Institutionen för tillämpad utbildningsvetenskap
Beteendevetenskapliga mätningar
Högskoleprovet

RAPPORT
2013-10-01

# Det nya högskoleprovet – Några valideringsstudier

# Förord

Hösten 2012 fick Institutionen för tillämpad utbildningsvetenskap (TUV) vid Umeå universitet ett uppdrag av högskoleprovets dåvarande huvudman, Högskoleverket, om utvärdering av resultaten på det nya högskoleprovet inom ramen för den försöksverksamhet som gällde för införandet av det nya provet. Denna rapport redovisar resultaten från denna utvärdering till den nuvarande huvudmannen för provet, Universitets- och högskolerådet.

Det övergripande ansvaret för att genomföra utvärderingen har legat på Gunilla Ögren och Per-Erik Lyrén, projektledare respektive vetenskaplig ledare för högskoleprovsprojektet vid Umeå universitet. Förutom dessa två har ett flertal andra personer varit inblandade i de studier som utgör utvärderingen. Studien om värdet i delprovspoäng i kapitel 2 är gjord av Jonathan Wedman, doktorand i beteendevetenskapliga mätningar vid TUV, och Per-Erik Lyrén. Studien om positionseffekter i kapitel 3 är gjord av Jenny Lindberg, provutvecklare för högskoleprovet vid TUV. Studien kopplad till kapitel 4 är gjord av Gunilla Ögren, och studien kopplad till kapitel 5 är gjord av Magnus Wikström, professor i nationalekonomi vid Umeå universitet, och Christina Wikström, universitetslektor i beteendevetenskapliga mätningar vid Umeå universitet. Övriga studier är gjorda av Per-Erik Lyrén.

Umeå 2013-10-01

Per-Erik Lyrén
Gunilla Ögren

# Innehåll

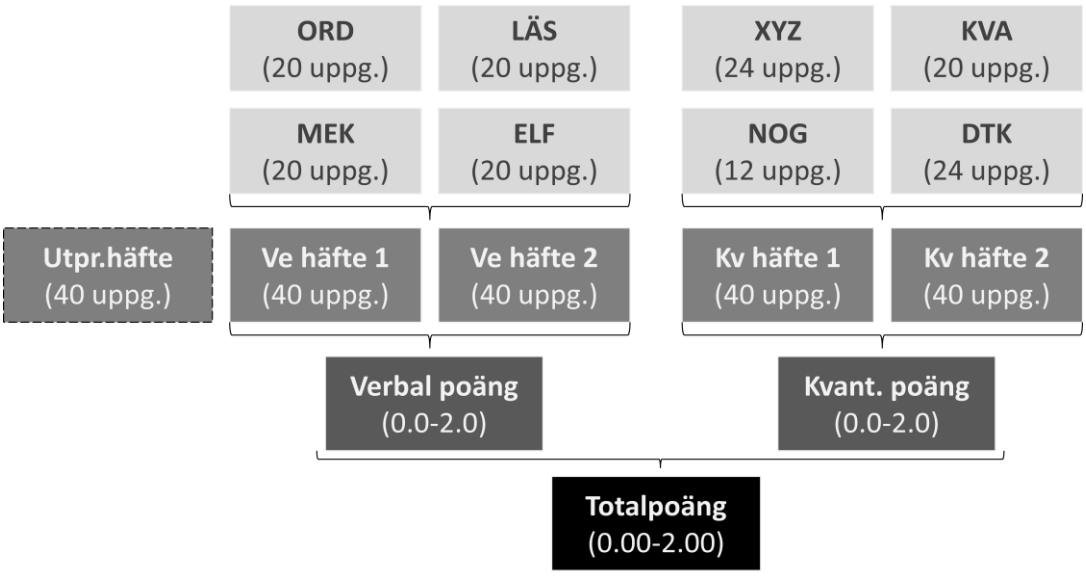
# 1. Inledning

## Rapportens upplägg

Rapporten innehåller ett antal studier och består av dels en huvudrapport med ett antal kapitel och dels en bilagedel. Mindre omfattande studier rapporteras i sin helhet i huvudrapporten medan mer omfattande studier sammanfattas i huvudrapporten och finns som komplett rapport i bilagedelen. Kapitel 1 innehåller en beskrivning av uppdraget, en bakgrund till det nya högskoleprovet, en beskrivning av det validitetsramverk som ligger till grund för vilka specifika studier som genomförts och hur dessa grupperats. Kapitel 2 redovisar studier om provpoängens reliabilitet och interna struktur samt mervärdet i delprovspoäng. Kapitel 3 redovisar studier om icke-svar och positionseffekter. Kapitel 4 redovisar provresultat för olika grupper av provtagare, och Kapitel 5 redovisar en studie om samband mellan provresultat och betyg från gymnasieskolan. Kapitel 6 är en sammanfattning med slutsatser och rekommendationer, och Kapitel 7 innehåller angivna referenser.

## En kort beskrivning av det nya högskoleprovet

Det nya högskoleprovet introducerades hösten 2011 och består av två provdelar, en kvantitativ och en verbal, och fyra delprov inom varje provdel. De kvantitativa delproven är XYZ (matematisk problemlösning), KVA (kvantitativa jämförelser), NOG (matematisk-logiska resonemang) och DTK (bearbetning och tolkning av information i diagram, tabeller, kartor). De verbala delproven är ORD (förståelse av isolerade ord och begrepp), LÄS (svensk läsförståelse), MEK (förståelse av ord och begrepp satta i ett sammanhang) och ELF (engelsk läsförståelse). Respektive provdel är uppdelad på två häften med 40 uppgifter vardera, där varje häfte innehåller alla fyra delprov i respektive provdel. Provtiden för varje häfte är 55 minuter, och på de olika delproven sker inte separat tidtagning. Alltså ska provtagaren under provtagaren under provdagen gå igenom fyra häften om totalt 160 uppgifter, samt även ett femte häfte med uppgifter som ska prövas ut och som därmed inte räknas in i provpoängen (ordningen på häftena varierar mellan olika provtillfällen). Antalet rätt på varje provdel omvandlas till en normerad poäng mellan 0.0 och 2.0, och medelvärdet av dessa två poäng utgör den totala normerade poängen (0.00–2.00) och det är denna totalpoäng som i dagsläget används i urvalet till högre utbildning. Figur 1:1 visar en översikt av sammansättningen på det nya högskoleprovet.




Figur 1:1. Översikt av sammansättningen på det nya högskoleprovet.

# En bakgrund till det nya högskoleprovet[1]

## Utvecklingsuppdraget

De förändringar som ledde fram till det nya högskoleprovet har sitt ursprung till stor del i en utredning av Högskoleverkets provråd (Högskoleverket, 2000), i en internationell utvärdering av provet (Högskoleverket, 2002) och i synpunkter från högskoleprovets internationella vetenskapliga råd. I de två rapporterna från Högskoleverket framförs att provets prognosförmåga skulle kunna förbättras om provpoängen differentierades genom separata provdelspoäng, exempelvis verbal respektive kvantitativ poäng. 2006 fick Umeå universitet i uppdrag att utveckla och pröva ut nya delprov för högskoleprovet som komplement till de delprov som redan fanns. Provet skulle bli mer flexibelt genom att ha åtskiljda provdelar, vilket kunde möjliggöra för olika utbildningar att lägga större vikt vid en provdel än en annan. Målsättningen var därmed att ha en jämn fördelning mellan antalet verbala och kvantitativa uppgifter. Ett annat mål med revideringarna, vilket var en framträdande synpunkt från det internationella vetenskapliga rådet, var att provet skulle bli mer effektivt i termer av antalet uppgifter per tidsenhet.

## Förändringar i den verbala provdelen[2]

En förändring som var önskvärd i det här sammanhanget var att minska antalet ORD-uppgifter, eftersom det utgjorde många gånger fler poäng per tidsenhet än de andra delproven och att det är ett format som inte är så verklighetsnära då det handlar om ordkunskap utan sammanhang. Analogier (ANA) och meningskomplettering (MEK) prövades ut i ett första skede, och båda proven uppvisade goda tekniska mätegenskaper. Dock ansågs ANA vara alltför konstlat med tvetydiga uppgifter, så arbetet med detta delprov avbröts i juni 2008. Utvecklingsarbetet med MEK fortsatte och finns med i det nya högskoleprovet, då det ansågs vara ett bra komplement till ORD och LÄS då provet mäter såväl ordkunskap som läsförståelse. I och med detta halveras antalet uppgifter i ORD.

## Förändringar i den kvantitativa provdelen

Som komplement till de kvantitativa delproven utvecklades ett prov, VAN, som mäter verbal analytisk resonemangsförmåga. Detta delprov var emellertid alltför resurskrävande och känsligt för coachning och avfärdades därför i ett tidigt skede. Begreppet verbal analytisk förmåga var dock fortfarande av intresse, och därför utvecklades uppgifter (benämnda AR) som ansågs mäta detta begrepp men i samma format som det redan befintliga NOG-provet. AR-uppgifter prövades först ut som en del av NOG-provet, men i samband med den första större utprövningen av de nya provsammansättningarna våren 2009 så prövades AR ut som ett eget delprov; här emellertid i ett häfte med verbala delprov (ORD, MEK och LÄS eller ELF). Anledningen till detta var att AR i teorin skulle kunna ha minst lika starka samband med de verbala delproven som med de kvantitativa, samt att ELFs berättigande i högskoleprovet ifrågasattes av det internationella vetenskapliga rådet och därmed skulle AR vara en möjlig ersättning till detta prov.

Förutom AR prövades provformatet kvantitativa jämförelser, KVA, vilket har haft en lång historia som del i SAT. Provformatet plockades bort från SAT i samband med de stora förändringarna som genomfördes efter kritiken från University of California, möjligen på grund av att den typen av uppgifter ansågs vara tvetydiga. Fördelen med KVA-provet är att det liksom NOG (och DTK) har en analytisk komponent, om än inte lika stor, och att det samtidigt är mindre tidskrävande än både NOG och DTK. I de första utprövningarna uppvisade KVA goda mättekniska egenskaper (reliabilitet, svårighetsgrad, och diskrimineringsgrad).

---

[1] Innehållet under denna rubrik är till stora delar hämtat från Stage & Ögren (2010).
[2] Här hänvisas till begreppet provdel även när det gäller det gamla provet, även om dessa provdelar inte sågs som separata delar på samma sätt som i det nya provet.

Det internationella vetenskapliga rådet efterlyste också ett mer generellt prov av matematisk problemlösningsförmåga. Utifrån detta utarbetades XYZ som innehåller uppgifter som rent innehållsmässigt har en nära koppling till kursplanerna för gymnasieskolans tidiga kurser i matematik och som uppgiftsmässigt ligger närmare undervisning och läromedel än vad de andra kvantitativa delproven gör. Likt uppgifterna i KVA är uppgifterna i XYZ mindre tidskrävande än uppgifterna i NOG och DTK, vilket skulle göra den kvantitativa provdelen betydligt mer effektiv (som definierat tidigare). Både KVA och XYZ prövades ut vid den första större utprövningen våren 2009 och mättekniskt visade de sig fungera tillfredsställande.

## Utprövning av ny provsammansättning hösten 2009

Efter utprövningen våren 2009 gjordes en del förändringar i provsammansättningen, och den provsammansättning som nu gäller för det nya provet prövades ut hösten 2009. När det gäller den kvantitativa provdelen konstaterades att sambanden mellan XYZ och KVA å ena sidan och NOG och DTK å andra sidan inte var så starka som önskvärt men ändå acceptabla. XYZ och KVA visade sig återigen fungera mättekniskt bra, och drygt 80 procent av dessa uppgifter uppfyllde de mättekniska kraven. Det konstaterades även att könsskillnaden var något lägre i genomsnitt per uppgift, men att den totala könskillnaden kommer att öka i det reguljära provet då antalet kvantitativa uppgifter ökar från 42 till 80.

I den verbala provdelen fungerade samtliga fyra delprov (ORD, LÄS, MEK, ELF) tillfredsställande rent mättekniskt. MEK-uppgifterna var överlag för lätta, men detta skulle kunna justeras till ett reguljärt prov. Det konstaterades även att skillnaderna i resultat för olika åldersgrupper torde bli något mindre med MEK-provet, men att rangordningen i totalgruppen inte torde påverkas nämnvärt och att könsskillnaderna på den verbala provdelen kan förväntas vara ungefär som i det gamla provet.

## Utvärdering av den nya provsammansättningen efter utprövningen hösten 2009

Efter utprövningen i samband med det reguljära högskoleprovet hösten 2009 gjordes en utvärdering av den nya provsammansättningen. Fokus var på gruppjämförelser, och de variabler som fanns tillgängliga (via SCB) var ursprung (svenskt/utländskt), socialgruppstillhörighet (SEI; 3 grupper), föräldrarnas utbildning, kön, ålder och utbildningsprogram på gymnasiet.

Gällande den kvantitativa provdelen konstateras följande:

- Liksom i tidigare analyser gäller att könsskillnaden per uppgift är mindre på försöksproven XYZ och KVA än på de reguljära delproven NOG och DTK. Den totala könsskillnaden kommer dock att bli större eftersom den kvantitativa delen i det närmast kommer att fördubblas.
- Avseende ålder är det gruppen 20 år eller yngre som presterar bäst. Skillnaden per uppgift är ungefär som för NOG och DTK i gamla provet, men den totala skillnaden kommer att bli större i och med att antalet uppgifter blir större.
- Skillnaden mellan grupper med olika ursprung är betydligt mindre på de nya delproven XYZ och KVA än på NOG och DTK.
- Skillnaden mellan olika socialgrupper är på samma nivåer som för tidigare reguljära prov. Föräldrarnas och provtagarnas utbildningsnivå har fortfarande stor betydelse för resultaten på den kvantitativa provdelen.
- Provtagare från naturvetenskapligt program på gymnasiet har som tidigare de högsta poängen på samtliga delprov på såväl försöksprov som reguljära prov.

Gällande den verbala delen konstateras följande:

- Skillnader mellan kvinnor och män torde fortsatt vara liten på det reguljära provet. ELF-provet står ensamt för merparten av könsskillnaden på den verbala provdelen.

- Skillnader mellan åldersgrupper kommer förmodligen att minska något då antalet ORD-uppgifter halveras.
- Skillnader mellan grupper med olika ursprung kommer troligen inte att påverkas nämnvärt med den nya verbala provsammansättningen.
- Skillnaden mellan olika socialgrupper är på samma nivåer och följer samma mönster som för tidigare reguljära prov. Föräldrarnas och provtagarnas utbildningsnivå har fortsatt stor betydelse för resultaten.
- Provtagare från naturvetenskapligt program på gymnasiet har som tidigare de högsta resultaten på samtliga verbala delprov på såväl försöksprov som reguljära prov.

# Utvärdering och validitet

Den aktuella utvärderingen har gjorts utifrån ett validitetsperspektiv och de syften och antaganden som låg bakom förändringarna. Enligt *Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], & National Council on Measurement in Education [NCME], 1999) bör en valideringsprocess göras utifrån fem källor till validitetsbevis:

1. bevis baserade på innehåll,
2. bevis baserade på svarsprocesser,
3. bevis baserade på intern struktur,
4. bevis baserade på relationen till andra variabler, och
5. bevis baserade på konsekvenser av testningen.

När det gäller den första källan till bevis (innehåll) så kan man konstatera att sådana bevis till stor del kan dras från det utvecklingsarbete som legat till grund till förändringarna med provet. Den här typen av bevis är för övrigt mindre viktiga (om än inte oviktiga) när det gäller ett prov som högskoleprovet, vilket inte är strikt baserade på exempelvis kursplaner eller andra styrdokument.

När det gäller den andra källan till bevis (svarsprocesser) så är det vanligtvis analyser av individuella provtagares svar, för att avgöra om de avger ett visst svar av rätt anledning (t.ex. genom tänka högt-studier). Det kan även handla om själva processen från det att provtagaren bestämt sig för ett svar till det att det tänkta svaret behandlats och blivit en del av provtagarens resultat, eller om svarshäftets utformning och den optiska läsningen av svarshäftena.

Validitetsbevis baserade på intern struktur kan komma från analyser av hur väl förhållandet mellan olika provuppgifter eller olika delar av provet stämmer överens med det teoretiska begrepp som avses att mätas.

Bevis baserade på relationen till andra (externa) variabler är exempelvis, när det handlar om prov för urval till högskolan, sambandet mellan provresultat och prestation i högskolan. I detta fall antas det finnas ett (starkt) positivt samband med den externa variabeln (prestation i högskolan), men det kan också handla om externa variabler som man antar att det finns ett negativt samband med eller inget samband alls. Externa variabler kan även vara olika typer av kategorivariabler, vilka exempelvis kan indikera grupptillhörighet (t.ex. kön och ursprung). När det gäller den typen av variabler kan det finns underliggande antaganden om att en viss grupptillhörighet ska påverka eller inte påverka prestationen på provet.

När det slutligen gäller bevis baserade på konsekvenser av testningen så handlar det om att undersöka såväl avsedda som ej avsedda konsekvenser, och då framförallt sådana som kan spåras till en källa av ”invaliditet” (*invalidity*, ogiltighet) som begreppsunderrepresentation eller begreppsirrelevanta komponenter. Ta konsekvenser av könsskillnader i provprestation som exempel. Anta att män har

större chans att bli antagen till högskolan för att de presterar bättre på provet än kvinnor. Om orsaken till prestationsskillnaden är att män är bättre på det som provet avser att mäta så är inte det en källa till invaliditet. Däremot kan det uppstå problem som en följd av detta, exempelvis att könsfördelningen på vissa utbildningar blir så pass sned att det får negativa följdkonsekvenser. Om orsaken däremot är att provet mäter något som är helt orelaterat till prestationer i högskolan eller om det mäter något irrelevant som just män är bättre på så är det en källa till invaliditet.

# Valideringsstudier

För den aktuella utvärderingen ansågs det rimligt, utifrån de syften och antaganden som låg bakom förändringarna samt de data som är tillgängliga, att i första hand fokusera på att undersöka samband mellan prov och externa variabler. Även svarsprocesser och intern struktur är i det här sammanhanget viktiga delar att studera. För svarsprocesser gäller det till exempel frågor om bortfall, eftersom vi vet från de första utprövningarna att provtagarna tyckte att tiden var för kort, speciellt för de kvantitativa häftena. Sådana studier är framförallt informativa när det gäller effekter av ”effektiviseringen” av provet. För intern struktur handlar det främst om att undersöka om provets teoretiska modell i termer av uppdelningen i en verbal och en kvantitativ del har stöd i empiriska data. Sådana studier är framförallt informativa när det gäller effekter av uppdelningen av en kvantitativ respektive verbal provdel i den nya provsammansättningen.

När det gäller externa variabler som är av intresse så hade det naturligtvis varit mycket värdefullt att ha tillgång till kriteriedata, exempelvis antal högskolepoäng efter första årets studier, för att kunna studera det nya provets prognosförmåga. Detta finns dock inte tillgängligt i någon större skala ännu eftersom merparten av de som tog det nya provet första gången hösten 2011 använde det vid antagningen till studier som startade hösten 2012. Det som ligger närmast en relevant kriterievariabel är någon annan form av prestationsmått, och där är betyg från grundskola och gymnasium relevanta variabler.

## Studier med fokus på samband med externa variabler

### *Samband mellan provresultat och betyg*

Här undersöks hur poäng på provet, såväl totalt som från de två provdelarna och delproven, är relaterade till provtagarnas betyg. Betygsvariablerna kan vara såväl betygsgenomsnitt från grundskola respektive gymnasium som kursbetyg i enskilda ämnen. Speciellt informativt är det att undersöka

(1) sambandet mellan totalpoäng på provet och jämförelsetalet från gymnasiet (men även meritvärdet från grundskolan), och
(2) sambandet mellan den kvantitativa provdelen, som nu inkluderar de nya och mer matematiskt inriktade delproven XYZ och KVA, och betyg i matematik och poäng på nationella prov i matematik.

När det gäller det förstnämnda så är det värt att poängtera att vi inte kan förvänta oss mycket höga samband mellan provpoäng och betyg, eftersom dessa mått i stor utsträckning mäter olika saker och under helt olika tidsramar. Det andra är intressant framförallt då de nya delproven, och därmed även den kvantitativa delen överlag, till viss del utgår från gymnasieskolans kursplaner i matematik. Inte heller här kan man förvänta sig ett alltför högt samband, då kursplanerna är generella och högskoleprovet inte har specifika läroböcker att rätta sig efter. Studierna under denna rubrik rapporteras i Kapitel 5 och Bilaga D.

### *Gruppskillnader i provresultat*

I data finns uppgifter om bakgrundsvariabler såsom ursprung, kön och föräldrars utbildningsnivå. Föräldrars utbildningsnivå kan ses som en proxyvariabel för socioekonomisk indelning (SEI), vilket

inte finns i dessa data. Det är relevant att undersöka resultat på hela provet samt på delprovs- och provdelsnivå utifrån grupptillhörighet för att se olika gruppers relativa prestationer. Det är även av intresse att undersöka hur sambandet mellan provresultat och betyg varierar mellan olika grupper. Studierna under denna rubrik replikerar till stor del de studier som gjordes på försöksproven hösten 2009 och rapporteras i Kapitel 4 och Bilaga C.

## Studier med fokus på svarsprocesser

***Icke-svar***
I och med att det nya provet kan antas vara mer speedat än det gamla är det relevant att undersöka förekomsten av icke-svar i provdata, både vad gäller frekvens och effekt på kvaliteten i data. Det är även relevant att undersöka samband mellan icke-svar och prestation: Är det provtagare med höga eller låga prestationer som lämnar många uppgifter obesvarade, eller finns det inget samband däremellan? Studierna under denna rubrik rapporteras i Kapitel 3 och Bilaga B.

***Positionseffekter***
Studier på det gamla provet har visat att vissa delprov är känsliga för positions- och ordningseffekter, t.ex. att provtagare som fått ett visst delprov som utprövning och samma delprov som reguljärt prov direkt efter kan ha fördel av detta. På delprovsnivå är det dock prioriterat att undersöka positionseffekter, dvs. huruvida provtagarnas prestation på vissa uppgifter är beroende av om uppgifterna kommer längre fram eller längre bak i häftet. Speciellt fokus i denna fråga ligger på DTK-provet, eftersom det finns starka indikationer på att de kvantitativa häftena är tidspressade och DTK ligger sist i dessa häften. Studierna under denna rubrik rapporteras i Kapitel 3.

## Studier med fokus på intern struktur

***Reliabilitet***
Reliabilitet handlar om mätningens tillförlitlighet, exempelvis att provtagare som tar två olika versioner av provet ska få överensstämmande resultat. Reliabiliteten är framförallt viktig för de poäng som direkt ligger till grund för någon typ av beslut, dvs. i första hand totalpoängen men även provdelspoängen eftersom tanken är att de ska kunna användas i urvalet någon gång i framtiden. Reliabiliteten på delprovsnivå är inte av lika avgörande betydelse, men det är ändå intressant att undersöka detta eftersom dessa poäng rapporteras till provtagarna. Studierna under denna rubrik rapporteras i Kapitel 2.

***Dimensionalitet***
Ett par antaganden bakom förändringen av provet är att de två provdelarna mäter olika begrepp (om än korrelerade sådana), och att det nya provet är mer tydligt tvådimensionellt än det gamla provet. Dimensionaliteten undersöks genom faktoranalyser där man jämför den teoretiska modellen med empiriska provdata. Studierna under denna rubrik rapporteras i Kapitel 2 och Bilaga A1.

***Mervärdet i delprovspoäng***
I det gamla provet hade delproven överlag moderata samband med de övriga delproven och de hade generellt sett hög reliabilitet, vilket gjorde att det fanns ett mervärde i att rapportera provtagarnas poäng på de olika delproven jämfört med att bara rapportera totalpoängen. Med det nya provet är läget annorlunda, eftersom delproven i respektive provdel har höga samband med varandra och det är i genomsnitt färre uppgifter per delprov, och det är därmed inte lika självklart att det finns ett mervärde i att rapportera delprovspoängen till provtagarna. Studierna under denna rubrik rapporteras i Kapitel 2 och Bilaga A2.

# 2. Reliabilitet, intern struktur och mervärdet i delprovspoäng

## Reliabilitet

Reliabiliteten på det gamla provet har varit genomgående hög. Medelvärdet för alfakoefficienten (eng. *coefficient alpha*; även *Cronbach's alpha*), vilket är ett mått på provpoängens interna konsistens, är 0,93 för proven 08:A-11:A[3]. Motsvarande värde för det nya provet 11:B-13:A är även det 0,93, vilket är över gränsen 0,9 som kan ses som ett idealt värde (se t.ex. Streiner, 2003). Det går även att skatta reliabiliteten genom parallella prov-metoden, vilket innebär att man tittar på korrelationen mellan poängen från två provomgångar. Korrelationen mellan den normerade poängen 11:B och motsvarande värde 12:A var 0,90, och samma värde erhölls för korrelationen mellan 09:A och 09:B. Dessa korrelationer är korrigerade för den beskärning i spridning som uppstår på grund av att det sker en viss selektion i provtagargruppen från första till andra provtillfället, och därför blir den okorrigerade korrelationen en underskattning av det faktiska sambandet. Det ska dock noteras att spridningen är marginellt lägre i den upprepande provtagargruppen, och effekten av korrigeringen (Thorndikes fall 2-korrigering; se t.ex. Wiberg & Sundström, 2009) är därför liten. En skattning ger att den korrigerade korrelationen är 0,90, dvs. marginellt högre än den icke-korrigerade. En anledning till att parallella prov-värdena blir något lägre än alfakoefficientvärdena är att de är baserade på olika typer av poäng, nämligen normerad poäng respektive råpoäng (antal rätt).

På provdelsnivå visar analyserna att reliabiliteten (alfakoefficienten) för den verbala provdelen (VERB) är i genomsnitt 0,91 för de fyra omgångarna av det nya provet, vilket är ett högt värde, medan det är något lägre på den kvantitativa provdelen (KVANT) med ett genomsnittligt alfavärde på 0,89. Ett annat sätt att skatta reliabiliteten på provdelarna är genom den s.k. *split half*-metoden där korrelationen mellan de två häftena inom respektive provdel skattas och sedan korrigeras med Spearman-Browns formel. Ett argument för att använda *split half*-metoden på det nya provet är att häftena inom respektive provdel är konstruerade för att vara så gott som strikt parallella, vilket gör att *split half*-metoden bör ge tillförlitliga reliabilitetsskattningar. Som kan ses i Tabell 2:1 så överensstämmer alfakoefficientvärdena och *split half*-värdena väl med varandra.

[3] Ett A efter årtalet betecknar ett vårprov och ett B betecknar ett höstprov.

Tabell 2:1. Reliabilitetsskattningar för det gamla och det nya högskoleprovet, totalt samt uppdelat på verbal och kvantitativ provdel.

| Prov | Totalt | | Verbal provdel | | Kvantitativ provdel | |
|---|---|---|---|---|---|---|
| | alfa | PP [1] | alfa | S-B[2] | alfa | S-B[2] |
| 08:A | ,92 | | | | | |
| 08:B | ,93 | | | | | |
| 09:A | ,93 | ,90 | ,91 | | ,86 | |
| 09:B | ,92 | | ,90 | | ,84 | |
| 10:A | ,93 | | ,92 | | ,86 | |
| 10:B | ,93 | | ,91 | | ,85 | |
| 11:A | ,93 | | ,91 | | ,83 | |
| Medelvärde gamla provet | ,93 | ,90 | ,91 | | ,85 | |
| 11:B | ,94 | ,90 | ,91 | ,91 | ,91 | ,91 |
| 12:A | ,93 | | ,91 | ,89 | ,89 | ,88 |
| 12:B | ,94 | | ,91 | ,91 | ,89 | ,87 |
| 13:A | ,94 | | ,90 | ,89 | ,89 | ,88 |
| Medelvärde nya provet | ,93 | ,90 | ,91 | ,90 | ,89 | ,89 |

[1] PP = Parallella prov. Korrelationen mellan normerade poäng, 09:A–09:B respektive 11:B–12:A. Värdena är korrigerade för beskuren spridning genom Thorndikes fall 2-korrigering (se t.ex. Wiberg & Sundström, 2009).
[2] S-B = Spearman-Brown-korrigerade split half-korrelationer.

En undersökning av reliabiliteten på delprovsnivå i den verbala provdelen visar att ORD, MEK och ELF har vad som kan anses vara en ”tillräcklig” reliabilitet, dvs. över 0,7 (Streiner, 2003; Nunnally & Bernstein, 1994), medan LÄS har en lägre reliabilitet än så. I den kvantitativa provdelen är det NOG och DTK som har reliabilitetsvärden som ligger på gränsen eller under 0,7. Att NOG har så lågt värde är inte oväntat, givet att provet har så pass få uppgifter. Att LÄS och DTK har relativt låg reliabilitet har förmodligen att göra med att dessa prov har en större variation i provinnehåll än många av de andra delproven. LÄS-texter och DTK-figurer kommer från olika ämnesområden och provtagarnas intresse och förförståelse för innehållet kan därmed variera påtagligt, vilket gör att urvalet av texter och figurer i dessa prov är en källa till variation som inte finns på samma sätt i de andra delproven.

Tabell 2:2. Interkorrelationer mellan delprov och provdelar hösten 2011 (kursiverade korrelationer är korrigerade för reliabilitetsbrister), samt reliabilitetsskattningar i form av alfakoefficienter (på diagonalen).

| 11:B | ORD | LÄS | MEK | ELF | XYZ | KVA | NOG | DTK | V1 | V2 | V | K1 | K2 | K |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORD | **,80** | *,81* | *,96* | *,76* | *,26* | *,29* | *,43* | *,49* | | | | *,39* | *,37* | *,38* |
| LÄS | ,58 | **,63** | *,90* | *,88* | *,47* | *,51* | *,62* | *,63* | | | | *,59* | *,58* | *,58* |
| MEK | ,75 | ,62 | **,76** | *,85* | *,42* | *,46* | *,58* | *,61* | | | | *,55* | *,53* | *,54* |
| ELF | ,59 | ,61 | ,64 | **,75** | *,48* | *,54* | *,62* | *,66* | | | | *,61* | *,60* | *,60* |
| XYZ | ,21 | ,34 | ,33 | ,38 | **,83** | *,95* | ,83 | *,66* | *,42* | *,42* | *,42* | | | |
| KVA | ,23 | ,35 | ,35 | ,41 | ,75 | **,76** | *,91* | *,74* | *,47* | *,47* | *,47* | | | |
| NOG | ,31 | ,39 | ,41 | ,44 | ,61 | ,64 | **,65** | *,80* | *,59* | *,58* | *,59* | | | |
| DTK | ,36 | ,41 | ,44 | ,47 | ,49 | ,53 | ,53 | **,68** | *,63* | *,62* | *,63* | | | |
| V1 | | | | | ,35 | ,37 | ,43 | ,47 | **,82** | *,99* | | *,56* | *,54* | *,55* |
| V2 | | | | | ,35 | ,38 | ,43 | ,47 | ,83 | **,85** | | *,55* | *,54* | *,55* |
| V | | | | | ,37 | ,39 | ,45 | ,49 | | | **,91** | *,56* | *,54* | *,55* |
| K1 | ,32 | ,43 | ,44 | ,48 | | | | | ,46 | ,47 | ,48 | **,83** | *1,00* | |
| K2 | ,30 | ,42 | ,42 | ,47 | | | | | ,44 | ,45 | ,47 | ,83 | **,83** | |
| K | ,32 | ,44 | ,45 | ,50 | | | | | ,47 | ,48 | ,50 | | | **,91** |

Not: V1 = 1:a Verbala häftet; V2 = 2:a Verbala häftet; V = Verbala provdelen; K1 = 1:a Kvantitativa häftet; K2 = 2:a Kvantitativa häftet; K = Kvantitativa provdelen.

Tabell 2:3. Interkorrelationer mellan delprov och provdelar våren 2012 (kursiverade korrelationer är korrigerade för reliabilitetsbrister), samt reliabilitetsskattningar i form av alfakoefficienter (på diagonalen).

| 12:A | ORD | LÄS | MEK | ELF | XYZ | KVA | NOG | DTK | V1 | V2 | V | K1 | K2 | K |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ORD | **,76** | *,82* | *,95* | *,75* | *,32* | *,31* | *,44* | *,52* | | | | *,42* | *,43* | *,43* |
| LÄS | ,57 | **,65** | *,91* | *,86* | *,53* | *,53* | *,64* | *,65* | | | | *,63* | *,65* | *,64* |
| MEK | ,71 | ,63 | **,74** | *,86* | *,46* | *,46* | *,57* | *,60* | | | | *,56* | *,57* | *,57* |
| ELF | ,56 | ,59 | ,63 | **,72** | *,50* | *,51* | *,59* | *,62* | | | | *,59* | *,61* | *,60* |
| XYZ | ,24 | ,37 | ,34 | ,37 | **,76** | *,99* | ,83 | *,64* | *,46* | *,49* | *,48* | | | |
| KVA | ,22 | ,36 | ,33 | ,36 | ,71 | **,68** | *,88* | *,64* | *,46* | *,49* | *,48* | | | |
| NOG | ,31 | ,42 | ,40 | ,41 | ,59 | ,59 | **,67** | *,74* | *,58* | *,59* | *,59* | | | |
| DTK | ,38 | ,44 | ,44 | ,44 | ,47 | ,45 | ,51 | **,71** | *,62* | *,63* | *,63* | | | |
| V1 | | | | | ,37 | ,35 | ,43 | ,48 | **,85** | *,97* | | *,57* | *,58* | *,57* |
| V2 | | | | | ,38 | ,36 | ,44 | ,48 | ,80 | **,81** | | *,59* | *,61* | *,60* |
| V | | | | | ,39 | ,37 | ,46 | ,50 | | | **,91** | *,58* | *,60* | *,59* |
| K1 | ,33 | ,45 | ,43 | ,45 | | | | | ,46 | ,47 | ,49 | **,79** | *,99* | |
| K2 | ,34 | ,47 | ,44 | ,46 | | | | | ,48 | ,49 | ,51 | ,79 | **,80** | |
| K | ,35 | ,49 | ,46 | ,48 | | | | | ,50 | ,51 | ,53 | | | **,89** |

Not: V1 = 1:a Verbala häftet; V2 = 2:a Verbala häftet; V = Verbala provdelen; K1 = 1:a Kvantitativa häftet; K2 = 2:a Kvantitativa häftet; K = Kvantitativa provdelen.

# Intern struktur

Begreppet intern struktur handlar om hur provets olika delar är relaterade till varandra. Provets interna struktur har undersökts genom att studera interkorrelationsmatrisen för delproven och provdelarna (Tabell 2:2 och 2:3), och genom konfirmatorisk faktoranalys (CFA; *confirmatory factor analysis*). Interkorrelationsmatriserna ovan visar några intressanta saker. Den korrigerade korrelationen mellan de två provhäftena inom respektive provdel är nära 1, vilket är en indikation på att de två häftena mäter samma sak. Det är även så att DTK har det högsta sambandet med den verbala provdelen, vilket är logiskt eftersom DTK är det kvantitativa prov som innehåller mest text och som ställer störst krav på läsförståelse.

När det gäller CFA-analyser visar dessa att den teoretiska modellens anpassning till empiriska provdata inte är perfekt. Figurer som visar modeller med latenta variabler/ faktorer (VERBAL_11B/VERBAL_12A respektive KVANT_11B/KVANT_12A) och delproven som indikatorer, samt parameterestimat och anpassningsindex finns i Bilaga A1. Enligt vissa index och tumregler (se t.ex. Schreiber et al., 2006) är anpassningen bra (SRMR = ,07 för både 11:B och 12:A; tumregeln är SRMR < ,08), enligt andra är den på gränsen till vad som anses som bra (CFI = ,94 respektive ,95; tumregeln är CFI > ,95) och enligt andra mindre bra (RMSEA = ,12 respektive ,11; en tumregel är RMSEA < ,08). Chi²-värdena visar på dålig modell–data-anpassning, men det är förväntat eftersom dessa värden är mycket känsliga för stickprovsstorleken och de data vi använder oss av har ett mycket stort antal provtagare. Vid en närmare titt på de så kallade modfikationsindexen så ser vi att om DTK tillåts vara en indikator även på den verbala faktorn så förbättras samtliga anpassningsindex så att både SRMR och CFI indikerar bra modell–data-anpassning (SRMR= ,06/,05; CFI = ,96/,97; RMSEA = ,10/,09). Detta bekräftar det som kunde utläsas från interkorrelationsmatriserna ovan, att DTK är en indikator på såväl en kvantitativ som en verbal variabel. Det ska dock understrykas att DTK-poängen har ett starkare samband med den kvantitativa faktorn än med den verbala, så det är rimligare att DTK ingår i den kvantitativa provdelen än i den verbala. Sammanfattningsvis, när det gäller modell–data-anpassningen sammantaget så kan vi dra slutsatsen att grundmodellen, den som ligger till grund för hur de olika provpoängen rapporteras och används, har möjligtvis en acceptabel anpassning till empiriska provdata.

# Mervärdet i delprovspoäng

CFA-analyserna visar som sagt att de två provdelarna inte är så distinkta och åtskilda som vi kan önska, men nackdelen är att dessa analyser inte tar hänsyn till reliabiliteten i provpoängen på olika nivåer. I bedömningen av på vilken nivå som provpoäng ska användas eller inte bör hänsyn tas till inte bara dimensionaliteten utan även reliabiliteten, eftersom beslut som baseras på mätningar med låg reliabilitet kommer automatiskt att ha en begränsad validitet i termer av samband med andra variabler. Det finns tidigare exempel på subjektiva kriterier för dimensionalitet, i termer av en högsta gräns för korrelationer mellan delprovet/provdelen av intresse och övriga sådana, och reliabilitet, i termer av en lägsta gräns för delprovs-/provdelspoängen (se t.ex. Lyrén, 2009), men för några år sedan presenterades en mer objektiv metod (Habmerman, 2008) där sådana subjektiva kriterier inte behöver fastställas i förväg. Metoden kallas allmänt ofta för Haberman's metod och utgår från en jämförelse mellan två kvantiteter: (1) reliabiliteten hos poängen som är av intresse, och (2) en kvantitet som kan ses som reliabiliteten hos poängen på en viss nivå som mått på det som poängen på den underliggande nivån är en indikator på. Som exempel kan vi då jämföra hur reliabla LÄS-poängen respektive Verbal-poängen är som indikatorer på läsförståelse. Är värdet högre för LÄS-poängen än för Verbal-poängen så finns det information i LÄS-poängen som är reliabel och distinkt från det som finns i Verbal-poängen, och därmed anses det finnas ett mervärde (*added value*) i LÄS-poängen utöver det som finns i Verbal-poängen.

I Bilaga A2 redogörs för en utförlig studie av mervärdet i poängen på delprovs- och provdelsnivå för proven 11:B och 12:A. Som kan ses i den sammanfattande tabellen nedan (vilket är en översatt och bearbetad version av Tabell 6 i Bilaga A2) verkar det inte finnas något mervärde i delprovspoängen utöver delprovspoängen, medan det är tydligt att det finns ett mervärde i provdelspoängen utöver totalpoängen. Detta innebär att poängen från de två provdelarna bär information som är distinkt och reliabel i jämförelse med den information som finns i totalpoängen.

Tabell 2:4. Mervärdet i provpoäng på delprovs- och provdelsnivå. Ett + indikerar mervärde medan ett – indikerar inget mervärde. Ett +/– eller ett –/+ indikerar olika resultat för 11:B och 12:A.

| Metod | Delprov | | | | | | | | Provdel | |
|---|---|---|---|---|---|---|---|---|---|---|
| | XYZ | KVA | NOG | DTK | ORD | LÄS | MEK | ELF | K | V |
| Habermans metod | +/– | – | – | + | –/+ | – | – | – | + | + |
| Faktoranalys | – | – | – | – | – | – | – | – | + | + |
| DIMTEST | + | – | + | + | + | + | – | + | + | + |
| DETECT | – | – | – | – | – | – | – | – | + | + |

# Sammanfattning av studierna om reliabilitet, intern struktur och mervärdet i delprovspoäng

Reliabiliteten för totalpoängen är ungefär som för det gamla högskoleprovet (cirka 0,93), och detsamma gäller för den verbala provdelen (0,90–0,91). Reliabiliteten på den kvantitativa provdelen är högre på det nya provet än vad den var på det gamla och når värden kring 0,90, vilket som tidigare sagts kan ses som ett idealvärde. Analysen av provets interna struktur visar att den underliggande teoretiska modellen där ORD, LÄS, MEK och ELF är indikatorer på en verbal faktor och där XYZ, KVA, NOG och DTK är indikatorer på en kvantitativ faktor inte har en perfekt anpassning till faktiska provdata, utan denna anpassning är möjligtvis acceptabel. Studien av mervärdet i delprovspoängen ger dock stöd för att rapportera en kvantitativ och en verbal poäng separat. Sammantaget ger studierna i detta kapitel stöd för att den nya provsammansättningen med separata verbala och kvantitativa delar fungerar, och att en flexibel användning av de två provdelspoängen torde kunna bringa ett mervärde till urvalsprocessen.

# 3. Icke-svar och positionseffekter

## Icke-svar

Med ett icke-svar menas den situation där provtagaren på en uppgift inte markerar något svar i svarshäftet. Att undersöka andelen icke-svar på uppgiftsnivå är relevant eftersom många provtagare som genomförde de första utprövningarna av det nya provformatet upplevde att framförallt den kvantitativa provdelen var tidspressad. En hög andel icke-svar skulle vara en tydlig indikation på tidspress, men däremot är det inte säkert att det omvända gäller. En relativt låg andel icke-svar skulle kunna vara en indikation på liten tidspress, men en markering i svarshäftet behöver inte innebära att provtagaren faktiskt hunnit gå igenom uppgiften och avlämnat ett väl grundat svar utan den kan lika gärna komma från en helt slumpmässig gissning. Att undersöka detta kräver dock vidare analyser som inte prioriterats inom ramen för det aktuella uppdraget.

