



SU:s svar på frågor från UHR inför slutrapport MBP

Texten i blått nedan är de frågor som Stockholms universitet (SU) har fått från Universitets- och högskolerådet (UHR) att besvara inför UHR:s slutrapport om medborgarskapsprovet till regeringen den 30 september 2026. SU:s svar och resonemang följer i svart text efter respektive del/frågebatteri.

Frågor från UHR:

Här följer frågor för klargöranden kring avvägningar, förutsättningar, tidplan och kostnader för att utforma provet i svenska språket i dess helhet och delar på kort och lång sikt. Frågorna ställs dels för att uppfylla rapporteringskravet som ställs i regeringsuppdragen till UHR och Stockholms universitet, dels för att få underlag till strategiska beslut som behöver fattas för att genomföra medborgarskapsprovsuppdragets olika steg. Flera av dessa strategiska beslut behöver fattas innan rapporteringen till regeringen den 30 september därför ber vi om återrapporing i två steg, maj och augusti.

Steg 2 – Läs- och hörförståelse, steg 2

(Enl. ändrat regeringsuppdrag: "Det andra steget ska omfatta prov i läs- och hörförståelse i svenska och ska genomföras senast den 1 oktober 2027." "Redovisningen ska innehålla en uppdaterad tidplan och kostnadsberäkning".)

Förutsättningar att uppfylla uppdraget att leverera ett prov i läs- och hörförståelse senast den 1 oktober 2027:

- Vad kan ni leverera till den 1 oktober 2027?
Utveckla genom att resonera kring begreppen i uppdragsbeskrivningen:
 - Ändamålsenlighet
 - Träffsäkert
 - Rättssäkert
- Redogör för processen för framtagande av provdelarna till färdigt prov, steg för steg
- Resonera om det finns något som kan skynda på framtagandet och vilka eventuella kostnadsmässiga konsekvenser det skulle få.
- I vilken mån påverkar omfattningen av provdelarna tiden som det tar att ta fram dem?
 - Går det snabbare att ta fram prov som genomförs på kortare tid?

Svar från SU:

Förutsättningar att uppfylla uppdraget att leverera ett prov i läs- och hörförståelse senast den 1 oktober 2027

I det följande redogör vi för förutsättningarna att leverera ett prov i läs- och hörförståelse i svenska språket som är ändamålsenligt, träffsäkert och rättssäkert senast den 1 oktober 2027.

Det vi kan leverera till den 1 oktober 2027 är en konstruktdefinition, utkast till testspecifikationer för hör- och läsförståelse, statistiska data från utprovning av olika typer av uppgifter från ett första utprovningstillfälle samt en första analys av provdelarnas validitet, reliabilitet och rättssäkerhet. Till detta datum kommer vi dock inte att kunna leverera ett färdigt prov för användning i skarpt läge som är ändamålsenligt, träffsäkert och rättssäkert utifrån uppdragsbeskrivningen. Vi bedömer att detta kan levereras tidigast 1 oktober 2028.

Validitet, reliabilitet och rättssäkerhet

Begreppen ändamålsenlighet och träffsäkerhet tangerar de i testsammanhang etablerade begreppen *validitet* och *reliabilitet*, och vi kommer även att använda de senare begreppen i vår redogörelse. Validitet handlar om i vilken utsträckning de tolkningar, beslut och konsekvenser som följer av ett provresultat är motiverade och ändamålsenliga. Reliabilitet handlar om hur tillförlitliga och konsekventa provresultaten är och är en nödvändig, men inte tillräcklig, förutsättning för validitet. Validitet är i sin tur en förutsättning för testtagarnas *rättssäkerhet*, det vill säga att individen skyddas mot godtycke från stat och myndigheter.

Validitetsforskaren Michael Kanen (2013) argumentbaserade valideringsmodell har de senaste decennierna fått ett betydande genomslag inom språktestning och vår grupp har därför valt att utgå från denna modell i valideringsarbetet med det aktuella

språkprovet. Enligt den argumentbaserade valideringsmodellen beror validiteten i en viss bedömning på styrkan i de argument som anförs för de slutsatser som dras av ett provresultat. Ett hållbart validitetsargument består av en sammanhängande kedja av inferenser från provinterna aspekter (poängsättning på provprestationer; provpoängens generaliserbarhet till andra liknande prov; förklaring av provpoängens innebörd) till provexterna aspekter (extrapolering av provresultat till språkdomäner utanför själva provsituationen; samt de konsekvenser och beslut som följer av ett provresultat).

I vår validering av om testet är tillräckligt ändamålsenligt och träffsäkert behöver provgruppen alltså kunna argumentera för att provprestationerna på hör- och läsförståelsetesten poängsätts och bedöms tillförlitligt och konsekvent (jfr ovan om generaliserbarhet). Detta innebär bland annat att säkerställa att resultat som provet genererar är konsistenta, det vill säga att en testtagare kan förväntas få samma resultat oavsett provversion, plats eller tid. Dessutom behöver frågan om extrapolering beaktas, vilket handlar om i vilken utsträckning provresultatet verkligen säger något om testtagarens faktiska språkkunskap utanför provsituationen.

Därutöver behöver validitetsargumentet visa att konsekvenserna av provresultaten är önskvärda och att de beslut som fattas utifrån en provpoäng, är rättvisa och rimliga. Idag märks ett uttalat fokus på etiska frågor om rättvisa i samband med testanvändning, och provgruppen avser att inkludera ett sådant perspektiv i hela valideringskedjan. Det etiska perspektivet på språktest passar väl in i Kanes valideringsmodell som beskrivits ovan. Exempelvis skiljer McNamara m.fl. (2019) mellan begreppen *fairness* och *justice*, där *fairness* syftar på rättvisa när det kommer till test-interna aspekter, såsom bedömning, poängsättning, kravgränser osv, medan *justice* handlar om test-extern rättvisa, till exempel huruvida testresultaten riskerar att leda till diskriminering av vissa grupper. Ett rättssäkert språkprov förutsätter att vi kan argumentera för validiteten i tolkningarna av provresultaten, med hänsyn till såväl test-intern som test-extern rättvisa, för att garantera att språkprovet är ändamålsenligt och rättssäkert.

En noggrann validering av provet förutsätter också att syftet med språkravet för medborgarskapet är tydligt definierat. I regeringens proposition *Skärpta krav för svenskt medborgarskap* (Regeringen, 2026, s. 81) anges flera olika skäl för införandet av ett språkrav för medborgarskap i Sverige, såsom att kunskaper i svenska språket är nödvändiga för att ”fullt ut kunna delta i samhället” och ger bättre förutsättningar att ”få ett arbete och ingå i samhällsgemenskapen”, ”förstå sina skyldigheter och ta tillvara sina rättigheter och möjligheter i samhället”. Vidare anges att språkraven syftar till att stärka ”den enskildes incitament att skaffa sig kunskaper i svenska”, vilket i sin tur antas leda till ”en förbättrad integration” och ”minskat utanförskap”. Andra skäl som anges för införandet av ett språkrav är att det har ”ett symboliskt värde” som ”kan förstärka medborgares känsla av delaktighet” och ”stärka medborgarskapets status”.

Att leverera ett ändamålsenligt prov i läs- och hörförståelse i svenska innebär således att vi behöver kunna argumentera för att provet har de kvalitéer som leder till ovanstående önskade konsekvenser. Detta innebär också att provet inte medför några icke avsedda negativa konsekvenser (se även Bachman & Palmer, 2010, s.177–192). Innan språkprovet ges i skarpt läge behöver man därför exempelvis kunna argumentera för att ett godkänt testresultat verkligen visar att testtagares språkkunskaper i svenska är tillräckliga för att de ska kunna delta i samhället och förstå sina rättigheter och skyldigheter, liksom att testet kan förväntas leda till ökade incitament att lära sig svenska språket och till en förbättrad integration. Exempel på icke avsedda konsekvenser kan vara att provet gynnar vissa grupper men missgynnar andra på sätt som inte var avsedda (se Bugge & Carlsen, 2025).

Ovanstående resonemang om provets önskvärda och icke-önskvärda konsekvenser visar att validering inte är en process som tar slut när provet är färdigkonstruerat och ges i skarpt läge, utan att det kräver ett kontinuerligt arbete, som det också måste finnas utrymme för så länge som provet är i bruk. Det handlar dels om att säkerställa att varje ny testversion möter kvalitetskraven, dels att kontinuerligt utvärdera vilka konsekvenser för såväl testtagare som samhället i stort som språkprovet får.

