

Hantering av bortfall i longitudinella studier

Ett exempel

SVEN-ERIC REUTERBERG

Institutionen för pedagogik och didaktik, Göteborgs universitet

Sammanfattning: *Bortfall är ett stort problem i praktisk taget all empirisk pedagogisk forskning. Vanligen hanteras bortfallet så att personer som saknar vissa uppgifter utesluts vid analyserna. Oftast leder detta till en snedvridning av resultaten och till att de statistiska analyserna förlorar i skärpa. I denna artikel diskuteras ett antal tekniker för hur bortfallet kan hanteras på ett bättre sätt än genom uteslutning. Vissa av dessa tekniker tillämpas på ett stort longitudinellt datamaterial och artikeln avslutas med en redogörelse för de konsekvenser bortfallshanteringen fått såväl för undersökningsgruppens storlek som för resultaten.*

Bortfall är ett vanligt problem vid pedagogisk forskning. Speciellt stort är problemet vid longitudinella undersökningar och i synnerhet i de fall där datainsamlingen sträcker sig över en lång tidsperiod eller där den omfattar ett stort antal variabler (Kim & Curry 1977, Laird 1988). De tre huvudproblem som bortfallet förorsakar är enligt Little och Schenker (1995) att:

- resultaten blir missvisande om bortfallet inte är slumpmässigt och såvida man inte vidtar några korrektioner för detta,
- vi förlorar information och de statistiska skattningarna blir osäkrare, samt
- de statistiska standardmetoderna förutsätter kompletta data och analyserna blir mer komplicerade om vissa data saknas.

Denna artikel handlar om den första punkten ovan, närmare preciserat om vilka åtgärder man kan vidta för att minska effekterna av ett inte slumpmässigt bortfall.

OLIKA TYPER AV BORTFALL

Eftersom bortfallets karaktär är avgörande för vilka åtgärder som är tillämpbara finns det skäl att först skilja mellan olika typer av bortfall. En grund för kategorisering av bortfall kan vara dess orsak och man skiljer då mellan strukturellt bortfall och tillfälligt bortfall. Med strukturellt bortfall menas att bortfallet är orsakat av undersökningsdesignen. Det kan gälla

undersökningar där man tillämpat så kallad matrissampling, vilket innebär att man i förväg bestämt att olika undergrupper lämnar delvis skilda uppgifter. Som exempel på sådana undersökningar kan nämnas IALS-undersökningen (Skolverket 1996), där man för att få en så bred täckning av variabler som möjligt lät olika undergrupper besvara enbart vissa uppgifter och där de utelämnade uppgifterna roterades mellan grupperna.

Ett annat exempel utgör de nya utprövningsrutinerna för Högskoleprovet där provtagarna, förutom det ordinarie provet också får besvara en utprövningsversion av något av delproven. Varje delprov har fem eller sex olika utprövningsversioner, vilka administreras i separata högskoleregioner (se Ögren 1998). Ytterligare ett exempel på strukturellt bortfall är då man i en longitudinell studie väljer att följa upp enbart en speciell del av totalgruppen. En sådan studie genomfördes av Hammarström (1996) då hon studerade motiv och hinder för val av högre utbildning bland den tredjedel av en ålderskohort som i trettonårsåldern presterat bäst på ett begåvningsprov. Uppgifter om motiv och hinder insamlades via enkät då deltagarna var 25 år gamla.

Den vanligaste typen av bortfall är dock det tillfälliga bortfallet där samtliga personer planeras att ha fullständiga uppgifter, men där vissa saknar uppgifter på en eller flera variabler beroende på att de ej kunnat nås, har vägrat att delta eller där det skett sådana kodningsfel att uppgiften måst utelämnas vid bearbetningarna.

En annan indelningsgrund för bortfall kan vara dess statistiska egenskaper. Little och Rubin (1987) skiljer här mellan tre olika typer av bortfall. Den minst allvarliga typen är den som de benämner »Missing Completely at Random» (MCAR). Med svensk terminologi skulle vi beteckna den som ett fullständigt slumpmässigt bortfall och den kännetecknas av att de individer som saknar en viss uppgift inte skiljer sig från dem som har uppgiften i fråga. Om detta är fallet kommer bortfallet inte att medföra några missvisande resultat, men däremot leder de minskade gruppstorlekarna till större medelfel vilket medför att de statistiska analyserna blir mer okänsliga (lägre »power») samt att vi inte kan utnyttja de statistiska standardmetoderna.

Något mer komplicerad är situationen om vi har ett bortfall som i Little och Rubins terminologi är »Missing at Random» (MAR). Denna typ av bortfall kännetecknas av att bortfallet inte är slumpmässigt totalt sett, men att det är MCAR inom olika undergrupper och att vi har tillgång till sådana uppgifter att totalgruppen kan delas in i dessa undergrupper. Bortfallet kan till exempel vara olika stort för pojkar och flickor men MCAR inom vardera gruppen. Har vi uppgift om kön för samtliga är således bortfallet MAR.

Om bortfallet inte uppfyller ens de mildare krav som MAR ställer har vi en bortfallstyp som betecknas som NonMAR (Little & Rubin 1987) eller »Non-ignorable» (Rubin 1987). Denna typ av bortfall kräver alltid att man vidtar speciella åtgärder för bortfallshantering och i dag finns ett antal sådana åtgärder tillgängliga. Tyvärr tillämpas dessa åtgärder alltför sällan utan vanligen betraktas bortfallet som om det vore MCAR eller att det avviker så lite från detta antagande att man kan bortse från konsekvenserna. Normalt är det emellertid så att om man har ett Nonignorable bortfall och man rätt tillämpar en metod för bortfallshantering uppnås säkrare resultat jämfört med om

man helt bortser från bortfallet (Gustafsson & Stahl 2000 s 87). Dock är ingen metod fullständigt tillförlitlig och man får därför inte låta sig invaggas i den falska tryggheten att metoden fullständigt korrigerar för den missvisning som bortfallet i sig ger upphov till. Anderson, Basilevsky och Hum (1983 s 417) säger att »the only true solution to the missing data problem is to not have any».

OLIKA METODER FÖR BORTFALLSHANTERING

De vanligaste metoderna för bortfallshantering kan delas in i följande fyra kategorier: (i) uteslutning av individer, (ii) imputation av enskilda värden, (iii) skattning av den kompletta kovariansmatrisen, samt (iv) modellering med inkompleta data. De flesta av dessa kan också delas in i underkategorier som framgår av Tabell 1 på nästa sida.

Den i särklass vanligaste metoden för bortfallshantering är uteslutning av individer och bland de två varianterna den som innebär att samtliga individer som saknar en viss uppgift utelämnas ur analyserna. En förutsättning för att denna metod inte skall ge missvisande resultat är att bortfallet är fullständigt slumpmässigt eller MCAR (Missing Completely at Random). Även om detta stränga krav är uppfyllt medför ändå metoden de två andra huvudproblemen som nämndes inledningsvis, nämligen att vi förlorar information och att vi inte kan nyttja de statistiska standardmetoderna.

Av dessa två är förlusten av information det allvarligaste. Särskilt stor blir denna då man väljer »listwise deletion» och då data omfattar många variabler (Kim & Curry 1977, Raymond & Roberts 1987). Detta leder i sin tur till att den information som faktiskt finns tillgänglig utnyttjas i ringa grad (Wotke 2000). En annan konsekvens är att analyserna kommer att baseras på en starkt reducerad sampelstorlek vilket medför stora medelfel för estimaten och därmed reduceras den statistiska skärpan (Bello 1993). För att metoden inte skall ge missvisande resultat bör man försäkra sig om att den återstående informationen faktiskt är representativ (Frane 1976).

