ARBETSFÖRMEDLINGEN
SWEDISH PUBLIC EMPLOYMENT SERVICE

Träffsäkerhet och likabehandling vid automatiserade anvisningar inom Rusta och matcha

En kvalitetsgranskning

Författare: Anders Böhlmark, Tom Lundström och Petra Ornstein
Arbetsförmedlingen Analys 2021:9

Författare: Anders Böhlmark, Tom Lundström och Petra Ornstein
Datum: 2021-04-12
Diarienummer: Af-2020/0046 7913

Innehåll

1 Sammanfattning

Arbetsförmedlingens reformering bygger på samarbete med fristående aktörer, digitalisering och automatisering. *Rusta och matcha* är den reformerade myndighetens nya stora tjänst för att stärka arbetssökandes möjligheter på arbetsmarknaden. En viktig del av Rusta och matcha är det statistiska bedömningsstödet. Det används för att bedöma huruvida inskrivna arbetssökande har ett tillräckligt stort stödbehov för att anvisas insatser. I denna rapport studeras kvaliteten på bedömningsstödets rekommenderade anvisningar. Det övergripande syftet är att bidra till det pågående utvecklingsarbetet av myndighetens statistiska bedömningsverktyg, med målet om en effektiv och rättssäker handläggningsprocess.

Analysen har genomförts på en testpopulation som skulle varit föremål för bedömning om reformen hade införts i hela landet i början av 2018. I analysen har vi låtit det statistiska bedömningsstödet rangordna de arbetssökandes stödbehov och generera anvisningar. Vi jämför sedan bedömningsstödets output med individernas faktiska stödbehov - i termer av deras verkliga jobbutfall - sex månader efter bedömningstillfället. Syftet är att studera kvaliteten på de rekommenderade anvisningarna avseende andelen korrekta anvisningar (träffsäkerhet) och likabehandling av sökande från olika grupper (exempelvis kvinnor respektive män).

Rapportens viktigaste slutsatser och rekommendationer redovisas nedan i denna sammanfattning. En liknande summering, med fler detaljer och diskussion, finns i den avslutande diskussionen i avsnitt 6.

1.1 Träffsäkerheten är god

Arbetsmarknadspolitiskt relevanta beslut tas när arbetssökande i behov av stöd också anvisas insatser, samtidigt som arbetssökande utan stödbehov inte anvisas insatser. Vi identifierar korrekta beslut när individer med ett faktiskt stödbehov (utan jobb efter sex månader) anvisas insatsen, samt när individer som saknar ett faktiskt stödbehov (de hittade jobb själva) *inte* anvisas insatsen. Resultaten visar att bedömningsstödets rekommendationer är korrekta för 68 procent av de inskrivna. Vi bedömer denna träffsäkerhet som god, baserat på bland annat jämförelser med utvärderingar av kvaliteteten på arbetsförmedlares bedömningar.

1.2 Likabehandling med förbättringspotential

Utgångspunkten i vår analys av likabehandling är principen att *sökande som liknar varandra i termer av faktiskt stödbehov ska behandlas på samma sätt, oavsett vilken grupp de tillhör* (exempelvis kvinnor respektive män). Vi studerar hur bedömningsstödet till Rusta och matcha presterar med avseende på två viktiga egenskaper, som ett bedömningsstöd enligt litteraturen bör ha, för att denna princip ska gälla.

Den första av dessa egenskaper innebär att: *vid en jämförelse av sökande som fått samma bedömning ska sannolikheten att den sökande har ett faktiskt stödbehov=1 (arbetslös sex månader eller längre) vara oberoende av den grupp som individen tillhör (till exempel kvinnor eller män)*. Denna egenskap är särskilt viktig för att arbetsförmedlare, som fattar besluten om anvisning, ska kunna lita på bedömningsstödets rekommendationer. Det vill säga att bedömningar betyder samma sak - i termer av genomsnittligt faktiskt stödbehov - för individer från olika målgrupper. Arbetsförmedlare som vet att det faktiska stödbehovet i genomsnitt är lika stort för sökande som fått samma rekommendation, saknar incitament att försöka korrigera rekommendationerna för enskilda sökande baserat på deras yttre karaktäristika. Vi finner att detta *likabehandlingskriterium* för bedömningsstödet är uppfylld i vissa gruppjämförelser, men inte i andra. Här identifierar vi alltså en förbättringspotential.

Den andra egenskapen innebär att: *sannolikheten att den sökande korrekt anvisas Rusta och matcha bland sökande med ett faktiskt stödbehov=1 är oberoende av den grupp som individen tillhör*. Vi finner att kvinnor och män har samma chans, 67 procent, att korrekt rekommenderas till Rusta och matcha, vid en jämförelse av sökande med ett dokumenterat faktiskt stödbehov. Bedömningsstödet genererar alltså anvisningar som förefaller likabehandla män och kvinnor med stödbehov=1. Det är ett viktigt resultat eftersom ett uttalat krav för Rusta och matcha är att kvinnor och män ges likvärdig tillgång till stöd. Motsvarande jämförelser för övriga målgrupper visar att chansen att rekommenderas anvisning till stöd är betydligt större för sökande som är äldre än 50 år (72 procent jämfört med 47 procent för ungdomar), för utrikes födda (76 procent jämfört med 54 procent för svenskfödda) och för sökande med funktionsnedsättningar (87 procent jämfört med 62 procent för övriga). Dessa resultat kan tolkas som att de ligger i linje med direktiven att rikta mer insatser till grupper som tenderar att stå längre ifrån arbetsmarknaden.

1.3 Bedömningsstödet kan bli ännu bättre

Sammantaget visar våra resultat att kvaliteten på de rekommenderade anvisningarna är relativt god, men vi finner också att det finns en tydlig utvecklingspotential. I rapporten pekar vi på vikten av att använda representativa urval av arbetssökande vid träning och testning av modellerna. Vi pekar också särskilt på behovet av att säkerställa likabehandling på ett mer systematiskt sätt än vad som görs i dagens bedömningsstöd. Vi rekommenderar också att den typ av analyser som ges exempel på i denna rapport genomförs regelbundet, med syfte att kontrollera kvaliteten, men även som ett verktyg i förbättringsarbetet. För detta ändamål har vi tagit fram ett förslag till en rutin för kvalitetssäkring.

2 Inledning

2.1 Bakgrund

Rusta och matcha är Arbetsförmedlingens nya stora tjänst för att stärka arbetssökandes möjligheter på arbetsmarknaden. Tjänsten är en viktig del av myndighetens reformering mot en hög grad av samarbete med fristående aktörer och av digitalisering och automatisering. Sedan mars 2020 pågår en försöksverksamhet med Rusta och matcha (i resten av rapporten: KROM) i 32 av landets kommuner, och nu ska tjänsten införas i hela landet. En viktig del av KROM är det statistiska bedömningsstödet. Det används för att bedöma om inskrivna arbetssökande bör anvisas till insatser eller till fortsatt eget jobbsökande. Inskrivna som bedöms ha goda jobbchanser ska anvisas till fortsatt eget jobbsökande, medan individer som bedöms ha ett tillräckligt stort stödbehov ska anvisas insatser inom KROM. Bedömningarna ska även ligga till grund för nivån på den ersättning som de olika fristående aktörerna får. Den uttalade ambitionen är att besluten om anvisning av arbetssökande i huvudsak ska följa de rekommendationer som bedömningsstödet genererar. Kvaliteten på rekommendationerna är därmed en nyckelfaktor bakom kvaliteten på de beslut som fattas för enskilda arbetssökande.

2.2 Syfte

Syftet med denna rapport är att studera kvaliteten på de automatiserade bedömningarna och rekommenderade anvisningarna i KROM avseende *träffsäkerhet* och *likabehandling*. Med träffsäkerhet menas att individer med verkliga behov av insatser också anvisas till insatser, medan individer som saknar tillräckliga behov inte anvisas till insatser. Likabehandling studeras utifrån principen att individer som liknar varandra i termer av faktiskt stödbehov ska behandlas på samma sätt, oavsett vilken grupp de tillhör (till exempel kvinna eller man).

2.3 Begränsning

I rapporten begränsar vi oss till att studera kvaliteten på rekommendationerna om anvisning för individen avseende fortsatt eget jobbsökande eller att bli anvisad insatser inom KROM av någon omfattning - eller till fördjupat stöd för dem som bedöms stå för långt ifrån arbetsmarknaden för KROM. Vi studerar alltså inte kvaliteten på rekommendationerna till olika nivåer inom KROM (som även ska styra den ersättning som de olika leverantörerna får).

2.4 Metod

Metoden vi använder kan beskrivas i följande steg. (1) Vi har använt data för alla inskrivna i hela landet som var öppet arbetslösa eller i arbetsmarknadspolitiska

program i mars 2018, en population som är representativ för de individer som besluten kom att beröra med start i mars 2020 i försökskommunerna. (2) Vi har låtit använda det statistiska bedömningsstödet på denna testpopulation. Detta genererar en rangordning för varje individs *bedömda stödbehov* som kan jämföras med ett tröskelvärde för anvisning till KROM i det pågående försöket. Med hjälp av detta kan vi alltså observera den rekommenderade anvisningen för varje individ i vår stora testpopulation. (3) Vi jämför sedan den information vi har om bedömt stödbehov och rekommenderad anvisning med hur det faktiskt gick för individerna på arbetsmarknaden inom sex månader. Vi använder alltså individernas realiserade arbetsmarknadsutfall som ett mått på deras *faktiska stödbehov*. På så sätt kan vi verifiera huruvida rekommendationerna som görs vid en viss tidpunkt visar sig vara korrekta i efterhand.

2.5 Relevans för Arbetsförmedlingens beslutfattande

I denna rapport undersöks i vilken grad de automatiserade rekommendationerna om anvisning till KROM kan förväntas följa två delar av det regelverk som reglerar Arbetsförmedlingens beslutsfattande: att beslut ska vara *arbetsmarknadspolitiskt motiverade* samt att myndighetsutövning ska vara *likvärdig*.

