellhuvud).

31 Vi antar då att As sanna poäng är normalfördelad runt hens observerade poäng (11) med standardavvikelsen (*SEM*) 2,4. Detta ger den sannolikhetsfördelning för olika sanna poäng som visas i tabell 3.

Tabell 3. Sannolikhetsfördelning av sanna poäng för tre elever med olika provpoäng (11, 21 respektive 29). Sannolikhet i procent.

	Elev	A	B	C
	Poäng	11	21	29
	SEM	2,4	2,6	2,2
	Prov-poäng	Sannolikhet (%)		
IG	0	0	0	0
	1	0	0	0
	2	0	0	0
	3	0	0	0
	4	0	0	0
	5	1	0	0
	6	2	0	0
	7	4	0	0
	8	8	0	0
	9	12	0	0
	10	15	0	0
	11	**17**	**0**	**0**
G	12	15	0	0
	13	12	0	0
	14	8	0	0
	15	4	1	0
	16	2	2	0
	17	1	5	0
	18	0	8	0
	19	0	11	0
	20	0	14	0
	21	**0**	**15**	**0**
VG	22	0	14	0
	23	0	11	0
	24	0	8	1
	25	0	5	3
	26	0	2	7
	27	0	1	12
	28	0	0	16
	29	**0**	**0**	**18**
MVG	30	0	0	16
	31	0	0	12
	32	0	0	7
	33	0	0	3
	34	0	0	1
	35	0	0	0
	36	0	0	0
	37	0	0	0
	38	0	0	0
Summa		100	100	100

Man kan av tabellen också notera att även elever som har fått poäng längre från betygsgränsen har en inte försumbar sannolikhet att av slumpskäl hamna på rätt eller fel sida om betygsgränsen i relation till sin sanna poäng.

Samma resonemang som för elev A gäller för de två andra eleverna B och C. Eftersom SEM varierar något beroende på poängnivån varierar också sannolikheterna något. Dock är de skillnader som beror på olika värden på *SEM* tämligen försumbara och den viktigaste allmänna slutsatsen blir att det finns ett inte försumbart slumpinflytande på enskilda provresultat.

Att en viss erhållen poäng ska vara ”sann” är således inte särskilt sannolikt (17, 15 respektive 18 procent för de tre exemplen). Däremot är det sannolikt att den sanna poängen ligger inom ett visst *poängintervall* som täcker in den erhållna poängen. Det intervallet kan man som framgått beräkna även om man ska vara medveten om att man arbetar med approximationer.

Hur gick det då för Stina? Tabell 4 illustrerar situationen. Stina har 23 poäng vilket ger VG på provet (utifrån villkor 1). Det finns dock en inte försumbar sannolikhet (27,9≈28 %, summan av de röda talen) att hennes sanna totalpoäng inte skulle räcka till VG. Sannolikheten för att Stinas sanna poäng å andra sidan skulle räcka till MVG på provet är dock mindre än 1 procent och kan försummas.

Tabell 4. Sannolikhetsfördelning för Stina som fått 23 poäng på provet Fy åk9.

Betyg	G						VG								MVG
Poäng	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
Sannolikhet (%)	0,4	1,1	2,4	4,7	7,9	11,4	14,2	15,3	14,2	11,4	7,9	4,7	2,4	1,1	0

Standardfel när olika villkor kombineras

I första avsnitt tittade vi endast på det slumpbaserade standardfelet *SEM* för villkor 1 i tabell 2, det vill säga då betygen enbart baserades på totala antalet poäng oberoende av typ av poäng. Frågan här är hur man ska se på slumpfelet när det finns flera villkor och om det finns någon rimlig metod att approximativt beräkna standardfelet för sådana kombinerade villkor. För det prov vi använder som exempel gäller som tabell 2 visar två villkor för både provbetyget VG och MVG. Hur beräknar vi *SEM* i sådana fall? Vi nöjer oss med att undersöka villkoren för VG. Villkoren för MVG kan hanteras på motsvarande sätt.

Villkor 1 som gäller totalpoängen har vi utrett och där har vi fått ett genomsnittligt standardfel på 2,46 vilket för enkelhetens skull avrundas till SEM_{tot}=2,5.

För betyget VG gäller som villkor 1 minst 22 poäng totalt. Villkor 2 innebär därutöver minst 8 VG- och/eller MVG-poäng (eller belägg om den benämningen används). För att hantera frågeställningen kan vi dela upp den totala poängskalan i *två* poängskalor, en för G-poäng och en för VG/MVG-poäng där totalpoängen utgör summan av poängen på de två delskalorna och motsvarar villkor 1 medan villkor 2 endast gäller den del av skalan som består av VG/MVG-poäng.

Vi kan beräkna SEM_G för den del av provet som endast gäller G-poäng och $SEM_{VG/MVG}$ för den del av provet som består av VG- och MVG-poäng. Om vi utgår från de poäng som finns på G respektive VG och MVG-nivå gäller nedanstående tabell.

