Befintliga och framtida provmodeller i vuxenproven

Marie Wiberg

Umeå Universitet

Innehållsförteckning

Inledning 1
Syfte 1
Utgångspunkter 2
Rättvisa, validitet och reliabilitet 2
Uppgiftsanalys 3
Bakgrundsinformation 3
Provanalys 4
Jämförelse av provversioner över tid 4
Metoder för att jämföra provversioner 5
Utformning av uppgifter 6
Utformning av prov 6
Uppgiftsbank 7
Datorbaserade adaptiva prov 7
Multistage testing 8
Principer för att konstruera likvärdiga prov 10
Programvaror för analys av prov och uppgifter 11

Nationella prov i svenska för invandrare (kurs B, C och D) 12

Nuvarande uppgiftsformat 12
Insamlad bakgrundsinformation 13
Nuvarande provanalys och uppgiftsanalys 13
Utprövning av nya uppgifter 14
Nuvarande programvara 15
Validitetshot med dagens prov 15
Förslag på förändringar av provmodell 16
Förslag på förändring A-N 16

Nationella prov i matematik (2b) som används till vuxna 21

Nuvarande uppgiftsformat 21
Insamlad bakgrundsinformation 21
Nuvarande uppgiftsanalys 22
Nuvarande provanalys 22
Utprövning av nya uppgifter 23
Nuvarande programvara 23
Validitetshot med dagens prov 23
Förslag på förändringar av provmodell 24
Förslag på förändring A-K 24

Nationella prov i engelska (6) som används till vuxna 27

Nuvarande uppgiftsformat 28
Insamlad bakgrundsinformation 28
Nuvarande uppgiftsanalys 28
Nuvarande provanalys 29
Utprövning av nya uppgifter 29
Nuvarande programvara 30
Validitetshot med dagens prov 30
Förslag på förändringar av provmodell 31
Förslag på förändring A-M 31

Generella förändringsförslag i provmodeller för vuxenprov 35

Litteraturförteckning 36

Inledning

Skolverket tar på uppdrag av regeringen fram nationella prov inom grundskola, gymnasieskola och kommunal vuxenutbildning. Dessa prov ges i dagsläget i pappersform. Enligt regeringsuppdraget U2017/03739/GV ska Skolverket utveckla och tillhandahålla digitala nationella prov och bedömningsstöd i grundskolan och på gymnasial nivå utifrån vad som anges i propositionen *Nationella prov - rättvisa, likvärdiga, digitala (prop.2017/18:14)*. Skolverket planerar därför att erbjuda digitala nationella prov i framtiden. I den kommunala vuxenutbildningen (komvux) används obligatoriska nationella prov i bland annat utbildning i svenska för invandrare (SFI) kurserna B, C, och D, utbildning på gymnasial nivå i svenska och svenska som andraspråk kurs 1 och 3, matematik kurserna 1-4 på gymnasial nivå samt engelska kurserna 5 och 6 på gymnasial nivå. Fokus i denna rapport kommer att ligga på några av dessa kurser.

Syfte

Projektet syftar till att:

a) formulera en eller flera modeller för provsystem för vuxenutbildning som bättre hanterar den frekventa användningen och exponeringen av nationella prov, med utgångspunkt i analyser av följande:
   - Skolverkets systemramverk för nationella prov,
   - nationella prov i svenska för invandrare, kurs B och D, och deras konstruktionsprinciper,
   - nationella prov i matematik 2b prov och dess konstruktionsprinciper,
   - nationella prov i svenska och svenska som andraspråk 1 och 3 och deras konstruktionsprinciper, och
   - nationella prov i engelska 6 prov och dess konstruktionsprinciper,

   I första hand studeras nationella prov i svenska för invandrare och matematik i andra hand proven i svenska/svenska som andraspråk och engelska.

b) presentera och diskutera utkast till provmodell (eller provmodeller) med provkonstruktörer på en workshop som Skolverket anordnar i oktober eller november 2018, (November)

c) vid behov delta på ytterligare en workshop/samråd med provkonstruktörer, som Skolverket tar initiativ till, i januari 2019.

d) samt leverera en rapport som beskriver provmodellen (eller modellerna), inklusive de principer som ingår för att konstruera likvärdiga prov.

Uppdraget ska organiseras och utföras i samarbete med Skolverkets ansvariga undervisningsråd. Utöver den kontinuerliga kontakten mellan Skolverkets och Lärosätets kontaktpersoner, genomförs en till två avstämningstillfällen som Skolverket tar initiativ till.

Utgångspunkter

Den här rapporten är upplagd så att gemensamma delar för uppgifter, prov och analys av dessa beskrivs i den första delen. Sedan diskuteras befintlig(a) provmodell(er) samt vilka problem som finns med den inklusive hot mot validitet som finns för dem samt principer för att konstruera likvärdiga prov. Utifrån eventuella problem eller utmaningar med validitet och eller reliabilitet, så ges förslag på alternativ(a) provmodell(er) i form av ett antal förslag på förändringar för respektive prov.

Rättvisa, validitet och reliabilitet

Standardiserade prov måste vara utformade så att de bedömer kunskap rättvist bland de elever som genomför dem. Med rättvisa avses att alla ska ges likvärdiga förutsättningar att genomföra proven. Prov måste vidare utformas så att de har hög reliabilitet och hög validitet. Med validitet avses att provet mäter i sammanhanget det ges vad det avser att mäta och med reliabilitet avses att provet mäter på ett tillförlitligt sätt, dvs. att mätningen kännetecknas av noggrannhet och pålitlighet. För ramverk inom validitet hänvisas historiskt till Messick (1989) och Kane (2006), samt Crooks, Kane and Cohen (2006) för användande av validitetsargument. Validitetsargumenten som har betydelse för provens validitet byggs upp utifrån följande åtta steg; (1) administration och genomförande, (2) bedömning, (3) aggregering (4) generalisering av provresultat till provdomän, (5) extrapolering av provresultat till måldomän, (6) värdering av innebörd i elevers skattade resultat i måldomän, (7) beslut, samt (8) provets konsekvenser. Det finns med andra ord många olika faktorer som kan påverka ett prov. Exempel på faktorer som påverkar provens validitet är om man har tydliga riktlinjer för att skapa proven och för att bedöma elevernas lösningar eller svar. Vidare att proven inte blir kända i förväg till grupper av elever.

Faktorer som påverkar provens reliabilitet är fysisk miljö som provet ges i, instruktioner och information innan och i samband med proven men även elevernas fysiska och psykiska form när de tar provet. Nervositet och stress kan påverka negativt elevens provresultat, samtidigt som det kan finns intressant information om man studerar elevernas svarstider på uppgiftsnivå och provnivå. Det är viktigt att man eftersträvar en hög reliabilitet genom att se till att likvärdiga bedömningar ges oavsett geografisk placering i landet, bakgrund på eleven samt vilken lärare som utfört bedömningen.

Uppgiftsanalys

Oavsett provmodell är det viktigt att undersöka uppgifternas egenskaper för att uppnå hög reliabilitet i proven. Detta kan göras både utifrån den klassiska testteorin (CTT; Crocker & Algina, 1986) och utifrån den moderna testteorin den s.k. item response theory (IRT; Hambleton & Swaminathan, 1985). En stor del av de nedanstående analyserna görs i proven idag – men det är viktigt att lyfta fram vikten att dessa analyser görs oavsett om det är ett prov i svenska, engelska eller matematik. Dessa analyser bör göras både vid utprövning, vid användandet av eventuella ankaruppgifter, skarpa uppgifter samt vid regelbundna uppföljningar av proven:

- Andel provtagare som besvarar en uppgift korrekt, dvs. lösningsproportionen eller s.k. p-värdet
- Uppgiftens svårighetsgrad
- Uppgiftens diskrimineringsförmåga
- Hur enkelt det är att gissa det korrekta svaret
- Uppgifternas medelfel och standardavvikelse
- Partiell bortfallsanalys
- Förhållande mellan provtagarnas förmåga och uppgiftens svårighetsgrad.
- Uppgiftens stabilitet över tid (utifrån utprövningar och skarpt användande).
- Om uppgifter fungerar på olika sätt om de ingår i olika provversioner.
- Resultatet på en uppgift utifrån elevers bakgrund eller enskilda grupper.
- Differential Item Functioning (DIF) analys (Dorans & Holland, 1993;), dvs. man undersöker hur en uppgift fungerar i olika elevgrupper.

I de fall man har uppgifter som kräver lärares bedömning (ex. inom språkprov) bör man även analysera:

- Interbedömarreliabilitet
- Gruppens bedömningsmönster

Bakgrundsinformation

För att ta reda på hur en uppgift, eller en provversion fungerar och för att säkerställa så att uppgiften eller provversionen inte gynnar eller missgynnar en enskild grupp så bör man relatera uppgiften eller provet mot extern information om provtagarna. Information om eleverna som kan vara användbart är:

- Kön
- Ålder
- Elevernas resultat på tidigare nationella prov eller utprövningar
- Utbildningsbakgrund
- Betyg i ämnet
- Moder respektive fader född i Sverige eller utomlands
- Modersmål
- Tid i Sverige

Provanalys

Utöver en uppgiftsanalys bör man även undersöka hur hela provet respektive varje delprov fungerar. Lämpliga mått kan återigen hittas både i den klassiska och den moderna testteorin och inkluderar:

- Cronbach's (1951) alfa för att mäta den interna konsistensen både inom ett delprov och på hela provet i förekommande fall.
- Delprovens stabilitet över tid, utifrån ankaruppgifter, samt om det finns information både från utprövning och skarpa prov.
- Provets stabilitet över tid, utifrån ankaruppgifter, samt om det finns information både från utprövning och skarpa prov.
- Bortfall (utifrån enskild provtagargrupp)
- Bortfall (om det sker exempelvis i slutet av ett delprov eller ett prov)
- Kontrollera innehåll och format utifrån testspecifikationerna om provet
- Kontrollera provet i relation till de styrdokument som finns (Systemramverket).
- Lärarnas uppfattning om provet i relation till kursplaner och kursinnehåll.
- Deskriptiv statistik över provpopulationen så att man vet om den förändras över tid.

Jämförelse av provversioner över tid

Man bör undersöka stabilitet över tid på uppgiftsnivå, delprovsnivå samt på hela provet. Vad gäller jämförelse mellan delprovsversioner och hela provet över tid så bör ekvivaleringsmetoder användas. Beroende på vilken information och vilken data man samlar in och därmed har tillgång till så kan olika ekvivaleringsmetoder användas. Inom ekvivaleringsteorin brukar man prata om datainsamlings designer. Se González & Wiberg (2017) för en översikt av alla olika designer. Om samma provgrupp får göra två prov som ska jämföras kallas det *Single group (SG) design*, om man dessutom vill ta hänsyn till att det kan finnas en ordningseffekt så kan man låta halva gruppen göra det ena provet först och det andra provet sedan, och den andra halvan av gruppen får göra proven i omvänd ordning, då har vi en *Counterbalanced (CB) design*. I dagsläget så används SG designen exempelvis när utprövningsdata jämförs med data från skarpa prov för samma elever. Om man anser att två grupper är likvärdiga eller relativt lika så kan man använda sig av *Equivalent Groups (EG) design*. Eftersom detta antagande inte alltid är uppfyllt även i standardiserade prov (se ex. Lyrén & Hambleton, 2011) så kan denna design vara problematisk att använda. Den design som framförallt förespråkas inom prov och forskning om prov är att använda sig av *Non-equivalent groups with anchor test (NEAT) design*, vilken innebär att man ger ett antal uppgifter som är samma i provversionerna, dvs. ankaruppgifter. Detta görs inte i dagsläget i alla de nationella proven, men det görs exempelvis i det standardiserade högskoleprovet i Sverige. Om man inte har möjlighet att använda sig av en NEAT design så finns en möjlighet att använda sig av en *Nonequivalent groups with covariates (NEC) design* (Wiberg & Bränberg, 2015). I en NEC design så använder man sig av bakgrundsinformation om provtagarna, ex. utbildningsbakgrund, betyg eller liknande.