Provkonstruktionsprocessen

Provkonstruktion är en dynamisk process som i mångt och mycket vilar på kvalitativa avgöranden i framtagandet av provmaterial. Processen fram till en första version av hör- och läsförståelseproven som kan ges i skarpt läge innefattar flera steg såsom behovsanalys, urval av GERS-deskriptorer, konstruktdefinition, operationalisering, testspecifikation, konstruktion av uppgifter (inklusive stimuli), granskning, utprovningar, statistisk och kvalitativ uppföljning samt kravgränssättning. En redogörelse för dessa steg i processen finns i tabell 1, och i det följande resonerar vi om de olika stegens betydelse för att vi ska kunna argumentera för validiteten i tolkningarna av provresultaten. I planeringen av konstruktions- och valideringsprocessen förhåller vi oss till rekommendationer för utveckling av GERS-baserade språktest framtagna av ALTE (Association of Language Testers in Europe) i samverkan med Europarådet (ALTE, 2011). Arbetsprocessen ligger även i linje med den som sedan länge gäller för andra typer av storskaliga prov nationellt och internationellt, t.ex. sfi-proven, Tisus, PIACC, PISA och PIRLS. Vi vill även understryka att det är en optimistisk tidsplan, som förutsätter att vi inte stöter på problem med sådant som delvis ligger utanför vår kontroll, exempelvis rekrytering av utprovningssdeltagare eller tillhandahållande av en digital plattform. Eventuella problem som vi i nuläget inte kan förutse riskerar att medföra att tidsplanen förlängs.

Tabell 1 Processen för framtagande av provdelarna för färdiga delprov i hör- och läsförståelse

2026	
Januari – Augusti	Rekrytering av konstruktionsgrupp och referensgrupp Analys av uppdraget och inläsning Internationell utblick (medborgarskapsprov i andra länder) Utbildning av provkonstruktörer för det specifika syftet
September	Behovsanalys och urval av relevanta GERS-deskriptorer Konstrukt/Operationalisering Planering för rekrytering av utprovningssgrupper Lära känna digital plattform (möjligheter, begränsningar av olika uppgiftsformat)
Oktober – November	Arbete med digital plattform – möjligheter och begränsningar av olika uppgiftsformat Prototypuppgifter och utkast till testspecifikationer Rekrytering av utprovningssdeltagare

December		Granskning av prototypuppgifter och utkast till testspecifikationer (Referensgrupp) Revidering av testspecifikationer
2027		
Januari – februari	–	Konstruktion Läsförståelse (motsvarande 2 prov, dvs. 2x30–50 item) Konstruktion Hörförståelse (motsvarande 2 prov, dvs. 2x30–50 item) Skapa enkäter till testtagare, lärare, provledare
Mars		Granskning av provuppgifter (Referensgrupp) Revidering av provuppgifter
April – maj		Utprovning 1 Enkäter till testtagare, lärare, bedömare Bedömning av utprovning 1
Juni – augusti		Statistisk analys och uppföljning av utprovning 1 Kvalitativ uppföljning av utprovning 1
September – oktober	–	Testspecifikation Läsförståelse, version 1 Testspecifikation Hörförståelse, version 1
November – december	–	Konstruktion Läsförståelse (150 item) Konstruktion Hörförståelse (150 item)
2028		
Januari – Februari	–	Konstruktion Läsförståelse (150 item) Konstruktion Hörförståelse (150 item)
Mars		Itemgranskning (Referensgrupp) Revidering av item
April – maj		Utprovning 2, helst ca 3000 deltagare Rekrytering och utbildning av panel till kravgränssättning Bedömning av utprovning 2
Juni		Statistisk analys och uppföljning av utprovning 2 De 50 mest välfungerande itemen per provdel går till slutversionen 30 välfungerande item används till exempelprov Testspecifikation slutversion
Juli		Exempelprov och instruktioner för genomförande läggs ut på webben
Augusti		Slutversion Läsförståelse, 50 item (20 item stryks efter resultaten analyserats) Slutversion Hörförståelse, 50 item (20 item stryks efter resultaten analyserats)

September	Kravgränssättning Efter genomförande: Statistisk analys De 20 minst fungerande itemen per provdel stryks efter resultaten analyserats
-----------	--

Beskrivning av arbetsstegen

En **behovsanalys** innebär en analys av språkprovets målgrupp och måldomän i förhållande till dess syfte. Ett språktest i medborgarskap skiljer sig från ett språkprov för högskolebehörighet som till exempel Tisus, som har en specifik målgrupp och måldomän (blivande studenter på universitet och högskola som ska uppfylla behörighetskraven i svenska). Måldomänen för språkprovet för medborgarskap är bredare och vagare i konturerna, och uppgiften för konstruktionsgruppen är i detta steg att definiera vilka språkliga krav som ställs för att delta fullt ut i samhället, ingå i samhällsgemenskapen, förstå sina rättigheter och skyldigheter och så vidare. Eftersom provet kommer att genomföras av personer mellan 16 och 67 år, och testtagarnas bakgrund gällande utbildnings- och yrkesbakgrund, tidigare språkkunskaper med mera kan variera kraftigt måste den tilltänkta målgruppen antas vara bred, vilket behöver beaktas under provkonstruktionsprocessen. Enligt utredningen (SOU 2021:2) kommer det krävas anpassningar för personer med funktionsnedsättning och andra med dispens, vilket också behöver utredas.

Av uppdraget framgår att provet i läs- och hörförståelse ska länkas till nivån B1 enligt GERS-skalan. Det innebär att vi behöver se till att de krav som måldomänen ställer på målgruppen ligger i linje med de nivåer för receptiva färdigheter som anges för nivå B1 i GERS. I regeringens uppdrag specificeras inte vidare vilka tablåer i GERS som ska användas. Vi utgår från den senaste versionen (Europarådet, 2020), där det finns ett stort antal tablåer som beskriver receptiva språkliga färdigheter vid olika kommunikativa aktiviteter och situationer, och ett relevant urval måste göras som fångar syftet med språkkravet så som det uttrycks för det svenska medborgarskapet ovan. Behovsanalysen och urvalet av GERS-deskriptorer ska sedan ligga till grund för definitionen av provets **konstrukt**, det vill säga en teoretisk beskrivning av de

färdigheter som språkprovet avser att testa. En tydlig konstruktdefinition är en nödvändig utgångspunkt för konstruktions- och valideringsprocessen för att kunna utvärdera i vilken utsträckning provet mäter det som avses och för att kunna skapa nya parallella test i framtiden.

När konstruktet är formulerat ska det **operationaliseras** i form av konkreta uppgifter i hör- och läsförståelse. Exakt hur dessa ska utformas och bedömas är en direkt funktion av hur konstruktet har formulerats och våra förslag på tidsåtgång för provets genomförande, antal item, frågeformat med mera ska således ses som preliminära.

Ett led i att säkerställa provets validitet/ändamålsenlighet, och i förlängningen dess rättssäkerhet, är att formulera en **testspecifikation**, det vill säga ett slags recept som i detalj beskriver vad varje version av provet ska innehålla i form av exempelvis teman, typer av instruktioner, stimuli, uppgiftsformat och bedömningssätt. En tydlig testspecifikation är avgörande för arbetet med att göra varje provversion likvärdig avseende svårighetsgrad och vad som testas. I konstruktionsprocessens inledande skede behöver därför testspecifikationer för respektive delprov formuleras parallellt med att prototypuppgifter tas fram. Testspecifikationerna kommer även att behöva revideras efter utprovningar av provuppgifterna.

Konstruktionsarbetet innefattar utöver konstruktion av item (frågor till testtagarna) även skapande av skriftliga och muntliga stimuli (texter och ljudinspelningar) i syfte att kunna testa färdigheter i läs- och hörförståelse. Alternativet att använda autentiska texter, radioinslag o.s.v. är sannolikt inte aktuellt, då dessa behöver vara anpassade efter B1-nivån vilket autentiska texter sällan är. Vi behöver därför ta höjd för detta relativt tidskrävande arbete i tidsplanen.

Såväl testspecifikation som utkast till uppgifter behöver granskas av en referensgrupp bestående av externa personer med relevanta erfarenheter och kompetenser. Hit hör expertis inom språkvetenskap, språktestning, bedömning, undervisning, med erfarenheter av GERS-skalan och särskilt B1-nivån. Detta är en central del i valideringen av provet för att säkerställa provets ändamålsenlighet och träffsäkerhet.

