Analys av flervalsuppgifter som examinationsform.

Jonas R. Persson

Program for lærerutdanning Norges Tekniske og Naturvitenskaplige Universitet, NTNU, 7491 Trondheim

Abstract

In higher education, multiple choice tests are becoming increasingly used. This poses a problem when applying the same marking scheme as for ordinary questions, as guessing becomes profitable. I examine the effect of guessing in multiple choice tests and compare different marking algorithms. Different algorithms are compared using a 100-item standard test, where the true knowledge levels are added to the result of guessing on other items. The result show that guessing increases the probability for a higher grade using a dichotomous algorithm.

Abstrakt

I högre utbildning har användandet av flervalsuppgifter för examination ökat. Detta medför problem när man använder samma betygsgränser som för textbaserade uppgifter, då gissning blir lönsamt. Här undersöks effekterna av gissning i fallet med flervalsuppgifter och jämförelser mellan olika poängsättningsalgoritmer. Olika algoritmer jämförs i fallet med ett standardtest med 100 frågor, där kunskapsnivån ger en poäng och resten av frågorna besvaras slumpmässigt för att imitera gissning. Resultatet är att gissning ger en icke-försumbar påverkan på sannolikheten att få ett betyg som inte svarar mot kunskapsnivån i fallet med en dikotom poängsättning.

Introduktion

Sedan flervalsuppgifter först utvecklades av Frederick Kelly 1915 och med senare förbättringar har flervalsuppgifter fått en ökad utbredning och användning inom bland annat intelligenstester och andra typer av tester med höga deltagarantal. Med ett ökande antal studenter är det frestande att välja att ha examination enbart eller till stora delar baserad på flervalsuppgifter för att bland annat minska arbetsbelastningen vid rättningen. Även det ökade bruket av digitala examina främjar flervalsuppgifter genom möjligheten till automatisk rättning, i tillägg kan flervalsuppgifter ha andra positiva effekter sett ur testdeltagarnas synvinkel. Man kan sammanfatta fördelarna med att använda flervalsuppgifter som:

- Det tar kortare tid för studenterna att besvara flervalsuppgifter, jämfört med utredande eller beräkningsfrågor. (Tidsaspekten)
- En test kan, genom att antalet frågor kan ökas, täcka pensum i högre grad. (Omfattningsaspekten)
- Genom att använda flervalsuppgifter minskar skrivarbetet för deltagarna. (Insatsaspekten)
- Det går fortare att värdera svaren. Om examen är digital kan resultatet erhållas i stort sett omedelbart. (Rättningsaspekten)
- Värderingen är helt objektiv. (Objektivitetsaspekten)

Men med dessa fördelar får man i tillägg ett antal nackdelar, som man måste beakta i motsvarande grad.

- Frågan är vad man testar med givna svarsalternativ. Det finns en risk att det inte är lärande som premieras i första hand utan logiskt tänkande eller rena faktakunskaper. (Testaspekten)
- Uppgifterna måste formuleras på ett genomtänkt och logiskt sätt. Använder man sig av ett rätt svar, så måste alla alternativen vara rimliga. Uppenbara fel i hur dom felaktiga svaren är formulerade ändrar förutsättningarna. (Svarsalternativaspekten)
- Det finns en risk för att studenterna gissar. (Gissningsaspekten)

Formuleringen av uppgifterna är sålunda lika viktig som formuleringen av svarsalternativen. Om man inte noga tänker igenom dessa kommer man i många fall att göra så att testdeltagarna lätt kan se vilka svarsalternativ som är fel (svarsalternativaspekten). I tillägg måste svarsalternativen vara så väl logiskt formulerade som att dom tar hänsyn till vanliga (förväntade) fel, baserade på missuppfattningar eller alternativa föreställningar (testaspekten). Testaspekten omfattar även problemet med att det inte är möjligt att ha utredande eller flerstegsfrågor (exempelvis problemlösning som kräver flera steg) där man vill se hur man går tillväga för att lösa en uppgift. Det som jag kommer att gå mer in i detalj på är fenomenet med gissning och vilka konsekvenser det får för resultaten på en test (gissningsaspekten). Detta kommer att höra starkt samman med hur svaren värderas, det vill säga hur poängsättningen ser ut. Där en viss typ av poängsättning gynnar en strategi med gissningar eller inte, och om testdeltagarna är medvetna om detta. Detta beroende på att medvetenhet om poängsättningsalgoritmen påverkar testdeltagarnas strategier.