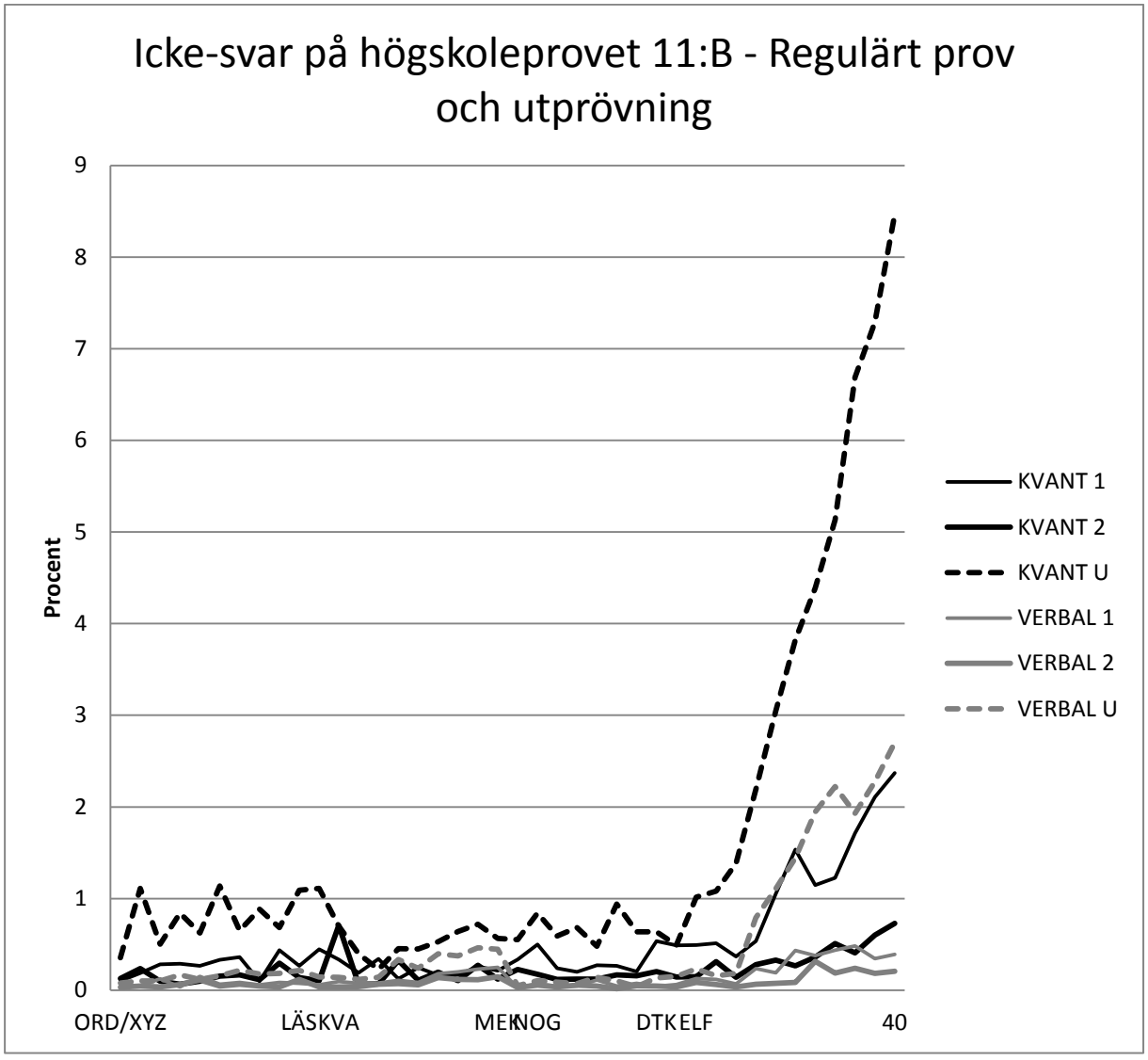
Förutom att andelen provtagare som lämnat icke-svar kan vara en indikator på tidspress på provet så skulle andelen uppgifter som en viss provtagare lämnat icke-svar på kunna vara en indikator på en provtagares snabbhet. Ett syfte med den nya provsammansättningen var som tidigare nämnts att göra provet mer effektivt i termer av antal provuppgifter per tidsenhet, och när den frågan diskuterades i högskoleprovets internationella vetenskapliga råd nämndes just risken att provet skulle bli för tidspressat. Detta sågs dock inte som något större problem med argumentet att även om provet blir mer tidspressat så borde förmågan att kunna lösa uppgifter snabbt vara relevant även vid urval till högre studier. Utifrån nämnda antaganden och detta argument blir det därför relevant att undersöka om det finns något samband mellan snabbhet och prestation i form av provpoäng. Om antagandena och argumentet är giltiga så förväntar vi oss ett positivt samband mellan snabbhet och prestation, eller mer specifikt att det finns ett negativt samband mellan andelen icke-svar och provpoäng.

### Andel icke-svar

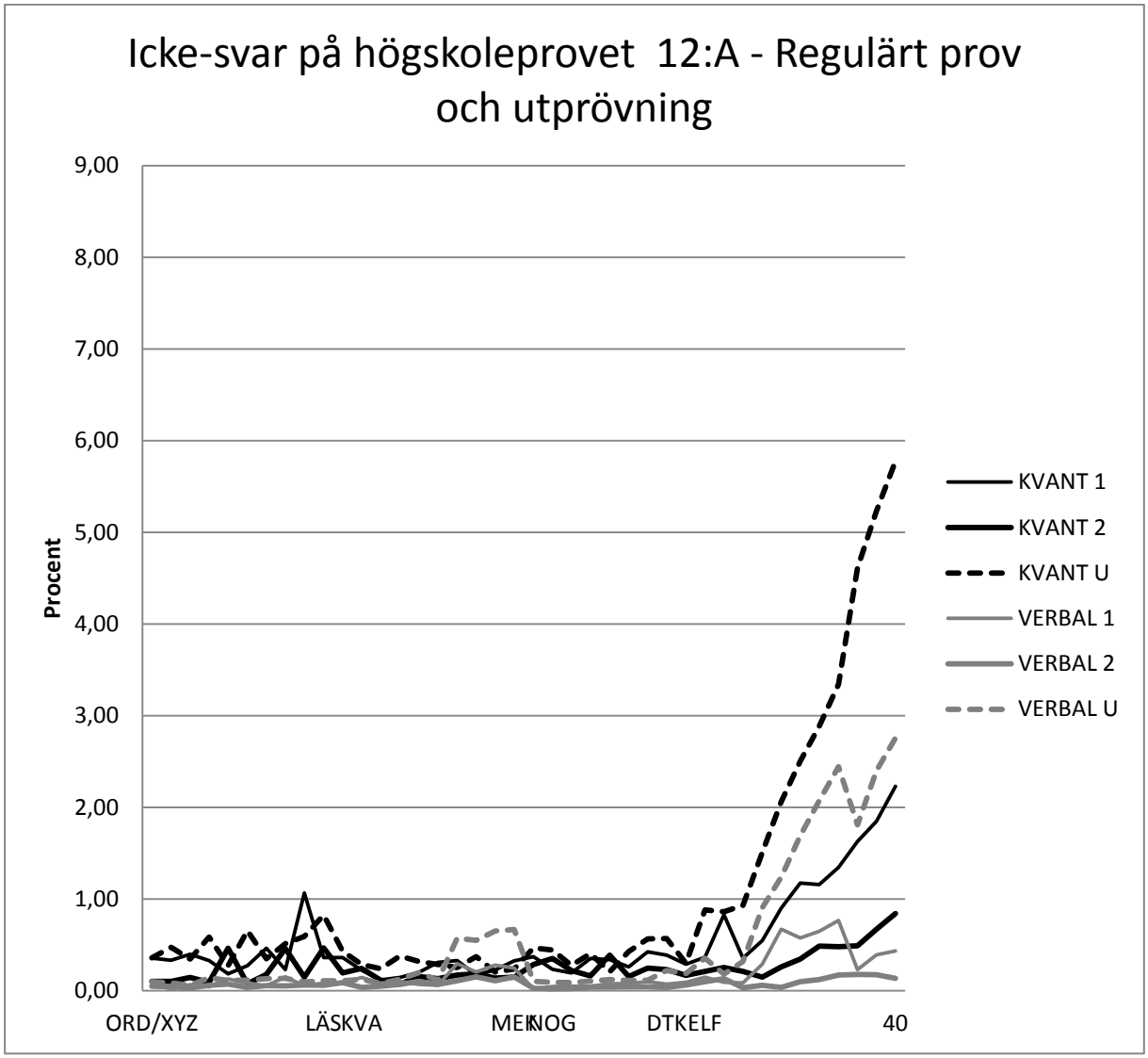
I Figur 3:1 och 3:2 redovisas för proven 11:B respektive 12:A andelen icke-svar på de reguljära provhäftena och utprövningshäftena. För tolkning av nedanstående figurer är det väsentligt att veta ordningen på provhäftena under provdagen, vilken var densamma för båda provtillfällena: Utprövning (Kvant/Verbal) – Kvant 1 – Verbal 1 – Kvant 2 – Verbal 2.

Andelen icke-svar stiger avsevärt mot slutet av häftena, speciellt de kvantitativa, men överlag är andelen icke-svar inte anmärkningsvärda utan på ungefär samma nivåer som på det gamla högskoleprovet. Som kan ses i figurerna så minskar den andel icke-svar som noteras i slutet av häftena allteftersom provdagen förlöper. Exempelvis är den genomsnittliga andelen icke-svar på den sista uppgiften i de kvantitativa utprövningshäftena drygt 8 procent, medan den är drygt 2 procent på det första reguljära kvantitativa häftet och knappt 1 procent på det andra reguljära kvantitativa häftet. Att andelen minskar så pass mycket torde inte bero på skillnad i häftenas svårighetsgrad eftersom denna skillnad är relativt liten och kontrolleras, åtminstone på de reguljära häftena, genom utprövningsvärdena. Då är det mer sannolikt att minskningen beror på att provtagarna blir mer bekant med provformatet och att de lär sig tidshanteringsstrategier, som att de inte ska dröja kvar vid vissa frågor utan gå vidare så att de hinner gå igenom hela häftet.

I Figur B:1 till B:8 i Bilaga B redovisas andelen icke-svar på de reguljära provdelarna efter vilken provdel som utgjorde utprövningshäftet, och det vi kan se i dessa figurer är dels att andelen icke-svar minskar från det första till det andra reguljära häftet i respektive provdel, och dels att utprövningshäftets innehåll verkar ha större inverkan på andelen icke-svar på de kvantitativa häftena än på de verbala häftena.




Figur 3:1. Andel icke-svar på de reguljära häftena och utprövningshäftena från högskoleprovet hösten 2011. Andelarna för utprövningshäftena är beräknat på samtliga kvantitativa respektive verbala utprövningshäften.




Figur 3:2. Andel icke-svar på de reguljära häftena och utprövningshäftena från högskoleprovet våren 2012. Andelarna för utprövningshäftena är beräknat på samtliga kvantitativa respektive verbala utprövningshäften.

## Samband mellan icke-svar och provprestation

Tabell 3:1 visar korrelationer mellan antalet icke-svar och provprestation. Antalet rätta svar används inte i detta fall som prestationsvariabel eftersom antalet rätta svar inte tar hänsyn till att felsvar kan bero på såväl felmarkerade svar som på just icke-svar. I stället definieras prestationsvariabeln som kvoten antalet rätta svar och antalet giltiga svar. Som vi ser i tabellen så är korrelationerna mellan icke-svar och prestation på den verbala provdelen signifikanta på 0,01-nivån (vilket inte är överraskande eftersom antalet provtagare är så stort), men de är samtidigt mycket svaga. På den kvantitativa provdelen är korrelationerna ännu svagare, och till och med obefintligt för provet 12:A.

Tabell 3:1. Korrelationer mellan antalet icke-svar och provprestation på respektive provdel. Värden för 11:B är under diagonalen och för 12:A ovanför diagonalen.

| Variabel | Icke-svar Verb | Icke-svar Kvant | p Verb | p Kvant |
|---|---|---|---|---|
| Icke-svar Verb | | 0,38** | -0,04** | -0,05** |
| Icke-svar Kvant | 0,44** | | 0,04** | -0,00 |
| p Verb | -0,04** | 0,02** | | 0,53** |
| p Kvant | -0,04** | -0,01* | 0,50** | |

Not: p = (antalet rätta svar)/(antalet giltiga svar)
**. Korrelationen är signifikant på 0,01-nivån.
*. Korrelationen är signifikant på 0,05-nivån.

I Tabell 3:2 ser vi motsvarande korrelationer när det gäller icke-svar och prestation utifrån totalpoängen på det gamla provet. I det här fallet är sambandet något starkare än på det nya provet, men det är ändå ett svagt samband. En möjlig orsak till skillnaderna skulle kunna vara att det gamla provet var mindre tidspressat och att högpresterande provtagare i mindre utsträckning än lågpresterande lämnade icke-svar jämfört med hur det ser ut på det nya provet.

Tabell 3:2. Korrelationer mellan antalet icke-svar och provprestation på det gamla högskoleprovet. Värden för 10:B under diagonalen och för 11:A ovanför diagonalen.

| Variabel | Icke-svar | p Total |
|---|---|---|
| Icke-svar | | -0,09** |
| p Total | -0,07** | |

Not: p = (antalet rätta svar)/(antalet giltiga svar)
**. Korrelationen är signifikant på 0,01-nivån.

# Positionseffekter

I den utprövning som gick 12:A sattes två häften samman för att undersöka effekterna av vilken position en figur och dess uppgifter har i ett häfte på uppgifternas svårighetsgrad i termer av lösningsproportion ($p$). Samma fyra figurer placerades i de två häftena (H1 och H2) men presenterades i omvänd ordning i förhållande till varandra. Den inbördes placeringen av uppgifter till en viss figur var densamma i båda häftena för att eventuella positionseffekter inom varje figur skulle vara konstant. Som kan ses i Tabell 3:3 så var effekten på lösningsproportionen av att ändra position från först till sist i häftet i genomsnitt −0,10 per uppgift om man tittar på den korrigerade skillnaden ($\Delta p$). Motsvarande effekt erhölls av att ändra position från sist till först, +0,10 i genomsnittlig skillnad för de tre uppgifterna.

Tabell 3:3. Skillnader i lösningsproportion (Δ*p*) för DTK-uppgifter med olika position i två kvantitativa utprövningshäften 12:A. Skillnaden i lösningsproportion redovisas också korrigerad utifrån de två utprövningsgruppernas olika medellösningsproportion på de tolv uppgifterna (Korr. Δ*p*).

| Figurtyp | Uppgift | Position | | | Δ*p* | Korr. Δ*p* |
|---|---|---|---|---|---|---|
| | | H1 | H2 | Δpos. | | |
| Tabell | A | 29 | 38 | +9 | −0,09 | −0,12 |
| | B | 30 | 39 | +9 | −0,08 | −0,11 |
| | C | 31 | 40 | +9 | −0,04 | −0,07 |
| Karta | D | 32 | 35 | +3 | −0,06 | −0,09 |
| | E | 33 | 36 | +3 | −0,05 | −0,08 |
| | F | 34 | 37 | +3 | −0,03 | −0,06 |
| Cirkel | G | 35 | 32 | −3 | ±0,00 | −0,03 |
| | H | 36 | 33 | −3 | −0,02 | −0,05 |
| | I | 37 | 34 | −3 | +0,02 | −0,01 |
| Stapel | J | 38 | 29 | −9 | +0,14 | +0,11 |
| | K | 39 | 30 | −9 | +0,10 | +0,07 |
| | L | 40 | 31 | −9 | +0,15 | +0,12 |

I samband med utprövningen 12:B sattes två häften ihop för att ytterligare undersöka positionseffekter. Vid denna undersökning var fokus mer på positionseffekter på figurnivå, utifrån att antalet uppgifter per figurer i DTK kan variera mellan två och fyra. I det ena häftet (H3) placerades figuren med fyra uppgifter först och figuren med två uppgifter sist. I det andra häftet (H4) placerades figurerna tvärtom. De två mittenfigurerna behöll samma placering i de två häftena. Som kan ses i Tabell 3:4 så var positionseffekterna generellt mindre i detta fall: −0,04 i genomsnitt för figuren med de fyra uppgifter som låg först i H3 och sist i H4, och +0,06 i genomsnitt för figuren med de två uppgifter som låg sist i H3 och först i H4.

Att effekten är något större för figuren med två uppgifter än för figuren med fyra uppgifter skulle kunna bero på att provtagaren tagit sig igenom fler uppgifter (92 % av DTK-uppgifterna) när figuren med två uppgifter ligger sist än vad som gäller för det omvända dvs. när den sista figuren har fyra uppgifter (då de gjort 83 % av DTK-uppgifterna). Detta kan bidra till att provtagarna upplever en större tidsbrist när de kommer fram till den sista figuren med två uppgifter, än vad de upplever när de kommer till en sista figur med fyra uppgifter.

Att positionseffekten är något mindre vid undersökningen 12:B än den vid 12:A beror mest troligt på utprövningspassets placering under provdagen. Vid provet 12:A låg utprövningspasset som tidigare nämnts först, men vid provet 12:B låg detta pass som nummer tre (Kvant 1 – Verb 1 – Utprövning – Kvant 2 – Verb 2). På samma sätt som andelen icke-svar är lägre för häftena som ligger senare under provdagen beror även denna skillnad i positionseffekt troligen på att provtagarna blir mer bekant med provformatet och att de lär sig tidshanteringsstrategier allteftersom under provdagen.

Tabell 3:4. Skillnader i lösningsproportion (Δ*p*) för DTK-uppgifter med olika position i två kvantitativa utprövningshäften 12:B. Skillnaden i lösningsproportion redovisas också korrigerad utifrån de två utprövningsgruppernas olika medellösningsproportion på de tolv uppgifterna (Korr. Δ*p*).

| Figurtyp | Uppgift | Position | | | Δ*p* | Korr. Δ*p* |
|---|---|---|---|---|---|---|
| | | H3 | H4 | Δpos. | | |
| Figur 1 | A | 29 | 37 | +8 | −0,01 | −0,03 |
| | B | 30 | 38 | +8 | −0,03 | −0,05 |
| | C | 31 | 39 | +8 | −0,03 | −0,05 |
| | D | 32 | 40 | +8 | −0,01 | −0,03 |
| Figur 2 | E | 33 | 31 | −2 | +0,01 | −0,01 |
| | F | 34 | 32 | −2 | +0,04 | +0,02 |
| | G | 35 | 33 | −2 | ±0,00 | −0,02 |
| Figur 3 | H | 36 | 34 | −2 | +0,04 | +0,02 |
| | I | 37 | 35 | −2 | +0,04 | +0,02 |
| | J | 38 | 36 | −2 | −0,02 | −0,04 |
| Figur 4 | K | 39 | 29 | −10 | +0,09 | +0,07 |
| | L | 40 | 30 | −10 | +0,07 | +0,05 |

Det bör nämnas i sammanhanget att studierna om positionseffekter skulle behöva upprepas för att det ska gå att dra ännu säkrare slutsatser om dessa effekter. Exempelvis har det i dessa studier inte kontrollerats för innehållet i de andra delproven (XYZ, KVA och NOG) i häftena.

# Sammanfattning av studierna om icke-svar och positionseffekter

Sammantaget ger studierna om icke-svar och positionseffekter indikationer om att framförallt den kvantitativa provdelen är tidspressad, och att förmågan att hantera denna tidspress inte har något samband med provprestation. Men, resultaten indikerar samtidigt att graden av tidspress minskar under provdagen vilket skulle innebära att den faktiska tidspressen, och därmed även den upplevda, kan lindras genom grundliga förberedelser från provtagarnas sida. Är man som provtagare väl införstådd med provets format och har klara tidshanteringsstrategier redan innan provdagen så borde vi inte se dessa mönster i icke-svar och positionseffekter.

# 4. Provresultat för olika grupper

I detta kapitel redovisas provdeltagarnas resultat på de reguljära proven hösten 2011 och våren 2012. Resultaten redovisas för män och kvinnor, för olika åldersgrupper, för provdeltagare indelade efter föräldrars utbildning, för provdeltagare som genomgått olika gymnasieprogram samt för provdeltagare med olika ursprung (svenskt/utländskt). Den fullständiga rapporten som merparten av detta kapitel är baserat på finns i Bilaga C.

I Tabell 4:1 (Tabell 2 i Bilaga C) redovisas provdeltagargruppens sammansättning utifrån kön, ålder och utbildningsnivå. Noterbart är att fördelningen män/kvinnor är helt jämn 12:A, vilket är det enda provtillfället den har varit det.

Tabell 4:1. Provdeltagarnas procentuella fördelning på kön, ålder och utbildning.

| Provtillfälle | 11:B | 12:A |
|---|---|---|
| Kön | | |
| Män | 49 | 50 |
| Kvinnor | 51 | 50 |
| Ålder | | |
| ≤ 20 år | 58 | 56 |
| 21–24 år | 25 | 28 |
| 25–29 år | 9 | 9 |
| 30–39 år | 6 | 5 |
| ≥ 40 år | 2 | 2 |
| Utbildning | | |
| Grund/realskola | 1 | 1 |
| Folkhögskola | 2 | 2 |
| Gymnasium ≤ 2 år | 3 | 3 |
| Gymnasium ≥ 3 år | 77 | 79 |
| Högre utbildning ≤ 80 p | 12 | 10 |
| Högre utbildning ≥ 80 p | 4 | 4 |
| Ej angivet | 1 | 1 |
| Totalt antal provtagare | 40 695 | 56 864 |

Nedan följer en redovisning av poängmedelvärde (m) och standardavvikelse (s) för de olika grupper som genomförde proven 11:B, dvs. hösten 2011, respektive 12:A, dvs. våren 2012. Medelvärdena på den verbala delen 12:A är genomgående lägre än vid provet 11:B, och anledningen är att fyra ORD-uppgifter i den verbala delen 12:A blev strukna på grund av att dessa hade lagts ut på ett publikt forum på internet. Istället för 20 ORD-uppgifter innehåller 12:A därför 16 ORD-uppgifter och därmed totalt 76 verbala uppgifter. Av utrymmesskäl redovisas inte antalet provtagare i respektive grupp i nedanstående tabeller, utan för denna information hänvisar vi till tabellerna i Bilaga C.

# Gruppskillnader med avseende på kön

Tabell 4:2 (se Tabell 3 och 14 i Bilaga C) visar resultat på de två provdelarna proven 11:B och 12:A. På den kvantitativa delen 11:B var medelvärdet 37,9 och spridningen var 13,4 för totalgruppen. Skillnaden mellan män och kvinnor var 7,1 till männens fördel, motsvarande 0,53 standardavvikelser. Spridningen var 1,7 högre för männen än för kvinnorna. Vid provet 12:A var medelvärdet 37,0 och spridningen 12,1. Skillnaden mellan män och kvinnor var 6,7 till männens fördel, vilket motsvarar 0,55 standardavvikelser. Spridningen var 1,6 högre för männen än för kvinnorna, och det är konsekvent så över alla delprov (se Tabell 4 i Bilaga C). På delprovsnivå varierar könsskillnaderna i termer av standardavvikelseskillnader (*d*) inte särskilt mycket, utan de är relativt konsekventa över delproven.

Vid provet 11:B var medelvärdet på den verbala delen för totalgruppen 41,6 och spridningen var 13,5. Skillnaden i medelpoäng mellan män och kvinnor var 2,3 till männens fördel, motsvarande 0,17 standardavvikelser. Spridningen var densamma för männen som för kvinnorna. Vid provet 12:A var medelvärdet 39,7 (baserat på 76 uppgifter) och spridningen var 12,5. Skillnaden mellan män och kvinnor var här 2,5 till männens fördel, motsvarande 0,20 standardavvikelser. Spridningen var något högre (0,3) för männen än för kvinnorna. På delprovsnivå står ELF för mer än hälften av könsskillnaden på den verbala delen (se Tabell 15 i Bilaga C); standardavvikelseskillnaden är 0,35 för ELF vid båda proven medan det är som högst 0,17 för något av de andra delproven (MEK 12:A). Delprovet ELF är även det verbala delprov som har den högsta lösningsproportionen per uppgift, 0,58, medan delprovet ORD har den lägsta lösningsproportionen per uppgift, 0,48.

Tabell 4:2. Resultat för kvinnor och män samt totalt på respektive provdel 11:B och 12:A.

| Kön | Kvantitativ provdel | | | | Verbal provdel | | | |
|---|---|---|---|---|---|---|---|---|
| | 11:B | | 12:A | | 11:B | | 12:A | |
| | m | s | m | s | m | S | m | s |
| Kvinnor | 34,4 | 12,1 | 33,7 | 10,8 | 40,5 | 13,5 | 38,5 | 12,3 |
| Män | 41,5 | 13,8 | 40,4 | 12,4 | 42,8 | 13,5 | 41,0 | 12,6 |
| Totalt | 37,9 | 13,4 | 37,0 | 12,1 | 41,6 | 13,5 | 39,7 | 12,5 |
| *d* | 0,53 | | 0,55 | | 0,17 | | 0,20 | |

Not*: d* = [(m för män)−(m för kvinnor)] / (s för totalgruppen)

Om vi jämför könsskillnaderna på det nya provet med hur de såg ut på det gamla provet (se *d*-värdena i Tabell 4:3) så framkommer att skillnaderna mellan män och kvinnor är mycket lika. På det gamla provet 10:B och 11:A var skillnaden i termer av *d* (vilket kan ses som ett mått på skillnaden i rangordning mellan de två grupperna) 0,51 respektive 0,52, och på det nya provet 11:B och 12:A var motsvarande värden, som tidigare nämnts, 0,53 respektive 0,55. Alltså är det ingen större skillnad mellan det nya och det gamla provet när det gäller hur män respektive kvinnor rangordnas på de två provdelarna.

Om vi gör motsvarande jämförelse men på totalpoängsnivå (Tabell 4:4) så ser vi att skillnaderna mellan det gamla och det nya provet är större. Orsaken till detta är en kombination av att könsskillnaderna är större på den kvantitativa delen än på den verbala och att den kvantitativa delen har relativt större vikt i det nya provet än den hade i det gamla.

Tabell 4:3. Resultat för kvinnor och män samt totalt på respektive provdel 10:B och 11:A.

| Kön | Kvantitativ provdel (NOG+DTK) | | | | Verbal provdel (ORD+LÄS+ELF) | | | |
|---|---|---|---|---|---|---|---|---|
| | 10:B | | 11:A | | 10:B | | 11:A | |
| | m | s | m | s | m | S | m | s |
| Kvinnor | 20,1 | 7,1 | 18,7 | 6,6 | 42,6 | 13,5 | 41,8 | 13,4 |
| Män | 24,0 | 7,7 | 22,4 | 7,1 | 45,0 | 13,5 | 44,1 | 13,2 |
| Totalt | 21,9 | 7,7 | 20,4 | 7,1 | 43,7 | 13,6 | 42,9 | 13,4 |
| *d* | 0,51 | | 0,52 | | 0,18 | | 0,17 | |

Not*: d* = [(m för män)−(m för kvinnor)] / (s för totalgruppen)

Tabell 4:4. Normerad poäng för kvinnor och män samt totalt på gamla och nya provet.

| Kön | Gamla provet | | | | Nya provet | | | |
|---|---|---|---|---|---|---|---|---|
| | 10:B | | 11:A | | 11:B | | 12:A | |
| | m | s | m | s | m | S | m | s |
| Kvinnor | 0,82 | 0,47 | 0,79 | 0,45 | 0,87 | 0,39 | 0,86 | 0,36 |
| Män | 0,98 | 0,47 | 0,94 | 0,45 | 1,03 | 0,39 | 1,02 | 0,38 |
| Totalt | 0,89 | 0,47 | 0,86 | 0,46 | 0,95 | 0,40 | 0,94 | 0,38 |
| *d* | 0,34 | | 0,33 | | 0,40 | | 0,42 | |

Not*: d* = [(m för män)−(m för kvinnor)] / (s för totalgruppen)

# Gruppskillnader med avseende på ålder

I Tabell 4:5 nedan (Tabell 5 och 16 i Bilaga C) redovisas resultat för olika åldersgrupper. Eftersom ungefär 85 procent av provdeltagargruppen är under 25 år redovisas i Tabell 6 och 17 i Bilaga C även resultat för varje ålder från –17 år till och med 24 år.

Det är den yngsta åldersgruppen som har det högsta resultatet på den kvantitativa delen, följt av den äldsta gruppen, dvs. 40 år eller äldre, som har det näst högsta resultatet. På delprovsnivå (se Tabell 7 och 18 i Bilaga C) ser vi att ju yngre provtagare desto högre poäng på delproven XYZ, KVA och NOG, medan skillnaden mellan olika åldersgrupper är försumbar på delprovet DTK. Den matematik som ligger till grund för de kvantitativa delproven är kurs A/1b i gymnasieskolan. En rimlig förklaring till att de yngre åldersgrupperna har högst resultat på de kvantitativa delproven torde vara att de är under utbildning eller nyss avslutat en utbildning och har därmed matematiken nära i tiden.

Poängen på den verbala delen varierar mycket med ålder, på så sätt att ju högre ålder desto högre poäng på den verbala delen. Det skiljer ungefär 15 poäng mellan den yngsta och äldsta gruppen, motsvarande drygt 1 standardavvikelse. Spridningen är störst i åldersgruppen 30–39 år och minst i gruppen som är 20 år och yngre. Även på delprovsnivå varierar poängen mellan olika åldersgrupper, där det genomgående mönstret är ju äldre desto högre poäng. Den största skillnaden mellan åldersgrupper finns på ORD med ett *d* på cirka 1,7, följt av MEK där *d* är cirka 1,3. Delproven LÄS och ELF uppvisar betydligt mindre skillnader med *d*-värden på 0,5 respektive 0,3.

Tabell 4:5. Resultat för olika åldersgrupper på respektive provdel 11:B och 12:A.

| Ålder | Kvantitativ provdel | | | | Verbal provdel | | | |
|---|---|---|---|---|---|---|---|---|
| | 11:B | | 12:A | | 11:B | | 12:A | |
| | m | s | m | s | m | S | m | s |
| –20 år | 39,7 | 13,5 | 38,3 | 12,3 | 39,2 | 12,4 | 37,7 | 11,5 |
| 21–24 år | 35,9 | 12,8 | 35,5 | 11,5 | 42,4 | 13,2 | 39,7 | 12,2 |
| 25–29 år | 34,3 | 12,4 | 34,9 | 11,5 | 46,5 | 14,3 | 44,1 | 13,2 |
| 30–39 år | 34,8 | 13,8 | 35,1 | 12,3 | 50,9 | 15,3 | 48,3 | 14,2 |
| 40– år | 36,1 | 15,0 | 35,9 | 13,4 | 55,2 | 13,8 | 52,0 | 13,7 |
| Totalt | 37,9 | 13,4 | 37,0 | 12,1 | 41,6 | 13,5 | 39,7 | 12,5 |

# Gruppskillnader med avseende på föräldrarnas utbildningsnivå

I Tabell 4:6 nedan (Tabell 8 och 19 i Bilaga C) redovisas resultat för olika grupper baserat på föräldrarnas utbildningsnivå. Det föreligger tämligen stora skillnader mellan de redovisade grupperna på den kvantitativa delen, där provtagare som har föräldrar med en hög utbildningsnivå har högre provpoäng. Provtagare med föräldrar som har eftergymnasial utbildning är de som har högst resultat på provdelen. Skillnaden i poäng mellan den grupp provtagare vars föräldrar har 2-årig gymnasial utbildning och den grupp vars föräldrar har eftergymnasial utbildning 5 år eller längre var 13,4 (*d* = 1,0) respektive 11,4 (*d* = 0,9) för de två provomgångarna. På delprovsnivå (se Tabell 9 och 20 i Bilaga C) finns den största skillnaden på XYZ (*d* = 0,9) och den minsta skillnaden på DTK (*d* = 0,6-0,7).

Det föreligger stora skillnader mellan de redovisade grupperna även på den verbala delen. Skillnaden i resultat mellan den grupp vars föräldrar har 2-årig gymnasial utbildning och den grupp vars föräldrar har minst 5 år eftergymnasial utbildning var 10,5 poäng (*d* = 0,8) på provet 11:B och 9,1 (*d* = 0,7) på provet 12:A. På delprovsnivå finns den största skillnaden på ELF (*d* = 0,8) och den minsta skillnaden på ORD (*d* = 0,5). På samtliga delprov är det gruppen vars föräldrar har en 2-årig gymnasial utbildning som har det lägsta poängmedelvärdet, förutom på ORD där det är gruppen vars föräldrar har en 3-årig gymnasial utbildning. Detta beror på att provtagare vars föräldrar har en 2-årig gymnasial utbildning är överrepresenterade bland äldre provtagare och att äldre provtagare har betydligt högre poäng på ORD än vad de yngre provtagarna har.

Tabell 4:6. Resultat för olika provtagargrupper indelade efter föräldrarnas utbildningsnivå på respektive provdel 11:B och 12:A.

| Utbildnings-nivå[1] | Kvantitativ provdel | | | | Verbal provdel | | | |
|---|---|---|---|---|---|---|---|---|
| | 11:B | | 12:A | | 11:B | | 12:A | |
| | m | s | m | s | m | S | m | s |
| Gy 2 år | 32,3 | 11,8 | 32,4 | 10,6 | 38,1 | 13,2 | 36,8 | 12,3 |
| Gy 3– år | 35,5 | 12,5 | 34,7 | 11,4 | 39,1 | 12,9 | 37,4 | 12,0 |
| E-Gy –2 år | 37,7 | 12,9 | 36,8 | 11,6 | 41,3 | 13,0 | 39,4 | 12,1 |
| E-Gy 3–4 år | 40,1 | 13,4 | 39,9 | 12,2 | 43,9 | 13,3 | 41,9 | 12,3 |
| E-Gy 5– år | 45,7 | 13,3 | 43,8 | 12,2 | 48,6 | 13,1 | 45,8 | 12,0 |
| Totalt | 37,9 | 13,4 | 37,0 | 12,1 | 41,7 | 13,5 | 39,8 | 12,5 |

[1] Gy = Gymnasial utbildning; E-Gy = eftergymnasial utbildning.

# Gruppskillnader med avseende på studieinriktning i gymnasiet

I Tabell 4:7 nedan (Tabell 10 och 21 i Bilaga C) redovisas resultat för grupper av provtagare med olika studieinriktning i gymnasiet.

De provdeltagare som har gått en naturvetenskapligt inriktad utbildning har högst poängmedelvärde på den kvantitativa provdelen medan de som gått en utbildning med inriktning mot barn/fritid, handel/administration, hotell/restaurang, livsmedel har lägst poängmedelvärde; skillnaden mellan grupperna är 19,6 respektive 16,4 poäng, eller sett i termer av *d* 1,5 respektive 1,4 för de två proven. Skillnaden uttryckt i *d* mellan de två till antalet största grupperna, de med samhällsvetenskaplig respektive naturvetenskaplig inriktning, är cirka 1, 0 respektive 0,9. Skillnaden i poäng mellan provdeltagare från olika inriktningar är mindre på NOG och DTK än på XYZ och KVA (se Tabell 11 och 22 i Bilaga C).

Även på den verbala provdelen har de som gått en naturvetenskapligt inriktad utbildning högst poängmedelvärde och de som gått en utbildning med inriktning mot barn/fritid, handel/administration, hotell/restaurang, livsmedel har lägst poängmedelvärde; skillnaden mellan grupperna är 12,7 respektive 12,5 poäng, eller sett i termer av *d* 1,0 på respektive provversion. Skillnaden uttryckt i *d* mellan provtagarna med samhällsvetenskaplig respektive naturvetenskaplig inriktning är 0,4 på både 11:B och 12:A. Skillnaden på delprovsnivå i räknat i standardavvikelser är störst på ELF (*d* = 1,0), något mindre på MEK (*d* = 0,9) och LÄS (*d* = 0,8) och minst på ORD (*d* = 0,6).

Tabell 4:7. Resultat för provtagare med olika studieinriktning på gymnasiet för respektive provdel 11:B och 12:A.

| Inriktning[1] | Kvantitativ provdel | | | | Verbal provdel | | | |
|---|---|---|---|---|---|---|---|---|
| | 11:B | | 12:A | | 11:B | | 12:A | |
| | m | s | m | s | m | S | m | S |
| Samhälls | 33,6 | 11,0 | 33,7 | 10,0 | 41,7 | 12,8 | 40,7 | 12,1 |
| Natur | 46,2 | 12,8 | 44,1 | 11,9 | 47,0 | 13,1 | 45,5 | 12,0 |
| Teknisk | 41,3 | 11,7 | 40,2 | 10,5 | 39,8 | 11,9 | 39,6 | 11,0 |
| Estetisk | 31,8 | 10,5 | 32,0 | 9,8 | 42,6 | 13,3 | 40,8 | 12,6 |
| Hemk. | 26,7 | 8,3 | 27,7 | 8,1 | 34,3 | 11,6 | 33,0 | 10,7 |
| Yrkes | 33,9 | 11,9 | 34,6 | 10,6 | 38,7 | 12,2 | 37,9 | 11,5 |
| Totalt | 37,2 | 13,2 | 36,3 | 11,7 | 42,4 | 13,3 | 41,0 | 12,4 |

[1] Samhälls = samhällsvetenskap, ekonomi; Natur = naturvetenskap, naturbruk, International Baccalaureate; Teknisk = teknikprogrammet; Estetisk = estetisk, hantverk, media; Hemk. = barn och fritid, handel och administration, hotell och restaurang, livsmedel; Yrkes = bygg, el, energi, fordon, industri.

# Gruppskillnader med avseende på ursprung

I Tabell 4:8 nedan (Tabell 12 och 23 i Bilaga C) redovisas resultat för provtagare med olika ursprung.

Ungefär 75 procent av provdeltagarna har ett ursprung som kan benämnas ”helsvenskt”, dvs. de är födda i Sverige och båda föräldrarna är också svenskfödda. Det är den gruppen och gruppen födda i Sverige med en svenskfödd och en utlandsfödd förälder samt den lilla gruppen provdeltagare (0,5 procent) födda utomlands med minst en förälder född i Sverige som har högst resultat på den kvantitativa delen. Den sistnämnda gruppen är dock för liten för att vi ska dra några långtgående slutsatser om denna. Skillnaden i *d* mellan provtagare med ”helutländskt” ursprung (födda utomlands

med båda föräldrarna utlandsfödda) och provtagarna med helsvenskt ursprung är 0,4 respektive 0,5 för de två provomgångarna. Skillnaden mellan grupperna på delprovsnivå (se Tabell 13 i Bilaga C) är klart minst på XYZ (*d* = 0,1–0,2), medan den är störst på DTK (d = 0,6–0,7).

Provtagarna med helsvenskt ursprung har högst poäng på den verbala provdelen, medan det på samma sätt som för den kvantitativa provdelen är provtagarna med helutländskt ursprung som har lägst poäng; skillnaden är 0,7 standardavvikelser mellan dessa grupper. På delprovsnivå (se Tabell 24 i Bilaga C) är skillnaden mellan dessa två grupper som störst på MEK, men skillnaderna mellan delproven är avsevärt mindre än vad de är på den kvantitativa delen.

Tabell 4:8. Resultat för olika provtagargrupper indelade efter ursprung för respektive provdel 11:B och 12:A.

| Ursprung[1] | Kvantitativ provdel | | | | Verbal provdel | | | |
|---|---|---|---|---|---|---|---|---|
| | 11:B | | 12:A | | 11:B | | 12:A | |
| | m | s | m | s | m | S | m | s |
| U-UU | 33,2 | 12,7 | 32,3 | 11,5 | 33,5 | 13,3 | 32,1 | 12,0 |
| S-UU | 35,1 | 12,8 | 33,6 | 11,3 | 34,8 | 12,4 | 33,5 | 11,3 |
| U-S? | 40,0 | 13,3 | 39,8 | 12,8 | 42,0 | 12,9 | 40,9 | 12,1 |
| S-SU | 37,5 | 13,2 | 36,9 | 12,0 | 42,5 | 13,4 | 40,7 | 12,3 |
| S-SS | 38,7 | 13,5 | 37,9 | 12,1 | 43,0 | 13,2 | 41,1 | 12,2 |
| Totalt | 37,9 | 13,4 | 37,0 | 12,1 | 41,6 | 13,5 | 39,7 | 12,5 |

[1] U-UU = utlandsfödd med båda föräldrarna utlandsfödda; S-UU = svenskfödd med båda föräldrarna utlandsfödda; U-S? = utlandsfödd med minst en svenskfödd förälder; S-SU = svenskfödd med en svenskfödd och en utlandsfödd förälder; S-SS = svenskfödd med båda föräldrarna svenskfödda.

Sammantaget för båda provdelarna är det tydligt att det är föräldrarnas ursprung som har den största betydelsen för provresultatet, större än om man är född utomlands eller i Sverige. Skillnaderna generellt sett är större på den verbala delen än på den kvantitativa, samtidigt som variationerna över delprov är större på den kvantitativa delen än på den verbala.

# Sammanfattning gällande provresultat för olika grupper av provtagare

Män har högre poäng än kvinnor på båda provdelarna. Skillnaden är cirka 0,5-0,6 standardavvikelser på den kvantitativa delen och cirka 0,2 standardavvikelser på den verbala delen. Könsskillnaden varierar inte så mycket över de olika kvantitativa delproven, medan den är dubbelt så stor på ELF som något av de andra verbala delproven. Könsskillnaderna på provdelsnivå har förändrats mycket lite från det gamla till det nya provet. På totalpoängsnivå är skillnaderna större, vilket beror på att könsskillnaderna är större på den kvantitativa delen än på den verbala och att den kvantitativa delen har relativt större vikt i det nya provet än den hade i det gamla.

Det är den yngsta åldersgruppen som har det högsta resultatet på den kvantitativa delen, och det är relativt stora skillnader mellan olika åldersgrupper på XYZ, KVA och NOG, medan skillnaden mellan olika åldersgrupper är försumbar på delprovet DTK. Poängen på den verbala delen varierar mycket med ålder, på så sätt att ju högre ålder desto högre poäng, och det skiljer drygt 1 standardavvikelse mellan den yngsta och äldsta gruppen. På ORD är skillnaden mellan de äldsta och de yngsta provtagarna cirka 1,7 standardavvikelser, och på MEK 1,3. Skillnaderna är betydligt mindre på LÄS och ELF med 0,5 respektive 0,3 standardavvikelser.

Skillnaderna mellan provtagare från olika utbildningsbakgrunder är avsevärda på båda provdelarna där provtagare som har föräldrar med en hög utbildningsnivå även har högre provpoäng. På den kvantitativa delen är skillnaden uppemot 1 standardavvikelse och på den verbala delen är skillnaden uppemot 0,8 standardavvikelser. På delprovsnivå är skillnaden störst på XYZ respektive ELF, och minst på DTK respektive ORD.