Misshushållningen med data blir givetvis mindre om vi i stället för »listwise deletion» väljer »pairwise deletion», där man utesluter enbart de individer som saknar uppgift i någon av de variabler som ingår i en och samma analys. Dock drabbas vi då av det problemet att olika analyser i en studie baseras på delvis olika grupper. Om bortfallet inte helt uppfyller kraven för MCAR kan det till exempel få konsekvenser för strukturen i kovariansmatriser vilket bland annat kan leda till att dessa inte blir »positivt definita» (Kim & Curry 1977, Malhotra 1987). Enligt Roth (1994) har simuleringsstudier visat att uteslutning av data ofta är en otillförlitlig metod för bortfallshantering och i synnerhet i de fall där en stor del av undersökningsgruppen utesluts.

Imputation innebär att vi ersätter den saknade uppgiften med ett värde som bedöms vara det mest sannolika för individen ifråga. På detta sätt kan datamatriken göras komplett och all den information som finns tillgänglig kan därmed utnyttjas i analyserna. Som framgår av Tabell 1 nedan kan man komma fram till detta, det mest sannolika värdet, på olika vägar.

Tabell 1. Olika metoder för bortfallshantering.

<i>Metod och varianter</i>	<i>Åtgärd</i>
<i>Uteslutning av individer</i>	
»Listwise»:	Samtliga individer som saknar data utesluts oavsett analys
»Pairwise»:	Endast individer som saknar data för variabler ingående i en och samma analys utesluts
<i>Imputation</i>	
Medelvärden för samtliga med fullständiga data:	Medelvärdet för bortfallsvariabeln beräknas för dem som har uppgift om variabeln i fråga
Medelvärden för individer i samma subgrupp:	Medelvärdet för bortfallsvariabeln beräknas för individer som tillhör samma subgrupp som den individ för vilken imputationen ska göras
Medelvärden för individen själv:	Medelvärdet för individen beräknas för de variabler som har nära anknytning till bortfallsvariabeln
Hot deck:	Man identifierar samtliga individer som tillhör samma subgrupp som den för vilken imputationen ska göras och väljer slumpmässigt en individ med uppgift på variabeln i fråga vars värde imputeras.
Enkel regression:	Det mest troliga värdet skattas utifrån individens värden på sådana variabler som uppvisar ett högt samband med bortfallsvariabeln.
Multipel regression:	Samma som enkel regression, men man väljer ut flera stickprov bland dem som har uppgift och skattar flera olika imputationsvärden. Alternativt man utgår från samtliga med uppgift i bortfallsvariabeln och skattar ett värde till vilket adderas ett slumpmässigt värde.
<i>Skattning av komplett kovariansmatris</i>	
EM-algoritmen:	Man utgår från den grupp som har fullständiga värden och skattar den kompletta kovariansmatrisen iterativt
<i>Modellering av inkompleta data</i>	
Matrisbaserad modellering:	Genom strukturell ekvationsmodellering av medelvärden och kovariansmatriser för varje grupp med ett givet bortfallsmönster görs Maximum-Likelihood-skattningar av den kompletta kovariansmatrisen
Rådatabaserad modellering:	Med utgångspunkt i rådata och med hjälp av strukturell ekvationsmodellering görs Maximum-Likelihood-skattningar av den kompletta kovariansmatrisen.

Den enklaste metoden för imputation är medelvärdesimputation och den innebär att vi ersätter det saknade värdet med medelvärdet för de individer som har information på variabeln i fråga. Som visas i Tabell 1 finns tre olika varianter för att beräkna det medelvärde som skall utgöra ersättning för det saknade värdet. Den enklaste varianten är att beräkna medelvärdet för samtliga som har information på variabeln ifråga. Raaijmakers (1999) kallar denna metod »total mean substitution» (TMS). TMS förutsätter emellertid att det inte finns skillnader mellan olika undergrupper på variabeln i fråga, det vill säga att bortfallet är MCAR. Om det finns sådana skillnader kommer denna variant av medelvärdesimputation att leda till missvisande resultat såtillvida att skillnaderna mellan undergrupperna underskattas. För att undvika detta kan man som imputationsvärde välja det medelvärde som gäller för den undergrupp som individen tillhör. Eftersom vi i detta fall tar hänsyn till grupptillhörighet förutsätter denna metod att bortfallet är MAR (Missing at Random) det vill säga att det är slumpmässigt inom respektive undergrupp.

Ytterligare en variant av medelvärdesimputation kan utnyttjas i de fall då vi har kluster av variabler som är så högt inbördes korrelerade att de kan ingå i en och samma skala. Då kan man nämligen beräkna individens eget medelvärde över de variabler som ingår i klustret och för vilka individen har uppgift. Raaijmakers (1999) kallar metoden »Valid Mean Substitution» (VMS). Detta »personliga» medelvärde får då utgöra imputationsvärdet. Precisionen hos det imputerade värdet är naturligtvis i detta fall beroende av de inbördes sambanden mellan variablerna i klustret samt antalet variabler som individen har uppgift för. Ju högre samband och ju fler variabler som individen har uppgift på desto bättre blir givetvis precisionen hos de imputerade värdena.

Om data har sådana mätegenskaper att medelvärdet inte bör beräknas kan naturligtvis detta ersättas av medianvärdet, även om medianen är ett grövre mått än medelvärdet. Är data av nominalkaraktär kan varken median eller medelvärde beräknas. Då kan man tillämpa så kallade »Hot deck»-imputation, vilket innebär att man söker fram en individ som har så många egenskaper gemensamt som möjligt med den för vilken uppgift saknas. Det värde som gäller för denne »dubbelgångare» får ersätta det saknade värdet (Ford 1983, Rizvi 1983). Ett alternativ till medelvärdesimputation och Hot deck-imputation är regressionsimputation varvid man utifrån en annan eller flera andra variabler som korrelerar med bortfallsvariabeln skattar ett sannolikt värde som skall ersätta det saknade.

Imputationsmetoderna har den fördelen framför uteslutning av individer att de leder fram till en komplett datamatrix varigenom man kan utnyttja all den information som faktiskt finns insamlad. Dock har de vissa brister och den allvarligaste av dessa är att de leder till underskattningar av varianser och kovarianser (Raymond 1986; Little & Rubin 1987; Roth, Champion & Jones 1996). I och med denna underskattning av varianserna kommer också medelfelen att underskattas vilket ökar risken för Typ I-fel vid signifikansprövningar (Heitjan 1997, Wotke 2000). Allra mest påtaglig blir underskattningen givetvis om vi imputerar totalmedelvärdet, då samtliga personer med bortfall tilldelas ett och samma imputationsvärde. Dock medför även regressionsimputationerna en viss underskattning eftersom de individer som har identisk

kombination av värden på de oberoende variablerna i regressionskvationen tilldelas identiska värden i bortfallsvariabeln.

För att undvika en underskattning av varianser och kovarianser har Rubin (1987) föreslagit multipel regressionsimputation. Med detta menas att man beräknar de värden som skall imputeras på flera olika stickprov av personer och att man tar hänsyn till den osäkerhet som imputationsmetoden medför genom att väga in också den varians som gäller för de upprepade imputationerna. Ett alternativt sätt att ta hänsyn till denna osäkerhet är att man till varje imputerat värde adderar ett slumpmässigt valt värde med en på förhand bestämd variation och med ett medelvärde lika med noll. En sådan funktion finns t ex i SPSS:s statistikpaket.

De olika metoderna för bortfallshantering har jämförts i en rad studier ofta av Monte Carlo-typ. Så har ett flertal studier visat att fullständig uteslutning av individer med bortfall är en mindre lämplig metod än parvis uteslutning då det gäller att skatta korrelationer och regressionskoefficienter (Beale & Little 1975, Gleason & Staelin 1975, Kim & Curry 1977, Raymond & Roberts 1987). Vad beträffar medelvärdesimputation är bilden inte entydig. Enligt Kim och Curry (1977) är denna metod sämre än uteslutning då det gällde att skatta en korrelationsmatris, medan Chan och Dunn (1972) liksom Raymond och Roberts (1987) kom till den motsatta slutsatsen. En möjlig förklaring till att resultaten varierat på detta sätt kan vara att studierna skiljer sig åt vad gäller bortfallsfrekvens och bortfallsmekanism. De divergerande resultaten kan också ses som uttryck för att ingen enskild metod är tillämplig på alla imputationsproblem (Little 1988).