Frågan om arbetsmarknadspolitisk relevans är kopplad till myndighetens förordning 9 § (2000:628) om den arbetsmarknadspolitiska verksamheten. Där framgår att anvisningar till arbetsmarknadspolitiska program ska vara arbetsmarknadspolitiskt motiverade. Frågan om likvärdig myndighetsutövning kopplar även till 5 § förvaltningslagen (2017:900), där krav ställs på att beslut ska motiveras sakligt. Likvärdighet i den arbetsmarknadspolitiska verksamheten är i tillägg särskilt reglerat för vissa målgruppstillhörigheter genom krav på att besluten inte ska medföra diskriminering av arbetssökande utifrån någon av diskrimineringsgrunderna. Bestämmelser om det finns i 9 § andra kapitlet i diskrimineringslagen (2008:567).[1]

Kvaliteten på de automatiserade rekommendationerna kan förväntas ha stor påverkan på Arbetsförmedlingens regelefterlevnad i beslutsfattandet inom KROM. Att säkerställa kvaliteteten på dessa rekommendationer och möjliggöra identifikation och åtgärder av eventuella problem är därmed en förutsättning för att myndigheten ska kunna nyttja dem som huvudsakligt beslutsunderlag i myndighetsutövningen. I och med att rekommendationerna kan förväntas ha en inte obetydlig inverkan på den enskildes liv kräver lagstiftningen även att Arbetsförmedlingen har testat och kvalitetskontrollerat att de medför en rättvis behandling.[2]

[1] Förbudet mot diskriminering ska dock inte hindra myndighetens möjligheter att fördela insatser utifrån arbetsmarknadspolitisk relevans eller att vidta åtgärder för att främja jämställdhet eller lika rättigheter och möjligheter utifrån etnicitet. Myndigheten har i tillägg möjlighet att fördela insatser utifrån ålder, givet att det kan motiveras.
[2] DSF 3 kap 13 artikel 2f, samt 3 kap 14 artikel 2f

Rapporten kan även ses som en del i Arbetsförmedlingens utvecklingsarbete med reformeringen av myndigheten, genom att möjliggöra en ökad integrering av statistiska verktyg i de arbetsmarknadspolitiska bedömningarna och vara en del i utvecklingsarbetet med att införa ett delvis automatiserat beslutsfattande. Rapporten kan utifrån detta med fördel ses som en del i en bredare och återkommande uppföljning. Givet att statistiska verktyg fortsätter att användas i allt högre utsträckning, kan denna första granskningsrapport förväntas följas av mer detaljerade analyser vilka även skulle kunna integreras i utvecklingen av nästa generations verktyg. Ett bisyfte är därmed att staka ut analysstrategier och analysområden för kvalitetssäkring av profileringsverktyg.

Rapporten kan också ses som ett led i Arbetsförmedlingens arbete med jämställdhetsintegrering och god förvaltning. Att Arbetsförmedlingen har i uppdrag både att fördela insatser utifrån arbetsmarknadspolitisk relevans och att vidta åtgärder för att främja mångfald och jämställdhet samt motverka diskriminering och den könsuppdelade arbetsmarknaden, innebär krav på stor medvetenhet kring dessa faktorer och hur de ofta omedvetet och indirekt påverkar myndighetens och andras agerande. Arbetsförmedlingens har till uppgift att motverka konsekvenser av diskriminering på arbetsmarknaden, att säkerställa att myndighetens bedömningar inte baseras på förutfattade meningar och fördomar, samt att vidta åtgärder för kvinnor och män att få likvärdig tillgång till stöd, särskilt bland utrikes födda.[3] Granskningar av myndighetens egen verksamhet sker utifrån detta perspektiv återkommande för att säkerställa att verksamheten följer och bidrar till uppsatta mål. I en allt mer digitaliserad myndighetsutövning behöver denna typ av granskningar i högre utsträckning fokusera på just digitala verktyg. Denna rapport, som är en kvalitetskontroll av besluten som följer av myndighetens algoritmer, kan därmed ses som en del av Arbetsförmedlingens utvecklingsarbete av statistiska bedömningsstöd.

2.6 Spårindelningen

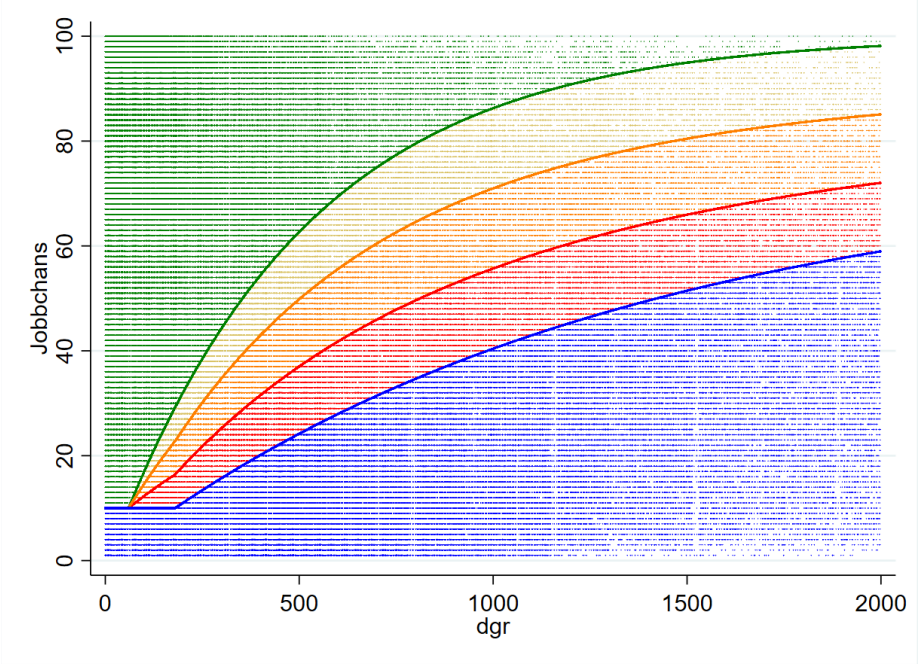
De automatiserade rekommendationerna för anvisning inom KROM är utformade som en spårindelning. Indelningen sker genom att en statistisk modell först bedömer sannolikheten att den arbetssökande får en osubventionerad anställning inom sex månader (jobbchans). Den bedömda jobbchansen baseras på individuella och ort-specifika uppgifter. Därefter vägs den arbetssökandes tid som inskriven på Arbetsförmedlingen in. Kombinationen av jobbchans och antal dagar som inskriven utgör ett mått på individens stödbehov. En individ som har en hög bedömd jobbchans, men som har en lång faktisk tid som arbetslös bakom sig, kan således ha samma stödbehov som en individ med en låg bedömd jobbchans men kort tid som inskriven.

Rekommendationerna om anvisning till KROM eller till eget jobbsökande sker genom att det bedömda stödbehovet jämförs med ett förbestämt tröskelvärde. Ett

[3] https://www.regeringen.se/pressmeddelanden/2019/12/uppdrag-till-arbetsformedlingen-2020/

tröskelvärde, eller beslutsgräns, utgörs av olika kombinationer av jobbchans och antal dagar som inskriven, vilka har bedömts utgöra samma nivå av stödbehov där insatser ska börja vara aktuella.

Figur 1: Bedömd jobbchans över inskrivningstid. Linjerna anger tröskelvärden för bedömt stödbehov.



Spårindelningen illustreras i figur 1. I figuren markeras ett stödbehov lägre än KROMs målgrupp med grön färg (fältet uppe till vänster). Sökande med denna nivå av stödbehov anvisas till eget jobbsökande De tre spåren inom KROM markeras i gult, orange och rött (fälten i mitten). Ett stödbehov som motiverar fördjupat stöd från Arbetsförmedlingen markeras i blått i figuren (fältet nere till höger).[4] I denna rapport studeras enbart rekommenderade anvisningar som följer av att bedömt stödbehov faller inom det gröna fältet (spåret eget jobbsökande) eller något av de andra fälten (spåret som vi förenklat kallar ”anvisad KROM”). Den gröna linjen (första linjen från vänster) är följaktligen det tröskelvärde/beslutsgräns som vi refererar till i resten av rapporten.

[4] Notera att denna beskrivning av skarpa beslutsgränser för anvisning är den som är tänkt att gälla vid en implementering av KROM i hela landet. I det pågående försöket i ”KROM-kommunerna” sker i utvärderingssyfte också en randomisering av individer till olika spår från urval av individer som befinner sig i zonerna mellan varje spår.

3 Testpopulationen

För att analysera träffsäkerhet och likabehandling i de rekommenderade anvisningarna behöver vi använda historiska uppgifter. För att kunna analysera mindre delpopulationer, t.ex. individer med funktionsnedsättningar, behöver vi dessutom ha ett dataunderlag bestående av en stor mängd individer. Vi använder därför en stor testpopulation bestående av alla individer i hela landet som var öppet arbetslösa eller i arbetsmarknadspolitiska program per 2018-03-01.

Våra analyser genomförs således på en testpopulation som skulle ha utgjort den relevanta populationen för urval till KROM om reformen hade införts i hela landet i början av 2018. För att kvalitetskontrollen ska ge information om de framtida besluten är det viktigt att de historiska uppgifterna kommer från individer och förhållanden som i så hög utsträckning som möjligt motsvarar de individer och förhållanden som besluten avser gälla.[5] En viktig jämförelse är därför om vår testpopulation liknar de arbetssökande som var inskrivna i ”KROM-kommunerna”, det vill säga de individer som ingick i det verkliga skarpa försöket två år senare (2020-03-18). Jämförelsen, som redovisas i Bilaga 7.1, visar att de inskrivna i KROM-kommunerna 2018 i allt väsentligt likar de inskrivna i samma kommuner 2020. Vidare kan vi se att de inskrivna i KROM-kommunerna liknar de inskrivna i landet i övrigt. Vi finner alltså att testpopulationen är representativ för de individer som besluten gäller i det pågående försöket (vilka i sin tur är representativa för de inskrivna i hela landet 2020).[6]

4 Arbetsmarknadspolitisk relevans

För att det statistiska bedömningsstödet ska vara relevant bör det i tillräckligt hög grad se till att individer med ett stort stödbehov anvisas KROM, och att individer med ett lågt stödbehov istället anvisas eget jobbsökande. För att bedöma den arbetsmarknadspolitiska relevansen studerar vi här träffsäkerheten, det vill säga i vilken grad korrekta anvisningar görs. De mått på träffsäkerhet som vi använder kallas i litteraturen för accuracy och F1-värde.[7]

Accuracy (tillförlitlighet) är andelen av alla individer som är korrekt klassificerade. Individer som är korrekt klassificerade i vår undersökning är antingen ”anvisade eget jobbsökande *och* fick ett jobb inom sex månader” eller

[5] Om möjligt ska undersökningen testas på flera typer av populationer och förhållanden – eftersom framtiden kan förväntas förändras, och det är viktigt att säkerställa att beslutsprocessen kan hålla en tillräcklig kvalitet även under förändringar, samt att utforska gränserna för vad det statistiska verktyget utan justeringar kan hantera.
[6] Detta kan troligen komma att ändras något till följd av den pandemiska kris som i skrivande stund drabbar den svenska ekonomin, men hur mycket går just nu inte att fastställa.
[7] Notera att denna undersökning inte är en fråga om att välja en modell över en annan, och därmed är valet av specifikt test inte kritiskt. Undersökningen är en övergripande kvalitetskontroll av den modell som är i bruk, och vi använder två olika mått på träffsäkerhet som kompletterar varandra.