De provdeltagare som har gått en naturvetenskapligt inriktad utbildning har klart högst poängmedelvärde på båda provdelarna, medan de som gått en utbildning med inriktning mot barn/fritid, handel/administration, hotell/restaurang, livsmedel har lägst poängmedelvärde på båda provdelarna. Skillnaden mellan grupperna är uppemot 1,5 standardavvikelser på den kvantitativa delen och 1,0 på den verbala delen. På delprovsnivå är skillnaden störst på XYZ och KVA respektive ELF, och minst på NOG och DTK respektive ORD.

Skillnaden mellan provtagare med helsvenskt ursprung och provtagare med helutländskt ursprung är uppemot 0,5 standardavvikelser på den kvantitativa provdelen och 0,7 standardavvikelser på den verbala provdelen. På den kvantitativa delen är skillnaden klart minst på XYZ och störst på DTK, medan den inte varierar särskilt mycket över de verbala delproven. Sammantaget för båda provdelarna är det tydligt att det är föräldrarnas ursprung som har den största betydelsen för provresultatet, större än om man är född utomlands eller i Sverige. Skillnaderna generellt sett är större på den verbala delen än på den kvantitativa, samtidigt som variationerna över delprov är större på den kvantitativa delen än på den verbala.

# 5. Samband mellan provresultat och betyg

Studien om samband mellan högskoleprovresultat och betyg och provresultat från gymnasieskolan är uppdelad på två olika grupper: de som gått ut gymnasieskolan (avgångna gymnasieelever) och de som fortfarande går i gymnasieskolan (innevarande gymnasieelever). Den fullständiga rapporten över studien finns i Bilaga D.

## Avgångna gymnasielever

Dessa provtagare är födda mellan 1986 och 1992, dvs. de var mellan 19 och 25 vid provet hösten 2011. De huvudsakliga variabler som studeras är högskoleprovsresultat, jämförelsetal från gymnasiet (betygsgenomsnitt), betyg från vissa kurser (Ma A, Ma B, En A, En B), samt bakgrundsvariabler som exempelvis kön, ålder, utbildningsprogram, föräldrars utbildning, och svenskt/utländskt ursprung. Antalet provtagare i denna grupp var 23 214 vid provet 11:B och 27 075 vid provet 12:A.

### Korrelationer

Korrelationer mellan högskoleprovspoäng och gymnasiebetyg i form av jämförelsetal och kursbetyg kan ses i Tabell 5:1 (motsvarande Table 1 i Bilaga D). Korrelationen mellan den totala normerade poängen på provet och jämförelsetalet är cirka 0,5, vilket är av samma magnitud som motsvarande korrelationer för det gamla högskoleprovet (se, t.ex., Stage, 1992, och Lexelius, 2004). Korrelationen mellan poäng på den kvantitativa provdelen och betygen i matematik är cirka 0,6, och mellan poäng på den verbala provdelen och betygen i engelska drygt 0,5, vilket är ungefär som på det gamla provet (t.ex., Lexelius, 2004). Korrelationerna är förhållandevis stabila över undergrupper (se Table A4-A6 i Bilaga D), förutom i två avseenden. Korrelationerna är något högre bland provtagare med högutbildade föräldrar, och bland provtagare på studieförberedande/teoretiska program jämfört med yrkesförberedande/praktiska program på gymnasiet.

Tabell 5:1: Korrelationer mellan högskoleprovspoäng och betyg från gymnasiet för utgångna elever.

| | Provversion | |
|---|---|---|
| | 11:B | 12:A |
| HP ↔ Jämförelsetal | 0,51 | 0,46 |
| KVANT ↔ Ma A | 0,66 | 0,61 |
| KVANT ↔ Ma B | 0,59 | 0,54 |
| VERB ↔ En A | 0,55 | 0,56 |
| VERB ↔ En B | 0,54 | 0,53 |

Not: Pearsonkorrelationskoefficienter. Alla korrelationer är signifikanta på 0,01-nivån.
HP = Normerad poäng på högskoleprovet; KVANT = Normerad poäng på kvantitativa provdelen; VERB = Normerad poäng på verbala provdelen; Ma = Matematik; En = Engelska.

### Avgörande faktorer för högskoleprovsresultat och kursbetyg

Denna del av studien bygger på linjära regressionsanalyser, och detaljerade resultat presenteras i tabellerna A7-A9 i Bilaga D. När det gäller totalpoängen på provet så har männen 0,13 högre normerad poäng än vad kvinnorna har, vilket motsvarar 0,35 standardavvikelser. På den kvantitativa delen är skillnaden 0,20 i normerad poäng vilket är 0,44 standardavvikelser, och motsvarande på den verbala delen är 0,05 i normerad poäng eller 0,11 standardavvikelser. När det gäller jämförelsetalet så har

kvinnorna 1 poäng bättre än männen, vilket är en skillnad på 0,38 standardavvikelser. Tittar vi däremot på specifika betyg i vissa kärnämnen så presterar männen bättre än kvinnorna i såväl matematik (kurs A; 0,37 poäng, motsvarande 0,1 standardavvikelser) som engelska (kurs A).

Det är värt att poängtera att provtagare skiljer sig från den generella gymnasieelevspopulationen i minst två avseenden. För det första är provtagarna mer högpresterande än populationen i helhet, och för det andra har en större andel av provtagarna examen från ett studieförberedande program. Att männen presterar bättre än kvinnorna i det här samplet beror det sistnämnda.

Ålder är också en förklaringsfaktor för betyg och provresultat. Äldre provtagare har lägre jämförelsetal, kursbetyg och poäng på den kvantitativa provdelen än yngre provtagare, medan de har högre poäng på den verbala provdelen.

När det gäller provtagarnas ursprung så har 2:a generationens invandrare[4] lägre kursbetyg, jämförelsetal och provpoäng än provtagare med svenskt ursprung, och 1:a generationens invandrare har lägre på dessa mått än 2:a generationens invandrare. Undantaget är Engelska A där 2:a generationens invandrare har samma betyg som svenskfödda provtagare. Skillnaderna på högskoleprovet är störst på den verbala provdelen. Resultaten visar även att faktorer som tidigare erfarenhet av högre utbildning, deltagande i studieinriktade program och föräldrar med högre utbildningsnivå har en tydlig positiv påverkan på kursbetyg, jämförelsetal och provpoäng.

## Poängskillnader

Analyserna av poängskillnader utgår från att jämförelsetal och provpoäng båda transformerats till en standardiserad skala med medelvärde 0 och standardavvikelse 1 (*z*-skala), vilket innebär att koefficienter tolkas i termer av standardavvikelser för endera mått. Det dessa poängskillnader visar är en relativ rangordning, och en positiv koefficient innebär en högre rangordning utifrån jämförelsetalet jämfört med provresultat, och en negativ koefficient innebär en högre rangordning utifrån provresultat jämfört med jämförelsetalet.

I Tabell 5:2 nedan (Table 2 i Bilaga D) ser vi att den relativa rangordningen påverkas av de flesta variablerna. I genomsnitt hamnar män (variabeln MALE) 0,7-0,8 standardavvikelser högre i provpoängsfördelningen än i fördelningen av jämförelsetalet. Äldre provtagare (variabeln AGE) rangordnas relativt högre av högskoleprovet än av jämförelsetalet, och detta på grund av att äldre provtagare har lägre jämförelsetal än yngre. Föräldrars utbildning (variabeln HIGHEDU) påverkar också skillnaden i rangordning, om än i liten grad, där provtagare vars föräldrar har en högre utbildningsbakgrund rangordnas högre av högskoleprovet än av jämförelsetalet.

Vid en undersökning av rangordningsskillnaderna mellan Ma A och KVANT respektive En A och VERB så är skillnaderna som förväntat mindre, men de är i många fall signifikanta. Män rangordnas högre på KVANT än Ma A, och även högre på VERB än En A men skillnaden är inte lika stor. Ålder påverkar skillnaden En A–VERB mer än skillnaden Ma A–KVANT, medan det omvända gäller med avseende på provtagarnas ursprung.

## Poängskillnader mellan män och kvinnor

Som redovisats ovan finns det betydande skillnader mellan män och kvinnor, både när det gäller betyg och högskoleprovsresultat (se Table A10 och A11 i Bilaga D). Vid djupare analyser av dessa skillnader i olika undergrupper framkommer att könsskillnaderna som finns i skillnaden mellan jämförelsetal och högskoleprovspoäng inte varierar med ålder och föräldrars utbildningsnivå. Det finns emellertid vissa

[4] Att de två grupperna med utlandsfödda föräldrar i denna studie benämns som invandrare beror på att de har gått svenskt gymnasium och därmed kan vi vara mer säkra på att de uppehåller sig mer permanent i landet.

indikationer att denna skillnad skulle variera med ursprung när det gäller jämförelsetalet, medan denna skillnad inte finns på högskoleprovspoängen.

Tabell 5:2. Variabler som påverkar poängskillnaden mellan jämförelsetalet och den normerande poängen på högskoleprovet.

| | Provversion | |
|---|---|---|
| Oberoende variabel | 11:B | 12:A |
| Konstant | 3,19** | 3,55** |
| MALE | -0,72** | -0,77** |
| AGE | -0,12** | -0,13** |
| FIRSTGEN | 0,07* | 0,09** |
| SECONDGEN | 0,05* | 0,03 |
| FORBORN | 0,15 | 0,17 |
| MIDEDU | -0,04 | -0,06 |
| HIGHEDU | -0,08* | -0,10** |
| ACTRACK | -0,25** | -0,21** |
| KOMVUX | -0,61** | -0,55** |
| HIGHEXP | -0,14** | -0,10** |
| D11B | - | -0,17** |
| $R^2$ | 0,20 | 0,22 |
| N | 19 708 | 22 030 |

Not: Den beroende variabeln är skillnaden mellan jämförelsetal och normerad högskoleprovspoäng, där båda måtten transformerats till en standardiserad *z*-skala.
** Koefficienten är signifikant på 0,01-nivån.
* Koefficienten är signifikant på 0,05-nivån.

# Innevarande gymnasieelever

Dessa provtagare är födda 1993 eller 1994, vilket innebär att de var 17 eller 18 vid höstprovet 2011. De variabler som studeras är till stor del samma som för de utgångna eleverna, men när det gäller gymnasievariablerna finns det skillnader. Eftersom alla dessa elever inte har de ovan nämnda kursbetygen så används istället resultat från nationella prov i matematik, engelska och svenska. Av samma anledning har eleverna inte något jämförelsetal, och därför fokuseras analyserna på högskoleprovets två provdelar och de nationella kursproven. Dock görs vissa jämförelser även med meritvärdet från grundskolans årskurs 9.

## Korrelationer

Korrelationen mellan den normerade poängen på högskoleprovet och meritvärdet i årskurs 9 är 0,53 respektive 0,54 för de två provomgångarna (Tabell 5:3; Table 4 i Bilaga D). Korrelationerna mellan provdelarna och betygen på nationella prov är generellt högre; uppåt 0,7 mellan KVANT och NP Ma A/B och kring 0,6 mellan VERB och NP En A/B. Korrelationen mellan VERB och Sv B är betydligt lägre, vilket är i linje med resultat från en tidigare studie (Lexelius, 2004).

Jämfört med resultaten från analyserna på de avgångna gymnasieeleverna så kan vi notera att sambanden är högre mellan provpoäng och betyg på nationella prov än mellan provpoäng och motsvarande kursbetyg. Detta är förväntat med tanke på att kursbetygen är baserade på andra former

av bedömningar än bara prov. Provbetygen däremot är baserade just på ett prov på samma sätt som högskoleprovet (även om innehåll och uppgiftsformat så klart skiljer sig åt) .

Tabell 5:3. Parvisa korrelationer mellan poäng på högskoleprovet och meritvärde från årskurs 9 respektive betyg från nationella kursprov i gymnasiet.

| | Provversion | |
|---|---|---|
| | 11:B | 12:A |
| HP ↔ Meritvärde åk 9 | 0,54 | 0,53 |
| KVANT ↔ NP Ma A | 0,68 | 0,64 |
| KVANT ↔ NP Ma B | 0,70 | 0,66 |
| VERB ↔ NP En A | 0,58 | 0,63 |
| VERB ↔ NP En B | 0,61 | 0,59 |
| VERB ↔ NP Sv B | 0,34 | 0,35 |

Not: Pearsonkorrelationskoefficienter. Alla korrelationer är signifikanta på 0,01-nivån.
HP = Normerad poäng på högskoleprovet; KVANT = Normerad poäng på kvantitativa provdelen; VERB = Normerad poäng på verbala provdelen; NP = Nationella kursprov; Ma = Matematik; En = Engelska; Sv = Svenska.

### Avgörande faktorer för högskoleprovsresultat och betyg på nationella kursprov

För denna grupp studerades avgörande faktorer för högskoleprovspoäng och betyg på nationella kursprov. Dessa resultat åter finns i tabellerna B5 och B6 i Bilaga D, och de överensstämmer väl med motsvarande resultat för de avgångna eleverna och kommenteras därför inte vidare.

### Poängskillnader

Givet att de nationella kursproven har högre samband med högskoleprovspoäng än vad kursbetygen har så kan vi förvänta oss mindre systematisk variation i poängskillnaderna mellan olika grupper av provtagare, och resultaten som visas i Table 5 i Bilaga D ger stöd för detta. Det finns en konsekvent skillnad mellan män och kvinnor när det gäller jämförelsen NP Ma A–KVANT, och denna skillnad (0,37 standardavvikelser) är i samma storleksordning som för de avgångna eleverna. Det finns ingen signifikant könsskillnad i jämförelsen NP En A–VERB, vilket det fanns bland de avgångna eleverna.

Vidare kan vi se att provtagarnas ursprung har betydelse för skillnaden NP En A–VERB, vilket det även hade bland de avgångna för skillnaden En A–VERB. Ursprunget har liten eller ingen betydelse för vare sig skillnaden NP Ma A–KVANT för denna grupp eller för skillnaden Ma A–KVANT för de avgångna eleverna. När det så gäller ålder går dessa skillnader i samma riktning för såväl de innevarande som de avgångna eleverna. Som nämnts tidigare beror dessa skillnader på att de äldre provtagarna har lägre betyg än de yngre, och sett ur det perspektivet verkar det som att provet faktiskt används som en andra chans.

## Sammanfattning och slutsatser gällande sambanden mellan provresultat och betyg

Den generella resultatbilden från denna studie stämmer väl överens med resultat från studier på det gamla högskoleprovet. Korrelationen mellan den totala normerade poängen på högskoleprovet och jämförelsetalet från gymnasiet är cirka 0,5, vilket den var även på det gamla provet. Korrelationerna mellan olika provpoäng å ena sidan och jämförelsetal, kursbetyg och betyg på nationella kursprov å andra sidan är i stort sett stabila över olika provtagargrupper.

Kön är en av de variabler som förklarar mest av skillnaden i rangordning mellan högskoleprov och jämförelsetal. Bland provtagarna har männen lägre jämförelsetal men högre högskoleprovspoäng än kvinnorna, och dessutom har männen högre kursbetyg och betyg på nationella prov. Dessa könsskillnader verkar vara konstanta över olika undergrupper.

Vid en närmare titt på nationella prov-betyg i hela provtagargruppen (vilket inte studeras i Bilaga D) framkommer att skillnaden mellan män och kvinnor uttryckt i *d* är cirka 0,2 för nationella proven på kurserna En A, En B, Ma A och Ma B. Om man ser prestationen på dessa nationella prov som ett mått på provtagarnas förkunskaper så kan vi dra slutsatsen att åtminstone en del av den könsskillnad som förekommer på högskoleprovet beror på skillnader i förkunskaper. Det faktum att båda dessa mått är baserat på samma typ av mätning, dvs. prov, spelar givetvis också roll. Men, det går ändå inte att bortse från att könsskillnaderna har samma riktning på såväl högskoleprovet som i de undersökta nationella proven.

Provtagare med utländsk bakgrund rangordnas lägre av högskoleprovet än av gymnasiebetyg, och då framförallt i jämförelsen mellan den verbala provdelen på högskoleprovet och betygen i engelska medan de rangordnas relativt lika av den kvantitativa provdelen och betygen i matematik. När det gäller olika provtagargrupper med avseende på föräldrarnas utbildning så är skillnaderna i rangordning mycket små, och den största skillnaden föreligger bland provtagare med högutbildade föräldrar där dessa rangordnas relativt bättre av betygen än av provpoängen. Provtagare med en huvudsakligen studieförberedande inriktning på gymnasiet rangordnas relativt bättre av högskoleprovet än av jämförelsetalet. Skillnader mellan olika åldersgrupper förekommer princip i alla jämförelser mellan provresultat och betyg, och överlag har äldre provtagare lägre betyg men högre provresultat.

Tittar vi närmare på provets roll som en andra chans i urvalet till högre utbildning så är en relevant fråga vilka provtagare det är som genom provresultatet faktiskt klättar i rangordning jämfört med det jämförelsetal de har (se diskussionsavsnittet i Bilaga D för mer detaljer kring definitioner, analys och diskussion gällande andra chans-fenomenet). I korthet visar resultaten att det är framförallt för män, äldre provtagare och i viss mån provtagare med svensk bakgrund som provet får rollen som en andra chans.

# 6. Sammanfattning, slutsatser och rekommendationer

## Syftet med revideringarna och de aktuella studierna

Det övergripande syftet med den nya provsammansättningen har varit att höja provets prognosförmåga, och detta skulle åstadkommas genom (1) att dela upp provet i en verbal och en kvantitativ provdel som kunde användas flexibelt i urvalet beroende på utbildningsinriktning, vilket i sin tur krävde att antalet kvantitativa uppgifter utökades markant, och (2) en effektivisering av provet i termer av antalet uppgifter per tidsenhet. När provet infördes hösten 2011 skulle detta ske inom ramen för en försöksverksamhet på två år, och provresultaten skulle därefter utvärderas. När det gäller utvärderingen av provet hade det optimala och logiska varit att undersöka just provets prognosförmåga, men detta är inte möjligt i någon större skala ännu eftersom det inte finns något prestationskriterium tillgängligt. Anledningen till detta är att det har gått för kort tid från det att det nya provet infördes till dess att de första provtagarna med resultat därifrån har genomgått ett år i högskolan. Därför inriktades denna utvärdering i stället på att utvärdera de två huvudsakliga åtgärderna i samband med revideringen: uppdelningen verbal-kvantitativ och effektiviseringen. Utvärderingen gjordes utifrån ett validitetsperspektiv (AERA, APA, & NCME, 1999), där fokus var på validitetsbevis från svarsprocesser, intern struktur, och samband med externa variabler.

## Resultat

### Validitetsbevis gällande intern struktur

Reliabiliteten för totalpoängen är ungefär som för det gamla högskoleprovet (cirka 0,93), och detsamma gäller för den verbala provdelen (0,90–0,91). Reliabiliteten på den kvantitativa provdelen är högre på det nya provet än vad den var på det gamla och når värden kring 0,90, vilket kan ses som ett tillräckligt högt och lämpligt värde. Analysen av provets interna struktur visar att den underliggande teoretiska modellen där ORD, LÄS, MEK och ELF är indikatorer på en verbal faktor och där XYZ, KVA, NOG och DTK är indikatorer på en kvantitativ faktor inte har en perfekt anpassning till faktiska provdata, utan denna anpassning kan snarare betecknas som acceptabel. Studien av mervärdet i delprovspoängen ger dock stöd för att rapportera en kvantitativ och en verbal poäng separat. Sammantaget ger studierna gällande provets interna struktur stöd för att den nya provsammansättningen med separata verbala och kvantitativa delar fungerar som avsett, och att en flexibel användning av de två provdelspoängen torde kunna bringa ett mervärde till urvalsprocessen.

### Validitetsbevis gällande svarsprocesser

Sammantaget ger studierna om icke-svar och positionseffekter indikationer om att framförallt den kvantitativa provdelen är tidspressad, och att förmågan att hantera denna tidspress inte har något samband med provprestation. Men, resultaten indikerar samtidigt att graden av tidspress minskar under provdagen vilket skulle innebära att tidspressen kan lindras genom grundliga förberedelser från provtagarnas sida. Om provtagarna vore väl införstådda med provets format, hade klara tidshanteringsstrategier och i övrigt var väl förberedda inför provdagen så borde vi inte se dessa mönster i icke-svar och positionseffekter.

## Validitetsbevis gällande samband med externa variabler

### *Gruppskillnader i provresultat*

När de gäller gruppskillnader i provresultat utifrån vissa bakgrundsvariabler är studieinriktning på gymnasiet den faktor som enskilt har störst betydelse, och det är elever på just studieförberedande inriktningar (ffa. en naturvetenskaplig sådan) som genomgående har de högsta provpoängen. Dock beror denna skillnad mycket på klassificeringen av studieinriktningar, och skulle vi välja ut och jämföra enskilda program skulle skillnaderna bli ännu större. Man ska också komma ihåg att det är många andra variabler som påverkar just valet av studieinriktning (socialgrupp, förkunskaper, kön, m.m.) och dessa är såklart viktiga för att förklara prestationsskillnaderna mellan de olika studieinriktningarna. Skillnaderna är större på den kvantitativa delen än på den verbala.

Föräldrarnas utbildning har också stor betydelse för provresultatet. Ju högre utbildning de har desto bättre provresultat, och skillnaderna är något högre på den kvantitativa delen än på den verbala. Även provtagarnas ålder har stor betydelse för provresultaten, framförallt på den verbala delen. Ju äldre provtagarna är desto högre poäng har de på den verbala delen, och skillnaderna är som allra störst på delproven ORD och MEK. Vidare har också provtagarnas ursprung betydelse för provresultatet, och då har föräldrarnas ursprung (svenskfödda eller utlandsfödda) större betydelse än var provtagarna själva är födda. Provtagare med ”helsvenskt” ursprung presterar bättre än provtagare med ”helutländskt” ursprung, och skillnaderna är något större på den verbala delen än på den kvantitativa. Skillnaderna är som minst på de nya kvantitativa delproven XYZ och KVA, medan de inte varierar så mycket över de verbala delproven.

När det så gäller kön så har män högre poäng än kvinnor på såväl det gamla som det nya högskoleprovet. Skillnaderna är större på den kvantitativa delen än på den verbala, vilket tillsammans med det faktum att den kvantitativa delen har relativt större vikt i det nya provet än i det gamla gör att skillnaderna på totalpoängsnivå är något större på det nya provet än på det gamla. På provdelsnivå är dock skillnaderna mellan det gamla och det nya provet mycket små. Skillnaden varierar inte mycket över de kvantitativa delproven, medan den är dubbelt så stor på delprovet ELF som på något av de andra verbala delproven.

### *Samband mellan provresultat och betyg*

Den generella resultatbilden när det gäller samband mellan provresultat och betyg stämmer väl överens med resultat från studier på det gamla högskoleprovet. Korrelationen mellan den totala normerade poängen på högskoleprovet och jämförelsetalet från gymnasiet är cirka 0,5, vilket den var även på det gamla provet. Korrelationerna mellan olika provpoäng å ena sidan och jämförelsetal, kursbetyg och betyg på nationella kursprov å andra sidan är i stort sett stabila över olika provtagargrupper.

Kön är en av de variabler som förklarar mest av skillnaden i rangordning mellan högskoleprov och jämförelsetal. Bland provtagarna har männen lägre jämförelsetal men högre högskoleprovspoäng än kvinnorna, och dessutom har männen högre kursbetyg och betyg på nationella prov. Dessa könsskillnader verkar vara konstanta över olika undergrupper. Vid en närmare titt på nationella prov-betyg i hela provtagargruppen framkommer att skillnaden mellan män och kvinnor uttryckt i *d* (antalet standardavvikelser) är cirka 0,2 för ett flertal nationella prov i engelska och matematik, vilket innebär att könsskillnaderna har samma riktning på såväl högskoleprovet som i de undersökta nationella proven. Om man antar att prestationen på dessa nationella prov är ett bra mått på provtagarnas förkunskaper så kan vi dra slutsatsen att åtminstone en del av den könsskillnad som förekommer på högskoleprovet beror på skillnader i förkunskaper.

Provtagare med utländsk bakgrund rangordnas lägre av högskoleprovet än av gymnasiebetyg, och då framförallt i jämförelsen mellan den verbala provdelen på högskoleprovet och betygen i engelska

medan de rangordnas relativt lika av den kvantitativa provdelen och betygen i matematik. När det gäller olika provtagargrupper med avseende på föräldrarnas utbildning så är skillnaderna i rangordning mycket små, och den största skillnaden föreligger bland provtagare med högutbildade föräldrar där dessa rangordnas relativt bättre av betygen än av provpoängen. Provtagare med en huvudsakligen studieförberedande inriktning på gymnasiet rangordnas relativt bättre av högskoleprovet än av jämförelsetalet. Skillnader mellan olika åldersgrupper förekommer princip i alla jämförelser mellan provresultat och betyg, och överlag har äldre provtagare lägre betyg men högre provresultat.

# Slutsatser och rekommendationer

Vår övergripande slutsats från valideringsstudierna är att de förändringar som infördes i högskoleprovet hösten 2011 hittills fungerar som avsett. De utförda studierna ger stöd för att den nya provsammansättningen har fått de önskvärda egenskaperna i termer av reliabilitet och dimensionalitet. Studierna som är kopplade till svarsprocesser ger skäl att ifrågasätta effektiviseringen av provet (dvs. fler uppgifter per tidsenhet), men mer djupgående studier måste göras innan vi kan ge några rekommendationer om att ändra antalet uppgifter i provet. Däremot rekommenderar vi att en översyn görs när det gäller informationen om provet som riktar sig till potentiella provtagare.

Gruppskillnaderna i provresultat är inte mer anmärkningsvärda än på det gamla provet. Speciellt skillnaden mellan män och kvinnor brukar vara i fokus när provet diskuteras i media och annorstädes, men det verkar som att denna skillnad till viss del kan förklaras av skillnader i förkunskaper. Överlag är vår rekommendation att inte göra några förändringar i provsammansättningen åtminstone till dess att provets prognosförmåga undersökts. Vi skulle i detta sammanhang återigen vilja lyfta fram behovet av att undersöka rimligheten i den nuvarande skalan för den normerade poängen och möjligheterna att revidera denna. Denna fråga lyftes gentemot provets tidigare huvudman, Högskoleverket, inför att det nya provet infördes men denne såg då ingen anledning att revidera skalan. Vi anser dock att de anledningar för att ändra skalan som då framfördes fortfarande är relevanta.

# 7. Referenser

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Fan, X., Thompson, B., & Wang, L. (1999). Effects of sample size, estimation method, and model specification on structural equation modeling fit indexes. *Structural Equation Modeling, 6*, 56-83.

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229.

Högskoleverket (2000). *Högskoleprovet. Gårdagens mål och framtida inriktning* (Högskoleverkets rapportserie nr. 2000:12 R). Stockholm: Högskoleverket.

Högskoleverket (2002). *The Swedish national aptitude test: a 25-year testing program* (Högskoleverkets rapportserie nr. 2002:22 R). Stockholm: Högskoleverket.

Lexelius, A. (2004). *Högskoleprovet och de målrelaterade betygen. En studie av de första eleverna med de nya gymnasiebetygen* (BVM nr. 8). Umeå: Umeå universitet, Institutionen för beteendevetenskapliga mätningar.

Lyrén, P.-E. (2009). Reporting subscores for college admission tests. *Practical Assessment, Research & Evaluation, 14*(4). Tillgänglig från: http://pareonline.net/getvn.asp?v=14&n=4

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Schreiber, J. B., Nora, A., Stage, F. K., Barlow, E. A., & King, J. (2006). Reporting structural equation modeling and confirmatory factor analysis results: A review. *The Journal of Educational Research, 99*(6), 323–337.

Stage, C. (1992). How important are age and education for gender differences in test results? *Scandinavian Journal of Educational Research, 36*(3), 223–235.

Stage, C., & Ögren, G. (2010). *Ett nytt högskoleprov. Bakgrund och konsekvenser* (BVM nr. 42). Umeå: Umeå universitet, Institutionen för tillämpad utbildningsvetenskap.

Streiner, D. L. (2003). Starting at the beginning: An introduction to Coefficient Alpha and internal consistency. *Journal of Personality Assessment*, *80*(1), 99-103.

Wiberg, M., & Sundström, A. (2009). A comparison of two approaches to correction of restriction of range in correlation analysis. *Practical Assessment, Research & Evaluation*, 14(5). Tillgänglig från: http://pareonline.net/getvn.asp?v=14&n=5

# Bilagor

Bilaga A1: Intern struktur (Kapitel 2)

Bilaga A2: Värdet i delprovspoäng (Kapitel 2)

Bilaga B: Icke-svar (Kapitel 3)

Bilaga C: Provresultat för olika grupper (Kapitel 4)

Bilaga D: Samband mellan provresultat och betyg (Kapitel 5)

# Intern struktur

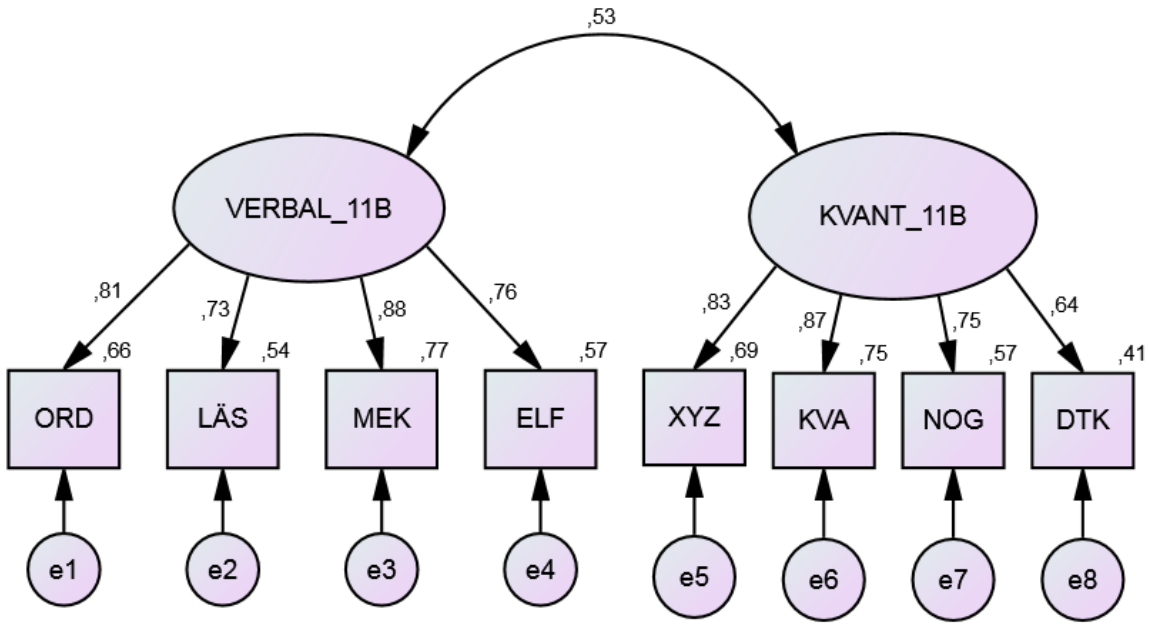
*Per-Erik Lyrén*

CFA-analyserna är gjorda i IBM SPSS Amos 21 med *maximum likelihood*-estimering.

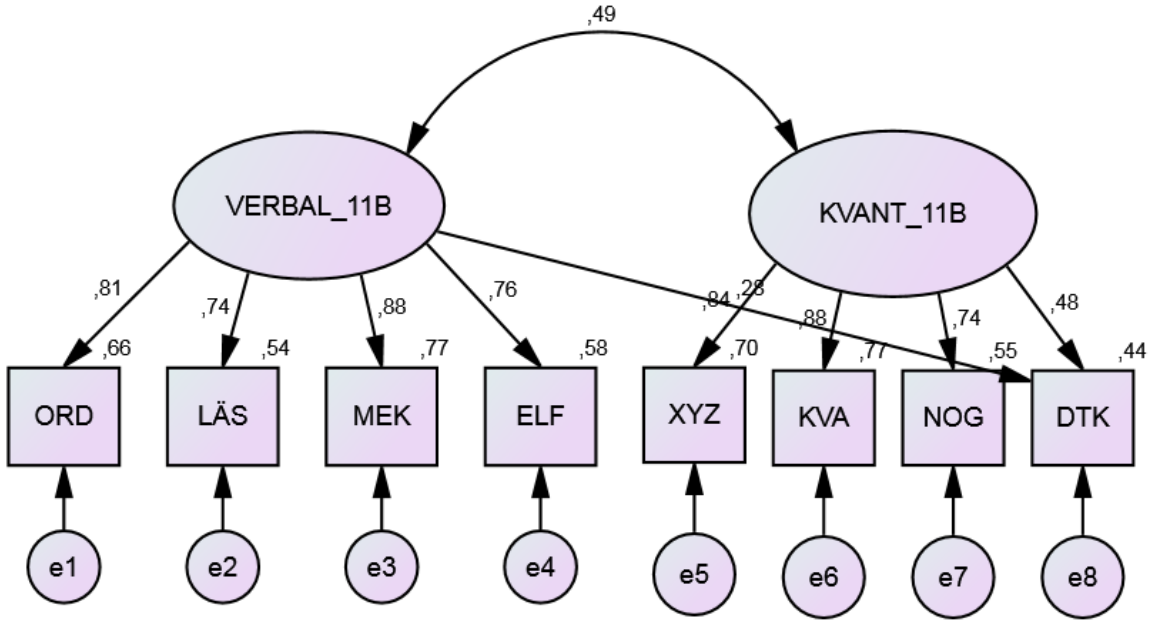
Tabell A1:1. Anpassningsindex från CFA-analyser av data från det nya högskoleprovet.

| Index | Grundmodell | | Variant VERB→DTK | |
|---|---|---|---|---|
| | 11:B | 12:A | 11:B | 12:A |
| $chi^2$ | 10314 | 12254 | 7254 | 7655 |
| df | 19 | 19 | 18 | 18 |
| $chi^2$/df | 543 | 645 | 403 | 425 |
| CFI | 0,94 | 0,95 | 0,96 | 0,97 |
| RMSEA | 0,12 | 0,11 | 0,10 | 0,09 |
| SRMR | 0,07 | 0,07 | 0,06 | 0,05 |

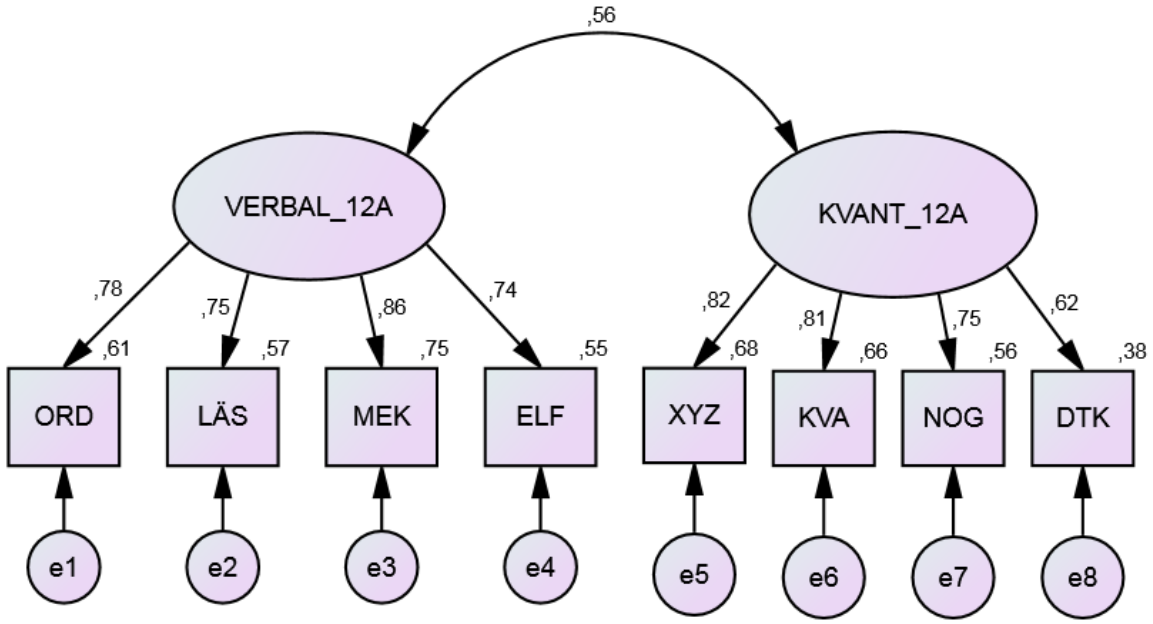
Not: Grundmodell: (VERB → ORD, LÄS, MEK, ELF), (KVANT → XYZ, KVA, NOG, DTK), VERB ↔ KVANT
Variant: (VERB → ORD, LÄS, MEK, ELF, DTK), (KVANT → XYZ, KVA, NOG, DTK), VERB ↔ KVANT



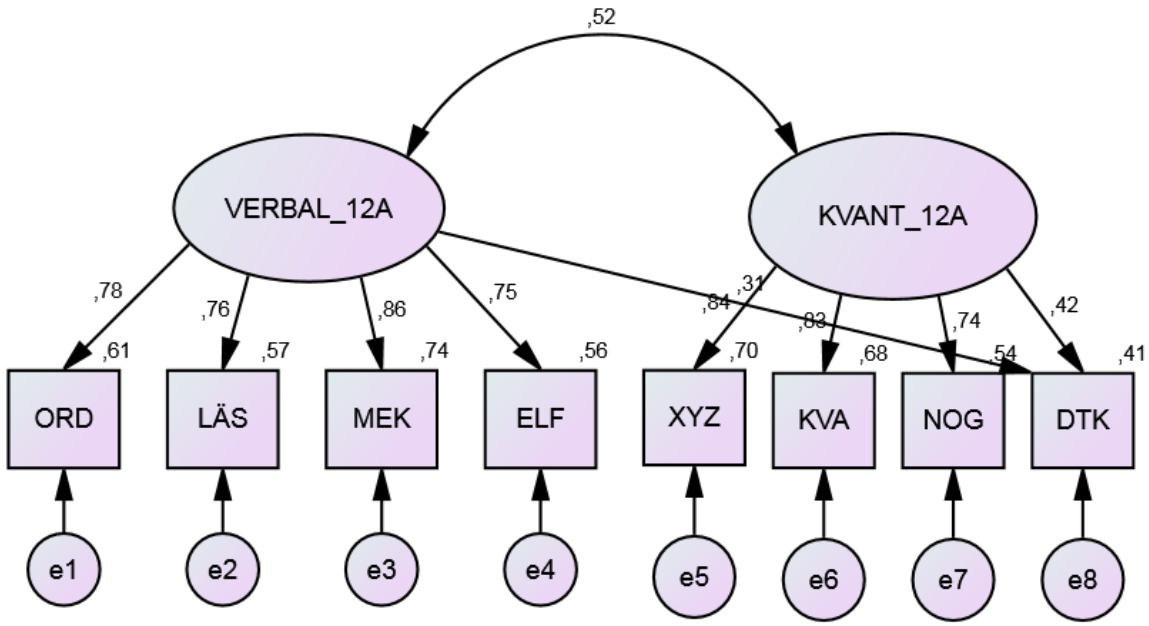

Figur A1:1. Mätmodell (grundmodell) med standardiserade parameterestimat från CFA-analys på provet 11:B. e1–e8 betecknar feltermer.




Figur A1:2. Mätmodell (variant med VERB → DTK) med standardiserade parameterestimat från CFA-analys på provet 11:B. e1–e8 betecknar feltermer.




Figur A1:3. Mätmodell (grundmodell) med standardiserade parameterestimat från CFA-analys på provet 12:A. e1–e8 betecknar feltermer.




Figur A1:4. Mätmodell (variant med VERB → DTK) med standardiserade parameterestimat från CFA-analys på provet 12:A. e1–e8 betecknar feltermer.

# A comparison of methods for examining the added value of subscores

Jonathan Wedman & Per-Erik Lyrén

Abstract

When subscores on a test are reported to the test taker, the appropriateness of reporting them depends on whether they provide meaningful information above what is provided by the total score. Subscores that fail to do so lack added value and should not be reported. There are several methods for examining the added value of subscores, and in this study seven such methods are compared, four of which are based on classical test theory and three of which are based on item response theory. The data consisted of test takers' scores on four test forms – two administrations of an old version of a college admission test and two administrations of a new version – and the analyses were carried out on the subtest and section levels. The two section scores were found to have added value with all methods used, whereas the results for subtest scores ranged from almost all scores providing added value to none providing added value. This highlights the problem of choosing an appropriate method for acquiring empirical evidence to support decisions concerning the reporting of subscores to test takers.

A test score is intended to reflect the test takers' knowledge in the domain purportedly measured by the test. When a test score is reported back to the test taker, those responsible for the testing programs should provide appropriate interpretations of the score. Also, when a test score is intended to be used for decisions as to which areas to focus on in remedial studies, those providing the score must take measures to ensure that the score is useful for such purposes (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999). The same is true for scores derived from any subset of a test, which are generally referred to as *subscores*. However, a specific requirement for subscores to be considered useful is that they should have *added* value over the total score (AERA et al., 1999). Subscores that lack added value should not be reported because they are considered indistinguishable from a less reliable version of the total score and, therefore, do not provide the test taker with any relevant information. The issue of whether subscores have added value or not is, on a test taker level, primarily a concern when the information is to be used as a basis for remedial study decisions (Monaghan, 2006) or is to be used for high-stakes decisions such as certification, placement, or college admission. It is important, therefore, to find empirical evidence for the added value of subscores that are to be reported to test takers. In this study, we explore different methods to find such evidence.

**Purpose**

There are two purposes of this study. The main purpose is to compare the results from the different methods of determining the value of a subscore. More specifically, we want to examine to what extent the conclusions that can be drawn when applying the different methods to operational tests are in agreement with each other. A secondary purpose is to examine the added value of the subscores on these operational tests and to determine if they should be reported to the test taker.

The remainder of the paper is structured as follows. We first present previous applications of methods for examining subscore value. This is followed by the method section that provides information about the tests, participants, hypotheses, software, and the subscore value methods. The results are then presented separately for each method. Finally, in the discussion section, the consequences of the results, the limitations of the study, and some thoughts on further research are presented.

**Previous applications of methods for examining subscore value**

Sinharay, Puhan, and Haberman (2011) reviewed several methods for examining the added value of subscores, and given the comprehensiveness of their review we have based our

own review of previous research on these methods. The methods they discussed were factor analysis, the beta-binomial model, multidimensional item response theory (MIRT), DIMTEST, DETECT, and Haberman's (2008) method based on classical test theory (CTT). In addition to these six methods, we also apply the Utility index, which is closely linked to Haberman's method.