Att använda en NEC design har visat sig ge mindre fel än en EG design men notera att en NEAT design alltid ger minst fel.

För att man ska kunna använda ankaruppgifter så krävs det att provtagarna inte är medvetna om att dessa uppgifter inte tillhör det skarpa provet. Utifrån hur de nationella proven är byggda idag så finns inte alltid utrymme för ankaruppgifter. Vid en förändring framförallt en digitalisering ges dock stora möjligheter att skapa en provmodell så att ankaruppgifter kan läggas in. Man kan då antingen lägga in dem blandat med de skarpa uppgifterna, alternativt lösa det på samma sätt som i det svenska högskoleprovet där provtagarna ges fyra skarpa delar och en utprövningsdel som antingen innehåller ankaruppgifter eller utprövningsuppgifter. Provtagarna är i dessa fall inte medvetna om vilken del av provet som är utprövningsdel och vilka delar av provet som är skarpa delar. Fördelen med detta är också att man skulle få bättre utprövningsdata vilket i förlängningen förenklar när man ska sätta ihop ett nytt prov. Nackdelen är att proven blir längre när man inkluderar en extra del.

Metoder för att jämföra provversioner

Det finns ett stort antal ekvivaleringsmetoder för att jämföra provversioner, och för en översikt se Kolen och Brennan (2014) samt hur de kan utföras i praktiken se González och Wiberg (2017). Generellt kan man dela upp dem i tre typer av metoder; traditionella metoder, kernelmetoder samt IRT metoder. Traditionella metoder är medelvärdesekvivalering, linjärekvivalering samt ekvipercentilekvivalering. I medelvärdesekvivalering så utgår man ifrån att enbart medelvärdet skiljer sig mellan provversionerna. I linjärekvivalering tar man hänsyn både till medelvärdet och till standardavvikelsen. Slutligen i ekvipercentilekvivalering så anses provpoäng på två provversioner vara lika om samma andel provtagare har besvarat dem korrekt. Eftersom provpoäng ofta är diskreta så måste man förändra poängen för att skapa en kontinuerlig fördelning, vilket görs inom de traditionella metoderna med hjälp av linjär interpolation. Kernelmetoder, använder istället en fördelning (ex. Gaussian, logistisk eller uniform) vid förändringen av diskreta provpoäng till kontinuerliga provpoäng. Oftast fungerar ekvipercentilmetoder tillfredsställande oavsett om traditionella eller kernelmetoder avvänds om grupperna är relativt lika (Liu & Low, 2008). Slutligen så bygger IRT metoder på olika IRT modeller och de används både på observerade eller (skattade) sanna provpoäng.

Utformning av uppgifter

Beroende på typ av kunskap eller färdighet som ska prövas så kan olika uppgiftsformat används. Uppgiftsformatet beror även på vilket medium som används för att förmedla provet. Vanliga uppgiftstyper vid papper och penna prov är:

- Flervalsfrågor: Flera svarsalternativ finns – ett svar är korrekt.
- Lucktexter: Man fyller i ett saknat ord med eller utan alternativ att välja mellan.
- Matchning: Para ihop två egenskaper eller begrepp.
- Korta svar: Korta skrivna texter eller lösningar.
- Uppsatser: Längre skrivna texter.

Uppgiftstyper som kan användas vid digitaliserade prov utöver de som nämns ovan för papper och penna prov är exempelvis:

- Flervalsuppgifter där svarsalternativens ordning slumpas eller att ordningen på uppgifterna slumpas
- Hot-Spot uppgift: Flervalsuppgift där provtagaren besvarar uppgiften genom att markera svaret på en text eller bild. Denna typ av flervalsuppgift ger möjlighet till stora valmöjligheter och fler svarsalternativ än vanliga flervalsuppgifter.
- Uppgifter ställs via en filmsekvens, ljudsekvens eller bildsekvens.
- Uppgifter som skall lösas i flera steg och där du enbart kommer vidare om du klarat steget innan.
- Uppgifter som kräver en hög grad av interaktion, exempelvis att man ska flytta eller organisera en miljö eller agera i en miljö.
- Drag and Drop uppgift: Frågestam i form av en instruktion till provtagaren och uppgiften är att organisera eller flytta objekt utifrån instruktionen.

Utformning av prov

Det finns en mängd olika provformat som kan antingen användas inom ett delprov eller appliceras på hela provet. I ett *prov med fixt antal uppgifter* så syftas i denna text på ”vanliga” prov som ofta ges på papper med ett givet antal uppgifter som alla provtagare får. En variant av prov med fixt antal uppgifter är sekventiella prov (Wald, 1966). I dessa prov så ges uppgifterna i en sekvens och man avbryter provet när man bedömer att provtagaren antingen kan tillräckligt mycket eller tillräckligt lite om ett område. Dessa prov är därför lämpliga vid exempelvis certifieringsprov. Notera dock att uppgifterna som ges inte är anpassad utifrån provtagarens förmåga utan enbart ges i en given förutbestämd följd. För en jämförelse av uppgifters egenskaper mellan prov med fixt antal uppgifter och sekventiella prov se Wiberg (2003). Andra provformat inkluderar datorbaserade adaptiva prov samt prov som ges i ett antal stadier. Om man vill kunna skapa en mängd olika provversioner underlättar det dock om man först skapar en uppgiftsbank.

Uppgiftsbank

En uppgiftsbank är en samling uppgifter med välkända karakteristika. Det kan även vara en samling av uppgifter som hör till en viss text eller ljudfil. Varje uppgift som prövas ut lagras tillsammans med all information om uppgiften. Information om uppgiften som är bra att lagra inkluderar; vem som skrivit uppgiften, när den skapades, hur den har använts (utprövningar, skarpa prov), hur uppgiften fungerat i utprövningen tillsammans med information om typ av uppgift, innehåll, samt vilket kursplaneinnehåll den förväntas mäta. Det är att föredra att spara så mycket information som möjligt eftersom det tillåter flexibel användning. Man bör även se till att man skapar utrymme i sin uppgiftsbank för att lägga in information om hur uppgiften fungerar i skarpt prov. Man bör lägga in olika statistiska uppgiftsanalysmått från utprövningar och när uppgiften används i skarpa prov. Här är att föredra att man använder mått både från den klassiska och den moderna testteorin. Exempel på statistiska mått är uppgiftens diskrimineringsförmåga och lösningsfrekvens samt hur elever med olika förmåga klarar att lösa uppgiften. För fler uppgiftsmått hänvisas till avsnittet om uppgiftsanalys. Vilka grupper av elever som ska undersökas beror på provens karaktär. Man bör även lägga in information om vilken typ av uppgift man har (dvs. uppgiftsformat), vilken del av kursplanen den tillhör och därmed vilka kunskapskrav som finns, hur uppgiften ska besvaras, eventuell koppling till andra uppgifter, om den innehåller bilder eller figurer, samt om eventuella verktyg får användas för att lösa uppgiften. Slutligen bör man addera information när uppgiften används i skarpa prov och hur den besvaras för att följa uppgiften. Om svarsfrekvensen ändras drastiskt kan man exempelvis misstänka att uppgiften blivit känd för provtagarna i förväg. För mer information om uppgiftsbanker generellt se exempelvis Umar (1997) eller Vale (2004), eller mer specifikt kopplat till det svenska körprovet (Wiberg. 2002).

Datorbaserade adaptiva prov

Ett papper och penna prov kan vara adaptivt, men det kräver en hel del administration och det är därför vanligare att man har ett datorbaserat adaptivt prov, s.k. *Computerized Adaptive Test* (CAT; van der Linden & Glas, 2000). Idén med ett adaptivt prov är att kontinuerligt skatta vilken nivå provtagaren har på sina kunskaper och anpassa vilka uppgifter denna får efter kunskapsnivån. När man har identifierat provtagarens kunskapsnivå så avbryts provet. I ett CAT så har man programmerat datorn för att välja en relevant fråga från en uppgiftsbank utifrån ett givet mönster (Sands, Waters & McBride, 1997). Ofta är den inledande uppgiften av medel svårighetsgrad vilket kan vara bra utifrån ett provtagarperspektiv. Utifrån ett uppgiftsexponeringsperspektiv är det dock bra att välja mellan ett flertal olika uppgifter så att inte samma uppgifter används hela tiden. När provtagaren besvarar en uppgift korrekt får denne en svårare uppgift, medan om denne besvarar en uppgift inkorrekt så ges en enklare uppgift. Med andra ord så görs ett beslut om provtagarens förmåga efter varje uppgift och provet anpassas efter provtagarens förmåga.

Ett CAT ställer höga krav på att provkonstruktören vet uppgifternas egenskaper i förväg och att provkonstruktören skapat en algoritm för att ge lämpliga uppgifter utifrån provtagarens svarsmönster. Det finns även en relativt hög risk att vissa uppgifter exponeras mer än andra om man använder sig av CAT. Tanken är dock att

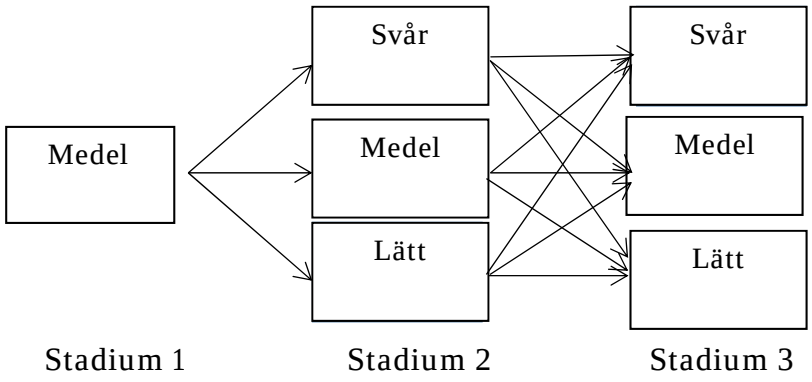
ett CAT ska vara mer relevant för varje enskild elev samtidigt som man sparar tid och pengar när alla provtagare inte får alla uppgifter (Drasgow & Olsen-Buchanan, 1999).

Även om CAT inte är en vanlig form av datorbaserade prov så finns ett antal kända exempel på CAT som getts till elever eller studenter. Dessa prov inkluderar av Educational Testing Service utvecklade GMAT (the Graduate Management Admission Test) samt GRE (the Graduate Record Examination) som mäter kunskaper i matematik och engelska i USA. USA:s försvarsdepartement har även använt sig av CAT i ASVAB (the Armed Services Vocational Aptitude Test Battery) som bland annat mäter allmänna kunskaper i naturkunskap, ordförståelse, matematisk/logisk förmåga, textförståelse, numeriska operationer, kodningshastighet samt mekanisk förståelse.

I Europa så har CITO i Nederländerna utvecklat adaptiva prov i språk och matematik för grundskolan, gymnasium och vuxenutbildning. Middlesex University har utvecklat CATE (the Computerized Adaptive Test of English) som mäter kunskaper i engelska som andraspråk. Vidare har University of Cambridge, Alliance Francais, Goethe-institutet och Universidad de Salamanco utvecklat BULATS (Business Language Testing Service) som mäter språkförmåga i flera olika språk och det består av flervalsuppgifter, höruppgifter samt läsuppgifter. För mer information om CAT se exempelvis Bunderson Inouye, och Olsen (1989), Wainer (2000), van der Linden och Glas (2000) och Weiss (2014).

Multistage testing

Multistage testing (MST; Yan, von Davier, & Lewis, 2014) är prov där moduler av uppgifter ges i två eller flera stadier till provtagarna. En modul består av ett mindre antal uppgifter. Det vanligaste är att ett MST ges i två eller tre stadier. Se figur 1 för en schematisk bild över ett MST där den generella svårighetsgraden på modulen är inskriven för varje modul. Utifrån provtagarens svar på uppgifterna i den första modulen blir de tilldelad någon av de tre modulerna i stadium 2. Utifrån provtagarens svar på uppgifterna i modulen i stadium 2 blir provtagaren sedan tilldelad en lämplig modul i stadium 3.