”anvisad KROM *och* ännu utan jobb efter sex månader”. Jobb inom sex månader innebär en osubventionerad anställning inom sex månader, vilket är samma definition som används när den statistiska modellen beräknar jobbchans för varje individ. F1-värdet är ett mått som vi använder för att studera skillnader i träffsäkerhet separat för individer som antingen anvisas till KROM eller till eget jobbsökande.

Sammanfattningsvis visar analysen att träffsäkerheten i sin helhet är god. Träffsäkerheten är hög för anvisningar till KROM: den statistiska bedömningen leder till att få individer med ett lågt faktiskt stödbehov felaktigt anvisas KROM. Träffsäkerheten är mycket lägre för anvisningar till eget jobbsökande: den statistiska bedömningen leder till att relativt fler individer med ett högt faktiskt stödbehov felaktigt anvisas till eget jobbsökande. Det senare gäller särskilt för personer med en dokumenterad funktionsnedsättning samt för personer födda utanför Sverige (grupper där få sökande har stödbehov=0).

4.1 Resultat från analys av träffsäkerheten

I tabell 1 kan vi se att träffsäkerheten mätt som accuracy är relativt hög och att den genomgående är högre än vad den hade varit om anvisningar hade skett slumpmässigt. Jämförelsen med slumpen kan verka konstig mot bakgrund av att myndigheten inte använder slumpen vid tilldelning av insatser. Jämförelsen kan motiveras dels med att slumpmässig anvisning kan ses som en absolut undre gräns för vad som är en godtagbar träffsäkerhet. Den kan även motiveras med att flera studier som undersöker träffsäkerheten i manuella bedömningar av arbetssökandes stödbehov använder slumpen som en relevant jämförelsenivå. En schweizisk studie av arbetsförmedlares träffsäkerhet i allokeringen av verksamma insatser (Lechner och Smith, 2007) indikerar att arbetsförmedlare inte ökar effektiviteten i anvisningarna utöver slumpen. En studie av arbetssökandes förmåga att bedöma sina egna jobbchanser kommer till samma slutsats (Smith med flera, 2020). En ny studie från Arbetsförmedlingen finner också att arbetsförmedlares förmåga att bedöma unga arbetssökandes stödbehov är signifikant lägre än om bedömningen görs av en statistisk modell (Ornstein & Thunström, 2021).

En viktig sak att ha med sig är att våra jämförelser är gjorda på ett sätt som ger slumpen samma grundförutsättningar som det statistiska verktyget. Bedömningsstödet har betydligt lättare att anvisa korrekt om det faktiska stödbehovet i populationen i genomsnitt antingen är högt eller lågt. På motsvarande sätt ger vi slumpen lika lätt att gissa rätt vid skeva fördelningar.[8]

[8] Med slumpmässig anvisning menas i denna rapport att anvisningen sker slumpmässigt och med en sannolikhet att ”anvisas till KROM” som är lika stor som andelen som anvisas till KROM av det statistiska verktyget. Antag att px är andelen som anvisas till KROM och att py är andelen som faktiskt inte övergår till jobb inom sex månader. Under dessa förutsättningar är sannolikheten för att av slumpen korrekt anvisas till KROM px * py (sannolikheterna får multipliceras eftersom de är oberoende av varandra på grund av den slumpmässiga tilldelningen). Sannolikheten för att korrekt bedömas stå för nära blir på samma sätt (1-px)(1-py). Träffsäkerheten vid slumpmässig anvisning är summan av dessa sannolikheter: px*py + (1-px)(1-py). Detta är ett sätt att försöka göra jämförelsen med slumpen på ett ”rättvist” sätt. Det medför också att slumpen kan prestera avsevärt bättre än en träffsäkerhet på 50 %.

Tabell 1. Träffsäkerhet i bedömningen av stödbehov

Population	**Accuracy** Anvisning baserad på statistisk bedömning	**Accuracy** Slumpmässig anvisning	Andel av förbättringspotential vid slump-anvisning som kan förklaras av statistisk anvisning	**F1** Anvisade KROM	**F1** Anvisade eget jobbsökande
Samtliga	0,68	0,57	0,25	0,77	0,44
Män	0,67	0,56	0,26	0,77	0,45
Kvinnor	0,68	0,57	0,25	0,78	0,42
Ej funktionsnedsatta	0,64	0,52	0,25	0,73	0,46
Funktionsnedsatta	0,81	0,80	0,07	0,89	0,17
Svenskfödda	0,61	0,48	0,25	0,67	0,51
Utrikes födda	0,73	0,67	0,18	0,83	0,32
Ej ungdom	0,70	0,59	0,26	0,79	0,43
Ungdom	0,56	0,44	0,20	0,61	0,48
50 år eller yngre	0,66	0,55	0,24	0,76	0,44
Äldre än 50 år	0,71	0,60	0,28	0,81	0,44
Ej KROM-kommun	0,67	0,56	0,25	0,77	0,44
KROM-kommun	0,70	0,59	0,27	0,79	0,45

En accuracy på 0,68 betyder att 68 procent av de arbetssökande får ett korrekt beslut, vilket kan jämföras med att 57 procent skulle ha fått korrekt beslut om beslutet hade skett slumpmässigt. Det statistiska bedömningsstödet ger därmed en träffsäkerhet som ligger 11 procentenheter över slumpmässig tilldelning av insatser. Ett annat sätt att beskriva träffsäkerheten relativt slumpen är att jämföra med den förbättringspotential (på 43 procentenheter) som finns vid slumpmässig anvisning: bedömningsstödet förbättrar med cirka en fjärdedel (11/43) av förbättringspotentialen vid slumpmässig anvisning. Eller enklare uttryckt: bedömningsstödet rättar till cirka 25 procent av det som slumpen gör fel.
En träffsäkerhet på 0,68 kan också jämföras med en träffsäkerhet på 0,82 vilket är det man får om alla individer hade placerats i KROM. Detta resultat följer av det faktum att de flesta individerna i testpopulationen har ett stort faktiskt stödbehov, så som vi kan mäta det. Att anvisa alla till KROM blir alltså mekaniskt ett rätt beslut i 82 procent av fallen och fel beslut i 18 procent av fallen. Att inte alla sökande anvisas till KROM beror främst på en budgetrestriktion som styr var gränsen för anvisning sätts.

Det statistiska bedömningsstödet kan alltså förväntas ha lättare att korrekt anvisa individer med ett stödbehov=1 (utan jobb) till KROM än att korrekt anvisa individer med ett stödbehov=0 (de hittade jobb) till eget jobbsökande. Detta bekräftas tydligt av F1-värdena: träffsäkerheten är betydligt högre för anvisningar till KROM (0,77) än för anvisningar till eget jobbsökande (0,44).

Separata analyser för olika undergrupper visar att accuracy och F1-värden är tämligen likvärdiga när man jämför män och kvinnor, olika åldersgrupper samt KROM-kommuner och icke KROM-kommuner. Däremot kan vi observera en accuracy på 0,81 för funktionsnedsatta jämfört med 0,64 bland ej funktionsnedsatta. Detta kan ses i ljuset av att en accuracy på 0,91 skulle uppnås genom att anvisa alla i gruppen funktionsnedsatta till KROM (endast 9 procent hittar jobb inom sex månader). Vi kan även se att träffsäkerheten i bedömningen för denna grupp bara förbättras med 1 procentenhet när bedömningsstödet används jämfört med slumpmässig bedömning. Följaktligen visar de båda F1-värdena att bedömningsstödet lyckas extremt bra på att träffa rätt i anvisningarna till KROM för funktionsnedsatta individer (F1=0,89), men dåligt med att träffa rätt i anvisningarna till eget jobbsökande (F1=0,17). På ett liknande sätt lyckas bedömningsstödet bättre med att korrekt anvisa utrikes födda till KROM, men sämre med att korrekt anvisa samma grupp till eget jobbsökande. F1-värdena är 0,83 respektive 0,32 för utrikes födda.

5 Likabehandling

I rapporten studerar vi likabehandling utifrån två olika kriterier, vilket vi redogör för nedan. Vi kan dock börja med att konstatera att det finns en stor och snabbt växande USA-dominerad litteratur som handlar om rättvisa/likabehandling vid automatiserade bedömningar. Några viktiga lärdomar från denna litteratur är att: (1) Det finns många sätt att se på och definiera likabehandling; (2) Det finns olika sätt att säkerställa att de automatiserade bedömningarna leder till önskvärda utfall enligt vald likabehandlingsdefinition; (3) Valet av definition och metod för implementering är kontext-specifik; (4) Litteraturen växte fram från en oro att automatiserade bedömningar riskerade förstärka redan orättvisa utfall, men idag pekar litteraturen på möjligheterna (och utmaningarna) med att säkerställa rättvisa utfall på ett förutbestämt sätt.

5.1 Kriterier för likabehandling

Många modeller för automatiserade bedömningar som används i praktiken världen över bedömer sannolikhet/risk för ett önskvärt eller icke önskvärt binärt utfall, till exempel sannolikheten att en individ ska klara av en utbildning eller att ett hushåll ska ställa in betalningarna på ett lån. Litteraturen beskriver olika kriterier för att bedömningarna görs på ett sätt så att individer som liknar varandra behandlas på samma sätt, oberoende av grupptillhörighet.