Det vanligaste och enklaste sättet att poängsätta är att göra det dikotomt, det vill säga rätt svar ger poäng och fel eller blankt svar ger inga poäng. Denna enkla metod har kritiserats på grund av sina inneboende svagheter. Abu-Sayf (1979) presenterar fyra argument mot dikotom rättning i förbindelse med gissning:

Det *psykometriska* argumentet är att denna poängsättning uppmuntrar gissande och att man inte tar direkt hänsyn till partiell kunskap.

Det *pragmatiska* argumentet är att testdeltagare som inte tar risker straffas gentemot risktagare, genom att gissningar och därmed ett riskbeteende belönas.

Det *moraliska* argumentet är att det är fel att gissa och att man därför inte skall uppmuntra till det.

Det *politiska* argumentet är att om man direkt eller indirekt uppmanar testdeltagare att gissa gör att de kommer att tappa tilltron till testmetoden.

Dikotom poängsättning har kritiserats för att den inte kan ge ett direkt samband mellan resultatet och den kunskap testdeltagarna har. Den information som man får är inte absolut utan relativ och ger en rankning. Detta gör det problematiskt att använda denna i fall där resultaten är viktiga, som exempelvis vid examination (Abu-Sayf, 1979).

Svagheterna har gjort att det har utvecklats en rad alternativa poängsättningsalgoritmer. Dessa poängsättningsalgoritmer försöker att lösa många av problemen med dikotom poängsättning och få en bättre bild av testdeltagarnas kunskaper och färdigheter. Dessa inkluderar algoritmer för att motverka gissning och algoritmer för att på ett mer effektivt sätt belöna partiell kunskap. I teorin skall dessa metoder ge en ökad validitet och pålitlighet för testresultatet och gynna de deltagare som annars straffas för att dom inte är risktagare eller strategiska i sitt testbeteende. Det finns ett antal publicerade artiklar om olika rättnings-algoritmer och deras påverkan på validiteten hos tester, men effekten på betygsgränser är ett område där det inte finns många publicerade studier, det som oftast anges är sannolikheten för att få godkänt genom ren gissning på alla uppgifter. Ett undantag är Burton (2001) som adresserar problemet med gissning och val av frågor i ett test med 60 uppgifter, och beräknar effekten av dessa på en testdeltagare på 50% kunskapsnivå. Campbell (2015) har gjort en liknande analys, om än inte lika detaljerat med avseende på betygsgränser, för tester i kemi.

I denna artikel kommer jag att diskutera några poängsättningsalgoritmer och göra en jämförelse mellan dikotom och gissningskorrigerad poängsättning för ett standardtest. Jämförelsen kommer att göras utgående från en antagen kunskapsnivå och gissningar på övriga uppgifter. Påverkan på betyget i en tänkt examenssituation genomförs också för att undersöka hur stor sannolikhet det är för olika kunskapsnivåer att uppnå ett visst betyg. Detta för att matematiskt testa påståendet: ”Risikoen for att en elev får en kunstig høy skåre på en test med 50 enkelt oppgaver eller mer, er meget liten.” (Sirnes, 2005, sid 45). Den övergripande frågan som skall besvaras är: Hur stor sannolikhet är det att en student som inte har tillgodogjort sig tillräckliga kunskaper skall få ett högre

betyg än som svarar till studentens kunskapsnivå vid bruk av olika poängsättningsalgoritmer som respons på gissning.

Gissningskorrigerad poängsättning

Genom att gissning är ett av problemen när det gäller flervalsuppgifter, är detta något som man i många fall vill undvika. Dock finns det inte någon riktigt bra lösning för att helt undvika gissning. I tillägg handlar det om att testdeltagare kommer att ha olika strategier och självförtroende. Om man inte vill avge ett felaktigt svar och avstår att svara straffas man i fallet med dikotom rättning, jämfört med om man chansar och gissar.

En metod för att korrigera för gissning är att betrakta sannolikheten för att avge rätt svar vid gissning och korrigera för detta. Om vi ansätter att antalet korrekta svar är R och antalet felaktiga svar är W, med c som antalet svarsalternativ på varje uppgift, kan vi beräkna den totala poängsumman S som:

$$S = R - \frac{W}{c-1}$$

Detta är den så kallade konventionella gissningskorrigerade metoden (Davies 1964). Här måste man observera att den fungerar bäst då alla felaktiga svar är baserade på gissning och att alla svarsalternativ är lika attraktiva för testdeltagarna. Detta vet vi dock inte är fallet vid många tillfällen, då svarsalternativen kan vara dåligt formulerade (svarsalternativaspekten). Med komplikationen att denna formel inte till fullo kan korrigera för gissning. Dock kommer information till testdeltagarna om rättningsalgoritmen inte uppmuntra till gissning utan att hellre avstå från att svara, för att undvika avdrag.