Figur 1. *Schematisk bild över ett trestegs MST med en modul i stadium 1, tre moduler i stadium 2, och tre moduler i stadium 3.*

Fördelar med MST jämfört med prov med fixt antal uppgifter där alla provtagare får samma uppgifter är att MST är mycket flexiblare i sin utformning. Genom att inte alla elever får samma uppgifter så får varje uppgift mindre uppgiftsexponering än i prov med fixt antal uppgifter där alla elever får samma uppgifter. MST ger precis information om provtagare på flera kunskapsnivåer samtidigt som proven kräver en mindre population för att genomföras än ett CAT. Nackdelen är att om man skapar många moduler kan det leda till vissa komplikationer när man ska bestämma vilka moduler som ska ges utifrån provtagarens resultat på tidigare moduler.

Exempel på MST som används runt om i världen är bland annat GRE (Graduate Record Examination), dvs. antagningsprov för doktorander i USA. MST är dessutom på förslag att användas vid NAEP (National Assessment of Educational Progress) i USA för att få högre precision i mätningen i urbana eller utsatta områden (där det är troligt att kunskapsnivån är lägre). Vidare föreslås det att MST ska användas vid internationella prövningar för att uppnå precision på alla kunskapsnivåer. Dessutom är MST föreslaget att användas för RTTA (Race to the Top Assessment). RTTA är ett prov med syfte att göra en bedömning av vilken kunskap elever har och mäter elevernas kunskap gentemot standarder som är designade för att säkerställa att alla elever har de kunskaper och färdigheter som krävs för att lyckas på ett arbete och eller på college i USA.

Principer för att konstruera likvärdiga prov

Oavsett hur den exakta slutgiltiga provmodellen är utformad så finns det ett antal principer som man kan använda för att konstruera likvärdiga prov samt för att säkerställa att de olika versionerna av provet fungerar som planerat och därmed är likvärdiga över tid.

Man bör eftersträva att man har samma proportion av uppgiftstyper i alla versioner av provet. Man bör skapa anvisningar över vilka typer av uppgifter som ska användas och vilken proportion av dessa som ska ingå i varje prov. Vidare hur svårighetsgraden ska vara på provet. För att förenkla uppgiftskonstruktionen kan man använda sig av skelettfrågestammar där samma typ av uppgift ges men där man byter ut exempelvis kontext eller siffror. Det är även viktigt att uppgifter konstrueras som tar ungefär samma tid att lösa om de ska vara utbytbara.

Eftersom det kan vara utmanande att göra många likvärdiga versioner av en uppgift så kan det underlätta om man istället skapar moduler av uppgifter i vilka 5-10 uppgifter ingår. En modul kan även bestå av en text med ett antal uppgifter kopplat till den. Idén är att sätta samman moduler för olika stadier som blir likvärdiga utifrån svårighet, sammansättning och utformning. En viktig aspekt när man tar fram likvärdiga moduler är att även tiden det tar att lösa alla uppgifter i modulen bör vara liknande mellan olika versioner av modulen.

När man har uppgifter som kräver en lärares bedömning så är det viktigt att man gör tydliga bedömningsanvisningar för alla olika poäng som uppgiften kan generera. För att ytterligare öka likvärdigheten mellan olika versioner av provet så bör man erbjuda lärare som rättar uppgifterna regelbunden bedömningsträning.

Även om man sätter kravgränser innan provet ges bör man genomföra en preekvivalering av nya provversioner. Denna kan sedan användas som jämförelse när provet har getts i skarpt läge. Man bör även genomföra postekvivaleringar med viss regelbundenhet om provet ges upprepade gånger för att säkerställa att provet fungerar som man planerat. Utöver detta bör man regelbundet undersöka olika provet med olika kvalitetskontrollmått under den tid som provet är giltigt för att säkerställa att elevernas svarsmönster inte plötsligt förändras eller att svårighetsgraden i en uppgift, en modul, delprov eller ett helt prov förändras. Detta är oerhört viktigt eftersom det kan indikera att provet eller delar av provet har blivit spritt och att eleverna därmed haft tillgång till det innan de skulle skriva provet.

Programvaror för analys av prov och uppgifter

Det finns ett antal programvaror som är lämpliga för att analysera prov och uppgifter. En del av dessa är framtagna för speciella prov eller mätningar. I detta avsnitt så summerar vi några av de mer spridda och användbara programvarorna för analys av prov och uppgifter.

- R (https://www.r-project.org/)

Ett fritt tillgängligt program som kan laddas ner från internet. Först laddar man ner huvudprogrammet, sedan laddar man ner olika programpaket beroende på vilka analyser som man vill göra. Nedan ges exempel på ett antal programpaket som är lämpliga att använda för uppgiftsanalys, provanalys, för att jämföra provversioner etc.

  - ShinyItemAnalysis (för att genomföra uppgiftsanalys)
  - CTT (för att genomföra analyser med klassisk testteori)
  - ltm och mirt (för att genomföra uppgiftsanalys med modern testteori)
  - equate (för att genomföra traditionell ekvivalering av provversioner)
  - kequate (för att genomföra kernelekvivalering av provversioner)
  - equateIRT (för att genomföra IRT ekvivalering av provversioner)

- SPSS

Statistikprogram som kan användas för att genomföra uppgiftsanalys och provanalys utifrån klassisk testteori. Man kan även undersöka DIF, dimensionalitet i prov, lösningsproportioner och liknande.

- Mintab

Statistikprogram som kan användas för att genomföra uppgiftsanalys och provanalys utifrån klassisk testteori. Man kan även undersöka DIF, dimensionalitet i prov, lösningsproportioner och liknande.

- Excel

Ett kalkylprogram som tillåter att man gör enklare uppgiftsanalyser såsom exempelvis lösningsfrekvens och lösningsproportion.

- winsteps (https://www.winsteps.com)

Ett program som möjliggör att man gör analyser utifrån en parameter logistiska IRT modellen, Raschmodellen, rating scale modellen, partial credit modellen och liknande.

- Bilog-MG 3 (http://www.ssicentral.com/irt/)

Ett program för att analysera dikotoma uppgifter och prov med modern testteori såsom exempelvis de logistiska IRT modellerna.

Nationella prov i svenska för invandrare (kurs B, C och D)

De nationella proven i svenska för invandrare (SFI) genomförs i dagsläget alltid i nära anslutning till att eleven slutfört aktuell kurs. Det finns dessutom ett stort antal aktörer både i offentlig och privat regi där det i nuläget inte sker samordning när proven erbjuds. Detta innebär att proven ges ett mycket stort antal gånger under ett år och proven har därmed en mycket hög exponeringsgrad. I dagsläget finns det dessutom relativt få likvärdiga provversioner som är aktuella samtidigt.

Utgångspunkterna för de befintliga proven är rättvisa, reliabilitet och validitet vid provens konstruktion, genomförande och bedömning. Varje kursprov är idag uppdelat i fyra delprov som prövar färdigheterna;

- *Delprov LÄSA – läsförståelse*
- *Delprov HÖRA – hörförståelse*
- *Delprov SKRIVA – skriftlig framställning*
- *Delprov TALA – muntlig produktion och interaktion*

Oftast delas delproven upp i två häften eftersom det kan vara tröttsamt och kräva en hög koncentration att kommunicera på ett andraspråk.

Nuvarande uppgiftsformat

De uppgiftsformat som används i dagsläget är;

- Flervalsfrågor
- Kortsvar
- Längre egenproducerade svar
- Matchning
- Muntliga uppgifter (används i TALA)

I HÖRA används enbart flervalsfrågor i B- och C-kurserna men matchning kan förekomma i kursprov D. I LÄSA används kortsvar, flervalsuppgifter samt matchning (textsvar/bildsvar). I kursprov B i TALA finns enbart en öppen muntlig uppgift per provversion. I SKRIVA i kurs B ges eleverna en skrivuppgift som eleverna besvarar genom att skriva för hand i ett pappershäfte. I dagsläget ges raka poäng (dvs. ingen viktning) och man ger ett poäng per uppgift undantaget SKRIVA och TALA som bedöms kvalitativt utifrån matriser (är dock under revidering för att bli poängbaserade men det blir inte ett poäng för hela skrivuppgiften).

Insamlad bakgrundsinformation

Den information som idag samlas in i samband med utprövning och det skarpa SFI provet är:

- Ålder
- Kön
- Utbildningsbakgrund (antal år)
- Språkbakgrund (frivillig uppgift om modersmål)
- Antal år i Sverige
- Tid på SFI utbildningen
- Läs/skrivkunnighet på modersmål (samlas enbart in vid utprövning).

Dessutom finns ytterligare tre frågor som rör elevernas bakgrundsinformation för det skarpa provet; deltagande i grundläggande läs- och skrivundervisning, om eleven är i närheten av kursmålen vid provtillfället samt en fråga om kursbetyg. De fyra första variablerna (ålder, kön, utbildningsbakgrund, och språkbakgrund) är de viktigaste och används vid analyser på uppgiftsnivå och resultatanalyser. De övriga variablerna används om det finns osäkerhet och det behövs ytterligare förklaringar till resultat på individ eller gruppnivå.

Nuvarande provanalys och uppgiftsanalys

Provets testspecifikationer reglerar innehåll, format och sammansättning. I de receptiva färdigheterna anges sammansättning av uppgifter utifrån svarsformat, aspekter av hörförståelse och läsförståelse, domäner, texttyper och relationer samt antal ord för delprovet som helhet. För de produktiva färdigheterna anges vilka språkhandlingar som prövas samt ramar för uppgifternas utformning och genomförande. Proven regleras även av Skolverket genom deras mallar för texter, terminologi och layout. Olika provversioner och dess uppgifter jämförs förutom kvalitativt även med hjälp av klassisk testteori samt modern testteori.

De skarpa proven följs upp efter ett år med hjälp av insamlat material i provgruppen och provresultat från SCB. Statistiska analyser görs av de receptiva delproven för varje provversion. För de receptiva delproven analyseras uppgifternas medelfel, standardavvikelse, lösningsproportion, diskriminationsindex, svårighetsgrad, diskriminationsförmåga, mätfel, bortfall, interna konsistens (Cronbachs alfa) samt förhållandet mellan svårighetsgrad och provtagarnas förmåga (variabel-karta). För de bedömningar som referenslärarna gör av SKRIVA analyseras reliabiliteten genom mått på interbedömarreliabilitet och genom analyser av bedömningsmönster hos gruppen. Provens stabilitet över tid undersöks som helhet, på delprovsnivå samt på uppgiftsnivå. De receptiva delprovens stabilitet över tid mäts med stöd av dubbla resultat och ankaruppgifter (båda dessa från utprövningar). Man undersöker även eventuella avvikelser i enskilda uppgifter i relation till elevernas bakgrund. Dessutom undersöker man lärares uppfattning om provet och provets relation till styrdokument.

Notera att i dagsläget finns inte möjlighet att samla in resultat på uppgiftsnivå för samtliga genomförda prov på nationell nivå utan enbart för ca var 5:e elev. Notera vidare att i dagsläget ges delprovsbetyg och inga ankaruppgifter ges i skarpa prov.

Utprövning av nya uppgifter

I dagsläget prövas inte uppgifter ut samtidigt med skarpa prov utan man anordnar speciella utprövningar för elever som lärare bedömer ligga på godkänd nivå och är i slutfasen av sin utbildning alternativt har slutfört utbildningen max fyra veckor innan utprövningstillfället. Med andra ord sker utprövningen i nära anslutning till det skarpa provet. Det är en utmaning att hitta utprövningsgrupper utifrån dessa förutsättningar. Fusk förekommer och i vissa fall så deltar elever som inte är på korrekt nivå (dvs. elever som är i princip klara med kursen). I dagsläget är det en förutsättning att utprövning genomförs i samarbete med konstruktionsgruppen då ett fokussamtal tillsammans med elever och lärare också kan hållas.