Flera studier har emellertid visat att regressionsimputation är en tillförlitligare metod än både uteslutning och medelvärdesimputation (Beale & Little 1975; Chan, Gilman & Dunn 1976; Little 1988). Dock bör observeras att regressionsimputationen förutsätter att bortfallet är åtminstone MAR och att man har tillgång till data som ger bra skattningar av de saknade uppgifterna (Bello 1993).

Samtliga de metoder som diskuterats så här långt innebär att man kompletterar en ofullständig datamatrix. De två övriga metoder som finns upptagna i Tabell 1 ovan, Expectation Maximization-algoritmerna (EM) (Dempster, Laird & Rubin 1977; Little & Rubin 1987) och modellering av inkompleta data innebär däremot att man skattar den kompletta kovariansmatrisen. EM-algoritmen innebär att man med utgångspunkt i de observerade data gör maximum-likelihood-skattningar av de data som saknas. Därefter sker parameterskattningar utifrån faktiska data och de skattade bortfallsvärdena och proceduren upprepas till dess att man når konvergens. Undersökningar baserade på Monte Carlo-teknik tyder på att EM ger säkrare skattningar än såväl parvis som fullständig uteslutning av individer (Malhotra 1987, Graham & Donaldson 1993). Dock är ett problem med metoden att den kan ta lång tid för att konvergera (Little & Rubin 1987, Laird 1988).

Modellering av inkompleta data innebär att skattningsalgoritmerna vid strukturell ekvationsmodellering anpassas för att hantera bortfallet. Procedurer för sådan anpassning finns tillgängliga i till exempel Amos (Arbuckle 1997), Mplus (Muthén & Muthén 1998) och i LISREL (Jöreskog & Sörbom

1993) med tekniker utvecklade av Allison (1987) och Muthén, Kaplan och Hollis (1987). Med dessa metoder korrigeras för den snedvridning som bortfallet förorsakar i skattningar av medelvärdesdifferenser och av alla de parametrar som kan skattas vid strukturell ekvationsmodellering.

Som framgått av Tabell 1 (ovan) finns två olika varianter, den matrisbaserade modelleringen i LISREL och den rådatabaserade modelleringen i Amos. Vid den matrisbaserade modelleringen skapas separata matriser för varje grupp med en specifik bortfalls kombination och genom att lägga på likhetsrestriktioner över de olika grupperna på skattningarna får man parameter-skattningar vilka är giltiga för det kompletta datamaterialet. Som namnet anger utgår den rådatabaserade modelleringen från att matrisen skapas utifrån rådata, där de saknade värdena ersatts av en konstant. Vid skapandet av matrisen tas automatiskt hänsyn till bortfallets struktur och parameterskattningarna blir därigenom giltiga för den kompletta undersökningsgruppen. En förutsättning för modellering av inkompleta data är att bortfallet är MAR, men även med ett bortfall som är Nonrandom ger de modellbaserade metoderna mer rättvisande resultat jämfört med om man inte tar hänsyn till bortfallet (Gustafsson & Stahl 2000, Wotke 2000).

Även om modellering av inkompleta data inneburit ett stort framsteg vad gäller att hantera bortfall har metoden vissa nackdelar. Såväl den rådatabaserade skattningen som sker i Amos och den matrisbaserade skattningen i LISREL är avsevärt mer beräkningsintensiva än de traditionella metoderna för att hantera bortfall. När det gäller stora stickprov gäller detta i synnerhet för den rådatabaserade skattningen och när vi har många bortfalls kombinationer blir den matrisbaserade proceduren omständlig eftersom beräkningarna sker utifrån separata matriser. Gustafsson och Stahl (2000) föreslår att man i sådana fall reducerar bortfalls kombinationerna med hjälp av till exempel imputation. Ett annat problem är att modellspecifikationen är omfattande och komplex, vilket har hindrat en mera generell tillämpning av metoden. Genom utvecklingen av STREAMS (Gustafsson & Stahl 2000) har emellertid modellspecifikationen förenklats avsevärt.

DATABASEN UGU-R

Vid Institutionen för pedagogik och didaktik vid Göteborgs universitet har man byggt upp en mycket stor databas kallad UGU-R. Avsikten med denna databas är att den skall kunna ligga till grund för bl a longitudinella studier av utbildningsval och studiekarriär för olika grupper. Databasen omfattar samtliga födda 1972–79, totalt 842 800 personer för vilka uppgifter sammanställts från följande register:

- FOB-registret med uppgifter avseende bl a social och etnisk bakgrund
- Årskurs 9-registret med kursval och betyg för samtliga ämnen
- Gymnasieregistret med uppgifter om sökt och genomgången utbildning samt alla avgångsbetyg
- Högskoleregistret med uppgifter om sökalternativ samt genomgången utbildning inklusive poängproduktion

- Högskoleprovet med uppgiftsdata från samtliga provtillfällen från våren 1991 till hösten 1996
- Militära inskrivningsuppgifter med data för samtliga begåvningsprov
- Sysselsättningsregistret med uppgift om sysselsättning under åren 1988–96

Genom att utnyttja dessa registerdata skapar vi således en longitudinell databas med vars hjälp vi kan beskriva personernas utbildningskarriär. Av central betydelse är då avgångsbetygen från Årskurs 9 eftersom de utgör hela registrets utgångspunkt och eftersom detta är en variabel som finns för nästan samtliga ingående i registret. Grundskolebetygen är också den variabel som är mest jämförbar mellan olika individer och därför kan den fungera som en kontrollvariabel för gruppjämförelser vad gäller senare utbildningskarriärer.

Tyvärr har betygen vissa brister vilka måste åtgärdas så långt möjligt för att de skall fungera väl som kontrollvariabel. För det första finns det två olika strukturella bortfall. Ett sådant är alternativkursbetygen i engelska respektive matematik där ett betyg avser allmän kurs och ett avser den mer krävande särskilda kursen. Varje person har följaktligen enbart ett av dessa två betygsalternativ och de är inte direkt jämförbara med varandra eftersom den grupp som valt en viss alternativkurs utgör en egen referensgrupp. Ett annat exempel är de så kallade blockbetygen. Somliga elever har nämligen blockbetyg i de naturvetenskapliga respektive de samhällsvetenskapliga ämnena, medan andra har separata betyg i varje ämne som ingår i respektive block.

Slutligen förekommer tillfälliga bortfall därigenom att enstaka personer saknar ett eller flera enskilda betyg. Detta bortfall är visserligen litet, omkring 5 procent i varje ämne, men med tanke på det stora antalet ämnen som ingår medför bortfallet att andelen personer med fullständiga betyg blir starkt reducerad.

BETYGEN I URSPRUNGSDATA

Den ursprungliga datauppsättningen omfattar totalt 27 olika betygsuppgifter och detta antal överstiger klart det som gäller för de skolämnen som eleverna har i Årskurs 9. Det finns flera anledningar till att så är fallet. En sådan är att såväl engelska som matematik vid den tid som gäller för våra betygsdata var uppdelade i två alternativkurser. Därtill kommer att en mindre del av eleverna följt en icke uppdelad kurs, en så kallad gemensam kurs. Vid betygssättningen utgjorde varje alternativkursgrupp sin egen referensgrupp vilket innebär att ett visst betyg från en av alternativkurserna inte är jämförbart med samma betyg från någon av de andra.

Ytterligare ett problem är de så kallade blockbetygen. I somliga skolor har ämnena geografi, historia, religion och samhällskunskap integrerats till ett enda ämne kallat samhällsorienterande ämnen (SO) och i dessa fall har ett enda betyg givits, medan övriga elever fått separata betyg i varje ämne. På motsvarande sätt har biologi, fysik, kemi och teknik slagits samman till ett ämne kallat naturorienterande ämnen (NO). Det faktum att vi har tre betygsvarianter i engelska respektive matematik och två betygsvarianter för de samhällsvetenskapliga respektive naturvetenskapliga ämnen gör att data

innefattar inte mindre än 36 olika betygskombinationer, vilket naturligtvis gör grundskolebetygen omöjliga som kontrollvariabel.