Ett alternativ till denna metod, är att ta hänsyn till obesvarade frågor, O, genom att även ge poäng för dessa (Gulliksen 1950):

$$S = R + \frac{O}{c}$$

Denna metod korrigerar strikt sett inte gissning , utan uppmuntrar istället att utelämna svaret när man inte vet eller är osäker. Detta genom att belöna utelämnade svar proportionellt mot sannolikheten att få rätt svar vid gissning, där man är garanterad poäng vid utelämnat svar. Här måste man vara observant på att förväntansvärdet ändras, att jämföra med ovanstående där förväntansvärdet är R.

Man kan även kombinera dessa och använda en algoritm som:

$$S = R + \frac{O}{c} - \frac{W}{c-1}$$

Här förstärker man ett beteende där gissningar inte lönar sig ytterligare. Men det är viktigt, som tidigare nämnts att alla svarsalternativ är lika attraktiva för att undvika en bias. Något som stället mycket stora krav på konstruktören. En felaktighet i ett svarsalternativ så som orimlig enhet eller felaktigt årtal gör att sannolikheten för att testdeltagarna väljer den minskar betydligt. Detta medför att olika algoritmer för poängsättning förlorar sin giltighet när det gäller att korrigera för gissning.

Ett alternativ till dessa metoder är att ta i bruk Item Respons Theory (IRT) (Crocker & Algina, 1986). Denna baseras på en testdeltagares sanna nivå och sannolikheter på att deltagaren skall svara rätt. Detta är dock svagheten i och med att vi inte vet den sanna nivån utan att man måste uppskatta värden på denna, vilket gör metoden osäker. Detta gör att IRT sällan används, då det inte är speciellt praktiskt att analysera svaren med IRT.

Metoder för belöning av partiell kunskap

Ett annat problem med flervalsuppgifter är att de inte direkt tar hänsyn till partiell kunskap. Eller rättare det finns en möjlighet att få fram det genom att partiell kunskap kan göra att man eliminerar felaktiga svarsalternativ. Med andra ord, man får fram en ökad sannolikhet, jämfört med ren gissning, att svara rätt på en fråga. Här handlar det om att deltagare bör få ett sätt att tydligare visa på sin partiella förståelse. Det finns ett antal föreslagna metoder för detta.

En metod är att till varje uppgift lägga till en gradering om hur säker testdeltagaren är på att svaret är korrekt. Här kommer då rätt svar med en hög grad av säkerhet ge mer poäng än rätt svar med en lägre grad av säkerhet. Här kommer då olika testdeltagare att få olika poäng på en uppgift beroende på hur säkra på svaret som dom är. I studier av denna metod (Ebel 1965) har man dock funnit att pålitligheten och validiteten i metoden inte ökar och att personliga variabler såsom självförtroende och risktagning påverkar mycket. Detta har medfört att denna metod inte har används i större grad. Den har dock en tillämpning i diagnostiska tester.

Med digitala hjälpmedel finns också möjligheten att införa en metod som baseras på att man kan välja svarsalternativ flera gånger tills man får rätt svar. Denna metod kallas Answer-Until-Correct (AUC) (Gilman & Ferry,1972) (Hanna 1975). Det finns ett antal uppenbara fördelar med denna typ av test: Testdeltagarna får en direkt respons som gör examen till en del av lärandeprocessen. Man kan också korrigera resultaten för gissning på ett bättre sätt. I tillägg får man en större graderingsskala att arbeta med om testdeltagarna avger svar flera gånger innan dom får rätt svar. I fallet med digital examen är detta ett alternativ som bör beaktas noga. Tidigare har höga kostnader gjort att metoden

inte använts, något som nu har ändrats med en utökad möjlighet till digital examen. Men här kommer andra aspekter när det gäller formulering och respons till testdeltagarna att vara viktig.

Man kan även ändra flervalsuppgifterna på så sätt att det kan finnas flera alternativ som är korrekta. Detta gör att partiell kunskap belönas på ett enklare sätt. Dock är detta format svårare att acceptera för många testdeltagare om de inte träffat på den metoden innan och den kan upplevas som både ologisk och orättvis. Det är därför viktigt att introducera och motivera metoden tidigt i undervisningen.

Här är även rankingsuppgifter möjliga som ett alternativ, där testdeltagarna skall rangera alternativ baserat på sina kunskaper. Här kan det röra sig om rangering om händelser och när dom inträffat eller rangering av exempelvis materialegenskaper eller olika förutsättningar. Även här är möjligheten med digital examen en fördel, då det blir lättare att besvara dom.

Man skall i tillägg vara medveten om att man kan kombinera flera av dessa metoder så att det passar ämnet på bästa sätt.