En fråga som också behöver utredas i ett tidigt skede är hur **bedömningen** av uppgifterna i hör- och läsförståelse ska organiseras. Om provet ges digitalt och består av flervalsfrågor finns möjligheten till automatisk rättning under förutsättning att vi har ett system som tillåter det. Om öppna frågor används kommer dock sannolikt även en manuell bedömning behöva göras vilket kräver rekrytering och utbildning av bedömare.

När ett första utkast till uppgifter i hör- och läsförståelse finns på plats behöver vi genomföra en **utprovning** (pilotstudie) på en representativ målgrupp, i åtminstone två omgångar. Till den första utprovningen ska med fördel uppgifter motsvarande (minst) två provversioner per provdel vara konstruerade. Detta ger möjlighet att testa huruvida samma deltagare i pilotstudien får liknande resultat på två olika prov som är konstruerade utifrån samma testspecifikation, vilket ger viktiga indikationer om provets reliabilitet. Fler item än vad som kommer att behövas när provet går i skarpt läge är också en fördel då fungerande item kan återanvändas i kommande versioner medan icke-fungerande item kan strykas. I samband med utprovningen behöver även en enkätundersökning om deltagares upplevelse av provet genomföras, liksom uppföljande intervjuer med deltagarnas lärare i svenska och provledare.

Eftersom den tilltänkta målgruppen är bred kommer rekryteringen av ett representativt urval av utprovningssdeltagare att vara utmanande. Målsättningen är att utprovningsgruppen ska vara bred när det gäller modersmål, tidigare litteracitet, skolbakgrund, utbildningsnivå, ålder och kön. Det är mycket önskvärt att utprovningsgruppen inkluderar deltagare med kort skolbakgrund. Detta kommer sannolikt att bli utmanande för projektet, då erfarenheter från utprovningar av de nationella proven för sfi visar att dessa grupper kan vara svåra att nå på grund av att de i mindre utsträckning deltar i utbildningar i svenska. Ett komplement kan vara att man i likhet med GU:s prov i samhällskunskap genomför utprovningen på personer som är i färd med att ansöka om medborgarskap. Resultat från en utprovningssversion

av provet kan dock inte ligga till grund för beslut om medborgarskap då uppgifterna i ett utprovningsskede inte är kvalitetssäkrade.

En stor utmaning vid utprovningen kommer vara att bibehålla sekretessen, då det sannolikt kommer finnas starka incitament för utomstående att ta del av och sprida information om provuppgifterna.

Efter utprovningen görs en **statistisk analys** av data som samlats in för att utvärdera provuppgifternas reliabilitet. Exempelvis kan en Rasch-modell (se t.ex. Fan m.fl. 2019) användas för att mäta svårighetsgraden hos enskilda item i relation till den aktuella testpopulationens språkliga nivå, samt huruvida varje item mäter samma underliggande språkförmåga. För att testa huruvida provuppgifterna är rättvisa i förhållande till olika grupper av deltagare kan en DIF-analys (Differential Item Functioning, se t.ex. McNamara & Roever, 2006, s. 81–128) göras, för att undersöka om två personer med lika hög språknivå har liknande sannolikhet att svara rätt, oberoende av gruppstillhörighet. Bakgrundsvariabler för olika grupper som kan vara relevanta i DIF-analysen kan vara kön, utbildningsbakgrund, ålder, modersmål och tid i Sverige. För att en sådan utvärdering ska vara tillförlitlig behövs en utprovningsgrupp som är tillräckligt stor för att kunna brytas ner i undergrupper baserade på sådana bakgrundsvariabler. Därför bedömer vi att en utprovning behöver göras på ca 3000 personer. Om det blir för utmanande skulle deltagarantalet vid utprovning 1 kunna vara mindre, medan utprovning 2 görs på en större grupp.

De statistiska analyserna behöver också följas upp av en kvalitativ utvärdering där konstruktionsgruppen analyserar orsakerna till utfallet och drar eventuella slutsatser om hur testspecifikationen behöver revideras. Efter utprovningen behöver vi även ta ställning till frågan om vilka anpassningar för personer med funktionsnedsättning som är lämpliga.

Innan vi har ett prov med tillräckligt hög grad av validitet, reliabilitet och rättssäkerhet för att kunna användas i skarpt läge behöver processen med konstruktion, granskning,

utprovning och analys upprepas. Till utprovningstillfälle 2 kan eventuellt vissa fungerande item från utprovning 1 återanvändas. De kommer då få en viktig funktion som så kallade **ankaritem**, som i den statistiska analysen möjliggör att utvärdera den andra utprovningens svårighetsgrad jämfört med den första. Det kommer dock vara nödvändigt att även konstruera nya item baserat på de slutsatser som dragits efter utprovning 1 och eventuella revideringar av testspecifikationerna. De item som fungerat bäst i utprovning 2 får sedan ingå i slutversionen.

Eftersom detta är ett nytt prov med stora konsekvenser för testtagarna är det nödvändigt för rättssäkerheten att tillhandahålla ett representativt **exempelprov** (som läggs på UHR:s webbsida) för blivande testtagare innan provet ges i skarpt läge. Därför behöver vi till utprovningstillfälle 2 konstruera minst 150 item per delprov, för att det efter strykningar av mindre fungerande item ska finnas ca 30–50 väl fungerande item till slutversionen och ytterligare 30–50 autentiska item till exempelprovet.

Före provets genomförande behöver även en **kravgränssättning** för hör- och läsförståelseprovet genomföras, i syfte att fastställa reliabla poänggränser för godkänt i respektive delprov. För detta behövs en panel bestående av externa experter, främst med kunskap om den relevanta nivån. Eftersom kravgränssättning enligt GERS-skalan är nytt för många i Sverige bör vi även ta höjd för att det behövs utbildning i det. Den metod som används i Tisus och som även kan tillämpas i detta fall är Bookmark-metoden (se Cizek & Bunch, 2007, s.155–191). Med utgångspunkt i de statistiska analyserna av frågornas svårighetsgrad görs en kvalitativ bedömning av vilka item som en svensk medborgare med minimalt acceptabla kunskaper i svenska språket motsvarande B1-nivån förväntas klara och därmed var gränsen för godkänt ska ligga. Kravgränssättning är således ett viktigt och tidskrävande steg i valideringsprocessen.

Övriga förutsättningar

Värt att poängtera är att tidsplanen i tabell 1 förutsätter att konstruktionsgruppen från och med hösten 2026 har tillgång till den digitala plattform som är tänkt att användas till provet, då den behöver finnas på plats under såväl konstruktionsarbetet som

utprovningen. Om UHR:s arbete med att tillhandahålla den digitala plattformen drar ut på tiden kommer även vår tidsplan att behöva revideras. Att i stället för ett digitalt prov ta fram ett pappersprov ser vi inte som ett alternativ, eftersom slutmålet enligt SOU 2021:2 är ett digitalt prov. Att först utveckla ett analogt prov, för att sedan börja om och skapa ett digitalt prov vore ineffektivt (se resonemang på s. 15).

Planen förutsätter också att utprovningar och analyser ger svar som vi kan arbeta vidare med. Vi måste reservera oss för att utvärderingarna av utprovningarna kan visa att mer omfattande revideringar och eventuellt ytterligare utprovningar behöver göras, och att tiden därför måste justeras.

Med tanke på det omfattande arbete det kräver för att ta fram ett prov som möter de höga kvalitetskrav som ett high stakes-test som detta ställer är det inte möjligt att skynda på framtagandet mer än vi har gjort. Tidsplanen är redan i dagsläget optimistisk, dels med tanke på att rekryteringen av utprovningsdeltagare kan dra ut på tiden, dels för att konstruktionsgruppen behöver utökas ytterligare om tidsplanen ska vara hållbar. Omfattningen av provdelarna och tidsinramningen i provet har endast marginell påverkan på framtagningsprocessen. De stora flaskhalsarna gäller snarare granskning, utprovning, analys och kravgränssättning, och att kompromissa med dessa är inte ett alternativ om vi ska kunna säkerställa provets validitet, reliabilitet och rättssäkerhet.

Något som kommer att påverka arbetet med alla delproven är också hur provet kommer att hanteras av UHR gällande sekretess och eventuell frisläppning av uppgifter. Om genomförda prov offentliggörs/frisläpps och/eller om sekretessen röjs kan detta förlänga tidsplanen och konstruktionsarbetet eftersom fler helt nya frågor då behöver provas ut.

Frågor från UHR:

Fördjupad beskrivning av utformning av prov i läs- och hörförståelse vid den tidpunkt det kan genomföras (oberoende er bedömning av vad som är möjligt att leverera till den 1/10-27).