Slutligen skall sägas att det föreligger ett visst bortfall vad gäller de övriga betygen. I första hand gäller detta betygen svenska som andra språk och hemspråk. Som framgår av Tabell 2 (nedan) har enbart 3 procent av totalgruppen det förstnämnda betyget och cirka 4 procent det sistnämnda.

Tabell 2. Andel elever med betyg i ursprungsdata. I procent av samtliga.

<i>Ämne</i>	<i>Andel med uppgift</i>
Barnkunskap	95,3
Bild	95,8
Hemkunskap	95,7
Hemspråk	4,2
Idrott	95,1
Musik	95,5
Slöjd	95,9
Svenska	94,2
Svenska 2	2,9
NO-blockbetyg	7,7
Biologi	87,7
Fysik	87,5
Kemi	87,5
Teknik	87,5
SO-block	11,7
Historia	83,9
Geografi	83,9
Religion	83,8
Samhällskunskap	84,0
Engelska, allmän kurs	27,9
Engelska, särskild kurs	65,5
Engelska, gemensam kurs	2,0
Matematik, allmän kurs	39,8
Matematik, särskild kurs	54,2
Matematik, gemensam kurs	2,1

På grund av dessa betygs speciella karaktär har inga bortfallsåtgärder vidtagits för dem. Det samma gäller också betygen i tillvalsämnen och anledningen är främst den att vi inte vet vilka ämnen som döljer sig bakom dessa

etiketter. Som framgår av Tabell 2 har även de övriga betygen drabbats av ett visst bortfall även om det är mycket måttligt för varje enskilt betyg, omkring 5 procent.

Eftersom eleverna enbart kan ha ett betyg i vardera engelska och matematik och då man antingen har blockbetyg eller ämnesbetyg i SO- respektive NO-ämnena kan det inte finns någon person med en komplett uppsättning av betyg. Det största bortfallet orsakar här uppdelningen på alternativkurser i engelska respektive matematik eftersom det vanligaste betyget i engelska (det i särskild kurs) gäller för enbart 65 procent och det vanligaste matematikbetyget (också det i särskild kurs) gäller för 54 procent. Blockbetyget får en något mindre effekt eftersom 84 respektive 87 procent av samtliga haft betyg i de enskilda ämnena. Även i detta fall är emellertid bortfallet oacceptabelt med tanke på betygens tilltänkta funktion i databasen.

IMPUTATION AV BASDATA

Syftet med denna artikel är att beskriva hur jag med olika tekniker för bortfallshantering kompletterat avgångsbetygen från grundskolan så att de kan fungera som en kontrollvariabel för olika studier baserade på databasen UGU-R. Som påpekas av Raaijmakers (1999) är i ett fall som detta imputation av basdata en lämplig metod för att hantera bortfallet eftersom det är opraktiskt att separata bortfallsåtgärder vidtas i varje enskild undersökning. Bland annat kan ett sådant förfarande medföra att resultatbilden blir olika från en undersökning till en annan. Det första steget blir därför att så långt möjligt reducera det bortfall som orsakas av alternativkursbetygen. Därefter kommer jag att eliminera det problem som orsakas av blockbetygen och slutligen skall bortfallet reduceras i de enskilda ämnen som har störst bortfall.

ALTERNATIVKURSBETYGEN I ENGELSKA OCH MATEMATIK

Huvudproblemet med alternativkursbetygen i engelska och matematik är att betygsskalorna inte är jämförbara. Uppgiften blir därför att finna en justeringsfaktor med vars hjälp de tre betygsskalorna kan transformeras till en och samma skala. Som framgår av Tabell 2 leder en sådan åtgärd till att vi får betygsuppgift i engelska för 95.4 procent av samtliga och betygsuppgift i matematik för 96.1 procent.

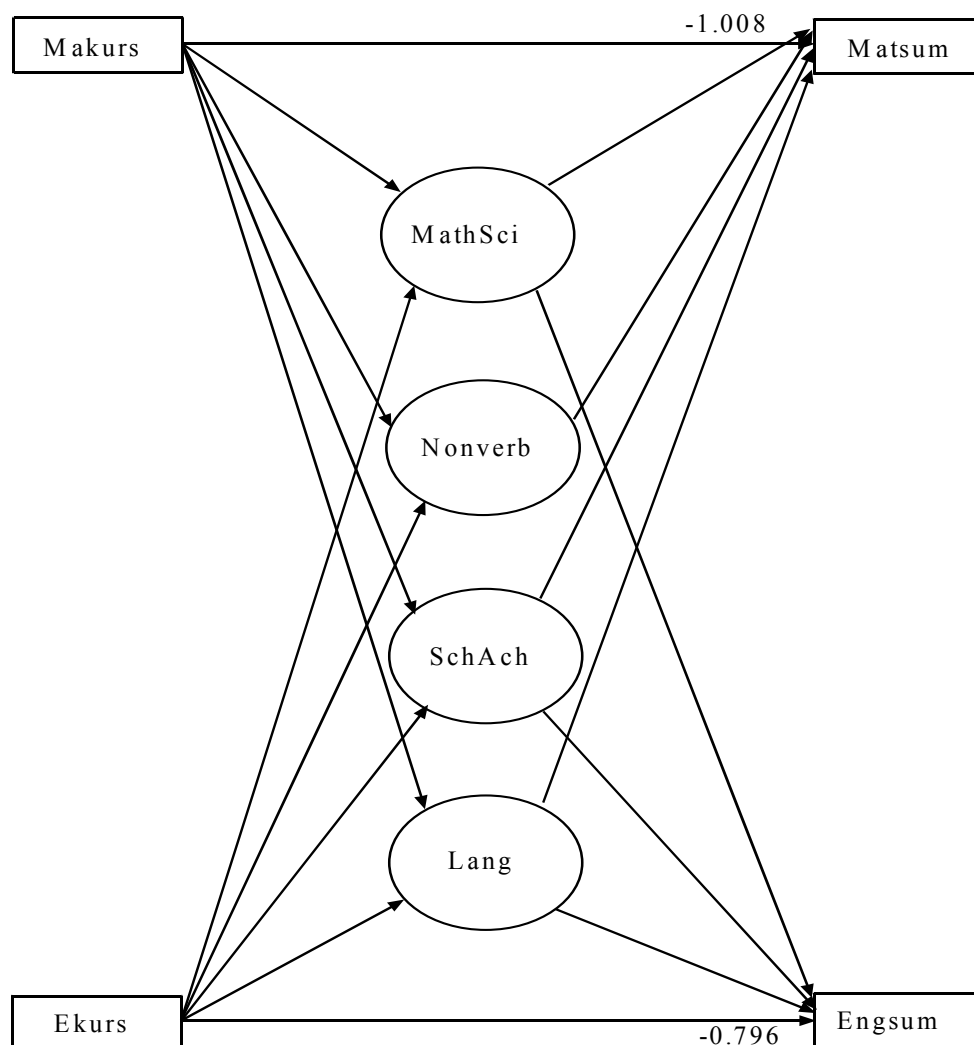
För att skatta denna justeringsfaktor har jag som en första åtgärd skapat en variabel av de tre alternativkursbetygen i vardera ämnet. Dessa två nya variabler kallas *Engsum* respektive *Matsum*. Vid skattningen av justeringsfaktorn har jag tagit dessa två variabler tillsammans med övriga betygsvariabler och satt in dem i en teoretisk modell över betygens dimensionalitet hämtad från Andersson (1998). Enligt denna modell utgör de enskilda betygen indikatorer på fyra olika faktorer eller latent variabler. Först har vi en generell skolprestationsfaktor, kallad »School achivement» (SchAch) i vilken samtliga

Tabell 3. Betygens standardiserade faktorladdningar i olika skolprestationsfaktorer.