**Haberman's method and Utility Index**

Haberman (2008) suggested a method for examining subscore value based on CTT and applied it to SAT I data from 2002. When applying his method, a subscore is considered to have added value if its reliability is greater than the reliability of the observed total score as a measure of the subtest construct. He found that none of the subtest scores in the SAT I had added value but that both section scores did. Haberman (2008) concluded that the relative value of subscores increases when their reliability increases, when the reliability of the total score decreases, and when the correlation between the true subscores and the true total score decreases.

A simulation study by Sinharay (2010) provided more details concerning the extent to which reliabilities and correlations affect subscore value. Haberman's method was further described in Sinharay, Haberman and Puhan (2007) and was applied to a test for prospective and practicing teacher's aides where the subscores were found to lack added value. Puhan, Sinharay, Haberman and Larkin (2010) applied the method to six tests used for educational certification and also found a lack of added value for all subscores in those six tests. Lyrén (2009) applied the method to a college admissions test and found, in contrast to the other studies, that four out of the five subscores and both section scores had added value.

Brennan (2011) proposed and applied a statistic called the Utility Index to the SAT I data analyzed by Haberman (2008). The Utility Index is algebraically equivalent to Haberman's method but is derived using different assumptions. Because of the algebraic equivalence of the methods, Brennan's results matched those of Haberman.

**Factor analysis**

Several researchers have used factor analysis when examining subscores (e.g. Sinharay et al., 2007; Stone, Ye, Zhu, & Lane, 2010). In this case the number of factors determines the number of subscores with added value. Sinharay et al. (2007) conducted an exploratory factor analysis on a test with six subtest scores given to prospective and practicing teacher's aides. The test was made up of two subtests in the areas of reading, writing, and mathematics, and no added subscore value was found because the results of their analysis suggested only one

dominant factor. Stone et al. (2010) used an exploratory factor analysis to determine the extent to which a one-factor model described a mathematics tests with four proposed subscores – numeric, algebraic, geometric, and quantitative – given to eighth-graders in the United States. They found one dominant factor and thus no added subscore value. They noted, however, that the correlations among the proposed subscores indicated that each subscore contributed some unique variance to the examinees' scores.

**Augmented subscores**

Wainer and colleagues (2001) applied a method intended to stabilize subscores "by augmenting data from any particular subscale with information obtained from other portions of the test" (p. 344) with the intention of increasing the subscore value to the test taker. They examined subscore value in two tests; the first test was composed of six subtests and the second test was composed of four subtests. All six augmented subtest scores in the first test had substantially smaller mean squared errors than the corresponding observed subtest scores and were, therefore, considered to have added value over the observed subtest scores. In the second test, using the same criterion, three of the four augmented subtest scores were found to have added value over the corresponding observed subtest scores. Haberman (2008) suggested a special case of augmented subscores, which he called weighted averages. This was further discussed by Haberman, Sinharay and Puhan (2009) in the context of reporting subscores to institutions. Sinharay (2010) conducted studies on operational and simulated data and found that weighted averages had added value more often than observed subscores. The simulation studies showed that observed subscores generally needed to consist of at least 20 items to have added value. Also, if the disattenuated intercorrelations were higher than 0.9, the observed subscores rarely had added value no matter what the test length was. Weighted averages, in contrast, often had added value even if the test length was 10 items as long as the disattenuated intercorrelations were less than 0.85. For test lengths of at least 20 items, weighted averages often had added value as long as the disattenuated intercorrelations were less than 0.95. Sinharay (2010) also found that Wainer and colleagues' augmented subscores and Haberman's weighted averages performed very similarly as predictors of the true subscores.

**Four-parameter beta-binomial model**

Lord (1965) proposed applying a four-parameter beta-binomial model (a strong true-score model) when lengthening and equating tests. The purpose of applying this model is to examine if different tests measure the same psychological construct, in which case their true

scores would have a "perfect functional relation for any group of examinees" (Lord, 1965, p. 245). Lord suggested that after fitting the four-parameter beta-binomial distribution to observed bivariate data, the presence of a functional relation of the true scores should be determined using a chi-squared test where rejection of the null hypothesis would indicate that the tests are measuring the same trait. Hanson (1989) used the four-parameter beta-binomial model to evaluate the assumption that the ACT and the P-ACT+ measured the same construct, that is, if their true scores were functionally related, and concluded that the assumption "appears to be reasonably well met" (p. 66). Harris and Hanson (1991) applied the model to the subtest scores within the two sections, English and Mathematics, of the P-ACT+ to examine if they provided information different from the section scores. They found that the fitted bivariate distributions provided an adequate fit and thus came to the same conclusion as Hanson (1989). This in turn led to the conclusion that the information provided by the subtest scores was not distinct from that of the section scores.

Although the beta-binomial model has been applied in previous studies of subscore value, we did not use it in this study. This was because this model has been applied only to tests with two subscores, and there is no guidance or theoretical basis in the literature for analyzing tests with more than two subscores.

**Multidimensional item response theory (MIRT)**

Haberman and Sinharay (2010) applied a method of using multidimensional item response theory (MIRT) to determine subscore value in data obtained from five tests used for teacher certification. The tests were made up of multiple-choice items in the areas of elementary education, reading, writing, mathematics, and foreign languages. Their findings showed that the use of MIRT provided overall better estimates of the true subscores than those estimates obtained from observed subscores. This does not imply that the observed subscores lacked added value, but rather that the MIRT-based subscores could be reported instead. In all cases, the subscores obtained using MIRT were more useful also better than those obtained using unidimensional IRT (UIRT).

**DIMTEST and DETECT**

Stout (1987) proposed a nonparametric IRT-based approach, DIMTEST, to assess latent trait unidimensionality. He conducted a 6,300 trial Monte Carlo simulation study to investigate his proposed significance test of the unidimensionality of a set of test items and found that the method was applicable for its intended use. DIMTEST and DETECT were used by Ackerman and Shu (as cited in Sinharay, Puhan, & Haberman, 2011) to examine the

usefulness of subscores on a fifth-grade assessment, and they found none of the subscores to be useful.

The DETECT procedure is a nonparametric MIRT-based method of assessing test dimensionality using factor analytical concepts (Zhang & Stout, 1999). DETECT searches for dimensionally homogenous clusters and produces an index value that indicates the amount of multidimensionality present in the test (Zhang & Stout, 1999). Based on arguments presented in Stout (1990), Stout et al. (1996) conducted preliminary simulation studies and found that DETECT values near or less than 0.1 were indicative of essential unidimensionality in a test and that values greater than 1.0 were indicative of sizeable multidimensionality. These results were corroborated in simulation studies by Zhang and Stout (1999).

The interpretation of DETECT values was modified by Roussos and Ozbek (2006) who used several simulation studies and real data analyses to create a similar, but more detailed, scale. They found that DETECT values below 0.2 indicate approximate unidimensionality, values between 0.2 and 0.4 indicate weak to moderate multidimensionality, values between 0.4 and 1.0 indicate moderate to strong multidimensionality, and values above 1.0 indicate strong multidimensionality. Zhang and Stout (1999) applied DETECT to four passages in two analytical reasoning sections of an administration of the Graduate Record Examinations and found four clusters that corresponded perfectly to the four passages. They also applied DETECT to four reading comprehension passages in an administration of the Law School Admission Test. There they found three clusters that corresponded to two passages requiring knowledge of science, one requiring knowledge of economics, and one requiring understanding knowledge of Navajo weaving. The DETECT index values obtained in the two studies were 0.799 and 0.709 indicating moderate to strong multidimensionality within the sections of both tests that were analyzed.

**Hypotheses**

The first hypothesis is that the compared methods will yield similar results when applied to the tests because all methods involve the assessment of test dimensionality to some extent. The second hypothesis is that the subtest scores will not have added value. This is because the subtest scores have fairly large correlations with the total score (Test A) and their respective section scores (Test B) as can be seen in Tables 1 and 2. The third hypothesis is that the two section scores (Test B) will have added value over the total score due to their relatively moderate intercorrelations (Table 2).

## Method

### Participants and Data

The two tests considered in this study were an old and a newly revised version of the Swedish Scholastic Assessment Test (SweSAT) that is used for selection and admission to higher education in Sweden. In many ways the old and the new versions of the test can be viewed as the same test because they have the same purpose, the same target construct, the same type of items (i.e., all multiple choice), and both are voluntary to take. However, we argue that the two versions are different enough to provide a basis for informative results in the type of study conducted here. The old version was designed to give only a total score that was a composite of five subtest scores. The newly revised version, which was first administered in October 2011, was designed to give a quantitative score and a verbal score (both of which are composites of four subtest scores) that are separately scaled and equated in a manner similar to that of the SAT. However, with both test versions the subtest scores have also been reported to test takers, and this is an action that has been supported by little empirical evidence and, therefore, should be examined more thoroughly.

The old version of the SweSAT (for sake of brevity hereafter referred to as Test A) consisted of the subtests DS (data sufficiency; 22 items), DTM (diagrams, tables, and maps; 20 items), WORD (vocabulary; 40 items), READ (Swedish reading comprehension; 20 items), and ERC (English reading comprehension; 20 items), which gave a total of 122 items. The new version of the SweSAT (hereafter referred to as Test B) had a quantitative section that consisted of the subtests DS (12 items), DTM (24 items), and the new subtests XYZ (mathematical problem solving; 24 items) and QC (quantitative comparisons; 20 items). The verbal section consisted of WORD (20 items), READ (20 items), ERC (20 items), and the new subtest SEC (sentence completion; 20 items). Each section thus had 80 items for a total of 160 items.

The participants in this study were test takers of Test A in the spring of 2010 (test form A1, $n_{A1}$ = 57,050) and in the fall of 2010 (test form A2, $n_{A2}$ = 40,662), and test takers of Test B in the fall of 2011 (test form B1, $n_{B1}$ = 40,431) and in the spring of 2012 (test form B2[1], $n_{B2}$ = 56,358). Of those taking test A1, 51.0% were female and the age range was 14–76 years with 80% being between the ages of 18 and 24. The other test forms, A2, B1, and B2, had gender and age distributions similar to A1. Table 1 and Table 2 show reliabilities ($\alpha$'s) and intercorrelations for A1and B1, respectively. The data used were the test takers' scores on each item.

Table 1. *Reliabilities (diagonal, in bold), intercorrelations (below the diagonal), and disattenuated intercorrelations (above the diagonal) between subtests in Test A1.*

| | DS | DTM | WORD | READ | ERC | Total test |
|---|---|---|---|---|---|---|
| DS | **.79** | .88 | .49 | .67 | .67 | .86 |
| DTM | .67 | **.74** | .57 | .70 | .71 | .92 |
| WORD | .41 | .46 | **.87** | .81 | .75 | .93 |
| READ | .51 | .52 | .65 | **.74** | .88 | .99 |
| ERC | .53 | .54 | .62 | .67 | **.79** | .97 |
| Total test | .74 | 76 | .84 | .82 | .83 | **.93** |

Table 2. *Reliabilities (diagonal, in bold), intercorrelations (below the diagonal), and disattenuated intercorrelations (above the diagonal) between subtests in Test B1.*

| | XYZ | QC | DS | DTM | WORD | READ | SEC | ERC | Q | V | Total test |
|---|---|---|---|---|---|---|---|---|---|---|---|
| XYZ | **.83** | .95 | .83 | .65 | .26 | .47 | .42 | .47 | 1.02 | .42 | .82 |
| QC | .75 | **.76** | .91 | .74 | .29 | .51 | .46 | .54 | 1.07 | .47 | .87 |
| DS | .61 | .64 | **.65** | .80 | .43 | .62 | .58 | .62 | 1.03 | .59 | .92 |
| DTM | .49 | .53 | .53 | **.68** | .49 | .63 | .61 | .66 | .96 | .63 | .91 |
| WORD | .21 | .23 | .31 | .36 | **.80** | .82 | .96 | .76 | .38 | 1.02 | .80 |
| READ | .34 | .35 | .39 | .41 | .58 | **.63** | .91 | .88 | .58 | 1.07 | .94 |
| SEC | .34 | .35 | .41 | .44 | .75 | .62 | **.76** | .85 | .54 | 1.07 | .92 |
| ERC | .38 | .41 | .44 | .47 | .59 | .61 | .64 | **.75** | .60 | 1.01 | .92 |
| Q | .89 | .89 | .79 | .76 | .32 | .44 | .45 | .50 | **.91** | .55 | .94 |
| V | .37 | .39 | .45 | .49 | .87 | .81 | .89 | .83 | .50 | **.91** | .94 |
| Total test | .72 | .74 | .72 | .72 | .69 | .72 | .78 | .77 | .87 | .87 | **.93** |

**Software**

Four different software programs were used for the analyses. *SPSS Statistics 20* was used to estimate the statistics necessary for the application of Haberman's method, Utility Index, and augmented subscores; *FACTOR* (Lorenzo-Seva & Ferrando, 2011) was used for the factor analyses; *DIMPACK* (William Stout Institute for Measurement, 2006) was used for DIMTEST and DETECT; and *MIRT* (Haberman, 2012) was used to perform the MIRT-based

analyses. *DIMPACK* had a restriction of 7,000 cases and 150 variables that led to randomly selected subsamples being used for the DIMTEST and DETECT analyses. We also used these samples for some of the analyses of Test A using the *MIRT* software to apply a five-factor MIRT to the data. The reason for this was that the analyses could not be completed for full datasets due to issues with memory allocation. Furthermore, we used the default settings in *MIRT* that included a two-parameter logistic model in both the MIRT and UIRT cases and a between-item model in the MIRT case.

When applying the DIMTEST and DETECT methods, 14 items from B1 and B2 were omitted from the analysis due to software limitations. The *DIMPACK* software allows a maximum of 150 items and this meant that at least five items from each section had to be omitted from the analysis. The omitted items were selected at random within each subtest. In the quantitative section two items were omitted from XYZ, DTM, and QC, and one item was omitted from DS. This was determined based on DS having considerably less items than the rest of the subtests in the quantitative section making the omission of items proportionally similar among the four subtests. To ensure comparable results for the two section scores, seven items were removed from the verbal section as well. Because the subtests in the verbal section had an equal number of items, one item was omitted from one subtest selected at random (WORD) and two items were omitted from the remaining subtests.

When applying the DIMTEST method, the mean values from the analyses of three random samples of only 500 cases were used instead of the maximum of 7,000 supported by the *DIMPACK* software. This was due to problems with type I error rates with larger sample sizes and because a sample size of 500 was used as a lower bound in a previous study (Seo & Roussos, 2010).

The equations used in each of the methods below have been intentionally omitted to improve the readability of the paper. The equations for each method can be found in the citations provided at the end of the description of each method.

**Haberman's method**

Haberman (2008) proposed a CTT-based method for examining whether subscores have added value over total scores. This method is based on the concept that there is value in reporting a certain subscore if the observed subscore is a more reliable predictor of the true subscore than the observed total score. Assuming that $S_s$ and $S_x$ are the relevant predictors based on the subscore and the total score, respectively, Haberman suggested using the proportional reduction in mean squared error (PRMSE) of the predictors compared to the

mean squared error (MSE) of the trivial predictor $E(S)$ as a criterion for comparing predictors of true subscores. The MSE of a predictor is reasonable to use in this context because a larger MSE implies more error in the decisions made on the basis of that score, and it is reasonable to use the trivial predictor $E(S)$ as a baseline because it gives the same prediction of the true subscore for all examinees (i.e. the average subscore). From this, it can be shown that the PRMSE for the predictor $S_s$, $PRMSE_s$, is simply the subscore reliability. Similarly, the PRMSE for the predictor $S_x$, $PRMSE_x$, is a quantity that can be thought of as the reliability of the observed total score as a measure of the subtest construct. For a subscore to have added value, therefore, $PRMSE_s$ must be larger than $PRMSE_x$. For more details about computation of the PRMSE's, consult Haberman (2008), Haberman (2009) or Lyrén (2009).

**Utility Index**

The Utility Index was suggested by Brennan (2011) as a measure of "the extent to which the borrowed information enhances reliability" (p. 1) using a method based on CTT that is algebraically equivalent to the method of Haberman (2008). The difference from Haberman's method is in the underlying assumptions and in the presentation of the results. The Utility ($U$) is the same as Haberman's $PRMSE_x$, which can be seen in Table 3. In addition to $U$, Brennan suggests using the Relative Utility ($\tilde{U}$), which is the proportional change in subtest length needed for the subscore reliability to equal $U$. Values >1 imply an increase in subtest length and values <1 imply a decrease in subtest length. Brennan also suggests using a $k$ statistic, which gives the number of items to add to or delete from a subtest to bring that subscore's reliability equal to $U$. Both $\tilde{U}$ and $k$ provide for a more detailed analysis of subscore value than just $U$. Because of the algebraic equivalence to Haberman's method, the results will always be identical as far as determining whether the subscores have added value or not. For details about the computation, consult Brennan (2011).

**Factor Analysis**

A factor analysis seeks to examine if the variables in a test or other instrument can be grouped together into a smaller number of variables, called factors, based on item covariance. The factor analysis can be either exploratory or confirmatory depending on the strength of the underlying theory from which the test is constructed. Exploratory factor analyses were carried out in this study in accordance with the method described by Sinharay et al. (2007) and Stone et al. (2010). Because Test A only had two hierarchical levels (subtest score and total score), test forms A1 and A2 were only analyzed on the item level. Because Test B had three

hierarchical levels (subtest score, section score, and total score), test forms B1 and B2 were analyzed on both the item level and the subtest level.

The number of factors indicated by a factor analysis represents the number of scores that can be considered to have value to the test taker. For a subscore to be considered to have added value, the items that make up the subscore need to coincide with the items that make up a factor (Sinharay et al., 2007). Both orthogonal and oblique rotations were performed, but the orthogonal results were used in accordance with Thompson (2005) who points out that "unless oblique results are substantially better than orthogonal results, orthogonal results may be preferred" (p. 33) because more parameters need to be estimated when using oblique rotations.

**Augmented Subscores**

The method of augmenting subscores involves estimating an individual's true subscore on a given subscale using information from not only the observed subscore but from all subscores in the test (Wainer et al., 2001). According to Wainer and colleagues, this stabilizes the estimate of the true subscore and the augmented subscore will, therefore, be a better estimate of the true subscore compared to the observed subscore. For augmented subscores to have added value, the PRMSE of the augmented subscores ($PRMSE_a$) should be substantially larger than both $PRMSE_s$ and $PRMSE_x$ (Haberman & Sinharay, 2012).

It should be noted that if the results indicate that augmented subscores have added value this does not imply that observed subscores lack added value (and vice versa) but rather that the augmented subscores provide more information than the observed subscores. For details about computation, consult Haberman (2008), Haberman (2009), and Haberman and Sinharay (2012).

**Multidimensional Item Response Theory (MIRT)**

Sinharay, Puhan, and Haberman (2011) report on two general approaches to the use of MIRT models in this study. One approach is to fit both a MIRT model and a UIRT model to the test score data and perform a statistical test of the relatively better fit of the MIRT model over the UIRT model. However, as is the case with all statistical tests this approach is sensitive to sample size. The other approach, which was the one used in this paper, was proposed by Haberman and Sinharay (2010) and can be viewed as an MIRT version of Haberman's CTT-based method. The general idea behind this approach is to compare the PRMSE's of predictors based on different types of scores. For comparison with CTT-based scores, they used the previously described $PRMSE_s$, $PRMSE_x$, and $PRMSE_a$. They then

proposed an MIRT-based PRMSE derived directly from scores on the theta-scale ($PRMSE_m$) and one derived from scores estimated from the test characteristic curve ($PRMSE_{mTCC}$). They also used the UIRT equivalents of these PRMSEs (here denoted as $PRMSE_u$ and $PRMSE_{uTCC}$, respectively). The difference between the UIRT- and the MIRT-based PRMSE's is that in the MIRT case the model is fitted to all subtests at once, but in the UIRT case the model is fitted to each subtest individually. The reason why $PRMSE_{mTCC}$ and $PRMSE_{uTCC}$ were proposed was a concern about comparing $PRMSE_m$ or $PRMSE_u$ with $PRMSE_s$, $PRMSE_x$, or $PRMSE_a$ because theta scores appear to be on a different scale than true scores (Haberman & Sinharay, 2010). However, for the application of the MIRT method in our study we chose to look only at the theta-derived PRMSE's. This was because when Haberman and Sinharay applied this methodology to real test data, $PRMSE_{mTCC}$ and $PRMSE_{uTCC}$ were always equal to or larger than $PRMSE_m$ and $PRMSE_u$. Given that one would usually want to have an indication of the added value of an IRT-based score over a CTT-based score, it seemed reasonable to use the more conservative IRT-based PRMSE. For details about computation of the PRMSE's, consult Haberman and Sinharay (2010).

Haberman and Sinharay explain that the estimated $PRMSE_m$ is similar in spirit to the reliability index for each of the dimensions described in Adams, Wilson, and Wang (1997), and that the estimated $PRMSE_u$ is similar to the marginal reliability described in Thissen, Nelson, and Swygert (2001). Furthermore, when it comes to comparing the different PRMSE's, Haberman and Sinharay note that (a) a comparison of $PRMSE_a$ with $PRMSE_m$ will reveal if the MIRT-based subscores are more accurate than the CTT-based subscores, (b) a comparison of $PRMSE_m$ and $PRMSE_u$ will reveal the gain from using an MIRT model over a UIRT model, and (c) a comparison of $PRMSE_s$ with $PRMSE_u$ will reveal the gain from using UIRT-based subscores rather than the observed subscores.

As with augmented subscores, it should be noted that if the results indicate that MIRT-based subscores have added value this does not imply that observed subscores lack added value (and vice versa). Instead this indicates that the MIRT-based subscores provide more information than the observed subscores.

**DIMTEST**

The DIMTEST procedure is a nonparametric, IRT-based method for investigating the assumption of unidimensionality in a test. It was originally developed by Stout (1987) and is nonparametric in the sense that it does not estimate parameters of item response functions. DIMTEST conducts hypothesis tests of two sets of items. One of the tests is called the

assessment subtest (AT), which is the subtest in focus, and one is called the partitioning subtest (PT), which is made up of all or some of the remaining items in the test. This method tests the null hypothesis that there is a dimensional similarity between the AT and the PT, and rejection of the null hypothesis indicate added value for the subscore that makes up the AT. To maximize the possibility of rejecting the null hypothesis, the items in the AT should be dimensionally similar to each other while being dimensionally different from the PT. In this study, on the subscore level, the AT is made up of the items in one of the subtests, and for Test A the PT is made up of all remaining items in the test. For Test B the PT is made up of all remaining items in the same section as the subtest. For the section level analyses, the AT is one section and the PT is the other section. For details about the computations, consult Stout (1987).

**DETECT**

The DETECT procedure, short for Dimensionality Evaluation To Enumerate Contributing Traits, is nonparametric in the same sense as DIMTEST and estimates the number of dimensions in a test using cluster analysis. It also estimates the "amount" of multidimensionality, measured as the distance of the dispersion of items away from the test composite, and states which dimension that each item predominantly measures (Zhang & Stout, 1999). The analyses can be either exploratory or confirmatory and exploratory analyses were applied in this paper. The number of dimensions considered to be measured by a test is the number that maximizes the DETECT statistic. According to Zhang and Stout (1999), only sizeable clusters should be judged to be actual dimensions and small clusters should be considered as random statistical noise and ignored. In their example, a cluster of two items in a 40-item test was considered random noise but a cluster of three was considered sizeable. They suggested that when specifying the number of items that make up a random noise cluster, the number should be proportional to the total number of items in the test. For details about computations used in the DETECT procedure, consult Stout (1990) and Zhang and Stout (1999).

## Results

The relevant statistics for Haberman's method, Utility Index, augmented subscores, DIMTEST, and MIRT are presented first because these results are provided in the same table (Table 3). The results for factor analysis and DETECT come at the end of this section.

### Haberman's Method and Utility Index

The PRMSE's based on Haberman's method for each subtest score and section score are shown in Table 3. It can be seen that in both test forms of Test A that all subtest scores except READ have a $PRMSE_s$ larger than $PRMSE_x$ and, therefore, have added value. In B1 the subtest scores for XYZ and DTM have added value and in B2 the subtest scores for DTM and WORD have added value. The quantitative and the verbal section scores have added value in both B1 and B2.

The results obtained with the Utility Index are also shown in Table 3. As stated previously, these are identical to those obtained with Haberman's method in terms of whether the subscores have added value or not. In this case a $\tilde{U}$ larger than 100 corresponds to $PRMSE_x$ being larger than $PRMSE_s$. In A1 and A2 it would take 7 and 13 additional READ items, respectively, as indicated by *k*, for READ to have added value. In Test B, the length of several subtests would need to be at least doubled for their corresponding subtest scores to have added value.

### Augmented Subscores

The $PRMSE_a$ values for each augmented subtest score and for each augmented section score are shown in Table 3. All of the augmented subtest scores in Test A had a substantially larger $PRMSE_a$ than both the $PRMSE_s$ and $PRMSE_x$ of the corresponding observed subscores. In Test B, all of the augmented subtest scores, except for QC, READ, and SEC, had a substantially larger $PRMSE_a$ than both the $PRMSE_s$ and $PRMSE_x$ of the corresponding observed subtest scores. The augmented section scores did not have a substantially larger $PRMSE_a$ than both $PRMSE_s$ and $PRMSE_x$ of the corresponding observed section scores. This implies that the augmented subtest scores in general can be viewed as containing more information than the observed subtest scores with the exception of QC, READ, and SEC in Test B. This also implies that the augmented section scores in general cannot be viewed as containing more information than the observed section scores.

### Multidimensional item response theory

$PRMSE_m$ and $PRMSE_u$ for each subscore are shown in Table 3. For all subtest scores in both Test A and Test B, $PRMSE_m$ was larger than $PRMSE_u$ and this means that the MIRT-

based subtest scores had added value over the UIRT-based subtest scores. Also, for all subtest scores in Test A, $PRMSE_m$ was the largest of the five PRMSE's and this means that the MIRT-based subtest score in this test had added value over any other type of subtest score. For the subtest scores in Test B, $PRMSE_a$ was occasionally equal to or larger than $PRMSE_m$ indicating that the advantage of MIRT-based subtest scores over CTT-based subtest scores was not as evident when looking at this test. In regards to the section scores in Test B, there was no major difference between $PRMSE_m$ and $PRMSE_u$ and these were only somewhat, if at all, higher than $PRMSE_s$.

In only a few instances was $PRMSE_s$ larger than or equal to $PRMSE_u$. These were ERC in A1, DS in A2, and WORD, SEC, and ERC in B2. This implies that IRT-based subtest scores in general can be viewed as containing more information than the CTT-based subtest scores.

**DIMTEST**

The *p*-values for all subtest scores with the DIMTEST method (Table 3) for Test A were significant ($\alpha = .05$), which means that these scores had added value. In Test B the subtests XYZ, DS, DTM, WORD, READ, and ERC as well as both section scores had added value, but QC and SEC showed a lack of added value.

Table 3. *Estimated PRMSE, Ũ (in percentages), k, and DIMTEST p-values for the subtest scores in all test forms and for the section scores in B1 and B2.*

| Test form | Statistic | Subtest scores | | | | | | | | Section scores | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | **XYZ** | **QC** | **DS** | **DTM** | **WORD** | **READ** | **SEC** | **ERC** | **Q** | **V** |
| A1 | $PRMSE_s$ | | | 79 | 74 | 87 | 74 | | 79 | | |
| | $PRMSE_x$ | | | 61 | 67 | 72 | 80 | | 77 | | |
| | $PRMSE_a$ | | | 82 | 81 | 88 | 85 | | 85 | | |
| | $PRMSE_u$ | | | 80 | 76 | 88 | 76 | | 79 | | |
| | $PRMSE_m$ | | | 86 | 85 | 89 | 88 | | 87 | | |
| | $\tilde{U}$ | | | 43 | 62 | 44 | 135 | | 86 | | |
| | $k$ | | | -13 | -8 | -23 | 7 | | -3 | | |
| | DIMTEST $p$-value | | | <.01 | <.01 | <.01 | <.01 | | <.01 | | |
| A2 | $PRMSE_s$ | | | 79 | 72 | 86 | 69 | | 79 | | |
| | $PRMSE_x$ | | | 61 | 64 | 74 | 79 | | 76 | | |
| | $PRMSE_a$ | | | 82 | 79 | 88 | 83 | | 85 | | |
| | $PRMSE_u$ | | | 79 | 73 | 87 | 70 | | 81 | | |
| | $PRMSE_m$ | | | 85 | 83 | 89 | 85 | | 88 | | |
| | $\tilde{U}$ | | | 43 | 69 | 48 | 165 | | 83 | | |
| | $k$ | | | -13 | -7 | -21 | 13 | | -4 | | |
| | DIMTEST $p$-value | | | <.01 | <.01 | <.01 | <.01 | | <.01 | | |
| B1 | $PRMSE_s$ | 83 | 76 | 65 | 68 | 80 | 63 | 76 | 75 | 91 | 91 |
| | $PRMSE_x$ | 81 | 87 | 80 | 65 | 81 | 81 | 88 | 77 | 73 | 73 |
| | $PRMSE_a$ | 87 | 88 | 83 | 76 | 86 | 82 | 89 | 83 | 91 | 91 |
| | $PRMSE_u$ | 84 | 77 | 66 | 70 | 81 | 67 | 78 | 78 | 92 | 92 |
| | $PRMSE_m$ | 89 | 89 | 83 | 78 | 88 | 85 | 90 | 85 | 92 | 92 |
| | $\tilde{U}$ | 86 | 203 | 210 | 85 | 111 | 255 | 243 | 107 | 27 | 26 |
| | $k$ | -4 | 21 | 14 | -4 | 3 | 31 | 29 | 2 | -59 | -59 |
| | DIMTEST | <.01 | .10 | <.01 | <.01 | <.01 | <.01 | .05 | <.01 | <.01 | <.0 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *p*-value | | | | | | | | | 1 |
| B2 | $PRMSE_s$ | 76 | 68 | 67 | 70 | 80 | 65 | 74 | 72 | 89 | 91 |
| | $PRMSE_x$ | 81 | 82 | 77 | 61 | 77 | 81 | 87 | 76 | 72 | 75 |
| | $PRMSE_a$ | 84 | 83 | 81 | 75 | 85 | 82 | 88 | 81 | 89 | 91 |
| | $PRMSE_u$ | 78 | 72 | 68 | 72 | 77 | 70 | 74 | 68 | 90 | 90 |
| | $PRMSE_m$ | 87 | 88 | 81 | 78 | 87 | 87 | 85 | 76 | 90 | 92 |
| | $\widetilde{U}$ | 131 | 217 | 165 | 70 | 84 | 221 | 233 | 121 | 34 | 31 |
| | *k* | 8 | 24 | 8 | -8 | -4 | 25 | 27 | 5 | -54 | -56 |
| | DIMTEST *p*-value | <.01 | .07 | <.01 | <.01 | <.01 | <.01 | .14 | <.01 | <.01 | <.01 |

*Note*. Strictly speaking, *k* is always a positive integer but because *k* indicates a decrease in subtest length when $\widetilde{U}$ <100 it is more descriptive to present *k* as a negative integer in these cases.

**Factor Analysis**

On the subtest level, the exploratory factor analysis yielded one factor in both forms of Test A and two factors in both forms of Test B. These factors were determined using the Minimum Average Partial procedure (Velicer, 1976) and confirmed via parallel analysis (Horn, 1965). In Test A, the one factor explained 64.7% (A1) and 63.4% (A2) of the variance, respectively. In Test B, the first factor explained 53.8% (B1) and 52.7% (B2) of the variance and the second factor explained an additional 18.2% and 17.3%, respectively. The rotated factor matrices showed that WORD, READ, SEC and ERC were associated with the first factor and XYZ, QC, DS, and DTM were associated with the second factor (Table 4). However, there was some ambiguity concerning DTM, which seemed to be associated with the first factor as well. Nonetheless, in general the results were in accordance with the theoretical model of Test B where WORD, READ, SEC, and ERC constituted one section (verbal) and XYZ, QC, DS, and DTM constituted another section (quantitative). From a factor analysis perspective, therefore, both section scores in both test forms were considered to have added value.

Table 4. *A varimax rotated loading matrix of a two-factor exploratory factor analysis on the subscore level in test forms B1 and B2.*

| | **Test form B1** | | **Test form B2** | |
|---|---|---|---|---|
| | Factor | | Factor | |
| Subscore | **1** | **2** | **1** | **2** |
| XYZ | .16 | **.82** | .18 | **.82** |
| QC | .17 | **.87** | .16 | **.81** |
| DS | .29 | **.70** | .31 | **.68** |
| DTM | .38 | **.56** | **.41** | **.50** |
| WORD | **.83** | .10 | **.80** | .11 |
| READ | **.68** | .28 | **.69** | .31 |
| SEC | **.84** | .24 | **.83** | .23 |
| ERC | **.69** | .34 | **.68** | .31 |

*Note*. Bold typeface indicates pattern/structure coefficients ≥.40 as suggested by Thompson (2005).

On the item level, the exploratory factor analysis yielded three factors in both forms of both tests. The factors in Test A corresponded to one quantitative factor (DS and DTM), and two verbal factors – one made up of WORD and part of READ, and one made up of ERC and the other part of READ. The factors in Test B corresponded to a quantitative factor (XYZ, QC and DS), a verbal factor (WORD, READ, and SEC), and a third factor consisting of DTM and ERC. The three factors together accounted for 16.4% (A1) and 15.7% (A2) of the explained variance in Test A and 22.0% (B1) and 20.1% (B2) in Test B. Because none of the factors coincided with the theorized subtests, all subtests in all test forms were considered to lack added value.

**DETECT**

On a subtest level, the exploratory DETECT analyses for A1 and A2 showed moderate multidimensionality (the DETECT indices for the total score on Test A were $Total_{A1} = 0.40$ and $Total_{A2} = 0.39$). For test B, the analysis showed weak multidimensionality within the test sections (the DETECT indices for the quantitative and the verbal section of Test B were $Quantitative_{B1} = 0.28$, $Quantitative_{B2} = 0.29$, $Verbal_{B1} = 0.26$, and $Verbal_{B2} = 0.21$), and the quantitative section seemed to be somewhat more multidimensional than the verbal section. However, the found clusters (dimensions) in both Test A and Test B did not correspond to the

theorized clusters, which are defined by each item's belonging to a specific subtest. Therefore, according to the DETECT analysis the subtest scores in both tests were considered to lack added value.

On the section level, the exploratory DETECT analyses showed moderate multidimensionality (the DETECT indices for the total score on Test B were $Total_{B1} = 0.59$ and $Total_{B2} = 0.46$). In all of the analyses, either three or four clusters maximized the DETECT statistic with one cluster containing mainly quantitative items, one cluster containing mainly verbal items, and one or two random noise clusters containing between one and four items. Therefore, according to the DETECT analysis the section scores of Test B were considered to have added value.

**Summary of Results**

An overview of the results from all methods is shown in Tables 5 (Test A) and 6 (Test B). The results are presented in terms of whether the subscores were found to have added value or not. It should be noted again that augmented subscores and MIRT-based subscores do not test if observed subscores have added value, but rather if there are better ways of reporting them. Therefore, the results using those methods are classified as either 'More' (the observed subscores contain more information than the augmented subscores or MIRT estimates) or 'Less' (the observed subscores contain less information).

Table 5. *The results from all methods showing whether the observed subscores in Test A have added value or not and whether they contain more information than the augmented subscores or MIRT estimates. The answers are presented as Yes (Y), No (N), More (M), or Less (L).*

| **Method** | **Subscores for Test A** | | | | |
|---|---|---|---|---|---|
| | **DS** | **DTM** | **WORD** | **READ** | **ERC** |
| Haberman's Method and Utility Index | Y | Y | Y | N | Y |
| Factor Analysis | N | N | N | N | N |
| Augmentation | L | L | L | L | L |
| MIRT | L | L | L | L | L |
| DIMTEST | Y | Y | Y | Y | Y |
| DETECT | N | N | N | N | N |

*Note*. For all methods, the same results were found for both test forms of both tests so only one answer is given.

Table 6. *The results from all methods showing whether the observed subscores and section scores in Test B have added value or not and whether they contain more information than the augmented scores or MIRT estimates. The answers are presented as Yes (Y), No (N), More (M), or Less (L).*

| **Method** | **Subtest scores for Test B** | | | | | | | | **Section scores** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | XYZ | QC | DS | DTM | WORD | READ | SEC | ERC | Q | V |
| Haberman's Method and Utility Index | Y/N | N | N | Y | N/Y | N | N | N | Y | Y |
| Factor Analysis | N | N | N | N | N | N | N | N | Y | Y |
| Augmentation | M | L | L | L | L | M | M | L | M | M |
| MIRT | L | L | L | L | L | L | L | L | L | L |
| DIMTEST | Y | N | Y | Y | Y | Y | N | Y | Y | Y |
| DETECT | N | N | N | N | N | N | N | N | Y | Y |

*Note*. Whenever the same result is obtained using the same method on both test forms only one answer is given. When different results were obtained, the first answer refers to test form B1 and the second to B2.

## Discussion

All of the methods used in this study suggest there is added value in reporting the two section scores in Test B to the test takers. In contrast, on the subtest level the results varied greatly for both tests. Factor analysis and DETECT indicated no added value for the subtest scores, but DIMTEST indicated added value for almost all subtest scores. These large differences appear to refute the first hypothesis that the methods would produce similar results. A possible explanation for this is that although the methods all involve assessment of dimensionality, they still have specific characteristics unique to that method or set of methods. Factor analysis and DETECT are methods developed specifically for assessing dimensionality. Haberman's method, Utility Index, augmented subscores, and MIRT, on the other hand, involve reliability estimation, and DIMTEST involves significance testing. Factor

analysis and DETECT could, therefore, be sensitive to the relatively high correlations between subscores (see Tables 1 and 2), while DIMTEST, like all significance tests, is sensitive to sample size and the PRMSE methods are highly dependent on score reliability. The differences in the results from the different methods illustrate the importance of caution when choosing which method to use. They also shed light on a potential risk for post hoc decisions if one were to test several methods and select the method that produced the desired result instead of selecting the most appropriate method given the data.

The second hypothesis, that the subtest scores would not have added value, was confirmed by most methods. Several methods showed a lack of added value for most subtest scores, but DIMTEST indicated added value for all subtest scores in A1 and A2 as well as for all subtest scores except QC and SEC in B1 and B2. The third hypothesis, that the section scores would have added value was confirmed by all methods, but the MIRT methodology indicated that IRT-based scores are preferable to CTT-based scores. Confirming this hypothesis might be considered an important result for Test B, the newly revised version of the SweSAT, because it provides empirical evidence for the validity of the decision to scale and equate the two sections of the SweSAT separately and to report both results to test takers. This result also supports a suggested legislative change to allow university programs the option of weighing the sections differently depending on which section is considered more important for that particular program.

When applying Haberman's method, the subscore value changed for some subtest scores between test forms in Test B. In B1, subtests DTM and XYZ provided added value but in B2 subtests DTM and WORD provided added value. This change was expected because subscore reliabilities vary between test forms due to sampling variability. This means that subscores with small differences between $PRMSE_s$ and $PRMSE_x$ (the $k$ values are close to zero) will sometimes be shown to provide added value and sometimes not. Therefore, no definitive conclusions can be drawn in regards to these subtest scores using Haberman's method. As a consequence, it would seem appropriate to recommend not reporting these subtest scores. Still, as pointed out by Haberman (2008), as long as the subtest scores are rather reliable there is no harm in reporting them. While they may not be very useful as a basis for decisions about remedial studies, they will not be detrimental either. This applies to IRT-based PRMSE's as well.

Factor analysis and DETECT are both used to determine the number of dimensions present in a test and this is how they were used in this study. They are also the only two

methods that suggest a lack of added value for all subtest scores. This might be due to all subtest scores truly lacking added value, but, because the other methods found added value, it might also be a sign that these methods are too strict when examining subscore values for tests where subtest scores are highly correlated with one another as is the case for subtest scores within each section of the SweSAT (Tables 2 and 3).

A problem we encountered in this study was that the DIMPACK software (William Stout Institute for Measurement, 2006) only allowed a maximum of 7,000 cases and 150 items, and this meant that subsamples had to be used for both the DIMTEST and the DETECT analyses. A related problem was that due to the risk of type I errors, only 500 cases per sample, less than two percent of the data, could be used for the DIMTEST analyses. Another issue was that the computational power needed to conduct the MIRT analyses seems to increase exponentially with the number of factors in the model. Due to the memory allocation errors that result from this, and were mentioned earlier, a five-factor analysis on 122 items could only be completed when the sample size was reduced significantly.

The findings in this study concerning the first hypothesis are of use to test users who want to examine subscore values before deciding which scores to report to the test takers. These findings are likely to be generalizable to other large scale tests with similar characteristics to the SweSAT. A possible limitation, and a basis for further research, is that the analyzed tests were not developed from a strict factor analytic perspective where highly separated factors are the primary goal. While internal structure is an important consideration for these tests, as well as for many other educational tests, content considerations are even more important. It is possible, or even likely, that the differences between the methods would have been different if we had analyzed tests that were developed in such a way that the different subtest scores were clearly distinct, which is the case with many psychological tests.

The findings concerning the second hypothesis primarily affect those who retake the SweSAT and who base their remediation strategy on their subtest scores from previous administrations of that test. Because the results indicated that few, if any, of the subtest scores provide added value, remediation strategies based on subtest scores might not lead to better section scores – contrary to what test takers might believe. The percentage of repeaters has increased over the past few years and repeaters constituted 46% of the test takers in A1 and A2, 47% in B1, and 49% in B2. This means that about 20,000 test takers or more per test form or at least 40,000 test takers per year are potentially affected if their remediation strategies are based on subtest scores from previous administrations of the test.

Those affected by the findings concerning the third hypothesis in this study are the policy makers, university programs, and other users of the admissions system that favors the option of differential weighting of the section scores. This study provides empirical evidence that supports their standpoint in the matter.

The results concerning the second and third hypothesis can be generalized to future administrations of the SweSAT but might not be generalized to other tests. This limitation should also be evident from the difference in results between the old and new version of the test. It is, however, important to remember that the results for ERC, when using PRMSE methods, might very well change at some point to indicate that ERC lacks added value as was explained previously.

Future research in this area might be to investigate how many of the repeated test takers actually pay attention to their subscores when deciding on remedial strategies. It might also be of interest to find out how many repeated test takers are aware of the implications of several subtest scores potentially not having added value and why these should not be used for remedial studies. A study of the effects of remedial studying using subscores that lack added value should also be considered.

## Notes

[1] The verbal section in B2 was scored using only 76 of the 80 items due to four WORD items being posted on an Internet forum before the items were to be administered. The items omitted from scoring were still used in this study because the item parameters appeared to be unaffected by the exposure of these items.

