Analys av rättningsalgoritmer för flervalsuppgifter som examinationsform.

Jonas R. Persson *Skolelaboratoriet, Institutt for lærerutdanning, Norges Tekniske og Naturvitenskaplige Universitet, NTNU, 7491 Trondheim*

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 görs mellan olika poängsättningsalgoritmer 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.

1 INTRODUKTION

Flervalsuppgifter har sedan de först utvecklades av Frederick Kelly i 1915 används bland annat intelligenstester och andra typer av tester med stora deltagarantal. Då antalet studenter ökar och med det arbetsbördan att rätta examina är det frestande att välja en examination enbart eller till stora delar baserad på flervalsuppgifter. Även ökade krav på digitalisering och objektiva bedömningar främjar flervalsuppgifter, genom bland annat möjligheten till automatisk rättning. Det finns således ett antal fördelar med bruken av flervalsuppgifter som exempelvis:

- Det tar kortare tid för studenterna att besvara flervalsuppgifter
- Större del av pensum kan täckas med fler frågor
- Insatsen med att skriva minskar
- Värderingen går fortare och är objektiv

Men det finns ett antal nackdelar, som måste beaktas.

- Frågan är vad man testar med givna svarsalternativ.
- Uppgifterna måste formuleras på ett genomtänkt och logiskt sätt.
- Det finns en risk för att studenterna gissar.

Här studerar vi speciellt fenomenet med gissning och vilka konsekvenser det får för resultaten på en test. Detta är en aspekt som är oberoende av kontext och ett inbyggt problem med flervalsfrågor. Effekten av gissning kommer att bero på hur uppgifterna poängsätts. Där olika rättningsalgoritmer kommer att få olika effekter av gissning. Har vi en viss typ av poängsättning som gynnar en strategi med gissningar och om deltagarna är medvetna om detta kommer detta påverka deras svarsstrategi.

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, bland annat för att det uppmuntrar gissning. I tillägg är det tveksamt om den kan ge ett direkt samband mellan resultatet och den kunskap testdeltagarna har. Informationen 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).

Genom att utveckla alternativa poängsättningsalgoritmer försöker man lösa många av problemen med dikotom poängsättning. Bland annat genom algoritmer som motverkar gissning och algoritmer för att på ett mer effektivt sätt kan 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.

Här diskuterar vi några poängsättningsalgoritmer och jämför olika algoritmer med ett standardtest. Jämförelsen görs utgående från en antagen kunskapsnivå och gissningar på övriga uppgifter. Effekten på betygen i en tänkt examenssituation erhålls genom att att undersöka hur stor sannolikhet det är för en kunskapsnivå att uppnå ett visst betyg. Den 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. En fullständig beskrivning finns i en kommande artikel [Persson 2017].

2 GISSNINGSKORRIGERAD POÄNGSÄTTNING

Gissning är ett problem vid flervalsfrågor och något som det är önskvärt att minimera. Det är dock inte möjligt att helt undvika gissning men effekterna kan minimeras. Men det handlar även om strategier och självförtroende, vill man inte gissa och riskera att avge ett felaktigt svar och avstår att svara straffas man i fallet med dikotom rättning, jämfört med om man chansar och gissar. Man kan korrigera för detta genom att betrakta sannolikheten för att avge rätt svar vid gissning och beräkna den totala poängsumman som

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

där R är antalet korrekta svar och W antalet felaktiga svar, och c antalet svarsalternativ på varje uppgift Denna kallas den 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. Information om algoritmen kan medföra att deltagarna undviker att gissa och hellre avstå från att svara, för att undvika avdrag.
Ett alternativ är att ta hänsyn till obesvarade frågor, O, genom att 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.
Dessa kan också kombineras:

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

Vilket förstärker ett beteende där gissningar inte lönar sig.

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. Svagheten är dock att vi inte vet den sanna nivån utan den måste uppskattas, vilket gör metoden osäker. Detta gör att IRT sällan används, då det inte är speciellt praktiskt att analysera svaren med IRT.

3 STANDARD TEST

För att studera effekterna används ett standardtest bestående av 100 uppgifter med 4 svarsalternativ. Sannolikheten och slumpmässighet i de uppgifter som inte motsvarar kunskapsnivån ger en binomialfördelning med 100 uppgifter och 4 alternativ. Sannolikheten för att uppnå ett visst antal rätta svar kan beräknas med olika spreadsheet program som exempelvis Excel, som har danna funktion inbyggd.
Till exempel kommer en kunskapsnivå på 40% motsvaras av 40 rätta svar på 40 uppgifter som adderas med en slumpmässig fördelning för resterande 60 uppgifter. Det gör det möjligt att beräkna sannolikheten för att en deltagare 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. 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.
Man kan då konstruera ett fullständigt set med olika kunskapsnivåer. Det är viktigt att poängtera att testen görs för en strategi där alla svar som inte kan besvaras utifrån kunskapsnivån besvaras slumpmässigt samt att alla uppgifter besvaras, vilket är ett extremfall. Då det i verkligheten är möjligt att utesluta svar så kommer detta vara det värsta scenariot.