Standard test

Den övergripande frågan som jag önskar att besvara är: *Hur stor sannolikhet är det att en student som inte har tillgodogjort sig tillräckliga kunskaper skall få ett högre betyg än som svarar till studentens kunskapsnivå*. Detta studeras med ett teoretiskt standard test bestående av 100 uppgifter med 4 svarsalternativ. Metoden som används är att ta de statistiska sannolikheterna med lika stor sannolikhet för alla alternativ och en slumpmässig fördelning av svaren. Det handlar med andra ord om en binominalfördelning med 100 uppgifter och 4 alternativ. Sannolikheten för att uppnå ett visst antal rätta svar, beräknas genom att beakta binomalfördelningen, med ett bestämt antal positiva utfall, antal försök och sannolikheten för ett positivt utfall. Beräkningarna är enkla att genomföra då funktionen finns inbyggt i många spreadsheet program som exempelvis Excel.

För att korrigera för en kunskapsnivå kommer en kunskapsnivå på 40% motsvaras av 40 rätta svar på 40 uppgifter som adderas med en slumpmässig fördelning för resterande 60 uppgifter. Det är från detta möjligt att beräkna sannolikheten för att en deltagare som svarar helt slumpmässigt på de uppgifter som testdeltagaren inte vet svaret på skall uppnå en viss poängsumma (angett i procent). Tittar vi på väntevärdet för en testdeltagare som har kunskapsnivå 0% så blir den (100*25%)=25 rätta svar, det vill säga att det är ca 50% sannolikhet att den testdeltagaren skall få mer än 25 rätta svar på hela testen. Väntevärdet anger gränsen för 50% sannolikhet, här kan de diskreta stegen göra att väntevärdet inte är ett heltal och man får då en avvikelse från exakt 50% sannolikhet. Har vi en

testdeltagare med en kunskapsnivå på 20%, blir väntevärdet för deltagaren 20+(80*25%)=40 rätta svar på hela testen, och så vidare.

Vi kan därmed konstruera ett fullständigt set med rätta svar för olika testdeltagare med olika kunskapsnivåer. Här är det viktigt att poängtera att testen som utförs är för en strategi där alla svar som inte kan besvaras utifrån kunskapsnivån besvaras slumpmässigt samt att alla uppgifter besvaras. Detta handlar sålunda om en testdeltagare som bara gissar och inte lämnar uppgifter obesvarade. Med andra ord är detta ett extremfall. Men då det i verkligheten är möjligt att utesluta svar så kommer detta samtidigt att vara det värsta scenariot.

Standard test med dikotom poängsättning

Genom att använda en dikotom poängsättning (rätt = 1p och fel =0p) är det möjligt att räkna ut sannolikheten för att en test deltagare skall få ett visst betyg. Jag har valt att använda den betygsskala som rekommenderas på NTNU (Tabell 1) som utgångspunkt men denna kan enkelt varieras.

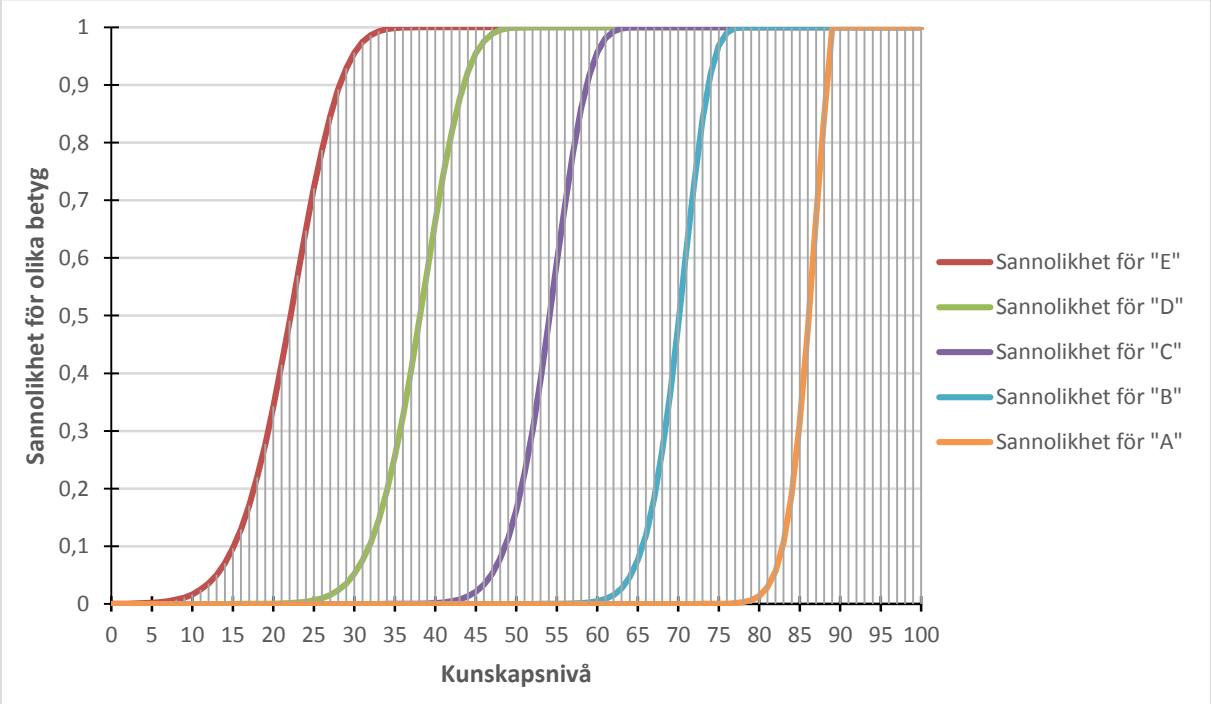
Tabell 1 Betygsgränser vid NTNU.