Tabell 5. Parametervärden för skalan för G-poäng respektive VG/MVG-poäng

Skala	Antal poäng	Medel	Stdav	Reliabilitet (α)	SEM
G-poäng	18	13,0	3,1	0,70	1,67
VG/MVG-poäng	20	7,2	4,6	0,86	1,72
Totalt	38	20,2	7,1	0,88	2,46

Där *SEM* liksom tidigare beräknas enligt

$$SEM = s_o\sqrt{1 - r_{xx'}}$$

För G-skalan blir $SEM_G = 3{,}1\sqrt{1 - 0{,}70} = 1{,}67 \approx 1{,}7$ och för VG/MVG-skalan $SEM_{VG/MVG} = 4{,}6\sqrt{1 - 0{,}86} = 1{,}72 \approx 1{,}7$. Vi kan notera att de två skalorna har i det närmaste samma slumpfel. Vi ser också i tabell 5 att felet för totalpoängskalan enligt tidigare är $SEM_{tot} = 7{,}1\sqrt{1 - 0{,}88} = 2{,}46 \approx 2{,}5$ dvs. mindre än summan av felet för de två delskalorna (1,7+1,7=3,4). Man kan alltså inte summera mätfelen rakt av, vilket också framgår av tabell 2.

Däremot kan man approximativt utgå från sambandet[32] $SEM_{tot}^2 = SEM_G^2 + SEM_{VG/MVG}^2$ dvs.

$SEM_{tot} = \sqrt{SEM_G^2 + SEM_{VG/MVG}^2}$ detta ger $SEM_{tot} = \sqrt{1{,}67^2 + 1{,}72^2} = 2{,}40 = 2{,}4$. Detta värde kan jämföras med värdet 2,46 som vi fick för SEM_{tot} ovan. Det vill säga ungefär samma värde.

I matrisform

För att illustrera de två villkoren för VG i exemplet kan man skapa en matris med G-poäng på den vertikala axeln och VG/MVG-poäng på den horisontella. Använder man de data som finns tillgängliga fås nedanstående matris (figur 6).

Totalt ingår 2345 elever. Hur de fördelar sig på den vertikala G-skalan anges i marginalen till höger. Och hur de fördelar sig på den horisontella VG/MVG-skalan anges i nedersta raden. Varje cell anger hur många elever som haft just den kombinationen av G-poäng och VG/MVG-poäng som axlarna anger.

För att försöka tydliggöra bilden har vår tidigare elev Stinas poäng lagts in. Stina hade 23 poäng som vid närmare granskning visade sig bestå av 13 G-poäng och 10 VG/MVG-poäng. Den mörkare gröna rutan anger 12 elever, inklusive Stina, som hade 23 totalpoäng bestående av 13 G-poäng och 10 VG/MVG-poäng.

32 Se t.ex. Kane (2008)

Figur 6. Antal elever med olika poängfördelning på G- och VG/MVG skalorna. Den trappformade linjen anger gränsen för villkor 1 för VG. Den vertikala orangea linjen anger gränsen för VG enligt villkor 2. Den mörkare turkosa rutan anger Stinas poäng (13 G- och 10 VG/MVG-poäng). De ljusare gröna rutorna illustrerar övriga kombina-tioner som ger 23 poäng.

G poäng	VG- och MVG-poäng																				Total
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
0	3																					3
1																						0
2	1																					2
3	3	1																				4
4	5	1																				6
5	9	5	1		1																	16
6	8	5	5	4	1		1															24
7	15	12	8	4	4	4	2	1			1				1							52
8	16	13	11	13	16	12	3	1														85
9	12	16	12	25	19	13	11	9	2	1	3	1	1									125
10	7	23	29	33	25	20	20	18	8	4		0	1	1								189
11	7	22	25	37	18	24	23	18	11	7	6	4	1									203
12	5	21	25	35	24	38	25	31	22	9	13	7	2	1								258
13	10	13	11	16	27	26	37	27	27	26	12	17	8	4	2	2	0	1				266
14		4	16	16	13	27	33	31	30	36	41	19	16	7	10	2	4	2				307
15	3	4	5	1	9	15	19	17	26	32	21	28	22	17	8	20	5	1	2	1		256
16	2	1	4	4	5	6	22	12	22	27	19	22	23	18	27	14	14	10	1	3	1	257
17		2	1	3	2	3	8	8	15	5	14	20	18	18	16	21	14	20	9	4		201
18						3		1		3	7	8	6	6	9	15	9	7	7	2	8	91
Total	106	143	153	191	164	191	205	174	163	150	137	126	98	72	73	74	46	41	19	10	9	2 345

Den trappformade linjen i figuren anger gränsen för villkor 1, minst 22 poäng totalt. De rutor som ligger ”under” trappan klarar villkor 1, de som ligger ovanför uppfyller inte villkor 1. Eftersom det finns många sätt att kombinera G- poäng med VG/MVG-poäng så att summan blir 22 blir gränsen trappformad. Längst ner till vänster i trappan består de 22 totalpoängen av 18 G-poäng och 4 VG/MVG-poäng, och i övre högra delen av trappan utgörs de av 2 G-poäng och 20 VG/MVG-poäng. Notera dock att inga elever hade dessa extrema värden.

Den vertikala röda linjen anger villkor 2, minst åtta VG/MVG-poäng. Villkoret uppfylls av de celler som ligger till höger om linjen.

De två villkorsgränserna delar in ytan i fyra fält. Det övre vänstra fältet representerar de elever som varken uppfyller villkor ett eller två. De uppfyller alltså inget villkor för VG. Den nedre vänstra delen under trappan representerar de elever som uppfyller villkor 1 men inte villkor 2 (de har minst 22 totalpoäng men inte åtta eller fler VG/MVG-poäng). De är förhållandevis få.

Eleverna i den övre högra delen är de elever som uppfyller villkor 2 men inte villkor 1. Även dessa elever är förhållandevis få. I nedre högra delen återfinns slutligen de elever som uppfyller båda villkoren och alltså får provbetyget VG (eller MVG om de också uppfyller de villkor som gäller för detta betyg). Man kan notera att en stor majoritet av eleverna antingen inte uppfyller båda

villkoren eller uppfyller båda villkoren. De som uppfyller endast ett av villkoren är förhållandevis få.