<i>Betyg</i>	<i>SchAch</i>	<i>Nonverb</i>	<i>Mathsci</i>	<i>Lang</i>	<i>Andel förklarad varians</i>
Barnkunskap	0.761	0.210	---	---	0.623
Bild	0.495	0.321	---	0.201	0.388
Hemkunskap	0.715	0.304	---	---	0.604
Idrott	0.393	0.231	---	---	0.208
Musik	0.626	0.189	---	0.188	0.464
Biologi	0.846	0.147	0.209	---	0.782
Fysik	0.810	0.120	0.383	---	0.817
Kemi	0.833	0.120	0.322	---	0.811
Teknik	0.582	0.249	0.300	---	0.491
Geografi	0.879	0.035	---	---	0.774
Historia	0.893	---	---	---	0.797
Religionskunskap	0.894	---	---	---	0.799
Samhällskunskap	0.894	---	---	---	0.800
Slöjd	0.469	0.470	---	---	0.441
Svenska	0.834	---	---	0.307	0.790
Engsum	0.586	---	0.090	0.301	0.442
Matsum	0.545	0.097	0.265	0.102	0.387

betyg laddar. Av de kovarianser som återstod efter detta skapades tre olika betygsfaktorer, en bred icke-verbal faktor (Nonverb) med sina starkaste relationer till slöjd, bild och hemkunskap, en matematisk-naturvetenskaplig faktor (Mathsci) med relationer till matematik, de naturvetenskapliga ämnena och engelska samt en språklig faktor (Lang) med de starkaste relationerna till svenska och engelska, men också med relationer till bild och musik. Betygens standardiserade faktorladdningar i de olika faktorerna framgår av Tabell 3 (ovan).

Med de fyra betygsfaktorerna i Tabell 3 kan skolprestationer i olika ämnen tämligen väl prediceras och detta är utgångspunkten för beräkningen av justeringsfaktorerna för betygen i engelska och matematik. Beräkningen sker med hjälp av regressionsmodell där de fyra betygsfaktorerna tillsammans med kursvalen i engelska och matematik (Ekurs respektive Makurs) utgör oberoende variabler och *Engsum* respektive *Matsum* utgör beroende variabler så som framgår av Figur 1 (nedan).



Figur 1. Skattning av justeringsfaktorer för kursval i engelska respektive matematik.

Ekurs och *Matkurs* utgörs av dikotoma variabler där elever som valt särskild kurs kodats som 1 och övriga som 0. Den relation som går från var och en av dessa variabler till respektive betyg anger vilken unik effekt kursvalet haft för betyget då skolprestationerna hålls under kontroll. Denna ostandardiserade regressionskoefficient utgör således justeringsfaktorn. Som grund för beräkningarna ligger totalgruppen, men eftersom somliga saknar vissa betyg kommer beräkningarna att baseras på 659 133 personer eller 78 procent av samtliga. Huvuddelen av bortfallet beror här på att elever med blockbetyg i SO och/eller NO måste uteslutas. Modellanpassningen är tämligen god vilket framgår av att det vanligaste anpassningsmättet, RMSEA, uppgår till 0.048.

Som framgår av Figur 1 uppgår den ostandardiserade regressionskoefficienten mellan *Ekurs* och *Engsum* till -0.796 . Detta betyder att de elever som valt särskild kurs i engelska har fått ett betyg som ligger nästan 0.8 betygssteg lägre än vad man skulle ha väntat sig utifrån deras övriga skolprestationer. Motsvarande koefficient för *Matsum* är -1.008 , vilket bety-

der att elever på särskild kurs i matematik har fått ett betyg som ligger ungefär ett betygssteg för lågt. För att utröna om kursvalet haft olika effekter på pojkarnas respektive flickornas betyg upprepades analyserna för vardera könet, men skillnaderna var obetydliga, varför vi kan använda samma justeringsfaktor för båda könen.

De koefficienter som analyserna resulterade i är de exakta justeringsvärdena, men om vi skulle använda dem skulle vi få en betygsskala som utgjordes av heltal för somliga elever och decimaler för somliga. För att undvika detta har justeringsfaktorn avrundats till närmaste heltal. Detta innebär att ett betygssteg har lagts till det ursprungliga betyget för dem som valt särskild kurs. För denna grupp kommer således betygsskalan i engelska och matematik att gå från 2 till 6, medan den varierar mellan 1 och 5 för de övriga. De nya betygsskalorna har betecknats *Enggem* respektive *Matgem* och dessa har följaktligen ett variationsområde mellan 1 och 6. Härigenom kommer dessa båda betyg att ha ett något högre medelvärde och en något större spridning än övriga betyg.

BLOCKBETYGEN I NO OCH SO

Som framgår av Tabell 1 (s 169) är andelen elever som haft blockbetyg i NO 7.7 procent och de med blockbetyg i SO utgör 11.7 procent av samtliga. Eftersom de två grupperna inte sammanfaller fullständigt skulle detta ge upphov till ett betydande bortfall vad gäller de enskilda ämnena inom respektive block. För att undvika detta har enskilda ämnesbetyg imputerats för dem med blockbetyg. Ett enkelt sätt att göra detta hade naturligtvis varit att enbart ersätta de enskilda betyg som saknas med blockbetyget, men ett sådant förfarande hade medfört stora fel.

För det första hade det lett till att kovarianserna mellan de olika betygen inom blocket blivit grovt överskattade, eftersom samma värde då imputeras i samtliga ämnen för varje individ. För det andra hade vi underskattat varianserna för de enskilda ämnena eftersom samtliga individer med ett visst blockbetyg hade fått samma betyg för det enskilda ämnet. Eftersom vi inte har ett perfekt samband mellan blockbetyg och de enskilda ämnesbetygen kommer olika individer med ett visst blockbetyg att få något varierande ämnesbetyg. En underskattning av varianserna medför i sin tur att de enskilda betygens medelfel kommer att underskattas. Slutligen hade könsskillnaderna snedvridits. Dessa skillnader varierar nämligen mellan olika ämnen inom ett block och i vissa fall avviker könsskillnader i ämnet avsevärt från det som gäller för blockbetyget.

För att undvika dessa fel har jag tagit blockbetyget som utgångspunkt för imputationen. Till detta betyg har adderats en normalfördelad slumpfaktor för att undvika en underskattning av variansen hos ämnesbetyget och en överskattning av kovarianserna mellan de olika ämnena inom ett block. Denna slumpfaktor har genomgående ett medelvärde som är 0 och en standardavvikelse, vilken bestäms av det enskilda betygets felvarians. Därefter har könsskillnaden justerats så att den blir lika stor för de imputerade värdena som den är bland dem som har betygsuppgift i ämnet. När det gäller

Tabell 4. Algoritmer för imputation av enskilda ämnesbetyg utifrån blockbetyget.

<i>Ämne</i>	<i>Kön</i>	<i>Beräkning</i>
<i>Biologi</i>	Män	NO-block + normal(0.438) – 0.078
	Kvinnor	NO-block + normal(0.432) + 0.079
<i>Fysik</i>	Män	NO-block + normal(0.428) + 0.069
	Kvinnor	NO-block + normal(0.415) – 0.069
<i>Kemi</i>	Män	NO-block + normal(0.424) + 0.002
	Kvinnor	NO-block + normal(0.415) – 0.003
<i>Teknik</i>	Män	NO-block + normal(0.552) + 0.130
	Kvinnor	NO-block + normal(0.505) – 0.131
<i>Historia</i>	Män	SO-block + normal(0.434) + 0.057
	Kvinnor	SO-block + normal(0.417) – 0.057
<i>Geografi</i>	Män	SO-block + normal(0.438) + 0.035
	Kvinnor	SO-block + normal(0.415) – 0.035
<i>Religion</i>	Män	SO-block + normal(0.423) – 0.053
	Kvinnor	SO-block + normal(0.422) + 0.054
<i>Samhällskunskap</i>	Män	SO-block + normal(0.422) + 0.039
	Kvinnor	SO-block + normal(0.415) + 0.039

justeringen för könsskillnaderna har halva justeringsfaktorn adderats för det ena könet och halva subtraherats för det andra. Härigenom bibehålls totalmedelvärdet oförändrat.