References

Adams, R. J., Wilson, M., & Wang, W. C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–23. doi: 10.1177/0146621697211001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (1999). *Standards for educational and psychological testing*. Washington, DC: AERA.

Brennan, R. L. (2011). *Utility indexes for decisions about subscores* (CASMA Report 33). Iowa City, IA: University of Iowa, Center for Advanced Studies in Measurement and Assessment. Retrieved from: http://www.education.uiowa.edu/centers/docs/casma-files/33utility-revised.pdf?sfvrsn=0

Haberman, S. J. (2008). When can subscores have value? *Journal of Educational and Behavioral Statistics, 33*(2), 204–229. doi: 10.3102/1076998607302636

Haberman, S. J. (2012). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm*. Manuscript submitted for publication.

Haberman, S. J., & Sinharay, S. (2010). Reporting of subscores using multidimensional item response theory. *Psychometrika, 75*(2), 209–227. doi: 10.1007/S11336-010-9158-4

Haberman, S. J., & Sinharay, S. (2012). Does subgroup membership information lead to better estimation of true subscores? *British Journal of Mathematical and Statistical Psycholgy*. Advance online publication. doi: 10.1111/j.2044-8317.2012.02061.x

Haberman, S. J., Sinharay, S., & Puhan, G. (2009). Reporting subscores for institutions. *British Journal of Mathematical and Statistical Psycholgy, 62*(1), 79–95. doi: 10.1348/000711007X248875

Hanson, B. A. (1989). Scaling the P-ACT+. In R. L. Brennan (Ed.), *Methodology used in scaling the ACT Assessment and P-ACT+* (pp. 57–73). Iowa City, IA: American College Testing.

Harris, D. J., & Hanson, B. A. (1991, March). *Methods of examining the usefulness of subscores*. Paper presented at the annual meeting of the National Council on Measurement in Education, Chicago, IL.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185.

Lord, F. M. (1965). A strong true-score theory, with applications. *Psychometrika, 30*(3), 239–270.

Lorenzo-Seva, U., & Ferrando, P. J. (2011). FACTOR (Version 8.02) [Computer software]. Tarragona, Spain: Rovira i Virgili University.

Lyrén, P.-E. (2009). Reporting subscores from college admission tests. *Practical Assessment, Research & Evaluation*, *14*(4), 1–10.

Monaghan, W. (2006). *The facts about subscores* (Report No. RDC-04). Princeton, NJ: Educational Testing Service.

Puhan, G., Sinharay, S., Haberman, S. J., & Larkin, K. (2010). The utility of augmented subscores in a licensure exam: an evaluation of methods using empirical data. *Applied Measurement in Education, 23*(3), 266–285. doi: 10.1080/08957347.2010.486287

Roussos, L. A., & Ozbek, O. (2006). Formulation of the DETECT population parameter and evaluation of DETECT estimator bias. *Journal of Educational Measurement, 43*(3), 215–243.

Seo, M., & Roussos, L. A. (2010). Formulation of a DIMTEST effect size measure (DESM) and evaluation of the DESM estimator bias. *Journal of Educational Measurement, 47*(4), 413–431.

Sinharay, S. (2010). How often do subscores have added value? Results from operational and simulated data. *Journal of Educational Measurement, 47*(2), 150–174.

Sinharay, S., Haberman, S. J., & Puhan, G. (2007). Subscores based on classical test theory: to report or not to report. *Educational Measurement: Issues and Practice, 26*(4), 21–28.

Stone, C. A., Ye, F., Zhu, X., & Lane, S. (2010). Providing subscale scores for diagnostic information: A case study when the test is essentially unidimensional. *Applied Measurement in Education, 23*(1), 63–86.

Stout, W. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*(4). 589–617.

Stout, W. (1990). A new item response theory modeling approach with applications to unidimensionality assessment and ability estimation. *Psychometrika, 55*(2), 293–325.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L. A., & Zhang, J. (1996). Conditional covariance-based nonparametric multidimensionality assessment. *Applied Psychological Measurement 20*(4), 331–354.

Thissen, D., Nelson, L., & Swygert, K. A. (2001). Item response theory applied to combinations of multiple-chioce and constructed-response items—approximation

methods for scale scores. In D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 293–342). Mahwah, NJ: Lawrence Erlbaum Associates.

Thompson, B. (2005). *Exploratory and confirmatory factor analysis: understanding concepts and applications*. Washington, DC: American Psychological Association.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika*, *41*(3), 321–327.

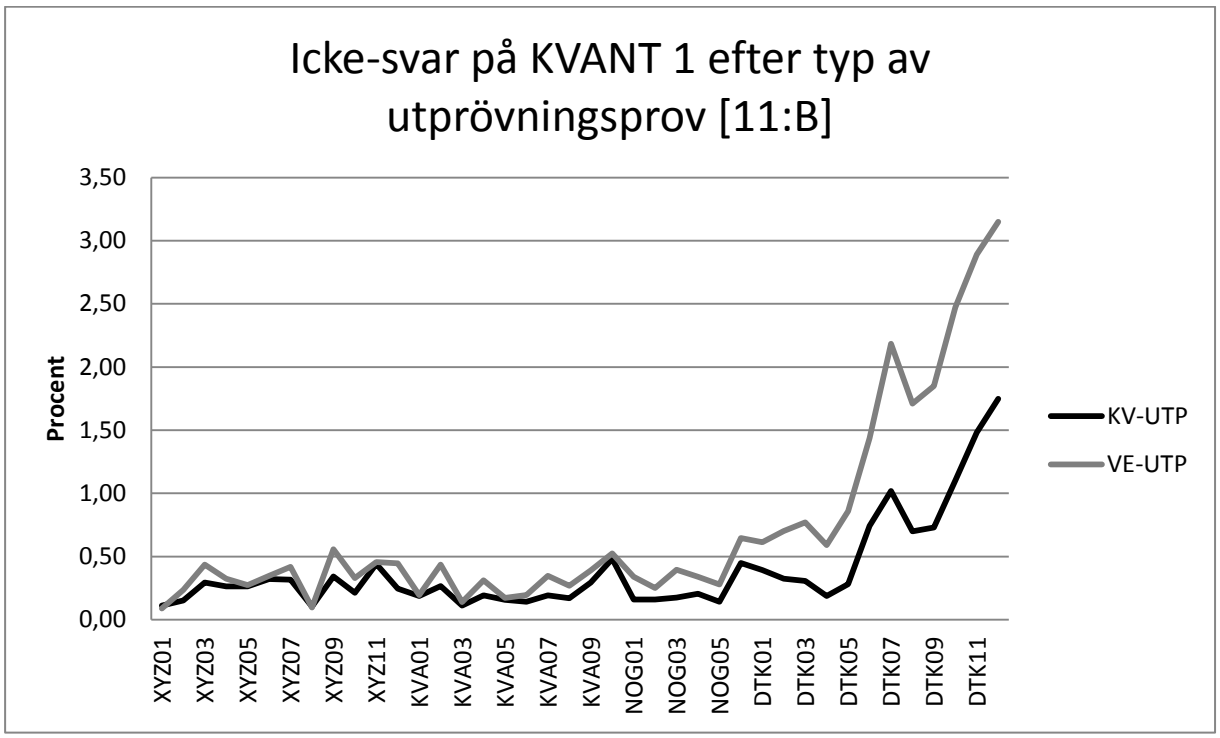
Wainer, H., Vevea, J. L., Camacho, F., Reeve, B. B., III, Rosa, K., Nelson, L., Swygert K. A., & Thissen, D. (2001). Augmented scores—"borrowing strength" to compute scores based on small number of items. In D. Thissen & H. Wainer (Eds.), *Test Scoring* (pp. 343–388). Mahwah, NJ: Lawrence Erlbaum Associates.

William Stout Institute for Measurement. (2006). Nonparametric dimensionality assessment package DIMPACK (Version 1.0) [Computer software]. St. Paul, MN: Assessment Systems Corporation.

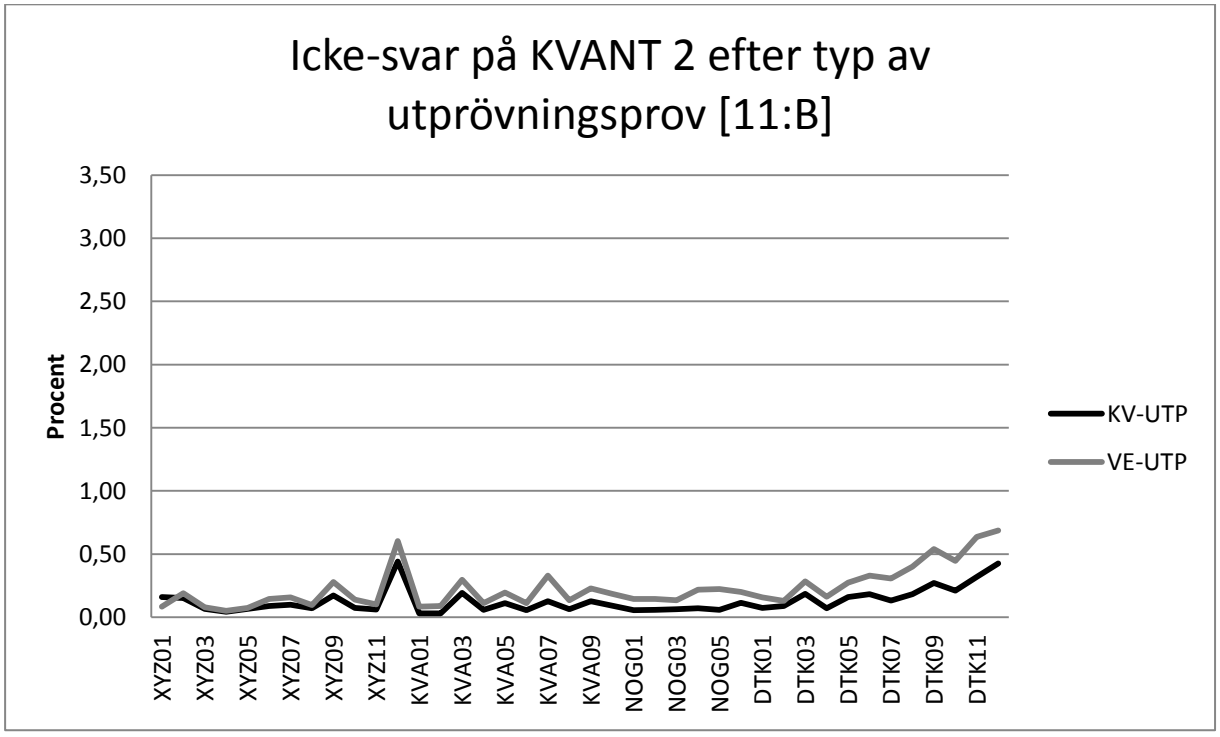
Zhang, J., & Stout, W. (1999). The theoretical DETECT index of dimensionality and its application to approximate simple structure. *Psychometrika*, *64*(2), 213–249.

# Icke-svar

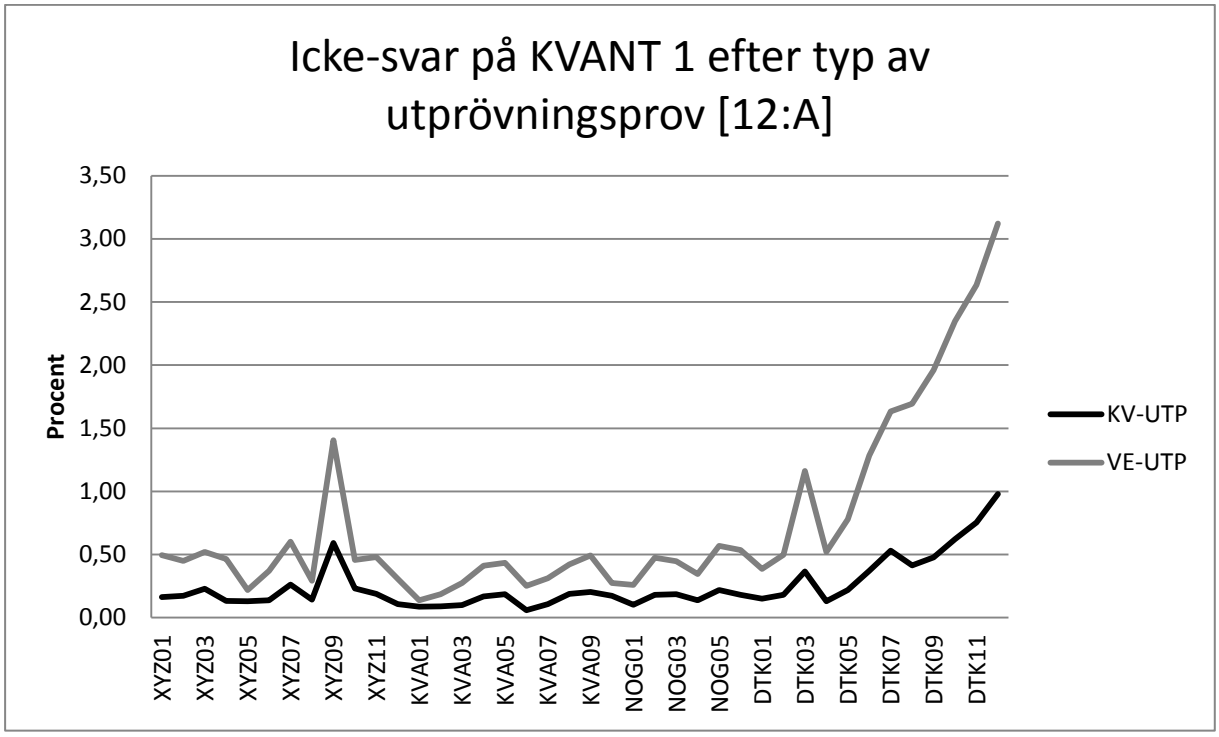
*Per-Erik Lyrén*



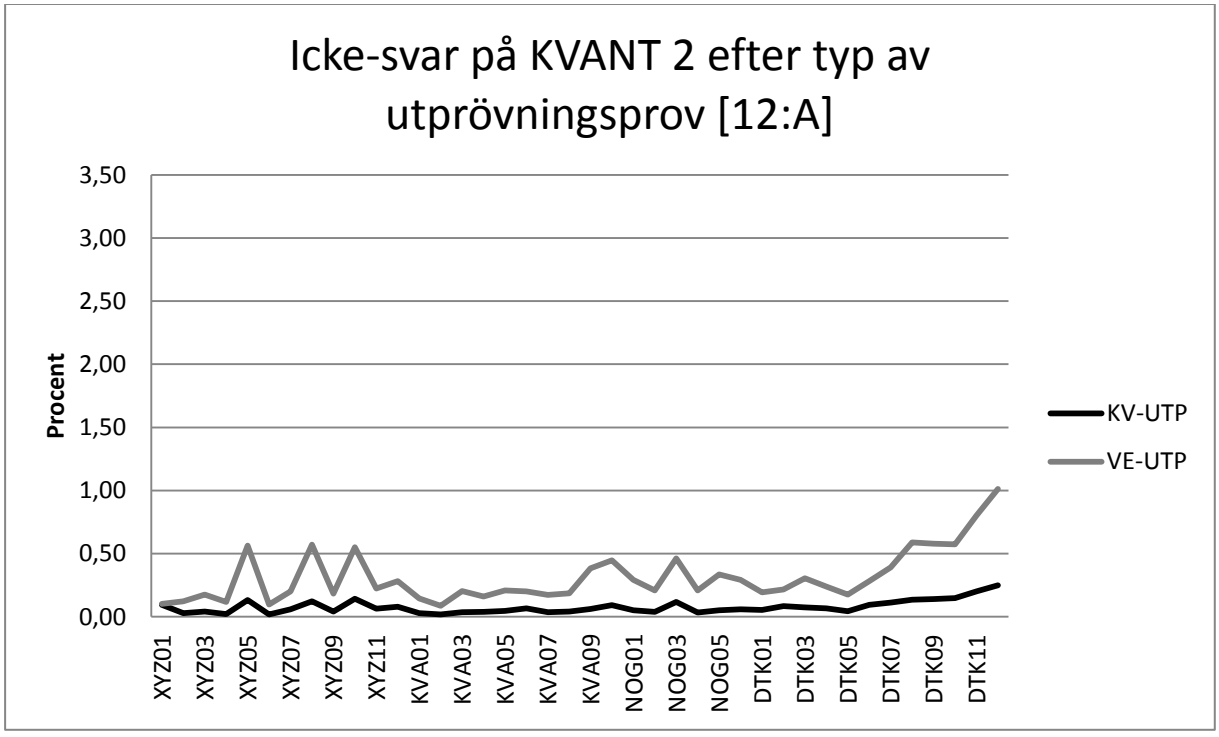

Figur B1. Andel partiellt bortfall på det *första* reguljära *kvantitativa* häftet hösten 2011 uppdelat efter vilken typ av utprövningsprov provtagare fick.



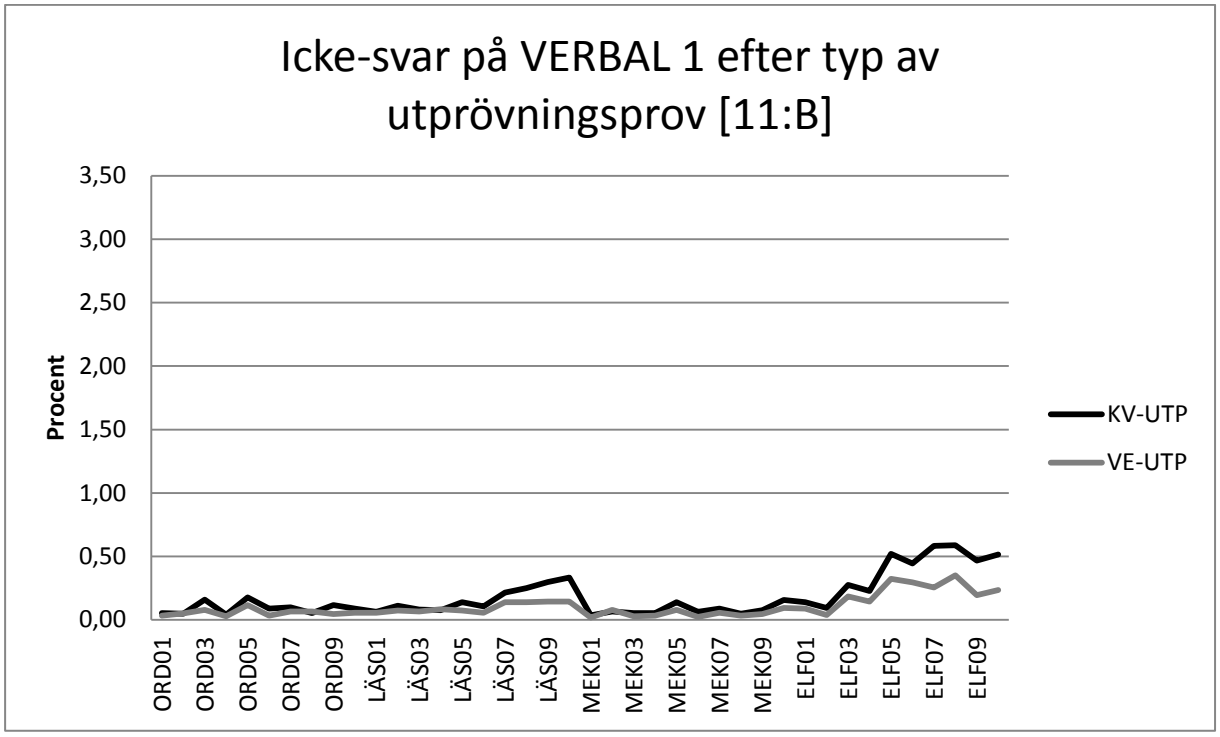

Figur B2. Andel partiellt bortfall på det *andra* reguljära *kvantitativa* häftet hösten 2011 uppdelat efter vilken typ av utprövningsprov provtagare fick.



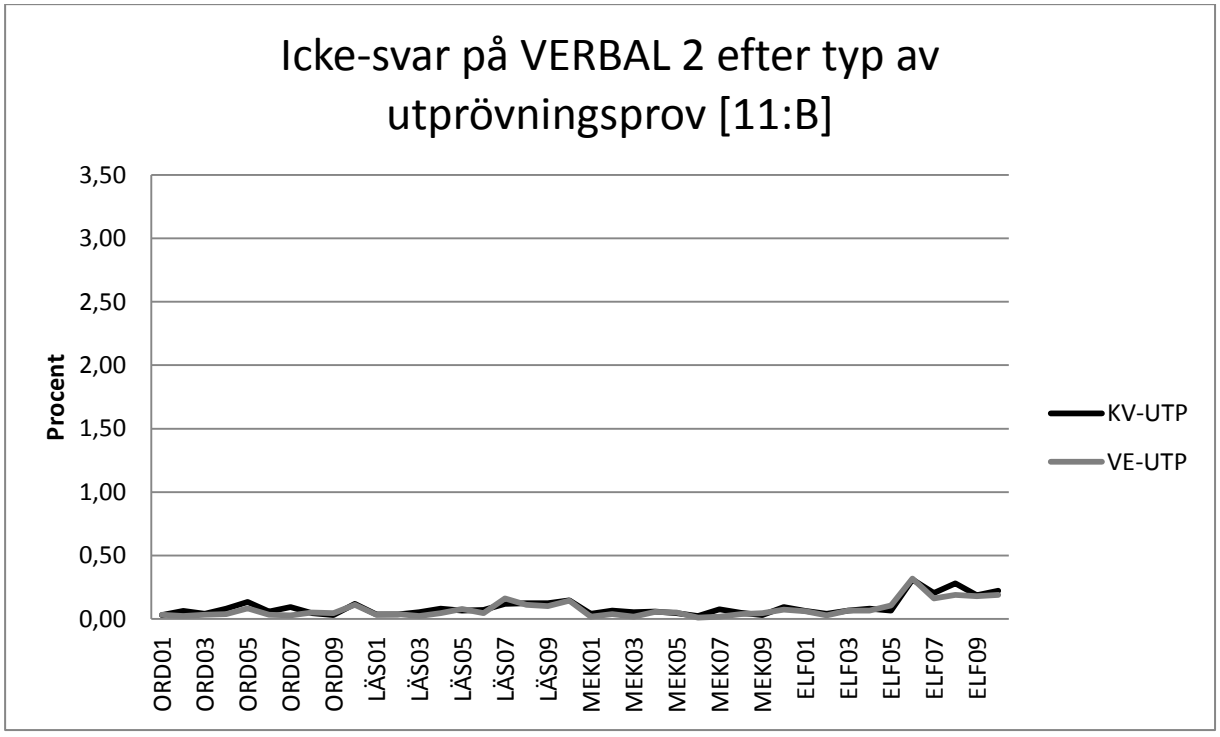

Figur B3. Andel partiellt bortfall på det *första* reguljära *kvantitativa* häftet våren 2012 uppdelat efter vilken typ av utprövningsprov provtagare fick.



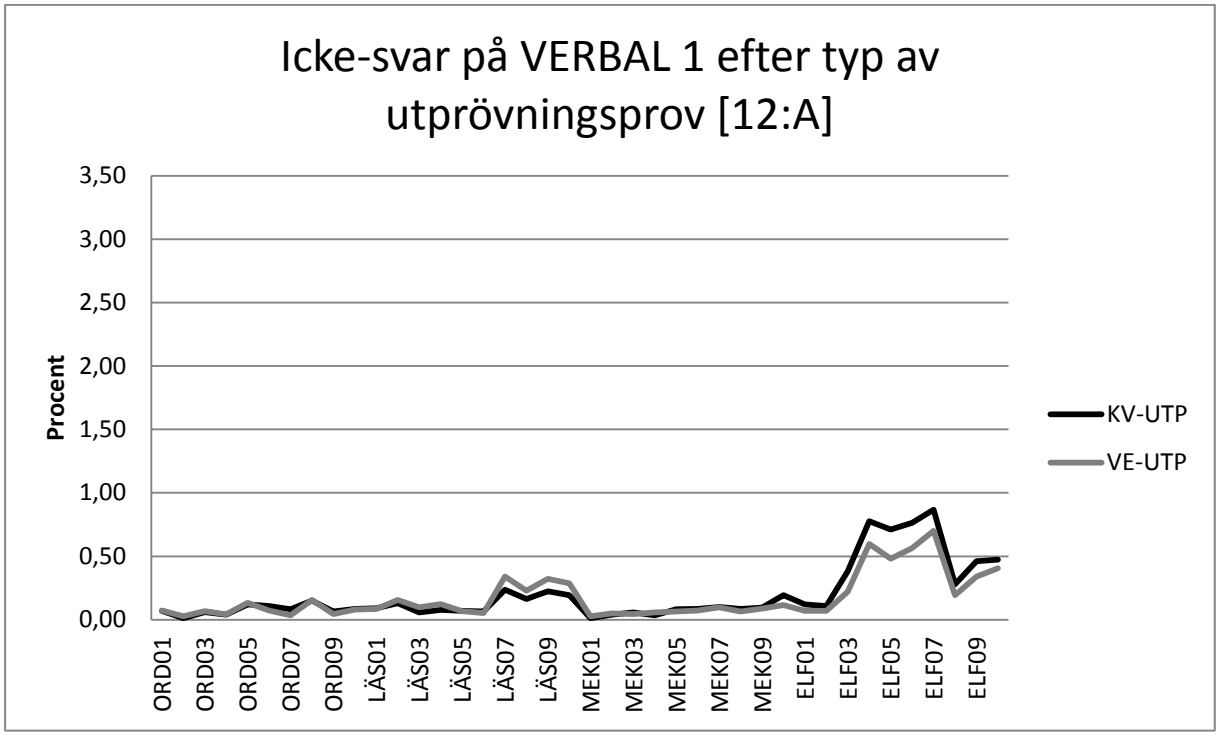

Figur B4. Andel partiellt bortfall på det *andra* reguljära *kvantitativa* häftet våren 2012 uppdelat efter vilken typ av utprövningsprov provtagare fick.



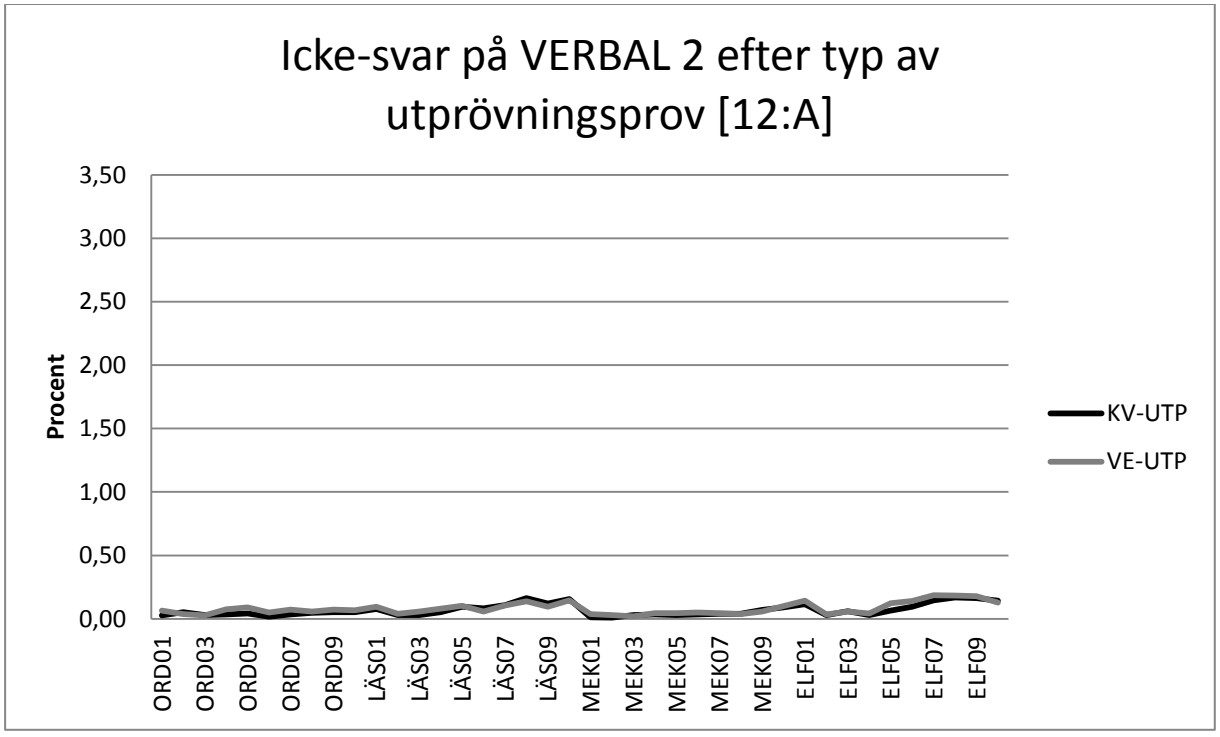

Figur B5. Andel partiellt bortfall på det *första* reguljära *verbala* häftet hösten 2011 uppdelat efter vilken typ av utprövningsprov provtagare fick.




Figur B6. Andel partiellt bortfall på det *andra* reguljära *verbala* häftet hösten 2011 uppdelat efter vilken typ av utprövningsprov provtagare fick.




Figur B7. Andel partiellt bortfall på det *första* reguljära *verbala* häftet våren 2012 uppdelat efter vilken typ av utprövningsprov provtagare fick.




Figur B8. Andel partiellt bortfall på det *andra* reguljära *verbala* häftet våren 2012 uppdelat efter vilken typ av utprövningsprov provtagare fick.

## Konsekvenser för olika grupper av provdeltagare av en ny sammansättning av högskoleprovet

*Gunilla Ögren*

### BAKGRUND

Högskoleprovet har sedan våren 1977 fungerat som urvalsinstrument för antagning till högre utbildning. Fram till våren 1991 kunde provet användas enbart av en begränsad grupp sökande nämligen de som tillhörde kvotgrupp fyra, d.v.s. var behöriga för högskolestudier genom att de var minst 25 år och hade minst fyra års arbetslivserfarenhet. År 1991 trädde nya regler för antagning till universitets- och högskolestudier i kraft. Dessa nya regler innebär att sedan höstterminen 1991 sker urvalet till studieplatser för samtliga sökande antingen utifrån betyg eller utifrån resultat på högskoleprovet.

Under åren har ett antal förändringar gjorts i provets sammansättning, den senaste hösten 2011. I tabell 1 redovisas hur provets sammansättning har förändrats från den första provversionen till den nu gällande.

**Tabell 1**. Förändringar i provets sammansättning. Delprov och antal uppgifter.

| Delprov | 1977-1979 | 1980-1991 | 1992-1995 | 1996-vt 2011 | ht 2011- |
|---|---|---|---|---|---|
| ORD | 30 | 30 | 30 | 40 | 20 |
| NOG | 20 | 20 | 20 | 22 | 12 |
| LÄS | 30 | 24 | 24 | 20 | 20 |
| DTK | 20 | 20 | 20 | 20 | 24 |
| STUF | 20 | 20 | - | - | - |
| AO | 30 | 30 | 30 | - | - |
| ELF | - | - | 24 | 20 | 20 |
| XYZ | - | - | - | - | 24 |
| KVA | - | - | - | - | 20 |
| MEK | - | - | - | - | 20 |
| Totalt | 150 | 144 | 148 | 122 | 160 |

För utförligare beskrivning av provets tidigare förändringar se t.ex. Stage & Ögren, 2001; Stage, 1995; Ögren, m.fl., 1995.

Från och med hösten 2011 består provet av två distinkta delar – en kvantitativ och en verbal – som består av 80 uppgifter vardera.

Den kvantitativa delen består av de fyra delproven XYZ, KVA, NOG och DTK. Delprovet XYZ med 24 uppgifter mäter förmågan att lösa matematiska problem; delprovet KVA med 20 uppgifter mäter förmågan att göra kvantitativa jämförelser; delprovet NOG med 12 uppgifter mäter förmågan att hantera matematiska och logiska problem; delprovet DTK med 24 uppgifter mäter förmågan att hämta och tolka information från grafiska beskrivningar.

Den verbala delen består av de fyra delproven ORD, LÄS, MEK och ELF. Delprovet ORD med 20 uppgifter mäter svensk ordförståelse; delprovet MEK med 20 uppgifter mäter förmågan att förstå ord och begrepp; delprovet LÄS med 20 uppgifter mäter svensk läsförståelse; delprovet ELF med 20 uppgifter mäter engelsk läsförståelse.

Syftet med den senaste förändringen har varit att 1) göra provet mer effektivt genom att utöka antalet uppgifter från 122 till 160, 2) få en bättre prognosförmåga genom en utökning av både antalet uppgifter och antalet delprov med kvantitativt innehåll, 3) möjliggöra en flexibel användning av provet för olika utbildningar. Detta förutsätter dock en förändring i högskoleförordningen.

Det nya högskoleprovet som togs i bruk hösten 2011 är ett resultat av ett utvecklingsuppdrag som högskoleprovsgruppen vid Institutionen för beteendevetenskapliga mätningar, Umeå universitet fick av Högskoleverket år 2006. I senare delen av uppdraget ingick även att göra en studie av konsekvenserna av förändringarna för olika undergrupper av provdeltagare. Resultatet finns redovisat i Stage & Ögren, 2010. I rapporten ges en utförlig beskrivning av det nya provet samt bakgrund och konsekvenser av dess införande.

Det nya provet har genomförts vid fyra provtillfällen och högskoleprovsgruppen har fått i uppdrag att göra en studie motsvarande den som genomfördes av försöksproven hösten 2009 och finns redovisad i Stage & Ögren, 2010. Skillnaden mellan dessa två studier är att vid förra tillfället var det fortfarande det gamla provet som gällde som reguljärt prov medan utprövningsuppgifterna utgjordes av de nya delproven. Det innebar att provdeltagarna visste vilken del som var utprövning och kunde inte förväntas ha maximal motivation. Det medförde osäkra data som i sin tur medförde osäkerhet vid analys av de nya uppgifterna/delproven. Den studie som redovisas i denna rapport är däremot en analys av det nya reguljära provet.

I föreliggande rapport redovisas provdeltagarnas resultat på de reguljära proven hösten 2011 och våren 2012. Resultaten redovisas för män och kvinnor, för olika åldersgrupper, för provdeltagare indelade efter föräldrars utbildning, för provdeltagare som genomgått olika gymnasieprogram samt för provdeltagare med svensk/utländsk bakgrund. Informationen om grupptillhörighet har hämtats från SCB. I följande redovisning anges provtillfället på hösten 2011 som 11B och på våren 2012 som 12A.

Som tidigare nämnts består provet från och med hösten 2011 av två distinkta delar – en kvantitativ och en verbal. Av den anledningen redovisas resultat för den kvantitativa delen och den verbala var för sig. Men först görs en redovisning av provdeltagargruppens sammansättning.

### *Provdeltagargruppernas sammansättning*

**Tabell 2.** Prodeltagarnas procentuella fördelning på kön, ålder och utbildning.

| **Provtillfälle** | **11B** | **12A** |
|---|---|---|
| **Kön** | | |
| Män | 49 | 50 |
| Kvinnor | 51 | 50 |
| **Ålder** | | |
| ≤ 20 år | 58 | 56 |
| 21–24 år | 25 | 28 |
| 25–29 år | 9 | 9 |
| 30–39 år | 6 | 5 |
| ≥ 40 år | 2 | 2 |
| **Utbildning** | | |
| Grund/realskola | 1 | 1 |
| Folkhögskola | 2 | 2 |
| Gymnasium ≤ 2 år | 3 | 3 |
| Gymnasium ≥ 3 år | 77 | 79 |
| Högre utbildning ≤ 80 p | 12 | 10 |
| Högre utbildning ≥ 80 p | 4 | 4 |
| Ej angivet | 1 | 1 |
| Totalt antal | 40 695 | 56 864 |

## RESULTAT

### *Resultat kvantitativ del hösten 2011 och våren 2012*

Nedan följer en redovisning av poängmedelvärde (M), genomsnittlig lösningsproportion (p) och standardavvikelsen (s) för de olika grupper som genomförde proven hösten 2011 respektive våren 2012. Även antalet provdeltagare (N) anges. Tabellerna 3 - 13 redovisar resultat för den kvantitativa delen medan tabellerna 14 - 24 redovisar resultat för den verbala delen.

**Tabell 3. Resultat för kvinnor och män samt totalt**

| **Kön** | **11B** | | | | **12A** | | | |
|---|---|---|---|---|---|---|---|---|
| | **M** | **p** | **s** | **N** | **M** | **p** | **s** | **N** |
| Kvinnor | 34.41 | .43 | 12.1 | 20 772 | 33.66 | .42 | 10.8 | 28 595 |
| Män | 41.51 | .52 | 13.8 | 19 923 | 40.38 | .50 | 12.4 | 28 269 |
| Totalt | 37.88 | .47 | 13.4 | 40 695 | 37.00 | .46 | 12.1 | 56 864 |

Vid provet hösten 2011 var medelvärdet på den kvantitativa delen för totalgruppen 37.88 och spridningen var 13.4. Skillnaden mellan män och kvinnor var 7.10 till männens fördel. Spridningen var 1.7 högre för männen än för kvinnorna.

Vid provet våren 2012 var medelvärdet 37.00 och spridningen 12.1. Skillnaden mellan män och kvinnor var 6.72 till männens fördel. Spridningen var 1.6 högre för männen än för kvinnorna.

I tabell 4 visas resultat för kvinnor och män på respektive delprov i den kvantitativa delen.

**Tabell 4. Resultat för kvinnor och män uppdelat på kvantitativa delprov**

| Kön | XYZ 24 uppgifter M | s | KVA 20 uppgifter M | s | NOG 12 uppgifter M | s | DTK 24 uppgifter M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| Kvinnor | 9.93 | 4.9 | 8.42 | 3.8 | 4.53 | 2.4 | 11.53 | 3.7 |
| Män | 11.91 | 5.3 | 10.56 | 4.5 | 5.69 | 2.7 | 13.35 | 4.1 |
| Totalt | 10.90 | 5.2 | 9.47 | 4.2 | 5.10 | 2.6 | 12.42 | 4.0 |
| | | | | | | | | |
| **12A** | | | | | | | | |
| Kvinnor | 9.33 | 4.1 | 8.04 | 3.2 | 4.47 | 2.4 | 11.82 | 3.9 |
| Män | 11.38 | 4.7 | 9.81 | 3.7 | 5.58 | 2.8 | 13.61 | 4.1 |
| Totalt | 10.35 | 4.5 | 8.92 | 3.6 | 5.02 | 2.7 | 12.71 | 4.1 |

Skillnaden mellan kvinnor och män på den kvantitativa delen vid provtillfället 11B var i genomsnitt 0.09 per uppgift till fördel för männen. Motsvarande skillnad var vid provtillfället 12A 0.08. På delprovsnivå uppvisar delprovet DTK en något mindre skillnad i resultat mellan kvinnor och män än de övriga kvantitativa delproven.

Spridningen var högst på delprovet XYZ. Sett över samtliga delprov var spridningen högre för männen än för kvinnorna.

**Tabell 5. Resultat för olika åldersgrupper**

| Ålder | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| - 20 | 39.65 | .50 | 13.5 | 23 653 | 38.31 | .48 | 12.3 | 31 629 |
| 21 - 24 | 35.93 | .45 | 12.8 | 10 371 | 35.52 | .44 | 11.5 | 15 764 |
| 25 - 29 | 34.28 | .43 | 12.4 | 3 640 | 34.94 | .44 | 11.5 | 5 309 |
| 30 - 39 | 34.75 | .43 | 13.8 | 2 233 | 35.08 | .44 | 12.3 | 2 951 |
| 40 - | 36.09 | .45 | 15.0 | 798 | 35.91 | .45 | 13.4 | 1 211 |
| Totalt | 37.88 | .47 | 13.4 | 40 695 | 37.00 | .46 | 12.1 | 56 864 |

Det är den yngsta åldersgruppen som har det högsta resultatet på den kvantitativa delen, följt av den äldsta gruppen, dvs. 40 år eller äldre som har det näst högsta resultatet. Skillnaden mellan dessa grupper var 3.56 råpoäng vid provet hösten 2011 och 2.40 råpoäng vid provet våren 2012. Eftersom mer än hälften (58 respektive 56 procent) av provdeltagargruppen är 20 år eller yngre och en fjärdedel är i åldern 21 till 24 år redovisas i tabell 6 resultat och antal provdeltagare per ålder upp till 24 år.

**Tabell 6. Resultat för åldrar och åldersgrupper**

| Ålder | 11B | | | | 12A | | | |
|---|---|---|---|---|---|---|---|---|
| | M | p | s | N | M | p | s | N |
| -17 | 43.40 | .54 | 13.1 | 2 066 | 41.24 | .52 | 12.2 | 875 |
| 18 | 40.27 | .50 | 13.5 | 10 236 | 40.84 | .51 | 11.9 | 5 732 |
| 19 | 39.10 | .49 | 13.6 | 6 424 | 37.99 | .47 | 12.3 | 17 131 |
| 20 | 37.52 | .47 | 13.1 | 4 926 | 36.86 | .46 | 12.2 | 7 891 |
| 21 | 36.64 | .46 | 12.9 | 3 794 | 36.01 | .45 | 11.6 | 5 748 |
| 22 | 35.89 | .45 | 12.7 | 2 865 | 35.42 | .44 | 11.6 | 4 507 |
| 23 | 35.43 | .44 | 12.7 | 2 114 | 35.20 | .44 | 11.3 | 3 159 |
| 24 | 34.96 | .44 | 12.6 | 1 598 | 34.93 | .47 | 11.2 | 2 350 |

Som framgår i tabell 6 utgör 18-åringarna störst andel av provgruppen på hösten 2011 och 19-åringarna utgör störst andel på våren 2012. Förklaringen till denna skillnad i andelen mellan höst och vår är att de som går ut gymnasieskolan på våren 2012 är 19 år och en stor del av dessa hade gjort provet på hösten 2011 och finns då i åldersgruppen 18-åringarna.

Det är de allra yngsta (17 år eller yngre) som uppvisar högst resultat på den kvantitativa delen. Denna grupp utgör en liten del av hela provdeltagargruppen, 5 procent vid provtillfället 11B och 1,5 procent vid provtillfället 12A. Det torde också vara de som är mest studieinriktade i sin åldersgrupp som gör provet.