En mycket uppmärksammad debatt om likabehandling i USA har handlat om ett bedömningsstöd för att bedöma återfallsrisk i brottslighet (*COMPAS*). Man har kunnat visa att svarta och vita åtalade med samma faktiska realiserade utfall (inte återfall i brottslighet *eller* återfall i brottslighet) i genomsnitt fick olika bedömd risk och därmed olika stränga påföljder. Många välgjorda studier hänvisar till COMPAS. En central studie är Kleinberg, Mullainathan och Raghavan (2016). Studien utgår från exemplet med COMPAS och sammanfattar några viktiga egenskaper som en riskbedömning kan behöva ha för att vara rättvis. Två viktiga definitioner att först ha med sig är:

Positiva fall: Individer med faktiskt utfall = 1
Negativa fall: Individer med faktiskt utfall = 0

Viktiga kriterier som lyfts fram i studien är:

(A) Bland individer som *bedöms* ha en viss sannolikhet *z* att vara positiva fall, ska ungefär en lika stor andel också vara positiva fall. För likabehandling på gruppnivå ska detta även gälla separat för olika grupper, tex för män respektive kvinnor. [9] Med andra ord: *vid en jämförelse av sökande som fått samma bedömning ska sannolikheten att individen är ett positivt fall vara oberoende av den grupp som individen tillhör.*

(B) Den genomsnittliga riskbedömningen för individer i grupp 1 (tex kvinnor) ska vara samma som den genomsnittliga riskbedömningen för individer i grupp 2 (tex män), bland individer som är *positiva fall*.

(C) Den genomsnittliga riskbedömningen för individer i grupp 1 (tex kvinnor) ska vara samma som den genomsnittliga riskbedömningen för individer i grupp 2 (tex män), bland individer som är *negativa fall*.

5.2 Likabehandling i dagens bedömningsstöd

Som beskrevs i avsnitt 2.1 ovan baseras de rekommenderade anvisningarna till KROM på ett system i två steg. Systemets första del är en statistisk modell som bedömer den sökandes jobbchans, det vill säga sannolikheten att den arbetssökande får en osubventionerad anställning inom sex månader. Denna del av bedömningsstödet är alltså utformad på samma sätt som många av de bedömningsstöd som beskrivs i litteraturen: Det som bedöms är individens

[9] I litteraturen beskrivs denna egenskap som en viktig garanti för likabehandling när de skattade sannolikheterna används som bedömningsstöd av mänskliga beslutsfattare (Kleinberg, Mullainathan och Raghavan, 2016; Noriega-Campero, Bakker, Garcia-Bulle och Pentland, 2018). I exemplet COMPAS handlar det om vikten av att skattade brottsåterfallssannolikheter betyder samma sak för svarta och vita åtalade. Om så inte är fallet riskerar domare, som fattar beslut med hjälp av COMPAS att systematiskt ge åtalade från de två grupperna olika stränga påföljder.

chans/risk att få ett av två möjliga utfall. I fallet KROM är det chansen till jobb inom sex månader, eller omvänt risken för långtidsarbetslöshet.[10]

Bedömningsmodellen, systemets första del, i KROM är designat för att uppfylla egenskap (A) som beskrevs i avsnitt 5.1 ovan. Denna egenskap innebär att de skattade sannolikheterna faktiskt menar vad de säger, och att de betyder samma sak för olika delgrupper. Om till exempel ett antal individer har fått sin jobbchans bedömd att vara 0.5 ska också 50 procent av dessa individer faktiskt hitta ett jobb inom sex månader. Och detta ska inte skilja sig för olika delgrupper, till exempel för kvinnor respektive män.[11]

I litteraturen beskrivs (A) som en viktig garanti för likabehandling när de skattade sannolikheterna används som bedömningsstöd av mänskliga beslutsfattare (Kleinberg, Mullainathan och Raghavan, 2016; Noriega-Campero, Bakker, Garcia-Bulle och Pentland, 2018). Anledningen är att om (A) inte är uppfyllt och en beslutsfattare ska fatta beslut om en individ, baserat på bedömningar som betyder en sak för en viss grupp och en annan sak för en annan grupp, så har beslutsfattaren incitament att väga in individens grupptillhörighet i beslutet. Resultatet *efter* beslutsfattarens beslut riskerar då bli att individer som är lika varandra i det relevanta avseendet (tex samma faktiska risk för långtidsarbetslöshet) får olika behandling (tex olika anvisning) beroende på sin grupptillhörighet (irrelevant egenskap).

I fallet med KROM fattar inte arbetsförmedlaren beslut baserat på individens bedömda jobbchans, utan enbart baserat på verktygets rekommendation vad gäller om den arbetssökande ska ta del av tjänsten (och i så fall vilket spår). Egenskap (A) förefaller ändå vara relevant för bedömningsstödet, även om besluten inte baseras på den typ av risk-/chansbedömning som litteraturen relaterar till. En liknande risk för systematisk olikbehandling borde föreligga när beslut baseras på rekommenderad anvisning: Om till exempel rekommenderad avisning till KROM innebär ett lägre faktiskt stödbehov i genomsnitt för en *grupp 1* jämfört med en *grupp 2* så kan arbetsförmedlare tendera att i högre utsträckning göra ”annan bedömning” för individer som tillhör grupp 1. Resultatet blir då att enskilda sökande med ett högt faktiskt stödbehov blir inkorrekt anvisade eget jobbsökande enbart på grund av sin grupptillhörighet.

5.3 Vår kvalitetsgranskning av likabehandling

I rapporten studerar vi likabehandling utifrån de två olika kriterierna (A) och (B) vilka beskrevs i avsnitt 5.1 ovan. Att vi utvärderar likabehandling enligt (A) faller sig naturligt av att det är det kriterium som Arbetsförmedlingen har fokuserat på i

[10] Notera att när bedömningsstödet tas fram så görs det på en population där man känner till de sökandes sanna jobbutfall inom sex månader.
[11] Se avsnitt 6.3.1 ”Analyser för att utesluta systematiska fel” i Arbetsförmedlingens rapport ” Beskrivning av arbetsmarknadspolitisk bedömning med ett statistiskt bedömningsstöd i Kundval Rusta och Matcha. Rättsliga ställningstaganden.”

konstruktionen av bedömningsmodellen. Vi utvärderar även utifrån kriterium (B) eftersom det framstår som relevant för KROM. Detta kriterium kan beskrivas som att ”individer med liknande faktiskt högt stödbehov ska ha lika möjligheter att korrekt anvisas KROM, oavsett deras grupptillhörighet”. Kriterium (C) är också potentiellt relevant (lika möjligheter att anvisas eget jobbsökande bland dem med lågt stödbehov), men vi har begränsat oss till att inte studera det. Anledningen är att kriteriet framstår som mer relevant om KROM skulle visa sig ha reella inlåsningseffekter, vilket KROM ska vara utformat för att undvika.

Vårt fokus när vi studerar likabehandling utifrån kriterium (A) respektive (B) är att utvärdera bedömningsstödet i sin helhet, det vill säga efter det att riskbedömningarna från systemets första del även kombinerats med individernas inskrivningstid. Vi menar att det mest relevanta att kvalitetsgranska är de rekommenderade anvisningarna, och dessa baseras på hela bedömningsstödet, inte bara jobbchansbedömningarna i systemets första del.

När vi studerar likabehandling utifrån kriterium (A) tittar vi på om den önskvärda egenskapen om inga gruppskillnader (till exempel mellan män och kvinnor) kvarstår när vi utvärderar hela bedömningsstödet. Om egenskap (A) kvarstår innebär det inga gruppskillnader i faktiskt stödbehov bland individer som bedömts ha ett lika stort stödbehov (samma scorevärde, vilket är en kombination av jobbchans och inskrivningstid).[12]

När vi studerar likabehandling utifrån kriterium (B) jämför vi enbart individer som är ”positiva fall”, det vill säga individer med stödbehov=1. Vi studerar då om individer - som liknar varandra i det relevanta avseendet att de har ett faktiskt stödbehov - har samma chans att korrekt anvisas KROM, oavsett vilken grupp de tillhör.

Notera att båda kriterierna (A) och (B) syftar till att säkerställa att sökande som liknar varandra i termer av stödbehov (den relevanta egenskapen) ska behandlas på samma sätt, oavsett vilken grupp de tillhör (till exempel kvinna eller man). (A) beskrivs i litteraturen som viktigt *indirekt,* genom att mänskliga beslutsfattare (arbetsförmedlare) måste kunna lita på att bedömningsstödets bedömningar är relevanta och betyder samma sak för individer från olika grupper. (B) är viktigt som ett direkt kriterium för att bedömningsstödets anvisningar likabehandlar sökande med ett faktiskt stödbehov. Vi studerar kvaliteten på bedömningsstödets rekommenderade anvisningar gällande både (A) och (B). Eftersom (A) har

[12] Genom extra analyser, som kan erhållas på begäran, har vi verifierat att kriterium (A) är uppfyllt i bedömningsmodellen (systemets första del) när vi testar för detta på ett urval av relativt nyinskrivna sökande. Detta är förväntat eftersom riskbedömningarna för långtidsarbetslöshet (eller omvänt jobbchans) är framtagna baserat på ett urval av nyinskrivna där man har säkerställt att kriterium (A) ska vara uppfyllt. Att vi kan replikera detta med nyinskrivna i vår stora testpopulation visar att skilda resultat från vår utvärdering inte kan förklaras med att vi använder ett annorlunda urval. Vidare finner vi att egenskap (A) inte längre är uppfyllt när vi testar för detta i bedömningsmodellen (systemets första del), men använder modellen för alla inskrivna och inte bara relativt nyinskrivna. Detta visar att bedömningsmodellen borde baseras (”tränas”) på ett urval av sökande som är representativt för de sökande som modellen kommer att användas för, det vill säga även för sökande med längre inskrivningstider.

betydelse indirekt för likabehandling i de slutliga besluten för verkliga sökande kan våra resultat för kriterium (A) inte jämföras med resultaten för kriterium (B).

Vi jämför följande målgrupper: *ungdomar* (under 25 år) och *ej ungdomar* (25 år eller äldre); *äldre än 50 år* och *50 år eller yngre*; *kvinnor* och *män*; *funktionsnedsatta* och *ej funktionsnedsatta*; *utrikes födda* och *svenskfödda*.