För somliga individer medför dessa beräkningar att betygsvärdet inte utgörs av ett heltal. I sådana fall har betyget avrundats. Vidare kan justeringen medföra att vi får betygsvärden som ligger utanför det normala variationsområdet, 1 – 5. I dessa fall har alla värden under 1 kodats om till 1 och alla värden över 5 har kodats till 5. De värden som använts för slumpfaktorn och för justeringen av könsskillnaderna framgår av ovanstående tablå. Normal (0.438) innebär att slumpfaktorn har medelvärde lika med 0 och standardavvikelsen 0.438.

IMPUTATION AV BETYG I SVENSKA, IDROTT OCH TEKNIK

Den metod som använts för att imputera betyg i ämnena svenska, idrott och teknik är multipel regression, där de 8 à 9 betyg som visade högst prediktionsförmåga för respektive ämnesbetyg utgjorde oberoende variabler. Vilka dessa

ämnen är framgår av nedanstående Tabell 5. Efter att ha justerat för kursval i matematik och engelska samt imputerat värden för dem som saknat enskilda betyg i SO- och NO-ämnen har 774 649 fullständiga betyg i alla ämnen utom svenska som andra språk, hemspråk och tillvalsämnena. Detta motsvarar 91.9 procent av samtliga. 3.4 procent saknar nu samtliga betyg, vilket innebär att det saknas grund för att regressionsimputera betygsuppgifter. Återstår således 4.7 procent av totalgruppen med ofullständiga betyg. Av dessa saknar 2.4 procent ett enda betyg och ytterligare 0.5 procent saknar betyg i två ämnen. De ämnen som nu har lägst andel med betygsuppgift är svenska, idrott och teknik i nu nämnd ordning. För att reducera betygsbortfallet ytterligare regressionsimputerades värden för dessa tre ämnen.

Tabell 5. Prediktorer för regressionsimputation av betyg i svenska, idrott och teknik.

	<i>Svenska</i>	<i>Idrott</i>	<i>Teknik</i>
Prediktionsbetyg	Enggem	Slöjd	Fysik
	Religion	Teknik	Kemi
	Barnkunskap	Geografi	Slöjd
	Samhällskunskap	Musik	Matgem
	Hemkunskap	Historia	Biologi
	Historia	Fysik	Idrott
	Musik	Hemkunskap	Samhällskunskap
	Biologi	Svenska	Barnkunskap
	Bild		Svenska
Multipl korrelation	0.843	0.453	0.687

Som de multipla korrelationerna visar görs den säkraste prediktionen för betyget i svenska och den minst säkra för betyget i idrott. I det förstnämnda fallet förklarar prediktionsbetygen tillsammans något mer än 70 procent av variansen, medan motsvarande andel för idrott uppgår till något mer än 20 procent. Den låga andelen förklarad varians för idrottsbetyget är helt i linje med vad som visats i Tabell 3, nämligen att just idrottsbetyget hade den lägsta andelen förklarad varians av faktormodellen för samtliga betyg.

Liksom vid imputationen av betyg i blockämnena måste hänsyn tas till att de imputerade värdena sänker variansen. Detta har jag gjort på samma sätt som tidigare, nämligen genom att införa en slumpfaktor som bestäms av det imputerade betygets felvariens. De standardavvikelser som på detta sätt bestämts är 0.427 för svenska, 0.764 för idrott samt 0.529 för teknik. Därefter har betygsvärdena avrundats till heltal. Även här medför slumpfaktorn och avrundningen att somliga imputerade värden hamnar utanför betygsskalan. Liksom tidigare har värden lägre än 1 kodats som 1 och värden över 5 har kodats som 5.

Efter det att jag imputerat betyg i svenska, idrott och teknik har antalet personer med fullständiga betyg ökat till 789 724 vilket motsvarar 93.7 procent av samtliga. Samtidigt har den andel som saknar ett enskilda betyg reducerats till 0.7 procent, vilket innebär att det skulle krävas ett tämligen omfattande arbete att med hjälp av multipel regressionsimputation ytterligare minska bortfallet.

VAD HAR BORTFALLSHANTERINGEN BETYTT FÖR BETYGSUPPGIFTERNA I UGU-R?

Ett första och självklart svar på den frågan är att uppsättningen av betygsdata blivit avsevärt enklare att hantera. I utgångsläget fanns totalt 23 olika betygsuppgifter om man bortser från de mycket lågfrekventa ämnena hemspråk, svenska 2 samt de båda ospecificerade tillvalsämnena. Bland dessa fanns tre alternativa betyg i vardera engelska och matematik samt två alternativa betyg för de ämnen som ingår i SO-blocket respektive NO-blocket. Med en datauppsättning innefattande alla dessa 23 ämnen fanns självfallet ingen person med fullständiga betygsuppgifter.

Genom att transformera de tre olika betyg som fanns i engelska och matematik till ett enda för vardera ämnet minskade antalet betygsuppgifter för varje person till 19. Dock var bortfallet fortfarande stort eftersom varje person hade antingen ett blockbetyg eller fyra olika ämnesbetyg inom varje block. Genom att ersätta de båda sammanfattande blockbetygen med enskilda betyg i varje ämne kom antalet betygsuppgifter att reduceras till 17 och därmed hade andelen med fullständiga betygsuppgifter ökat från 0 procent till nästan 92 procent.

Denna andel höjdes till 93.7 procent genom att imputera betygsvärden i vart och ett av de ämnen som hade störst bortfallsandel, nämligen svenska, idrott och teknik. Därmed återstår 6.3 procent av totalgruppen som saknar betyg i ett eller flera ämnen. För 3.4 procent gäller att de saknar samtliga betyg och för dem återstår ingen annan metod för bortfallshantering än medelvärdesimputation. Jag har emellertid avstått från att göra en sådan imputation. För de återstående 2.9 procenten gäller att de saknar en eller flera betygsuppgifter. I princip kan naturligtvis denna grupp reduceras genom att göra multipla regressionsimputationer, men det skulle kräva ett mycket omfattande arbete, varför jag avstått från det. Ett alternativt förfarande skulle här kunna vara att komplettera även dessa saknade uppgifter med medelvärdesimputation. Om detta leder till en sådan ökning av precisionen att det kompenserar för det arbete som blir följderna är emellertid en bedömning jag överlåter åt kommande användare av databasen att göra.

Det finns också skäl att närmare granska vilka effekter vi fått av bortfallshanteringen för medelvärden, standardavvikelser och kovarianser. I Tabell 6 (nedan) redovisar vi effekterna på medelvärdena och standardavvikelseerna för de betygsuppgifter som varit föremål åtgärder. Som framgår av Tabell 6 har bortfallshanteringen inte inneburit några nämnvärda förändringar av de enskilda betygs medelvärde och standardavvikelse utom i två fall, nämligen då det gäller engelska och matematik. I båda dessa fall har medel-

värdet och standardavvikelsen höjts, vilket är den naturliga konsekvensen av att betygsskalan utökats med ett betygssteg för dem som följt den särskilda kursen och att vi därigenom har en skala för dem som går från 1 till 6. Däremot har andelen med betygsuppgifter höjts väsentligt för samtliga övriga ämnen i tabellen. Allra störst höjning uppvisar de samhällsvetenskapliga ämnena där uppgiftsfrekvensen ökat från omkring 84 procent till mer än 95 procent. Antalsmässigt motsvarar denna höjning mer än 90 000 personer.

Tabell 6. Medelvärden, standardavvikelser samt uppgiftsfrekvens för de ämnesbetyg som varit föremål för bortfallshantering.