Betyg	Gränser (%)
A	≥89
B	77-88
C	65-76
D	53-64
E	41-52

Om vi tittar på sannolikheten att få godkänt (E) och räknar ut sannolikheten finner man att en testdeltagare med en kunskapsnivå på 23% har 57% chans att få godkänt. Uträknat för olika kunskapsnivåer ser vi sannolikheten för godkänt (”E”) i tabell 2 och figur 1.

Tabell 2 Sannolikhet för att få "E" på standardtesten.

Kunskapsnivå	Sannolikhet för «E»
2	0,000
3	0,001
4	0,001
5	0,002
6	0,003
7	0,005
8	0,007
9	0,011
10	0,017
11	0,025
12	0,035
13	0,050
14	0,070
15	0,096
16	0,129
17	0,170
18	0,220
19	0,277
20	0,343
21	0,415
22	0,491
23	0,570
24	0,647
25	0,721
26	0,787
27	0,845
28	0,892
29	0,929
30	0,956
31	0,975
32	0,986
33	0,993
34	0,997
35	0,999
36	1,000
37	1,000
38	1,000
39	1,000
40	1,000
41	1,000



Figur 1 Sannolikhet att få olika betyg mot kunskapsnivå.

Sannolikheten för att få E (godkänt) överstiger 50% redan vid 23% kunskapsnivå, medan den i praktiken är 100% (>98%) redan vid 32% kunskapsnivå (Tabell 2 och Figur 1). Här handlar det om en i realiteten sänkt gräns för godkänt. Detta innebär ett relativt stort problem genom att sannolikheten att få godkänt utan ”nödvändiga” kunskaper är relativt stor. Tittar vi på sannolikheten att få ”D” (Figur 1) är sannolikheten över 50% vid en kunskapsnivå på ca: 38%, och vid en kunskapsnivå på ca: 47% är sannolikheten i praktiken 100% (>98%).

Motsvarande siffror för 50% sannolikhet för ”C”, ”B” och ”A” är respektive ca:54 %, ca:70% och ca:86% (Figur 1). Med andra ord så ger gissning en mindre effekt vid hög kunskapsnivå. Det vill säga i fallet med NTNUs betygsgränser kommer sannolikheten att öka mest för de lägre betygen, ”E”, ”D” och till viss del även ”C”. ”A” och ”B” kommer i stort sett vara oberörda och ge en allmän indikation på svårighetsgraden på testen. I praktiken innebär detta att betygsgränserna effektivt sett kommer att ändras vilket visas i tabell 3. För att anpassa dessa till kunskapsnivån bör man då korrigera betygsgränserna enligt tabell 3.

Tabell 3 Effektiva och korrigerade betygsgränser vid dikotom rättning.