Sammanfattningsvis visar analysen av kriterium (A) inga systematiska skillnader avseende ålder. Och vi finner mindre skillnader med avseende på kön. Vi finner vidare endast mindre skillnader för alla gruppjämförelser bland individer med ett högt bedömt stödbehov, det vill säga bland dem som är tydligt över beslutsgränsen för att anvisas till KROM. Vi observerar dock större systematiska skillnader vid några gruppjämförelser (funktionsnedsättning respektive utländsk bakgrund) när vi beaktar individer med ett bedömt stödbehov just omkring och under gränsen för anvisning till KROM. Kriterium (A), med fokus på gruppskillnader, kvarstår alltså delvis men inte helt när vi utvärderar hela bedömningsstödet. Analysen av kriterium (B) visar att kvinnor och män med stödbehov=1 har samma chans att korrekt rekommenderas KROM. Vi finner vidare att de grupper som i genomsnitt står längre ifrån arbetsmarknaden (äldre, utrikes födda och funktionsnedsatta) har en högre chans att korrekt rekommenderas anvisning till KROM än andra grupper.

5.4 Resultat från analys av likabehandling

Vi börjar med analysen av likabehandlingskriterium (A). Vi gör jämförelserna utmed hela skalan av de bedömningar som det statistiska verktyget genererar. Detta är av intresse för att kunna avgöra om eventuella avvikelser uppträder bland sökande med lågt bedömt stödbehov, runt tröskelvärdet, eller bland dem med högt bedömt stödbehov.

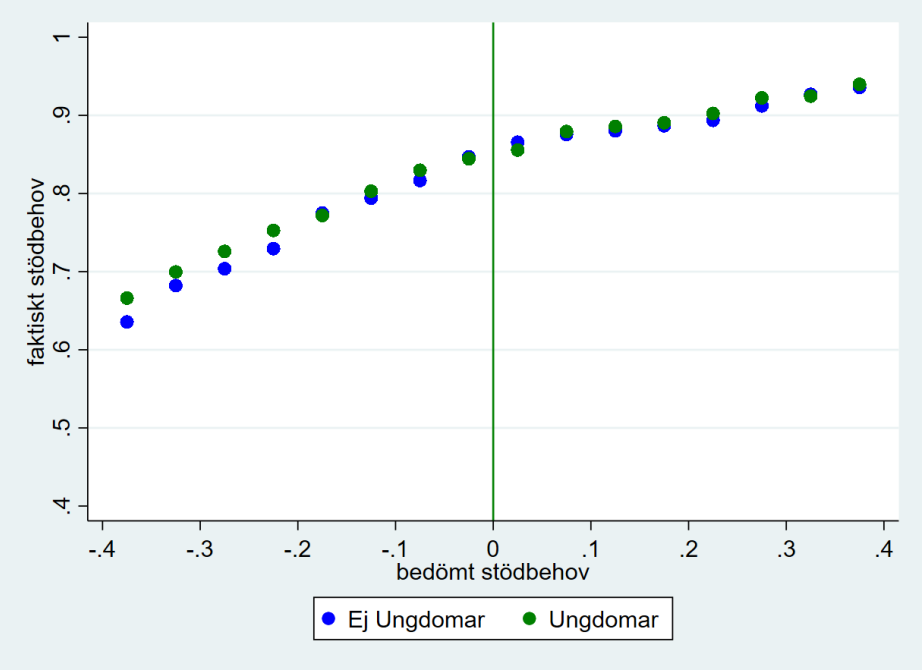
I figurerna nedan visas individens bedömda stödbehov på x-axeln. Det underliggande bedömda stödbehovet är ett värde utan uppenbar tolkning, eftersom det är en kombination av estimerad jobbchans och tid som inskriven (se figur 1 ovan). Vi har valt att uttrycka det bedömda stödbehovet som *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*.

Positiva värden indikerar att individen har ett stödbehov som är bedömt att vara större än tröskelvärdet. Tröskelvärdet markeras med ”0”. Individer till höger om tröskelvärdet i figurerna är alltså de som också får rekommendationen att anvisas till KROM. Större positiva värden innebär ett större bedömt stödbehov. Negativa värden indikerar att individen har ett stödbehov som är bedömt att vara mindre än tröskelvärdet. Individer till vänster om tröskelvärdet i figurerna är alltså de som rekommenderas till eget jobbsökande. Större negativa värden innebär ett mindre bedömt stödbehov.

På y-axeln visas genomsnittligt faktiskt stödbehov för respektive målgrupp. Det är andelen individer som inte har hittat en osubventionerad anställning inom sex

månader efter bedömningstillfället. Det är alltså genomsnittet för individer i respektive grupp med ett faktiskt stödbehov=1 (de utan jobb efter sex månader) och de som har ett faktiskt stödbehov=0 (de hittade jobb).

Figur 2. Andel utan jobb sex månader efter bedömningstillfället

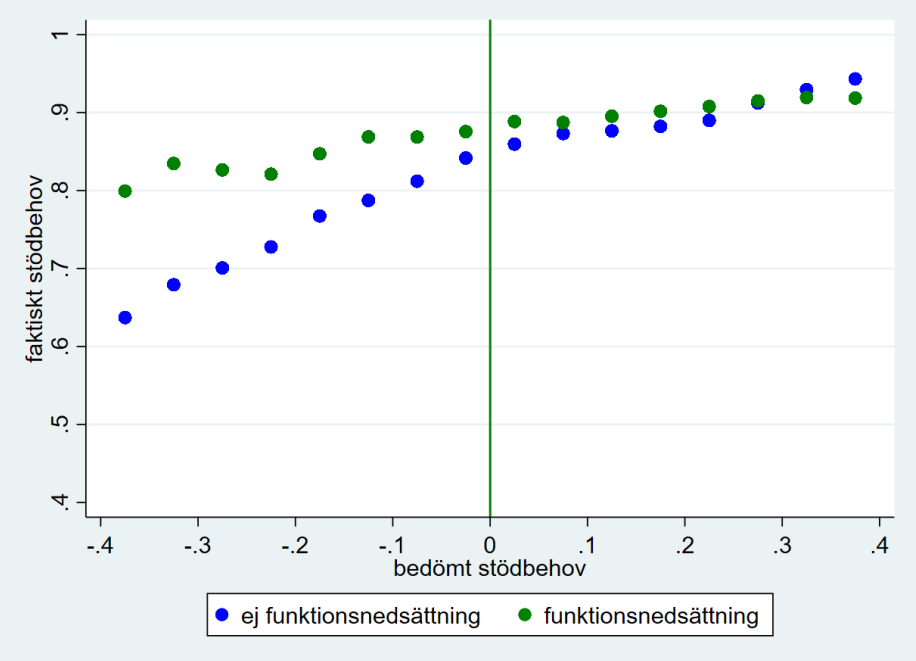


Notera: *faktiskt stödbehov* är andelen individer i respektive grupp utan jobb (utan osubventionerad anställning) inom sex månader efter bedömningstillfället. *bedömt stödbehov* är *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*. *tröskelvärdet* är bedömningsgränsen för anvisning till KROM och markeras med den gröna vertikala linjen vid 0 på det bedömda stödbehovet. Individer till höger om den gröna linjen (med ett bedömt stödbehov>0) anvisas till KROM. Individer till vänster om den gröna linjen (med ett bedömt stödbehov<0) anvisas till eget jobbsökande.

Figur 2 visar jämförelsen mellan ungdomar och övriga. Vi kan inte observera några noterbara skillnader i genomsnittligt faktiskt stödbehov för de två grupperna, varken bland individer som anvisas eget jobbsökande (till vänster om 0) eller bland de som anvisas KROM (till höger om 0). Vi får samma resultat vid en jämförelse av sökande över 50 år och övriga, se figur i bilaga. Vi drar slutsatsen att kriterium (A) – som modellen som estimerar jobbchans är konstruerad för att uppfylla – kvarstår när vi utvärderar bedömningsstödet i sin helhet med avseende på olika åldersgrupper.

Figur 3 visar jämförelsen mellan personer med en dokumenterad funktionsnedsättning och övriga. Vi kan inte observera några noterbara skillnader i genomsnittligt faktiskt stödbehov bland de som anvisas KROM. Däremot kan vi se ett gap mellan grupperna till vänster om tröskelvärdet, som blir större ju lägre individernas bedömda stödbehov är. Vi drar slutsatsen att kriterium (A) – som modellen som estimerar jobbchans är konstruerad för att uppfylla – inte kvarstår när vi utvärderar bedömningsstödet i sin helhet då vi jämför funktionsnedsatta individer mot övriga.

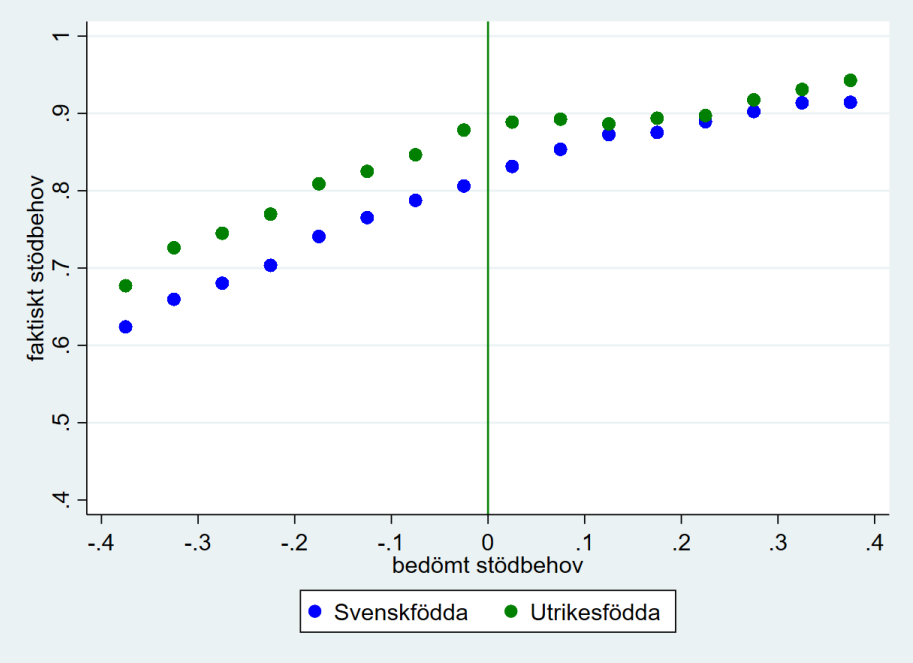
Figur 3. Andel utan jobb sex månader efter bedömningstillfället



Notera: *faktiskt stödbehov* är andelen individer i respektive grupp utan jobb (utan osubventionerad anställning) inom sex månader efter bedömningstillfället. *bedömt stödbehov* är *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*. *tröskelvärdet* är bedömningsgränsen för anvisning till KROM och markeras med den gröna vertikala linjen vid 0 på det bedömda stödbehovet. Individer till höger om den gröna linjen (med ett bedömt stödbehov>0) anvisas till KROM. Individer till vänster om den gröna linjen (med ett bedömt stödbehov<0) anvisas till eget jobbsökande.