Det vi främst är intresserade av i det här sammanhanget är vilken betydelse slumpfelen (*SEM*) har vid en villkorad bedömning. Eftersom det finns två villkor finns det två slumpfel inblandade, ett för G-skalan och ett för VG/MVG-skalan. Enligt ovan råkar båda felen bli ungefär lika stora 1,7 poäng, vilket får ses en tillfällighet. Eftersom *SEM* antas representera standardavvikelsen för ett normalfördelat slumpfel innebär det att Stinas sanna poäng med 68 procent sannolikhet ligger mellan 13±1,7 poäng på G-skalan och med samma sannolikhet mellan 10±1,7 poäng på VG/MVG-skalan. Vill man ha 95 procents säkerhet blir felet 1,7*1,96poäng. Tabell 6 sammanfattar resultaten

Tabell 6. Konfidensintervall för respektive poängskala.

Skala	Nedre 95 %	Nedre 68 %	Obs poäng	Övre 68 %	Övre 95 %
G	9,7	11,3	13	14,7	16,3
VG/MVG	6,7	8,3	10	11,7	13,3

Lägger man in dessa gränser i figur 6 fås nedanstående figur

Figur 7. Poängfördelning, betygsgränser samt felintervall. Heldragen linje 68 procent konfidensintervall, streckad 95 procent. Talen anger antal elever med respektive poäng i det använda materialet.

G poäng	VG- och MVG-poäng																					Total
	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
0	3																					3
1																						0
2	1																					2
3	3	1																				4
4	5	1																				6
5	9	5	1		1																	16
6	8	5	5	4	1		1															24
7	15	12	8	4	4	4	2	1			1				1							52
8	16	13	11	13	16	12	3	1														85
9	12	16	12	25	19	13	11	9	2	1	3	1	1									125
10	7	23	29	33	25	20	20	18	8	4		0	1	1								189
11	7	22	25	37	18	24	23	18	11	7	6	4	1									203
12	5	21	25	35	24	38	25	31	22	9	13	7	2	1								258
13	10	13	11	16	27	26	37	27	27	26	12	17	8	4	2	2	0	1				266
14		4	16	16	13	27	33	31	30	36	41	19	16	7	10	2	4	2				307
15	3	4	5	1	9	15	19	17	26	32	21	28	22	17	8	20	5	1	2	1		256
16	2	1	4	4	5	6	22	12	22	27	19	22	23	18	27	14	14	10	1	3	1	257
17		2	1	3	2	3	8	8	15	5	14	20	18	18	16	21	14	20	9	4		201
18						3		1		3	7	8	6	6	9	15	9	7	7	2	8	91
Total	106	143	153	191	164	191	205	174	163	150	137	126	98	72	73	74	46	41	19	10	9	2 345

Figur 7 ger en bild av säkerheten i provresultaten när de är uppdelade på två poängskalor. Den heldragna röda ellipsen representerar *SEM*=1,7 poäng och den streckade 1,96**SEM*, dvs. den linje som med 95 procents sannolikhet anger gränsen för Stinas sanna poäng. Vi ser att här liksom i det fall då vi endast utgick från Stinas totalpoäng att det finns en betydande sannolikhet för att hennes sanna poäng inte uppfyller villkoren för VG (den del som ligger ovanför trappan). Det är dock svårt att mer exakt uppskatta hur stor sannolikheten är ur figur 7. För att göra detta kan man i stället utgå från figur 8.

Här är sannolikhetsfördelningarna angivna (i procent) som marginalfördelningar för respektive poängskala och även illustrerade i två stolpdiagram. Liksom tidigare är det Stinas poäng som är markerad, vilket ger medelvärdet 13 på G-skalan och 10 på VG/MVG-skalan.[33]

För båda skalorna gäller enligt tidigare att standardavvikelsen (*SEM*) är 1,7 poäng. De två skalorna betraktas som oberoende och sannolikheten för Stinas sanna poäng fås genom att multiplicera sannolikheten för G-poängen med sannolikheten för VG/MVG-poängen för respektive ruta.

Vi ser att sannolikheten för att Stinas sanna poäng ska vara 23 poäng utifrån den kombination av poäng hon har på provet är 5,4 procent[34]. Tidigare (tabell 4) fann vi att sannolikheten för 23 poäng var 15 procent baserat på den sammanslagna skalan för totalpoäng. Enligt figur 6 finns det dock flera olika kombinationer av G- och VG/MVG-poäng som ger 23 totalpoäng, t.ex. 14+9, 15+8, 12+11, 10+12 osv.

Lägger vi samman sannolikheterna för de rutor som ger 23 poäng i figur 8 får vi i procent 0,3+1,4+3,9+5,4+3,9+1,4+0,3= 16,6%. Detta kan jämföras med 15,4 procent som var sannolikheten enligt tabell 4. Värdena är i princip samma, men olika avrundningsfel ger en skillnad på drygt en procentenhet. Att de blir samma värden är inte förvånande då matrisen endast är en uppdelning av totalpoängen i två dimensioner.