Betyg	Effektiv (%)	Korrigerade (%)
A	≥89	≥89
B	76-88	78-88
C	61-75	68-77
D	47-60	59-67
E	32-46	49-58

Standard test med gissningskorrigerad poängsättning

När det gäller betygen för gissningskorrigerad poängsättning med

$$S = R - \frac{W}{c-1}$$

så tillkommer en dimension, genom att studenterna kan lämna en uppgift obesvarad. Det är därför inte lika lätt att illustrera sannolikheterna för ett visst betyg. Om en deltagare gissar på alla uppgifter som inte kan besvaras utifrån kunskapsnivån så kommer väntevärdet för resultatet att vara samma som antalet korrekta svar. Detta för att väntevärdet för rena gissningar ($\frac{W}{c-1}$) alltid kommer att vara noll med denna algoritm. Detta innebär att för en deltagare som gissar på alla uppgifter kommer sannolikheten att klara betygsgränsen att vara 50% på kunskapsnivån. Se tabell 4. Detta för att det finns en sannolikhet att alla gissningarna totalt ger upphov till ett negativt bidrag som då subtraheras

från kunskapsnivån. En testdeltagare som inte gissar utan bara besvarar dom som svarar mot kunskapsnivån kommer då att ha 100% sannolikhet att få betyg svarande till kunskapsnivån. Med andra ord de testdeltagare som gissar kommer inte att tjäna på det utan riskerar att missa gränsen. Betygsgränserna blir med detta oförändrade.

I fallet med belöning för obesvarade uppgifter blir problematiken också mer komplicerad. Om vi tar en testdeltagare med en kunskapsnivå på 50% så innebär det att den testdeltagaren som inte gissar kommer att få en poäng på 62,5% som resultat enligt formeln:

$$S = R + \frac{O}{c}$$

I detta fall lönar det sig inte att gissa i lika hög grad, men väntevärdet blir det samma om man gissar på alla man inte kan svaret på. Här är det tydligt att gissning inte lönar sig då man är garanterad en poäng vid en obesvarad uppgift, men det finns en risk att man gissar fel och inte får poäng. Detta gör dock att den effektiva betygsgränsen ändras och man bör korrigera för detta. Om vi antar att man har 4 svarsalternativ, vilket då medför att om en testdeltagare svarar rätt på uppgifterna svarande mot kunskapsnivån och lämnar resten obesvarade kommer den effektiva gränsen för godkänt (E) att vara 22%. (22 + 78/4=41,5). De effektiva betygsgränserna för denna algoritm ges i tabell 5. De korrigerade betygsgränserna utgående från kunskapsnivån ges i tabell 6. Man bör dock notera att detta gäller om deltagarna inte gissar på några uppgifter. Samma resultat får man för den kombinerade algoritmen $S = R + \frac{O}{c} - \frac{W}{c-1}$ under förutsättning att deltagarna inte gissar.

Tabell 4 Sannolikhet för att få "E" på standardtesten med gissningskorrigerad rättning ($S = R - \frac{W}{c-1}$) och gissning på alla frågor som inte besvaras från kunskapsnivån..

Kunskapsnivå	Sannolikhet för «E»
24	0,001
25	0,001
26	0,002
27	0,004
28	0,006
29	0,010
30	0,016
31	0,025
32	0,038
33	0,056
34	0,081
35	0,113
36	0,156
37	0,209
38	0,274
39	0,348
40	0,431
41	0,520
42	0,610
43	0,697
44	0,777
45	0,845
46	0,899
47	0,939
48	0,967
49	0,984
50	0,993
51	0,997
52	0,999
53	1,000
54	1,000
55	1,000

Tabell 5 Effektiva betygsgränser vid rättning enligt algoritmen S=R+O/c utan gissningar.

Betyg	Gränser (%)
A	≥86
B	70-85
C	54-69
D	38-53
E	22-37

Tabell 5 Korrigerade betygsgränser vid rättning enligt algoritmen S=R+O/c utan gissningar.

Betyg	Korrigerade (%)	Ursprunglig (%)
A	≥91,75	≥89
B	82,75-91,50	77-88
C	73,75-82,50	65-76
D	64,75-73,50	53-64
E	55,75-64,50	41-52

Konsekvenser

Man ställs med flervalsuppgifter inför ett antal problem som bör åtgärdas. I fallet med dikotom rättning, kommer de effektiva betygsgränserna att ändras så mycket att dom inte längre är giltiga, utan borde justeras uppåt. Frågan hur man skall ställa sig till gissning är dock mer fundamental och här får man fråga sig om man vill undvika eller minska förekomsten av gissning genom att anpassa rättningsalgoritmen eller justera betygsgränserna. Ur rättvisesynpunkt är möjligheten att få ett högre betyg genom en gissningsbaserad strategi tveksam, speciellt då den i stort sett bara gynnar testdeltagare med lägre kunskapsnivå. I tillägg belönas ett risktagarbeteende som i många fall är könsrelaterat, vilket gör att kvinnor missgynnas genom ett generellt mindre risktagarbeteende.

Analysens svagheter

Denna metod att testa konsekvenserna av olika rättningsalgoritmer har ett antal svagheter.