Utprövningarna för de receptiva delproven HÖRA och LÄSA görs genom att först genomföra en pilotstudie (ca 50 elever), sedan genomföra ytterligare en utprövning till 200-250 elever. I vissa fall ges sedan en extra omutprövning (ca 50 elever). I utprövningen av LÄSA ges ankaruppgifter för att undersöka provet över tid. Ankaruppgifterna ska representera en lämplig och normerande svårighetsgrad för respektive kursprov och är utvalda efter diskussioner i konstruktionsgruppen. I dagsläget använder man cirka 8-10 ankaruppgifter och ankaruppgifterna är en miniversion av delprovet och placerad i början av delprovet. Det har dock ibland resulterat i uttröttning av eleverna och det har blivit ett högre bortfall på de sista uppgifterna i utprövningen. I framtiden skulle det eventuellt vara intressant att istället ankra mot prov som har standardsatts enligt gängse principer för standardsättning.

Av praktiska skäl prövas skriv- och talfärdigheter i mer begränsad utsträckning än hörförståelse och läsförståelse. Cirka 100 texter samlas in per uppgift och version för SKRIVA. Det är en utmaning att få in texter på samtliga nivåer eftersom de elever som deltar i utprövningen har bedömts av lärare att de klarat kursen (minst betyg E) och därmed gått vidare till nästa kurs. Delprovet TALA prövas ut av ca 25 elever och fokus ligger på om instruktionerna är tydliga och begripliga och om uppgiften och ämnet ger deltagaren möjlighet att visa språklig bredd.

Efter utprövningen sammanställs detaljerad statistik över utprövningsurvalet, deltagande skolor, resultat kopplat till tidigare provversioner, ankaruppgifter, nationell statistik från Skolverket, reliabilitetsmått över tid, dubbla resultat, analys av resultat på uppgiftsnivå i förhållande till de bakgrundsvariabler som samlas in och Raschanalyser på uppgiftsnivå. Även misstänkt fusk eller andra analyser som är relevanta för den aktuella utprövningsversionen kan ske här.

När man gör den stora utprövningen undersöks relationen mellan provpoäng, delprovspoäng och uppgifter med elevernas bakgrundsvariabler för att undvika att en uppgift inte gynnar/missgynnar en viss grupp. Att använda dubbla resultat ger stöd för att jämföra svårighetsgraden hos det skarpa provet med den nya utprövningsversionen. Vidare går det att följa hur resultaten på utprövningarna samvarierar med resultaten på de skarpa proven över tid.

I dagsläget sätts poänggränserna för de receptiva delproven i två steg. Detta görs genom att referenslärare vid sammanställning av utprövningsversionen skattar uppgifter och genom noggranna urval av uppgifter mellan utprövningsversionen och den skarpa versionen. Representativa uppgifter väljs ut så att slutversionens

svårighetsgrad blir jämförbar med tidigare provversioner. Antalet uppgifter inom delproven varierar med 1-2 uppgifter mellan olika provversioner och därmed varierar gränsen för E endast med ett fåtal poäng över tid. I dagsläget genomförs inga regelrätta kravsättningsmöten där referenslärare bjuds in för att sätta kravgränser på ett fastställt prov. Detta håller dock på att förändras och under våren 2019 kommer alla delprov kravgränser sättas enligt Bookmark metoden (se ex. Karantonis & Sireci, 2006). Notera att kravgränssättning för SKRIVA görs under bedömarmötet för varje provversion.

Nuvarande programvara

SPSS och QUEST (för Raschanalys) används i dagsläget för att analysera uppgifterna och proven. Planen är att i framtiden använda Winsteps istället för Quest.

Validitetshot med dagens prov

I dagsläget sprids i princip alla skarpa prov på sociala medier så snart de används. Proven används regelbundet i ett antal år eftersom de ofta finns kvar på skolorna. Att provet sprids på sociala medier är något man måste göra något åt eftersom ett av de största validitetshoten är att eleven känner till provet eller delar av provet i förväg. I framtiden bör man därför försöka minska exponeringen av proven och uppgifterna och skapa förutsättningar för att proven är hemliga innan eleverna skriver dem. För att säkerställa provens kvalitet över tid bör man även fundera vilka kvalitetsmått som man ska använda och hur man får tillräcklig information om eleverna, uppgifterna och proven som är tillförlitlig.

Förslag på förändring av provmodell

Man kan tänka sig att man förändrar provet på olika sätt och här beskrivs möjliga förändringar. Med andra ord beskrivs inte ett förslag på provmodell utan snarare olika möjliga förändringar som är olika enkla att genomföra. Vissa förändringar hänger samman men flera av förslagen är fristående. Alla förslag på förändringar syftar dock till att skapa bra prov i framtiden med hög validitet och reliabilitet och som inte är känd för eleven innan eleven skriver provet samt där uppgifternas exponering begränsas.

Förslag på förändring A

Det största problemet i dagsläget är att provet ges kontinuerligt (i princip de flesta dagarna i veckan året runt) trots att man enbart har tillgång till ett fåtal provversioner. Detta är inte att rekommendera och bör förändras så snart som möjligt. I framtiden bör man skapa möjlighet att dels begränsa antalet tillfällen som provet ges, genom att exempelvis enbart tillåta att prov ges under bestämda veckor eller vissa datum - oavsett vilken aktör som genomför dem. Ett första steg skulle kunna vara att man begränsade möjligheten till att ge ett prov till en vecka per månad under terminstid. Detta skulle resultera i ca 8-9 veckor där det ges möjlighet att genomföra ett prov istället för dagens möjligheter som är att genomföra prov över mer än 40 veckor. Man bör också tydligt definiera vilka provversioner som får användas vid vilka provtillfällen.

Förslag till förändring B

Man bör skapa möjligheter och underlätta att ta fram fler provversioner som kan fungera samtidigt för att garantera ett rättssäkert prov för alla elever. Om man skapar ett större antal provversioner så minskar risken att eleven fått tag i aktuell provversion innan eleven har genomfört det skarpa provet. En nackdel med att skapa fler provversioner av dagens typ är att det är en relativt kostsam åtgärd. I dagsläget förekommer det på en del skolor som har kvar gamla provversioner som fortfarande faller under sekretess att dessa återanvänds om aktuell provversion blivit känd i förväg för eleverna.

Förslag på förändring C

Man bör påbörja skapandet av en uppgiftsbank så att man kan återanvända väl fungerande uppgifter – se avsnittet om uppgiftsbanker vad man bör tänka på. Vid införandet av en uppgiftsbank är det bra att kategorisera uppgifterna på olika sätt men också skapa ett system där man kan följa en uppgift från utprövning till hur den fungerar i skarpa prov (se avsnitt om uppgiftsbanker). Genom olika kvalitetskontrollmått (se ex. Wiberg & von Davier, 2017) kan man se om en uppgift slutar fungera som man vill att den ska fungera. Detta skapar möjligheter att säkerställa att provversionerna är jämförbara över tid. Om en uppgiftsbank byggs upp på lämpligt sätt så blir det enkelt att leta efter jämförbara uppgifter som man kan byta ut de uppgifter som eventuellt har blivit kända bland eleverna i förtid. Att ha en uppgiftsbank underlättar även om man väljer att skapa moduler av uppgifter som man

sedan kombinerar på olika sätt. Om man ska aktivt kunna använda en uppgiftsbank så bör man se över sekretessreglerna så att sekretessen gäller under en mycket längre tid än idag.

Förslag på förändring D

Man bör skapa moduler av uppgifter med kända egenskaper. Om man bygger upp modulerna utifrån en given modell skulle man kunna byta ut en modul om den skulle bli känd. Det bästa vore om varje provversion består av ett antal moduler som man sedan kan kombinera på en mängd olika sätt. Detta minskar risken att eleverna i förväg vet vilka moduler som används i dennes prov även om någon modul blivit känd i förväg. Det skapar också stora möjligheter att generera ett större antal provversioner istället för ett fåtal. Redan idag borde det inte vara några problem att ta exempelvis delprovet LÄSA från en provversion och kombinera med delprovet SKRIVA från en annan provversion. På så sätt får man ytterligare en provversion. Detta kräver dock att man ser över sekretessreglerna så att sekretessen gäller under en mycket längre tid.

Förslag på förändring E

Man bör digitalisera provet. Fördelen med en digitalisering av provet är att man får möjlighet att samla in delprovsresultat och uppgiftsresultat från alla eleverna. I dagsläget försöker man samla in resultat från ungefär var femte elev men i realiteten får man in betydligt färre provresultat. En annan fördel är att man får en enklare tillgång till information om elevernas uppgiftsresultat och (del)provresultat samt man får utökade möjligheter till att följa uppgifter från utprövning till skarpt prov. En tredje fördel är att man kan samla in elevernas svarstider på uppgifterna – vilket kan ge mycket information om eleven och den informationen kan användas på en mängd olika sätt (Lee & Chen, 2011). Slutligen så förenklas utbytet av uppgifter i ett skarpt prov om någon uppgift kommer ut på sociala medier. En nackdel kan vara att olika delprov kan vara olika enkla att digitalisera.

Förslag på förändring F

Man bör fundera över att göra delar av provet (eller ett visst delprov) till ett MST för att minska exponeringen av provet (se avsnittet om MST). Detta är kanske framförallt relevant för delprov LÄSA och HÖRA. Detta förutsätter dock en digital lösning som tillåter detta. Det förutsätter också att man bygger ihop lämpliga moduler av uppgifter som kan fungera tillsammans.

Förslag på förändring G

Man bör försöka skapa möjligheter att genomföra utprövningen i samband med det skarpa provet. Fördelen med att pröva uppgifterna i samband med ett skarpt prov är att det blir en korrekt provtagargrupp som får göra utprövningen och det då blir en högre kvalitet på utprövningen i och med att man får bättre skattningar på uppgifternas egenskaper. Man kan även testa fler uppgifter om alla elever gör utprövningsdelar. Utprövning i samband med skarpt prov kan göras på olika sätt men beror delvis på hur provet distribueras. Om provet distribueras via papper så är det att föredra att vika en del till utprövningsdel, dvs. att eleven får genomföra en extra del. Man bör dock vara noga när man utformar den och provet så att eleven inte vet vilken av delarna som är utprövningsdel. Man kan exempelvis ge två LÄSA delar till en elev varav en del är utprövningsdel och en del tillhör det skarpa provet. Att lägga in utprövningsdelar underlättas om man har moduler av uppgifter som kan läggas in. Om proven inte är digitala ställs dock höga krav på administrationen kring proven för att säkerställa att utprövningsuppgifterna inte kommer på villovägar. Det ställer även höga krav på hur återrapportering av utprövningsresultatet ska ske. Alternativt, om provet levereras digitalt så har man även möjlighet att blanda in utprövningsuppgifter bland de skarpa uppgifterna. Om provet sker digitalt så underlättas administrationen och man behöver enbart testa en uppgift tills man har tillräcklig information om uppgiften. Man kan då begränsa exponeringen av utprövningsfrågorna eftersom man kan byta ut en uppgift eller modul av uppgifter när man har tillräcklig information och pröva ut andra uppgifter istället. Oavsett om provet sker med penna och papper eller digitalt kan man låta elever som får olika provversioner få olika utprövningsdelar. Nackdelar med att pröva ut uppgifter i samband med ett skarpt prov är att provet blir längre och det finns en risk att uppgifterna blir kända innan de används i ett skarpt prov. Vidare kan det vara svårare att skapa möjlighet att diskutera utprövningsuppgifterna med eleverna.

Förslag på förändring H

Om man kan skapa möjligheter att utföra utprövning av uppgifter i samband med skarpt prov så bör man även undersöka möjligheten att använda sig av ankaruppgifter för att kunna undersöka hur proven fungerar över tid. Ankaruppgifter ger även information om kunskapsnivån på den aktuella elevgruppen. Viktigt är att det tas fram ett antal ankaruppgifter om någon av dessa blir kända. Notera dock att om man ger proven digitalt så behöver enbart en del av eleverna få ankaruppgifterna, andra elever kan få utprövningsuppgifter som dessutom kan skilja sig åt mellan provorterna. På det sättet minskar man uppgiftsexponeringen.