Figur 4 visar jämförelsen mellan personer födda i Sverige och personer födda i ett annat land. Vi kan inte observera några noterbara skillnader i genomsnittligt faktiskt stödbehov bland individer med högt bedömt stödbehov. Däremot kan vi se ett gap mellan grupperna precis runt och under bedömningsgränsen. Vi drar slutsatsen att kriterium (A) – som modellen som estimerar jobbchans är konstruerad för att uppfylla – inte kvarstår till fullo när vi utvärderar bedömningsstödet i sin helhet då vi jämför utrikes födda individer mot övriga.

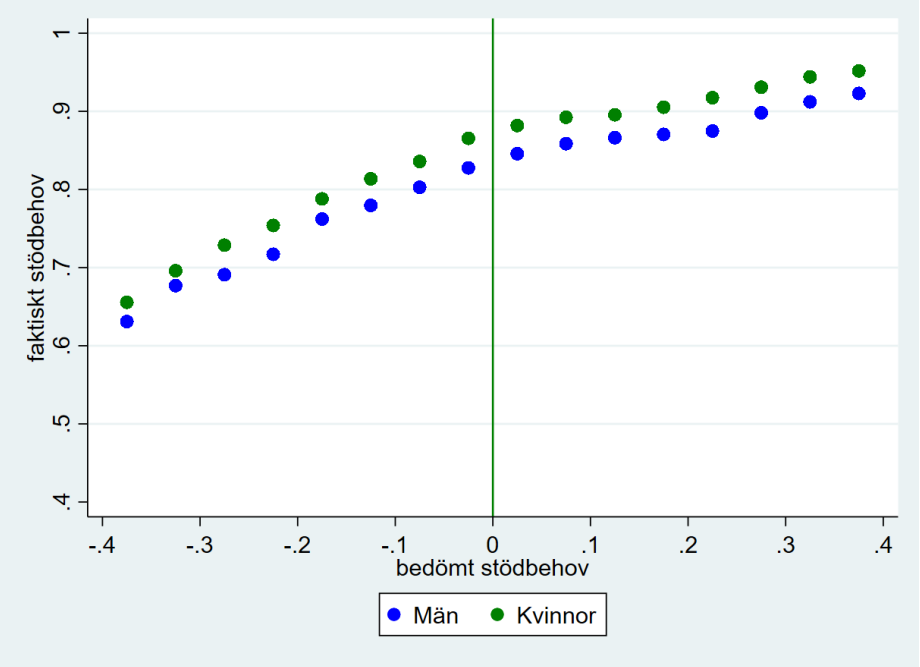
Figur 4. Andel utan jobb sex månader efter bedömningstillfället



Notera: *faktiskt stödbehov* är andelen individer i respektive grupp utan jobb (utan osubventionerad anställning) inom sex månader efter bedömningstillfället. *bedömt stödbehov* är *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*. *tröskelvärdet* är bedömningsgränsen för anvisning till KROM och markeras med den gröna vertikala linjen vid 0 på det bedömda stödbehovet. Individer till höger om den gröna linjen (med ett bedömt stödbehov>0) anvisas till KROM. Individer till vänster om den gröna linjen (med ett bedömt stödbehov<0) anvisas till eget jobbsökande.

Figur 5 visar jämförelsen mellan kvinnor och män. I figuren kan vi avläsa en relativt konstant skillnad på ca 3 procentenheter i genomsnittligt faktiskt stödbehov mellan kvinnor och män utmed hela skalan av bedömt stödbehov. Kvinnor förefaller alltså att ha ett något högre faktiskt stödbehov i genomsnitt än män vid varje givet värde av bedömt stödbehov. Vi drar slutsatsen att kriterium (A) – som modellen som estimerar jobbchans är konstruerad för att uppfylla – inte fullt ut kvarstår när vi utvärderar bedömningsstödet i sin helhet då vi jämför kvinnor och män.

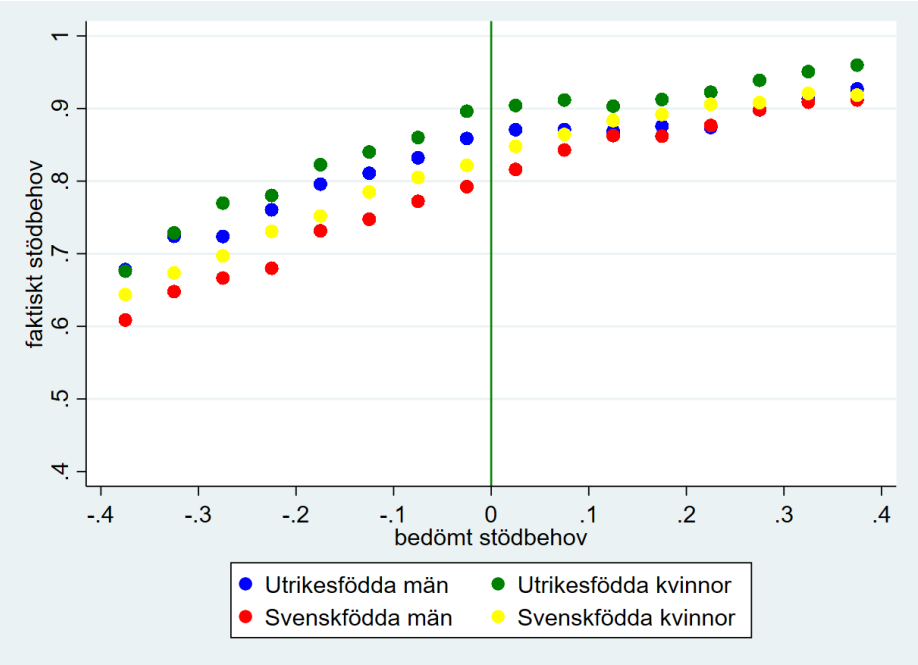
Figur 5. Andel utan jobb sex månader efter bedömningstillfället



Notera: *faktiskt stödbehov* är andelen individer i respektive grupp utan jobb (utan osubventionerad anställning) inom sex månader efter bedömningstillfället. *bedömt stödbehov* är *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*. *tröskelvärdet* är bedömningsgränsen för anvisning till KROM och markeras med den gröna vertikala linjen vid 0 på det bedömda stödbehovet. Individer till höger om den gröna linjen (med ett bedömt stödbehov>0) anvisas till KROM. Individer till vänster om den gröna linjen (med ett bedömt stödbehov<0) anvisas till eget jobbsökande.

I figur 6 visas jämförelser mellan kvinnor och män födda i Sverige och kvinnor och män födda i andra länder. Jämförelsen mellan alla kvinnor och män i figur 5 visade att kvinnor förefaller ha ett något högre faktiskt stödbehov i genomsnitt än män vid varje givet värde av bedömt stödbehov. Huvudintrycket i figur 6 är att detta mönster ser snarlikt ut även inom respektive födelsebakgrundsgrupp.

Figur 6. Andel utan jobb sex månader efter bedömningstillfället



Notera: *faktiskt stödbehov* är andelen individer i respektive grupp utan jobb (utan osubventionerad anställning) inom sex månader efter bedömningstillfället. *bedömt stödbehov* är *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*. *tröskelvärdet* är bedömningsgränsen för anvisning till KROM och markeras med den gröna vertikala linjen vid 0 på det bedömda stödbehovet. Individer till höger om den gröna linjen (med ett bedömt stödbehov>0) anvisas till KROM. Individer till vänster om den gröna linjen (med ett bedömt stödbehov<0) anvisas till eget jobbsökande.

Vi avslutar med analysen av likabehandlingskriterium (B). Här studerar vi om sannolikheten att anvisas till insatser inom KROM skiljer sig åt för sökanden från olika grupper, om man jämför sökande som faktiskt inte hittar jobb inom sex månader. I denna analys studerar vi alltså enbart sökande med ett stödbehov=1. I beräkningarna ingår alltså alla individer till höger om tröskelvärdet, samt de individer som felaktigt bedömts till vänster om tröskelvärdet (de som har ett faktiskt stödbehov, men som rekommenderats till eget jobbsökande). Resultaten visas i tabell 2.

Tabell 2. Likabehandling enligt kriterium B

Population	Chans att korrekt anvisas KROM om stödbehov=1
Män	0,67
Kvinnor	0,67
Ungdomar	0,47
Äldre än 50 år	0,72
Svenskfödda	0,54
Utrikes födda	0,76
Ej funktionsnedsatta	0,62
Funktionsnedsatta	0,87

Vi finner att kvinnor och män har samma chans, 67 procent, att korrekt rekommenderas anvisning till KROM, vid en jämförelse av sökande med ett dokumenterat stort faktiskt stödbehov. Det nuvarande bedömningsstödet genererar alltså rekommendationer som förefaller likabehandla män och kvinnor med stödbehov=1 (alltså enligt definition (B) ovan). Det är ett viktigt resultat eftersom ett uttalat krav för KROM är att kvinnor och män ges likvärdig tillgång till stöd.

Motsvarande jämförelser för övriga målgrupper visar att chansen att rekommenderas stöd är betydligt större för sökande som är äldre än 50 år (72 procent jämfört med 47 procent för ungdomar), för utrikes födda (76 procent jämfört med 54 procent för svenskfödda) och för sökande med funktionsnedsättningar (87 procent jämfört med 62 procent för övriga). Dessa resultat kan tolkas som att de ligger i linje med direktiven att rikta mer insatser till grupper som tenderar att stå längre ifrån arbetsmarknaden. I sammanhanget kan det vara viktigt att notera att individernas faktiska stödbehov troligen är ojämnt fördelat mellan olika grupper, även i en jämförelse av individer med ett dokumenterat stödbehov=1. Den genomsnittliga tiden till jobb, efter sex månaders arbetslöshet, är rimligen längre bland individer i grupper för vilka ett stödbehov=1

är vanligare.[13] För det fortsatta utvecklingsarbetet av Arbetsförmedlingens statistiska bedömningsstöd är detta en fråga att titta närmare på.