En av dom är baserad på att gissningen som sker är helt slumpmässig. Detta är dock inte fallet i realiteten, partiell kunskap kommer att möjliggöra uteslutning av en eller flera svarsalternativ. På samma sätt kan dåligt konstruerade frågor och svarsalternativ påverka sannolikheten för valet av svarsalternativ. Här kommer då sannolikheten för att rätt svarsalternativ hittas att öka och därmed kommer även sannolikheten att få fler poäng också att öka. Analysen är på grund av detta konservativ och visar på lägsta nivån. Det vill säga att sannolikheterna i realiteten är högre.

Antalet svarsalternativ påverkar i tillägg sannolikheterna. Med fem svarsalternativ istället för fyra blir sannolikheten att svara rätt 20%, och väntevärdet sjunker i motsvarande grad. Men påverkan på betygsgränserna blir förhållandevis begränsad när man ökar antalet svarsalternativ. I tabell 6 visas dom effektiva betygsgränserna för dikotom rättning och olika antal svarsalternativ. Men detta gäller enbart om svarsalternativen är lika attraktiva. Det grundläggande problemet kvarstår fortfarande men blir mindre allvarligt, dock inte i så stor omfattning som vore önskvärt.

Tabell 6 Effektiva betygsgränser vid dikotom rättning och olika antal svarsalternativ.

Betyg	4 svarsalternativ (%)	5 svarsalternativ (%)	6 svarsalternativ (%)
A	≥89	≥89	≥89
B	76-88	77-88	77-88
C	61-75	63-76	64-76
D	47-60	49-62	51-63
E	32-46	35-48	37-50

Standardtesten som använts innehåller 100 uppgifter, detta är dock inte möjligt att använda i realiteten, där antalet frågor av praktiska skäl är begränsat. Detta gör att sannolikheterna kommer att ändras på grund av de diskreta stegen i poängsättningen. Effekten av detta kommer dock vara begränsad och är relativt lätt att korrigera för i exempelvis ett spreadsheet program med statistik funktioner.

Det är möjligt, om än svårt, att på ett enkelt sätt visualisera effekterna som man får för olika strategier att besvara frågor, gissa eller lämna obesvarade, då det rör sig om ytterligare en dimension. Dock är detta inte en så viktig fråga och det är också möjligt att få fram informationen med ett spreadsheet program med statistik funktioner.

Svagheterna i metoden medför att analysen i sin helhet är att betrakta som konservativ och ger en underskattning av problematikens reella effekter.

Diskussion och slutsatser

Flervalsuppgifter har i sig fördelar, men samtidigt har de inbyggda svagheter när det gäller vad som testas. De fördelar som finns är många gånger knutna till ämnet, medan det är svårt att se hur flervalsuppgifter skall kunna visa på färdigheter som exempelvis problemlösning och med det problemlösningsteknik och räknefärdigheter. Av samma skäl är utredande uppgifter uteslutna. Om man använder sig av flerstegsfrågor blir även det komplicerat att få till med rena flervalsuppgifter. Detta medför att flervalsuppgifter kan förväntas att vara ”enklare” än utredande frågor eller frågor baserade på lösning av ett komplicerat problem.

Men även ett bra verktyg kan användas på ett sätt som ger ett olyckligt resultat, och det gäller främst problematiken med gissning i fallet med flervalsuppgifter. Som visats här så innebär möjligheten till gissning i kombination med en dikotom rättning att betygsgränserna effektivt sett sänks. Men denna sänkning gynnar de deltagare som har en större benägenhet till att ta risker och deltagare som har en lägre kunskapsnivå. Genom att man inte på något sätt förlorar på att gissa gör detta att sannolikheten för att få fler poäng ökar utan att det avspeglar färdigheter eller kunskaper. Med en godkänt-gräns på nominellt 41% kommer den effektiva godkänt-gränsen med dikotom rättning att vara cirka 32%. En sänkning av godkänt-gränsen på 9% enheter! Då analysen som gjorts baseras på fyra svarsalternativ som skall vara lika attraktiva, betyder detta att om svarsalternativ kan uteslutas, genom dålig formulering eller fel, så kommer sänkningen i realiteten att vara större. Detta innebär att om man administrerar en examen med enbart flervalsuppgifter så kommer andelen som inte

klarar examen automatiskt att minska, i större eller mindre grad, om man inte justerar betygsgränserna. Men en sådan justering riskerar att straffa de deltagare som inte är benägna att gissa. Så den grundläggande problematiken kvarstår. Det finns även en faktor när det gäller svårighetsgraden som gör att man kan få ett högre medel än förväntat på grund av flervalsuppgifts-formatet.