Förslag på förändring I

Man bör utöka dagens insamling av hur uppgifterna och proven fungerar över tid genom att även använda olika statistiska kvalitetskontrollmått samt möjliggöra statistisk ekvivalering av proven. Om man använder provversioner kontinuerligt över en längre tid är det viktigt att undersöka att uppgifternas och provens egenskaper är likadana som när provet började användas. Vidare är både preekvivalering och postekvivalering viktiga instrument att undersöka för nya versioner av prov gentemot befintliga versioner av prov. Preekvivalering kan ges innan provet för att säkerställa jämförbarheten med tidigare provversioner. Även om man behåller fasta kravgränser som publiceras innan provet är genomfört så kan man genomföra en postekvivalering för att säkerställa att provversionerna verkligen är jämförbara över tid. Om resultatet av ekvivaleringen förändras över tid är det viktigt att undersöka vad det beror på och om det då innebär att delar av provet eller hela provet har blivit känd innan provet ges till eleverna.

Förslag på förändring J

Man bör se över poängsättningen. I en framtida provmodell så bör delprovsbetygen tas bort (de strider mot systemramverket) och eleverna bör erhålla raka poäng inom varje delprov (vilket detta prov tidigare haft med ett välfungerat resultat) vilka kan summeras till en totalpoäng. För att ge eleverna återkoppling på provet så kan man tänka sig ge dem feedback om deras kunskaper inom varje prov. Det kan dock vara viktigt att undersöka hur mycket de olika delprovspoängen bidrar och bara rapportera om de har adekvat psykometrisk kvalitet (Sinharay, Puhan, & Haberman, 2011). För eleverna kan det vara viktigt med information om delprovspoäng för att veta mer om vilka delar av deras språkutveckling som är stark eller svag. Genom att använda rak poängsättning och inte vikta provpoäng skapar man ett enklare sätt att sätta samman prov och moduler av uppgifter.

Förslag på förändring K

Man bör fundera på om man ska införa möjlighet till bedömningsträning för undervisande lärare samt erbjuda utbildning i bedömningsmatriserna. I dagsläget ges regelbunden utbildning enbart till lärarna i bedömargrupperna. Bedömningsträning skulle troligtvis öka samstämmigheten i bedömning mellan lärarna.

Förslag på förändring L

I dagsläget genomförs inga regelrätta kravsättningsmöten där referenslärare bjuds in för att genomföra en kravgränssättning ett fastställt prov. Man kan fundera på om man ska genomföra den typen av möten i framtiden för att säkerställa kunskapsnivån på hela provet. Detta är något som i dagsläget håller på att utformas och ska användas från och med våren 2019.

Förslag på förändring M

Man kan fundera på vilka aktörer som ska ha rätt att administrera proven och vilka krav som de ska uppfylla. Viktigt är att man studerar hur proven fungerar på olika platser och att alla riktlinjer följs samt att prov byts ut när de kan anses vara kända. Man skulle kunna fundera på om det krävs någon form av kvalitetsstämpel liknande som krävs för testcentra i Danmark. Oavsett om man behöver en kvalitetsstämpel eller inte bör man kontinuerligt följa upp hur de riktlinjer man gett uppfylls.

Förslag på förändring N

Man bör göra en undersökning bland tidigare elever för att se på vilket sätt de använder provresultaten eller provbetygen i praktiken. Används de enbart för att bedöma kunskapsnivån eller används de på andra ställen i samhället. Om de används på andra ställen (ex. vid arbetsansökningar) bör man fundera på om det påverkar provet och om det påverkar vilka åtgärder man måste ta.

Nationella prov i matematik (2b) som används till vuxna

Ett flertal nationella prov i matematik för vuxna används, fokus i denna rapport är dock framförallt på kurs 2b. De nationella proven i matematik för vuxna genomförs oftast efter att en elev har avslutat en kurs. Det finns ett stort antal aktörer både i offentlig och privat regi som ger proven och det finns i dagsläget inte någon samordning på samma ort när provet ges. De nationella proven i matematik används mycket och exponeras därmed i hög grad. Alla olika kursprov ges till exempel i Stockholmsområdet under en och samma vecka. Det finns alltid två provversioner tillgängliga (höstprovet samt vårprovet, som är utvecklad för gymnasieskolan). En elev gör hela kursprovet utom eventuell muntlig del (kurs 1 och kurs 3) på samma dag. Proven kan både genomföras i grupp i klassrum och enskilt eftersom en del elever kan sitta på distans och därför göra provet på ett speciellt ställe. Alla poäng på provet är i dagsläget värda lika mycket. Provpoängen vägs inte samman mellan delarna utan man får en totalpoäng och utifrån totalpoängen kan man få ett provbetyg. Totalpoängen kan variera med enstaka poäng mellan åren och därmed också kravgränserna. Kravgränserna sätts innan provet ges med hjälp av en utökad Angoff procedur, för detaljer om metoden hänvisas till Hambleton och Plake (1995) samt Lind Pantzare (2017).

Utgångspunkterna för de befintliga proven är rättvisa, reliabilitet och validitet vid provens konstruktion, genomförande och bedömning. Proven består av lite olika delprov beroende på vilken kurs som avses. Gemensamt är dock att det ges möjlighet både till flervalsfrågor och fullständiga lösningar. Delproven testar en blandning av olika förmåga och innehåll. I vissa delprov får man använda digitala hjälpmedel och i andra inte. Kurs 1 skiljer sig mest från kurs 2-4, så fokus är här på de senare kurserna där det finns tre delprov B-D. Delprov B innehåller uppgifter som kräver kortsvar och flervalsfrågor där räknare inte är tillåten. I delprov C finns uppgifter som kräver fullständiga lösningar men där räknare inte är tillåtet. Slutligen, i delprov D ges uppgifter som kräver fullständiga lösningar och räknare är tillåtet.

Nuvarande uppgiftsformat

Beroende på delprov och vilken typ av uppgift så används ett antal olika uppgiftstyper i de olika delproven och dessa är i dagsläget;

- Flervalsfrågor med ett korrekt svarsalternativ
- Korta svar
- Långasvar: Fullständiga lösningar krävs för full poäng.

Insamlad bakgrundsinformation

I dagsläget anger lärarna information om elevernas bakgrund på följande variabler både vid utprövningar och vid skarpa prov;

- Kön
- Modersmål, om det är möjligt samlas information in om huruvida de har ett annat modersmål eller inte (men inte vilket modersmål de då har).
- Preliminära kursbetyg.

Nuvarande uppgiftsanalys

Provet består av allt från flervalsfrågor till fullständiga lösningar. När det gäller uppgifter som kräver lösningar så är det lärare som bedömer lösningarna. Ibland är det elevens lärare, ibland är det en annan lärare och ibland är det elevens lärare med stöd av andra lärare som bedömer elevernas lösningar. Det finns dock ett bedömningsunderlag, i form av uppgiftsspecifika bedömningsanvisningar samt bedömda elevsvar till vissa uppgifter, som ska följas vid rättning och bedömning av elevernas lösningar. I dagsläget används en mängd olika mått för att undersöka uppgifternas kvalitet vid utprövning och skarpt prov. Dessa inkluderar exempelvis;

- Lösningsfrekvens, dvs. sannolikheten att en provtagare klarar att lösa uppgiften.
- Lösningsproportion för alla elever.
- Uppgifternas diskrimineringsförmåga.
- Modellering av hur en uppgift fungerar för olika totalpoäng med icke-linjär regressionsanalys (för detaljer se Ramstedt, 1996).
- Hur olika elever med olika betyg klarar uppgifterna.
- Differential Item Functioning (DIF) undersöks med hjälp av Mantel-Hanszel samt logistisk regression där hela prov undersöks i utprövning av hela prov, och i skarpa prov.

Eftersom provutvecklarna i matematik använder sig av en uppgiftsbank så sparar de en mängd information om uppgifterna som inte enbart är statistiska mått. I nuläget kategoriseras därför uppgifterna utifrån exempelvis

- Förmåga
- Centralt innehåll
- Kunskapskrav
- Om det finns en figur eller inte.
- Om det finns en kontext eller inte.
- Kortsvar/långsvar
- Om digitala verktyg används eller inte.

Nuvarande provanalys

En relativt omfattande analys av provet görs med hjälp av statistiska mått från den klassiska testteorin. Provens reliabilitet beräknas med hjälp av Cronbach's alfa och man undersöker medelfel och testinformationsfunktionen, vilken anger hur mycket information provet ger för olika totalpoäng. För att undersöka hur proven fungerar så undersöker man ibland klassificeringskorrekthet i provbetygen. För att skapa likvärdiga provversioner används skelettfrågestammar för att kunna skapa likvärdiga uppgifter och på så sätt skapa likvärdiga provversioner över tid. Dessutom innehåller alla provversioner några standarduppgifter med kända egenskaper för att få information om elevgruppens sammansättning.

Utprövning av nya uppgifter

I dagsläget använder man både en mindre utprövning och en helprovsutprövning. I den mindre utprövningen så utprövas 5-6 uppgifter på ca 100 elever och de analyseras utifrån betygsnivågrupp. Den mindre utprövningen genomförs av elevernas lärare.

Helproven utprövas i normalfallet på cirka 200 elever men varje enskild uppgift ska i slutändan vara utprövad av 300-400 elever. Man använder ca 15 skolor/kunskapsprov per termin. Vid helprovsutprövningarna används kända lärare (ca 20-50 st.) och därmed utförs utprövningen i deras klasser. Den information som man får in om uppgifterna används sedan för att applicera Ramstedts (1996) metod. Metoden bygger på att man för varje uppgift beräknar medelvärdet av lösningsproportionen för elever med respektive totalpoäng. Därefter anpassas en logistisk regressionskurva till dessa punkter vilket ger en matematisk beskrivning över hur uppgiften fungerar för olika totalpoäng på provet. De allra flesta uppgifterna ingår i en utprövning av ett helt prov innan de används i ett skarpt prov.

Enbart klassisk testteori används vid uppgiftsanalysen i utprövningarna. Man undersöker testinformationen och medelfelet samt klassificeringskorrektheten, hur den utsatta tiden fungerar samt lärarnas förståelse av bedömningsanvisningar. Hur uppgiften fungerar i utprövningen kopplas sedan ihop med information om hur uppgiften fungerade i ett skarpt prov. I dagsläget är det dock inte möjligt att koppla elevers utprövningsresultat med deras skarpa provresultat. Lärare bedömer elevens lösningar (både i skarpa provet och i utprövningsprovet) och oftast är elevens lärare inblandad, men inte alltid enligt tidigare beskrivning.

Nuvarande programvara

I dagsläget görs provanalyser och uppgiftsanalyser med hjälp av SPSS och Excel.

Validitetshot med dagens prov

Ett av de största validitetshoten är att provet blir känt i förväg för de elever som ska skriva provet eftersom det blivit allt vanligare att prov sprids på sociala medier. I framtiden bör man därför försöka minska exponeringen av proven och uppgifterna och skapa förutsättningar för att proven är hemliga innan eleverna skriver dem. I dagsläget undersöks en mängd mått både på provnivå och uppgiftsnivå och för att säkerställa kvalitet över tid om fler versioner används parallellt så kan man fundera på om man ska använda något fler uppföljningsmått.

Förslag på förändringar av provmodell

De nationella proven i matematik som används i vuxenutbildningarna kan tänkas förändras på olika sätt och i det här avsnittet beskrivs några möjliga förändringar. Precis som i tidigare delar av rapporten så ges inte ett nytt förslag på provmodell utan istället lyfts ett antal möjliga förändringar fram. En del av förändringarna bygger på andra förändringar men ett antal av förslagen är fristående. Syftet med förslagen på förändringar är att ge förutsättningar för att ge bra prov med avseende på validitet och reliabilitet och som inte blir tillgängliga innan eleven skriver provet och där uppgifterna inte exponeras i samma grad som de gör i dagens prov.