6 Avslutande diskussion

6.1 Vad har vi lärt oss?

Vi har studerat kvaliteten på de automatiserade anvisningarna till KROM, med fokus på träffsäkerhet och likabehandling i beslutsfattandet. Vi har funnit att besluten överlag har en god träffsäkerhet. Resultaten visar att bedömningsstödet ger korrekta anvisningar till nära 70 procent av de inskrivna arbetssökande.

Likabehandling är en mer komplex fråga. Utgångspunkten i vår analys har varit principen att sökande som liknar varandra i termer av faktiskt stödbehov ska behandlas på samma sätt, oavsett vilken grupp de tillhör. Vi har studerat hur bedömningsstödet till KROM presterar med avseende på två viktiga egenskaper, som ett bedömningsstöd enligt litteraturen ska ha, för att denna princip ska gälla.

Den första av dessa egenskaper innebär att *vid en jämförelse av sökande som fått samma bedömning ska sannolikheten att den sökande har ett faktiskt stödbehov=1 (arbetslös sex månader eller längre) vara oberoende av den grupp som individen tillhör* (egenskap A). Det betyder att arbetsförmedlare som använder bedömningsstödet får rätt incitament att behandla sökande med samma bedömning (eller rekommenderad anvisning) på samma sätt, snarare än att behandla sökande med samma bedömning (eller rekommenderad anvisning) olika baserat på den grupp de tillhör. Det senare kan vara en risk om bedömningsstödet inte uppfyller detta likabehandlingskriterium. Till exempel genom att oftare gå emot den rekommenderade anvisningen som bedömningsstödet ger när de möter sökande från en viss grupp.

Arbetsförmedlingen har säkerställt att bedömningsmodellen som bedömer de sökandes jobbchans uppfyller egenskap (A) i den driftsatta lösningen. Det framstår som ett viktigt första steg för att säkerställa likabehandling. Två faktum innebär dock att det finns utvecklingspotential: (1) bedömningsmodellen är framtagen för relativt nyinskrivna (men används även för de med längre inskrivningstider) och (2) i ett separat andra steg vägs den viktiga variabeln inskrivningstid in manuellt. I rapporten studerar vi om egenskap (A) kvarstår efter dessa steg, vi studerar alltså egenskap (A) för bedömningsstödet i sin helhet. Vi finner att egenskap (A) kvarstår i vissa gruppjämförelser, men inte fullt ut i andra gruppjämförelser. Hur stort problem det senare är för likabehandlingen av

[13] Stödbehov=1 (utan jobb efter sex månader) är med andra ord ett trubbigt mått på högt stödbehov. Det sanna stödbehovet är troligen högre för äldre, utrikes födda och i gruppen med funktionsnedsättningar, inom hela populationen med ett dokumenterat stödbehov=1. Om så är fallet skulle de skillnader i chanser till stöd som vi observerar vid gruppjämförelser av individer med stödbehov=1 minska om vi hade ett mått som mäter individernas sanna stödbehov perfekt.

sökande som idag anvisas KROM kan vi inte avgöra. Det beror på att vi studerar kvaliteten på bedömningarna och de rekommenderade anvisningarna med avseende på (A). Vi studerar inte de slutliga besluten om anvisning som fattas av arbetsförmedlare.

Den andra egenskapen innebär att *sannolikheten att den sökande korrekt anvisas KROM bland sökande med ett faktiskt stödbehov=1 är oberoende av den grupp som individen tillhör* (egenskap B). Vi finner att kvinnor och män har samma chans, 67 procent, att korrekt rekommenderas anvisning till KROM, vid en jämförelse av sökande med ett dokumenterat faktiskt stödbehov. Det nuvarande bedömningsstödet genererar alltså rekommendationer som förefaller likabehandla män och kvinnor med stödbehov=1. Det är ett viktigt resultat eftersom ett uttalat krav för KROM är att kvinnor och män ges likvärdig tillgång till stöd.

Motsvarande jämförelser för övriga målgrupper visar att chansen att rekommenderas stöd är betydligt större för sökande som är äldre än 50 år (72 procent jämfört med 47 procent för ungdomar), för utrikes födda (76 procent jämfört med 54 procent för svenskfödda) och för sökande med funktionsnedsättningar (87 procent jämfört med 62 procent för övriga). Dessa resultat kan tolkas som att de ligger i linje med direktiven att rikta mer insatser till grupper som tenderar att stå längre ifrån arbetsmarknaden (men det är inte strikt likabehandling enligt (B)). Notera igen att våra resultat gäller rekommenderade anvisningar, inte slutliga beslut om anvisning.

Det är också viktigt att notera att beslutsgränsen för anvisning till KROM i bedömningsstödet har flyttats två gånger efter att våra analyser genomfördes.[14] Resultaten kan se annorlunda ut efter dessa förändringar, särskilt som vilken hänsyn som tas till variabeln inskrivningstid i spårindelningen också har förändrats. En annan sak att notera är att vi förenklat studerar anvisningar till eget jobbsökande eller till KROM, vilket i vår studie även inkluderar anvisning till fördjupat stöd. Ett utvecklingsområde för fortsatta analyser är att ta hänsyn till distinktionen mellan KROM och fördjupat stöd.

Rapporten är en första kvalitetskontroll av anvisningarna inom KROM, och är en del av ett pågående utvecklingsarbete. Vi har lärt oss att den driftsatta lösningen av bedömningsstödet har tydlig utvecklingspotential, men också att kvaliteteten på de rekommenderade anvisningarna som bedömningsstödet genererar trots detta överlag är relativt god. I de efterföljande avsnitten ger vi förslag för det fortsatta utvecklingsarbetet och kvalitetsuppföljningen av Arbetsförmedlingens automatiserade bedömningar.

[14] Förändringarna har gjorts för att öka inflödet av sökande till KROM. Beslutsgränserna har alltså flyttats till vänster i figur 1.

6.2 Ändamålsenliga garantier för likabehandling

Likabehandling kan definieras på flera olika sätt, men bygger ofta på tanken att individer som liknar varandra i relevanta avseenden ska behandlas lika oavsett deras grupptillhörighet. En fördel med automatiserade bedömningar jämfört med manuella bedömningar är att det är möjligt att konstruera modellen så att den likabehandlar individer på ett förutbestämt sätt. Arbetsförmedlingen behöver göra ytterligare arbete för att implementera ändamålsenliga garantier för likabehandling i bedömningsstödet.

Notera att en ändamålsenlig garanti för likabehandling vid automatiserade anvisningar till KROM knappast kan vara att alla individer får en korrekt anvisning. Det skulle kräva att modellen inte gör några felbedömningar alls, vilket i dagsläget inte är realistiskt. En garanti för likabehandling på gruppnivå kan på motsvarande sätt knappast vara lika representation av olika grupper till varje typ av anvisning. En sådan form av likabehandling kan vara ändamålsenlig i flera sammanhang. Men den är inte ändamålsenlig för Arbetsförmedlingens verksamhet så länge som olika grupper skiljer sig åt i genomsnittligt stödbehov. Ändamålsenliga garantier handlar snarare om att säkerställa att modellens felbedömningar inte är systematiskt relaterade till individers grupptillhörighet. En rimlig ambition kunde vara att garantera ”att arbetssökande med liknande stödbehov ska ha lika möjligheter till en korrekt anvisning, oavsett grupptillhörighet”. Likabehandlingskriterier A, B och C som vi har studerat och diskuterat i rapporten är alla nära relaterade till en sådan ambition.

Det finns en stor litteratur om hur man praktiskt går till väga för att implementera likabehandlingsgarantier i modellerna. I den beskrivs också vad som är möjligt att uppnå, samt vilka avvägningar mellan olika garantier som behöver göras. Till exempel är det välbelagt att olika garantier för att säkerställa likabehandling alltid kommer att stå delvis i konflikt med varandra när man jämför grupper som skiljer sig avsevärt åt i genomsnittliga utfall.[15]

6.2.1 Fortsatta analyser behövs för att säkerställa likabehandling

För att säkerställa måluppfyllelse om likabehandling behöver myndigheten bestämma vilka målen är. Det handlar om att definiera principerna för likabehandling i anvisningarna till stöd, och att identifiera de garantier i litteraturen som kan användas för att säkerställa likabehandling enligt den önskade definitionen. Analyser kan vara nödvändiga för att studera vad som är möjligt att implementera utan att till exempel träffsäkerheten i anvisningarna försämras alltför mycket.

[15] Se Kleinberg, Mullainathan och Raghavan, 2016. Liknande målkonflikter har även uppmärksammats i en svensk myndighetskontext av Inspektionen för Socialförsäkringen (2018). I en granskning av Försäkringskassans riskbaserade kontroller av felaktiga utbetalningar diskuteras en rad olika definitioner av likabehandling som alltid står delvis i konflikt med varandra. I granskningen betonas behovet av rättsliga ställningstaganden i samband med användning av statistiska verktyg för urval.

6.3 Representativitet och modellering

6.3.1 Modellen bör tränas på den population den är tänkt att användas på

Varje bedömningsmodell kommer att fånga mönster i den population den är tränad/framtagen på. För att modellen ska prestera väl är det därför nödvändigt att kontrollera överrensstämmelsen mellan utvecklingspopulationen och tillämpningspopulationen redan vid utveckling av modellen. När samhället och arbetsmarknaden förändras behöver modellen också anpassas regelbundet med hjälp av en representativ population.

6.3.2 Överväg fördelarna med att inkludera inskrivningstid i modellen istället för separat korrigering

Det bedömningsstöd som används i det pågående försöket med KROM, och som beskrivs i figur 1 i denna rapport, består av ett system med två delar. Först sker en sannolikhetsbedömning av individens jobbchans utifrån ett trettiotal olika faktorer. Därefter görs en separat regelbaserad korrigering som tar hänsyn till individens tid som inskriven på Arbetsförmedlingen. Vid praktisk tillämpning av de garantier för likabehandling, som beskrivs i litteraturen, förutsätts det oftast att bedömningarna uttrycks som en sannolikhetsbedömning. Om Arbetsförmedlingen vill implementera garantier för likabehandling i bedömningar och anvisningar skulle det vara en fördel om bedömningsstödet var konstruerat på ett mer standardmässigt sätt. Det skulle i så fall innebära att inskrivningstid inkluderades från början i skattningen av individens jobbchans/stödbehov. Detta förslag bör vägas mot eventuella fördelar med den nuvarande konstruktionen av bedömningsstödet med två separata delar.