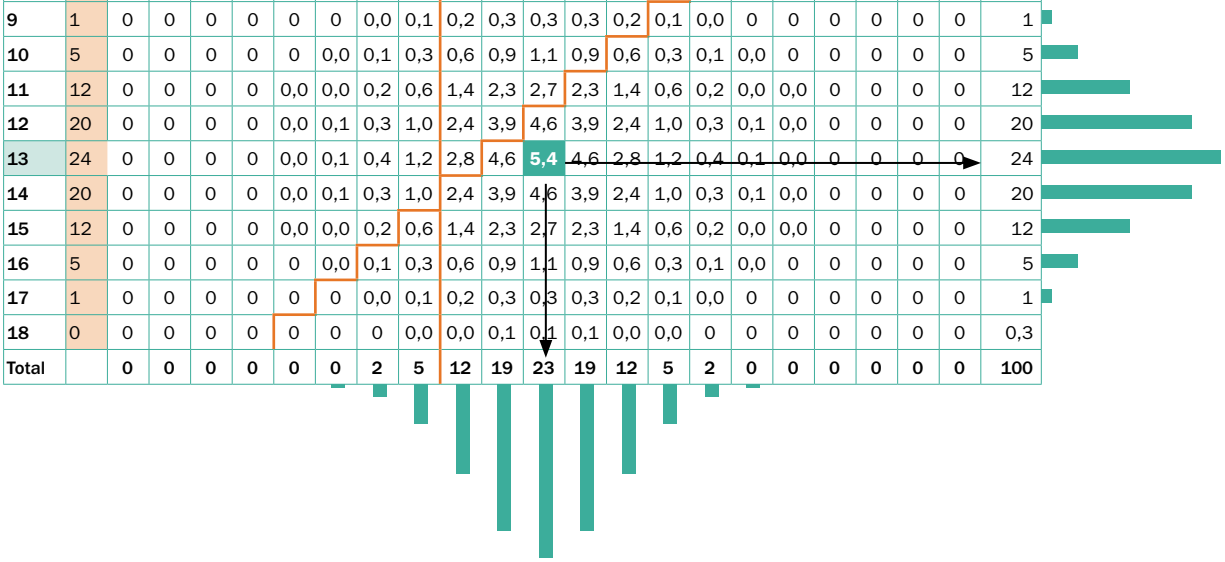
---

33 Eller någon annan av de 12 elever som hade 23 poäng i Stinas kombination. Eftersom vi utgår för att samma standardfel gäller för hela skalan för båda poängtyperna kan sannolikhetsmönstret förflyttas till vilken cell som helst i matrisen.

34 Under antagande om oberoende gäller att sannolikheten fås som 0,24 * 0,23 = 0,0555,5%. (Avrundningsfel ger 5,5% jämfört med 5,4% som figuren anger baserat på fler decimaler.). Oberoende innebär att en viss poäng på skala G-poäng inte är beroende av en viss poäng på skala VG/MVG-poäng och omvänt. Poängen antas således vara *lokalt* oberoende (jämför antagande vid användning av IRT). Däremot finns förstås en positiv korrelation mellan de två skalorna.

Figur 8. Sannolikhetsfördelning i procent för en elev som har 13 G-poäng och 10 VG/MVG-poäng, dvs. Sannolikheten för att en elev med 23 poäng har sin sanna poäng i rutan som består av 13 G- och 10 VG/MVG-poäng är 5,4 procent.

G poäng		VG- och MVG-poäng																					Total
		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	
	p	0	0	0	0	0	0	2	5	12	19	23	19	12	5	2	0	0	0	0	0	0	100
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0,0	0,0	0,0	0,0	0,0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0,0	0,0	0,1	0,1	0,1	0,0	0,0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	0	0	0,0	0,1	0,2	0,3	0,3	0,3	0,2	0,1	0,0	0	0	0	0	0	0	1
10	5	0	0	0	0	0	0,0	0,1	0,3	0,6	0,9	1,1	0,9	0,6	0,3	0,1	0,0	0	0	0	0	0	5
11	12	0	0	0	0	0,0	0,0	0,2	0,6	1,4	2,3	2,7	2,3	1,4	0,6	0,2	0,0	0,0	0	0	0	0	12
12	20	0	0	0	0	0,0	0,1	0,3	1,0	2,4	3,9	4,6	3,9	2,4	1,0	0,3	0,1	0,0	0	0	0	0	20
13	24	0	0	0	0	0,0	0,1	0,4	1,2	2,8	4,6	**5,4**	4,6	2,8	1,2	0,4	0,1	0,0	0	0	0	0	24
14	20	0	0	0	0	0,0	0,1	0,3	1,0	2,4	3,9	4,6	3,9	2,4	1,0	0,3	0,1	0,0	0	0	0	0	20
15	12	0	0	0	0	0,0	0,0	0,2	0,6	1,4	2,3	2,7	2,3	1,4	0,6	0,2	0,0	0,0	0	0	0	0	12
16	5	0	0	0	0	0	0,0	0,1	0,3	0,6	0,9	1,1	0,9	0,6	0,3	0,1	0,0	0	0	0	0	0	5
17	1	0	0	0	0	0	0	0,0	0,1	0,2	0,3	0,3	0,3	0,2	0,1	0,0	0	0	0	0	0	0	1
18	0	0	0	0	0	0	0	0	0,0	0,0	0,1	0,1	0,1	0,0	0,0	0	0	0	0	0	0	0	0,3
Total		0	0	0	0	0	0	2	5	12	19	23	19	12	5	2	0	0	0	0	0	0	100



En intressant fråga i sammanhanget är om sannolikheten för att Stina ska ha en sann poäng som *inte* uppfyller kraven för VG blir olika om provet baseras enbart på totalpoäng (villkor 1) jämfört med om det baseras på två villkor (dels totalpoäng dels att av dessa visst antal av dessa är VG/MVG-poäng).