Förslag på förändring A

Flera vuxenprov har i dagsläget problem med att uppgifterna blir kända i förväg i och de har en relativt hög exponering av uppgitferna. Det innebär att jämförbarheten skadas när delar av provet eller enstaka uppgifter blir kända för eleverna innan provet ges (via ex. sociala medier). I dagsläget ges de få existerande provversionerna kontinuerligt de flesta vardagarna i veckan året runt. Detta bör förändras så snart som möjligt. Man bör begränsa antalet tillfällen som provet ges. Detta kan göras genom att prov enbart får ges vissa datum eller under vissa perioder. Ett första steg skulle kunna vara att man begränsade möjligheten till att ge proven till en vecka per månad under terminstid. Detta skulle resultera i att man kan genomföra ett prov under ca 8-9 veckor till skillnad mot dagens möjligheter att genomföra prov över mer än 40 veckor. Man bör också tydligt definiera vilka provversioner som får användas vid vilka provtillfällen.

Förslag på förändring B

Något som påverkar att proven är kända i förtid är att det är relativt få provversioner som är aktuella vid samma tidpunkt. Genom att använda sig av den redan existerande uppgiftsbanken skulle man kunna skapa ett större antal provversioner som är giltiga under samma tid. Detta ställer dock krav på förlängd sekretesstid för uppgifterna. Om det finns ett större antal provversioner skulle det bli mycket svårare för eleverna att i förväg känna till vilken provversion, och därmed vilka provuppgifter, de kommer att få. Om en provversion eller delar av en provversion blir känd är det enklare att byta ut den om det redan finns flera likvärdiga provversioner. En utmaning med detta är att i dagsläget så sätts ett prov samman så att hela provet är balanserat innehållsmässigt. Man kan därför i framtiden behöva begränsa innehållsmässigt vad som prövas i olika delar. En nackdel med att skapa fler provversioner är att det är relativt kostsamt.

Förslag på förändring C

För att utöka antal provversioner relativt snabbt så skulle man kunna använda alla sekretessbelagda prov under en provperiod. Detta skulle utöka dagens två provversioner till 16 provversioner om det görs två provversioner per år och proven är sekretessbelagda i åtta år. Om man vill få tillgång till ytterligare provversioner skulle man kunna utöka sekretesstiden.

Förslag på förändring D

I dagsläget finns redan en uppgiftsbank för de nationella proven i matematik och detta borde utnyttjas mer. Detta kräver dock att man förlänger sekretesstiden på uppgifterna. En möjlig utveckling för att skapa fler provversioner är att sätta ihop ett antal uppgifter i moduler som är jämförbara. Man kan sedan bygga ihop ett flertal skarpa prov med olika kombinationer av olika moduler och på så sätt skapa en större mängd provversioner. Man skulle även kunna återanvända moduler vid senare tillfällen om man vet att dessa moduler fungerat tillfredsställande och att de inte blivit spridda på sociala medier.

Förslag på förändring E

Man bör digitalisera proven, i alla fall delar av proven. De delar som är mest relevanta att digitalisera är de som har kortsvar och flervalsfrågor (dvs. de som utgör ca 50% av provet). Man bör kunna effektivisera bedömningen genom att automaträtta fler frågor. De uppgifter där det krävs fullständiga lösningar är det mer tveksamt hur mycket man skulle vinna på en digitalisering eftersom det inte är så enkelt för en elev att svara på uppgiften på en dator. Möjligtvis om eleven kan svara på en läsplatta så att det de facto fungerar likvärdigt som papper och penna. Fördelar med digitalisering är dock att det blir enklare att samla in uppgiftssvar, delprovssvar och provresultat för alla elever som kan kopplas på övrig information i uppgiftsbanken. En annan fördel är att man kan samla in elevernas svarstider på uppgifterna – vilket skulle kunna ge mer information om elevens kunskap (Lee & Chen, 2011). Slutligen så förenklas utbytet av uppgifter i ett skarpt prov om provet är digitalt om någon uppgift kommer ut på sociala medier eftersom det skulle vara enklare att byta ut enstaka prov än i ett papper och penna prov.

Förslag på förändring F

Om man skapar moduler av uppgifter såsom det beskrevs i förslag på förändring E så kan man i ett nästa steg göra provet (eller delar av provet) till ett MST. Fördelen med att använda MST är att eleverna får uppgifter på sin svårighetsgrad och att man inte överexponerar uppgifter som är för svåra eller för enkla till provgrupper som inte klarar dessa eller är överkvalificerade för att klara dessa. Man bör dock inte använda för många stadier (max tre) eller moduler i varje prov. Detta kräver dock att modulerna som används i varje stadium är byggda utifrån en tydlig modell.

Förslag på förändring G

Man bör fundera över om man kan genomföra utprövningar i samband med det skarpa provet eftersom man då når den grupp man är intresserad av, och gruppen är motiverad. Nyckeln ligger dock i att eleverna inte ska veta vilka uppgifter som är utprövningsuppgifter och vilka som är skarpa uppgifter. Genom att låta utprövningarna ske i samband med det skarpa provet så kan man om man vill få ett större underlag på uppgifterna. För flervalsfrågorna underlättas utprövning i samband med skarpt prov om provet är digitaliserat eftersom det är enklare att blanda in

utprövningsuppgifterna bland de skarpa uppgifterna. Man kan då även pröva eventuella ordningseffekter av uppgifterna och svarsförslagen. Alternativt kan man skapa speciella utprövningsdelar i likhet med hur det görs i det svenska högskoleprovet där provtagarna ges fyra skarpa delar och en utprövningsdel. En nackdel med att lägga utprövningarna i samband med det skarpa provet är dock att provtiden kommer att utökas vilket kan leda till uttröttningseffekter hos eleverna.

Förslag på förändring H

För att säkerställa reliabiliteten över tid kan man införa ankaruppgifter. Dessa kan användas både för att undersöka hur provet fungerar samt elevgruppens sammansättning. I dagsläget finns inga regelrätta ankaruppgifter som man kan använda för att jämföra provversioner med varandra över tid. Däremot finns nästan alltid någon standarduppgift med som man känner till hur den fungerar vilket kan ge en fingervisning om elevernas kunskap. I dagsläget samlas även information in om elevernas preliminära betyg för att på så sätt få en indikation om elevgruppens sammansättning.

Förslag på förändring I

Man bör utöka sitt kvalitetskontrollsystem så att man kan jämföra hur provversioner eller delprov fungerar över tid. Innan ett prov ges kan man genomföra en preekvivalering för att se hur provet ska ankras. När det skarpa provet har getts så kan man göra en postekvivalering för att undersöka hur det fungerar. Detta är lämpligt att göra också när provet gått en tid för att se att det fortfarande fungerar som det är tänkt att det ska fungera. Denna förändring underlättas om man har ankaruppgifter.

Förslag på förändring J

För att säkerställa reliabiliteten skulle man kunna erbjuda lärarna att delta i bedömningsträning. Även om det finns bedömningsanvisningar där det finns generell information om hur vissa typer av fel ska hanteras och hur bedömningsanvisningarna ska läsas samt bedömda elevsvar till vissa av uppgifterna så kan det vara lämpligt att erbjuda en kortare utbildning där bedömande lärare får möjlighet att diskutera olika lösningsförslag för att skapa en ännu högre samstämmighet.

Förslag på förändring K

Man kan fundera på vilka aktörer som ska ha rätt att administrera proven och vilka krav som de ska uppfylla. Viktigt är att man studerar hur proven fungerar på olika platser och att alla riktlinjer följs samt att prov byts ut när de kan anses vara kända. Man skulle kunna fundera på om det krävs någon form av kvalitetsstämpel liknande som krävs för testcentra i Danmark. Oavsett om man behöver en kvalitetsstämpel eller inte bör man kontinuerligt följa upp hur de riktlinjer man gett uppfylls.

Nationella prov i engelska (6) som används till vuxna

Det finns flera kurser i engelska som ges till vuxna men här utgår vi från kurs 6. De nationella proven i engelska för vuxna genomförs mot slutet av en kurs och det finns ett stort antal aktörer både i offentlig och privat regi. I dagsläget finns det ingen någon samordning på samma ort när ett prov ges. Dessa prov ges regelbundet och används oerhört mycket och exponeras därmed i hög grad. Första dagen en ny provversion kan ges är samma dag som provet ges i ungdomsskolan, därefter kan provet användas under ett år. Tre nya provversioner finns tillgängliga varje år och varje provversion har en sekretesstid på åtta år. Delprov kan dock återanvändas under sekretesstiden av Skolverket. Totalpoängen på en provversion är i stort sett samma oavsett provversion och kravgränserna sätts innan provet ges med hjälp av en Angoff-relaterad modell med stöd från de empiriska utprövningarna med drag av Bookmark-metoden (Angoff, 1971; Kaftandjieva, 2004; Karantonis & Sireci, 2006).

Utgångspunkterna för de befintliga proven är rättvisa, reliabilitet och validitet vid provens konstruktion, genomförande och bedömning. De nationella proven i engelska består av tre delprov. Två av delproven är produktiva och interaktiva och fokuserar på talande och skrivande. Det tredje delprovet är receptivt och fokuserar på att lyssna och läsa. Olika delprov görs i regel på olika dagar i klassrum och därmed kan det vara olika antal kursdeltagare vid tillfällena. I delprov A (Muntlig produktion och interaktion) ges en uppgift där eleven samtalar med en partner och utvecklar ett innehåll och uttrycker sina åsikter. För att kunna genomföra bedömningen har lärarna tydliga bedömningsanvisningar med exempel på olika nivåer.

I delprov B (reception) prövas läsförståelse och hörförståelse. Lärare rättar elevernas svar på proven. I delprov B används uppgifter med raka poäng och flera av uppgifterna är av dikotom typ (rätt/fel). Om mer komplexa svar efterfrågas finns utförliga bedömningsanvisningar med kommenterade exempel på olika nivåer. Efter avslutad bedömning av delprovet så överförs lärarnas sammanräknade poäng till en skala som är identisk med alla delprov. De olika färdigheterna som testas i de tre delproven tillmäts lika värde. Eftersom reception innehåller både muntlig och skriftlig förmåga dubbleras värdet på detta delprov när man räknar samman till ett provbetyg. Varje färdighet och uppgift redovisas också för sig vid sidan av det sammanvägda resultatet.

Nuvarande uppgiftsformat

I delprov A (Muntlig produktion och interaktion) så ges en större muntlig samtalsuppgift där eleven ska samtala med en partner. I delprov B1 (Reception - läsförståelse) samt delprov B2 (Reception – hörförståelse) används

- Flervalsfrågor med ett rätt svar.
- Kortsvar, eleverna skriver ett eller flera ord.
- Lucktexter, eleverna fyller i ett saknat ord i en text genom att antingen skriva själv eller välja ett ord från fyra alternativa ord.
- Info seek uppgifter, eleverna ska översiktsläsa en text och svara i en del fall med öppna svar och i andra fall med flervalssvar.
- Matchning, eleverna kombinerar texter med passande fraser eller
- Flervalsfrågor med ett rätt svar.
- Öppna frågor med konstruerade svar
- Frågor som ger partial credit.

Fördelningen mellan flervalsfrågor och korta svar är ungefär 50/50 med en större andel flervalsfrågor för hörförståelsen. I delprov C (produktion och interaktion) används en längre skrivuppgift.

Insamlad bakgrundsinformation

I samband med utprövningarna samlas information om elevens kön, betyg i engelska på föregående nivå, när de började läsa kursen, vilken ort utprövningen skett på, program i gymnasieskolan samt eventuella lärares kommentarer. Vid det skarpa provet samlas information in om skola, program i gymnasieskolan, födelsedatum, när de började läsa den engelska kursen samt vilken skola eleven tillhör.