	<i>Efter summering av alternativkursbetyg</i>			<i>Efter justering av alternativkursbetyg och imputation</i>		
	M	S	<i>Andel med uppg. i %</i>	M	S	<i>Andel med uppg. i %</i>
Engsum	3.22	0.89	95,4	---	---	---
Enggem	---	---	---	3.90	1.06	95,4
Matsum	3.16	0.93	96,1	---	---	---
Matgem	---	---	---	3.72	1.10	96,1
Biologi	3.19	0.96	87,7	3.18	0.97	95,4
Fysik	3.15	0.99	87,5	3.15	0.99	95,3
Kemi	3.11	1.00	87,5	3.11	1.00	95,2
Teknik	3.17	0.78	87,5	3.17	0.79	95,5
Geografi	3.20	0.93	83,9	3.20	0.95	95,6
Historia	3.19	0.98	83,9	3.19	0.99	95,6
Religion	3.20	0.96	83,8	3.20	0.98	95,6
Samhällsk	3.19	0.96	84,0	3.19	0.97	95,6
Svenska	3.18	0.89	94,2	3.18	0.89	95,5
Idrott	3.31	0.91	95,1	3.31	0.91	95,6

I Tabell 7 nedan redovisas betygens faktorstruktur efter justering för alternativkursbetygen i engelska och matematik, och att blockbetygen ändrats till betyg i enskilda ämnen och betygsuppgifter imputerats för betygen i svenska, idrott och teknik. Genom att jämföra faktorladdningarna i Tabell 7 med dem i Tabell 3 får man veta vilka konsekvenserna blivit av bortfallshanteringen.

Som framgår av jämförelsen mellan Tabell 3 och Tabell 7 har bortfallshanteringen inte heller fått några nämnvärda konsekvenser för betygens faktorladdningar utom då det gäller betygen i engelska och matematik. Genom att korrigera dessa båda betyg för kursvalet har deras laddningar i den generella faktorn, SchAch höjts för engelska från 0.59 till 0.73 och för matematik från 0.55 till 0.73. För betyget i engelska gäller vidare att dess laddningar i Mathsci och i Lang blivit något starkare efter korrigeringen. På samma sätt har mate-

Tabell 7. Betygens standardiserade faktorladdningar i olika skolprestationsfaktorer efter bortfallshantering.

<i>Betyg</i>	<i>SchAch</i>	<i>Nonverb</i>	<i>Mathsci</i>	<i>Lang</i>	<i>Andel förklarad varians</i>
Barnkunskap	0.756	0.237	---	---	0.628
Bild	0.495	0.324	---	0.179	0.382
Hemkunskap	0.709	0.330	---	---	0.612
Idrott	0.394	0.225	---	---	0.206
Musik	0.623	0.205	---	0.209	0.473
Biologi	0.832	0.167	0.234	---	0.775
Fysik	0.800	0.130	0.406	---	0.822
Kemi	0.821	0.138	0.339	---	0.807
Teknik	0.579	0.241	0.311	---	0.490
Geografi	0.879	0.018	---	---	0.774
Historia	0.891	---	---	---	0.794
Religionsksp	0.893	---	---	---	0.797
Samhällsksp	0.893	---	---	---	0.798
Slöjd	0.470	0.451	---	---	0.424
Svenska	0.829	---	---	0.306	0.781
Enggem	0.729	---	0.106	0.373	0.681
Matgem	0.734	0.125	0.324	0.098	0.669

matikbetyget fått något starkare laddningar i såväl Nonverb som Mathsci, medan dess laddning i Lang är i stort sett oförändrad. Som en konsekvens av dessa starkare laddningar har andelen förklarad varians också höjts påtagligt för båda ämnena.

AVSLUTANDE KOMMENTARER

Avslutningsvis kan vi konstatera att de förhållandevis enkla bortfallsåtgärder som vidtagits på UGU-R:s betygsdata fått mycket positiva effekter. Från att ha en databas som innehåller en mängd informationsluckor och där ingen av de 842 800 personerna har en komplett uppsättning av data har vi nu tillgång till en datauppsättning, där nästan 94 procent av samtliga har fullständiga uppgifter (Tabell 7). Den mest betydelsefulla åtgärden har här varit att transformera de tre alternativkursbetygen i engelska och matematik till ett enda betyg i vardera ämnet. I utgångsläget utgjorde den största gruppen med jämförbara betyg i matematik något mer än 50 procent av totalgruppen och de med jämförbara betyg i engelska omkring två tredjedelar av samtliga. I båda fallen rörde det sig om de elever som valt särskild kurs. Som Tabell 7 ovan visar har nu mer än 95 procent av samtliga jämförbara betyg i båda ämnena.

Tabell 8. Andel personer med betygsuppgifter totalt samt med uppdelning på årskull, i procent.

	<i>Födelseår</i>								
	1972	1973	1974	1975	1976	1977	1978	1979	Totalt
Antal	113230	111510	112648	106635	101565	99308	96914	100990	842800
Barnkunskap	94.9	95.3	95.4	95.5	95.5	95.9	95.6	95.0	95.4
Bild	95.3	95.8	95.8	95.8	95.8	96.4	96.1	95.7	95.8
Hemkunskap	95.2	95.5	95.7	95.7	95.7	96.2	95.9	95.5	95.7
Idrott	95.2	95.6	95.6	95.6	95.7	96.2	95.9	95.4	95.6
Musik	95.0	95.3	95.5	95.5	95.5	96.0	95.7	95.2	95.5
Slöjd	95.4	95.8	95.8	95.8	95.9	96.4	96.1	95.7	95.9
Svenska	95.2	95.6	95.6	95.5	95.6	96.0	95.6	95.1	95.5
Biologi	95.1	95.5	95.4	95.4	95.5	96.0	95.7	95.2	95.4
Fysik	94.9	95.3	95.3	95.2	95.3	95.7	95.5	95.0	95.3
Kemi	94.9	95.3	95.2	95.2	95.3	95.7	95.5	95.0	95.3
Teknik	95.1	95.5	95.5	95.5	95.5	95.9	95.6	95.0	95.5
Historia	95.1	95.6	95.6	95.6	95.6	96.1	95.8	95.3	95.6
Geografi	95.2	95.7	95.6	95.6	95.6	96.1	95.8	95.4	95.6
Religion	95.1	95.5	95.5	95.5	95.5	96.0	95.7	95.3	95.5
Samhällsksp.	95.2	95.7	95.7	95.7	95.7	96.2	95.9	95.4	95.7
Enggem	95.0	95.5	95.4	95.4	95.4	95.9	95.6	95.1	95.4
Matgem	95.6	96.1	96.1	96.1	96.1	96.6	96.3	95.9	96.1
Fullständ.betyg	93.6	93.4	93.7	93.8	93.8	94.3	94.0	93.1	93.7
Saknar samtliga betyg	4.0	3.4	3.3	3.3	3.4	3.0	3.2	3.5	3.4

Överföringen av de skilda betygen till en och samma skala har medfört att man kan studera nästan samtliga individer i en och samma analys. Om alternativkursbetygen hade bibehållits hade vi varit tvungna att hantera nio olika undergrupper för att uppfylla kravet på jämförbara betygsdata. Av betygens laddningar i de olika skolprestationsfaktorerna framgår också att denna metod för att skapa jämförbarhet varit relevant. Andelen förklarad varians är nämligen betydligt större för de justerade betygen (se Enggem och Matgem i Tabell 6) jämfört med för de betyg som byggde på en enkel sammanslagning (Engsum och Matsum i Tabell 3). Det pris vi fått betala för att få jämförbarhet i betygen över alternativkurserna är att de justerade betygen uttrycks i en skala med ett skalsteg mer än övriga betyg och detta medför att betygen i engelska och matematik har ett högre medelvärde och en något större standardavvikelse jämfört med övriga betyg. Enligt min bedömning är detta pris mycket rimligt.

Den andra större bortfallsåtgärden som vidtogs var att med hjälp av regressionsimputation ersätta blockbetygen i SO och NO med betyg i de

enskilda ämnena. På detta sätt minskades bortfallet i de enskilda ämnesbetygen från 12 procent i NO-ämnena och 16 procent i SO-ämnena till ungefär 4 procent för vardera gruppen. Bortfallet har således reducerats påtagligt och detta har vidare skett utan några nämnvärda effekter på vare sig medelvärden, standardavvikelser eller på betygens faktorstruktur.