**Tabell 7. Resultat för olika åldersgrupper uppdelat på de kvantitativa delproven**

| Ålder | XYZ 24 uppgifter | | KVA 20 uppgifter | | NOG 12 uppgifter | | DTK 24 uppgifter | |
|---|---|---|---|---|---|---|---|---|
| | M | s | M | s | M | s | M | s |
| **11B** | | | | | | | | |
| -20 år | 11.82 | 5.2 | 10.14 | 4.2 | 5.22 | 2.6 | 12.48 | 4.0 |
| 21-24 år | 9.77 | 4.9 | 8.78 | 3.9 | 5.03 | 2.6 | 12.35 | 4.0 |
| 25-29 år | 9.11 | 4.8 | 8.15 | 3.8 | 4.80 | 2.6 | 12.23 | 4.0 |
| 30-39 år | 9.54 | 5.2 | 8.04 | 4.1 | 4.76 | 2.6 | 12.41 | 4.4 |
| 40 år- | 10.26 | 5.7 | 8.47 | 4.4 | 4.86 | 2.6 | 12.50 | 4.9 |
| Totalt | 10.90 | 5.2 | 9.47 | 4.2 | 5.10 | 2.6 | 12.42 | 4.0 |
| | | | | | | | | |
| **12A** | | | | | | | | |
| -20 år | 11.08 | 4.5 | 9.46 | 3.6 | 5.14 | 2.7 | 12.64 | 4.0 |
| 21-24 år | 9.53 | 4.2 | 8.37 | 3.4 | 4.93 | 2.6 | 12.69 | 4.0 |
| 25-29 år | 9.16 | 4.3 | 8.01 | 3.4 | 4.82 | 2.6 | 12.95 | 4.1 |
| 30-39 år | 9.30 | 4.6 | 8.00 | 3.5 | 4.75 | 2.6 | 13.03 | 4.4 |
| 40 år- | 9.89 | 5.1 | 8.20 | 3.8 | 4.74 | 2.8 | 13.09 | 4.5 |
| Totalt | 10.35 | 4.5 | 8.92 | 3.6 | 5.02 | 2.7 | 12.71 | 4.1 |

I tabellerna 5-7 framgår att ju yngre desto högre resultat på delproven XYZ, KVA och NOG, de mest matematiskt inriktade delproven. Skillnaden i resultat mellan olika åldersgrupper är däremot försumbart för delprovet DTK som mäter förmågan att hämta och tolka information från grafiska framställningar. Det är också det delprov i den kvantitativa delen som är det minst matematiskt inriktade delprovet.

Den matematik som ligger till grund för de kvantitativa delproven är kurs A/1b i gymnasieskolan. En förklaring till att de yngre åldersgrupperna har högst resultat på de kvantitativa delproven torde vara att de är under utbildning eller nyss avslutat en utbildning och har därmed matematiken nära i tiden.

**Tabell 8. Resultat för provdeltagare i relation till föräldrarnas utbilning**

| Föräldrars utbildning | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| Gymnasium 2 år | 32.30 | .40 | 11.8 | 8 944 | 32.42 | .41 | 10.6 | 12 605 |
| Gymnasium 3 år | 35.51 | .44 | 12.5 | 5 695 | 34.73 | .43 | 11.4 | 8 193 |
| Eftergymn.utb. -2 år | 37.67 | .47 | 12.9 | 9 018 | 36.77 | .46 | 11.6 | 12 699 |
| Eftergymn.utb. 3-4 år | 40.99 | .51 | 13.4 | 13 042 | 39.86 | .50 | 12.2 | 18 142 |
| Eftergymn.utb 5 år - | 45.70 | .57 | 13.3 | 3 321 | 43.79 | .55 | 12.2 | 4 320 |
| Totalt | 37.91 | .47 | 13.4 | 40 020 | 37.04 | .46 | 12.1 | 55 959 |

Det föreligger ett tämligen starkt samband mellan föräldrars utbildning och provdeltagarnas resultat på den kvantitativa delen. Provdeltagare med föräldrar som har eftergymnasial utbildning är de som har högst resultat på provet. Skillnaden i resultat mellan den grupp provdeltagare vars föräldrar har 2-årig gymnasial utbildning och den grupp vars föräldrar har eftergymnasial utbildning 5 år eller längre var 13.40 råpoäng (0.17 per uppgift) vid provet hösten 2011 och 11.37 (0.14 per uppgift) vid provet våren 2012. I tabell 9 redovisas resultat på respektive delprov i den kvantitativa delen. Där framgår att det är de mest matematiskt inriktade delproven XYZ, KVA och NOG som uppvisar de starkaste sambanden med föräldrars utbildning. Delprovet DTK uppvisar ett något svagare samband.

**Tabell 9. Resultat på de olika kvantitativa delproven i relation till föräldrarnas utbildning**

| Utbildning | XYZ 24 uppg. M | s | KVA 20 uppg. M | s | NOG 12 uppg. M | s | DTK 24 uppg. M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| Gymnasium 2 år | 8.87 | 4.6 | 7.86 | 3.7 | 4.26 | 2.4 | 11.30 | 3.8 |
| Gymnasium 3 år | 10.04 | 4.9 | 8.85 | 3.9 | 4.65 | 2.5 | 11.98 | 3.9 |
| Eftergymnutb. -2 år | 10.74 | 5.0 | 9.42 | 4.1 | 5.05 | 2.6 | 12.46 | 3.9 |
| Eftergymnutb 3-4 år | 11.98 | 5.2 | 10.34 | 4.2 | 5.63 | 2.7 | 13.03 | 4.0 |
| Eftergymnutb. 5 år - | 13.78 | 5.2 | 11.59 | 4.1 | 6.32 | 2.7 | 14.01 | 4.0 |
| Totalt | 10.88 | 5.2 | 9.47 | 4.2 | 5.11 | 2.6 | 12.45 | 4.0 |
| | | | | | | | | |
| **12A** | | | | | | | | |
| Gymnasium 2 år | 8.77 | 4.0 | 7.71 | 3.2 | 4.21 | 2.4 | 11.72 | 3.9 |
| Gymnasium 3 år | 9.57 | 4.3 | 8.36 | 3.4 | 4.58 | 2.5 | 12.22 | 3.9 |
| Eftergymnutb. -2 år | 10.20 | 4.3 | 8.82 | 3.5 | 5.01 | 2.6 | 12.75 | 4.0 |
| Eftergymnutb 3-4 år | 11.32 | 4.6 | 9.65 | 3.6 | 5.53 | 2.7 | 13.37 | 4.0 |
| Eftergymnutb. 5 år - | 12.74 | 4.7 | 10.68 | 3.6 | 6.26 | 2.6 | 14.11 | 4.0 |
| Totalt | 10.34 | 4.5 | 8.92 | 3.6 | 5.03 | 2.7 | 12.75 | 4.0 |

I tabell 10 redovisas resultat för provdeltagare från olika gymnasieprogram.

**Tabell 10. Resultat för provdeltagare från olika gymnasielinjer/program**

| Utbildning[1]/ prov | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| 1 Samhälls | 33.58 | .42 | 11.0 | 9 064 | 33.70 | .42 | 10.0 | 9 861 |
| 2 Natur | 46.24 | .58 | 12.8 | 6 626 | 44.06 | .55 | 11.9 | 6 674 |
| 3 Teknisk | 41.31 | .52 | 11.7 | 1 544 | 40.16 | .50 | 10.5 | 1 806 |
| 4 Estetisk | 31.82 | .40 | 10.5 | 1 887 | 32.03 | .40 | 9.8 | 2 551 |
| 5 Hemkunskaps | 26.68 | .33 | 8.3 | 1 874 | 27.71 | .35 | 8.1 | 2 224 |
| 6 Yrkesinr | 33.93 | .42 | 11.9 | 1 323 | 34.58 | .43 | 10.6 | 1 801 |
| Totalt | 37.16 | .46 | 13.2 | 22 318 | 36.30 | .45 | 11.7 | 24 917 |

De provdeltagare som har gått en teoretisk utbildning på gymnasiet har högre resultat på de kvantitativa delproven än de som gått en praktiskt inriktad utbildning. De som gått naturvetenskapligt inriktat program/linje har högst resultat medan de som gått gymnasialutbildning med inriktning mot barn/fritid, handel/administration, hotell/restaurang, livsmedel har lägst resultat. Skillnad i resultat mellan provdeltagare från olika program är mindre på delprovet DTK än på de övriga kvantitativa delproven, vilket framgår i tabell 11.

**Tabell 11. Resultat för provdeltagare från olika gymnasielinjer/program uppdelat på kvantitativa delprov**

| Utbildning/ prov | XYZ 24 uppgifter M | s | KVA 20 uppgifter M | s | NOG 12 uppgifter M | s | DTK 24 uppgifter M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| 1 Samhälls | 8.82 | 4.1 | 8.08 | 3.5 | 4.61 | 2.4 | 12.07 | 3.7 |
| 2 Natur | 13.99 | 4.8 | 11.78 | 3.9 | 6.45 | 2.6 | 14.02 | 4.0 |
| 3 Teknisk | 11.80 | 4.7 | 10.59 | 3.7 | 5.73 | 2.5 | 13.20 | 3.8 |
| 4 Estetisk | 8.12 | 4.0 | 7.63 | 3.4 | 4.40 | 2.4 | 11.66 | 3.7 |
| 5 Hemkunsk | 6.72 | 3.0 | 6.27 | 2.8 | 3.55 | 2.0 | 10.14 | 3.4 |
| 6 Yrkesinr | 8.91 | 4.6 | 8.33 | 3.8 | 4.74 | 2.5 | 11.95 | 3.8 |
| Totalt | 10.33 | 5.0 | 9.18 | 4.1 | 5.14 | 2.6 | 12.52 | 4.0 |
| | | | | | | | | |
| **12A** | | | | | | | | |
| 1 Samhälls | 8.68 | 3.6 | 7.71 | 3.1 | 4.68 | 2.4 | 12.64 | 3.9 |
| 2 Natur | 12.70 | 4.2 | 10.71 | 3.5 | 6.34 | 2.6 | 14.31 | 4.1 |
| 3 Teknisk | 11.05 | 3.9 | 9.81 | 3.3 | 5.65 | 2.5 | 13.65 | 3.8 |
| 4 Estetisk | 8.22 | 3.6 | 7.31 | 3.0 | 4.34 | 2.4 | 12.15 | 3.8 |
| 5 Hemkunsk | 6.96 | 3.0 | 6.30 | 2.5 | 3.50 | 2.1 | 10.95 | 3.7 |
| 6 Yrkesinr | 9.02 | 3.9 | 8.13 | 3.3 | 4.66 | 2.5 | 12.77 | 3.8 |
| Totalt | 9.75 | 4.3 | 8.53 | 3.5 | 5.05 | 2.6 | 12.97 | 4.0 |

[1] 1 samtliga samhällsinriktade program
2 naturvetenskap, naturbruk, IB
3 tekniskt program
4 estet, hantverk, media
5 barn/fritid, handel/adm, hotell/rest., livsmedel
6 bygg, el,energi,fordon,industri

I genomsnitt är skillnaden mellan det program/linje som har högst resultat och det program/linje som har lägst resultat 0.22-0.30 per uppgift för delproven XYZ, KVA och NOG medan skillnaden mellan dessa grupper är ungefär 0.15 per uppgift för delprovet DTK.

**Tabell 12. Resultat för provdeltagare med svensk/utländsk[2] bakgrund**

| Sv/utl bakgrund | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| 1 | 33.24 | .42 | 12.7 | 2 903 | 32.27 | .40 | 11.5 | 4 219 |
| 2 | 35.14 | .44 | 12.8 | 3 466 | 33.59 | .42 | 11.3 | 5 239 |
| 3 | 40.01 | .50 | 13.3 | 236 | 39.76 | .50 | 12.8 | 350 |
| 4 | 37.54 | .47 | 13.2 | 3 916 | 36.92 | .46 | 12.0 | 5 487 |
| 5 | 38.67 | .48 | 13.5 | 30 138 | 37.90 | .47 | 12.1 | 41 529 |
| Totalt | 37.88 | .47 | 13.4 | 40 659 | 37.00 | .46 | 12.1 | 56 824 |

Ungefär 75 procent av provdeltagarna är födda i Sverige och har svenskfödda föräldrar. Det är den gruppen och gruppen födda i Sverige med en svenskfödd och en utrikes född förälder samt den lilla gruppen provdeltagare (0.5 procent) födda utomlands med minst en förälder född i Sverige som har högst resultat på den kvantitativa delen. Som framgår i tabell 13 är skillnaden mellan grupperna minst på delprovet XYZ, därefter på delprovet KVA. Skillnaden mellan gruppen födda i Sverige med svenskfödda föräldrar och gruppen utrikes födda med båda föräldrarna födda utomlands är endast 0.02-0.03 per XYZ-uppgift. För delproven NOG och DTK är skillnaden större mellan de ovan jämförda grupperna. Skillnaden ligger mellan 0.10-0.12 per uppgift.

[2] 1 utrikes födda med två utrikes födda föräldrar
2 inrikes födda med två utrikes födda föräldrar
3 utrikes födda med minst en utrikes född förälder
4 inrikes född med en inrikes och en utrikes född förälder
5 inrikes född med två inrikes födda föräldrar

**Tabell 13. Resultat för provdeltagare med svensk/utländsk bakgrund för olika delprov**

| | XYZ 24 uppgifter | | KVA 20 uppgifter | | NOG 12 uppgifter | | DTK 24 uppgifter | |
|---|---|---|---|---|---|---|---|---|
| | M | s | M | s | M | s | M | s |
| **11B** | | | | | | | | |
| 1 | 10.44 | 5.1 | 8.35 | 4.0 | 4.11 | 2.5 | 10.34 | 4.0 |
| 2 | 10.78 | 5.1 | 8.84 | 4.0 | 4.41 | 2.5 | 11.11 | 3.8 |
| 3 | 11.85 | 5.2 | 10.11 | 4.2 | 5.26 | 2.4 | 12.79 | 4.0 |
| 4 | 10.74 | 5.2 | 9.39 | 4.1 | 5.07 | 2.6 | 12.35 | 4.0 |
| 5 | 10.97 | 5.2 | 9.65 | 4.2 | 5.28 | 2.6 | 12.77 | 4.0 |
| Totalt | 10.90 | 5.2 | 9.47 | 4.2 | 5.10 | 2.6 | 12.42 | 4.0 |
| | | | | | | | | |
| **12A** | | | | | | | | |
| 1 | 9.76 | 4.5 | 8.31 | 3.4 | 4.00 | 2.4 | 10.21 | 4.0 |
| 2 | 9.99 | 4.4 | 8.44 | 3.4 | 4.24 | 2.5 | 10.93 | 3.9 |
| 3 | 11.35 | 4.9 | 9.71 | 3.6 | 5.55 | 2.8 | 13.14 | 4.2 |
| 4 | 10.35 | 4.5 | 8.90 | 3.5 | 5.02 | 2.6 | 12.66 | 4.0 |
| 5 | 10.45 | 4.5 | 9.03 | 3.6 | 5.22 | 2.7 | 13.19 | 3.9 |
| Totalt | 10.35 | 4.5 | 8.92 | 3.6 | 5.02 | 2.7 | 12.71 | 4.1 |

## Sammanfattning

Könsskillnaden på den kvantitativa delen är ungefär som förväntad. Innan antalet kvantitativa uppgifter utökades från 42 till 80 uppgifter var könsskillnaden cirka 4 råpoäng på 42 kvantitativa uppgifter, vid provet 11B och 12A med 80 kvantitativa uppgifter var den ungefär 7 råpoäng till männens fördel. Spridningen var högre för männen än för kvinnorna, något som även varit tidigare. På delprovsnivå uppvisar delprovet DTK något mindre skillnad i resultat mellan män och kvinnor än de övriga kvantitativa delproven.

När det gäller åldersgrupper är det de yngsta som uppnår de bästa resultaten på delproven XYZ, KVA och NOG medan skillnaden i resultat på delprovet DTK är små mellan åldersgrupperna. Generellt kan man säga att ju yngre desto högre resultat på delproven XYZ, KVA och NOG med undantag för gruppen 40 år eller äldre som har något högre resultat än provdeltagare i åldern 21-39 år. Den äldsta gruppen är dock en liten grupp, endast 2 procent av provdeltagargruppen. I den gamla provsammansättningen var mönstret i resultat mellan åldersgruppera mer tydligt, dvs. ju yngre desto bättre på delprovet NOG och ju äldre desto bättre delprovet DTK. Den nya provsammansättningen med alla fyra kvantitativa delprov i ett och samma provhäfte och att provdeltagarna själva får disponera tiden kan göra att resultatbilden mellan åldersgrupperna har förändrats och speciellt för delprovet DTK, som ligger sist i provhäftet.

Föräldrars utbildning har stor betydelse för resultatet på de kvantitativa delproven i högskoleprovet. Provdeltagare vars föräldrar har en eftergymnasial utbildning 5 år eller längre är de som har högst resultat på högskoleprovet. Även provdeltagarnas egen utbildning har betydelse för resultat på de kvantitativa delarna. De som gått naturvetenskapligt program/linje har de högsta resultaten och de som gått tekniskt program har de näst högsta resultaten.

Den lilla gruppen provdeltagare (0.5 % procent) som är utrikes födda med minst en utrikes född förälder är de som har högst resultat på såväl den kvantitativa delen som de enskilda kvantitativa delproven. För delprovet XYZ, det mest matematiskt inriktade delprovet, är skillnaden mellan de övriga grupperna med svensk eller utländsk bakgrund mycket liten. Skillnaden är något större på delprovet KVA. Störst skillnad återfinns på delprovet DTK, som kräver mest läsförståelse i jämförelse med de övriga kvantitativa delproven.

### *Resultat verbal del hösten 2011 och våren 2012*

Nedan följer en redovisning av poängmedelvärde (M) och genomsnittlig lösningsproportion (p) för de olika grupper som genomförde proven hösten 2011 och våren 2012. Även antalet provdeltagare (N) anges. Medelvärdet på den verbala delen 12A är genomgående lägre än vid provet 11B. Anledningen är att fyra ORD-uppgifter i den verbala delen blev strukna på grund av att dessa hade lagts ut på nätet under provdagens lunchrast. Istället för 20 ORD-uppgifter innehåller 12A därför 16 ORD-uppgifter och därmed totalt 76 verbala uppgifter.

**Tabell 14. Resultat för kvinnor och män samt totalt**

| **Kön** | **11B** | | | | **12A** | | | |
|---|---|---|---|---|---|---|---|---|
| | **M** | **p** | **s** | **N** | **M** | **p** | **s** | **N** |
| Kvinnor | 40.48 | .51 | 13.5 | 20 772 | 38.45 | .51 | 12.3 | 28 595 |
| Män | 42.76 | .53 | 13.5 | 19 923 | 40.97 | .54 | 12.6 | 28 269 |
| Totalt | 41.59 | .52 | 13.5 | 40 695 | 39.70 | .52 | 12.5 | 56 864 |

Vid provet hösten 2011 var medelvärdet på den verbala delen för totalgruppen 41.59 och spridningen var 13.5. Skillnaden mellan män och kvinnor var 2.28 till männens fördel. Spridningen var densamma för männen som för kvinnorna.

Vid provet våren 2012 var medelvärdet 39.70 (76 uppgifter) och spridningen 12.5. Skillnaden mellan män och kvinnor var 2.52 till männens fördel. Spridningen var 0.3 högre för männen än för kvinnorna.

Delprovet ELF står för mer än hälften av könsskillnaden på den verbala delen, vilket framgår av tabell 15. Delprovet ELF är även det verbala delprov som har den högsta lösningsproportionen, 0.58 per uppgift medan delprovet ORD har den lägsta lösningsproportionen, 0.48 per uppgift.

**Tabell 15. Resultat för kvinnor och män uppdelat på verbala delprov**

| Kön | ORD 20 uppgifter M | s | LÄS 20 uppgifter M | s | MEK 20 uppgifter M | s | ELF 20 uppgifter M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| Kvinnor | 9.68 | 4.4 | 9.64 | 3.3 | 10.16 | 4.1 | 11.00 | 3.9 |
| Män | 9.69 | 4.3 | 10.04 | 3.4 | 10.65 | 4.1 | 12.38 | 3.9 |
| Totalt | 9.68 | 4.4 | 9.83 | 3.4 | 10.40 | 4.1 | 11.68 | 3.9 |
| | | | | | | | | |
| **12A** | (16 uppgifter) | | | | | | | |
| Kvinnor | 7.63 | 3.6 | 10.05 | 3.5 | 9.88 | 3.8 | 10.89 | 3.7 |
| Män | 7.84 | 3.6 | 10.36 | 3.5 | 10.56 | 4.0 | 12.20 | 3.7 |
| Totalt | 7.74 | 3.6 | 10.20 | 3.5 | 10.22 | 3.9 | 11.54 | 3.7 |

I tabellerna 16-18 redovisas resultat för olika åldersgrupper.

**Tabell 16. Resultat för olika åldersgrupper**

| Ålder | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| - 20 | 39.15 | .49 | 12.4 | 23 653 | 37.69 | .50 | 11.5 | 31 629 |
| 21 - 24 | 42.38 | .53 | 13.2 | 10 371 | 39.73 | .50 | 12.2 | 15 764 |
| 25 - 29 | 46.54 | .58 | 14.3 | 3 640 | 44.08 | .58 | 13.2 | 5 309 |
| 30 - 39 | 50.85 | .64 | 15.3 | 2 233 | 48.25 | .64 | 14.2 | 2 951 |
| 40 - | 55.24 | .69 | 13.8 | 798 | 52.00 | .68 | 13.7 | 1 211 |
| Totalt | 41.59 | .52 | 13.5 | 40 695 | 39.70 | .52 | 12.5 | 56 864 |

Den verbala delen har ett starkt samband med ålder, dvs. ju högre ålder desto högre resultat på den verbala delen. Det skiljer ungefär 15 råpoäng mellan den yngsta och äldsta gruppen. Spridningen är störst i åldersgruppen 30-39-år och minst i gruppen som är 20 år och yngre.

Eftersom ungefär 85 procent av deltagargruppen är yngre än 24 år redovisas, på samma sätt som för den kvantitativa delen, antal provdeltagare per ålder upp till 24 år.

**Tabell 17. Resultat för olika åldersgrupper**

| Ålder | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| -17 | 39.33 | .49 | 12.0 | 2 066 | 37.44 | .49 | 11.2 | 875 |
| 18 | 38.53 | .48 | 12.2 | 10 236 | 38.22 | .50 | 11.2 | 5 732 |
| 19 | 39.11 | .49 | 12.5 | 6 424 | 37.43 | .49 | 11.5 | 17 131 |
| 20 | 40.43 | .51 | 12.6 | 4 926 | 37.89 | .50 | 11.7 | 7 891 |
| 21 | 40.91 | .51 | 12.8 | 3 794 | 38.79 | .51 | 11.9 | 5 748 |
| 22 | 42.54 | .53 | 13.2 | 2 865 | 39.26 | .52 | 12.2 | 4 507 |
| 23 | 43.04 | .54 | 13.4 | 2 114 | 40.56 | .53 | 12.2 | 3 159 |
| 24 | 44.70 | .56 | 13.6 | 1 598 | 41.83 | .55 | 12.7 | 2 350 |

Även om de enskilda åldrarna, upp till och med 24 år, redovisas var för sig blir resultatbilden den densamma, dvs. åldern avspeglar sig i resultat på den verbala delen. Med undantag för de allra yngsta vid provtillfället 12A, som har ett högre resultat än 18- och 19-åringarna. Som tidigare nämnts är det troligen de allra mest studieinriktade i den åldersgruppen som genomför provet. För att undersöka om sambandet med ålder är lika starkt för de olika delproven redovisas i tabell 18 resultat för de olika delproven uppdelat på motsvarande åldersgrupper som i tabell 16.

**Tabell 18. Resultat för olika åldersgrupper uppdelat på de verbala delproven**

| Ålder | ORD 20 uppgifter M | s | LÄS 20 uppgifter M | s | MEK 20 uppgifter M | s | ELF 20 uppgifter M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| -20 år | 8.46 | 3.8 | 9.55 | 3.2 | 9.73 | 3.8 | 11.41 | 3.9 |
| 21-24 år | 10.14 | 4.1 | 9.97 | 3.4 | 10.45 | 4.0 | 11.82 | 4.0 |
| 25-29 år | 12.14 | 4.2 | 10.37 | 3.6 | 11.77 | 4.3 | 12.26 | 4.3 |
| 30-39 år | 14.18 | 4.2 | 10.83 | 3.8 | 13.29 | 4.6 | 12.54 | 4.6 |
| 40 år- | 16.23 | 3.5 | 11.11 | 3.7 | 15.18 | 4.0 | 12.72 | 4.5 |
| Totalt | 9.68 | 4.4 | 9.83 | 3.4 | 10.40 | 4.1 | 11.68 | 4.0 |
| | | | | | | | | |
| **12A** | (16 uppgifter) | | | | | | | |
| -20 år | 6.84 | 3.2 | 9.97 | 3.4 | 9.57 | 3.6 | 11.30 | 3.6 |
| 21-24 år | 7.78 | 3.4 | 10.21 | 3.5 | 10.20 | 3.8 | 11.55 | 3.8 |
| 25-29 år | 9.65 | 3.5 | 10.67 | 3.7 | 11.60 | 4.1 | 12.16 | 4.0 |
| 30-39 år | 11.48 | 3.6 | 11.19 | 4.0 | 13.01 | 4.4 | 12.57 | 4.2 |
| 40 år- | 12.96 | 3.2 | 11.74 | 3.9 | 14.70 | 4.2 | 12.59 | 4.3 |
| Totalt | 7.74 | 3.6 | 10.20 | 3.5 | 10.22 | 3.9 | 11.54 | 3.8 |

Samtliga delprov har ett positivt samband med ålder, dvs. ju äldre desto högre resultat på delprovet. Det starkaste sambandet uppvisar delprovet ORD, följt av delprovet MEK. Delproven LÄS och ELF uppvisar ett betydligt svagare samband. I siffror räknat är skillnaden i resultat mellan den yngsta och den äldsta gruppen vid provet 11B 7.77 råpoäng för delprovet ORD, 5.45 råpoäng för delprovet MEK, 1.56 råpoäng för delprovet LÄS och 1.31 råpoäng för delprovet ELF.

**Tabell 19. Resultat för provdeltagare i relation till föräldrarnas utbildning**

| Föräldrars utbildning | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| Gymnasium 2 år | 38.10 | .48 | 13.2 | 8 944 | 36.75 | .48 | 12.3 | 12 605 |
| Gymnasium 3 år | 39.08 | .49 | 12.9 | 5 695 | 37.39 | .49 | 12.0 | 8 193 |
| Eftergymn.utb. -2 år | 41.26 | .52 | 13.0 | 9 018 | 39.38 | .52 | 12.0 | 12 699 |
| Eftergymn.utb. 3-4 år | 43.87 | .55 | 13.3 | 13 042 | 41.88 | .55 | 12.3 | 18 142 |
| Eftergymn.utb 5 år - | 48.57 | .61 | 13.1 | 3 321 | 45.82 | .60 | 12.0 | 4 320 |
| Totalt | 41.70 | .52 | 13.5 | 40 020 | 39.80 | .52 | 12.5 | 55 959 |

Det föreligger ett starkt samband mellan föräldrars utbildning och provdeltagarnas resultat även på den verbala delen. Skillnaden i resultat mellan den grupp vars föräldrar har 2-årig gymnasialutbildning och den

grupp vars föräldrar har eftergymnasialutbildning 5 år eller längre var 10.47 råpoäng (0.13 per uppgift) hösten 2011 och 9.07 (0.12 per uppgift) våren 2012.

I tabell 20 redovisas motsvarande värden för respektive delprov. Där framgår att resultat på delprovet ELF har något högre samband med föräldrars utbildning än de övriga delproven.

**Tabell 20. Resultat på de olika verbal delproven i relation till föräldrarnas utbildning**

| Föräldrars utbilning | ORD 20 uppg. M | s | LÄS 20 uppg. M | s | MEK 20 uppg. M | s | ELF 20 uppg. M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| Gymnasium 2 år | 9.16 | 4.5 | 8.98 | 3.1 | 9.54 | 4.0 | 10.42 | 4.0 |
| Gymnasium 3 år | 9.01 | 4.3 | 9.35 | 3.2 | 9.71 | 4.0 | 11.01 | 3.9 |
| Eftergymnutb. -2 år | 9.52 | 4.2 | 9.83 | 3.3 | 10.32 | 4.0 | 11.60 | 3.9 |
| Eftergymnutb 3-4 år | 10.12 | 4.3 | 10.33 | 3.4 | 10.99 | 4.0 | 12.43 | 3.9 |
| Eftergymnutb. 5 år - | 11.27 | 4.3 | 11.36 | 3.4 | 12.21 | 4.0 | 13.73 | 3.6 |
| Totalt | 9.71 | 4.4 | 9.86 | 3.3 | 10.43 | 4.1 | 11.70 | 4.0 |
| | | | | | | | | |
| **12A** | (16 uppg.) | | | | | | | |
| Gymnasium 2 år | 7.45 | 3.8 | 9.37 | 3.3 | 9.41 | 3.9 | 10.53 | 3.7 |
| Gymnasium 3 år | 7.19 | 3.5 | 9.66 | 3.4 | 9.58 | 3.9 | 10.96 | 3.7 |
| Eftergymnutb. -2 år | 7.66 | 3.5 | 10.15 | 3.4 | 10.13 | 3.8 | 11.44 | 3.7 |
| Eftergymnutb 3-4 år | 8.07 | 3.5 | 10.78 | 3.5 | 10.80 | 3.9 | 12.22 | 3.7 |
| Eftergymnutb. 5 år - | 8.85 | 3.5 | 11.74 | 3.5 | 11.94 | 3.7 | 13.29 | 3.5 |
| Totalt | 7.77 | 3.6 | 10.23 | 3.5 | 10.25 | 3.9 | 11.56 | 3.7 |

I tabell 21 redovisas resultat för provdeltagare från olika gymnasieprogram.

**Tabell 21. Resultat för provdeltagare från olika gymnasielinjer/program**

| Utbildning[3]/ prov | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| 1 Samhälls | 41.66 | .52 | 12.8 | 9 064 | 40.65 | .53 | 12.1 | 9 861 |
| 2 Natur | 46.99 | .59 | 13.1 | 6 626 | 45.47 | .60 | 12.0 | 6 674 |
| 3 Teknisk | 39.82 | .50 | 11.9 | 1 544 | 39.61 | .50 | 11.0 | 1 806 |
| 4 Estetisk | 42.55 | .53 | 13.3 | 1 887 | 40.80 | .54 | 12.6 | 2 551 |
| 5 Hemkunskaps | 34.33 | .43 | 11.6 | 1 874 | 32.98 | .43 | 10.7 | 2 224 |
| 6 Yrkesinr | 38.69 | .48 | 12.2 | 1 323 | 37.87 | .50 | 11.5 | 1 801 |
| Totalt | 42.40 | .53 | 13.3 | 22 318 | 40.99 | .54 | 12.4 | 24 917 |

Även på de verbala delproven har de provdeltagare som gått en teoretisk utbildning på gymnasiet högre resultat än de som gått en utbildning med

[3] 1 samtliga samhällsinriktade program
2 naturvetenskap, naturbruk, IB
3 tekniskt program
4 estet, hantverk, media
5 barn/fritid, handel/adm, hotell/rest., livsmedel
6 bygg, el,energi,fordon,industri

praktisk inriktning. De som gått en naturvetenskaplig utbildning har högst resultat medan de som gått en gymnasial utbildning mot barn/fritid, handel/administration, hotell/restaurang, livsmedel har lägst resultat. Skillnaden i resultat mellan provdeltagare från dessa är minst på delprovet ORD och störst på delprovet ELF, vilket framgår i tabell 22.

**Tabell 22. Resultat för provdeltagare från olika gymnasielinjer/program uppdelat på verbala delprov**

| Utbildning/ prov | ORD 20 uppgifter M | s | LÄS 20 uppgifter M | s | MEK 20 uppgifter M | s | ELF 20 uppgifter M | s |
|---|---|---|---|---|---|---|---|---|
| **11B** | | | | | | | | |
| 1 Samhälls | 9.86 | 4.2 | 9.88 | 3.2 | 10.20 | 3.9 | 11.71 | 3.9 |
| 2 Natur | 10.96 | 4.2 | 10.96 | 3.4 | 11.94 | 3.9 | 13.13 | 3.8 |
| 3 Teknisk | 8.87 | 3.9 | 9.56 | 3.1 | 9.76 | 3.6 | 11.64 | 3.7 |
| 4 Estetisk | 10.60 | 4.3 | 9.96 | 3.3 | 10.41 | 4.0 | 11.58 | 3.9 |
| 5 Hemkunsk | 8.64 | 4.0 | 8.19 | 2.9 | 8.49 | 3.5 | 9.01 | 3.8 |
| 6 Yrkesinr | 8.97 | 4.0 | 9.18 | 3.2 | 9.49 | 3.7 | 11.05 | 3.8 |
| Totalt | 10.03 | 4.2 | 10.00 | 3.3 | 10.52 | 4.0 | 11.85 | 4.0 |
| | | | | | | | | |
| **12A** | (16 uppgifter) | | | | | | | |
| 1 Samhälls | 8.09 | 3.5 | 10.37 | 3.4 | 10.51 | 3.8 | 11.68 | 3.7 |
| 2 Natur | 9.07 | 3.5 | 11.55 | 3.5 | 11.87 | 3.7 | 12.99 | 3.6 |
| 3 Teknisk | 7.52 | 3.2 | 10.25 | 3.2 | 10.13 | 3.5 | 11.71 | 3.5 |
| 4 Estetisk | 8.47 | 3.6 | 10.39 | 3.5 | 10.55 | 4.0 | 11.38 | 3.8 |
| 5 Hemkunsk | 6.95 | 3.4 | 8.50 | 3.0 | 8.37 | 3.5 | 9.16 | 3.4 |
| 6 Yrkesinr | 7.55 | 3.4 | 9.44 | 3.2 | 9.66 | 3.8 | 11.22 | 3.7 |
| Totalt | 8.21 | 3.5 | 10.44 | 3.5 | 10.60 | 3.9 | 11.74 | 3.8 |

**Tabell 23. Resultat för provdeltagare med svensk/utländsk[4] bakgrund**

| Sv/utl bakgrund | 11B M | p | s | N | 12A M | p | s | N |
|---|---|---|---|---|---|---|---|---|
| 1 | 33.53 | .42 | 13.3 | 2 903 | 32.11 | .42 | 12.0 | 4 219 |
| 2 | 34.77 | .43 | 12.4 | 3 466 | 33.49 | .44 | 11.3 | 5 239 |
| 3 | 41.96 | .52 | 12.9 | 236 | 40.89 | .54 | 12.1 | 350 |
| 4 | 42.48 | .53 | 13.4 | 3 916 | 40.74 | .54 | 12.3 | 5 487 |
| 5 | 43.03 | .54 | 13.2 | 30 138 | 41.10 | .54 | 12.2 | 41 529 |
| Totalt | 41.59 | .52 | 13.5 | 40 659 | 39.70 | .52 | 12.5 | 56 824 |

Den största delen av provdeltagaran (75 %) är födda i Sverige av svenskfödda föräldrar. Denna grupp har det högsta resultatet på den verbala delen av provet.

[4] 1 utrikes födda med två utrikes födda föräldrar
2 inrikes födda med två utrikes födda föräldrar
3 utrikes födda med minst en utrikes född förälder
4 inrikes född med en inrikes och en utrikes född förälder
5 inrikes född med två inrikes födda föräldrar

**Tabell 24. Resultat för provdeltagare med svensk/utländsk bakgrund för olika delprov**

| | ORD 20 uppgifter | | LÄS 20 uppgifter | | MEK 20 uppgifter | | ELF 20 uppgifter | |
|---|---|---|---|---|---|---|---|---|
| | M | s | M | s | M | s | M | s |
| **11B** | | | | | | | | |
| 1 | 7.52 | 4.0 | 8.19 | 3.3 | 7.95 | 4.1 | 9.86 | 4.2 |
| 2 | 7.33 | 3.9 | 8.56 | 3.1 | 8.26 | 3.7 | 10.61 | 4.0 |
| 3 | 8.96 | 4.2 | 9.91 | 3.3 | 10.24 | 4.1 | 12.85 | 3.7 |
| 4 | 9.81 | 4.4 | 10.02 | 3.3 | 10.50 | 4.1 | 12.15 | 3.9 |
| 5 | 10.15 | 4.3 | 10.11 | 3.3 | 10.86 | 4.0 | 11.90 | 4.0 |
| Totalt | 9.68 | 4.4 | 9.83 | 3.4 | 10.40 | 4.1 | 11.67 | 4.0 |
| | | | | | | | | |
| **12A** | (16 uppgifter) | | | | | | | |
| 1 | 5.71 | 3.2 | 8.31 | 3.4 | 8.14 | 3.8 | 9.95 | 4.0 |
| 2 | 5.85 | 3.0 | 8.70 | 3.3 | 8.25 | 3.7 | 10.69 | 3.7 |
| 3 | 7.51 | 3.7 | 10.39 | 3.4 | 10.21 | 3.9 | 12.77 | 3.7 |
| 4 | 7.85 | 3.6 | 10.38 | 3.5 | 10.51 | 3.9 | 12.00 | 3.7 |
| 5 | 8.17 | 3.6 | 10.56 | 3.4 | 10.64 | 3.8 | 11.74 | 3.7 |
| Totalt | 7.74 | 3.6 | 10.20 | 3.5 | 10.22 | 3.9 | 11.54 | 3.8 |

Gruppen svenskfödda med svenskfödda föräldrar har de högsta resultaten på delproven ORD, LÄS och MEK, men har lägre resultat på delprovet ELF än både gruppen födda i Sverige med en inrikes och en utrikes född förälder och gruppen födda utomlands med minst en utrikes född förälder.

## Sammanfattning

Skillnaden mellan det nya och det gamla provet är att 20 ORD-uppgifter har ersatts med 20 MEK-uppgifter (meningskompletteringsuppgifter). Denna förändring har dock inte påverkat könsskillnaden nämnvärt. Det är delprovet ELF som svarar för mer än hälften av denna skillnad.

När det gäller åldersgrupper ser man att det är ett starkt samband med ålder, det är starkast för delprovet ORD, följt av delprovet MEK. Skillnaden i resultat mellan åldersgrupperna är betydligt mindre för delproven LÄS och ELF. Sett över alla 80 uppgifter har åldersskillnaden minskat något genom att 20 ORD-uppgifter har ersatts med 20 MEK-uppgifter. Det tyder på att yngre provdeltagare har lättare att avgöra betydelse av ord och uttryck när de presenteras i sammanhang än när de står enskilt.

Föräldrars utbildning har även stor betydelse för resultat på de verbala delproven. Provdeltagare vars föräldrar har en eftergymnasial utbildning 5 år eller längre är de som har högst resultat på högskoleprovet. Störst skillnad återfinns på delprovet ELF. Även provdeltagarnas egen utbildning har betydelse för resultat på den verbala delen i högskoleprovet. De som gått naturvetenskapligt program/linje har de högsta resultaten.

Inrikes född med antingen en eller två inrikes födda föräldrar har högst resultat på den verbala delen. Vid en uppdelning i delprov framgår det att det är den lilla gruppen provdeltagare (0.5 % procent) som är utrikes födda med minst en utrikes född förälder som har det högsta resultatet på delprovet ELF följt av gruppen inrikes född med en inrikes och en utrikes född förälder.

### *Sammanfattning av konsekvenser för olika grupper av provdeltagare av den nya sammansättningen av högskoleprovet*

Hösten 2011 infördes ett förändrat högskoleprov som innebar att provet sedan dess består av två distinkta delar – en kvantitativ och en verbal – som vardera innehåller 80 uppgifter, totalt 160 uppgifter. Dessförinnan bestod provet av 80 verbala och 42 kvantitativa uppgifter, totalt 122 uppgifter.

Syftet med föreliggande rapport är att redovisa provdeltagarnas resultat hösten 2011 och våren 2012, de två första provtillfällena med det nya provformatet. I denna sammanfattande del kommer även en jämförelse med resultat från tidigare provsammansättning att göras.

#### Kvinnor och män

Resultatskillnader mellan kvinnor och män har följts noggrant sedan provets begynnelse. Det är i huvudsak de kvantitativa delproven som svarat för den största könsskillnaden, cirka 3.5 – 4 råpoäng, dvs. 0.08-0.09 per uppgift. I den nya provsammansättningen, där antalet kvantitativa uppgifter har utökats från 42 till 80 uppgifter var könsskillnaden 7.10 vid provet hösten 2011 och 6.72 vid provet våren 2012, dvs. 0.09 respektive 0.08 per uppgift. Storleken per uppgift är densamma men eftersom antalet kvantitativa uppgifter har utökats med 38 uppgifter har den totala könsskillnaden för den kvantitativa delen ökat.

På den verbala delen är könsskillnaden på samma nivå som den var i den tidigare provsammansättningen, 2-2.5 råpoäng, dvs. 0.03 per uppgift till männens fördel. Delprovet ELF – som mäter engelsk läsförståelse svarar för drygt hälften av denna skillnad.

#### Åldersgrupper

Generellt kan man säga att ju yngre desto högre resultat på den kvantitativa delen, det gäller framför allt de två nya delproven XYZ – som mäter förmågan att lösa matematiska problem och KVA – som mäter förmågan att göra kvantitativa jämförelser. Delprovet DTK – som mäter förmågan att hämta och tolka information från grafiska beskrivningar uppvisar däremot små skillnader mellan åldersgrupperna.

Den verbala delen har däremot ett klart samband med ålder, dvs, ju högre ålder desto högre resultat på den verbala delen. Delprovet ORD – som mäter svensk ordförståelse är det delprov som alltid stått för en stor del av denna skillnad i resultat (cirka 13-14 råpoäng). I och med den senaste förändringen har antalet ORD-uppgifter halverats och ersatts av delprovet MEK (meningskomplettering) – som mäter förmågan att förstå ord och begrepp, vilket har inneburit att skillnaden har minskat med ungefär 2 råpoäng. De två övriga verbala delproven LÄS (svensk läsförståelse) och ELF (engelsk läsförståelse) uppvisar däremot små skillnader.

#### Föräldrarnas utbildning

Föräldrarnas utbildning har stor betydelse för provdeltagarnas resultat på högskoleprovet det gäller såväl den verbala som den kvantitativa delen. Det är de provdeltagare vars föräldrar har eftergymnasial utbildning som har högst resultat medan det är provdeltagare vars föräldrar har 2-årig gymnasial

utbildning som har lägst resultat. Föräldrars utbildning har störst betydelse för de tre kvantitativa delproven XYZ, KVA och NOG medan föräldrars utbildning har något mindre betydelse för delprovet DTK och även de fyra verbala delproven. Skillnaden mellan provdeltagare vars föräldrar har eftergymnasial utbildning och de provdeltagare vars föräldrar har 2-årig gymnasial utbildning var för hela provet 23.87 råpoäng hösten 2011 och 20.37 råpoäng våren 2012.