Tabell 4, där vi enbart utgår från villkor 1, anger att sannolikheten för att Stinas sanna poäng ska ligga under 22 poäng är 27,9 ≈ 28 procent. Räknar vi i figur 8 samman sannolikheterna för de rutor som *inte* uppfyller *båda* villkoren (det vill säga rutorna till vänster om trappan *eller* tillvänster om den vertikala linjen) blir summan 27,6 ≈ 28 procent, det vill säga samma sannolikhet som i det endimensionella fallet ovan. *Sannolikheten för att Stinas sanna poäng ska ligga utanför gränsen för VG blir således densamma oberoende av om man enbart använder villkor 1 eller om man tillämpar både villkor 1 och 2*. Detta kan måhända framstå som förvånande eftersom man i det senare fallet kan säga att man har två skalor med två mätfel att ta hänsyn till. Som vi sett tidigare är mätfelen (*SEM*) för två skalor inte möjliga att addera rakt av utan *SEM* för en skala som

består av två (eller flera) olika delskalor får beräknas som kvadratroten ur summan av kvadraten på *SEM* för var och en av de två (eller flera) delskalorna.[35]

Samma tillvägagångssätt som tillämpats för kravgränsen för VG kan användas för MVG. Villkoren är förstås något annorlunda, men felmarginaler och sannolikheter kan bestämmas på motsvarande sätt som för VG. Även för MVG gäller att sannolikheten för att någon elevs sanna poäng ska ligga under eller över någon kravgräns inte påverkas av om man utgår från totalskalan eller någon form av uppdelning av skalan i delskalor. Även om andelen elever som når eller inte når en viss kravgräns är lika i de två fallen är det inte troligt att grupperna är identiska.

Den kanske viktigaste lärdomen av den något omständliga beskrivningen av konsekvenserna av att tillämpa dubbla villkor för olika provbetyg är att de ur ett mätperspektiv knappast tillför någon information. Slumpfaktorernas inverkan och därmed mätsäkerheten blir på det hela taget densamma om man använder en sammanhållen skala eller delar upp den i flera. Dock kan det finnas didaktiska och andra skäl till uppdelning.

Slumpfelens betydelse för kravgränser och provbetyg

Den enskilda mätningens standardfel *SEM* är som framgått en skattning av inom vilket intervall en enskild elevs sanna poäng ligger. Det är alltså ett mått på individnivå. När en poänggräns för ett visst provbetyg fastställs kommer vissa individer i närheten av gränsen att med viss sannolikhet hamna över eller under densamma. Om man har provresultat för en uppsättning elever är det då möjligt at skatta hur många individer eller hur stor andel av individerna som hamnar inom rätt betygsintervall, för högt eller för lågt. Däremot är det som tidigare nämnts inte möjligt att *identifiera* vilka enskilda individer det handlar om.

En approximativ beräkning

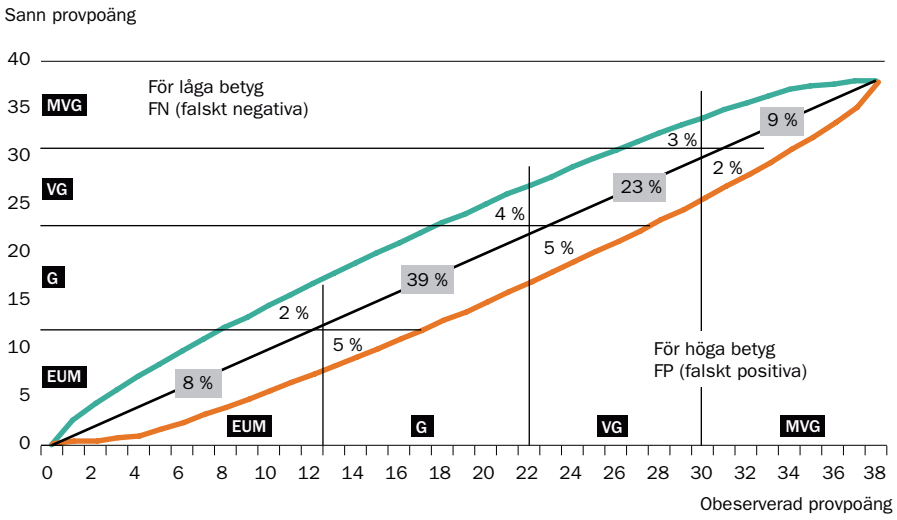
Låt oss än en gång utgår från de elever som gjort provet i fysik. I figur 9 visas observerad provpoäng på den horisontella skalan och sann provpoäng på den vertikala. Här görs approximationen att medelvärdet för den sanna poängen är lika med den observerade poängen (dvs. ingen regression mot medelvärdet antas). Detta illustreras av den räta linjen i figuren. Den blå och röda linjen anger inom vilka gränser (1,96**SEM*) som den sanna poängen fördelar sig för den aktuella provgruppen.[36] De vertikala och horisontella linjerna markerar poänggränserna för olika provbetyg.

Figuren visar a) de områden där observerad och sann provpoäng faller inom samma betygsgränser sammanfaller, b) de områden där observerad provpoäng är lägre än sann provpoäng och ger ett för lågt betyg, t.ex. EUM enligt observerad poäng, men G enligt sann poäng (*falskt negativ* eftersom eleven får för lågt betyg) och c) de områden där eleven får för högt betyg (*falskt positiv*).