Den mest begränsade åtgärd som vidtagits här är imputationen av betygsvärden för dem som saknat betyg i svenska, idrott och teknik och den har naturligtvis inte fått lika stora effekter på bortfallet som de två tidigare nämnda åtgärderna. Det kan ändå vara värt att uppmärksamma att inte heller denna åtgärd fått några nämnvärda konsekvenser för medelvärdena, standardavvikelseerna eller för betygens faktorladdningar.

I den ursprungliga datauppsättningen fanns totalt 27 olika betygsuppgifter. Genom att utesluta de två ospecificerade tillvals-betygen och de mycket lågfrekventa betygen i hemspråk och svenska 2 reducerades antalet till 23. Överföringen av alternativkursbetygen i engelska och matematik till ett enda betyg i vardera ämnet reducerade antalet betygsuppgifter till 19 och efter att ha ersatt blockbetygen i SO- och NO-ämnena till betyg i de enskilda ämnena återstår 17 betygsuppgifter. Som framgår av Tabell 8 (ovan) har mellan 93 och 94 procent av samtliga personer nu en komplett datauppsättning och mellan 3 och 4 procent av samtliga saknar helt betygsuppgifter. Självfallet kan täckningsgraden höjas ytterligare genom regressionsimputation eller medelvärdessimputation, men det innebär ett omfattande arbete och det är inte självklart att de åtgärderna medför sådana förbättringar av datauppsättningen att det kompenserar för arbetsinsatsen. Detta är en bedömning som får överlämnas till varje enskild användare av databasen.

Avslutningsvis skall sägas att inga betygsuppgifter uteslutits ur databasen, utan originalbetygen finns tillgängliga tillsammans med de betygsvariabler som innefattar imputerade och justerade värden.

LITTERATUR

- Allison, P.D. 1987: Estimation of linear models with incomplete data. I C. Clogg (red): *Sociological methodology*. San Francisco: Jossey Bass.
- Anderson, A.B., Basilevsky, A. & Hum, D.P.J. 1983: Missing data: A review of the literature. I P.H. Rossi, J.D. Wright & A.B. Anderson (red): *Handbook of survey research*. New York: Academic Press.
- Andersson A. 1998: The dimensionality of the leaving certificate in the Swedish compulsory school. *Scandinavian Journal of Educational Research*, 42(1), 25–40.
- Arbuckle, J.L. 1997: *Amos User's Guide. Version 3.6*. Chicago: Small Waters Corporation.
- Beale, E.M.L. & Little, R.J.A. 1975: Missing values in multivariate analysis. *Journal of the Royal Statistical Society, Series B*, 37, 129–145.
- Bello, A.L. 1993: Choosing among imputation techniques for incomplete multivariate data: A simulation study. *Communications in Statistics – Theory and Methods*, 22(3), 853–877.
- Chan, L.S. & Dunn, O.J. 1972: The treatment of missing values in discriminant analysis – I. The sampling experiment. *Journal of the American Statistical Association*, 67, 473–477.

- Chan, L. S. Gilman, J.A. & Dunn, O.J. 1976: Alternative approaches to missing values in discriminant analysis. *Journal of the American Statistical Association*, 71, 842–844.
- Dempster, A.P., Laird, N.M. & Rubin, D. 1977: Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.
- Ford, B.L. 1983: An overview of hot-deck procedures. I W.G. Madow, I Olkin & D.B Rubin (red): *Incomplete data in sample Surveys. Volume II: Theory and bibliographies*. New York: Academic Press.
- Frane, J.W. (1976) Some simple procedures for handling missing data in multivariate analysis. *Psychometrika*, 41(3), 409–415.
- Gleason, T.C. & Staelin, R. 1975: A proposal for handling missing data. *Psychometrika*, 40(2), 229–252.
- Graham, J.W. & Donaldson, S.I. 1993: Evaluating interventions with differential attrition: The importance of nonresponse mechanisms and use of follow-up data. *Journal of Applied Psychology*, 78(1), 119–128.
- Gustafsson, J-E. & Stahl, P.A. 2000: *STREAMS. User's Guide. Version 2.5 for Windows 95/98/NT/2000*. Mölndal: Multivariate Ware.
- Hammarström, M. 1996: *Varför inte högskola? En longitudinell studie av olika faktorerers betydelse för studiebegåvade ungdomars utbildningskarriär*. (Göteborg Studies in Educational Sciences 107) Göteborg: Acta Universitatis Gothoburgensis.
- Heitjan, D.E. 1997: Annotation: What can be done about missing data? Approaches to imputation. *American Journal of Public Health*, 87(4), 548–550.
- Jöreskog, K.G. & Sörbom, D. 1993: *LISREL 8: Structural Equation Modeling with the SIMPLIS Command Language*. Hillsdale, NJ: Erlbaum.
- Kim, J.O. & Curry, J. 1977: The treatment of missing data in multivariate analysis. *Sociological Methods & Research*, 6(2), 215–240.
- Laird, N.M. 1988: Missing data in longitudinal studies. *Statistics in Medicine*, 7, 305–315.
- Little, R.J.A. 1988: Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287–296.
- Little, R.J.A. & Rubin, D.B. 1987: *Statistical analysis with missing data*. New York: Wiley.
- Little, R.J.A. & Schenker, N. 1995: Missing data. I G. Arminger, C.C. Clogg & M.E. Sobel (red): *Handbook of statistical modeling for the social and behavioral sciences*. New York: Plenum.
- Malhotra, N.K. 1987: Analyzing marketing research data with incomplete information on the dependent variable. *Journal of Marketing Research*, 24, 74–84.
- Muthén, B.O., Kaplan, D. & Hollis, M. 1987: On structural equation modeling with data that are not missing completely at random. *Psychometrika*, 52, 431–462.
- Muthén, L.K. & Muthén, B.O. 1998: *Mplus User's Guide*. Los Angeles: Muthén & Muthén.
- Raaijmakers, Q.A.W. 1999: Effectiveness of different missing data treatments in surveys with Likert-type data: Introducing the relative mean substitution approach. *Educational and Psychological Measurement*, 59(5), 725–749.
- Raymond, M.R. 1986: Missing data in evaluation research. *Evaluation & the Health Profession*, 9, 395–420.
- Raymond, M.R. & Roberts, D.M. 1987: A comparison of methods for treating incomplete data in selection research. *Educational and Psychological Measurement*, 47, 13–26.
- Rizvi, M.H. 1983: Hot-deck procedures: Introduction. I W.G. Madow, I. Olikon & D.B. Rubin (eds): *Incomplete data in sample surveys. Volume III: Proceedings*

- of the symposium*. New York: Academic Press.
- Roth, P.L. 1994: Missing data: A conceptual review for applied psychologists. *Personnel Psychology*, 47, 537–560.
- Roth, P.L., Campion, J.E. & Jones, S.D. 1996: The impact of four missing data techniques on validity estimates in human resource management. *Journal of Business and Psychology*, 11, 101–112.
- Rubin, D.B. 1987: *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Skolverket 1996: *Grunden för fortsatt lärande. En internationell jämförande studie av vuxnas förmåga att förstå och använda tryckt och skriven information*. (Skolverkets rapport nr 115) Stockholm: Skolverket.
- Wotke, W. 2000: Longitudinal and multigroup modeling with missing data. I T.D. Little, K.U. Schnabel & J. Baumert (red): *Modeling longitudinal and multilevel data. Practical issues, applied approaches, and specific examples*. New Jersey: Lawrence Erlbaum.
- Ögren, G. 1998: *Utprovning av uppgifter till högskoleprovet. Utvärdering av försöksverksamhet med en ny utprovningsmodell*. (PM 140) Umeå: Umeå universitet, Enheten för pedagogiska mätningar.