Här ber vi er resonera kring hur ni tänker att provdelarna bör utformas när de ska genomföras i skarpt läge.

Omfattning

- Er uppfattning om optimal omfattning i tid för provtagaren att genomföra respektive provdel (läs- resp hörförståelse) för att kunna bedöma den språkliga nivån som stipuleras i uppdraget.

Framtagande och utformning: Digitalt/analogt prov resp. Fast/frågebank

Om vi utgår ifrån att det i huvudsak finns tre möjliga provtyper:

1. Analogt prov som genomförs på papper. Ett prov tas fram till varje provtillfälle. Alla provtagare vid samma provtillfälle får identiska prov.
 2. Fast digitalt prov som genomförs på dator. Ett prov tas fram till varje provtillfälle. Alla provtagare vid samma provtillfälle får identiska prov.
 3. Digitalt prov med frågor som "slumpas" fram ur en frågebank. Provtagarna får individuellt utformat prov.
- Redogör för speciella förutsättningar för framtagande av digitalt (fast) prov respektive analogt, går någon av dessa snabbare att ta fram än det andra?
 - Resonera kring för-, nackdelar, möjligheter och utmaningar med att inleda med analogt prov och att senare övergå till digitalt (fast) prov.
 - Hur många provtillfällen bedömer ni att ni kan ta fram underlag till per år från det första provtillfället och 5 år framåt?
 - Från och med när är det realistiskt att införa ett digitalt prov där frågor "slumpas" ur en frågebank?
 - Hur många provtillfällen behöver genomföras för att det ska finnas tillräckligt många provfrågor för att göra digitalt prov där frågor tas ut en frågebank.

Tidplan

Gör två scenarion; det ena för digitalt (fast) prov från start, det andra med ett inledande analogt prov som övergår till digitalt

- För framtagande av prov till det första provtillfället och 5 år fram
 - Resonera om en tidplan för en övergång från fasta prov till digitalt prov där frågorna "slumpas" ur en frågebank i de bägge scenariona.
-

Bedömning

- För att kunna uppskatta antalet provtagare, inklusive de som behöver göra om provdelar: Är det möjligt att uppskatta hur många man kan förvänta sig klarar de olika provdelarna? (I USSMA:n gjorde man uppskattningen att andelen personer som behöver göra om provet är 20 procent.)

Kostnadsberäkningar (behövs först till slutrapporteringen i augusti)

- Framtagande
- Genomförande per år
- Helårsarbeten – initialt framtagande och drift över tid

Svar från SU:

Omfattning

När det gäller en mer exakt beskrivning av hur provet kommer att vara uppbyggt i skarpt läge, exempelvis antal texter, antal uppgifter per delprov samt den exakta fördelningen mellan olika uppgiftstyper, är detta i nuläget inte möjligt att fastställa i detalj. Skälet är att dessa beslut behöver baseras på resultat från utprovningar, itemanalys och kalibrering av uppgifter (ALTE, 2011). Detta gäller även tidsåtgången, varför vi inte i nuläget kan uttala oss om en optimal tidsåtgång för genomförandet. Vid en första utprovning är det brukligt att inte ha någon tidsbegränsning alls för att kunna undersöka hur mycket tid testtagarna använder till de olika uppgifterna. En ytterst preliminär uppskattning är att det krävs åtminstone 45–60 minuter per delprov.

Vid framtagande av provet är en viktig princip att det ska innehålla en bred representation av ämnen, texttyper och uppgiftstyper. En sådan variation bedöms öka möjligheten att provet blir rättvist och speglar den språkliga bredd som definieras för B1-nivån (ALTE, 2011). I nuläget utgår vi ifrån att provet kommer att bestå av flervalsfrågor eller ja-/nej-frågor kopplade till cirka tre till fem olika texttyper respektive inspelningar av varierande längd. Texterna och ljudinspelningarna ska spegla olika typer av autentiskt och vardagsnära språkbruk på B1-nivå. För läsförståelse innebär denna breda representation att provet behöver pröva exempelvis förmågan att förstå huvudbudskap, identifiera specifik information, tolka enklare

resonemang och förstå autentiska texter av relevans för vardags-, samhälls- och arbetsliv på B1-nivå (se GERS, Europarådet, 2020). För hörförståelse behöver provet på motsvarande sätt mäta förmågan att förstå vanliga kommunikativa situationer, såsom instruktioner, samtal och informationsmeddelanden (Europarådet, 2020).

Framtagande och utformning: Digitalt/analogt prov resp. Fast/frågebank

Vad gäller provformat framstår ett analogt prov på papper inte som ett realistiskt alternativ och är något vi avstyrker eftersom proven ska genomföras digitalt (SOU 2021:2). Övergången från ett provformat till ett annat riskerar att påverka provets reliabilitet, vilket i sin tur äventyrar validitet och rättssäkerhet. Studier av hur övergången från analoga till digitala läsförståelseprov påverkar testtagares prestationer har gett blandade resultat, där vissa visar svag eller ingen effekt (Buerger m.fl. 2019; Grammatikopoulou m.fl. 2026), medan andra visar att läsförståelse påverkas negativt vid läsning på skärm (Mangen, 2013; Delgado, 2019). Vi kan således inte ta för givet att en övergång från analogt till digitalt prov inte får effekt på testtagarnas prestationer.

En övergång från ett provformat till ett annat innebär ökade kostnader och längre tidsfrister, särskilt om målet ändå är ett digitalt format. Ett analogt prov skulle behöva följa samma steg i processen som beskrivs i tabell 1 och innebär ingen tidsvinst. Om man börjar med ett analogt prov för att sedan lägga om till ett digitalt prov behöver processen startas om med konstruktdefinition, testspecifikationer, provuppgifter för digitalt format, utprovningar och analyser under tiden som man samtidigt fortsätter att ta fram analoga prov parallellt för att säkerställa tillgången på prov. Under ett antal år skulle det då behövas fler konstruktörer eftersom två versioner av alla delproven behöver tas fram, dvs. både analoga och digitala versioner.

Att genomföra ett hörförståelseprov på papper i sal innebär vidare en teknisk utmaning vad gäller ljudutrustning. Det skulle kräva ett betydande arbete att säkerställa att alla provutförare i landet håller likvärdig ljudkvalitet samt att varje testtagare (inklusive den som har någon form av hörsel- eller funktionsnedsättning) uppfattar det som sägs lika väl; ett arbete som sedan, vid övergången till digitala prov (där ljudet kan spelas i

hörlurar anpassat efter varje enskild testtagare), går om intet. Ytterligare tidskrävande är att rätta och bedöma prov på papper, medan det digitala formatet kan utformas för att ge poängsättning direkt. Ett analogt prov skulle med andra ord bromsa utvecklingen av provet i sin helhet. Sammantaget talar detta för att det är mer effektivt, både ekonomiskt och kvalitetsmässigt, att utveckla ett digitalt prov från start.

Det alternativ som enligt vår bedömning förefaller bäst lämpat för ändamålet är således ett digitalt prov, där alla provtagare får identiska uppgifter vid samma tillfälle.

Varje provtillfälle bör även utformas så att de genererar data för framtida prov och på sikt eventuellt möjliggör uppbyggnaden av en prov- eller frågebank. Utprovningen integreras alltså i ordinarie provtillfällen, antingen i respektive provdel eller som en egen del. Målgruppen för denna utprovningssdel kommer då att vara de som önskar medborgarskap och provet genomförs i en autentisk provsituation, något som ytterligare kan stärka validiteten hos kommande prov.

Ett digitalt prov där uppgifter slumpas fram ur en frågebank ser vi inte som ett alternativ från start, eftersom det kräver konstruktion av ett mycket stort antal uppgifter. Vi är i ett för tidigt skede för att ta ställning till om en ett digitalt prov med slumpade frågor är realistiskt på sikt. Något man behöver beakta är att ett sådant format kan äventyra likvärdigheten mellan olika provversioner. En uppgift i läs- och hörförståelse består i regel av en text eller ljudinspelning med flera tillhörande item vars svårighetsgrad kan vara känslig för vilken position i provet de har och i vilken ordning de kommer, varför ett format där frågor slumpas fram kan vara särskilt utmanande.