Nuvarande uppgiftsanalys

I alla delprov används raka poäng med samma poängskala för delprovsresultat. De flesta uppgifterna är dikotoma (rätt/fel vilket ger 0/1 poäng) men det finns även enstaka uppgifter som ger flera poäng (och som bedöms enligt partial credit). Uppgifterna undersöks med en stor mängd av kvantitativa och kvalitativa analyser. I dagsläget används framförallt mått från den klassiska testteorin men dessa kompletteras med mått från den moderna testteorin från och med 2019. Exempel på mått som används i dagsläget är;

- Lösningsfrekvenser för alla uppgifter.
- Lösningsproportion för alla elever.
- Uppgiftens svårighetsgrad.
- Uppgiftens diskrimineringsförmåga (punktbiserial korrelation)

Vid utprövningarna undersöker man uppgifterna med ovanstående mått men man har även tillgång till mer bakgrundsinformation om eleverna. Även kvalitativa metoder används för att undersöka språk, diskurs och uppfattningar. I framtiden planerar man även att genomföra Raschanalyser på utprövningsdata och skarpa uppgifter.

Nuvarande provanalys

De resultat som idag kommer in för engelska 6 från vuxna räcker inte till för att göra analyser i en postvalidering. De provanalyser som beskrivs här gäller därför för hela provet inklusive de som ges till ungdomsskolan. I dagsläget undersöks regelbundet det aktuella provets resultat i jämförelse med tidigare års resultat. Reliabiliteten på provet undersöks med hjälp av Cronbach's alfa och KR-20. Vidare undersöks bedömarnas överensstämmelse med hjälp av korrelationsanalyser (Kendall, Pearson samt Spearman), medelvärden och standardavvikelser. Medelfel undersöks samt korrekthet i klassifikationer av betygssteg. Elevernas produktiva svar bedöms av lärare, och till sin hjälp har de utförliga bedömningsanvisningar. Bedömningsanvisningarna innehåller både övergripande principer samt specifika anvisningar. Framförallt handlar detta om hänvisning till kunskapskrav samt generiska bedömningsfaktorer.

För att undersöka provens stabilitet används i varje utprövningsprov en ankaruppgift. Ankaruppgiften kan vara i form av en text som i sin tur ofta består av 10-12 frågor. Vidare återanvänds alltid en gammal uppgift i ett skarpt prov för att kunna koppla olika provversioner mellan varandra. I dagsläget görs ingen regelbunden statistisk ekvivalering före eller efter provens genomförande. Vid efteranalysen av det skarpa provet används data från SCB med avseende på elevens kön, samt program i gymnasieskolan. För kursen engelska 6 som även ges till ungdomsskolan studeras utfall i relation till utprövningsdata, ankaruppgifter och gjorda prediktioner. Samband mellan uppgifter och delprov beräknas liksom medelfel, grad av korrekthet i klassifikationer etc.

Utprövning av nya uppgifter

Nya uppgifter prövas först ut i miniutprövningar vid ett antal skolor. Utifrån resultaten på miniutprövningarna modifieras uppgifterna och sedan går uppgifterna till storskaliga utprövningar i slumpvis utvalda elevgrupper i landet. Omkring 400 elever används för att testa varje uppgift och i samband med de storskaliga utprövningarna så administreras även en ankaruppgift samt enkäter till elever och lärare. En ankaruppgift (ex. en text omfattande ca 10-12 frågor) ingår alltid i en utprövning. Flera av ankaruppgifterna är s.k. lucktexter där betydelsebärande ord tagits bort ur en kortare dialog eller mening. Ankaruppgifterna bidrar till analysen och utvärderingen av utprövningsresultatet samt ger möjlighet till trendstudier. Utprövningsversionerna sätts samman utifrån standardiserade mallar med avseende på antal ord, tid för avlyssning och placering av ankaruppgift.

Preliminära bedömningsanvisningar till lärarna prövas i samband med utprövningen av uppgifter. Bedömningen av utprövningsuppgifterna är frivillig och ingen feedback ges till lärarna. Enkäterna till lärare och elever fokuserar på uppgifternas relevans i relation till kursplanen respektive undervisningsgruppen/individen, deras innehåll och svårighetsgrad, tidsaspekter, upplevd rättvisa och nytta samt lärares och elevers allmänna eller känslomässiga reaktioner.

Nuvarande programvara

I dagsläget används EXCEL och SPSS för att analysera uppgifter och prov i utprövningsproven och i de skarpa proven.

Validitetshot med dagens prov

Ett av de största validitetshoten mot dagens prov är att uppgifterna blir kända för eleverna innan de genomför provet. Detta problem kännetecknar alla vuxenproven och man måste därför i framtiden skapa bättre förutsättningar för att minska exponeringen av proven och uppgifterna samt att säkerställa att proven är hemliga innan eleverna skriver dem. I dagsläget är det problematiskt att ungdomsskolan enbart ger provet en viss dag och en viss tid medan vuxenskolan sedan kan använda provet regelbundet under en längre tid. Ett annat validitetshot är att delproven inte används och genomförs eller kombineras enligt anvisningarna Detta är framförallt ett hot när proven används i en flexibel verksamhet såsom vuxenverksamheten.

Förslag på förändringar av provmodell

Provet och provmodellen skulle kunna förändras på olika sätt och här beskrivs några möjliga förändringar. Notera att detta inte är ett samlat förslag på provmodell utan istället ett antal möjliga förändringar som är olika svåra att genomföra och förknippade med olika kostnader. Vissa förändringar bygger på varandra medan andra är fristående. Vidare så kan vissa av förslagen på förändringar passa bra till ett eller flera delprov men kanske inte lika bra till hela provet. Det övergripande syftet med förslagen på förändringar är att ge förutsättningar för att ge bra prov med avseende på validitet och reliabilitet och som inte blir tillgängliga innan eleven skriver provet och där uppgifterna inte exponeras i samma grad som de gör idag.

Förslag på förändring A

Ett av de största problemen med dagens vuxenprov är att proven och uppgifterna exponeras oerhört mycket i och med att de administreras de flesta dagarna i veckan under hela året trots att det enbart finns ett mycket begränsat antal giltiga provversioner. Detta gör att det är stor risk att uppgifterna sprids i sociala medier och att provtagarna därmed känner till dem innan de genomför provet. Att provet ges hela tiden året runt bör förändras så snart som möjligt. I framtiden bör man kraftigt begränsa antal provtillfällen, genom att exempelvis enbart tillåta att prov ges under bestämda veckor eller under vissa datum. Ett första steg kan vara att begränsa möjligheten till att skriva ett prov till en vecka per månad under terminstid. Detta skulle minska antalet tillfällen drastiskt. Man bör även bestämma vilka provversioner som får ges vid de olika tillfällena. Begränsningen av antal provtillfällen bör gälla oavsett vilken aktör som ger proven.

Förslag till förändring B

För att undvika att eleverna får kännedom om provuppgifterna innan de skriver en provversion bör man underlätta att ta fram fler provversioner som kan fungera samtidigt. Om man skapar ett större antal provversioner som är giltiga samtidigt så minskar risken att eleven har tagit del av aktuell provversion innan eleven har genomfört det skarpa provet. En nackdel med att skapa fler provversioner av dagens typ är att det är en relativt kostsam åtgärd.

Förslag på förändring C

Även om man regelbundet återanvänder uppgifter idag så bör man i framtiden fundera på att systematiskt bygga upp en uppgiftsbank så att det förenklar återanvändandet av väl fungerande uppgifter – se avsnittet om uppgiftsbanker. Även om detta framförallt är applicerbart på delprov B så bör man även samla uppgifterna från delprov A och delprov C i uppgiftsbanker så att de skulle kunna användas vid ett senare tillfälle. Detta bygger dock på att man ser över dagens sekretesstid om åtta år och utökar den så att det blir meningsfullt att använda sig av uppgiftsbanker. Vid införandet av en

uppgiftsbank är det bra att kategorisera uppgifterna på olika sätt men också skapa ett system där man kan följa en uppgift från utprövning till hur den fungerar i skarpa prov (se avsnitt om uppgiftsbanker). Genom olika kvalitetskontrollmått (se ex. Wiberg & von Davier, 2017) kan man se om en uppgift slutar fungera som man vill att den ska fungera. Detta skapar möjligheter att säkerställa att provversionerna är jämförbara över tid. Om en uppgiftsbank byggs upp på lämpligt sätt så blir det enkelt att leta efter jämförbara uppgifter som man kan byta ut om man har uppgifter som man misstänker har blivit kända bland eleverna i förtid.

Förslag på förändring D

För att öka antal provversioner kan man fundera på att skapa moduler av uppgifter med givna egenskaper som tillsammans bildar ett delprov. För engelska 6 är detta relevant för delprov B medan det är mindre relevant för delprov A (muntlig uppgift) och delprov C (längre skrivuppgift). Om man bygger upp modulerna av uppgifter utifrån en given modell skulle man kunna byta ut en modul om den skulle bli känd utan att nödvändigtvis behöva byta ut hela delprovet. Detta ställer dock stora krav på hur modulerna är skapade och att modulerna görs likvärdiga utifrån innehåll och val av uppgifter. Om man lyckas skapa moduler av uppgifter så ges möjligheten att generera ett flertal delprovsversioner istället för ett fåtal. Om varje version av delprovet består av ett antal moduler minskar risken att eleverna i förväg vet vilka moduler som kommer används i dennes delprov. Skapandet av moduler underlättas om man har en uppgiftsbank som man kan välja lämpliga uppgifter från och kombinera dessa på lämpligt sätt.

Förslag på förändring E

Provet bör digitaliseras i framtiden. Det finns redan idag ett antal digitala verktyg för att kunna automaträtta flervalsuppgifter, kortsvar och längre skrivuppgifter. En fördel med en digitalisering av provet är att det ges möjlighet att samla in delprovsresultat och uppgiftsresultat från alla elever. En annan fördel med en digitalisering är att det blir enklare att samla in uppgiftssvar, delprovssvar och provresultat för alla elever som kan kopplas på en uppgiftsbank. En tredje fördel är att man får utökade möjligheter att följa uppgifter från utprövning till skarpt prov. Slutligen så förenklas utbytet av uppgifter i ett skarpt prov om någon uppgift kommer ut på sociala medier. En nackdel kan vara att olika delprov kan vara olika enkla att digitalisera. Speciellt den muntliga delen kan vara en utmaning att digitalisera på ett relevant sätt.

Förslag på förändring F

Eventuellt kan man fundera på om man ska göra delprov B till ett MST för att minska exponeringen av provet (se avsnittet om MST). Detta förutsätter dock en digital lösning som tillåter detta. Det förutsätter också att man bygger ihop lämpliga moduler av uppgifter som kan fungera tillsammans och att man har tydliga riktlinjer för vad som ska prövas i respektive stadie.

Förslag på förändring G

För att säkerställa att eleverna som genomför utprövningsuppgifterna är motiverade och att vi därmed får hög kvalitet på utprövningen bör man fundera på att genomföra utprövningar av nya uppgifter i samband med att det skarpa provet ges. Detta gäller kanske främst delprov B men skulle troligtvis gå att genomföra med delprov A. Delprov C kan vara svårt att pröva ut eftersom det enbart är en längre skrivuppgift och det finns en stor risk för uttröttning hos eleverna. En fördel med att pröva uppgifter i samband med skarpt prov är att man når rätt målgrupp och det blir högre kvalitet på utprövningen än om man har separata utprövningar. Man får oftast bättre skattningar på uppgifternas egenskaper och man kan testa fler uppgifter om alla elever gör utprövningsdelar. Nackdelen är att proven blir längre, det finns risk för uttröttningseffekter hos eleverna och att det finns en högre risk att uppgifterna blir kända. Vidare kan det vara svårare att skapa möjlighet att diskutera utprövningsuppgifterna med eleverna.

Utprövning av uppgifter i samband med skarpt prov kan göras på olika sätt. Eftersom många av uppgifterna i delprov B i engelska 6 är en samling av uppgifter kopplat till en text eller till en ljudfil så är det mest relevant att pröva ut ett helt delprov eller en modul av uppgifter istället för enskilda uppgifter. Det är dock oerhört viktigt att eleven inte vet vilken del (eller modul) som är utprövning och vilka delprov eller moduler som är skarpa. Detta är viktigt så att eleven är lika motiverad att svara bra på alla delprov (eller moduler). Man kan tex ge två hörförståelse delprov, där en är utprövning och en är ett skarpt delprov. Elever som får olika provversioner kan även ges olika utprövningsdelar eller olika utprövningsmoduler. Om provet administreras digitalt kan man välja att testa en modul eller ett delprov tills man har tillräcklig information om modulen/delprovet och sedan byter man ut den färdig utprövade modulen eller delprovet till andra utprövningsmoduler eller andra utprövningsdelprov.