6.4 Förslag på rutin för kvalitetssäkring

Vi rekommenderar att en rutin för kvalitetssäkring införs. Syftet är att säkerställa arbetsmarknadspolitisk relevans (träffsäkerhet) och likabehandling vid användningen av automatiserade bedömningar och anvisningar. En sådan rutin bör innebära regelbunden kontroll och redovisning av att representativa urval används vid träning och testning av modellen. Samt kontroll och redovisning av måluppfyllelse av fastställda kriterier för träffsäkerhet och likabehandling. Rutinen bör användas både under utvecklandet och under det fortsatta användandet av automatiserade rekommendationer. Det senare bör vara särskilt viktigt i tider då ekonomin och arbetsmarknaden förändras i snabb takt. Ett förslag till rutin finns bilagd denna rapport (Bilaga 2)

Referenser

William Dieterich, Christina Mendoza and Tim Brennan (2016), *COMPAS risk scirkales: Demonstrating accuracy equity and predictive parity*. Technical report, Northpointe, July 2016. http://www.northpointeinc.com/northpointe-analysis.

Jon Kleinberg, Sendhil Mullainathan and Manish Raghavan (2016), "Inherent Trade-Offs in the Fair Determination of Risk Scores", arXiv:1609.05807

Alejandro Noriega-Campero, Michiel A. Bakker, Bernardo Garcia-Bulle och Alex Pentland (2018), "Active Fairness in Algorithmic Decision Making", arXiv:1810.00031v2

ISF – Inspektionen för Socialförsäkringen (2018): *Profilering som urvalsmetod för riktade kontroller*, Rapport 2018:5.

Lechner, M., & Smith, J. (2007), "What is the value added by caseworkers?" Labour Economics, 14(2), 135-151.

Ornstein P. & Thunström H. (2021). *Träffsäkerhet i bedömningen av arbetssökande. En jämförelse av arbetsförmedlare och en statistisk modell.* Arbetsförmedlingen.

Smith, J. A., Whalley, A., & Wilcox, N. (2020). "Are program participants good evaluators?" Working paper, University of Michigan.

Bilaga 1

Resultat från analys av testpopulationen

Tabell B1 visar fördelningen av inskrivningstid angiven i procent och i antal. I Panel A ser vi testpopulationen uppdelad i KROM-kommuner och övriga kommuner fördelad över tiden som inskriven på Arbetsförmedlingen. Vid en jämförelse av de procentuella fördelningarna ser vi att KROM-kommunerna har cirka 3 procentenheter högre andel individer med inskrivningstider över 2 år. Detta är ingen stor skillnad i sammanhanget och i övrigt ses inga skillnader att tala om. Fördelningarna är alltså nästan desamma i KROM-kommunerna som i övriga kommuner. Panel B visar KROM-kommunerna per 2020-03-18. Vi kan här se att fördelningen av inskrivningstider i KROM-kommunerna är mycket lik vad den var 2018-03-01. Det man dock notera att andelen nyinskrivna är något högre för

2020 års population och att de med en inskrivningstid mellan 1 och 2 år är cirka 6 procentenheter lägre.

Tabell B1

Panel A			**Tid inskriven**			
Ej KROM-kommum	0-3 Månader	3-6 Månader	6-12 Månader	1-2 år	Över 2 år	Totalt
Andel	16%	12%	15%	20%	38%	100%
Antal	61 388	43 980	54 562	74 082	141 243	375 255
KROM-kommun						
Andel	14%	11%	14%	20%	41%	100%
Antal	5 098	3 963	4 882	7 026	14 292	35 261
Total						
Andel	16%	12%	14%	20%	38%	100%
Antal	66 486	47 943	59 444	81 108	155 535	410 516
Panel B KROM-kommun 2020						
Andel	17%	13%	15%	14%	42%	100%
Antal	5 873	4 395	5 161	5 045	14 637	35 111

Tabell B2 visar andelarna och antalen individer som är kvinnor, har funktionsnedsättning, utrikes födda, unga (definierade som individer under 25 år gamla) och äldre (definierade som över 50 år gamla). I panel A ser vi återigen testpopulationen uppdelad i KROM-kommuner och övriga kommuner. Andelarna är tämligen likvärdiga mellan KROM-kommunerna och övriga kommuner. Vi kan dock notera att andelen funktionshindrade är cirka. 3 procentenheter högre i KROM-kommunerna och att andelen utrikes födda är cirka 4 procentenheter lägre i KROM kommunerna. I panel B visas samma uppgifter för KROM-kommunerna

per 2020-03-18. Vi kan här se att andelarna är närmast identiska för KROM-kommunerna över de båda tidpunkterna.

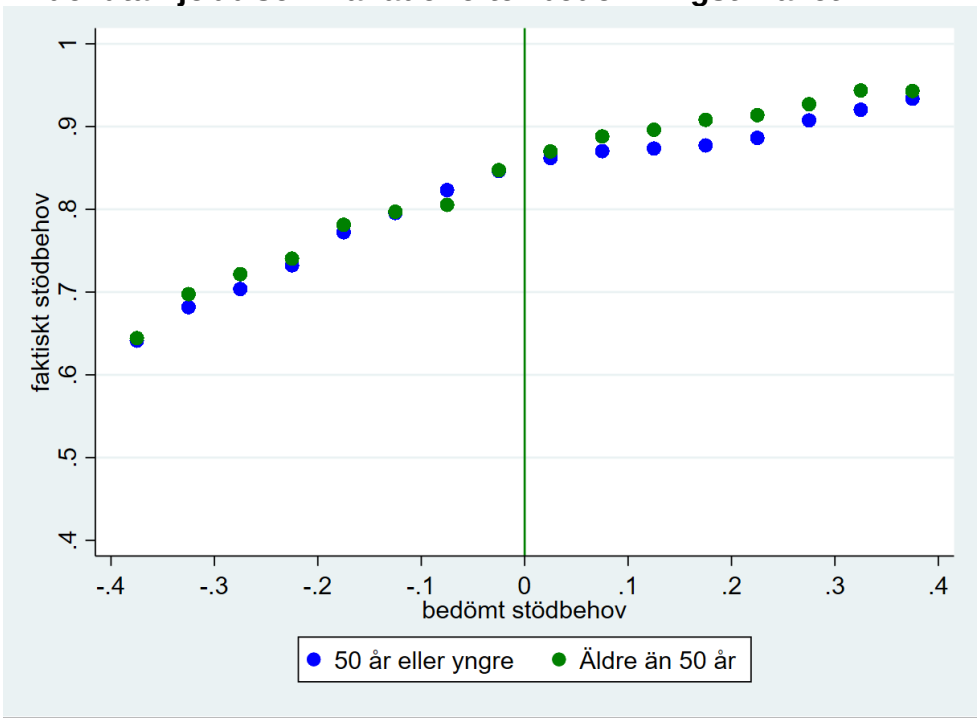
Tabell B2

Panel A					
Ej Kromkommun	Kvinnor	Funktionshindrade	Utrikesfödda	Ungdomar	Över 50 år
Andel	48%	19%	56%	14%	28%
Antal	178 571	70 912	211 384	52 907	105 939
Kromkommun					
Andel	46%	22%	52%	17%	26%
Antal	16 379	7 604	18 416	5 847	9 270
Total					
Andel	47%	19%	56%	14%	28%
Antal	194 950	78 516	229 800	58 754	115 209
Panel B					
Kromkommun 2020					
Andel	46%	20%	53%	15%	29%
Antal	16 259	6 934	18 491	5 213	10 031

Likabehandling, fler åldersgrupper

Figur B1 Personer äldre än 50 år jämfört med övriga

Andel utan jobb sex månader efter bedömningstillfället



Notera: *faktiskt stödbehov* är andelen individer i respektive grupp utan jobb (utan osubventionerad anställning) inom sex månader efter bedömningstillfället. *bedömt stödbehov* är *bedömt avstånd till arbetsmarknaden* minus *tröskelvärdet*. *tröskelvärdet* är bedömningsgränsen för anvisning till KROM och markeras med den gröna vertikala linjen vid 0 på det bedömda stödbehovet. Individer till höger om den gröna linjen (med ett bedömt stödbehov>0) anvisas till KROM. Individer till vänster om den gröna linjen (med ett bedömt stödbehov<0) anvisas till eget jobbsökande.

Bilaga 2: Utkast till kvalitetssäkringsrutin

1. När ska kvalitetsredovisning ske?
    a. Alltid inför nya tillämpningar av profilering.
    b. Med regelbundna tidsintervall vid kvarstående tillämpningar. (exempelvis var 6e till 12e månad)
    c. Alltid vid förändring av gränsvärde för rekommendation till anvisning.
    d. När modellen görs om.
2. Säkerställande av att testpopulation är så lik som möjligt den population som ska profileras.
    a. Relevant testpopulation innehåller information om arbetsmarknadsutfall samt är så stor som möjligt. Om testpopulationen understiger 20 000 individer ska motivering till detta bifogas.
    b. Relevant testpopulation har en sammansättning av individer som överensstämmer med den population som kommer vara föremål för profileringen. Fokus för dokumentationen om testpopulationen är på överensstämmelse avseende andelen i de målgrupper som analyseras i 3b.
        i. Målgrupp: under 25 år; över 54 år; kvinnor; utrikesfödda; dokumenterad funktionsnedsättning; arbetslöshetstid <90 dagar; arbetslöshetstid>365 dagar.
        ii. Tillåten avvikelse i andelar per målgrupp: mindre än 25% skillnad.
3. Säkerställande och dokumentation av kvalitet med avseende på:
    a. **Träffsäkerhet** mäts som summan av andelen som fått ett lyckat utfall och *inte* anvisats till insats, samt andelen som *inte* fått ett lyckat utfall och anvisats till insats.
        i. Lyckat utfall definieras som avaktualiserad till arbete och reguljära studier inom sex månader från profileringsdatumet. Om en annan definition av lyckat utfall används ska motivering till detta bifogas. Alla andra utfall ska definieras som inte lyckade, med undantag av avaktualiserade med orsak okänd (avors=6), som kan exkluderas ur undersökningen.
        ii. Godkänd uppmätt träffsäkerhet ska på totalen ligga på minst 65% samt överstiga den nivån som skulle nås vid slumpmässig tilldelning med samma andel positiva beslut.
    b. **Likabehandling** mäts utifrån principen att sökande som liknar varandra i termer av faktiskt stödbehov ska behandlas på samma sätt, oavsett vilken grupp de tillhör (till exempel kvinna eller man).