Om vi trots dessa svårigheter skulle välja att röra oss mot ett format med slumpade uppgifter behöver vi successivt utveckla en text- och frågebank där texter och uppgifter kategoriseras utifrån bland annat nivå, texttyp, ämnesområde och prövad delförmåga i syfte att över tid möjliggöra konstruktion av flera olika provkombinationer (ALTE, 2011). För att detta ska vara möjligt krävs en omfattande mängd kvalitetssäkrade uppgifter, vilket i sin tur förutsätter flera omgångars utprovningar/provgenomförande

och statistiska analyser. Hur snabbt en sådan frågebank kan byggas upp beror i hög grad på antalet, och kategorier av, provtagare vid varje tillfälle samt antalet tillfällen provet ges. Med ett stort antal provtagare som representerar olika typgrupper kan tillräckliga data samlas in relativt snabbt, medan ett mindre antal provtagare innebär att processen tar längre tid. Att redan från start genomföra provet digitalt underlättar denna datainsamling och utprovningensarbete.

Tidsplan

När det gäller produktionstakt bedöms utvecklingen gå relativt långsamt till en början. År 2028 skulle ett första provtillfälle i hör- och läsförståelse kunna ges, parallellt med ytterligare utprovningar inom dessa provversioner. År 2029 kan detta kompletteras med muntliga och skriftliga delar i utprovningssyfte. År 2030 kan ett mer komplett prov (hör-, läs-, skriv- och muntliga delar) genomföras som underlag för en helhetsbedömning av medborgarskapskraven, fortfarande i kombination med utprovning. Under perioden 2031–2032 kan provtillfällena öka till ett eller två prov per år, och från 2033 kan två eller fler provtillfällen per år bli möjliga, eventuellt i kombination med en etablerad frågebank. Även om utvecklingen inledningsvis är resurskrävande och tidsintensiv, finns potential för ökad effektivitet när digitala system, rutiner och datamängder väl är på plats.

2028 – 1 prov (hör och läs) + utprovning extra hör och läs

2029 – 1 prov (hör och läs) + utprovning muntligt och skriftligt + utprovning hör och läs

2030 – 1 prov (hör, muntl, läs, skriv) + utprovning alla färdigheter

2031 – 1-2 prov? (hör, muntl, läs, skriv) + utprovning alla färdigheter

2032 – 1-2 prov? (hör, muntl, läs, skriv) + utprovning alla färdigheter

2033 – 2 eller fler prov? Ev. frågebank

Ett förbehåll är dock att den information som utprovningarna ger oss måste vägleda arbetet, där provets validitet måste styra vår verksamhet. Ovanstående tidsplan är alltså endast en uppskattning och gäller endast under förutsättningen att utprovningen ger önskvärda och tillräckliga data (se tidigare resonemang kring tidsplanen).

Bedömning

Det är naturligtvis oerhört svårt att sja om hur många provtagare som kommer att få godkänt resultat vid ett givet provtillfälle när provet ännu inte är utvecklat. I USSMA (SOU 2021:2, s. 372) uppskattades att andel underkända skulle kunna vara i storleksordningen 20%, baserat bland annat på uppgifter om hur många personer som gör omprov i det norska språkprovet. Vi kan i nuläget varken ge stöd åt eller ifrågasätta denna bedömning.

Frågor från UHR:

Steg 3 – skriftlig och muntlig färdighet:

Enligt regeringsuppdraget: "... redovisa en analys av förutsättningarna för genomförandet av det tredje steget i ett medborgarskapsprov och hur prov i skriftlig och muntlig färdighet ska utformas, genomföras och bedömas. Redovisningen ska innehålla en uppdaterad tidplan och kostnadsberäkning..." Det tredje steget ska "genomföras så skyndsamt, ändamålsenligt och kostnadseffektivt som möjligt."

Skriftlig färdighet

– förutsättningar för utformning, genomförande och bedömning

Provutformning

- Beskriv hur provdelen bör utformas. Resonera tex utifrån:
 - Frågeformat och uppgiftstyper
 - Omfattning text (antal tecken eller ord) och tid för provtagare att genomföra provdelen
 - Hur provutformningen påverkas om provet skrivs på papper respektive skrivs digitalt
 - Utprovning av provuppgifterna

Genomförande

- Utifrån er erfarenhet: Är det några praktiska förutsättningar för provgenomförandet – digitalt resp. analogt – som ni vill delge UHR, exempelvis gällande lokaler, teknik, personal mm?

Bedömning

- Beskriv bedömningsarbetet. Resonera tex utifrån:
 - Hur bedömningen påverkas om provet skrivs på papper respektive skrivs digitalt
 - Bör en eller flera examinatorer bedöma varje individuellt prov?
 - Tidsåtgång för bedömning av varje skriftligt prov per provtagare
 - Behoven av rekrytering och utbildning av examinatorer
 - Bakgrund/kompetens examinatorer
 - Möjligheten att använda AI i bedömning av svaren

Svar från SU:

Provutformning

Enligt GERS (Europarådet, 2020) har den som befinner sig på nivå A2 ett totalt produktivt (s.k. aktivt eller expressivt) ordförråd på cirka 1000 ord. Detta omfång anses vara tillräckligt för att en testtagare t.ex. ska kunna skriva korta, enkla meddelanden såsom mycket enkelt hållna och personliga brev där man till exempel tackar för något.

Ett möjligt provformat skulle kunna vara att testtagarna ska visa sina skriftliga färdigheter genom två olika uppgiftstyper på A2-nivå, en som testar skriftlig *interaktion* och en som testar skriftlig *produktion*. Skriftlig *interaktion*¹ kan t.ex. handla om att kunna besvara kortare mejl eller sms. Detta kan t.ex. göras genom lucktexter, där testtagaren ska kunna fylla i relevanta repliker/kommentarer i en tänkt dialog. Skriftlig *produktion*² kan handla om att testtagaren ska kunna presentera sig själv, berätta något eller kunna föra enklare anteckningar om ett område som står testtagaren nära och som rör testtagarens omedelbara behov. Vad gäller den skriftliga

¹ 3.3.1.2 Overall written interaction A2:

Can compose short, simple formulaic notes relating to matters in areas of immediate need.

² 3.2.1.2 Overall written production A2:

Can produce a series of simple phrases and sentences linked with simple connectors like “and”, “but” and “because”.

produktionen, tänker vi oss att testtagaren ska kunna välja mellan två och tre olika teman. Dessa kan presenteras i form av en fråga, med korta stödjande underfrågor, eller av en uppgift att t.ex. beskriva en bild eller kort bildserie.

Tittar man på hur omfattande skriftliga svar på olika typer av uppgifter kan vara, kräver t.ex. British Council att testtagare på A2-nivå i *interaktion* kan skriva lämpliga svar på mellan 20 och 40 ord. Vid skriftlig *produktion*, där testtagaren t.ex. ska beskriva en bild, föreslår samma institution beskrivningar på mellan 20 och 50 ord (<https://www.englishscore.com/wp-content/uploads/2024/12/British-Council-EnglishScore-Writing-Test-Guide-Edited.pdf>).

Vi tänker oss att en liknande omfattning för att besvara den givna uppgiften är lämpligt, d.v.s. 20-40 ord i skriftlig interaktion och 20-50 ord i skriftlig produktion. Tiden för uppgiften bör ges med viss generositet, med tanke på att grupper som är skrivovana inte ska hindras av för korta tidsramar. Utprovningen bör utvisa hur lång tid denna grupp kan behöva, men ett maxtak på t.ex. 30 minuter per provdel kan prövas.

Genomförande

För att undvika s.k. ”halo-effect” eller ”legibility effect” (Chase 1968, Markham 1976), och för att i stället garantera en rättssäker bedömning är det viktigt att texten är datorskriven så att testtagarnas handstil inte påverkar bedömningen. Det är dock viktigt att tänka på att det i vår målgrupp finns testtagare med låg digital litteracitet (utöver mobiltelefon). Med tanke på just den populationen kan det vara värdefullt att navigeringen går att göra med pekskärm, då den här testtagargruppen kan ha svårigheter att navigera sig fram med datormus.

Bedömning

En rättssäker bedömning av skriftlig färdighet kräver att det finns tydliga bedömningskriterier. En del av framtagningsprocessen kommer därför att vara att utforma och kvalitetssäkra bedömningskriterierna. Vi behöver säkerställa att kriterierna motsvarar A2-nivån i GERS och att de är så pass tydliga att olika bedömare

tolkar dem på ett likvärdigt sätt. Därför behöver vi efter utprovningen av det skriftliga delprovet genomföra en utvärdering och eventuell revidering av kriterierna, samt en kravgränssättning där en panel av experter fastställer var gränsen mellan underkänt och godkänt ska gå.

Av rättssäkerhetsskäl liksom av arbetsmiljömässiga skäl, bör det vara minst två bedömare per prestation. Vid osäkra fall bör det också finnas en tredjebedömare att tillgå. För att hela bedömningsprocessen ska vara tillförlitlig – både vad gäller interreliabilitet (mellan olika bedömare) liksom intrareliabilitet (för varje enskild bedömare) - bör det finnas en sambedömningsprocess med statistisk analys såväl som kvalitativa inslag för samtliga bedömare inför varje nytt prov som ska ges i skarpt läge.

I sambedömningsprocessen bör också löpande fortbildning samt kalibrering ingå, så att bedömarna har en samsyn och en färsk uppdatering av just det prov som står på tur att ges. För att säkerställa att s.k. social kalibrering (jfr Sadler,2005) äger rum, är det viktigt att bedömare med viss regelbundenhet också får träffas och tala om bedömning. I och med att det här blir ett storskaligt prov, kan det vara svårt att samla alla bedömare på en och samma plats under fortbildningstillfällena. Det kan därför bli aktuellt att sätta samman fortbildningsgrupper som får möta varandra med viss rotation, så att alla får en möjlighet att träffa varandra och diskutera bedömningen.

Våra erfarenheter från liknande samarbeten kring t.ex. Tisus visar det som flera studier har pekat på, nämligen att den höga graden av såväl interreliabilitet liksom intrareliabilitet hänger ihop just med bedömargruppens möjligheter att regelbundet få träffa varandra och utbyta tankar och erfarenheter kring aktuellt prov. Bedömargrupper som inte möter andra bedömargrupper, riskerar att få bristande samsyn kring bedömningen, vilket i sin tur äventyrar provets rättssäkerhet.

När en rättssäker bedömarrutin väl är på plats, uppskattar vi att bedömningen av varje skriftlig prestation, interaktion respektive produktion, kan uppskattningsvis ta 5-10 minuter per del och bedömare.

Utifrån USSMA:s prognos om 18 000 testtagare i en första omgång, är vår uppskattning att behovet av att utbilda, fortbilda och fortlöpande kalibrera bedömare kommer att vara omfattande. För att provet ska vara rättssäkert bör bedömarna vara legitimerade lärare i svenska som andraspråk med dokumenterad god erfarenhet av undervisning och bedömning. Vi förespråkar central bedömning, vilket kräver att bedömarna blir bekanta med uppgiftstyp, bedömningskriterier och datorsystem. Därefter måste bedömarna löpande fortbildas och kalibreras, med hjälp av statistisk analys av deras bedömningar.

Eftersom provet har stora konsekvenser för testtagarna är det nödvändigt för rättssäkerheten att tillhandahålla ett representativt ”exempelprov”. Med detta menas ett konkret test att öva på som innehåller exempel på provets upplägg och uppgiftstyper så att testtagarna har möjlighet att förbereda sig innan de gör provet i skarpt läge.

Frågor från UHR:

Muntlig färdighet

– förutsättningar för utformning, genomförande och bedömning

I USSMA:n bedöms prov i muntlig färdighet kunna genomföras genom inläsning på ljudfil som får bedömas och värderas individuellt i efterhand. I rapporten behöver vi därför resonera kring denna möjlighet och andra sätt att genomföra provdelen.

Provutformning

- Beskriv hur provdelen bör utformas. Resonera tex utifrån:
 - Frågeformat och uppgiftstyper
 - Omfattning tid för provtagare att genomföra provdelen
 - Hur framtagande och provutformningen påverkas om provet spelas in och bedöms i efterhand eller om provet genomförs inför examinator
 - Utprovning av provuppgifterna

Genomförande

- Utifrån er erfarenhet av prov: Är det några praktiska förutsättningar, typ av lokaler, teknik, personal mm – som behövs/underlättar när prov genomförs inför examinator, respektive inspelad ljudfil, som ni vill delge UHR?

- I USSMA:n antas att inspelning av en ljudfil borde kunna ske utan att provtagaren behöver byta rum, dvs provtagaren sitter kvar i en provsal med andra provtagare. Resonera kring möjligheterna och vilka förutsättningar som skulle behövas för att det skulle fungera.
- Resonera kring olika sätt att ge provuppgiften – provuppgift/er ges av en provledare/examinator på plats, respektive provuppgift/er ges genom instruktioner på dator
- Om provdelen genomförs på plats inför examinator – en eller fler provtagare som examineras samtidigt?

Bedömning

- Beskriv bedömningsarbetet. Resonera tex utifrån:
 - Hur bedömning av provet påverkas om provet spelas in (USSMA:n förordar ju ljudfiler), och bedöms i efterhand eller om provet genomförs inför examinator
 - Bedömning, en eller flera examinatore. Om flera – kan en vara med vid provtillfället och andra lyssna på inspelning?
 - Tidsåtgång för bedömning av varje muntligt prov/individ?
 - Behoven av rekrytering och utbildning av examinatore
 - Bakgrund/kompetens examinatore

Tidplan – skriftlig och muntlig färdighet (gärna med detaljer kring steg i utvecklingsprocessen för klargörande)

- Tidplan för framtagande av provdelarna till första provtillfället med angivelser om processteg och ungefärliga tidsangivelser
- Tidplan från första provtillfället och 5 år fram
 - Antal möjliga prov per år

Kostnadsberäkningar för prov i skriftlig och muntlig färdighet (behövs först till slutrapporteringen i augusti)

- Redogör både för faktiska kostnader och i termer av helårsarbeten
 - För framtagande
 - För genomförande per år
-

Svar från SU:

Provutformning

Muntliga prov i stor skala innebär utmaningar när det kommer till genomförandet och en central fråga kommer att vara hur vi skapar ett genomförbart prov utan att validiteten, reliabiliteten och rättssäkerheten äventyras. Enligt GERS deskriptorer³ kan talare på A2-nivå främst interagera med andra personer och med deras stöd men även producera⁴ eget mycket enkelt monologiskt tal.

En förutsättning för en hög grad av reliabilitet är att se till att alla testtagare får likvärdiga förutsättningar vad gäller instruktioner, genomförande och bedömning. Ett sätt att skapa en likvärdig testsituation kan vara att ge det muntliga provet digitalt. Fördelarna med det är att testtagarna kan genomföra provet lokalt på många platser i landet. De kan då bli placerade i ett rum med inspelningsutrustning och får följa instruktioner på en skärm, för att sedan spela in sig själva. Ljudfiler kan därefter skickas till en central och till antalet relativt begränsad bedömaregrupp. Ett problem med digitala muntliga prov är dock bristen på autenticitet i situationen, vilket kan försvaga bedömningens validitet. Att som testtagare sitta och tala till en skärm är inte en helt autentisk eller vanlig samtalssituation, även om sådana situationer blivit vanligare i exempelvis mobiltelefon. Risker är att främst enkelriktad muntlig produktion kommer att prövas, men i mindre omfattning muntlig interaktion, trots att interaktion troligen är den vanligaste muntliga färdigheten på denna färdighetsnivå.

³ 3.3.1.1

Overall oral interaction A2

Can interact with reasonable ease in structured situations and short conversations, provided the other person helps if necessary. Can manage simple, routine exchanges without undue effort; can ask and answer questions and exchange ideas and information on familiar topics in predictable everyday situations.

Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters to do with work and free time. Can handle very short social exchanges but is rarely able to understand enough to keep conversation going of their own accord

⁴ 3.2.1.1 Overall oral production A2

Can give a simple description or presentation of people, living or working conditions, daily routines. likes/dislikes, etc. as a short series of simple phrases and sentences linked into a list.

Ett annat sätt är att pröva den muntliga färdigheten i realtid genom att låta testtagare genomföra provet enskilt med en provledare, i par eller grupp tillsammans med andra testtagare under ledning av en lokal bedömare. En fördel med detta är att ett samtal återspeglar en mer autentisk situation om interaktion ska prövas och bedömas. Då förståelse i samtal och interaktion bygger på annat än verbala uttryck, som kompletterande blickar, gester, kroppsspråk, strategier för att visa förståelse etc., blir situationen mer valid och autentisk. En nackdel med test i par eller grupp är svårigheten att forma jämbördiga grupper med likvärdiga förutsättningar. Ur bedömarperspektiv är det svårare att avgöra en svagare testtagares självständighet om vederbörandes talprestation beror på starkare gruppdeltagares insatser. Sett ur bedömarsynpunkt, riskerar gruppvis genomförande också att behöva hanteras av ett stort antal bedömare, vilket i sin tur riskerar att försämra likvärdigheten, och att reliabiliteten påverkas negativt. En nackdel med ett enskilt samtal med provledare kan vara att det innebär att det innebär ett asymmetriskt samtalsmönster där t.ex. bara en ställer frågor och den andra svarar (Rydell, 2016).