Ett alternativ till dikotom rättning är att använda sig av en gissningskorrigerande algoritm, där felaktiga svar ger avdrag. Här försvinner då belöningen i att gissa. Dock kvarstår problematiken om svarsalternativ kan uteslutas, med då detta i många fall är baserad på partiell kunskap, belönas denna indirekt. Det finns dock en psykologisk dimension, både hos testdeltagare som testkonstruktörer, där negativ poängsättning fungerar demotiverande. Ur rättvisesynpunkt är det fel att belöna en viss typ av beteende på bekostnad av kunskap och handlar i mycket om att uppmuntra lärande framför att gissa.

Det finns andra sätt att korrigera för gissning, bland annat en metod att ge poäng för obesvarade uppgifter, men detta medför dock att poänggränserna måste justeras i motsvarande grad. Även en kombination av poäng för obesvarade uppgifter och avdrag för fel: $S = R + \frac{O}{c} - \frac{W}{c-1}$ medför att betygsgränserna bör justeras. I hur stor grad måste beräknas i de enskilda fallen baserat på antal uppgifter och svarsalternativ.

För att minska sannolikheten för att gissning lönar sig, kommer också en ökning i antalet svarsalternativ att fungera om än i begränsad omfattning. Här krävs det dock att man har svarsalternativ som är lika attraktiva och inte innehåller uppenbara logiska brister eller fel. Detta innebär också mer arbete med att konstruera testerna, då det är speciellt viktigt att alla svarsalternativ är logiska och lika attraktiva.

Man kan i tillägg även beakta möjligheten med flera rätta svar eller ”Answer-Until-Correct” som i motsvarande grad minskar vinsten med gissning, samtidigt som man på olika sätt testar och belönar partiell kunskap. Genom bruk av digital examen är de tidigare hindren i fråga om kostnader i administrationen och rättning mindre, vilket gör denna typ av flervalsuppgifter intressanta för implementering.

Som alternativ kan man även använda sig av graderad rättning, där svarsalternativen simulerar ”free-response”(Lin & Singh 2012) som bestämts utifrån vanliga fel (ex. Räknefel, fel formel och så vidare). De olika svarsalternativen ger då olika poäng baserat på de misstag som ger de olika alternativen. Detta löser i sig inte problematiken med gissning men är ett alternativ värt att beakta.

Slutsatsen av analysen är att man måste vara medveten om och vara beredd att justera rättningsalgoritmen för de problem som flervalsuppgifter ger. Det är inte möjligt att direkt applicera rena flervalsexamina utan att först analysera följderna av ett utbrett gissande, då det i praktiken innebär en sänkning av betygsgränserna.

Sannolikheten för att en testdeltagare skall få ett betyg som inte svarar mot kunskapsnivån är ganska stor i fallet med dikotom rättning, i alla fall när det gäller dom lägre betygsgraderna. Då dikotom rättning effektivt sett innebär en sänkning av godkänt-gränsen är detta ett problem som bör bemötas på olika sätt. Det finns olika lösningar att använda men detta kräver en medvetenhet om problemet och dess lösningar med sina för- och nackdelar.

Svaret på frågan om risken att en testdeltagare skall få en konstlad hög nivå på grund av gissning är att sannolikheten inte är försumbar, speciellt för låga nivåer.

Litteratur

Abu-Sayf, F. K. (1979). The scoring of multiple choice tests: A closer look. Educational Technology, 19, 5–15.

Burton, R. F. (2001). Quantifying the effects of chance in multiple choice and true/false tests: question selection and guessing of answers. *Assessment & Evaluation in Higher Education*, *26*(1), 41-50.

Campbell, M. L. (2015). Multiple-Choice Exams and Guessing: Results from a One-Year Study of General Chemistry Tests Designed To Discourage Guessing. *Journal of Chemical Education*, *92*(7), 1194-1200.

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

Davis, F. B. (1964). Educational measurements and their interpretation. Belmont, Calif.: Wadsworth.

R.L. Ebel (1968) Blind guessing on objective achievement tests Journal of Educational Measurement, 5, pp. 321–325

Gilman, D.A. and Ferry, P. (1972). Increasing test reliability through self-scoring procedures. Journal of Educational Measurement, 9, 205-207

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

Hanna, G.S. (1975). Incremental reliability and validity of multiple choice tests with an answer-until-correct procedure. Journal of Educational Measurement, 12, 175-178

Lin, S. Y., & Singh, C. (2012). Can multiple-choice questions simulate free-response questions?. In *2011 PHYSICS EDUCATION RESEARCH CONFERENCE* (Vol. 1413, No. 1, pp. 47-50). AIP Publishing.

Sirnes, S.M. (2005). Flervalgsoppgaver – konstruktion og analyse. Fagboksforlaget, Bergen