35 Dvs. $SEM_{tot} = \sqrt{SEM_G^2 + SEM_{VG\_MVG}^2} \approx \sqrt{SEM_{G\_VG}^2 + SEM_{MVG}^2} \approx \sqrt{SEM_G^2 + SEM_{VG}^2 + SEM_{MVG}^2} \approx \sqrt{SEM_{Tot}^2} \approx 2{,}4$ poäng då $SEM_G = 1{,}68; SEM_{VG\_MVG} = 1{,}74;\ SEM_{G\_VG} = 2{,}30;\ SEM_{MVG} = 0{,}78;\ SEM_{VG} = 1{,}54;\ SEM_{MVG} = 0{,}78$

36 Här används villkorad *SEM* med Keats korrektion

Figur 9. Samband mellan observerad och sann provpoäng samt konfidensintervall för standardfelet. Betygsgränser och andel elever som fått korrekta betyg, för låga betyg respektive för höga betyg.



Som underlag för skattningen av den procentuella fördelningen av elever som fått betyg inom respektive område används den elevgrupp som genomfört det aktuella provet.

Som figuren visar kan cirka 79 procent av eleverna anses ha fått korrekta provbetyg, cirka 9 procent för låga provbetyg (summan av de falskt negativa) och cirka 12 procent för höga (summan av falskt positiva), det vill säga sammantaget har cirka 21 procent av eleverna fått fel provbetyg beroende på slumpmässiga mätfel.

En mer noggrann skattning

Figur 9 baseras på den mest approximativa modellen där ingen hänsyn tas till regression mot medelvärdet och *SEM* används i sin enklaste form som ett medelvärde över hela poängskalan.

Vilken blir då skillnaden om man i stället använder en mer fullständig metod med regression mot medelvärdet (se fig. 3). Figur 10 visar utfallet

Här anges för varje observerad poäng motsvarande sann poäng beräknad med Kelleys formel.[37] För varje poäng ges då en sannolikhetsfördelning med den sanna poängen som medelvärde och *SEM* som standardavvikelse. Genom att multiplicera med det antal elever som har respektive provpoäng fås då den frekvensfördelning som illustreras i figur 10.

37 $\hat{x}_S = r_{xx'}(x_O - \bar{x}) + \bar{x}$; där $\hat{x}_S$ är den estimerade sanna poängen, $r_{xx'}$ är reliabiliteten, x_O den observerade poängen och $\bar{x}$ medelvärdet av den observerade poängen för alla i gruppen som gjort provet.

Figur 10. Elevfördelning utifrån observerad (vågrät) och sann provpoäng (lodrät) med betygsgränser. De dubbla strecken vid 13 och 29 poäng på skalan för sann poäng anger betygsgränserna för G respektive MVG uttryckt i sann poäng.

Sann provpoäng: MVG	38																																							
MVG	37																																							
MVG	36																																			1	1	1		6
MVG	35																																	1	1	2	2	3	1	1
MVG	34																															1	1	2	3	5	5	4		
MVG	33																														1	2	2	4	6	7	5	3		
MVG	32																												1	1	2	5	4	7	9	8	4	1		
MVG	31																											1	1	3	4	8	7	9	9	5	2			
MVG	30																									1	1	2	3	6	6	12	8	9	7	3				
MVG	29																								1	2	3	4	6	10	9	14	8	7	4	1				
VG	28																							1	2	4	5	6	9	13	10	13	6	4	1					
VG	27																						1	2	4	8	9	9	11	14	9	10	3	2						
VG	26																				1	1	2	5	7	12	12	11	12	12	7	6	2	1						
VG	25																			1	2	3	5	8	11	16	14	12	10	9	4	3	1							
VG	24																		1	1	3	5	8	12	14	19	15	10	8	6	2	1								
VG	23																	1	2	3	7	9	11	16	16	19	13	8	5	3	1									
VG	22																1	2	4	5	11	13	14	18	16	16	9	5	2	1										
G	21															1	2	4	8	9	15	16	16	17	13	12	6	2	1											
G	20														1	2	4	7	12	12	18	17	15	14	10	7	3	1												
G	19													1	2	4	7	11	17	14	19	16	12	10	6	4	1													
G	18													2	5	8	10	15	20	15	18	13	8	6	3	2	1													
G	17												1	4	9	12	14	17	20	13	14	9	5	3	1	1														
G	16											1	3	8	14	16	16	17	17	10	9	5	3	1	1															
G	15										1	2	5	12	18	18	16	15	13	7	6	3	1	1																
G	14									1	2	3	7	15	20	18	13	11	9	4	3	1																		
G	13									2	3	5	9	17	18	15	10	7	5	2	1																			
EUM	12								1	3	5	6	10	15	15	10	6	4	2	1																				
EUM	11							1	2	5	7	7	9	12	10	6	3	2	1																					
EUM	10							1	3	6	7	6	7	8	6	3	1	1																						
EUM	9							2	4	6	6	4	4	4	3	1	1																							
EUM	8					1	2	3	4	5	4	3	2	2	1	1																								
EUM	7					1	2	3	3	3	2	1	1	1																										
EUM	6				1	1	2	2	2	2	1	1																												
EUM	5				1	1	1	1	1	1																														
EUM	4				1	1	1																																	
EUM	3	1																																						
EUM	2	2																																						
EUM	1																																							
EUM	0																																							
Poäng		0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38
Provbetyg		EUM												G										VG								MVG								
		Observerad provpoäng																																						
Antal elever		3	0	0	2	5	8	12	20	33	37	38	58	101	121	116	102	114	130	97	126	112	101	115	105	121	91	71	69	78	53	75	42	45	40	31	18	12	1	8

Som framgår av figuren ändras ett par betygsgränser på skalan för sann poäng (EUM/G och VG/MVG) då regression mot medelvärdet tillämpas.[38] Man ser också att den sanna poängskalan trängs samman mot mitten i enlighet med figur 4 (sid. 14). På samma sätt som för figur 9 kan man räkna samman hur många elever som har korrekta betyg, falskt positiva betyg respektive falskt negativa betyg. Tabell 7 visar resultatet.