Om den muntliga provdelen kommer att ges enskilt till varje testtagare framför en skärm, blir konstruktionsgruppens uppgift att konstruera lämpliga prompter för att elicitera tal (produktion och interaktion) på A2-nivå. Prompterna kan exempelvis ges i form av bilder/bildsekvenser, där testtagaren ges möjlighet att välja en av flera bilder att tala om. Prompterna kan också ges i form av enklare frågor eller uppmaningar som testtagaren får upp på skärmen, helst i både muntligt och skriftligt format. Även korta filminspelningar kan fungera som prompter.

Utmaningen för konstruktörerna ligger i att prompta uppgifter kring teman som ligger testtagarna nära och som ger möjligheter att kommunicera i tillräcklig omfattning för att kunna bedömas samtidigt som uppgifterna inte får avkräva information som blir för personlig. Att berätta till en bild/bildsekvens kan vara ett sätt att elicitera tal kring välbekanta teman, utan att man tvingas prata om sig själv. Testtagaren kan även få viss information om en person på en bild eller i en filminspelning och ställa frågor till

personen i fråga. Ytterligare en utmaning som hänger samman med det digitala provformatet är att inspelningssituationen kan hämma testtagaren, som kan känna sig besvärad och möjligtvis hämmad av att prata utan en verklig samtalspartner. Detta skulle kunna påverka provets validitet negativt. Vi bedömer dock att inspelningssituationer rent generellt är ganska vanligt förekommande i dagens digitala värld. Det är viktigt att frågan följs upp när testtagarna utvärderar provsituationen.

Vid ett digitalt genomförande av den muntliga provdelen skulle deltagarna kunna ges möjlighet till minst två inspelningar och därefter själv kunna välja vilken inspelning som ska bedömas. För att ge testtagaren tillräckligt med tid för att välja en bild eller att välja mellan flera frågor/uppmaningar som ska besvaras/genomföras, så måste testtagaren få minst 15 minuter till sitt förfogande. Även här kommer utprovningsen att ge viktiga ledtrådar till vilken tid som behövs för provdelarna för att ge bedömnarna tillräcklig mycket tal och samtal för en reliabel bedömning av den muntliga prestationen.

Om den muntliga provdelen kommer att ges i form av ett samtal, ligger utmaningen för konstruktörernas del i att hitta enkla men varierade och likvärdiga samtalsteman och uppgifter som kan appellera till den breda domänpopulationen. Denna utmaning gäller även om provet ges digitalt, men blir mer påtaglig och ger större konsekvenser för den enskilde testtagaren om provet ges i realtid med flera gruppdeltagare samt examinator närvarande.

Genomförande

Ett muntligt high-stakes-prov ställer höga krav på den lokal som behövs, liksom på den utrustning som testtagarna behöver. Logistiken kommer troligen inte tillåta att testtagarna genomför provet enskilt i egna rum. Inspelning i datorsal kommer att kräva brusreducerande hörlurar och skärmar mellan testtagarna, men det faktum att testtagare kan komma att störas och/eller stressas av att andra pratar intill, kommer ändå att kunna utgöra ett problem som kan påverka testtagarnas prestationer och därmed provets likvärdighet och rättssäkerhet. Synnerligen höga krav kommer också

att ställas på att det finns snabba lösningar och personligt stöd att tillgå vid eventuella teknikproblem. Utöver detta måste hänsyn tas till den ökade risk för fusk som uppstår då inspelning görs i en datorsal tillsammans med flera testtagare, som givetvis kan lyssna på andra som gör provet.

Bedömning

Precis som för det skriftliga delprovet behöver det finnas tydliga bedömningskriterier för det muntliga delprovet. Framtagningsprocessen för dessa (utprovning, analys, revidering, kravgränssättning) kommer behöva se ut på samma sätt som för det skriftliga delprovet (se föregående avsnitt).

Inspelade prestationer möjliggör central bedömning som inte kommer i direkt anslutning till provets genomförande. Detta ökar provets rättssäkerhet, eftersom bedömarna ges möjlighet att lyssna flera gånger på testtagarnas prestationer. Inspelade prestationer utgör också en säkerhet för bedömarna och för provets rättssäkerhet vid eventuell begäran om omprovning från testtagarens sida.

Vi utgår från att två bedömare, enskilt och utan inverkan på varandra lyssnar på varje prestation. Om bedömningarna skiljer sig åt, går prestationerna vidare till en tredjebedömare. Inför och under bedömningsperioden ska det också finnas bedömar- och kalibreringsträffar. I det här sammanhanget blir det extra viktigt med kravgränssättning och utbildning i GERS-skalan för bedömare. Dessa tillfällen kan samtidigt fungera som upplärningstillfällen för nya bedömare som ska slussas in i systemet.

Vad gäller de krav som måste ställas på bedömarna, ser vi samma bild som vid bedömning av skriftlig färdighet, d.v.s. bedömning bör genomföras av legitimerade lärare i svenska som andraspråk med dokumenterad erfarenhet av bedömning och storskaliga prov. Det kommer också att vara av största vikt att bedömarna löpande genomgår obligatorisk fortbildning och kalibrering med statistiska metoder för att säkerställa likvärdigheten och reliabiliteten. Detta gäller både *interreliabilitet*, d.v.s.

mellan de olika bedömarna och *intrareliabilitet*, d.v.s. mellan den enskilda bedömarens syn på likvärdiga prestationer från ett provtillfälle till ett annat.

Tidsplan

Hur lång tid valideringsprocessen tar är i hög grad avhängigt utprovningen av provet. Vi vet att utprovningsfasen, där vi kommer testa olika muntliga format, kommer att bli tidskrävande. Utprovningen kan visa att provdelar fungerar så som det är tänkt, men den kan också visa motsatsen. Vi måste ta höjd för det senare alternativet, vilket innebär att provet kan behöva modifieras och prövas ut på nytt.

Eftersom provet har stora konsekvenser för testtagarna är det nödvändigt för rättssäkerheten att tillhandahålla ett representativt exempelprov innan provet ges i skarpt läge.

Kostnadsberäkningar

Se bilaga 2 med separat kalkyl som kan behöva revideras.

Referenser/källor:

- ALTE. (2011). *Manual for Language Test Development and Examining*:
https://www.ecml.at/Portals/1/documents/CoE-documents/ManualLanguageTest-Alte2011_EN.pdf
- Bachman, L. F., & Palmer, A. S. (2010). *Language assessment in practice : developing language assessments and justifying their use in the real world*. Oxford University Press.
- Becker, B., Van Rijn, P., Molenaar, D., & Debeer, D. (2022). Item order and speededness: Implications for test fairness in higher educational high-stakes testing. *Assessment & Evaluation in Higher Education*, 47(7), 1030-1042.
- Buerger, S., Kroehne, U., Koehler, C., & Goldhammer, F. (2019). What makes the difference? The impact of item properties on mode effects in reading assessments. *Studies in Educational Evaluation*, 62, 1–9.
<https://doi.org/10.1016/j.stueduc.2019.04.005>
- Bugge, E. & Carlsen C. (2025). *Språk- og samfunnskunnskapskrav for permanent oppholdstillatelse og statsborgerskap. Hvordan påvirkes innvandrere med lite*
-

skolegang av kravene? Hovedfunn og policy-anbefalinger fra IMPECT-prosjektet, HVL-rapport.