### Gymnasial utbildning

De provdeltagare som har gått naturvetenskapligt gymnasieprogram har de högsta resultaten på såväl den kvantitativa som den verbala delen. Det är ett resultat som har varit gällande sedan provets begynnelse. På den kvantitativa delen är det provdeltagare från tekniskt gymnasieprogram som har det näst högsta resultatet medan på den verbala delen är det provdeltagare från estetiskt program som har det näst högst resultatet, vilket är endast något högre än för provdeltagare från samhällsvetenskapligt program.

### Svensk/utländsk bakgrund

De provdeltagare som är inrikes födda har högst resultat på såväl den kvantitativa som den verbala delen. Skillnaden mellan den grupp inrikes födda som har högst resultat och den grupp utrikes födda som har lägst resultat är ungefär hälften så stor för den kvantitativa delen som för den verbala delen, ungefär 5 råpoäng respektive 10 råpoäng. Den provsammansättning som gällde före hösten 2011 - med endast 42 kvantitativa uppgifter - hade även då ungefär 5 råpoäng skillnad mellan den grupp inrikes födda med högst resultat och den utrikes födda med lägst resultat. Införandet av två kvantitativa delprov ökar inte denna skillnad utan den är fortfarande ungefär 5 råpoäng.

## REFERENSER

Stage, C. (1995). *Utprövning av provuppgifter* (Pm nr 106). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.

Stage, C. & Ögren, G. (2001). *Högskoleprovets utveckling under åren 1977–2000. Provets sammansättning och provdeltagargruppens sammansättning och resultat* (Pm nr 169). Umeå: Umeå universitet, Enheten för Pedagogiska mätningar.

Stage, C. & Ögren, G. (2010). *Ett nytt högskoleprov. Bakgrund och konsekvenser* (BVM 42:2010). Umeå: Umeå universitet, Institutionen för tillämpad utbildningsvetenskap, Beteendevetenskapliga mätningar.

Ögren, G., Stage, C., Åström, K., & Lexelius, A. (1995). *Försök inför förändring av högskoleprovets utprövningsrutiner* (Pm nr 107). Umeå: Umeå universitet, Enheten för pedagogiska mätningar.

# Who benefits from university admissions tests? - A comparison between grades and test scores as selection instruments to higher education

Magnus Wikström, Department of Economics, Umeå University, Sweden[1]
Christina Wikström, Department of Applied Educational Science, Umeå University, Sweden[2]

## Abstract

In Sweden, there are two separate instruments for ranking applicants in the admission to higher education; the GPA from upper secondary school and the Swedish Scholastic Assessment Test (the SweSAT). A problem in the selection is that different groups of students perform differently on the instruments. Also, while the GPA is regarded more valid but with reliability problems, the test is seen as reliable but with lower predictive validity. Hence, in 2011, the test was revised, with the purpose to increase its relevance for university studies. New item types and new subtests were introduced, and the weights of the verbal and quantitative parts of the test were made more balanced. This study compares how students are ranked on the basis of the new test compared to their GPA, to find out if previous group differences still remain. The data consists of test participants in the autumn of 2011 and spring of 2012 at the ages 17 to 25. The data is analysed mainly through correlations and regression analysis. The results show that the correlation between test scores and GPA is approximately the same as before the revision. It is also found that there are still group differences in terms of boys performing better on the test and girls on the GPA. However, when studying separate sub-tests and grades and national course tests from isolated subjects, the students seem to be ranked more similarly than in the overall SweSAT-GPA comparison. An interesting finding is that although students with a non-Swedish background are performing lower than other students on both instruments, boys in this group seem to be graded more leniently than the girls with a similar background in mathematics, and the opposite is the case in verbal subjects.

[1] Magnus Wikström,
Department of Economics
Umeå School of Business and Economics
Umeå University
SE-901 87 Umeå
Sweden
email: magnus.wikstrom@econ.umu.se
phone: +46 90 786 56 40

2 Christina Wikström,
Department of Applied Educational Science
Educational Measurement Unit
Umeå University
SE-901 87 Umeå
Sweden
email: christina.wikstrom@edusci.umu.se
Phone: +46 90 786 55 70

## Table of Contents

# 1 Introduction

When there are more eligible applicants to higher education than there are study places, a selection must take place. Different systems have different strategies for making this selection. In Sweden, there are two main routes and two selection instruments, by which students are ranked; the grade point average based on the upper secondary school grades and an admissions test called the SweSAT. The SweSAT has been introduced as an option to the GPA, to function as a "second chance". The test has many similarities to other well-known admissions tests such as the American SAT and the Israeli PET, in both content and format. It consists of a number of subtests measuring verbal and quantitative skills that are expected to be important for academic performance, and the expectation is that the higher the score, the more likely is the test taker to be successful in higher education. The main difference between the Swedish test and other similar tests is generally that the Swedish admissions test is not used in combination with grades, but as an optional alternative for students who compete for study positions.

It is well known that the admission to higher education is not without problems, often raising questions about fairness, but also about practical usefulness and potential consequences when it comes to selecting students. By having two instruments it is possible to make comparisons, but if the instruments rank students in different ways it is not always clear what is to be preferred. Who "deserves" to be first in line? Should it be the highest performing students, or should other aspects be regarded equally or more important?

In the research on selection instruments, it is generally found that there are systematic differences in the information that is given by grades and tests, as will be discussed in this paper. Research has also shown that the upper secondary school grades are slightly better at predicting academic performance, compared to the test. One hypothesis often discussed is that the higher predictive strength can be explained by the GPA representing cognitive as well as non-cognitive aspects, both important in education, while the test does not capture motivation, industriousness, social skills etc. to the same extent. Although this hypothesis seems plausible, it has been difficult to provide

evidence that this is the reason for the difference in predictive strength between the instruments.

The test has also been criticized on other grounds, most often with the argument that the predictive validity could be improved with a change of content. It has been claimed that the strong emphasis on verbal items has been negative, especially when the test is used in the admission of students in technical or mathematically oriented study programmes (see, for instance, Svensson, 2004). Consequently In 2011, the SweSAT was revised. The main changes were that the verbal and the quantitative parts of the tests were made more balanced. New subtests were added, and the test now consists of more items (38 more items, making a total of 160 instead of 122). The main purpose of adding items was to make it possible to calculate a normed score for the verbal and the quantitative parts separately. The test score is now reported as a total score, as well as two separate scores. The idea with this is to make it possible to put extra weight on either the verbal or the quantitative part in the admissions process, if this is regarded important for a specific programme.

## 1.1 Aim of the study

This study investigates how applicants to higher education in Sweden are ranked on the basis of the revised SweSAT in comparison with the upper secondary GPA. The purpose is to find out if groups of students scores higher or lower depending on the instrument, and to see if recent revisions of the test have changed the patterns that have been found in previous research. The main focus will be on gender differences, since this has been a recurring topic in previous research, but we also discuss differences that are related to the applicants' socio-economic and ethnical background, as well as educational experience.

# 2 The admission system

Since this study focuses on analysing the selection instruments and how they work in terms of rank ordering students, it may be useful for the reader to have an understanding of how the Swedish admissions system works and what characterises the two selection instruments have. For a more detailed description, see Stage (2004c).

## 2.1 Eligibility vs. selection

In order to get admitted to higher education in Sweden, the applicant needs to fulfil the requirements for previous education, which generally is to have graduated from upper secondary school. For some university programmes or courses, there are special requirements regarding advanced courses. However, if there are more applicants than study places, a selection must take place. There are mainly two instruments used in this selection: the GPA from upper secondary school, and the SweSAT (in Swedish: *Högskoleprovet*). The instruments are never combined as in many other countries; instead the applicants are ranked in different quota groups depending on which instrument he or she is applying with. An applicant with both a GPA and a test score from the SweSAT will be admitted on the basis of the instrument that is most favourable. Universities are required to admit at least 30 per cent of their students from each group. The universities usually admit just above this limit from the SweSAT group, and the majority of the students from the GPA group.

Since the SweSAT is optional to take, but grades are necessary for eligibility, basically all students will apply with their grades and consequently their GPA, but only the students who have taken the SweSAT and submitted their results will be placed in the SweSAT group. The selection group based on grades is therefore considerably larger than the SweSAT group.

## 2.2 The upper secondary grades and GPA

Since 1994, the Swedish grading system has been criterion-referenced. Upper secondary school is not compulsory, but most students continue to this level after compulsory school, at the age of 16. After each course students are graded from A-F, where E is the lowest approved grade, and A is the highest. Prior to 2012, the grading scale had only

four grades, and ranged from Fail (IG) to Pass (G), Pass with distinction (VG) and Pass with special distinction (MVG). The GPA is the average from all upper secondary school grades, where also the length of the course determines its weight. Some courses also give extra credit. In order to make the calculation, the letter grades are converted to a numerical measure (IG=0, G=10, VG=15, MVG =20). Although the letter grades are criterion referenced, the GPA (0-20) should be regarded norm referenced, with the sole purpose of ranking students (Wikström, 2005).

## 2.3 The SweSAT

As previously described, The SweSAT is a norm-referenced test, used for ranking students who apply for higher education. The test is administered twice a year, and simultaneously at a number of places throughout Sweden and also in some places abroad. The SweSAT is optional and open for anyone to take. There is no limit when it comes to how many times the test can be taken, and the test taker has little to lose by retaking the test since it is the best result that counts. For most test takers, retaking the test generally gives a positive effect on the test score, especially between the first and second time. The assumption is that this is a learning effect that mainly has to do with becoming more familiar with the test format and better at allocating time (Henriksson & Bränberg, 1992; Henriksson & Wolming, 1998).

Since the autumn of 2011, a number of changes have been made to the test. The number of items in the test has changed from 122 to 160, and the test takers are given 25 extra minutes to compensate for this. In order to avoid making the test more speeded, some of the old items that need longer time to answer have been removed or changed. Still, the allotted time is not extensive considering the length of the test.

Prior to the autumn of 2011, the SweSAT consisted of six subtests: Swedish Reading comprehension (READ), English Reading comprehension (ERC), Vocabulary (WORD), Diagrams, Tables and Maps (DTM), and Data Sufficiency (DS). In the autumn of 2011, three new subtests were added: Sentence complementation (SC), Mathematics (XYZ) and Quantitative comparisons (QC), which means that the SweSAT now includes eight subtests. Consequently, the quantitative part of the test consists of 80 items and includes

the subtests DS, DTM, XYZ and QC, and the verbal part of the test consists of the subtests READ, WORD, ERC and SC (Stage & Ögren, 2010).

A test score is valid for five years and it is the best score that counts. This means that the student has little to lose by re-taking the test. The results are public, which means that the items and the test forms are changed for each administration. To ensure comparability, the test is standardised and equated over time (Lyrén, 2009).

## 2.4 Previous research on the selection instruments

Internationally, a relatively large amount of studies, focusing on the selection to higher education, have previously been carried out. The main part of this research has focused on the instrument's predictive validity, i.e. their ability to predict performance in higher education. Other studies have often focused on group differences, i.e. differences in performance on the selection instruments that systematically can be related to the test taker's gender, age, and socio-economic or ethnical background.

### 2.4.1 Research programmes

In Sweden, the main part of the research on the selection to higher education has been carried out within three research programmes. The first programme focused on the validity of the admission system. Among the studies, Henrysson (1984a) analysed methodological problems relating to how to analyse the predictive validity in selection instruments, and studied the impact of previous knowledge, study motivation and general aptitude (Henrysson, 1984b). Henrysson & Henrysson (1984) studied the effect of giving extra credits for work experience. Another study by Henrysson, Kriström, & Lexelius (1984) aimed to analyse the importance of characteristics relating to the students' background, which resulted in a discussion regarding the problem of indentifying criteria for study success.

A second research programme was characterised by a lot of focus on the criterion for academic success, as it had been found to be problematic when carrying out studies on prediction (Andersson & Grysell, 1992, 1995; Wester, 1995). Lexelius & Wolming (1995) discussed the problems with restriction of range in such studies, and found this

to be a problem that is difficult to solve, especially when studying university programmes where most students perform equally well on the criterion. Henriksson & Wolming (1998) studied group differences in academic performance, with special focus on students who had been given extra credits due to previous work experience, and found that these students performed lower than traditional students. Another study in this programme focused on how students in higher education performed on the SweSAT, compared to test takers without academic education, and found that further education generally improves performance on the test (Allergren, 1995). An interpretation of this is that the test seems to measure knowledge and skills that are relevant for higher education.

A few years later, the third and largest research programme was carried out (for an overview of this project, see Löfgren, 2005). The programme had the purpose of validating the selection system in a wider perspective. Different validity perspectives were considered; Åberg-Bengtsson (2005), Carlstedt & Gustafsson (2005), Cliffordson, (2006) and Cliffordson & Askling (2006) studied the characteristics of the SweSAT for construct validation, and results from their studies showed, for instance, that the test was dominated by *Gc* factors (skills, knowledge, and experience) rather than what is interpreted as *g*, or general intelligence. The grades were also in focus in this programme. Reuterberg & Hansen (2001) and Reuterberg (2003) studied group differences and found that test takers with a non-Swedish background performed lower on the test compared to test takers with a Swedish background. Studies by Stage (2004a; 2004b) focused on the gender differences on the test, where male test takers on average performed higher than female test-takers. The gender differences have been thoroughly investigated, both within this research programme, but also before and after. One example is a study by Reuterberg (1997) who analysed all subtests within the SweSAT with regards to gender differences but could not find any other explanation than that male students are higher performing, and especially in comparison with the grades where the differences were reversed, with female students performing higher than male students. The effect of repeated test taking was investigated also in this programme (Henriksson & Törnkvist 2002; Törnkvist & Henriksson 2004a, 2004b, 2006). Still, the main focus of the research within this project was the predictive validity of the selection

instruments. Studies by Cliffordson (2004a, 2004b, 2004c); Cliffordson and Askling (2006), Löfgren & Törnkvist (2004a, 2004b), Svensson (2004), Svensson & Nielsen, (2005) all showed that students selected by their GPA performed slightly higher at university compared to students selected by their test scores. For an overview of prediction studies, see Lyrén (2008). However, Wikström & Wikström (2005) and Wikström (2005a) showed that the grades had problems with reliability, and could not be strictly compared between schools or over time. The criterion problem was discussed also in this programme (Wolming, 2001, Andersson, 2003).

From an international perspective, research generally supports the findings described above. Similar group differences with regards to ethnic background and gender have been found in the American SAT (Zwick, 2002, 2004), but also that underperforming groups on the SAT (non-whites) are over predicted in terms of academic success. The latter has not been investigated in a Swedish context. Research on the admissions test in Israel, PET, has to a large extent also focused on performance differences relate to socio-economic background (Beller, 2001). As has also been confirmed by a number of other studies, student background is important for explaining both study success and performance on high stakes tests.

### 2.4.2 Evaluations

Since the Swedish model differs from most other countries by having two separate selection instruments that are also used in a centralised admission system, the instruments need to be validated separately, but also together, to see how students are ranked on the basis of each instrument.

Over the years, there a few studies have focused on making such comparisons. The most recent is Lexelius (2004), who studied correlations between the upper secondary grades in the basic and compulsory courses in English, Mathematics and Swedish and the SweSAT. The timing of this study is important, since it focuses on the upper secondary grades just after the Swedish grade system had been changed from being norm referenced to criterion-referenced (or “goal referenced”). The analysis shows a general positive correlation between the grades and the test. There is a weaker correlation

between the grades in Swedish and the SweSAT, in comparison with the grades in English and Mathematics. The second level course in English (B course) had the highest correlation with the verbal subtest and the subtest DTM (Diagrams, Tables and Maps) and also the total score of the test. The grades in mathematics had the highest correlation with the quantitative subtests DTK and NOG (mathematical reasoning). Lexelius also studied the correspondence between grades and subtests by factor analysis, which confirmed the relationships between the quantitative tests and upper secondary mathematics, and also the verbal tests and upper secondary English. The relationship between Swedish and the verbal tests was considerably weaker. Similar results were found in a study by Wikström & Lyrén (2009), who discussed if the course grades in Swedish are more sensitive to construct irrelevance and measure aspects such as industriousness and motivation, compared to other course grades. If so, this could explain why they show lower correlations with tests scores or other grades.

## 3 Data

In this study, the SweSAT results from the autumn of 2011 (2011B) and spring of 2012 (2012A) are used. In the autumn of 2011, the total number of test takers was 41,247, and in the spring of 2012 the number of test takers was 57,634. 16,049 took both tests and can therefore be observed in both groups. Since the test is open for all, some of the test takers are unlikely to apply to university studies, since they are too young or already enrolled. From the total number of test takers, we select two groups of potential university students; (i) students who had finished upper secondary school by the time the test was administered, and (ii) a group of students who had not yet finished upper secondary school in 2011-2012.

### 3.1 Upper secondary school graduates

The first part of this study contains individuals born between 1986 and 1992. This means that they were between 19 and 25 years old in 2011. The major reason for restricting the upper limit is to exclude individuals who are 25 years or older, since their upper secondary grades may be problematic in terms of making comparisons. In 1997 a new criterion referenced grading system was introduced, and it is generally known that

grade point averages increased dramatically for a number of years, meaning that it is difficult to compare GPA for different cohorts (Wikström, 2005). By 2005, the tendency for grades to be inflated over time had decreased considerably. For this reason, we therefore choose to include individuals born from 1986 and onwards.

The key variables concerning the SweSAT used in this study are the normed SweSAT test score (SWESAT), the normed test score on the quantitative parts of the test (SWESATQ), and the corresponding score for the verbal parts (SWESATV). The test scores are compared with the upper secondary grade point average (GPA). Also, since the SweSAT is supposed to capture basic verbal and quantitative skills corresponding to an achievement level at the first years of upper secondary school, school grades in the first two mathematics courses (MATA and MATB) and English courses (ENGA and ENGB) are compared with the quantitative and verbal parts of the SweSAT, respectively.

Test scores and grades are complemented with a number of background variables aimed at controlling for differences between groups of individuals. The dummy variable MALE controls for gender differences and AGE for age differences. Immigrant status is known to affect test scores as well as grades. We impose controls for first generation immigrants, i.e. those born abroad with both parents born abroad (FIRSTGEN). Second generation immigrants are defined as those born in Sweden with both parents born outside of Sweden (SECONDGEN). Finally, we also control for individuals born outside of Sweden with at least one parent born in Sweden (FORBORN). Other important information is what upper secondary program the applicant has graduated from. Students who are participating in a theoretically oriented programme in upper secondary school on average perform better on the SweSAT and are more prone to advance into higher education. Furthermore, students in theoretically oriented programmes are not necessarily comparable to students who have participated in other programmes, since the courses vary, and especially between theoretical and practical programmes. Therefore, we define the variable ACTRACK to include the students who participated in a theoretically oriented upper secondary programme. Another difference between individuals is that some students who did not finish upper secondary school regularly did so by participating in the municipal adult education (kommunala

vuxenutbildningen). These are controlled for with the variable KOMVUX. To identify the students with previous experience with higher education (university level) the variable (HIGHEXP) is used. In the analysis of SweSAT 2012A, we also control for if they participated in the round of 2011B (PART11).

Finally, we control for the test takers' socioeconomic background. This is often measured by parental education and/or occupation. In this paper, we use parental education since it appears to be the most important factor by which education is transferred over generations. We define the parent with the highest level of education into two dummy variables; HIGHEDU if the parent with the highest education has some experience with higher education (post upper secondary school) and MIDEDU if the parent has upper secondary education. The reference group consists of those whose parents have secondary education or less. Variable definitions and data sources are presented in Table A1 in the Appendix A. Tables A2 and A3 presents descriptive statistics for the 2011B and 2012A test administrations respectively. The number of individuals selected is 23,214 and 27,075 for the 2011B and 2012A round, respectively. Missing information pertains to GPA and subject grades.

## 3.2 Upper secondary school students

The second part of the study contains individuals that have not yet left upper secondary school at the time the tests were administered. We select students born in 1993 and 1994, meaning that they were between 17 and 18 years old in 2011.[3] The number of test takers in this selection is 12,272 from the 2011B administration and 23,113 from 2012A. The background control variables are the same as in the first part of the study; the difference is in the upper secondary school variables used. For students currently participating in upper secondary school, subject grades cannot be observed. Therefore, we choose as outcome variables the national course (curriculum) tests that are administered at various times during the upper secondary school. These tests are used for information about student and school achievement and for grade calibrating

[3] As some students leave upper secondary school before the age of 19, this selection criterion implies that a few individuals included will have left the upper secondary school at the time the SweSAT was taken. In the 2011B administration 170 individuals have a final GPA from upper secondary school, and in the 2012A administration this number is 189.

purposes (in Sweden, there are no examination tests as grades are set by the teachers). The SweSAT normed quantitative and verbal scores (SWESATQ and SWESATV) are used together with the national test scores in mathematics (NMATA and NMATB), English (NENGA and NENGB) and Swedish (NSVB). Although these tests are administered to all students taking the relevant courses, only a random sample were collected from the schools at the time this data was collected. This means that the number of observations that can be used is restricted by the information collected centrally. The number of observations used varies between 1,076 and 6,821 depending on the test and SweSAT administration. In addition, previous evaluations have used secondary school grade point averages to study those who have not completed upper secondary school. Therefore, secondary elementary school GPA (GGPA) is used in this evaluation as well. Definitions of the national course test variables are given in Table A1 in Appendix A, and descriptive statistics are presented in Tables B1 and B2 in Appendix B.

## 4 Methodology

The main purpose of this evaluation is to study how graduates and students with different characteristics perform on different instruments. Previous evaluations have mainly concentrated on the covariance between SweSAT scores and grades to see to what extent one can predict the other. In order to make comparisons with these earlier evaluations, we will also report correlations between grade points and test results.

One way of studying if two instruments treat individuals differently is to form score differences, see e.g. Wikström and Wikström (2005), and to correlate the score difference with key characteristics. This will be the main focus of the present evaluation. We use as dependent variable the difference between the grade point average (or separated grade points) and the SweSAT score (or scores on the verbal and quantitative parts), where both scores are placed on the same scale (in the present case a standard normal scale). Key characteristics include gender, socioeconomic background, and ethnicity. If a coefficient associated with a particular trait is positive, we say that individuals in this group benefits from being graded relative to the SweSAT and vice versa.

One should, however, note that there are several possible interpretations as to why a particular difference exists. To mention a few, the instruments may firstly differ in what they measure. This is true in the present case, since the SweSAT should measure knowledge assumed important for success in higher education, which to a large extent is assumed to be related to core subject knowledge (reading, writing, arithmetics etc.), while the GPA is made up by a large number of subject grades to reflect a student's knowledge from certain criteria. As such, the GPA measures a broader spectrum of knowledge, and has a multiplicity of purposes. Secondly, differences can depend upon format and practices. Grades are aggregates of written tests, tasks performed in and out of classroom etc., while the SweSAT is a single multiple choice test. Thirdly, differences may arise because of selection of the group of students taking the SweSAT. For example, older students may perform better on the SweSAT other things being held constant, since they have greater experience, e.g. when it comes to vocabulary. Older students are also more likely to have performed worse in upper secondary school, which is one of the reasons why they have not already entered higher education. For these reasons one should not necessarily interpret differences as unjust or as a bias in one or the other instrument.

To facilitate interpretations, score differences are complemented with regression models where the instruments are studied separately. Here we study if the SweSAT (and grades) can be considered as being neutral over key characteristics. As above, it is natural to study how disaggregated scores (a quantitative and a verbal part of the SweSAT) vary with key characteristics, and we will therefore study separate grades and the verbal and quantitative parts of the SweSAT. Finally, although this evaluation uses a large amount of data, occasionally one cannot reject a particular hypothesis although it is not true. To avoid misinterpretations as far as possible, the study is conducted for two different SweSAT administrations, and we can check for consistency of the results over the two administrations. Also, since the observable instruments vary with the SweSAT test taker, and as mentioned above, samples with younger and older individuals are studied separately. We find it to be plausible to look for systematic differences that remain over administrations and across samples.

# 5 Results

## 5.1 Upper secondary school graduates

As mentioned above, this part of the study concerns the relationship between the SweSAT results and the grade point average and basic level course grades in upper secondary school for individuals who have graduated from upper secondary school. This section is divided into four different parts; the first concerns the covariance between SweSAT scores and GPA/course grades, the second concerns the determinants of grades and SweSAT scores, the third score differences, and the final part discusses differences between male and female participants. The final part is a natural step in the evaluation because of the gender differences that have been observed in previous studies.

### 5.1.1 Correlations

In Table 1, correlations between SweSAT scores and grade points for the two administrations are presented. The SweSAT/GPA correlations are 0.51 and 0.46 for the 2011B and 2012A respectively. These correlations can be compared to previous studies. Stage (1992) studied test takers in 1990 that graduated from upper secondary school the same year, and found an overall correlation coefficient of 0.51. In a similar study by Lexelius (2004), focusing on test scores and grades of students leaving upper secondary school in 1997 and taking the test in the autumn of 1996 or spring of 1997, the overall correlations were 0.52-0.53. Thus, the results from previous studies are quite similar to the results we obtain. Comparing the two administrations, we note that the overall correlation is slightly higher in the 2011B administration.

**Table 1: Partial correlations between SweSAT scores and grade points; upper secondary school graduates**

| | Test administration | |
|---|---|---|
| | 2011B | 2012A |
| SweSAT/GPA | 0.51 | 0.46 |
| SweSATQ/MATA | 0.66 | 0.61 |
| SweSATQ/MATB | 0.59 | 0.54 |
| SweSATV/ENGA | 0.55 | 0.56 |
| SweSATV/ENGB | 0.54 | 0.53 |

Note: Pairwise correlation coefficients. All coefficients are significantly determined on the 1 per cent level.

In the overall comparison, the correlation between the quantitative part of the SweSAT and the mathematics A-course (A representing the first course in upper secondary school) is the highest (0.66 and 0.61 for the 2011B and 2012A administrations respectively). Correlations between the verbal parts of the test and English grades are somewhat lower (0.55 and 0.56 when using English A-course grades). The subtest-grade correlations are higher than the test-GPA correlations. This is also what we expect, since the SweSAT is designed to capture core subject knowledge. Lexelius (2004) carried out similar estimations to those presented here, for subtests and separate grades regarding the students of 1997. The highest correlations were found between the English B-course grade and ERC (the English reading comprehension part of the SweSAT) (0.65), the English B-course grade and WORD (Swedish vocabulary subtest) (0.61), and mathematics A and DS (data sufficiency) (0.58). Although the studies are not entirely comparable, the comparison nevertheless shows no major differences over time.

We also estimate correlation coefficients for different subgroups; gender, immigration status, parental education, and educational experience. These are presented in tables A4 to A6. When studying the different subgroups, we note that correlation coefficients are reasonably stable over different subsamples. However, in two dimensions there are differences that we want to highlight. First, correlations differ conditional on parental education. The correlation coefficients are slightly higher if we restrict the sample to the group of students with highly educated parents. Second, the correlation is higher for students on academic track compared to students that participated in non-academic school programmes. These differences are more pronounced for the overall GPA/SweSAT comparison than for subtests/grades, although they largely remain when comparing mathematics A with SweSATQ.

### 5.1.2 Determinants of SweSAT scores and grade points

Next, we turn to the question if one or the other instrument favours different subgroups taking the test. To this end, we start by presenting results from regression analysis performed using the different instruments as dependent variables. The results are presented in tables A7 to A9.

It is generally known that male test takers perform better on the SweSAT than what females test takers do. The results in tables A7 and A8 confirm this. For the overall test result, males score on average 1.3 points higher than females. This corresponds to about 0.35 standard deviations of the overall SweSAT score. For the different parts, we note that the difference is the largest in the quantitative part; 2 points or 0.44 standard deviations, while it is smaller in the verbal part (0.5 points or 0.11 standard deviations). Table A7 also reveals another well-known fact; that females have higher GPA (1 point or 0.38 standard deviations). However, when looking at results regarding the core subject grades, one can note that males perform better in both mathematics and English (A-courses). The difference is larger in mathematics; 0.37 grade points in the 2011B sample corresponding to approximately 0.1 standard deviations. It is important to note that SweSAT test takers differ from the general student population in at least two respects. Firstly, they are on average more high achieving than the student population as a whole, and, secondly, a larger proportion have graduated from an academically oriented programme. The fact that we observe positive male-female grade differences in this sample is due to the latter fact. In academically oriented programmes the gender differences are small, and in some programmes males have higher grades in e.g. English and mathematics.[4]

Age contributes to explain differences in grade points and SweSAT results. First, older persons have lower GPA and course grades than younger. These results are moreover stable across the test administrations. The yearly average age difference is about 0.3 GPA points. Second, with respect to the SweSAT, the results are different. The overall SweSAT score does not depend on age. But if we look at the two parts of the SweSAT, older individuals do better on the verbal part and worse on the quantitative part. The yearly average age difference is about 0.28 in the verbal part, meaning that a 24 year old individual scores approximately 1.4 points (0.33 standard deviations) higher than a 19 year old individual other characteristics held constant. The (negative) age difference is less stable in the quantitative part, but of similar magnitude as the verbal part in the

[4] Information on grades, aggregated as well as for different programmes are available at the National Board of Educations web site; see http://www.jmftal.artisan.se/

2011B administration. These differences are qualitatively the same as those found in previous studies (Stage, 1992, Lexelius, 2004).

Previous studies found large differences in SweSAT results with respect to immigration status (Reuterberg & Hansen, 2001, 2003). We also find differences, both when it comes to the verbal as well as the quantitative part of the SweSAT. Second generation immigrants have, on average, lower scores than test takers with a Swedish origin, and first generation immigrants have lower scores than second-generation immigrants. The differences are the largest in the verbal part; a first generation immigrant has on average a 2.9 point lower score than a test taker with Swedish background, other things held constant. Similar differences exist when comparing grades and GPA, except for the English A-course. A second-generation immigrant has, on average, the same English A-course grade as a native Swede.

Educational background is here measured by the variables ACTRACK and HIGHEXP, the former distinguishing students who graduated from a theoretical upper secondary school programme, and the latter those who had some experience with higher education at the time the test was administered. Core grades and GPA, as well as SweSAT results, are higher for academic track graduates and those with higher education experience. Finally, socioeconomic background is measured by parental education, influences grades and SweSAT results. This is especially important for those individuals where at least one parent has post upper secondary education.

### 5.1.3 Score differences

One central question when comparing grades and SweSAT results concerns how the different instruments rank students. As mentioned in section 4, we form the difference between the GPA and the SweSAT, both measured on a standard normal scale. This means that coefficient estimates are interpreted as standard deviations of either instrument. Although the GPA-SweSAT difference is the relevant measure for discussing how the two instruments function in the admission of university students, we nevertheless make similar comparisons using results for the SweSAT parts and the a-course grade points. The results are presented in tables 2 and 3.

**Table 2: Score differences GPA versus SweSAT; upper secondary school graduates**

| | Test administration | |
|---|---|---|
| Independent variable | 2011B | 2012A |
| Constant | 3.188** | 3.551** |
| MALE | -0.716** | -0.773** |
| AGE | -0.123** | -0.133** |
| FIRSTGEN | 0.073* | 0.085** |
| SECONDGEN | 0.051* | 0.028 |
| FORBORN | 0.149 | 0.170 |
| MIDEDU | -0.036 | -0.060 |
| HIGHEDU | -0.082* | -0.104** |
| ACTRACK | -0.250** | -0.206** |
| KOMVUX | -0.611** | -0.552** |
| HIGHEXP | -0.138** | -0.098** |
| D11B | - | -0.167** |
| R-squared | 0.203 | 0.216 |
| N | 19,708 | 22,030 |

Note: The dependent variable is the difference between GPA and SweSAT, both measured on a standard normal scale. A positive coefficient means that an individual scores higher on GPA, and a negative coefficient that the SweSAT score is higher. ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

Let us start by presenting the results concerning the overall GPA-SweSAT difference. As can be seen from table 2, the relative rankings vary over most of the characteristics that are included. The estimated coefficients in most cases show differences that are consistent over the two test administrations. The largest effects on the relative rankings come from gender and age. Males do relatively better on the SweSAT compared to the GPA. On average, a male student is placed 0.7-0.8 standard deviations higher in the overall distribution of the SweSAT. Older students also do relatively better on the SweSAT. This is entirely due to the fact that older students participating in the SweSAT have lower GPA than younger students. The same goes for students who finished upper secondary school through the municipal adult education. One can interpret the effects of AGE and KOMVUX as a second chance mechanism that makes it possible for these groups to get admission into higher education.

We also note a significant but small difference for first generation immigrants. The GPA ranks this group slightly higher than natives. The SweSAT and the GPA rank second-

generation immigrants the same way as natives. Educational experience affects the relative rankings. First, individuals that graduated from academic track programmes do better on the SweSAT than on the GPA compared to students with other educational experience. Academic track students are likely to be better prepared for the test, explaining their higher scores on the SweSAT. However, as we mentioned earlier, grades are not necessarily comparable between different educational programmes, and the estimated differences in table 1 may be explained by more lenient grading in non-academic track programmes. Second, students with experience from higher education do better on the SweSAT. Social background also has explanatory power on the GPA-SweSAT difference. Individuals from a high education environment fare better by the SweSAT. Although this difference is small, it may indicate that the SweSAT contributes to the social recruitment bias.

**Table 3: Score differences grades versus SweSAT quantitative and verbal parts; upper secondary school graduates**

| | Dependent variable/Test administration | | | |
|---|---|---|---|---|
| | MATA/SWESATQ | | ENGA/SWESATV | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| Constant | 1.382** | 1.723** | 3.213** | 3.184** |
| MALE | -0.379** | -0.430** | -0.074** | -0.111** |
| AGE | -0.053** | -0.065** | -0.148** | -0.139** |
| FIRSTGEN | -0.033 | 0.024 | 0.249** | 0.237** |
| SECONDGEN | -0.068** | -0.030 | 0.395** | 0.388** |
| FORBORN | -0.115 | -0.137 | 0.425 | 0.336 |
| MIDEDU | 0.013 | -0.007 | -0.014 | -0.013 |
| HIGHEDU | -0.027 | -0.017 | -0.083* | -0.070* |
| ACTRACK | -0.081** | -0.019 | -0.037** | -0.040** |
| KOMVUX | -0.383** | -0.154 | 0.006 | -0.108 |
| HIGHEXP | -0.150** | -0.146** | -0.247** | -0.223** |
| D11B | - | -0.198** | - | -0.113** |
| R-squared | 0.078 | 0.097 | 0.121 | 0.111 |
| N | 19,704 | 22,026 | 19,698 | 22,015 |

Note: The dependent variables are constructed as the difference between the course grade and the SWESATQ/V, both measured on a standard normal scale. A positive coefficient means that an individual scores higher on the course grade, and a negative coefficient that the SweSAT part has a higher score. ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

Let us also briefly discuss the results from the comparison of core grades and the SweSAT parts. Given that we believe that the core grades better should correspond to the SweSAT parts than the GPA does, it is natural to expect that the estimated differences are smaller in this case. Indeed, looking at the results in table 3, the estimated coefficients are generally of smaller magnitude than the corresponding estimates in table 2. This is also evident by looking at the degree of explanation (R-squared), which is considerably lower in the SweSAT/course grade comparisons. Comparing estimates between the quantitative and verbal rankings (MATA/SWESATQ versus ENGA/SWESATV), it is notable that the estimated coefficients are generally larger in the verbal ranking. The exception to this is the male-female difference, which is relatively large in the quantitative comparison and small (but still statistically significant) in the verbal comparison. In the quantitative comparison, it should be noted that many of the estimated coefficients are not consistent comparing the two test administrations. In fact, only three explanatory variables have coefficients that are consistent over the administrations (statistically significant of the same sign and of similar magnitude) in the quantitative comparison; MALE, AGE, and HIGHEXP. Therefore, it is in the verbal part that the major differences arise.

### 5.1.4 Male-female score differences

As previously shown, there are important gender differences both when it comes to grades and SweSAT results. We can potentially obtain a better understanding of these differences by looking at male-female differences within different groups. To accomplish this, the regression models of the previous subsection are extended by including interaction terms. In appendix A, the results concerning the relative ranking of the instruments are presented in tables A10 and A11. The results regarding the determinants of the instruments can be obtained from the authors upon request.

Comparing the GPA-SweSAT relative rankings (table A10), we find very limited evidence of interaction effects. Male-female differences do not vary with age and social background. In the 2012A administration, first- and second generation male immigrants are ranked relatively higher than females by GPA. The SweSAT score is, however, independent of the male immigrant interaction variables, meaning that male-female differences are entirely due to differences in grading. We also pick up a positive

coefficient from higher education experience for males in the 2011B sample, but this is close to zero in the 2012A sample.

The comparison of course grades and the SweSAT parts yield different results from the overall comparison (table A11). First, looking at the quantitative comparison, we note that there is a rather large difference for natives; the male coefficient is larger in absolute value than the corresponding estimate in table 3. However, the immigrant interaction effects are positive, meaning that, among the immigrant groups, males are either graded more lenient than females or score worse on the quantitative parts of the SweSAT. Looking closer at these relations, it appears to be mainly driven by differences in grading. The exception to this is first generation immigrant males who score lower on the SweSAT in the 2012A administration. There is also some (although not consistent over the two administrations) evidence of a positive age difference.

While immigrant males appear to fare better being graded in the quantitative comparison, the opposite holds in the verbal comparison. Immigrant males do relatively better on the SweSATV. The estimated differences are mainly due to differences in grading. In the verbal comparison, there is a negative age effect, meaning that older male students do relatively better on the SweSATV. However, the male-female difference among natives is close to zero, and the estimated difference in table 3 is therefore mainly due to the differences in the immigrant groups.[5] To sum up, male-female differences remain largely unexplained (constant over different characteristics) in the quantitative comparison, while it is to a large extent explained by differences among the immigrant groups in the verbal comparison.

[5] Assuming that statistically insignificant coefficients are zero, the average male-female difference for 20-year old natives is 0.508-0.026*20=-0.012 using the point estimates in column 3 (2011B), table A11.

## 5.2 Upper secondary school students

As mentioned in the data section, it is not possible to observe individual grades (and the GPA) for those who have not yet finished upper secondary school. This section therefore concentrates on comparing the SweSAT quantitative and verbal parts with national course tests that were carried out during upper secondary school.

### 5.2.1. Correlations

As with upper secondary school graduates, the different instruments are compared by pairwise correlations. As mentioned above, the quantitative (verbal) part of the SweSAT is correlated with the grades obtained in the national tests in key subjects. For the purpose of comparison with previous studies we also report the correlation between the overall SweSAT score and the GPA from secondary school (GGPA). The results are presented in table 4 and tables B2 and B3 in appendix B. The SweSAT/GGPA correlations are 0.53-0.54 in the two administrations, which is well in line with previous evaluations. Correlations between SweSATQ/V and national test grades are generally higher; the exception is the Swedish B-level national course test, where the correlation is considerably lower (0.34-0.35). This supports previous findings since Lexelius (2004) also found lower correlation between course grades in Swedish and the SweSAT in comparison with other core subjects.

**Table 4: Pairwise correlations between SweSAT scores and grade points/ national test scores; upper secondary school students**

| | Test administration | |
|---|---|---|
| Group | 2011B | 2012A |
| SweSAT/GGPA | 0.54 | 0.53 |
| SweSATQ/NMATA | 0.68 | 0.64 |
| SweSATQ/NMATB | 0.70 | 0.66 |
| SweSATV/NENGA | 0.58 | 0.63 |
| SweSATV/NENGB | 0.61 | 0.59 |
| SweSATV/NSVB | 0.34 | 0.35 |

Note: Pearson correlation coefficients. All coefficients are significantly determined on the 1 per cent level.

Tables B2 and B3 present correlations for different subgroups for the national test grades in core courses. The correlations vary some over subgroups, but there does not seem to be any particular group where the correlations are systematically higher or

lower. Comparing the results presented here with those concerning graduates, we note that the correlations between national test grades and SweSAT scores are slightly higher than those between course grades and SweSAT scores. One plausible explanation for this is that the core grades are determined on the basis of a combination of tests and other forms of assessment, while the national course test is more similar in format to the SweSAT (although the types of tasks in the tests are different).

### 5.2.2 Score differences

In Table 5 below, estimates of score differences between the SweSAT and the national course tests (A-courses) are presented. Given that national course tests are more similar in format to the SweSAT than the assessments underlying the course grades, we expect to find less systematic variation in the score differences between different groups of test takers. The degree of explanation (R-squared) varies between 0.04 and 0.07 in table 5. Although not perfectly comparable, the corresponding range in table 3 (course grades/SweSAT) is 0.08-0.12. Thus, our conjecture seems to be consistent with the data at hand and what was found from studying the pairwise correlations.

There is a consistent male-female difference in the quantitative comparison, but there is no difference in the verbal comparison. In the verbal comparison, immigrations status explains a part of the difference. First- and second generation immigrants appear to score higher on the national test, although the coefficient associated with first generation immigrants fails to reach significance in the 2011B-administration. There is also an inconsistency in the verbal comparison regarding socioeconomic background. In the 2011B sample, students from low-educated homes perform lower on the SweSAT than on the national test compared to other students. In the 2012A sample this effect is not present. The difference observed in the 2011B administration may be a coincidence, but it may also be due to an identification problems or a problem with outliers as the number of observations covering students coming from low-educated homes is very small in this case. We also note that experience with SweSAT matters and that younger students have higher national test grades than SweSAT scores.