38 På grund av regression mot medelvärdet ger alla observerade poäng en sann poäng som avviker från den observerade, men eftersom provpoängen anges i heltal avrundas den sanna poängen också till heltal

Tabell 7. Antal och andel elever som fått korrekta respektive felaktiga provbetyg.

Kategori	Antal	Procent
Sanna betyg	1 803	77
Falskt negativa betyg	234	10
Falskt positiva betyg	307	13
Totalt	2 345	100

En jämförelse med resultaten för den mer approximativa metod som lågt till grund för figur 9 visar att andelen elever med felaktiga betyg skattas något högre med den mer fullständiga metoden. Dock är skillnaden tämligen marginell och i relation till felets storlek förefaller den försumbar. Den väsentliga slutsatsen är att för det aktuella provet med fyra betygssteg är att *cirka 20 procent av eleverna kan antas få felaktiga provbetyg* på grund av slumpmässiga mätfel.[39]

En något större andel tycks få för höga än för låga provbetyg. Om det är en tillfällighet för just det här provet eller ett mer generellt fenomen kan inte avgöras utan närmare analys av flera prov. Det bör för övrigt påpekas att det aktuella provet väl uppfyller gällande kriterier för den aktuella typen av prov så slumpfelets storlek kan ses som normalt för nationella och andra prov av motsvarande typ.

Om syftet inte är att betygssätta enskilda elever utan handlar om resultat på gruppnivå i form av till exempel betygsmedelvärden blir som tidigare nämnts betydelsen av slumpmässiga mätfel avsevärt mindre eftersom positiva och negativa fel tar ut varandra vid medelvärdesberäkningar. I det här aktuella fallet blir skillnaden mellan andelen falskt positiva och falskt negativa ett par tre procentenheter.

39 I engelskspråkig litteratur tala man om att provets ”accuracy” är 80 procent eller 0,80. Accuracy är svåröversatt, möjligen kan man tänka sig ”riktighet” eller ”träffsäkerhet” för att inte blanda samman med redan befintliga begrepp.

Sammanfattande kommentar

Den här promemorian har behandlat slumpens betydelse för resultat på poängbaserade prov med utgångspunkt i den klassiska testteorin. Denna bygger på antaganden och förenklingar som innebär att resultaten måste betraktas som approximativa. Olika korrigeringar kan användas för att minska graden av approximationer men vinsterna är förhållandevis små och slutsatsen blir att det för normala behov är fullt tillräckligt att utgå från grundformlerna. *Den viktigaste slutsatsen är att en enskild individs provpoäng alltid är behäftad med ett slumpmässigt* mätfel. Mätfelets storlek för en specifik individ är alltid okänt, men kan med hjälp av statistik baserad på resultaten för den deltagande gruppen tilldelas en sannolikhet som kan användas vid bedömningen av den enskilda elevens resultat. Sådana bedömningar görs inte idag regelmässigt för nationella prov i Sverige och behöver kanske inte införas i den offentliga resultatredovisningen, men standardfelen *SEM* bör alltid finnas tillgängliga i tekniska redovisningar av provresultaten så att den som så önskar kan bedöma provpoängens och provbetygens tillförlitlighet.[40] Inte minst finns det anledning att ge alla dem som berörs av nationella prov möjlighet att få en realistisk uppfattning om hur tillförlitliga, eller otillförlitliga, provresultat är.

En intressant slutsats ur ett mer tekniskt perspektiv är att uppdelning av provpoängen i olika villkor för olika betyg till exempel si och så många G-poäng respektive VG/MVG-poäng inte tycks påverka mätfelets storlek nämnvärt. Vi såg också (figur 7) att det var förhållandevis få elever som påverkades av villkoren. Den absoluta majoriteten av elever uppfyllde inget av villkoren eller båda. Fördelen med att införa villkor av den typ som finns i det prov som använts som exempel ligger knappast i att de ökar mätsäkerheten. Däremot kan de kanske ha pedagogiska eller didaktiska fördelar, vilket inte bedöms i det här sammanhanget. Den preciserade bedömningen i belägg på itemnivå tycks dock bidra till att höja reliabiliteten och innebär därmed ur mätsäkerhetssynpunkt en fördel gentemot en mer helhetlig bedömning med flera poäng på uppgiftsnivå.

Det bör observeras att de ämnen som tagits upp här endast utgör en liten del av de frågor som måste hanteras vid konstruktion och analys av prov. Till exempel ingår inga frågor om validitet, om de uppgifter som ingår på ett lämpligt sätt representerar kursplanens kunskapsområden och de förmågor som kunskapskraven föreskriver, vilka slutsatser man kan dra av en elevs provpoäng osv. Detta kräver en egen promemoria.

Inte heller ingår den för provbetygens del mycket viktiga frågan om hur kravgränserna konstrueras och bestäms. Hur kunskapskravens verbala föreskrifter ska transformeras till poängbaserade prov är en annan knäckfråga. De fel som behandlats här baseras på slump. Fel i kravgränssättningen däremot är systematiska och får effekt både på individ- och gruppnivå. Validitet och kravgränssättning är centrala frågor som behöver diskuteras och till dem får vi återkomma. Men då är Kahnemans ekvation värd att bära med sig.

framgång = skicklighet + tur

40 Se också AERA, APA & NCME. (1999).

Referenser

AERA, APA & NCME. (1999). *Standards for educational and psychological testing.* Washington: AERA.