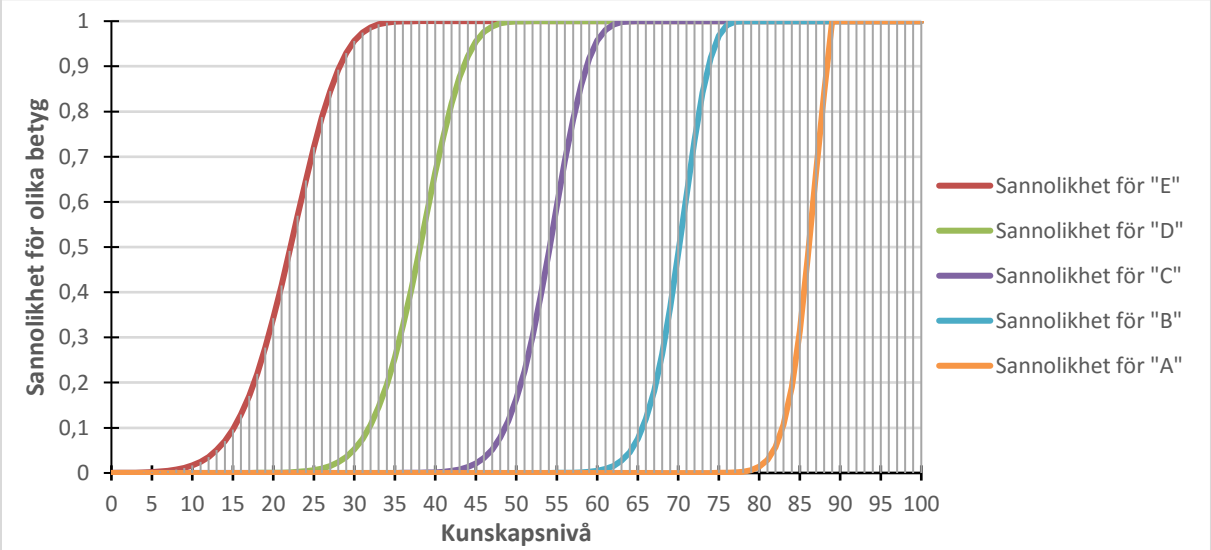
3.1 Standard test med dikotom poängsättning

Med utgångspunkt i betygsskala som rekommenderas på NTNU (Tabell 1) är det möjligt att beräkna sannolikheterna att få ett visst betyg för olika kunskapsnivåer.

TABELL 1 BETYGSGRÄNSER VID NTNU.

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

Beaktas sannolikheten att få godkänt (E) för en testdeltagare med en kunskapsnivå på 23% har denne 57% chans att få godkänt. Sannolikheterna för olika betyg mot kunskapsnivå ges i figur 1.



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. Vilket i realiteten är en sänkt gräns för godkänt. Tittar vi på sannolikheten att få ”D” 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 har gissning en mindre effekt vid hög kunskapsnivå. I praktiken innebär detta att betygsgränserna effektivt sett kommer att ändras och behöver korrigeras enligt tabell 2.

TABELL 2 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

3.2 Standard test med gissningskorrigerad poängsättning

Vid gissningskorrigerad poängsättning med

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

tillkommer en dimension, då det är möjligt att lämna uppgifter obesvarade. Vilket gör att det blir svårare att illustrera sannolikheterna för ett visst betyg (som blir två-dimensionell). Gissar man på alla uppgifter som inte kan besvaras utifrån kunskapsnivån blir väntevärdet för resultatet samma som antalet korrekta svar. Detta för att väntevärdet för rena gissningar ($\frac{W}{c-1}$) alltid kommer att vara noll med denna algoritm. Med andra ord kommer de testdeltagare som gissar inte att tjäna på det. Betygsgränserna blir med detta oförändrade.

I fallet med poäng 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:

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

Här är det tydligt att gissning inte lönar sig då man är garanterad en poäng vid en obesvarad uppgift, medan det finns en risk att man gissar fel. Här kommer den effektiva betygsgränsen ändras och man bör korrigera för detta. Vid 4 svarsalternativ medför detta 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 olika betygsgränserna utgående från kunskapsnivån ges i tabell 3. 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 3 OLIKA BETYGSGRÄNSER VID RÄTTNING ENLIGT ALGORITMEN S=R+O/C UTAN GISSNINGAR.

Betyg	Effektiva (%)	Korrigerade (%)	Ursprunglig (%)
A	≥86	≥91,75	≥89
B	70-85	82,75-91,50	77-88
C	54-69	73,75-82,50	65-76
D	38-53	64,75-73,50	53-64
E	22-37	55,75-64,50	41-52

4 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 bör 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.

5 DISKUSSION OCH SLUTSATSER

Flervalsuppgifter har fördelar, men samtidigt har de inbyggda svagheter. 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.

Problematiken med gissning i fallet med flervalsuppgifter är något som bör uppmärksammas. Som visats innebär gissning i kombination med dikotom rättning att betygsgränserna effektivt sett sänks. Sänkning gynnar deltagare som har en större benägenhet till att ta risker och deltagare med en lägre kunskapsnivå. Genom att man inte förlorar på att gissa ökar sannolikheten att få fler poäng utan att det avspeglar färdigheter eller kunskaper. Med en godkänt-gräns på nominellt 41% kommer den effektiva godkänt-gränsen vid dikotom rättning att vara cirka 32%. Då analysen som gjorts baseras på fyra svarsalternativ som skall vara lika attraktiva, betyder detta att dålig formulering eller fel med för en större sänkning. Detta innebär att en examen med enbart flervalsuppgifter kommer andelen som inte klarar examen automatiskt att minska om man inte justerar betygsgränserna.

Ett alternativ 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.

Att ge poäng för obesvarade uppgifter, medför också att poänggränserna måste justeras i motsvarande grad. 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 är det möjligt att öka antalet svarsalternativ, men effekten är relativt liten. Här krävs det dock att man har svarsalternativ som är lika attraktiva och inte innehåller uppenbara logiska brister eller fel.

Man kan också beakta möjligheten med flera rätta svar eller ”Answer-Until-Correct” (Hanna 1975) 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.

Ett alternativ speciellt för beräkningsbaserade flervalsfrågor är att man kan 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.

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. 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.

6 LITTERATUR

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

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.

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.

Persson, J.R. (2017) UNIPED accepterad.