- Chase, C. I. (1968). The impact of some obvious variables on essay test scores. *Journal of Educational Measurement*, 5(4), 315-8. Retrieved from <http://www.jstor.org/>
- Cizek, G. J., & Bunch, M. B. (2007). *Standard Setting: A Guide to Establishing and Evaluating Performance Standards on Tests (First Edition)*. SAGE Publications, Inc.
- Delgado, P., Vargas, C., Ackerman, R., & Salmerón, L. (2018). Don't throw away your printed books: A meta-analysis on the effects of reading media on reading comprehension. *Educational Research Review*, 25, 23–38. <https://doi.org/10.1016/j.edurev.2018.09.003>
- Grammatikopoulou, E., Johansson, S., & Rosén, M. (2026). Paper-based and digital reading in 14 countries: exploring cross-country variation in mode effects. *Educational Review*, 78(1), 61–79. <https://doi.org/10.1080/00131911.2025.2452236>
- Europarådet (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment. Companion Volume*.
- Fan, J., Knoch, U. & Bond, T. (2019). Fairness in language assessment: What can the Rasch model offer? *Papers in language testing and assessment*, 8(2), 117–142.
- Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1–73. <https://doi.org/10.1111/jedm.12000>
- Mangen, A., Walgermo, B. R., & Brønnick, K. (2013). Reading linear texts on paper versus computer screen: Effects on reading comprehension. *International Journal of Educational Research*, 58, 61–68. <https://doi.org/10.1016/j.ijer.2012.12.002>
- Markham, L. R. (1976). Influences of handwriting quality on teacher evaluation of written work. *American Educational Research Journal*, 13(4), 277–283. <https://doi.org/10.2307/1162390>
- McNamara, T. & Roever, C. (2007). *Language testing: the social dimension*. Blackwell.
- McNamara, T., Knoch, U., & Fan, J. (2019). *Fairness, justice, and language assessment: the role of measurement*. Oxford University Press.
-

Regeringen. (2026). Skärpta krav för svenskt medborgarskap. Prop. 2025/26:175.
[Skärpta krav för svenskt medborgarskap - Regeringen.se](#)

Rydell, M. (2016). Konsten att föra ett samtal framåt – om bedömning av muntlig interaktion på sfi. I: B. Kindenberg (red.), *Flerspråkighet som resurs. Symposium 2015*. Stockholm. Liber. 236–247.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, 30(2), 175–194. doi:10.1080/0260293042000264262

SOU 2021:2. *Krav på kunskaper i svenska och samhällskunskap för svenskt medborgarskap*. [Krav på kunskaper i svenska och samhällskunskap för svenskt medborgarskap - Regeringen.se](#)

<https://europass.europa.eu/en/common-european-framework-reference-language-skills>

<https://www.englishscore.com/wp-content/uploads/2024/12/British-Council-EnglishScore-Writing-Test-Guide-Edited.pdf>

Bilagor:

Bilaga 1 Testordlista

Bilaga 2 Kostnadsberäkning – sammanställning prognos Medborgarskapsprovet
SU

Bilaga 1

Testordlista

ankaritem (ankaruppgift)	En uppgift/ett item som är gemensam(t) för två eller flera test. Ankaruppgifter har kända egenskaper och ingår som en del av ett nytt test för att ge upplysning om det nya testet och testtagarna, dvs. man kalibrerar det nya testet till en känd skala.
bedömare	En person som bedömer/poängsätter testtagarens prestation på ett test genom subjektiv bedömning. Bedömaren är normalt väl förtrogen med området och utbildad för uppgiften. I muntliga test gör man ibland skillnad mellan rollen som bedömare (bisittare) och samtalspartner.
bedömning	Mätning av en eller flera språkfärdighetsaspekter med hjälp av test eller annat förfarande.
behovsanalys	En analys av testets målgrupp och måldomän i förhållande till dess syfte.
delprov	Del av ett provbatteri, ofta i form av ett separat test, med eget instruktionshäfte och egen tidsgräns. Delproven utgår ofta från olika färdigheter och benämns t.ex. hörförståelse eller skriftlig framställning.
deskriptorer	Standardiserade beskrivningar av hur språkförmåga (språkliga handlingar och prestationer) yttrar sig på olika nivåer.
exempelprov	Med exempelprov menas i detta sammanhang att tillhandahålla ett representativt exempel på provets upplägg och uppgiftstyper som testtagarna kan använda för att förbereda sig innan de gör provet i skarpt läge.
GERS-deskriptorer	GERS-deskriptorer är nivåbeskrivningar av språkfärdighet som används inom Gemensam europeisk referensram för språk (GERS) (eng. Common European Framework of Reference for Languages, CEFR). Deskriptorerna anger vad en språkanvändare typiskt kan göra på olika färdighetsnivåer, från A1 till C2. De formuleras som ”kan-

	göra-påståenden” (can-do statements) som anger vilka kommunikativa handlingar en individ förväntas kunna utföra på en viss språkfärdighetsnivå.
high-stakes-prov	Ett prov vars resultat används för att fatta viktiga beslut om individers möjligheter, rättigheter eller framtida handlingsutrymme.
item	Varje uppgift i ett test som bedöms separat. Den minsta poänggivande enheten i ett test.
itemanalys (uppgiftsanalys)	En statistisk beskrivning av resultatet på såväl enskilda uppgifter som hela provet. Man kan titta på svårighet (p-värden), standardavvikelse, diskriminationsförmåga (r-bis), korrelationer, homogenitet (KR-20) m.m.
interbedömarreliabilitet	Ett mått på reliabilitet som utgår från graden av överensstämmelse mellan olika personers bedömningar av samma testprestation.
intrabedömarreliabilitet	Ett mått på reliabilitet som utgår från graden av överensstämmelse mellan samma persons bedömningar av samma testprestation olika tillfällen
kalibrera	Att beräkna svårighetsgraden på ett antal testuppgifter (inom item-responsteorin).
kalibrering	Processen att fastställa skalan på ett eller flera test. Kalibreringen kan göras med hjälp av ankaruppgifter från olika test till en gemensam skala (theta-skala). När ett test konstrueras med hjälp av ankaruppgifter (kalibrerade) visar poängen testtagarnas förmåga, dvs deras placering på theta-skalan (abilitetsskala).
konstrukt	Den teoretiskt definierade språkförmåga som språktestet avser mäta.
konstruktör (provkonstruktör)	Den som arbetar med att ta fram item/uppgifter till ett prov/test genom hela provkonstruktionsprocessen.
kravgräns	Lägsta poäng (minimipoäng) som krävs på ett test för att erhålla ett givet omdöme. Omdömet kan vara exempelvis godkänd, väl godkänd etc.

kravgränssättning	Systematisk process för att fastställa vilken provprestation/poänggräns som krävs för att uppnå en viss kravgräns på ett test.
operationalisering	Operationalisering av ett konstrukt för språktest innebär att bestämma vilka konkreta testuppgifter och bedömningskriterier som användas för att mäta den aktuella språkförmågan.
population	En definierad grupp individer (testtagare) med någon gemensam egenskap.
prompt	Grafiskt material eller texter, t.ex. bilder, kartor, diagram som används i muntliga eller skriftliga test för att få testtagaren att producera text.
provkonstruktör (konstruktör)	Se konstruktör ovan.
provtagare (testtagare)	Se testtagare nedan.
reliabilitet	Anger konsekvens eller stabilitet i provpoäng. Ju mer reliabelt ett test är desto färre slumpmässiga mätfel förekommer. Ett test som innehåller systematiska fel, t.ex. bias mot någon grupp, kan vara reliabelt men testet saknar validitet.
sambedömning	Innebär att varje testtagares språkprestation bedöms av flera bedömare oberoende av varandra och att bedömningarna sedan jämförs och diskuteras.
test (prov)	Ett förfarande för att pröva olika aspekter på kunskaper/färdigheter.
testtagare (provtagare)	En person som deltar i ett test/prov.
testspekifikation	Detaljerat dokumentationsmaterial, som normalt utarbetas under konstruktion av ett nytt test eller i samband med revision av ett gammalt. Specifikationen ger upplysning om design, innehåll, nivå, uppgiftstyp, målpopulation,

	användning av testet etc. Prov på olika uppgiftstyper ingår ofta.
uppgiftsanalys (itemanalys)	Se itemanalys ovan.
uppgiftstyp	Uppgifter benämns med beskrivande namn, beroende på vad och hur de testas, t.ex. flervalsuppgifter i läsförståelse, styrd skrivning etc.
utprovning (utprovning)	Ett steg i utvecklingen av ett nytt testmaterial, där uppgifter prövas ut på representativa sampel av målpopulationen för att undersöka uppgifternas svårighetsgrad och diskriminationsförmåga. Efter statistisk analys kan de uppgifter som uppfyller ställda kriterier användas i ett verkligt test. Används också med subjektivt bedömda delprov som muntlig och skriftlig framställning.
validitet	Ett tests förmåga att mäta vad det avser mäta och i vilken mån poäng på ett test gör det möjligt att dra slutsatser angående testets lämplighet, användbarhet etc. Innehållsvaliditet, begreppsvaliditet och kriterierelaterad validitet är olika aspekter på validitet. Tillsammans ger de möjlighet att bedöma den allmänna validiteten hos ett test som är konstruerat för ett bestämt syfte.

Flera av definitionerna är helt eller delvis hämtade från:

Corell, M., & Sundberg, G. (Red.). (2000). *Testordlista: terminologi inom språkfärdighetstestning*.

ALTE. (2011). *Manual for Language Test Development and Examining*:
https://www.ecml.at/Portals/1/documents/CoE-documents/ManualLanguageTest-Alte2011_EN.pdf