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

Feldt, L.S. & Brennan. R.L.(1989). Reliability i Linn R.L. Educational measurement. Third edition.New York: Macmillan Publishing Company

Feldt, L.S., Steffen, M. & Gupta, N.C. (1985). A comparison of five methods for estimating the standard error of measurement at specific score levels. *Applied Psychological Measurement Vol. 9, pp. (351-361)*.

Harvill, L. M. (1991). Standard error of measurement. *Educational Measurement: Issues and Practice*, v10 n2 p33-41. http://ncme.org/linkser-vid/6606715E-1320-5CAE-6E9DDC581EE47F88/showMeta/0/

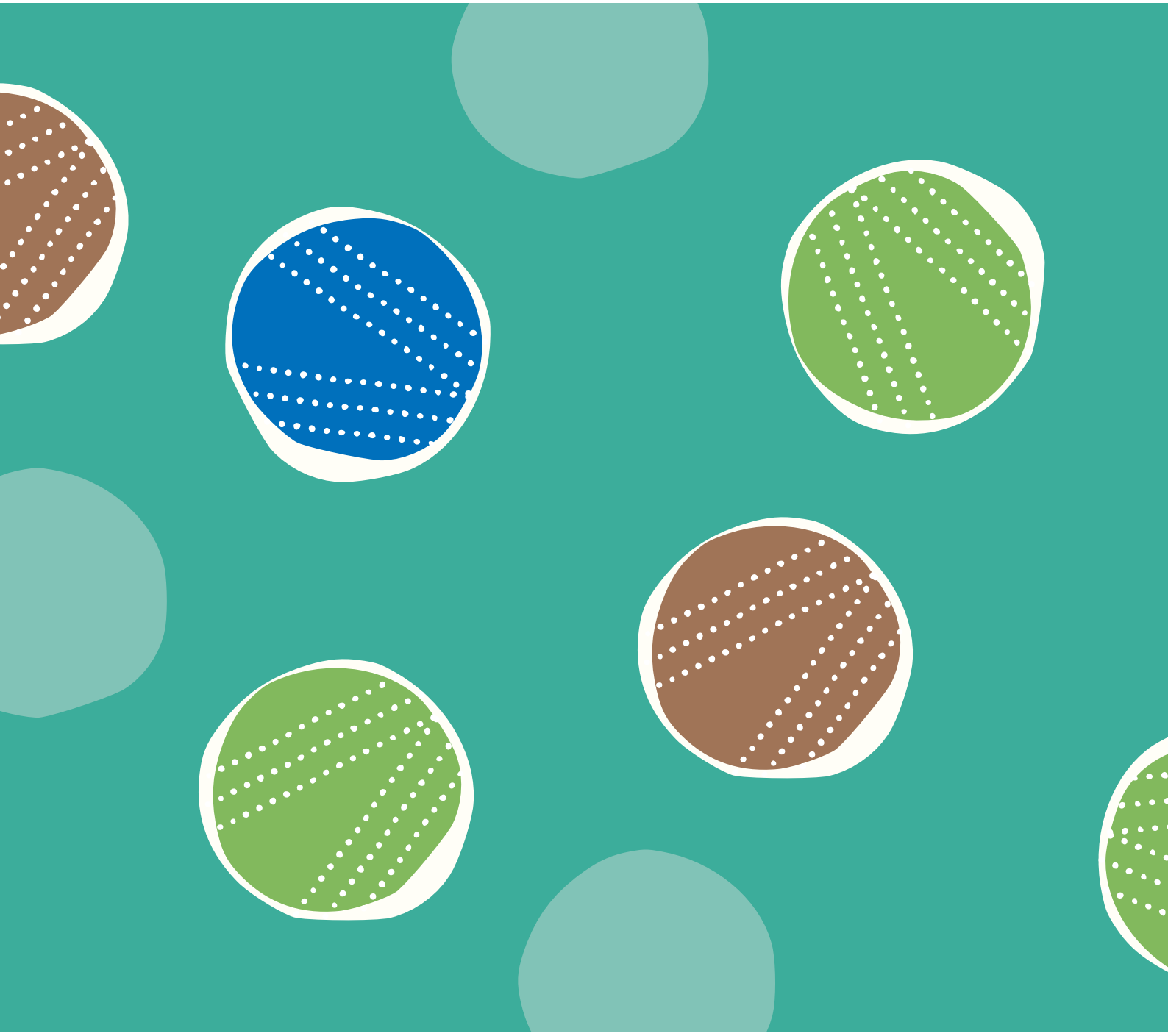
Kahneman, D. (2012). *Att tänka, snabbt eller långsamt.* Stockholm: Volante

Kane, M. (2008). *Errors of measurement, theory, and public policy*. Princeton: ETS http://www.ets.org/research/policy_research_reports/publications/publication/2008/hsbr

Kane, M. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement Spring 2013, Vol. 50, No. 1, pp. 1–73.*

Lord, F.M. & Novick, M. R. (1967). *Statistical theories for mental test scores.* London: Addison-Wesley Publishing Company.

Skolinspektionen (2013). *Olikheterna är för stora.* Stockholm: Skolinspektionen. http://www.skolinspektionen.se/Documents/omrattning/omrattning-nationella-prov-2013.pdf



Skolverket

www.skolverket.se