**Table 5: Score differences between national course test grades and SweSAT quantitative and verbal scores; upper secondary school students**

| | Dependent variable/Test administration | | | |
|---|---|---|---|---|
| | NMATA/SWESATQ | | NENGA/SWESATV | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| Constant | 0.517** | 0.527** | 0.596** | 0.044 |
| MALE | -0.374** | -0.374** | 0.050 | 0.029 |
| BORN94 | 0.195** | 0.086 | 0.288** | 0.181** |
| FIRSTGEN | -0.143 | -0.083 | 0.091 | 0.221** |
| SECONDGEN | -0.020 | -0.040 | 0.493** | 0.343** |
| FORBORN | 0.115 | -0.072 | 0.234 | 0.099 |
| MIDEDU | 0.008 | -0.070 | 0.636** | -0.067 |
| HIGHEDU | -0.002 | 0.009 | 0.468** | -0.066 |
| ACTRACK | -0.146* | -0.077 | -0.109 | -0.075 |
| HIGHEXP | -0.704* | -0.513* | -0.034 | 0.063 |
| D11B | - | -0.129** | - | -0.171** |
| R-squared | 0.074 | 0.062 | 0.061 | 0.040 |
| N | 1,076 | 2,120 | 1,565 | 3,056 |

Note: The dependent variables are constructed as the difference between the national test score and the SWESATQ/V score, both measured on a standard normal scale. A positive coefficient means that an individual is ranked higher by the national test, and a negative coefficient that the SweSAT ranks the individual higher. ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

Tables B5 and B6 present the estimation results regarding the determinants of SweSAT and national course tests. As the results are very similar (at least qualitatively) to those regarding upper secondary school graduates, we do not comment upon them further. However, let us briefly compare the score difference results across the student and the graduate samples. Most of the differences found in the sample of graduates are also present in the student sample. The size of the gender difference in the quantitative comparison is 0.37 standard deviations (see table 5), which is more or less the same as the difference found in the study of graduates. There is no gender difference found in the verbal part in the student sample, although a small difference in the graduate sample. The results also show that immigration status matters. The parameter estimates are qualitatively the same in both samples. Comparing the estimates regarding socioeconomic background, we found no or small differences in the graduate sample,

which is why we are reluctant to interpret the corresponding estimate in the verbal comparison of 2011B as a true effect.

Finally, age differences go in the same direction in both samples, but likely for different reasons. The age differences observed in table A8 imply that older individuals do better on the verbal part of the SweSAT and worse on the quantitative part. However, older individuals have lower grades (table A9), and this is what drives the estimate of the score differences. So, in this respect, the SweSAT as a second chance instrument appears to be working. In the student sample, the selection mechanism of individuals taking part of the SweSAT appear to be different; in this sample, those born 1994 have higher national test scores as well as SweSAT scores, indicating that younger students in this comparison are slightly more high-achieving.

## 6 Conclusions and discussion

This study has focused on how students are ranked on the basis of the two selection instruments to higher education in Sweden; the grade point average from upper secondary school (GPA), and SweSAT, an admissions test that serves as the alternative to the GPA. The main findings are in line with previous research, especially similar evaluations of previous forms of tests and previous forms of grades (Lexelius, 2004, Stage, 1992). In the present evaluation a large number of results have been presented. The conclusions that we draw are based on general patterns found across the two administrations and across the two samples. Our main conclusions are:

- The correlations between the SweSAT and the different grading instruments are at the same level as in previous studies. The correlation coefficient between the SweSAT and the GPA is approximately 0.5. The correlation between the course grades (A-courses) and the different parts of the SweSAT are higher, indicating that the SweSAT corresponds more closely to the core courses than the GPA. There are no systematic differences in the correlations over subgroups when comparing all available instruments.
- Gender is the most important trait when it comes to score differences between grades and the SweSAT. In the samples of SweSAT test takers, males have lower GPA but perform better on the SweSAT than females. Males also perform better on separate grades and on the national course tests. Males improve their position in the overall distribution by the SweSAT compared to grades and the GPA. There appear to be little systematic variation in the male-female differences over different traits. In the verbal comparison (SweSAT verbal part versus English A-course grade) females with immigrant background benefit from being graded, while the opposite is true for the quantitative comparisons.
- Test takers with immigrant background do better on the grade scale than on the SweSAT. This is due to differences in the verbal comparison, while the quantitative parts tend to treat natives and immigrants more symmetric.
- We find only partial and quantitatively small differences with respect to social background measured by parental education. Foremost individuals coming from high education backgrounds do slightly better on the GPA.

- Academic track students do better on the SweSAT compared to the GPA. They also perform better on the verbal part of the SweSAT than the English A-course grade, while there is no difference in the quantitative comparison. Thus, it seems that there are other grades where non-academic track students are being graded more highly than academic track students.
- Selection into the pool of SweSAT test takers may play an important role in the comparison of different instruments. In the data, we observe consistent differences with respect to the age of the test takers. Older students have lower grades but perform better on the SweSAT.

The results above, and their similarities to findings in previous evaluations, could be interpreted as a status quo, with no changes to the recruitment regardless of how the selection instruments are modified; certain groups of students perform better on grades than others, some perform better on tests, and some perform equally good on both. However, the selection model is complex and there are some interesting aspects to the results that should be discussed, and also investigated further.

The main purpose of the SweSAT is to function as a second chance in the selection into higher education. Students who have the required knowledge but, for some reason, cannot compete with their GPA, can take this alternative route to attractive university programmes. The intention is to ensure fairness (everybody should be given the opportunity of a second chance) although the main reason for introducing the test was to broaden the recruitment to higher education. For obvious reasons, the task of investigating to what extent the SweSAT serves these purposes is a difficult task that goes far beyond the purpose of this evaluation. However, one necessary condition for the SweSAT to serve the purpose of function as a second chance is that potential entrants into higher education who, for whatever reason, did not perform well in upper secondary school, de facto have the opportunity and the knowledge required to gain access into higher education. This means that at least some individuals must “climb the ladder” to be among those selected. Is this likely to happen?

The test is not given exclusively to underrepresented groups or students with low grades. Originally such an intention existed, when the test was open only to older students without a GPA from upper secondary school. For some reason the eligibility requirements were not so emphasised back then, and the test was expected to indirectly give similar information. Nowadays anyone can take the SweSAT and, if he or she meets the eligibility requirements and achieves a test score that is high enough, may gain access to higher education through this quota group. This also means that a student who can compete with his or her GPA can take the test and compete in two quota groups. It is well known that high performing students are more prone to take the test than other students. This is natural, since students on academic track often aim for selective university programmes, and for this reason may see the test as a way of maximising his or her opportunity to get admitted. An important question is if this means that students are ranked similarly in both admission groups, and historically weak students therefore are prevented from gaining access into higher education. This study shows that there is a positive correlation between the instruments, although less than perfect. A less than perfect correlation is good in the sense that the ranking of students otherwise would have been preserved and the second chance would not have had any practical meaning. But even so, the correlation coefficient between the instruments is not a sufficient condition to gain knowledge of the de facto opportunities of those with low GPA.

To shed some light on the possibility to improve one's position in the ranking of the SwSAT versus the GPA, we can make a crude estimate of the number of weak students (according to the GPA) that will be in the pool of applicants admitted by the SweSAT by using the data at hand. Suppose that half of the test takers will gain access to higher education and the other half will not. According to the general rules a university must admit at least 30 per cent of the students by the SweSAT. Under this assumption 15 per cent of the test takers, assuming that all test takers will apply to a similarly selective programme, will be admitted from the SweSAT group. The question we pose is: what is the proportion of students in the bottom half of the grade distribution that will score well enough to be among the top 15 per cent in the SweSAT distribution? This will then form a lower bound on the number of "weak" applicants who will get access to higher education. Using the 2011B (2012A) administration we calculate that about 500 (600)

students out of the 9500 (11000) students in the bottom half will score well enough to be among the top 15 per cent. This corresponds to about 5.3 to 5.5 per cent of the bottom half. Of course, there are several complications that make this figure unreliable such as repeated test taking and variable application rates among programmes and universities. Nevertheless, the example shows that some students actually do improve their position more than marginally when taking the SweSAT.

We can also say something about those students who improve their position in the way described above by studying their characteristics. Three of the traits stand out. First, in the 2011B (2012A) administration 84 (87) per cent of them are men. This is also what we expect given that males outperform females on the SweSAT and have lower GPA. In the high GPA group (those who would be accepted into higher education in the absence of the SweSAT), the proportion of men is about 43 per cent. The SweSAT therefore contributes to even out male-female representation in the group of admitted students. Second, first- and second generation immigrants are slightly underrepresented in this group. As we have shown above, immigrants benefit from being graded relative to the SweSAT. In this case 7.8 (6.9) are first- or second generation immigrants in comparison to 9.1 (8.6) per cent in the high GPA group. Finally, students who improve are about one year older on average than the average age in the top half with respect to the GPA.
The conclusions from this study, and the reasoning above is that there are some group differences in terms of how students are ranked on the basis of the two selection instruments. There is some uncertainty to the extent of these group differences since we know that there is a selection to the test taking population corresponding to groups who are higher performing on the test (male students, students with a high socioeconomic background, and student with a Swedish origin).

A fair number of students do change their rank in the selection to higher education by taking the test, and actually uses this opportunity as a second chance. Some of these would otherwise have limited possibility to get admitted to any university education where there is a selection. Another interesting aspect is that the second chance also makes the gender differences in higher education smaller, since it is foremost beneficial for male applicants. However, how this second chance really should be viewed – if it

should be reserved for students who have underperformed on grades, or is also relevant for already high performing students – is something that should be discussed from a policy perspective. Future research should focus on investigating selection effects further, and also on how students who have used the second chance perform once admitted.

## References

Allergren, B. (1995). *Testprestation vid början och slutet av högskoleutbildning. En tvärsnittsstudie med studerande vid två utbildningslinjer* [Test performance at the start, the middle and at the end of higher education. A cross-sectional study featuring students of two study-programs] (Licentiate dissertation). Umeå University: Department of Education.

Andersson, E. (2003). Who is a successful student from the perspective of university teachers in two departments? *Scandinavian Journal of Educational Research, 47*(5), 543-559.

Andersson, E. (2005). Högskolestudenters syn på högskoleprovet och strategier för att antas till högre utbildning [University students' views on the Swedish Scholastic Assessment Test (SweSAT) and strategies for admission to higher education]. *Pedagogisk Forskning i Sverige, 10*(2), 81-98.

Andersson, E., & Grysell, T. (1992). *Studenters syn på studieframgång. En pilotstudie inför studier av studieframgångskriteriet inom ramen för prognosprojektet vid Högskoleprovet* [Students opinion on the concept of academic success] (PM Nr 67). Umeå University: Department of Educational Measurement.

Andersson, E., & Grysell, T. (1995). *En beskrivning av fem utbildningar vid Umeå universitet. En genomgång av ekonomlinjen, sociala linjen, teknisk fysik, läkarlinjen samt ämneslärarlinjen med avseende på organisation och utformning* [A description of five study-programmes at Umeå University] (PM Nr 104). Umeå University: Department of Educational Measurement.

Carlstedt, B., & Gustafsson, J.-E. (2005) Construct validation of the Swedish Scholastic Aptitude Test by means of the Swedish Enlistment Battery. *Scandinavian Journal of Psychology, 46*(1), 31-42.

Cliffordson, C. (2004a) Betygsinflation i de målrelaterade gymnasiebetygen [Inflation in goal-related grades from upper-secondary school]. *Pedagogisk Forskning i Sverige, 9*(1), 1-14.

Cliffordson, C. (2004b) Effects of practice and intellectual growth on performance on the Swedish Scholastic Aptitude Test (SweSAT). *European Journal of Psychological Assessment, 20*(3), 192-204.

Cliffordson, C. (2004c) De målrelaterade gymnasiebetygens prognosförmåga. [The predictive validity of goal-related grades from upper secondary school]. *Pedagogisk Forskning i Sverige, 9*(2), 129-140.

Cliffordson, C. (2006). .Selection effects on applications and admissions to medical education with regular and step-wise admission procedures. *Scandinavian Journal of Educational Research, 50*(4), 463-482.

Cliffordson, C., & Askling, B. (2006) Different grounds for admission: Its effects on recruitment and achievement in medical education. *Scandinavian Journal of Educational Research, 50*(1), 45-62.

Henriksson, W., & Bränberg, K. (1992). *Effekten av upprepat provtagande. En studie av poängförändringar från första till andra provgenomförandet*. (Pm nr 70) Umeå: Umeå universitet, Enheten för pedagogiska mätningar

Henriksson, W., & Wolming, S. (1998). Academic performance in four study programmes: A comparison of students admitted on the basis of GPA and SweSAT scores, with and without credits for work experience. *Scandinavian Journal of Educational Research, 42*(2), 135-150.

Henriksson, W., & Törnkvist, B. (2002). *The effects of repeated test taking in relation to the test taker and the rules for selection to higher education in Sweden* (EM No 41). Umeå: Umeå University, Department of Educational Measurement.

Henrysson, S. (1984a). *Prognos av framgång i högre studier, metoder och problem* [Prediction of success in higher education. Methods and problems] (Nr 12). Umeå: Umeå University, Department of Education.

Henrysson, S. (1984b. *Studenterna och antagningen till högskolan. Intervjuer med lärare inom grundutbildningen* [Admission to higher education – teachers views on the admission] (Nr 13). Umeå: Umeå University, Department of Education.

Henrysson, S., & Henrysson, S. (1984). *Poängsättning av relevant arbetslivserfarenhet* [Work experience as merits in admission to higher education] (Nr 18). Umeå: Umeå University, Department of Education.

Henrysson, S., Kriström, M., & Lexelius, A. (1984) *Meritvärdering och studieprognos. Några undersökningar om antagningssystemets effekter* [Evaluation of merits and academic performance. Studies of the effects of selection regulations] (Nr 21). Umeå: Umeå University, Department of Education.

Lexelius, A. (2004). *Högskoleprovet och de målrelaterade betygen. En studie av de första eleverna med de nya gymnasiebetygen.[*SweSAT and the goal referenced gardes. A study of the first students with the new upper secondary grades]. (BVM No.8). Umeå: Umeå University, Department of Educational Measurement.

Lexelius, A., & Wolming, S. (1995). Högskoleprovets prediktiva förmåga. En studie av fyra utbildningslinjer i högskolan [The predictive valididty of the SweSAT] (unpublished manuscript). Department of Education: Umeå University.

Lyrén, P-E. (2008). Prediction of Academic Performance by Means of the Swedish Scholastic Assessment Test, *Scandinavian Journal of Educational Research, 52*(6), 565-581.

Lyrén, P-E. (2009). *A perfect score. Validity arguments for college admissions tests.* (Thesis). Umeå: Umeå University, Department of Educational Measurement.

Löfgren, K. (2005). *Validation of the Swedish university entrance system. Selected results from the VALUTA-project 2001-2004* (EM No. 53). Umeå: Umeå University, Department of Educational Measurement.

Löfgren, K., & Törnkvist, B. (2004a) *Jämförelser mellan studeranden i olika antagningsgrupper som har registrerats på socionomprogrammet* [Comparisons between social work-students from different admission groups] (PM No. 191). Umeå: Umeå University, Department of Educational Measurement.

Löfgren, K., & Törnkvist, B. (2004b) *Jämförelser mellan studeranden i olika antagningsgrupper som har registrerats på ekonomprogram* [Comparisons between business administration-students from different admission groups] (PM No. 198). Umeå University: Department of Educational Measurement.

Reuterberg, S-E. (1997). *Gender differencs on the Swedish scholastic aptitude test.* (Report 1997:2). Gothenburg University: Department of Education and Education Research.

Reuterberg. S-E. (2003) *Vilken betydelse har utländsk bakgrund för resultatet på Högskoleprovet: Del II* [What meaning has national background on the SweSAT scores: Part II] (2003:23R). Stockholm: National Agency for Higher Education.

Reuterberg. S-E., & Hansen, M. (2001) *Vilken betydelse har utländsk bakgrund för resultatet på Högskoleprovet: Del I* [What meaning has national background on the SweSAT scores: Part I] (2001: 3 R). Stockholm: National Agency for Higher Education.

Stage, C. (1992). Skillnader mellan betyg och högskoleprovsresultat 1991. (PM Nr 62). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (2004a). *Gruppskillnader i resultat på Högskoleprovet* [Group differences in the SweSAT scores] (PM Nr 192). Umeå University: Department of Educational Measurement.

Stage, C. (2004b). *Socialgruppskillnader i resultat på Högskoleprovet* [Social-group differences in SweSAT-scores] (BVM Nr 11). Umeå: Umeå University, Department of Educational Measurement.

Stage, C. (2004c). *Entrance to Higher Education in Sweden.* (Em 51:2004). Umeå: Umeå University, Department of Educational Measurement.

Stage, C., & Ögren, G. (2010). *Ett nytt högskoleprov: bakgrund och konsekvenser* (BVM Nr. 42). Umeå: Umeå University, Department of Educational Measurement.

Svensson, A. (2004). Gymnasiebetyg eller högskoleprov som urvalsinstrument? Fallet civilingenjörsutbildningarna [Grades or achievement test for selection to civil engineer education? The master of engineering case]. *Pedagogisk Forskning i Sverige, 9*(1), 15-36.

Svensson, A., & Nielsen, B. (2005). Vilka kommer in på juristutbildningen och hur klarar man studierna? [Law education in Sweden: Who is admitted and who is successful] *Pedagogisk Forskning i Sverige, 10*(1), 1-23.

Törnkvist, B. & Henriksson, W. (2004a) *SweSAT repeat* (Em No 46). Umeå: Umeå University, Department of Educational Measurement.

Törnkvist, B. & Henriksson, W. (2004b) *Repeated test taking. Differences between social groups* (Em No 47). Umeå: Umeå University, Department of Educational Measurement.

Törnkvist, B., & Henriksson, W. (2006) *Validity issues concerning repeated test taking of the SweSAT* (EM No. 56). Umeå: Umeå University, Department of Educational Measurement.

Wester, A. (1995). *Ambitions and attitudes toward studies and study results. Interviews with students of the Business Administration study program in Umeå* (EM No 14). Umeå: Umeå University, Department of Educational Measurement.

Wikström, C. (2005a). Criterion-referenced measurement for educational evaluation and selection (Doctoral dissertation). Department of Educational Measurement: Umeå University.

Wikström, C. (2005b). Grade stability in a criterion referenced grading system: The Swedish example. *Assessment in Education: Principles, Policy and Practice, 12*(2), 125-144.

Wikström, C., & Wikström, M. (2005). Grade inflation and school competition: An empirical analysis based on the Swedish upper secondary schools. *Economics of Education Review, 24,* 309-322.

Wikström, C. Wikström, M. & Lyrén (2009). *Prediction of study success: should selection instruments measure cognitive or non-cognitive factors?* Paper presented at the 35th annual IAEA conference, Brisbane, Australia 13-18 September 2009.

Wolming, S. (2001) Att värdera urvalsinstrument: Några reflektioner over begränsningar och möjligheter [To evaluate selection instruments: Some reflections over restrictions and possibilities]. *Pedagogisk Forskning i Sverige, 6*(2), 122-130.

Zwick, R. (2002). *Fair Game? The use of standardised admissions tests in Higher Education.* New York: RoutledgeFarmer

Zwick, R. (Ed.) (2004). *Rethinking the SAT. The Future of Standardised Testing in University Admission.* New York: Routledge

Åberg-Bengtsson, L. (2005). Separating quantitative and analytic dimensions in the Swedish Scholastic Aptitude Test (SweSAT). *Scandinavian Journal of Educational Research, 49*(4), 359-383.

## Appendix A: Upper secondary school graduates

### Table A1: Variable definitions

| Variable | Description |
|---|---|
| **SWESAT** | The normed score on the SweSAT |
| **SWESATQ** | The normed score on the SweSAT quantitative part (includes the subtests DS, XYZ, QC, DTM) |
| **SWESATV** | The normed score on the SweSAT quantitative part (includes the subtests WORD, READ, ERC, SC) |
| **GPA** | Grade point average, upper secondary school |
| **MATA** | Course A grade in mathematics, upper secondary school |
| **MATB** | Course B grade in mathematics, upper secondary school |
| **ENGA** | Course A grade in English, upper secondary school |
| **ENGB** | Course A grade in English, upper secondary school |
| **MALE** | Indicator variable that takes the value one if the participant is male |
| **AGE** | Age of the test participant defined as the year of the test administration less the birth year |
| **BORN94** | Indicator variable that takes the value one if the test participant is born in 1994 and zero if he/she is born in 1993 |
| **FIRSTGEN** | Indicator variable that takes the value one if the test participant and his/hers parents are born outside of Sweden |
| **SECONDGEN** | Indicator variable that takes the value one if the test participant is born in Sweden and his/hers parents are born outside of Sweden |
| **FORBORN** | Indicator variable that takes the value one if the test participant is born outside of Sweden and at least one of the parents is born in Sweden |
| **MIDEDU** | Indicator variable that takes the value one if the parent of the test participant with the highest education has upper secondary school education |
| **HIGHEDU** | Indicator variable that takes the value one if the parent of the test participant with the highest education has post-upper secondary school education |
| **ACTRACK** | Indicator variable that takes the value one if the test participant graduated from a theoretical upper secondary school programme |
| **KOMVUX** | Indicator variable that takes the value one if the test participant completed upper secondary school within the municipal adult education (kommunala vuxenutbildningen) |
| **HIGHEXP** | Indicator variable that takes the value one if the test participant had experience from university by the time the test was administered |
| **PART11** | Indicator variable that takes the value one if the test participant participated in both test administrations |

**Table A2: Descriptive statistics. Participants in the SweSAT autumn 2011, upper secondary school graduates**

| Variable | Nobs | Mean | Std.Dev. | Min | Max |
|---|---|---|---|---|---|
| SWESAT | 23214 | 9.24 | 3.95 | 0 | 20 |
| SWSSATQ | 23214 | 8.98 | 4.56 | 0 | 20 |
| SWESATV | 23214 | 9.01 | 4.41 | 0 | 20 |
| GPA | 19708 | 14.78 | 2.66 | 1.1 | 20.6 |
| MATA | 19704 | 14.73 | 4.00 | 0 | 20 |
| MATB | 18130 | 12.67 | 4.81 | 0 | 20 |
| ENGA | 19698 | 15.15 | 3.73 | 0 | 20 |
| ENGB | 18153 | 14.49 | 4.04 | 0 | 20 |
| MALE | 23214 | 0.502 | | 0 | 1 |
| AGE | 23214 | 20.94 | 1.81 | 19 | 25 |
| Firstgen | 23214 | 0.070 | | 0 | 1 |
| Secondgen | 23214 | 0.074 | | 0 | 1 |
| FORBORN | 23214 | 0.005 | | 0 | 1 |
| MIDEDU | 23214 | 0.316 | | 0 | 1 |
| HIGHEDU | 23214 | 0.646 | | 0 | 1 |
| ACTRACK | 23214 | 0.648 | | 0 | 1 |
| KOMVUX | 23214 | 0.007 | | 0 | 1 |
| HIGHEXP | 23214 | 0.181 | | 0 | 1 |

**Table A3: Descriptive statistics. Participants in the SweSAT spring 2012, upper secondary school graduates**

| Variable | Nobs | Mean | Std.Dev. | Min | Max |
|---|---|---|---|---|---|
| SWESAT | 27075 | 9.16 | 3.750 | 0 | 20 |
| SWSSATQ | 27075 | 8.71 | 4.25 | 0 | 20 |
| SWESATV | 27075 | 9.12 | 4.24 | 0 | 20 |
| GPA | 22030 | 14.51 | 2.64 | 2.9 | 20 |
| MATA | 22026 | 14.43 | 3.97 | 0 | 20 |
| MATB | 19981 | 12.29 | 4.83 | 0 | 20 |
| ENGA | 22015 | 14.97 | 3.71 | 0 | 20 |
| ENGB | 20084 | 14.26 | 4.02 | 0 | 20 |
| MALE | 27075 | 0.511 | | 0 | 1 |
| AGE | 27075 | 21.87 | 1.79 | 20 | 26 |
| Firstgen | 27075 | 0.081 | | 0 | 1 |
| Secondgen | 27075 | 0.073 | | 0 | 1 |
| FORBORN | 27075 | 0.005 | | 0 | 1 |
| MIDEDU | 27075 | 0.334 | | 0 | 1 |
| HIGHEDU | 27075 | 0.621 | | 0 | 1 |
| ACTRACK | 27075 | 0.601 | | 0 | 1 |
| KOMVUX | 27075 | 0.010 | | 0 | 1 |
| HIGHEXP | 27075 | 0.203 | | 0 | 1 |
| PART11B | 27075 | 0.297 | | 0 | 1 |

**Table A4: Pairwise correlations between GPA and the SweSAT score, upper secondary school graduates**

| | Test administration | |
|---|---|---|
| Group | 2011B | 2012A |
| All | 0.508 | 0.463 |
| Gender | | |
| Men | 0.542 | 0.508 |
| Women | 0.582 | 0.541 |
| Immigration status | | |
| First generation | 0.480 | 0.471 |
| Second generation | 0.530 | 0.487 |
| Natives | 0.500 | 0.444 |
| Parental education | | |
| Low | 0.407 | 0.396 |
| Medium | 0.424 | 0.380 |
| High | 0.507 | 0.460 |
| Educational experience | | |
| Academic track | 0.522 | 0.482 |
| No academic track | 0.427 | 0.381 |
| Higher education | 0.502 | 0.443 |
| No higher education | 0.481 | 0.422 |

Note: Pairwise correlation coefficients. All coefficients are significantly determined on the 1 per cent level.

**Table A5: Pairwise correlations between course grades and the SweSAT scores, upper secondary school graduates, autumn 2011**

| | SWESATQ | | SWESATV | |
|---|---|---|---|---|
| Group | MATA | MATB | ENGA | ENGB |
| All | 0.656 | 0.591 | 0.549 | 0.535 |
| Gender | | | | |
| Men | 0.664 | 0.605 | 0.517 | 0.512 |
| Women | 0.664 | 0.618 | 0.579 | 0.557 |
| Immigration status | | | | |
| First generation | 0.633 | 0.574 | 0.567 | 0.537 |
| Second generation | 0.664 | 0.599 | 0.552 | 0.496 |
| Natives | 0.652 | 0.587 | 0.548 | 0.538 |
| Parental education | | | | |
| Low | 0.572 | 0.511 | 0.547 | 0.508 |
| Medium | 0.621 | 0.533 | 0.518 | 0.494 |
| High | 0.651 | 0.598 | 0.534 | 0.531 |
| Educational experience | | | | |
| Academic track | 0.657 | 0.603 | 0.534 | 0.537 |
| No academic track | 0.580 | 0.503 | 0.546 | 0.508 |
| Higher education | 0.637 | 0.595 | 0.532 | 0.519 |
| No higher education | 0.653 | 0.580 | 0.539 | 0.521 |

Note: Pairwise correlation coefficients. All coefficients are significantly determined on the 1 percent level.

**Table A6: Pairwise correlations between course grades and the SweSAT scores, upper secondary school graduates, spring 2012**

| | SWESATQ | | SWESATV | |
|---|---|---|---|---|
| Group | MATA | MATB | ENGA | ENGB |
| All | 0.614 | 0.540 | 0.556 | 0.525 |
| Gender | | | | |
| Men | 0.629 | 0.567 | 0.525 | 0.499 |
| Women | 0.619 | 0.613 | 0.586 | 0.553 |
| Immigration status | | | | |
| First generation | 0.587 | 0.535 | 0.542 | 0.501 |
| Second generation | 0.597 | 0.531 | 0.541 | 0.506 |
| Natives | 0.611 | 0.536 | 0.557 | 0.527 |
| Parental education | | | | |
| Low | 0.512 | 0.486 | 0.484 | 0.435 |
| Medium | 0.583 | 0.491 | 0.528 | 0.490 |
| High | 0.606 | 0.543 | 0.545 | 0.520 |
| Educational experience | | | | |
| Academic track | 0.610 | 0.557 | 0.544 | 0.531 |
| No academic track | 0.553 | 0.446 | 0.544 | 0.491 |
| Higher education | 0.591 | 0.544 | 0.528 | 0.507 |
| No higher education | 0.604 | 0.521 | 0.546 | 0.511 |

Note: Pairwise correlation coefficients. All coefficients are significantly determined on the 1 percent level.

## Table A7: Determinants of SweSAT scores and GPA, upper secondary school graduates

| | Test adm./Dependent variable | | | |
|---|---|---|---|---|
| | 2011B | | 2012A | |
| Independent variable | SWESAT | GPA | SWESAT | GPA |
| Constant | 6.676** | 20.64** | 5.076** | 20.32** |
| MALE | 1.302** | -1.006** | 1.304** | -1.084** |
| AGE | -0.019 | -0.325** | 0.059** | -0.300** |
| FIRSTGEN | -2.065** | -1.216** | -2.238** | -1.313** |
| SECONDGEN | -1.448** | -0.775** | -1.675** | -0.983** |
| FORBORN | -0.029 | 0.264 | 0.206 | 0.592* |
| MIDEDU | 0.413** | 0.466** | 0.232* | 0.269** |
| HIGHEDU | 1.903** | 1.291** | 1.499** | 1.019** |
| ACTRACK | 1.097** | 0.418** | 0.911** | 0.355** |
| KOMVUX | 0.394 | -1.670** | 0.527** | -1.227** |
| HIGHEXP | 2.672** | 1.396** | 2.367** | 1.412** |
| D11B | - | - | 1.350** | 0.427** |
| R-squared | 0.210 | 0.172 | 0.259 | 0.187 |
| N | 23,214 | 19,708 | 27,075 | 22,030 |

Note: ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

**Table A8: Determinants of SweSAT quantitative and verbal scores, upper secondary school graduates**

| | Dependent variable/Test adm. | | | |
|---|---|---|---|---|
| | SWESATQ | | SWESATV | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| Constant | 12.36** | 9.231** | 0.510 | 0.496 |
| MALE | 2.103** | 2.045** | 0.489** | 0.564** |
| AGE | -0.316** | -0.164** | 0.278** | 0.279** |
| FIRSTGEN | -1.232** | -1.496** | -2.911** | -2.960** |
| SECONDGEN | -1.079** | -1.431** | -1.802** | -1.932** |
| FORBORN | -0.175 | 0.219 | 0.221 | 0.213 |
| MIDEDU | 0.049 | -0.101 | 0.767** | 0.553** |
| HIGHEDU | 1.509** | 1.085** | 2.284** | 1.896** |
| ACTRACK | 1.369** | 1.052** | 0.810** | 0.766** |
| KOMVUX | 0.508 | 0.557* | 0.230 | 0.486* |
| HIGHEXP | 2.486** | 2.257** | 2.880** | 2.477** |
| D11B | - | 1.751** | - | 0.944** |
| R-squared | 0.190 | 0.237 | 0.192 | 0.214 |
| N | 23,214 | 27,075 | 23,214 | 27,075 |

Note: ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

## Table A9: Determinants of mathematics and English grades, upper secondary school graduates

| | Dependent variable/Test administration | | | |
|---|---|---|---|---|
| | MATA | | ENGA | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| Constant | 21.68** | 20.72** | 19.23** | 18.47** |
| MALE | 0.366** | 0.269** | 0.152** | 0.116* |
| AGE | -0.456** | -0.398** | -0.308** | -0.261** |
| FIRSTGEN | -1.569** | -1.604** | -1.236** | -1.337** |
| SECONDGEN | -1.147** | -1.310** | 0.045 | 0.091 |
| FORBORN | -0.900* | -0.550 | 1.799** | 1.603** |
| MIDEDU | 0.608** | 0.367* | 0.773** | 0.654** |
| HIGHEDU | 1.651** | 1.373** | 1.770** | 1.616** |
| ACTRACK | 1.401** | 1.288** | 0.906** | 0.809** |
| KOMVUX | -1.198* | -0.125 | -0.618 | -0.233 |
| HIGHEXP | 1.538** | 1.546** | 1.468** | 1.343** |
| D11B | - | 0.735** | - | 0.316** |
| R-squared | 0.138 | 0.141 | 0.096 | 0.091 |
| N | 19,704 | 22,026 | 19,698 | 22,015 |

Note: ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

## Table A10: Gender differences in ranking of GPA versus SweSAT, upper secondary school graduates

| | Test administration | |
|---|---|---|
| Independent variable | 2011B | 2012A |
| MALE | -0.540** | -0.504** |
| AGE*MALE | -0.006 | -0.010 |
| FIRSTGEN*MALE | 0.075 | 0.219** |
| SECONDGEN*MALE | 0.083 | 0.107** |
| FORBORN*MALE | -0.276 | 0.231 |
| MIDEDU*MALE | -0.107 | -0.123 |
| HIGHEDU*MALE | -0.100 | -0.100 |
| ACTRACK*MALE | 0.021 | 0.035 |
| KOMVUX*MALE | 0.179 | -0.375 |
| HIGHEXP*MALE | 0.090** | 0.004 |
| D11B*MALE | - | 0.068* |
| R-squared | 0.204 | 0.218 |
| N | 19,708 | 22,030 |

Note: The dependent variable is the difference between GPA and SweSAT, both measured on a standard normal scale. Coefficients in the table are difference-in-difference estimates. A positive coefficient means that males are ranked higher by the course grade than females, and a negative coefficient that the SweSAT part ranks males higher. ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

**Table A11: Gender differences in ranking of course grades versus SweSAT quantitative and verbal parts, upper secondary school graduates**

| | Dependent variable/Test administration | | | |
|---|---|---|---|---|
| | MATA/SWESATQ | | ENGA/SWESATV | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| MALE | -0.774** | -0.616** | 0.508** | 0.414** |
| AGE*MALE | 0.019** | 0.012 | -0.026** | -0.022** |
| FIRSTGEN*MALE | 0.124** | 0.252** | -0.145* | -0.131** |
| SECONDGEN*MALE | 0.139** | 0.187** | -0.213** | -0.104** |
| FORBORN*MALE | -0.216 | 0.107 | -0.142 | -0.158 |
| MIDEDU*MALE | -0.010 | -0.072 | -0.002 | -0.047 |
| HIGHEDU*MALE | 0.006 | -0.008 | -0.014 | -0.029 |
| ACTRACK*MALE | 0.015 | -0.025 | -0.026 | 0.017 |
| KOMVUX*MALE | -0.110 | -0.301 | 0.072 | 0.100 |
| HIGHEXP*MALE | -0.099** | -0.130** | 0.059 | -0.041 |
| D11B*MALE | - | 0.068** | - | -0.025 |
| R-squared | 0.079 | 0.097 | 0.123 | 0.113 |
| N | 19,704 | 22,026 | 19,698 | 22,015 |

Note: The dependent variables are constructed as the difference between the course grade and the SWESAT part, both measured on a standard normal scale. Coefficients in the table are difference-in-difference estimates. A positive coefficient means that males are ranked higher by the course grade than females, and a negative coefficient that the SweSAT part ranks males higher. ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

## Appendix B: Upper secondary school students

### Table B1: Descriptive statistics. Participants in the SweSAT autumn 2011, upper secondary school students

| Variable | Nobs | Mean | Std.Dev. | Min | Max |
|---|---|---|---|---|---|
| SWESAT | 12272 | 9.45 | 3.83 | 0 | 20 |
| SWSSATQ | 12272 | 10.24 | 4.51 | 0 | 20 |
| SWESATV | 12272 | 8.16 | 4.15 | 0 | 20 |
| GGPA | 12116 | 265.8 | 38.0 | 20 | 320 |
| NMATA | 1076 | 14.18 | 4.75 | 0 | 20 |
| NMATB | 1342 | 13.09 | 5.40 | 0 | 20 |
| NENGA | 1565 | 16.28 | 3.50 | 0 | 20 |
| NENGB | 1841 | 14.71 | 3.60 | 0 | 20 |
| NSVB | 4304 | 14.72 | 3.83 | 0 | 20 |
| MALE | 12272 | 0.502 | | 0 | 1 |
| BORN94 | 12272 | 0.155 | | 0 | 1 |
| Firstgen | 12272 | 0.045 | | 0 | 1 |
| Secondgen | 12272 | 0.122 | | 0 | 1 |
| FORBORN | 12272 | 0.008 | | 0 | 1 |
| MIDEDU | 12272 | 0.292 | | 0 | 1 |
| HIGHEDU | 12272 | 0.684 | | 0 | 1 |
| ACTRACK | 12272 | 0.872 | | 0 | 1 |
| KOMVUX | 12272 | 0.007 | | 0 | 1 |
| HIGHEXP | 12272 | 0.007 | | 0 | 1 |

**Table B2: Descriptive statistics. Participants in the SweSAT spring 2012, upper secondary school students**

| Variable | Nobs | Mean | Std.Dev. | Min | Max |
|---|---|---|---|---|---|
| SWESAT | 23113 | 9.16 | 3.750 | 0 | 20 |
| SWSSATQ | 23113 | 8.71 | 4.25 | 0 | 20 |
| SWESATV | 23113 | 9.12 | 4.24 | 0 | 20 |
| GGPA | 22844 | 260.2 | 39.2 | 30 | 320 |
| NMATA | 2120 | 13.63 | 5.07 | 0 | 20 |
| NMATB | 2477 | 12.73 | 5.58 | 0 | 20 |
| NENGA | 3056 | 15.82 | 3.60 | 0 | 20 |
| NENGB | 2745 | 14.39 | 3.83 | 0 | 20 |
| NSVB | 6821 | 14.33 | 3.87 | 0 | 20 |
| MALE | 23113 | 0.484 | | 0 | 1 |
| BORN94 | 23113 | 0.250 | | 0 | 1 |
| Firstgen | 23113 | 0.049 | | 0 | 1 |
| Secondgen | 23113 | 0.122 | | 0 | 1 |
| FORBORN | 23113 | 0.007 | | 0 | 1 |
| MIDEDU | 23113 | 0.296 | | 0 | 1 |
| HIGHEDU | 23113 | 0.679 | | 0 | 1 |
| ACTRACK | 23113 | 0.844 | | 0 | 1 |
| HIGHEXP | 23113 | 0.007 | | 0 | 1 |
| PART11B | 23113 | 0.291 | | 0 | 1 |

**Table B3: Pairwise correlations between national test grades and the SweSAT scores upper secondary school students, autumn 2011**

| | SWESATQ | | SWESATV | | |
|---|---|---|---|---|---|
| Group | NMATA | NMATB | NENGA | NENGB | NSVB |
| All | 0.684 | 0.704 | 0.582 | 0.609 | 0.338 |
| Gender | | | | | |
| Men | 0.677 | 0.687 | 0.581 | 0.600 | 0.344 |
| Women | 0.702 | 0.712 | 0.571 | 0.608 | 0.386 |
| Immigration status | | | | | |
| First generation | 0.667 | 0.663 | 0.477 | 0.615 | 0.372 |
| Second generation | 0.638 | 0.686 | 0.545 | 0.590 | 0.315 |
| Natives | 0.687 | 0.706 | 0.595 | 0.605 | 0.322 |
| Parental education | | | | | |
| Low | 0.631 | 0.701 | 0.471 | 0.526 | 0.298 |
| Medium | 0.647 | 0.665 | 0.557 | 0.592 | 0.305 |
| High | 0.695 | 0.689 | 0.573 | 0.596 | 0.331 |
| Educational experience | | | | | |
| Academic track | 0.690 | 0.672 | 0.581 | 0.608 | 0.328 |
| No academic track | 0.644 | 0.712 | 0.560 | 0.597 | 0.383 |

Note: Pairwise correlation coefficients. All coefficients are significantly determined on the 1 percent level.

**Table B4: Pairwise correlations between national test grades and the SweSAT scores upper secondary school students, spring 2012**

| | SWESATQ | | SWESATV | | |
|---|---|---|---|---|---|
| Group | NMATA | NMATB | NENGA | NENGB | NSVB |
| All | 0.636 | 0.656 | 0.628 | 0.592 | 0.353 |
| Gender | | | | | |
| Men | 0.636 | 0.659 | 0.605 | 0.585 | 0.361 |
| Women | 0.635 | 0.650 | 0.639 | 0.592 | 0.419 |
| Immigration status | | | | | |
| First generation | 0.607 | 0.646 | 0.615 | 0.568 | 0.407 |
| Second generation | 0.603 | 0.629 | 0.594 | 0.553 | 0.326 |
| Natives | 0.631 | 0.654 | 0.625 | 0.593 | 0.336 |
| Parental education | | | | | |
| Low | 0.487 | 0.363 | 0.651 | 0.573 | 0.416 |
| Medium | 0.574 | 0.636 | 0.608 | 0.598 | 0.337 |
| High | 0.654 | 0.645 | 0.610 | 0.560 | 0.336 |
| Educational experience | | | | | |
| Academic track | 0.649 | 0.647 | 0.626 | 0.590 | 0.344 |
| No academic track | 0.513 | 0.609 | 0.610 | 0.592 | 0.340 |

Note: Pairwise correlation coefficients. All coefficients are significantly determined on the 1 percent level.

**Table B5: Determinants of SweSAT quantitative and verbal scores; upper secondary school students**

| | Dependent variable/Test adm. | | | |
|---|---|---|---|---|
| | SWESATQ | | SWESATV | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| Constant | 6.263** | 5.586** | 5.979** | 5.840** |
| MALE | 2.329** | 2.246** | 0.690** | 0.666** |
| BORN94 | 0.846** | 1.118** | 0.170 | 0.317** |
| FIRSTGEN | -1.808** | -1.710** | -2.77** | -2.303** |
| SECONDGEN | -1.776** | -1.972** | -2.481** | -2.212** |
| FORBORN | 0.110 | 0.242 | -0.762 | -0.156 |
| MIDEDU | 0.038 | 0.122 | 0.169 | 0.454** |
| HIGHEDU | 1.765** | 1.505** | 1.970** | 1.907** |
| ACTRACK | 2.100** | 1.680** | 0.994** | 0.925** |
| HIGHEXP | 0.920* | 1.504** | 0.798 | 1.110 |
| D11B | - | 2.093** | - | 1.320** |
| R-squared | 0.177 | 0.228 | 0.125 | 0.136 |
| N | 12,272 | 23,113 | 12,272 | 23,113 |

Note: ** Coefficient significantly determined on the 99% level, * significant on the 95% level.

**Table B6: Determinants of mathematics and English national test grades; upper secondary school students**

| | Dependent variable/Test administration | | | |
|---|---|---|---|---|
| | NMATA | | NENGA | |
| Independent variable | 2011B | 2012A | 2011B | 2012A |
| Constant | 12.82** | 9.965** | 13.21** | 12.84** |
| MALE | 0.671** | 1.192** | 0.807** | 0.862** |
| BORN94 | 1.836** | 2.147** | 1.019** | 1.087** |
| FIRSTGEN | -2.237** | -2.380** | -1.695** | -1.382** |
| SECONDGEN | -1.940** | -2.044** | -0.454 | -0.924 |
| FORBORN | -1.217 | -0.131 | 0.891 | 0.855 |
| MIDEDU | -1.117 | 0.589 | 1.378* | 0.718 |
| HIGHEDU | 0.106 | 2.082** | 2.526** | 1.979** |
| ACTRACK | 1.791** | 1.457** | 0.552* | 0.747** |
| HIGHEXP | -5.019* | -2.549 | -0.662 | 0.726 |
| D11B | - | 1.517** | - | 0.788** |
| R-squared | 0.093 | 0.146 | 0.084 | 0.107 |
| N | 1,076 | 22,026 | 1,565 | 22,015 |

Note: ** Coefficient significantly determined on the 99% level, * significant on the 95% level.