Förslag på förändring H

Ankaruppgifter ger information om kunskapsnivån på den aktuella elevgruppen. Man bör därför fundera på att skapa fler ankaruppgifter. Om man enbart använder en ankaruppgift så kan man få problem om den blir känd bland eleverna i förväg. Detta blir speciellt viktigt om man börjar använda utprövning i samband med skarpt prov.

Förslag på förändring I

För att få fler indikationer om hur uppgifterna och provversioner fungerar över tid kan man utöka kvalitetskontrollsystemet med statistisk ekvivalering av provversionerna. Innan ett prov ges kan man genomföra en preekvivalering för att se hur provet ska ankras samt för att säkerställa jämförbarhet med tidigare provversioner. När sedan det skarpa provet har getts så kan man göra en postekvivalering för att undersöka hur det fungerar. Detta är lämpligt att göra också när provet gått en tid, exempelvis efter sex månader, för att undersöka provens stabilitet över tid och att provet fortfarande fungerar som det är tänkt att det ska fungera. Om resultatet av ekvivaleringen

förändras över tid är det viktigt att undersöka vad det beror på och om det då innebär att delar av provet eller hela provet har blivit känd innan provet ges till eleverna.

Förslag på förändring J

Man bör undersöka om det finns ett behov hos lärarna (eller hos nya lärare) att få utbildning i bedömningsanvisningarna. Även om bedömningsanvisningarna innehåller generell information om hur vissa typer av fel ska hanteras och hur de ska läsas samt bedömda elevsvar till vissa av uppgifterna så kan det vara lämpligt med en kortare utbildning där de som ska bedöma får möjlighet att diskutera olika elevsvar för att säkerställa samstämmighet.

Förslag på förändring K

För att kunna få större kontroll över giltiga provversioner så kan man fundera på att skilja provet från ungdomsskolan. Eftersom provversioner som ges till ungdomsskolan i sig exponeras mycket så skulle det underlätta om man enbart använder speciella provversioner för vuxenproven – givetvis utifrån samma kravspecifikationer. Detta löser inte problemet med spridning av vuxenproven men skulle kunna leda till att man tillåts ha en annan provmodell för vuxenproven om man måste fortsätta ge många prov under ett stort antal tillfällen varje år.

Förslag till förändring L

I dagsläget så finns det en stor risk att aktuella provversioner eller delprov blivit kända i förväg. Inom en flexibel verksamhet såsom vuxenutbildningen finns det risk att någon aktör därför försöker lösa detta genom att skapa egna provversioner som inte är kända för eleverna genom att kombinera tidigare delprov som inte är skapade för att prövas tillsammans. Genom att ge förslag på riktlinjer om vilka kombinationer av delprov från olika provversioner som får sättas samman om sådana tillfällen uppstår skulle man kunna säkerställa validiteten. Detta förutsätter dock att delproven (eller moduler av dessa) är skapade så att de går att kombinera på ett lämpligt sätt.

Förslag på förändring M

Skolverket bör fundera på vilka aktörer som ska ha rätt att administrera proven och vilka krav som de ska uppfylla. Viktigt är att man studerar hur proven fungerar på olika platser och att alla riktlinjer följs samt att prov byts ut när de kan anses vara kända bland provtagarna i förväg. Man skulle kunna fundera på om det krävs någon form av kvalitetsstämpel liknande som krävs för testcentra i Danmark. Oavsett om man behöver en kvalitetsstämpel eller inte bör man kontinuerligt följa upp hur de riktlinjer man gett uppfylls.

Generella förändringsförslag i provmodeller för vuxenprov

Fokus i den här rapporten har varit på att formulera ett antal förslag på förändringar i provmodellerna som används för vuxenutbildningen för att bättre kunna hantera den frekventa användningen och exponeringen av nationella prov. Detta har gjorts utifrån analys av Skolverkets systemramverk (Skolverket, 2017) för nationella prov, nationella prov i svenska för invandrare, kurs B och D, och deras konstruktionsprinciper, nationella prov i matematik 2b prov och dess konstruktionsprinciper, samt nationella prov i engelska 6 prov och dess konstruktionsprinciper. Resultatet av analyserna finns under respektive prov. Det finns dock ett flertal av förändringsförslagen som passar alla de undersökta proven.

En viktig förändring för att minska exponeringen av de nationella proven och dagens frekventa användning är att starkt begränsa administrationen av prov inom vuxenutbildningen till ett fåtal tillfällen varje termin oavsett vilken aktör som genomför proven. Det är även lämpligt om det sätts upp tydliga riktlinjer vilka provversioner som får användas vid dessa provtillfällen. Vidare bör man försöka skapa möjlighet att använda ett större antal provversioner, vilket antingen kan göras genom att ta fram fler provversioner och eller genom att se över sekretesskraven. Genom att förlänga sekretesstiden så kan ett större antal provversioner används. Att använda sekretessbelagda prov bör inte göras slentrianmässigt utan som en tydlig strategi.

Ett sätt att ta fram fler provversioner är att möjliggöra att man kan kombinera olika delprov med varandra. Detta förutsätter dock att de är skapade utifrån samma modell vad gäller innehåll, struktur och svårighet. För att skapa fler varianter av delprov skulle vissa delprov kunna skapas utifrån ett antal moduler av uppgifter. Notera dock att inte alla delprov är lämpliga att dela upp i moduler. Men om vi tänker att varje delprov som är möjligt att dela upp består av tre moduler av uppgifter skulle man istället för två delprov kunna kombinera moduler så att man kunde få fram sex delprov som skiljer sig åt med en tredjedel av uppgifterna. Om man vill begränsa exponeringen av de svårare uppgifterna till de elever som har mest kunskap så skulle man kunna använda sig av MST för vissa delprov eftersom de med mindre kunskap aldrig kommer få en svårare modul. Slutligen om man vill ha en större kontroll på hur proven genomförs så kan man fundera på att skapa någon typ av kvalitetsstämpel som ges till de aktörerna som har rätt att ge proven liknande som görs i exempelvis Danmark.

I dagsläget är det viktigt att man genomför förändringar för att minska exponeringen av proven och uppgifterna. De olika förändringsförslagen har olika kostnad och är olika enkla att genomföra och sålunda beror det på hur mycket man vill och har råd att förändra. Den i särklass viktigaste förändringen och troligtvis de minst kostsamma är att begränsa de tillfällen som proven ges till ett fåtal tillfällen varje termin oavsett vilken aktör som genomför proven. Vidare att utöka sekretesstiden och sätta i system att återanvända prov som en tydlig strategi är en annan viktig föreslagen förändring som troligtvis inte kostar så mycket jämför med övriga föreslagna förändringar.

Litteraturförteckning

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement (2nd ed.)*, pp. 508-601. Washington, DC: American Council of Education.

Bunderson, C., Inouye, D. K. & Olsen, J. B. (1989). The Four Generations of Computerized Educational Measurement. I R. Linn. *Educational Measurement, (3rd ed.), pp. 367-407,* Ace-Macmillan, New York.

Crocker, L. & Algina, J,. (1986*). Introduction to classical and modern test theory*. Orlando, FL: Harcourt Brace Jovanovich.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), pp. 297-334.

Crooks, T. J., Kane, M. T., & Cohen, A. S. (1996). Threats to the valid use of assessments. *Assessment in Education: Principles, Policy & Practice, 3*(3), pp. 265-286.

Dorans, N. J., & Holland, P. W. (1993). DIF detection and description: Mantel-Haenzel and standardization. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning*, pp. 35-66. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Drasgow, F., & Olsen-Buchanan, J.B. (Eds.). (1999). *Innovations in Computerized Assessment*. Mahwah NJ: Erlbaum associates.

González, J. & Wiberg, M. (2017). *Applying test equating methods using R*, Springer: Cham.

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied measurement in education, 8*(1), pp. 41-55.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: principles and applications*. Boston: Kluwer-Nijhoff Publishing.

Kaftandjieva, F. (2004). Standard setting. I Council of Europe, *Reference supplement to the preliminary pilot version of the manual for relating language examinations to the Common European Framework of Reference for Languages* (Section B). https://rm.coe.int/1680667a1d

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational measurement (4th ed)*. pp. 17-64. Westport: ACE/Praeger.

Karantonis, A. & Sireci, S.G. (2006). The Bookmark Standard Setting Method: A Literature Review. *Educational Measurement: Issues and Practice 1*, 4-12.

Kolen, M. J., & Brennan, R. L. (2014). *Test equating, scaling and linking: methods and practices. (3rd ed.).* New York: Springer.

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling, 53*(3), pp. 359-379.

Lind Pantzare, A. (2017). Validating Standard Setting: Comparing Judgmental and Statistical Linking. In S. Blömeke & J.-E. Gustafsson (Eds.), *Standard Setting in Education: The Nordic Countries in an International Perspective*, pp. 143-160, Springer.

Liu, J. & Low, A. C. (2008). A comparison of the kernel equating method with traditional equating methods using SAT data. *Journal of Educational Measurement, 45*(4), pp. 309-323.

Lyrén, P-E., & Hambleton, R. K. (2011). Consequences of violated the equating assumptions under the equivalent group design. *International Journal of Testing, 36*, 308-323.

Magis, D., Yan, D., and von Davier, A. (2017). *Computerized adaptive and multistage Testing with R*. Springer.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement (3rd ed.)*, pp. 13-103. New York: American Council on Education & Macmillan.

Ramstedt, K. (1996). *Elektriska flickor och mekaniska pojkar. Om gruppskillnader på prov – en metodutveckling och en studie av skillnader mellan flickor och pojkar på centrala prov i fysik.* Akademisk avhandling. Pedagogiska institutionen. Umeå universitet.

Sands, W.A., Waters, B.K., McBride J.R (Red.). (1997). *Computerized Adaptive Testing: From Inquiry to Operation*. Mahwah NJ: Erlbaum associates.

Sinharay, S., Puhan, G., & Haberman, S. J. (2011). An NCME instructional module on subscores. *Educational Measurement: Issues and Practice, 30*(3), pp. 29-40.

Skolverket. (2017). *Skolverkets systemramverk för nationella prov*. Skolverket, Sverige. Nedladdat från www.skolverket.se i oktober 2018.

Umar, J. (1997). *Item banking*. In J. P. Keeves (Ed.), Educational Research, Methodology, and Measurement: An international handbook, Elsevier Science Ltd, Cambridge.

Vale, C.D. (2004). Computerized item banking. In Downing, S.D., & Haladyna, T.M. (Eds.) *The Handbook of Test Development*. Routledge.

van der Linden, W. J. & Glas, C. A. W. (Eds.) (2000). *Computerized Adaptive Testing: Theory and Practice*, Kluwer Academic Publishers, Dordrecht.

Wald, A. (1966). *Sequential analysis*. John Wiley & Sons: New York.

Wainer, H. (2000). *CATs: Whither and Whence.* Educational Testing. Service research report. Princeton, New Jersey.

Weiss, D. J. (Ed.). (2014). *New Horizon Testing: Latent Trait Test Theory and Computerized Adaptive Testing*. Elsevier.

Wiberg, M. (2002). *Uppgiftsbank för körkortsprovets teoretiska prov. Relationen mellan utformning, exponering och provtypen. PM Nr 173.* Enheten för pedagogiska mätningar, Umeå Universitet.

Wiberg, M. (2003). *Computerized Achievement Tests – sequential and fixed length tests.* Statistical Studies No. 29, Statistiska institutionen, Umeå universitet.

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied psychological measurement, 39*(5), pp. 349-361.

Wiberg, M. & von Davier, A. A. (2017). Examining the impact of covariates on anchor tests to ascertain quality over time in a college admissions test. *International Journal of Testing. 17*, pp. 105-126.

Yan, D., von Davier, A., and Lewis, C (2014). *Computerized multistage testing: theory and applications.* Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences.