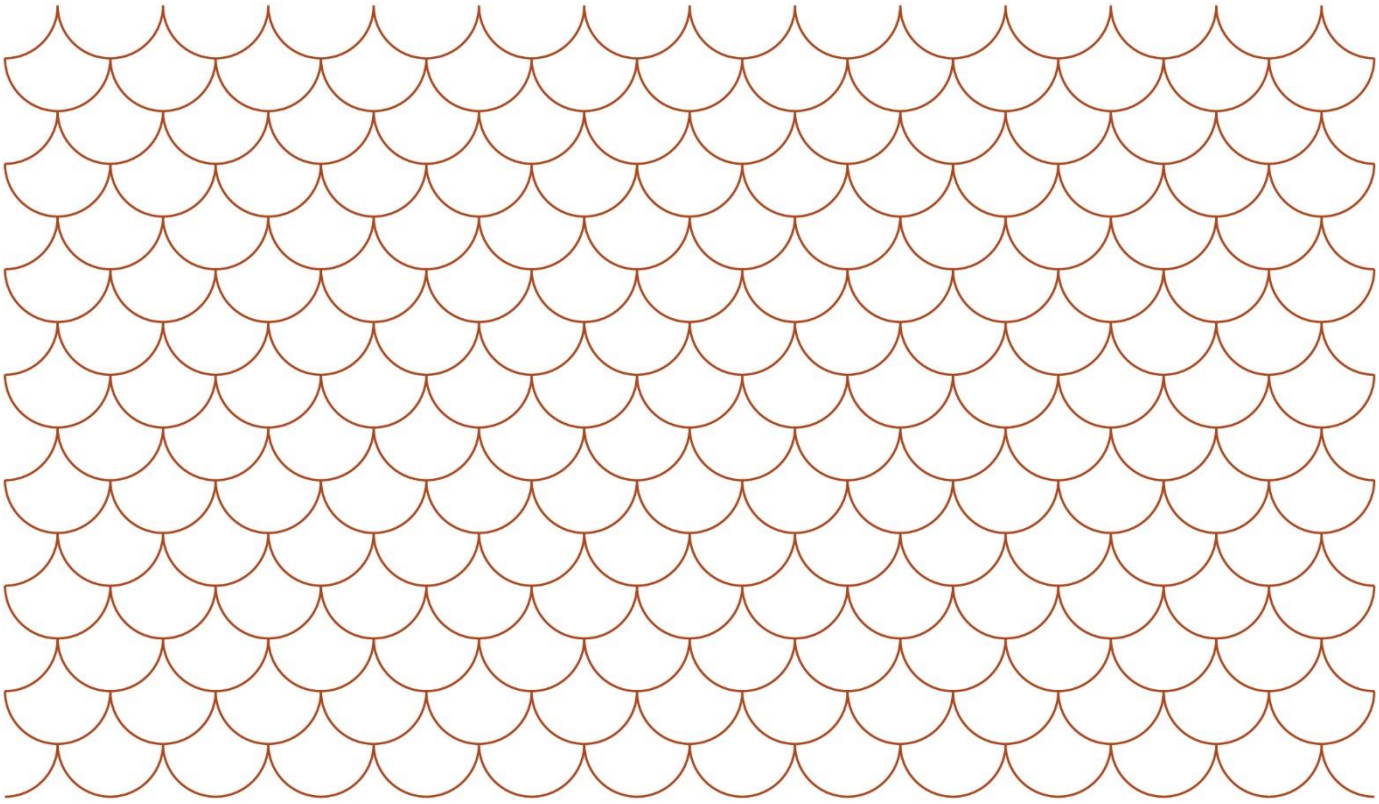
Utvärdering av den centrala rättningens och den externa bedömningens effekter på likvärdigheten i bedömningen

En del av försöksverksamheten med datorbaserade nationella prov, central rättning och extern bedömning



Skolverket

Publikationen finns att ladda ner som kostnadsfri
PDF från Skolverkets webbplats:
www.skolverket.se/publikationer

Dokumentdatum: 2024-02-07
Diarienummer: 2021:1673

Skolverket, Stockholm 2024

Innehåll

1. Sammanfattning

Skolverket har fått i uppdrag av regeringen att utvärdera hur central rättning och extern bedömning kan påverka likvärdigheten i bedömningen av de nationella proven inom ramen för försöksverksamheten.

Bakgrunden till uppdraget är dels rapporter från Skolverket om brister i likvärdigheten av betygssättningen, dels skillnader mellan lärares bedömning i Skolinspektionens ombedömningar av nationella prov. Skolverket har rapporterat att kraven som ställs för de olika betygsstegen skiljer sig åt mellan lärare. Ombedömningarna visade att olika lärare hade olika uppfattningar om vilka krav som skulle ställas på en elevs svar. Om någon annan än elevens betygssättande lärare bedömde elevernas svar på de nationella proven skulle de nationella provresultaten kunna utgöra ett externt utlåtande om elevens betygsnivå. Central rättning och extern bedömning är exempel på metoder där någon annan än elevens lärare bedömer elevens svar.

Central rättning innebär att bedömningen organiseras av en statlig myndighet och att bedömarna måste ha lärarlegitimation samt vara behöriga för ämnet och kursen eller årskursen. Extern bedömning innebär att någon annan än elevens betygssättande lärare bedömer de nationella proven. För att central rättning och extern bedömning skall fungera som ett externt utlåtande måste tillförlitligheten av bedömningen vara hög, vilket innebär att de centrala- eller externa bedömarna i hög grad är samstämmiga om hur elevsvaren skall bedömas. Skolverket utvärderade därför hur samstämmiga centrala respektive externa bedömare var i förhållande till lärare på några skolor när de bedömde elevers uppsatser i nationella prov.

Data till utvärderingen samlades in under januari och februari 2023. En förfrågan gick ut till rektorerna på de 100 försöksskolorna om de var intresserade av att delta i utvärderingen. Bland de skolor som tackade ja valdes tolv skolor ut för att delta med elever. Skolorna valdes ut eftersom de kunde bidra med tillräckligt många elever. Cirka 60 skolor bidrog med bedömare. Samtliga skolor som valdes ut deltog i försöket. Det förekom dock ett visst bortfall av bedömare.[1] En central bedömare och en extern bedömare för årskurs 9 saknades för att bilda fulla grupper. Två centrala bedömare för gymnasieskolan slutförde inte sin bedömning. En central bedömare för årskurs 6 genomförde inte sambedömningen. En extern bedömare för årskurs 9 genomförde sin individuella bedömning efter sambedömningen och kunde därför inte räknas med.

För varje årskurs bedömde ett tjugotal bedömare lika stor andel av elevernas uppsatser. Bedömarna var indelade i tre grupper:

[1] En central bedömare och en extern bedömare för årskurs 9 saknades för att bilda fulla grupper. Två centrala bedömare för gymnasiet slutförde inte sin bedömning. En central bedömare för årskurs 6 genomförde inte sambedömningen. En extern bedömare för årskurs 9 genomförde sin individuella bedömning efter sambedömningen och kunde därför inte räknas med.

- elevernas egna betygssättande lärare
- lärare från andra skolor – extern bedömning
- legitimerade och behöriga lärare som gått en bedömarutbildning hos Skolverket – central rättning.

Förutom sin andel fick alla bedömarna samma tio så kallade bedömarkontroller, det vill säga uppsatser som blandades med elevsvaren från skolorna. Dessa var uppsatser från elever som genomfört proven tidigare.

Bedömarna fick först bedöma sin andel och de tio bedömarkontrollerna individuellt. Sedan bedömde de om de tio texterna i par där de skulle komma överens om vilka poäng uppsatserna skulle få (sambedömning). Bedömarnas poäng på bedömarkontrollerna jämfördes med avseende på hur överens de var.

Bedömarna fick även svara på en webbenkät om deras upplevelse av utvärderingen. Skolverket intervjuade dessutom fyra av elevernas betygssättande lärare om tilliten till den centrala och externa bedömningen.

Utvärderingen visar:

- ingen statistiskt signifikant skillnad i likvärdighet mellan bedömningen av elevernas egna lärare och lärare från andra skolor
- ingen statistiskt signifikant skillnad i likvärdighet mellan bedömningen av bedömare som fått bedömarutbildning och de som inte fått det
- ingen statistiskt signifikant skillnad i likvärdighet mellan lärarnas individuella bedömning och deras sambedömning
- att bedömarnas upplevelse av central rättning genomgående var positiv
- att bedömarna genomgående ansåg att digital bedömning var enkelt och gick snabbt
- att flera bedömare menade att bedömningen skulle ha känts säkrare om man hade fått möjlighet att skriva ut texterna eller föra anteckningar i texten, även om tidsåtgången för bedömningen då skulle ha ökat.

Skolverkets utvärdering ger alltså inget stöd för att bedömningen av uppsatser i nationella prov blir mer likvärdig av att bedömaren är någon annan än elevens betygsättande lärare, det vill säga extern bedömning. Utvärderingen ger inte heller något stöd för att bedömningen blir mer likvärdig om bedömningen genomförs av legitimerade, behöriga lärare som genomgått en extra bedömarutbildning av Skolverket, det vill säga central rättning.

En förklaring till resultaten kan vara att de centrala bedömarna som ingick i studien inte kunde organiseras på samma sätt som Skolverket har föreslagit i en regeringsredovisning om hur central rättning kan organiseras.[2] Det vill säga genom team med tio bedömare som handleds och övervakas av en huvudbedömare, med en särskild utbildning. I utvärderingen delades bedömarna i

[2] Skolverket (2022). Redovisning av uppdrag om att införa central rättning av nationella prov. Dnr 2021:1559.

stället in i par som sambedömde uppsatserna utan någon handledning av en huvudbedömare.

En viktig erfarenhet från studien är att lärarna som ingick i studien uttryckte att de ser positivt på central rättning och att få samarbeta med kollegor från andra delar av landet och från skolor med olika elevunderlag. Sammantaget bedömde Skolverket att det borde vara möjligt att utveckla ett system med central rättning som både har en god likvärdighet i bedömningen och en hög legitimitet bland lärare i landet.

2. Inledning

I samband med att Skolverket fick i uppdrag av regeringen att digitalisera de nationella proven 2017 fick myndigheten även i uppdrag att upprätta en försöksverksamhet[3]. Försöksverksamheten skulle bestå av cirka 100 skolor, fördelade mellan olika skolformer. I försöksskolorna skulle olika moment för digitala nationella prov testas. Försöken skulle bland annat omfatta tester av extern bedömning[4], där någon annan än den undervisande läraren skulle bedöma elevens svar. Testerna skulle även omfatta så kallad medbedömning, där två bedömare oberoende av varandra bedömer samma elevs svar.

När Skolverket senare lämnade förslag på hur central rättning av digitala nationella prov skulle kunna införas ändrades uppdraget från regeringen.[5] Försöken i försöksverksamheten skulle nu omfatta tester av central rättning[6] medan tester av medbedömning togs bort.[7]

> 12 § Statens skolverk ska utvärdera försöksverksamheten. Av utvärderingen ska, så långt möjligt, den centrala rättningens och den externa bedömningens effekter på likvärdigheten i bedömningen av proven framgå …[8]

Inom ramen för försöksverksamheten med datorbaserade prov har Skolverket utvärderat om central rättning och extern bedömning har någon effekt på likvärdigheten[9] i bedömningen av uppsatsdelarna i de nationella proven i svenska

[3] Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning.
[4] Extern bedömning definieras i förordningen (2021:1061): bedömning som innebär att en lärare som inte är elevens undervisande lärare bedömer elevens lösning av ett nationellt prov, där resultatet av bedömningen ingår som ett underlag för den betygssättande läraren vid betygssättningen.
[5] U2021/03346.
[6] Central rättning definieras enligt förordningen (2021:1061): rättning som innebär att elevens lösning bedöms av en legitimerad och behörig lärare på uppdrag av en statlig myndighet, där resultatet av rättningen ingår som ett underlag för den betygssättande läraren vid betygssättningen
[7] Förordning (2021:1214) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning.
[8] Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning.

[9] I betydelsen av interbedömaröverensstämmelse, även kallad reliabilitet.

och engelska i grundskolan och i gymnasieskolan. I den här rapporten redovisas resultatet av utvärderingen.

Rapporten inleds med ett bakgrundskapitel, kapitel 2, vilket också innehåller en beskrivning av problemet med att uppnå en likvärdig bedömning och vilka frågor som behöver besvaras i utvärderingen. Därefter följer kapitel 3, med en detaljerad genomgång av metoder för datainsamlingen samt ett avsnitt om analysmetoder. Resultaten från utvärderingen presenteras i kapitlen 4 till 7 och diskuteras sedan i kapitel 8. Rapporten avslutas med kapitel 9 där resultatens implikationer för bedömningen av nationella prov och betygssättning diskuteras.

3. Bakgrund

I detta kapitel beskrivs det problem som central rättning eller extern bedömning skulle kunna lösa samt begränsningen av och motiveringen till de frågor som utvärderingen vill besvara.

De nationella proven har ett syfte: att stödja en likvärdig och rättvis betygssättning.[10] I Skolverkets allmänna råd om betyg och prövning[11] framgår att de nationella proven spelar en viktig roll när läraren analyserar och kalibrerar den egna betygssättningen. De nationella proven ger lärarna en gemensam referensram. Referensen består framför allt av resultat i form av nationella provbetyg. En lärare förväntas med andra ord kalibrera[12] sin betygssättning mot elevernas nationella provbetyg. Analyser ska även göras av såväl lärare som rektorer efter att betygen har satts. Ju större avvikelserna är, framför allt på gruppnivå, mellan betyg och nationella provbetyg, desto större är skälen att analysera om kravnivån för de olika betygsstegen som en lärare sätter upp behöver sänkas eller höjas. Lärare behöver vara observanta på avvikelser mellan provbetyg och betyg, särskilt systematiska avvikelser på gruppnivå. Sådana systematiska avvikelser kan indikera att resultat från de nationella proven inte särskilt har beaktats eller att läraren ställer krav för olika betygssteg som är högre eller lägre än de nationella provens.

Skolverket har i flera studier[13] visat att lärares betygssättning i förhållande till resultat på nationella prov skiljt sig åt beroende på skolornas prestationsnivå på proven. För elever som gick på skolor som presterade lågt på nationella prov var sannolikheten att få ett högre betyg än provbetyget större än för elever som gick på skolor som presterade högre på de nationella proven. Betygssättningen skiljde

[10] Förutom i årskurs 3 där syftet är att stödja bedömningen av uppfyllda kriterier för bedömning.
[11] Skolverkets allmänna råd (SKOLFS 2022:417) om betyg och prövning.
[12] Läraren sätter högre eller lägre betyg än vad hen hade tänkt göra från början, eftersom de nationella provresultaten indikerar att lärares krav varit för låga eller höga
[13] Skolverket (2019a). *Analyser av likvärdig betygssättning mellan elevgrupper och skolor*. Rapport 475. Skolverket (2020a). *Analyser av likvärdig betygssättning i gymnasieskolan. Jämförelser mellan kursbetyg och kursprov*. Rapport 2020:3.

sig även olika mycket åt beroende på ämne och framför allt för olika betygsgränser.

För att minska skillnaden mellan betyg och provbetyg infördes nya bestämmelser i skollagen 2018.[14] De innebär att om en elev genomfört ett nationellt prov ska resultatet från provet särskilt beaktas i betygssättningen. Begreppet ”särskilt beakta” betyder att resultatet på ett nationellt prov aldrig kan vara det enda underlaget vid betygssättningen, men det går inte att bortse från resultatet om det inte finns särskilda skäl. Även om resultat från ett nationellt prov har större betydelse än andra enskilda underlag när läraren värderar underlagens relevans kan det betyg som den enskilde eleven får skilja sig från provbetyget.

Hur väl ett nationellt prov fungerar för att stödja lärare att sätta mer likvärdiga betyg beror inte bara på hur de beaktar provresultatet utan också hur tillförlitligt provresultatet är. Skolinspektionen har, genom ombedömningar av nationella prov, visat att lärare kan vara oense om hur elevsvar på proven ska bedömas.[15] En lärare kan, till exempel, ha lägre krav för de olika betygsstegen än andra lärare i landet. Trots att bedömningen av nationella prov görs utifrån samma bedömningsanvisningar kan läraren även bedöma en elevs nationella prov till ett högre resultat än vad andra lärare skulle ha gjort. Resultatet från det nationella provet hjälper då inte läraren att hitta betygsnivåer som mer överensstämmer med de andra lärarnas krav. För att provbetygen ska kunna utgöra ett tillförlitligt underlag som är giltigt för kalibrering av lärares betygssättning, krävs alltså att de nationella proven bedöms på samma sätt, oavsett vem som bedömer dem.

Inom psykometri eller mätlära används begreppet *reliabilitet* för att beskriva tillförlitligheten i ett provresultat. De olika måtten på reliabilitet beskriver graden av *konsistens*, det vill säga i vilken utsträckning elevens resultat är oberoende av förändringar av till exempel provgenomförandet, provutgåvan[16] eller bedömaren. Ett provs giltighet, eller *validitet*, för ett visst syfte är alltid beroende av reliabiliteten. Om samma elev skulle få kraftigt varierande resultat vid olika provgenomföranden, på olika provutgåvor eller av olika bedömare vet vi inte vad provresultatet visar och det blir därför inte användbart för sitt syfte. Hög reliabilitet i nationella prov är därför viktigt för att provresultatet ska kunna användas för att betygssättningen mellan lärare i landet ska bli mer jämförbar och säker.

En aspekt av ett provs reliabilitet[17], är hur säker man kan vara på att bedömningen är rimlig. Om flera oberoende bedömare är samstämmiga i sin bedömning anses bedömningen vara mer reliabel. En huvudman eller rektor kan arbeta för att öka reliabiliteten för bedömning på skolan, till exempel genom att organisera sambedömning av nationella prov. Däremot ligger det utanför huvudmannens

[14] Proposition 2017/18:14
[15] T.ex. Skolinspektionen: Ombedömning av nationella prov 2019 Diarienummer 2019:503
[16] Begreppet ”prov” menas det som är kopplat till en kurs eller årskurs till exempel Engelska kurs 6. Begreppet ”provutgåva” avser en specifik version av provet till exempel Engelska kurs 6 vårterminen 2023. För de flesta elever är endast en provutgåva aktuell, men i kommunal vuxenutbildning har lärare möjlighet att välja mellan fler olika provutgåvor.
17 Ibland används *reliabilitet* för att beskriva ett provs inre konsistens. Begreppet beskriver då till vilken grad det är samma elever som presterar högre respektive lägre på olika uppgifter i provet, och blir därmed ett mått på hur konsistent provet mäter samma sak, det vill säga det som kallas provets *konstrukt*..

eller skolans kapacitet att harmonisera bedömningen över hela landet. En ökad reliabilitet i bedömningen av nationella prov på skolan betyder alltså inte att lärarnas bedömning stämmer överens med genomsnittet för alla Sveriges skolor. Lärarna på en skola skulle till exempel kunna enas om ovanligt låga krav när de bedömer elevsvaren, till exempel för att få ett visst antal poäng på en uppgift. En tänkbar åtgärd för att motverka detta skulle kunna vara att låta lärare från andra skolor och huvudmän bedöma elevernas nationella prov. Då finns möjligheten att lösningarna och svaren skulle bedömas av bedömare med högre krav för de olika nivåerna. Provresultaten skulle i dessa fall också kunna medverka till att lärarna på skolan blir uppmärksamma på att de ställer lägre krav för nivåerna än lärare på andra skolor.

När det ställs krav på att provresultatet särskilt ska beaktas vid betygssättningen kan lärare också ha intresse av att eleverna får provresultat som stämmer överens med de betyg som de hade tänkt sätta. Eftersom kravgränserna för de olika betygsstegen är kända för läraren vid bedömningen av elevernas lösningar och svar och eftersom lärarna själva ska sammanställa resultatet till ett provbetyg så har de möjlighet att leta efter ytterligare förtjänster i elevens lösningar och svar så att elevens resultat precis hamnar över gränsen för ett provbetyg.[18] Om man skulle låta någon annan än elevens egen betygssättande lärare bedöma provet skulle möjligheten att styra elevens resultat mot ett visst provbetyg försvinna.

I teorin skulle central rättning och extern bedömning kunna vara sätt att komma åt problemen med att krav för de olika betygsstegen för olika lärare skiljer sig åt. I teorin skulle även den möjliga intressekonflikten mellan myndighetsutövningen å ena sidan och lärarens intresse för att eleverna ska uppnå ett visst provbetyg å den andra, kunna hanteras. Metoderna skiljer sig dock åt, och därför behöver Skolverket utvärdera vinsterna i reliabilitet för metoderna separat. Följande definitioner av central rättning och extern bedömning används i förordningen[19]:

- central rättning: rättning som innebär att elevens lösning bedöms av en legitimerad och behörig lärare på uppdrag av en statlig myndighet, där resultatet av rättningen ingår som ett underlag för den betygssättande läraren vid betygssättningen,
- extern bedömning: bedömning som innebär att en lärare som inte är elevens undervisande lärare bedömer elevens lösning av ett nationellt prov, där resultatet av bedömningen ingår som ett underlag för den betygssättande läraren vid betygssättningen.

Extern bedömning innebär bara att någon annan än elevens betygssättande lärare bedömer elevens lösningar och svar. Det skulle kunna vara en kollega på samma skola som den betygssättande läraren. Så fungerar till exempel bedömningen i Skolverkets digitala provplattform; i plattformen samlas alla elevers svar och lösningar i en gemensam pool på skolan och bedömarna plockar slumpvis ut

[18] Diamond, R., and P. Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests". (No. w22207). National Bureau of Economic Research.
[19] Förordning (2021:1061) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning.

avidentifierade svar och lösningar för en uppgift i taget. Läraren vet då inte vilken elevs svar eller lösning som bedöms eller hur de poäng som läraren tilldelar svaret förhåller sig till poäng på andra lösningar och svar i elevens prov. Lärare på samma skola kan dock ha utvecklat ett liknande generöst bedömningsförfarande eller dela samma intresse för att eleverna på skolan ska få höga resultat. I dessa fall borde effekten av en sådan extern bedömning inom skolan bli liten. Den externa bedömningen skulle även kunna genomföras av lärare på andra skolor eller till och med hos andra huvudmän där chansen är mindre att bedömaren delar den betygssättande lärarens eventuellt avvikande krav eller intressen för att eleverna ska få specifika provresultat.

En ny utmaning som uppstår är att det inte finns något i beskrivningen av extern bedömning som reglerar den externa bedömarens kompetens, till exempel krav på legitimation eller behörighet. Om bedömningen genomförs av lärare på en annan skola eller hos en annan huvudman, vet den betygssättande läraren alltså inget om kompetensen eller noggrannheten hos bedömaren. Därför kan läraren bli osäker på hur tillförlitlig bedömningen blivit.

Den centrala rättningen ska vara organiserad av en statlig myndighet, enligt regeringsuppdraget om att införa central rättning. Då kan staten ställa krav på att bedömarna är legitimerade samt behöriga för ämnet och kursen eller årskursen. Myndigheten har även möjlighet att garantera viss kvalitet på bedömningen genom att erbjuda utbildning för och kvalitetskontroller av bedömarna.

3.1 Frågor att besvara inom ramen för uppdraget

Enligt uppdraget ska Skolverket utvärdera, så långt möjligt, den centrala rättningens och den externa bedömningens effekter på likvärdigheten i bedömningen av proven.

Det är viktigt att påpeka att Skolverket inte har utvärderat effekterna av central rättning och extern bedömning på *betygssättningen*. Skollagen ger läraren utrymme att själv bedöma hur det nationella provresultatet ska användas som betygsunderlag. Hur likvärdig betygssättningen blir till följd av en mer reliabel bedömning av proven beror alltså på hur provresultaten beaktas av den betygssättande läraren. Det skulle i och för sig vara möjligt att mäta om avvikelserna mellan provbetyg och betyg förändras efter ett införande av central rättning eller extern bedömning. Det går dock inte att urskilja orsaken till förändringen, exempelvis om den har berott på att lärarnas tillit till den centrala rättningen och externa bedömningen har gjort dem mer eller mindre benägna att beakta provbetyget i sin betygssättning. Särskilt inte då även annat ändras vid digitaliseringen av de nationella proven

Utvärderingen har inte haft som mål att uttala sig om hur skolors egenskaper påverkar reliabiliteten i lärarnas bedömning, Till exempel skulle en stor andel av landets skolor behövt delta i utvärderingen för att undersöka om storlek på

huvudman eller typ av huvudman spelade roll för reliabiliteten i bedömning av nationella prov.

Utvärderingen omfattar enbart uppgifter i skriftlig produktion i engelska och svenska. Resultatet från delprovet som prövar skriftlig produktion utgör endast en fjärdedel av provresultatet i engelska, en tredjedel av provresultatet i svenska och svenska som andraspråk i årskurs 6 och årskurs 9 samt lite mer än hälften av provresultatet i svenska och svenska som andraspråk i gymnasieskolans kurs 3. Variationen i bedömningen av detta delprov får därför även olika effekt för olika provbetyg.

Eftersom provresultat från central rättning och extern bedömning ska utgöra ett externt utlåtande till den betygssättande läraren[20], behövde reliabiliteten för central rättning och extern bedömning utvärderas i relation till den undervisande lärarens egen bedömning. Därför ingick en grupp lärare som bedömer sina egna elevers lösningar och svar också i utvärderingen.

Det är viktigt att komma ihåg att vad som är det sanna resultatet på ett prov i uppsatsskrivning är omöjligt att avgöra. Ofta brukar man anse att ju fler som bedömer lika, desto bättre är reliabiliteten i den samlade bedömningen. Om bedömningarna varierar mycket mellan bedömare är reliabiliteten lägre än om bedömarna är mer samstämmiga. Utvärderingen omfattar alltså en jämförelse av reliabilitet inom central rättning, inom extern bedömning samt mellan bedömning gjord av elevernas egna lärare.

Frågor att besvara i utvärderingen:

1. Vilken skillnad i reliabilitet går det att se mellan central rättning, extern bedömning och bedömning av elevernas betygssättande lärare?
2. Vilken effekt skulle eventuella skillnader mellan olika bedömares bedömning av uppsatserna kunna få på elevers provbetyg?
3. Hur uppfattar centrala bedömare respektive externa bedömare att det är att bedöma jämfört med lärare som bedömer sina egna elevers lösningar?
4. Vilka skillnader finns i upplevelserna av bedömningen mellan lärare som bedömde svenska årskurs 6, engelska årskurs 9 respektive svenska i gymnasiets kurs 3?

Vilken tillit hade de betygssättande lärarna till den bedömning som genomförts som central rättning och extern bedömning?

[20] Läraren förväntas kalibrera sin betygssättning utifrån ett externt provbetyg.

4. Metod för insamling av data

I det här kapitlet beskrivs hur data samlats in för att besvara de olika frågorna i utvärderingen.

Data till utvärderingen samlades in under januari och februari 2023. Arbetet genomfördes av två arbetsgrupper. Den första arbetsgruppen bestod av personer från enheten för allmändidaktik och skolans digitalisering på Skolverket. Gruppen ansvarade för att kontakta och informera deltagare från olika skolor om utvärderingens olika delar och vilka förväntningar som Skolverket hade på deltagarna. Arbetsgruppen var också tillgänglig för att besvara deltagarnas frågor under genomförandet. Den andra arbetsgruppen bestod av personer från enheten för nationella prov och enheten för internationella studier. Gruppen ansvarade för att utforma och genomföra utvärderingen. Den har också gjort beräkningarna och sammanställt den här redovisningen.

Eftersom de nationella proven inte ska genomföras digitalt förrän höstterminen 2024 och under 2025 var det inte möjligt att utvärdera bedömningen av obligatoriska nationella prov. I stället behövde provsituationen modelleras.

Tabell 1. Forskningsfrågor och datainsamlingsmetod

Utvärderingsfråga	Datainsamlingsmetod
1. Vilken skillnad i reliabilitet går det att se mellan central rättning, extern bedömning och bedömning av elevernas betygssättande lärare?	Insamling av bedömda uppsatser för jämförelse av olika bedömares poäng på samma uppsats
2. Vilken effekt skulle eventuella skillnader mellan olika bedömares bedömning av uppsatserna kunna få på elevers provbetyg?	Utifrån skillnader mellan olika bedömares poäng resonera om konsekvenser för elevens provbetyg
3. Hur uppfattar centrala bedömare respektive externa bedömare att det är att bedöma jämfört med lärare som bedömer sina egna elevers lösningar? 4. Vilka skillnader finns i upplevelserna av bedömningen mellan lärare som bedömde svenska årskurs 6, engelska årskurs 9 respektive svenska i gymnasiets kurs 3?	Enkät till samtliga bedömare

Utvärderingsfråga	Datainsamlingsmetod
5. Vilken tillit hade de betygssättande lärarna till den bedömning som genomförts som central rättning och extern bedömning?	Intervjuer med fyra undervisande lärare som deltagit med sina elever i utvärderingen

4.1 Datainsamling för utvärdering av reliabilitet

För att kunna utvärdera skillnaden i reliabilitet mellan central rättning, extern bedömning och bedömning av elevens betygssättande lärare behöver bedömarna bedöma samma elevlösningar eller elevsvar. I förordningen för försöksverksamheten[21] fanns emellertid ett krav om att resultatet av bedömningen skulle ingå som underlag i betygssättningen. Kravet kan dels motiveras med möjligheten att kunna utvärdera undervisande lärares tillit till resultatet som betygsunderlag, dels med att det borde öka intresset för elevernas resultat bland bedömarna. Utvärderingen behövde därför bestå av unika elevlösningar och svar från eleverna som skulle få betyg samt lösningar och svar som alla bedömare skulle bedöma för utvärderingens jämförelser.

Bedömningen genomfördes i Skolverkets provplattform, för att utvärderingen skulle bli så autentisk som möjligt i förhållande till den miljö som lärarna sedan kommer att bedöma digitala nationella prov i. Alla bedömare som deltog fick en utbildning i hur bedömningsmodulen i Skolverkets digitala provplattform fungerade[22] och de hade tillgång till en manual för bedömningsgränssnittet.

4.1.1 Urval av ämne, årskurser och kurser

I Skolverkets redovisning av uppdraget att införa central rättning av nationella prov[23] föreslår myndigheten att endast de provdelar som prövar skriftlig produktion i engelska, svenska och svenska som andraspråk i årskurs 9 samt kurs 6 i engelska och kurs 3 i svenska och svenska som andraspråk i gymnasiet ska bedömas centralt. Därför valde Skolverket ut den del av provet där eleverna får skriva en uppsats, i engelska för årskurs 9 samt svenska för gymnasiets kurs 3. Svenska som andraspråk valdes inte ut, främst för att elevunderlaget bedömdes vara för litet bland de försöksskolor som deltog. Avgörande var även att utvärdering av bedömning både i svenska och svenska som andraspråk skulle innebära praktiska utmaningar med parallella provutgåvor och bedömningsanvisningar och därmed en omständlig administration.

Som jämförelse valde Skolverket även ut uppsatsskrivning i svenska i årskurs 6. Det motiverades dels med att Skolverket behövde sprida deltagandet i

[21] 2017:1106.
[22] Provgenomförandet administrerades av Skolverket och lärarna på skolan bedömdes inte behöva någon instruktion i genomförande av provet. Eleverna fick se en film om hur elevgränssnittet för provuppgiften fungerade och hade en manual under genomförandet.
[23] Skolverket: Redovisning av uppdrag om att införa central rättning av nationella prov U2021/03346.

utvärderingen över många försöksskolor, dels med att Skolverket i redovisningen resonerat om att inkludera årskurs 6 i central rättning.

4.1.2 Urval av skolor

Skolverkets försöksverksamhet för digitala nationella prov omfattar 100 skolor. De utsedda skolorna får statsbidrag för att delta i olika försök under digitaliseringen av det nationella provsystemet. Försöksskolorna utgör exempel på olika skolformer och årskurser och består av skolenheter med olika storlek. En skola kan bestå av flera skolenheter där endast en av enheterna är utsedd för försöksverksamheten. Eftersom regeringen ställt krav på representation av olika egenskaper för skolan är de inte helt slumpvis utvalda. De består dock inte av skolor som frivilligt anmält sig för deltagande.

Eftersom utvärderingen skulle genomföras inom ramen för försöksverksamheten, var det inte möjligt ställa krav på att urvalet skulle representera den variation av skolor som finns i Sverige eller frekvensen av skolor av en viss typ[24] som finns i Sverige.

4.1.3 Urval av lärare till lokal bedömning

En förfrågan gick ut till rektorerna på de 100 försöksskolorna om vilka av skolorna som var intresserade av att delta i utvärderingen. Bland dessa valdes sedan fyra skolenheter ut för varje prov. De utvalda skolorna bedömdes kunna bidra med tillräckligt många elever för att fördelas i lagom stora kvoter mellan olika bedömare.[25] Sammanlagt behövdes cirka 250 uppsatser per prov[26] för att det skulle finnas utrymme för bortfall. En tredjedel av uppsatserna på skolan skulle bedömas av elevernas egna lärare på skolan. Dessa skolor ålades därför även att vardera ställa upp med två av elevernas lärare för bedömning. Denna bedömning av elevernas egna lärare benämns i fortsättningen *lokal bedömning*. Det fanns inga krav på legitimation eller behörighet för att delta i den lokala bedömningen.[27]

4.1.4 Urval av lärare till central rättning

För varje prov valdes sexton skolor ut bland de övriga intresserade för att vardera bidra med en bedömare till central rättning eller extern bedömning. Åtta av dessa bedömare placerades i en central rättningsgrupp. Dessa var alla legitimerade samt behöriga för undervisning i det aktuella ämnet och kursen eller årskursen. De genomgick även en utbildning en dag, utformad efter den norska *Sensorskoleringen* för bedömning av deras examensprov. I utbildningen fick bedömarna ett antal elevtexter från den aktuella provuppgiften att bedöma inför utbildningen. Utfallen från dessa bedömningar samlades in och sammanställdes av Skolverket. Sedan höll ett undervisningsråd från Skolverket en kort föreläsning om bedömning av det aktuella provet. Efter det delades bedömarna in i grupper

[24] Högpresterande – lågpresterande, storstad - landsbygd o.s.v.
[25] Se avsnittet datainsamlingsmetod.
[26] Ibid.
[27] Behörighet och legitimation varierade mellan skolform och ämne, men underlaget var för litet för att analyseras med avseende på denna faktor.

där de fick diskutera sina bedömningar av elevtexterna från den aktuella provuppgiften och jämföra dem med bedömningen som en referensgrupp, på uppdrag och inom provets framtagning av lärosätet, gjort. Passet avslutades med en gemensam diskussion om bedömningarna av de aktuella elevtexterna. En tredjedel av uppsatserna gick till dessa bedömare som alla var legitimerade, behöriga och särskilt utbildade för uppgiften. Bedömningen som genomförs av lärarna i denna grupp benämns i fortsättningen *central rättning*.

4.1.5 Urval av lärare till extern bedömning

De åtta återstående bedömarna placerades i en grupp som i fortsättningen benämns *extern bedömning*. I regeringens definition av extern bedömning sägs bara att någon annan än elevens undervisande lärare ska bedöma elevens lösningar och svar. All bedömning i Skolverkets digitala provplattform är dock avidentifierad och slumpvis fördelad mellan bedömare på skolenheten. Därigenom skulle all bedömning på skolan i provplattformen kunna sägas fungera som extern bedömning.[28] För att skilja extern bedömning från bedömningen som utfördes av lärarna på skolan fördelades den sista tredjedelen av uppsatserna till bedömare från andra skolor och huvudmän.[29] Det fanns inga krav på legitimation eller behörighet för att delta i den externa bedömningen.[30]

4.1.6 Insamling av uppsatser

De prov som användes var frisläppta nationella provuppgifter. Eleverna var medvetna om att det inte var ett riktigt nationellt prov, men att deras resultat skulle ingå i lärarens betygsunderlag. Eleverna fick inte välja uppsatsämne i svenska 3, som de normalt kan göra, eftersom alla bedömare behövde bedöma samma uppsatsämne. Provuppgifterna och uppsatsämnet valdes ut i samråd med Göteborgs universitet och Uppsala universitet, som utvecklar proven i engelska respektive svenska.

De inlämnade uppsatserna fördelades sedan jämnt mellan bedömarna. Förutom sin kvot av inlämnade uppsatser fick alla bedömare även tio uppsatser, så kallade *bedömningskontroller*. Bedömningskontrollerna var autentiska uppsatser som skickats in av lärare efter genomförandet av nationella prov. De olika uppsatserna representerade texter bedömda från låga till höga nivåer. Fördelningen var jämn över olika nivåer, snarare än att ha tyngdpunkten på de vanligaste förekommande nivåerna. Bedömningskontrollerna blandades in i lärarnas kvot, och även om lärarna hade informerats om att det skulle förekomma bedömningskontroller i kvoten hade de ingen möjlighet att identifiera dem i flödet.[31] Det var bedömningen av dessa bedömningskontroller som användes för att utvärdera reliabiliteten.

[28] Visserligen kan en bedömare slumpvis tilldelas sin egen elevs svar, men eftersom bedömaren saknar möjlighet att identifiera eleven bakom svaret blir det samma sak som om hen hade bedömt sin kollegas elevs svar.
[29] Ingen huvudman deltog med flera skolenheter i samma årskurs och skolform.
[30] Behörighet och legitimation varierade mellan skolform och ämne, men underlaget var för litet för att analyseras med avseende på denna faktor.
[31] Två lokala bedömare på samma skola skulle i och för sig kunna sätta sig och jämföra alla tilldelade uppgifter för att hitta de som var gemensamma, men då frångick de instruktionen.

En faktor som i forskning har visat sig påverka reliabilitet är om bedömarna sitter ensamma eller om det förekommer något slags sambedömning mellan dem.[32] Det fanns en risk att de lokala bedömarna skulle sambedöma sina elevsvar medan de externa bedömarna skulle bedöma texterna enskilt och att jämförelsen mellan central, extern och lokal bedömning därför inte skulle bli likvärdig. För att kontrollera för denna variabel delades datainsamlingen in i två delar. Samtliga bedömare fick först bedöma och lämna in sin kvot med inskickade elevuppsatser och bedömningskontroller enskilt. Sedan gjordes bedömningen av bedömningskontrollerna om, genom sambedömning i par. Den lokala bedömningen i par genomfördes av två bedömare från samma skola. Inom grupperna extern bedömning respektive central rättning delades de åtta bedömarna in i fyra par. Centrala bedömare parades endast ihop med andra centrala bedömare och externa bedömare endast med andra externa bedömare. Sambedömningen förenklades genom att bedömarna redan hade bedömt uppsatserna en gång, men det innebar också att det uppstod ett beroende mellan sambedömningen och den individuella bedömning som föregått sambedömningen. Sambedömningen blev på så sätt en förhandling om lärarna skulle ändra sin första poängsättning snarare än hur paret förhöll sig till kvaliteten av en helt ny uppsats.

I Skolverkets förslag om central rättning beskrivs en organisation med team handledda och övervakade av en huvudbedömare. Hur sambedömningen i teamen går till skulle säkert variera från team till team, men det borde vara vanligt förekommande att bedömare inom teamet väljer att dela svårbedömda uppsatser med en teammedlem eller huvudbedömaren efter att ha gjort ett försök till bedömning själva. Troligtvis är det i dag också vanligt att lärare på skolor ber en kollega om hjälp att medbedöma en svårbedömd text. Upplägget med parvis sambedömning efter individuell bedömning borde därför vara autentisk.

Det var nödvändigt att göra vissa förändringar av uppgifter och bedömnings-anvisningar i förhållande till det frisläppta nationella prov som uppgiften utgick ifrån. I originalen av bedömningsanvisningen används nämligen betygsbeteckningar för att bedöma kvaliteten av texterna. Detta skapade både problem för det digitala formatet, och för beräkningarna av reliabiliteten. Dessutom stämmer inte praktiken med att beskriva kvalitet på svar i uppgifter som betyg med Skolverkets uppdaterade allmänna råd om betygssättning[33], betygsbeteckningar bör endast användas vid betygssättning. De nivåer av kvalitet som beskrevs i bedömningsanvisningarna omvandlades därför till poäng.

När utvärderingen genomfördes var det inte helt klart hur poängmodellerna för digitala nationella prov i engelska respektive svenska skulle utformas. Därför skapades enkla modeller inför utvärderingen[34]:

[32] Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education, 41*(4), 638–653.
[33] Skolverkets allmänna råd (SKOLFS 2022:417) om betyg och prövning.
[34] Omvandlingen av betyg till poäng kan ha påverkat reliabiliteten. Under bedömarutbildningen uttryckte bedömarna ibland hur de använt sin egen uppfattning av betygsnivåerna, snarare än att referera till bedömningsanvisningens beskrivning av nivåerna eller de bedömda exempeltexterna de hade tillgång till. Risken var därmed att de tidigare har

- I svenska genomfördes, en så kallad, analytisk bedömning där läraren gav noll till tre poäng på tre till fyra olika aspekter. Dessa poäng lades sedan ihop automatiskt av provtjänsten till en delprovspoäng. För varje aspekt uppgavs det en beskrivning för varje poängnivå och bedömaren hade också tillgång till ett antal bedömda och kommenterade exempeltexter som stöd.
- I engelska genomfördes, en så kallad, holistisk bedömning där bedömaren satte en sammanfattande delprovspoäng från ett till nio direkt på elevtexten utifrån jämförelser med ett antal redan bedömda och kommenterade elevtexter.

De deltagande bedömarna informerades om att modellerna endast gällde för bedömning av prov inom ramen för utvärderingen.

4.1.7 Begränsningar vid datainsamlingen av uppsatser

Från gymnasieskolan skickade 303 elever in uppsatser i svenska kurs 3. Från årskurs 9 skickade 315 elever in uppsatser i engelska. Det var i båda fallen tillräckligt för att fördela uppsatserna mellan de 24 bedömarna och blanda in de tio bedömningskontrollerna i varje lärares kvot. Från årskurs 6 lyckades Skolverket endast få in 166 uppsatser i svenska. Det innebar att det blev för få uppsatser[35] att inkludera alla tio bedömningskontroller och bedömarna från årskurs 6 fick därför bara bedöma fem bedömningskontroller.

Tilldelningsfunktionen i Skolverkets provplattform fungerade inte riktigt som det var beskrivet i användarmanualen från leverantören. Det innebar att antalet uppsatser inte alltid fördelades jämnt mellan bedömarna. På grund av den ojämna fördelningen fick fyra bedömare för engelska årskurs 9, bara sex av de tio bedömningskontrollerna.

Det förekom ett visst bortfall av bedömare i utvärderingen. En central bedömare och en extern bedömare för årskurs 9 saknades för att bilda fulla grupper. Två centrala bedömare för gymnasiet slutförde inte sin bedömning. En central bedömare för årskurs 6 genomförde inte sambedömningen. En extern bedömare för årskurs 9 genomförde sin individuella bedömning efter sambedömningen och kunde därför inte räknas med.

Antalet bedömare per grupp som till slut ingick i analysen redovisas i resultatkapitlet.

4.1.8 Begränsningar i utvärderingens representativitet

Eftersom bedömningsanvisningarna behövde justeras för det digitala formatet kan inte reliabilitetsmåtten i utvärderingen användas för att säga något om hur reliabiliteten på de frisläppta nationella proven, som låg till grund för uppgifterna,

reproducerat en relativ uppfattning av betygsnivåerna i sin bedömning av uppsatser i nationella prov. Den modell med poäng som användes i utvärderingen var ny för alla. Det var dock inget som talade för att de tre grupperna, centrala-, externa- respektive lokala bedömare, skulle tolka poängen olika.

[35] De tio bedömningskontrollerna skulle komma att utgöra för stor del av varje bedömares kvot och skulle komma med en frekvens som var högre än varannan uppsats.

skulle kunna se ut.[36] I utvärderingen användes inte samma bedömningsskala[37] som användes i de prov som uppgifterna ingick i eller den poängskala[38] som kommer användas vid bedömning av skriftlig produktion i de kommande digitala nationella proven.

Uppdraget innebar vidare att Skolverket var tvungen att göra ett bekvämlighetsurval av såväl bedömare som omfattningen av bedömningskontroller. Hur väl urvalet representerade lärarkåren och elevtexter i Sverige går inte att uppskatta. Det finns till exempel studier som argumenterar för att en bedömarutbildning inte borde vara lika för alla, utan borde utformas för att rikta sig mot bedömare med olika erfarenhet, till exempel avseende hantering av specifika utmaningar med bedömning av texter av olika kvalitet.[39]

Bedömningen gjordes slutligen inte heller i samband med obligatoriskt genomförda nationella prov. Såväl elever som lärare kan därmed ha upplevt att de inte hade så mycket att förlora på dåliga resultat på uppgiften. Skolverkets bedömning är att bedömarna i utvärderingen ansträngde sig för att göra en noggrann bedömning av uppsatserna. Om eleverna ansträngde sig för att skriva bra uppsatser spelar däremot ingen större roll för utvärderingen, eftersom alla mätningar gjordes på bedömningskontrollerna, det vill säga på inlämnade uppsatser från ett genomfört nationellt prov, inte på uppsatserna skrivna av försöksskolorna.

4.2 Datainsamling om upplevelsen av bedömning

Skolverket skulle även undersöka hur centrala bedömare respektive externa bedömare uppfattade hur det var att bedöma, jämfört med lärare som bedömde sina egna elevers lösningar. Alla bedömare fick en webbenkät med frågor om lärarnas upplevelse av att genomföra bedömning centralt, externt respektive lokalt. Den innehöll också frågor om lärarnas upplevelse av att bedöma texter i kontexten för utvärderingen.

4.2.1 Datainsamling genom enkät till samtliga bedömare

Lärares erfarenheter av att bedöma uppsatser från elever som de inte själva undervisar varierade. I några skolor förekom organiserad sambedömning där alla elevers texter blandas och fördelas slumpvis mellan bedömare. På andra skolor bedömde lärare sina egna elevers uppsatser och diskuterade endast svårbedömda texter med en kollega.

[36] Gustafsson, J. E., & Erickson, G. (2013). To trust or not to trust? - Teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*, 69–87.
[37] Kvalitet beskrevs då direkt i termer av betyg.
[38] Var inte känd för Skolverket under utvärderingen.
[39] Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors? *Language Testing, 37*(3), 311–332.

Det var första gången lärarna var inne i den digitala provplattformen och förmodligen även första gången de bedömde uppsatser med poäng i stället för betyg.

De olika organisationerna av central och extern bedömning samt förändringarna med nya digitala bedömningsförfaranden, som byggde på poäng, skulle kunna påverka deltagarnas upplevelse av likvärdighet i sin bedömning. I direkt anslutning till att bedömningen av elevtexterna var genomförd ombads lärarna därför att besvara en enkät om dessa faktorer.

Valet av kvantitativ enkätundersökning som metod motiverades av behovet att få svar på frågor[40], både av fakta- och åsiktskaraktär, från så många av de deltagande lärarna som möjligt. De flesta frågorna hade givna svarsalternativ där läraren skulle ta ställning till ett påstående. Samtidigt skulle ett par frågor också besvaras med egen text.[41] Där kunde bedömarna dels ange positiva och negativa erfarenheter av att genomföra bedömningen digitalt, dels skriva ner övriga synpunkter. De centrala bedömarna fick också tre frågor om utbildningen som de besvarade innan bedömningen startade.

60 av de 70 bedömare som genomfört försöket i sin helhet besvarade enkäten. De fick ange om de hade varit en lokal, extern eller central bedömare och i vilken årskurs/kurs och ämne som de genomfört bedömningen i. De fick även tala om hur lång erfarenhet de hade av att bedöma nationella prov i det aktuella ämnet och årskursen eller kursen.

4.3 Datainsamling om tillit till bedömningen

Elevernas uppsatser som bedömts skulle användas som underlag vid betygssättning. Det var därför möjligt att utvärdera vilken tillit de undervisande lärarna kände till bedömningen som genomförts av de centrala bedömarna, lärare på andra skolor och av dem själva och en kollega på skolan.

I provplattformens resultatmodul kunde de undervisande lärarna se vilka poäng som eleverna hade fått på sina uppsatser. För proven i svenska kunde de även se hur de olika bedömarna hade poängsatt olika aspekter. Lärarna hade även möjlighet att öppna upp den bedömda uppsatsen och läsa den själv. Eleverna hade inte deltagit med sina egna namn, men lärarna fick föra egna listor på vilken elev som använde vilken inloggning så att de kunde koppla ihop elevens resultat och uppsats med rätt elev. Det fanns ingen möjlighet att se vem som bedömt elevens uppsats. Läraren kunde därmed inte skilja resultaten från central, extern och lokal bedömning.

4.3.1 Datainsamling genom intervjuer

För att få indikationer om vilken tillit lärare känner till att använda resultat de fått från central, extern och lokal bedömning intervjuades de undervisande lärarna. Lärarna

[40] Se enkätfrågorna i bilaga 1.
[41] Ibid.

visste inte vilka uppsatser som var bedömda av en kollega på skolan, en central eller en extern bedömare. De kunde dock se om resultaten skiljde sig från vad de var vana vid att eleverna brukade prestera. Eftersom de själva bedömt uppsatser från samma uppgift kunde de även resonera om olika utmaningar med att bedöma uppsatserna.

4.3.2 Urval vid intervjuer

Eftersom endast en tredjedel av de deltagande bedömarna hade ställt upp med elever var antalet så litet att det blev möjligt att genomföra fördjupande kvalitativa intervjuer. Lärarna valdes ut genom att de undervisande lärarna, det vill säga de lokala bedömarna, fick frågan om de ville delta i en intervju. Endast fyra av de 24 lärarna hade möjlighet att bli intervjuade online. Samtliga intervjuade lärare undervisade i svenska kurs 3. De arbetade för en kommunal respektive en enskild huvudman. Underlaget blev därmed för litet och för begränsat för att det skulle bli möjligt att kunna dra några slutsatser av lärarnas utsagor. Intervjuerna genomfördes ändå då de skulle kunna bidra med vissa insikter om utvärderingen som inte synliggjorts i bedömningarna eller enkäterna. Intervjuerna genomfördes vid två tillfällen som semistrukturella samtal och utgick från en lista med frågor som lärarna fick ta del av i förväg.[42] De båda lärare som intervjuades ihop var kollegor med varandra och hade genomfört sambedömningen tillsammans.

4.4 Analysmetoder

Man kan tala om tre sätt att studera överensstämmelse vid bedömning, som i litteraturen[43] brukar benämnas konsensus[44], konsistens och mätningsestimat (*measurement estimates)*. De har alla sina för- och nackdelar när det gäller kommunicerbarhet och komplexitet. Detta gör dem olika lämpliga för olika sammanhang av reliabilitet.

I den här utvärderingen har reliabilitet studerats utifrån de två perspektiven konsensus och konsistens. Konsensus lämpar sig bättre när kategorierna ska mäta innehållsliga aspekter där en viss poäng på en uppgift avspeglar en klar och tydlig innehållslig kvalitet, eller där en uppnådd kategori har en avgörande inverkan på ett beslut (därav ibland kallat *decision consistency*). Konsistens lämpar sig mer när man har en skala med flera skalsteg där skalans vidd i huvudsak syftar till att skilja individer åt. Det vill säga i de fall där rangordningen av individer är viktigare än den exakta kategorin som individen placerats i.

De mått som brukar användas för att redovisa konsensus och som använts i den här utvärderingen är absolut överensstämmelse, närliggande överensstämmelse[45], Cohens kappa och *Quadratic Weighted Kappa* (QWK). De två förstnämnda

[42] Se bilaga 2.
[43] Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. *Best practices in quantitative methods*, 29–49
[44] Ibland även kallad *decision consistency*.
[45] *Adjacent agreement*

konsensusmåtten utgörs av procentsatser där absolut överensstämmelse beräknas utifrån hur många elevlösningar, relativt sett, som har bedömts med samma poäng av bedömarna. Närliggande överensstämmelse är däremot lite mer förlåtande vad det gäller huruvida en elevlösning har bedömts. Enkelt uttryckt innebär närliggande överensstämmelse att två bedömningar som har hamnat intill varandra i till exempel antalet poäng betraktas som om bedömningarna har tilldelats ”lika många poäng”.

Cohens kappa som mått[46] på överensstämmelse. Som betraktare av två överensstämmande bedömningar kan man inte, utan ytterligare kunskap, avgöra om de två bedömningarna stämmer överens enkom som ett utfall av slumpen. Till exempel kan den ena eller båda bedömarna ha chansat[47] så mycket i sina bedömningar att resultatet av den parvisa jämförelsen dem emellan inte kan ses som äkta överensstämmelse. För att hantera slumpens inverkan på bedömares utfall är praxis i reliabilitetsstudier att beräkna så kallade kappavärden. Det finns mer än ett kappavärde att beräkna och i den här utvärderingen har Cohens kappa och QWK beräknats. Skillnaden mellan dessa är att det senare kvadrerar differenserna i de parvisa jämförelserna. QWK är således mer förlåtande mot mindre diskrepanser och mer straffande för större oenigheter i bedömning jämfört med kappa.

Om skalan är längre och utgörs av flera, snarare än färre, skalsteg minskar varje skalstegs innehållsliga relevans samtidigt som sannolikheten för att två bedömningar blir desamma minskar. I detta fall brukar konsistens vara ett mer relevant reliabilitetsperspektiv att lägga an.

Eftersom de möjliga poängen en uppsats kan få, är ganska många, kan man argumentera för att det är mindre relevant att bedömarna är överens om nivån för varje poäng. Det kan i stället anses viktigare att varje bedömare själv är konsekvent med hur hen tolkar varje poängnivå när hen bedömer uppsatserna. För att estimera i vilken grad bedömarna är konsekventa, i förhållande till varandra, brukar så kallade korrelationsmått användas. Främst är det Pearsons produktmomentskorrelationskoefficient som är vanligt att använda för estimerandet av konsistens.

Ett annat angreppssätt när det gäller konsistens är att beräkna det som kallas intraklasskorrelation (ICC). En fördel med ICC är att ett sammantaget mått erhålls, sett över alla de ingående bedömarna och dessas bedömningar. Det är alltså inte nödvändigt att räkna ut ICC baserat på parvisa jämförelser. ICC baseras på, så kallad, variansdekomposition där hänsyn tas till den variation av den totala variationen som går att tillskriva inom- respektive mellan bedömarna. Således ges ett mått på hur stor reliabiliteten i bedömarnas bedömningar är i relation till den totala variationen.

[46] Kappa kan anta värden mellan -1 och 1. Ett kappavärde på 1 innebär att de två bedömarna är helt eniga i sina bedömningar, -1 att de är helt oeniga och 0 att den parvisa överensstämmelsen i praktiken är helt slumpmässig.

[47] D.v.s. slumpvisa faktorer som bygger på andra grunder än de som ska användas vid bedömningen.

En annan fördel är att det utifrån olika antaganden om bedömarna är möjligt att beräkna olika ICC baserat på hur empirin ser ut samt vad som efterfrågas.[48] Det första antagandet rör huruvida det är olika uppsättningar av bedömare som bedömer respektive elevlösning eller inte. Då det är samma urval av bedömare som bedömer alla de i respektive delprov ingående elevlösningarna görs antagandet att modellen för utvärderingen är av typen tvåvägs (*two way)*. Det andra antagandet rör huruvida bedömarnas genomsnittliga bedömningar är av relevans för det som efterfrågas. I och med att denna utvärdering endast intresserar sig för konsistens, är den genomsnittliga bedömningen inte relevant. Det tredje nödvändiga antagandet är att bestämma huruvida det mått på reliabilitet som ska estimeras ska spegla den enskilda bedömarens bedömningar eller bedömarna som grupp. Sett ur ett elevrättssäkerhetsperspektiv är det främst den enskilda bedömarens bedömningar som är av intresse inom ramen för detta arbete. Till sist behövs ett antagande om huruvida bedömarna har blivit slumpmässigt utvalda eller om de har blivit utvalda på andra bevekelsegrunder.

Systematiska effekter som leder till att olika bedömare bedömer samma uppsats olika, s.k. *bedömareffekter*, utgörs av mätfel som går att tillskriva enskilda bedömare. Med anledning av att detta mätfel har sitt ursprung i ett mönster, en viss systematik, går inte detta nödvändigtvis att upptäcka i reliabilitetsstudier av typen konsensus eller konsistens. Men likväl har dessa systematiska mätfel ofta en negativ inverkan på ett provs validitet, det vill säga vilka slutsatser de kan användas till att dra samt deras syfte.

I analysarbetet har SPSS version 28 samt R med de två paketen *irr* (version 0.84.1) och *psych* (version 2.2.9) använts.

5. Redovisning och analys av reliabiliteten

I detta kapitel beskrivs analysen av de bedömningar som samlats in. Bedömningarna har analyserats utifrån de två reliabilitetsperspektiven konsensus och konsistens.

I analysen har bedömarna delats upp i olika bedömartyper. De olika bedömartyperna var centrala, externa och lokala bedömare. Dessa redovisas genomgående tillsammans med en fjärde bedömarkategori, sambedömare. I de påföljande analyserna redovisas dessa sambedömningar som om de var gjorda av olika individer när de i själva verket bestod av par av sambedömare.

[48] Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2), 420.

Tabell 1. Antalet bedömare per ämne, årskurs och bedömartyp.

	Åk6 svenska	Åk9 engelska	Gy3 svenska
Central rättning	7	7	6
Extern rättning	8	5	8
Lokal rättning	6	5	7
Sambedömning	10	10	11

5.1 Konsensus

Det kanske vanligaste och mest intuitiva sättet att se på och uppfatta överensstämmelse är om det råder konsensus eller inte, det vill säga om två eller fler bedömare har landat i samma bedömning av en specifik elevlösning. I de här påföljande konsensusanalyserna estimeras graden av konsensus utifrån ett par bedömare i taget. Därefter beräknas medelvärden för de parvisa bedömningarna.

För att värdera resultaten av de estimerade konsensusmåtten har en gräns på 0,7 använts. Det är det värde som i litteraturen brukar användas som gränsvärde för vad som är acceptabel reliabilitet.[49]

5.1.1 Absolut och närliggande överensstämmelse

Det mått på reliabilitet som är enklast att beräkna är det som kallas absolut överensstämmelse och som återger hur stor andel av ett antal elevlösningar som bedömarna har bedömt med samma poäng. Absolut överensstämmelse beräknas för ett bedömarpar i taget och har här redovisats som den genomsnittliga överensstämmelsen för respektive delprov och aspekt.

Tabell 2. Aspektbedömning i svenska årskurs 6 och gymnasiets år 3, absolut överensstämmelse.

5 uppsatser						**10 uppsatser**		
Åk6 svenska		*% absolut*				Gy3 svenska		
	antal bedömare	*Innehåll*	*Struktur*	*Språk*	*Skrivregler*	*antal bedömare*	*Innehåll och källhantering*	*Disposition och sammanhang*
Central rättning	7	0,52	0,58	0,5	0,6	6	0,47	0,55
Extern rättning	8	0,54	0,47	0,5	0,53	8	0,47	0,55
Lokal rättning	6	0,45	0,64	0,51	0,68	7	0,39	0,6

I tabellen ovan återges den genomsnittliga absoluta överensstämmelsen så som de estimerats för respektive aspekt i svenskuppsatserna i årskurs 6 respektive gymnasiets kurs 3. I både årskurs 6 och gymnasiet kunde bedömaren utifrån tillhörande bedömningsanvisningar tilldela en aspektpoäng som sträckte sig från 0 till och med 3, undantaget aspekten *disposition och sammanhang* i gymnasiet där

[49] Stemler, S. E., & Tsai, J. (2008). Best practices in interrater reliability: Three common approaches. *Best practices in quantitative methods*, 29–49

maxpoängen var 2 poäng. Oavsett om utfallet för årskurs 6 eller gymnasiet studeras, så såg det i huvuddrag ut som att ungefär hälften av de parvisa bedömningarna stämde helt överens. Viss variation mellan de olika aspekterna fanns, men det såg inte ut att finnas något tydligt mönster mellan de olika bedömartyperna.

Tabell 3. Aspektbedömning i svenska årskurs 6 och gymnasiets år 3, närliggande överensstämmelse.

5 uppsatser						**10 uppsatser**			
Åk6 svenska		*% närtilliggande*				**Gy3 svenska**			
	antal bedömare	*Innehåll*	*Struktur*	*Språk*	*Skrivregler*	*antal bedömare*	*Innehåll och källhantering*	*Disposition och sammanhang*	*Språk och stil*
Central rättning	7	0,94	0,98	0,98	0,98	6	0,89	0,99	1
Extern rättning	8	0,98	0,99	1	0,94	8	0,92	0,99	0,9
Lokal rättning	6	1	1	1	1	7	0,83	0,98	0,92
Sambedömning	10	0,98	0,98	0,98	0,95	11	0,93	0,99	0,96

I avsnittet *analysmetoder* beskrivs konsensusmåttet närliggande överensstämmelse och att vad som är att betrakta som närliggande beror på sammanhang och skala. I det här arbetet har endast den närmast underliggande eller överliggande poängen betraktats som närliggande i beräkningarna. Det vill säga att bedömningarna fick skilja sig med 1 poäng. När dessa avgränsningar gjordes blev det tydligt, vilket går att utläsa ur tabellen, att så gott som alla aspekter fick ett utfall på upp emot 100 procents överensstämmelse. Aspekten *innehåll och källhantering* i gymnasiets uppsats avvek dock från det övergripande mönstret med en lägre överensstämmelse, där det varierade mellan de lokala rättarnas 83 procent och sambedömarnas 93 procent.

Konsensus som reliabilitetsperspektiv lämpar sig, som tidigare nämnts, bättre då bedömningskategorierna är färre till antalet och där de är tydligt definierade i bedömningsanvisningarna. När bedömningarna av de i uppsatsen ingående aspekterna summerades ihop för att utgöra ett sammantaget resultat var bedömningskategorierna inte längre få, och framför allt var de summerade poängen, från aspekterna, inte tydligt definierade i hur de rent kvalitativt skiljde sig åt. Detta speglades i motsvarande analyser som gjorts på delprovstotalerna. Där sträckte sig den genomsnittliga absoluta överensstämmelsen från 20 till 40 procent samt mellan 50 och 80 procent vid närliggande överensstämmelse.

5.1.2 Cohens kappa och Quadratic Weighted Kappa

För att hantera och kompensera för att det finns en möjlighet att slumpen ger upphov till att två bedömare gör samma bedömning, alltså att två bedömare av ren tillfällighet sätter samma poäng på en viss aspekt på en viss elevlösning, studerades bedömningarna genom reliabilitetsmåttet Cohens kappa.

Tabell 4. Kappa för svenska årskurs 6 och gymnasiets år 3.

5 uppsatser						**10 uppsatser**			
Åk6 svenska		***Kappa***				**Gy3 svenska**			
	antal bedömare	*Innehåll*	*Struktur*	*Språk*	*Skrivregler*	*antal bedömare*	*Innehåll och källhantering*	*Disposition och sammanhang*	*Språk och stil*
Central rättning	7	0,3	0,36	0,33	0,35	6	0,19	0,14	0,27
Extern rättning	8	0,4	0,31	0,32	0,27	8	0,18	0,29	0,27
Lokal rättning	6	0,26	0,46	0,31	0,4	7	0,15	0,36	0,28
Sambedömning	10	0,32	0,35	0,33	0,21	11	0,22	0,34	0,33

Det som är mest slående när de värden som estimerats för Cohens kappa studeras, var att ingen bedömargrupp, vare sig i årskurs 6 eller gymnasiets kurs 3, i utvärderingen, hamnade på en genomsnittlig kappa som låg över den gräns på 0,7 som brukar användas som gräns för vad som är acceptabelt. Som bäst blev den genomsnittliga kappan i årskurs 6, som låg i det intervall på 0,4–0,6. Detta intervall beskrev Cohen[50] som måttlig överensstämmelse.

När det gäller bedömartyp fanns i övrigt inga tydliga mönster som det gick att dra några slutsatser av. Eventuellt pekade resultaten på en något högre överensstämmelse i årskurs 6 jämfört med gymnasiet. Åtminstone såg det ut som att aspekten *innehåll och källhantering* i gymnasiets uppsats i svenska, liksom i fallet för den närliggande överensstämmelse, stack ut något negativt.

Tabell 5. *Quadratic Weighted Kappa*.

5 uppsatser						**10 uppsatser**			
Åk6 svenska		***QWK***				**Gy3 svenska**			
	antal bedömare	*Innehåll*	*Struktur*	*Språk*	*Skrivregler*	*antal bedömare*	*Innehåll och källhantering*	*Disposition och sammanhang*	*Språk och stil*
Central rättning	7	0,48	0,6	0,72	0,53	6	0,35	0,27	0,48
Extern rättning	8	0,78	0,67	0,74	0,46	8	0,38	0,52	0,49
Lokal rättning	6	0,6	0,65	0,69	0,4	7	0,33	0,51	0,35
Sambedömning	10	0,7	0,66	0,71	0,51	11	0,51	0,54	0,57

Vid studier av måttet *Quadratic Weighted Kappa* (QWK) syns en tydligare skillnad dels mellan årskurs 6 och gymnasiet. Speciellt syns detta mellan aspekterna inom årskurs 6 respektive gymnasiet. Som tidigare nämnts är QWK, visavi Cohens kappa, mer förlåtande mot mindre diskrepanser och mer straffande mot större diskrepanser.

Ur tabell 4 går det att utläsa att några bedömarkategorier för ett par aspekter i årskurs 6 har ett genomsnittligt QWK på 0,7 eller över. Det vill säga, enligt QWK har bedömningen erhållit en acceptabel nivå på överensstämmelse avseende dessa. Sett över alla de fyra aspekterna i årskurs 6 är det dock svårt att dra några särskilda slutsatser utifrån bedömartyp. Däremot indikerar resultaten, liksom vad

[50] McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3), 276–282.

som observerades för Cohens kappa, att reliabiliteten per aspekt är högre i årskurs 6 jämfört med gymnasiet.

5.2 Konsistens

Konsistens lämpar sig mer när man har en skala med flera skalsteg där skalans vidd syftar till att skilja individer åt och där rangordningen av individer är viktigare än den exakta kategorin som individen placerats i. Fokus ligger på hur en och samma bedömare gör samma bedömning över tid eller att flera bedömare i förhållande till varandra gör samma bedömningar så att deras bedömningar i relation till varandra är konsekventa. Eller annorlunda uttryckt är konsistenta. Ju fler skalsteg – ju naturligare är det att konsensusmåtten får låga värden. Detta faktum bör beaktas apropå att fler skalsteg införts i proven i svenska när poäng läggs samman jämfört med när de skulle sammanfattas till ett betygssteg.

För att värdera resultaten som erhållits från konsistensanalyserna användes den gräns på 0,7 som är praxis att använda sig av som gränsvärde för vad som är acceptabelt eller inte.[51]

5.2.1 Korrelation

Liksom var fallet i avsnittet för konsensus beräknas konsistensmått, som utgörs av traditionella korrelationsmått av parvisa beräkningar. I regel är traditionella korrelationsmått parametriska och kan därmed vara känsliga för om data inte är normalfördelat. Till exempel, om data har en skev fördelning trots att den verkliga bakomliggande fördelningen är att betrakta som normal kan parametriska korrelationsmått komma att underskatta den verkliga korrelationen. Här har därför även redovisats estimat för det icke-parametriska Spearmans rangkorrelationskoefficient, utöver det parametriska Pearsons produktkorrelationskoefficient.

Tabell 6. Korrelationskoefficienter…, både Pearson och Spearman.

	Åk6 svenska			**Åk9 engelska**			**Gy3 svenska**		
	antal bedömare	*Pearson*	*Spearman*	*antal bedömare*	*Pearson*	*Spearman*	*antal bedömare*	*Pearson*	*Spearman*
Central rättning	7	0,794	0,769	7	0,904	0,842	6	0,498	0,431
Extern rättning	8	0,893	0,903	5	0,922	0,9	8	0,684	0,688
Lokal rättning	6	0,806	0,814	5	0,928	0,911	7	0,574	0,567
Sambedömning	10	0,819	0,814	10	0,929	0,931	11	0,802	0,764

Ur tabellen ovan går det att utläsa att bedömarna i engelskans årskurs 9 i snitt har varit mer konsistenta än motsvarande i svenskans årskurs 6 och, framför allt, gymnasiets kurs 3. Det blir synligt att de genomsnittliga korrelationerna ligger kring 0,9 i årskurs 9. Motsvarande genomsnitt i årskurs 6 ligger mellan 0,8 och 0,9 samt i gymnasiet mellan 0,4 och 0,7 sambedömarna undantaget.

[51] Barrett, P. (2001). Assessing the reliability of rating data.

Vid studier av de parvisa korrelationerna för bedömarna i gymnasiets kurs 3 i svenska på en, mer finkornig nivå, syns en variation mellan ett lägsta värde på mindre än 0,1 och ett högsta värde på nästan perfekt korrelation 1,0. Detta tyder på att det finns bedömare som är helt konsistenta i relation till varandra medan det finns bedömare som är nästintill helt inkonsistenta. Sådan inkonsistens kan i hög utsträckning påverka vad den enskilda eleven får för slutresultat på det enskilda delprovet. Man kan då säga att om elevens resultat hamnar ovanför eller under en kravgräns, till stor del utgörs av vilken lärare som gjort bedömningen och hur denne har bedömt elevens delprov.

5.2.2 Intraklasskorrelation

En stor fördel med intraklasskorrelation (ICC) som konsistensmått är att det går att beräkna ett sammanfattande reliabilitetsestimat för den grupp bedömare då det fanns fler än två bedömare att ta hänsyn till. Men där de tidigare redovisade korrelationsmåtten är enkla att rättframt estimera behöver man vid skattandet av ICC veta hur man ska förhålla sig till några antaganden för att bestämma den modell av ICC som är lämplig att använda.

Den modell som ställts upp för att estimera ICC utgår från att det var samma bedömare, givet delprov, bedömartyp och årskurs, som genomgående har bedömt alla elevlösningar och att dessa inte var slumpmässigt utvalda. Att utvärderingens primära fokus ligger på konsistens i bedömarnas bedömningar samt att det främst är den enskilda bedömarens bedömningar som är av intresse. Då det även är intressant att se utfallet på gruppnivå, framför allt om de två kontrasterande måtten visar på en stor diskrepans, har också analysen gjorts utifrån detta antagande.

Tabell 7. Intraklasskorrelation

	Åk6				**Åk9**				**Gy3**			
	Central	**Extern**	**Lokal**	**Sambedömare**	**Central**	**Extern**	**Lokal**	**Sambedömare**	**Central**	**Extern**	**Lokal**	**Sambedömare**
ICC3	0,79	0,86	0,74	0,79	0,9	0,91	0,92	0,92	0,52	0,63	0,56	0,77
ICC3K	0,96	0,98	0,95	0,97	0,98	0,98	0,98	0,99	0,87	0,93	0,9	0,97

Tabell 7 ovan återger de två varianterna av ICC som estimerats i utvärderingen. På den första raden finns storleken på den reliabilitet som gäller utifrån ett individperspektiv medan rad två återger reliabiliteten utifrån bedömarna som grupp.

Om resultaten först betraktas utifrån det individperspektiv som har betydelse för den enskilda eleven blir det synligt att de olika bedömargrupperna ligger väldigt högt i engelskans åk 9. Alla grupper uppnår en reliabilitet på cirka 0,9. Något sämre ser reliabiliteten ut att vara i årskurs 6 medan det urval som fanns att arbeta med för gymnasiet visar den lägsta intraklasskorrelationen. Sambedömarna i kurs 3 uppvisar dock en acceptabel reliabilitet. Skillnaden är inte så stor att det går att dra slutsatsen att sambedömning ökade reliabiliteten i bedömningen. Framför allt

då inte samma positiva effekt av sambedömning observerats jämfört med övriga bedömarkategorier i årskurs 6 eller 9.

Genom att i stället betrakta reliabilitetstalen för respektive bedömargrupp och årskurs, erhåller samtliga en väldigt hög reliabilitet. Detta innebär att även eleverna i gymnasiet skulle åtnjuta en hög reliabilitet ifall det fanns resurser för att låta en grupp bedömare bedöma varje elevs uppsats, i stället för av en enskild bedömare.

Avslutningsvis kan det konstateras att inte heller ICC, som speglar reliabilitet på gruppnivå, visar några tydliga tecken på att någon av grupperna centrala bedömare, externa, lokala eller sambedömare skulle vara mer lämplig än övriga grupper för bedömning av de nationella proven. De ligger alla på samma ungefärliga höga nivå.

6. Resonemang om hur reliabiliteten påverkar elevens provbetyg

I detta kapitel förs ett resonemang om vilken effekt skillnader i bedömning av uppsatser skulle få på elevernas provbetyg.

Som beskrivits i avsnittet Material och metod, hade inte Skolverket och de lärosäten som utvecklar digitala nationella prov i engelska och svenska helt beslutat hur poängmodellen konkret skulle se ut för uppsatsskrivning vid studiens genomförande. Det var därför svårt att säga något om hur stor roll de poäng som genererades i utvärderingen skulle spela för elevens provbetyg. I såväl proven i engelska som svenska kommer resultatet från uppsatserna att presenteras dels i form av en delprovspoäng, dels i form av en sammanlagd provpoäng. Kravgränserna för provbetyget anges för den totala provpoängen. Redan en poängs skillnad mellan två bedömare skulle kunna resultera i att en elev trillar över eller hamnar under en kravgräns för ett visst provbetyg beroende på hur elevens övriga delprovsresultat ser ut.

Minst problematiskt är det att resonera om konsekvenserna av skillnader i bedömning av engelska årskurs 9. Det är högst troligt att de digitala proven i engelska för årskurs 9 kommer fortsätta att ha en modell där varje delprov kan resultera i 1 till 9 delprovspoäng. Proven består i dag av fyra delprov[52] där uppsatsdelen står för 25 procent av provpoängen. Maxpoängen för provet som helhet borde ligga på 36 poäng.[53] Den största avvikelsen som uppmätts mellan två bedömare är fyra poäng vilket är en ganska liten andel, cirka 11 procent, av de

[52] Delproven i hörförståelse och läsförståelse räknas ihop men dubblas vid sammanräkningen.
[53] I det frisläppta prov som användes angavs delprovsresultat som betyg och delprovsresultaten blev, förenklat uttryckt, ett medelbetyg av delprovsbetygen.

totala möjliga poängen. Även om det varit känt vilka delprovsresultat eleven hade på övriga delprov går det inte säga hur många procent som skulle riskera att få ett annat provbetyg eftersom kravgränserna för provpoäng till provbetyg[54] inte var kända.

För proven i svenska årskurs 6 kan delprovsresultaten i försöket resultera i 0 till 9 delprovspoäng och i det digitala nationella provet kommer delprovsresultatet på de två uppsatserna som eleverna skriver generera 0 till 18 poäng. I utvärderingen kunde två bedömares divergerande poängsättning på aspektnivå ta ut varandra vid sammanläggningen, men den största avvikelsen låg ändå på sex poäng.

Skillnaderna mellan två bedömare skulle kunna få störst konsekvenser för proven i svenska för gymnasiets kurs 3. Uppsatsskrivningen ger i dag mellan 0 till 24 delprovspoäng och står för 60 procent av provresultatet. Övriga 40 procent av provresultatet kommer från en muntlig presentation som är minst lika svår att bedöma reliabelt. I utvärderingen skiljde sig bedömarna mest från varandra i bedömningen av uppsatserna från Svenska kurs 3. Den största skillnaden mellan bedömarna på en uppsats låg på 7 totalpoäng. Denna skillnad riskerar att bli betydligt större vid viktning och omskalning till 0 till 24 poäng på delprovet med uppsatsskrivning.

7. Redovisning samt analys av bedömarnas uppfattning av att bedöma

I detta kapitel redovisas utfallet av den enkät som deltagande bedömare fick svara på. Enkäten innehöll frågor om upplevelsen av att bedöma som central, extern och lokal bedömare i Skolverkets provplattform. Först redovisas likheter i svaren mellan de olika bedömningsgrupperna, lokala, externa och centrala bedömare samt mellan bedömare från proven i engelska årskurs 9 respektive svenska i årskurs 6 och kurs 3. Sedan redovisas skillnaderna i svar mellan grupperna och bedömarna från de olika proven. Slutligen kommer en sammanställning av vanliga kommentarer i bedömarnas fritextsvar.

Skillnaderna mellan gruppernas kompetens i bedömning bestod, som beskrivits tidigare, av att de centrala bedömarna fick en utbildning i bedömning. De var också legitimerade och behöriga att undervisa i ämnet och i den aktuella årskursen/kursen. Övriga bedömare kunde vara obehöriga, vilket naturligtvis, kan ha påverkat hur de uppfattade att det var att bedöma lokalt eller externt. I tabellen

[54] Se kommentar ovan om sammanställning av provresultat från det använda frisläppta provet.

sammanfattas hur bedömarna uppfattade att bedöma uppsatser i Skolverkets digitala provtjänst.

21 centrala bedömare besvarade enkäten. 3 centrala bedömare hade ett till två års erfarenhet av att bedöma nationella prov. 8 hade tre till nio års erfarenhet och övriga 10 hade mer än tio års erfarenhet.

19 externa bedömare besvarade enkäten. 3 hade mindre än två års erfarenhet av att bedöma nationella prov, 8 av dem hade tre till nio års erfarenhet och 8 av dem hade mer än tio års erfarenhet. De externa bedömarna hade bedömt uppsatser som kom från en okänd skola och sambedömde med någon de aldrig hade träffat förr.

20 lokala bedömare besvarade enkäten. Ingen av dem hade mindre än tre års erfarenhet av bedömning av nationella prov och 13 av dem hade mer än tio års erfarenhet. De lokala bedömarna hade, precis som alla andra, bedömt avidentifierade elevtexter från den egna skolan och de hade inte kunnat identifiera vilken elev som skrivit uppsatsen. Till skillnad från centrala och externa bedömare, hade de sambedömt med en kollega på skolan.

Tabell 8. Antalet bedömare som höll med om ett positivt påstående om digital bedömning i de olika grupperna

	Att bedöma uppsatser digitalt i stället för på papper gick snabbare eller mycket snabbare	**Att klicka på knappar i stället för att dokumentera på papper gick snabbare eller mycket snabbare**	**Att läsa elevtexter på dator var enklare eller mycket enklare**	**Att sätta poäng i stället för betyg var enklare eller mycket enklare**
20 lokala bedömare	18	19	12	13
19 externa bedömare	14	16	8	15
21 centrala bedömare	14	14	7	12
Sambedömning	**Det var inga problem att komma överens**	**Jag ändrade mina poäng i inget eller i något enstaka fall**	**Min uppfattning om elevtexters kvalitet påverkades i ganska hög eller hög grad**	
20 lokala bedömare	13	11	6	
19 externa bedömare	16	7	8	
21 centrala bedömare	17	4	16	

Eftersom reliabiliteten skiljer sig något åt mellan de olika proven som bedömts, se ovan kapitel 4, kan det också vara intressant att jämföra enkätsvaren för dessa. Av de som besvarade enkäten bedömde 17 lärare svenska åk 6, 22 lärare bedömde engelska åk 9 och 21 lärare bedömde svenska kurs 3.

Bedömarnas erfarenhet av att bedöma nationella prov var ungefär densamma i svenska åk 6 och kurs 3, där samtliga bedömare utom en hade minst tre års erfarenhet av bedömning av nationella prov. Bedömarna i engelska var något mindre erfarna. Där fanns ett par nybörjare och ett par med ett till två års erfarenhet. Minst hälften av bedömarna hade dock mer än tio års erfarenhet av bedömning i samtliga grupper. Det kan vara värt att påpeka att de erfarna bedömarna i åk 9 och kurs 3 har bedömt avidentifierade elevtexter skrivna på dator tidigare eftersom uppsatsdelarna i svenska och engelska i åk 9 och i gymnasieskolan ska skrivas på dator sedan 2018.[55] Det är emellertid inte säkert att bedömningen genomförts digitalt; många kan ha valt att skriva ut elevtexterna vid bedömningen.

[55] Skolförordning (2011:185). 9 kap. 21a § och 22 §.

Tabell 9. Antalet bedömare som höll med om ett positivt påstående om digital bedömning i de olika grupperna

	Att bedöma uppsatser på papper i stället för digitalt gick snabbare eller mycket snabbare	**Att klicka på knappar i stället för att dokumentera på papper gick snabbare eller mycket snabbare**	**Att läsa elevtexter på dator var enklare eller mycket enklare**	**Att sätta poäng i stället för betyg var enklare eller mycket enklare**
Svenska åk 6 17 bedömare	12	14	7	15
Engelska åk 9 22 bedömare	20	18	13	8
Svenska kurs 3 21 bedömare	14	14	7	12
Sambedömning	**Det var inga problem att komma överens**	**Jag ändrade mina poäng i inget eller i något enstaka fall**	**Min uppfattning om elevtexters kvalitet påverkades i ganska hög eller hög grad**	
Svenska åk 6 17 bedömare	17	6	8	
Engelska åk 9 22 bedömare	15	12	13	
Svenska kurs 3 21 bedömare	14	4	9	

7.1 Likheter mellan bedömares uppfattning om bedömningen

En majoritet av bedömarna som besvarade enkäten tyckte att det gick snabbare att bedöma digitalt jämfört med på papper och tyckte att det varit enklare att bedöma genom att klicka på knappar i stället för att dokumentera på papper. Mer än hälften tyckte att bedömningen hade blivit enklare eller mycket enklare när de läste texterna på datorn och tyckte att det hade blivit enklare eller mycket enklare att bedöma med poäng i stället för med betyg. Det hade också funnits möjlighet att skriva en kommentar om upplevelsen av den digitala bedömningen. De som kommenterade, oavsett om de tyckte att de blev enklare eller svårare att genomföra bedömningen, ansåg att överblicken över texten blev sämre och att de hade saknat möjligheten att markera språkfel, stryka under och kommentera i texterna under bedömningen.

En majoritet av bedömarna för samtliga prov tyckte både att det hade gått snabbare eller mycket snabbare att bedöma digitalt jämfört med på papper, samt att det hade varit enklare eller mycket enklare att bedöma genom att klicka på knappar i stället för att dokumentera på papper.

När det gäller sambedömning och hur samarbetet med kollegan fungerade ansåg samtliga som bedömde svenska åk 6 att det inte hade varit några problem alls att komma överens. Men även en majoritet av bedömarna i engelska åk 9 och svenska kurs 3 uppgav att de inte hade haft några problem alls att komma överens. På frågan om i vilken grad bedömarna ändrade sina poäng under sambedömningen var resultatet likartat för samtliga prov; förutom ett par personer för varje prov, menade alla att det bara hade inträffat i enstaka fall eller i färre än hälften av bedömningarna. På frågan om sambedömningen hade påverkat bedömarens uppfattning om elevtexternas kvalitet, menade ungefär hälften av bedömarna för samtliga prov att den hade påverkat dem i mindre grad eller inte alls.

7.2 Skillnader mellan bedömares uppfattning om bedömningen

Trots att bedömarna från de olika grupperna var eniga i svaren på frågan om hur bedömningen påverkades av att texterna lästes på datorn, går det att se en skillnad mellan bedömare i svenska och engelska. Av de som bedömde i svenska, i åk 6 och kurs 3 tyckte färre att bedömningen hade blivit enklare eller mycket enklare när de läste texterna på datorn, än de som bedömde i engelska åk 9. När det gäller frågan om att bedöma med poäng i stället för med betyg ansåg de flesta av bedömarna i svenska åk 6 att det hade blivit enklare eller mycket enklare att bedöma med poäng. I svenska kurs 3 ansåg hälften av bedömarna att det hade blivit enklare eller mycket enklare att bedöma med poäng i stället för med betyg. I engelska var det däremot mindre än hälften av bedömarna som tyckte att det blivit enklare eller mycket enklare.

När det gäller lokal sambedömning och hur samarbetet med kollegan fungerade ansåg över hälften av de lokala bedömarna att de inte hade några problem alls att komma överens. Medan övriga ansåg att de hade fått kompromissa om poängen eller behövt ha en omfattande diskussion för att komma överens. På frågan om i vilken grad bedömarna ändrade sina poäng under sambedömningen menade ungefär varannan lokal bedömare att det hade inträffat i inget eller i något enstaka fall. Endast en bedömare hade ändrat sina poäng i mer än hälften av fallen. Enligt lärarnas svar borde alltså utfallet från sambedömningen ha varit det samma som från deras individuella bedömning[56]. På frågan om sambedömningen hade påverkat bedömarens uppfattning om elevtexternas kvalitet menade en majoritet att den inte hade påverkat dem alls. De lokala bedömarna var kollegor och verkade redan ha en samsyn om hur de skulle bedöma uppsatserna.

När det gäller extern sambedömning och hur samarbetet med den andra externa bedömaren fungerade ansåg majoriteten att det inte hade varit några problem alls att komma överens. På frågan om i vilken grad bedömarna ändrade sina poäng under sambedömningen menade mindre än hälften att det hade inträffat i inget

[56] I kapitel 4 redovisas utfallet från sambedömningen som en gemensam grupp och ingen skillnad görs på sambedömningen från centrala, externa respektive lokala bedömare.

eller i något enstaka fall medan mer än hälften menade att det hade inträffat i färre än hälften av bedömningarna. De externa bedömarna, som inte kände varandra, ändrade alltså sin bedömning något oftare än vad de lokala bedömarna gjorde. Över hälften av de externa bedömarna menade dock att sambedömningen inte hade påverkat dem alls eller i mindre grad som bedömare.

När det gäller central sambedömning och hur samarbetet med den andra centrala bedömaren fungerade, ansåg majoritet att det inte hade varit några problem alls att komma överens. På frågan om i vilken grad bedömarna ändrade sina poäng under sambedömningen menade majoriteten att det hade inträffat i färre än hälften av bedömningarna. Till skillnad från de lokala och externa bedömarna menade dock en majoritet av de centrala bedömarna att hade påverkats i ganska eller i hög grad i sin uppfattning om bedömning av elevtexternas kvalitet.

7.3 Centrala bedömares uppfattning om utbildningen i bedömning

De centrala bedömarna hade genomgått en utbildning i bedömning av det aktuella uppsatsämnet, som anordnades av Skolverket. I enkäten ställdes tre frågor om utbildningen.

En majoritet av de centrala bedömarna i alla prov upplevde att de hade blivit en säkrare eller mycket säkrare bedömare tack vare utbildningen medan övriga ansåg att de hade varit lika säkra som före utbildningen. Att bygga utbildningen på bedömning och diskussion av elevexempel tyckte nästan alla fungerade bra eller mycket bra. Endast en bedömare tyckte att detta upplägg hade fungerat ganska dåligt. Bedömaren skrev i en kommentar att hen i stället hade önskat fler föreläsningar om hur man bedömde i olika nivåer. Majoriteten av de centrala bedömarna tyckte att det var viktigt att bygga en bedömarutbildning utifrån den aktuella provutgåvan.[57] Övriga tyckte att det hade varit mindre viktigt och två att det inte hade varit viktigt alls.

7.4 Fritextsvar om upplevelsen av bedömningen

Bedömarna fick också besvara två frågor med fritext, dels om vad de enligt deras erfarenhet upplevde som positivt med att bedöma i en digital plattform, dels vad de upplevde som negativt. De fick även möjlighet att skriva ner övriga synpunkter.

Positiva omdömen, som återkom för många av bedömarna, var att det var smidigt att bedöma digitalt, att det gick snabbare och att det var enklare. Det var också

[57] Begreppet ”prov” menas det som är kopplat till en kurs eller årskurs till exempel. Engelska kurs 6. Begreppet ”provutgåva” avser en specifik version av provet till exempel Engelska kurs 6 vårterminen 2023. För de flesta elever är endast en provutgåva aktuell, men i kommunal vuxenutbildning har lärare möjlighet att välja mellan fler olika provutgåvor.

skönt att ha allt samlat på ett ställe och att man lätt kunde gå tillbaka till de texter man var osäker på.

När det gällde de negativa aspekterna var återkommande synpunkter att de gärna hade velat anteckna i elevens text samt att någon tyckte att överblicken blev sämre. Generellt var bedömarna i svenska åk 6 och engelska åk 9 mer positiva till bedömningen i provplattformen. Bedömarna i kurs 3 på gymnasiet var i högre grad negativt inställda. De betonade problemen med bristande översikt och att möjligheten att kommentera och markera i texterna saknades i betydligt större utsträckning än övriga. Texterna från eleverna i kurs 3 skrev var också längst.

I den sista enkätfrågan kunde bedömarna uttrycka övriga synpunkter. De allra flesta av bedömarna kommenterade sådant som hade med försökens upplägg och genomförande samt med provplattformens funktionalitet att göra. Den informationen har tagits tillvara av Skolverkets försöksverksamhet och projektet Digitala nationella prov. I den här rapporten lyfts ett par synpunkter som är av intresse för resultatet av reliabiliteten.

De som bedömde svenska åk 6 var i högre grad positiva till försöket i sina övriga kommentarer än de andra två grupperna med bedömare. Här följer ett par exempel på kommentarer:

> Mycket smidigt bedömningssätt. Väldigt bra med anonyma texter. Tror starkt på sambedömning.
>
> Mycket intressant och bra fortbildning. Det blev tydligt att det finns stora vinster med sambedömning för att säkra kvalitén på bedömningarna.

Bedömarna i engelska åk 9 fokuserade i fler fall på vikten av utbildning och dess kvalitet. Här följer ett par exempel:

> Jag tycker att det är synd att vi inte fick mer tid till diskussion inför, under eller efter sambedömningen. Ett forum där man är många lärare från olika skolor och olika platser i Sverige är få förunnat, och jag önskar att det hade kunnat utnyttjats mer för att diskutera bedömning och nationella prov. Det hade varit önskvärt med en återsamling efter sista sambedömningen där vi diskuterar hur det gått och hur man upplever att plattformen och bedömningen funkar.
> Tack för att jag fick delta!
>
> Skulle önskat mer av föreläsningar kring bedömning och hur man differentierar olika nivåer.
>
> Det var bra att kunna vara en del av detta. Jag kände mig tryggare med mina egna bedömningar och jag lärde mig lite mer om vad är viktigaste för att få de bästa betygen.

De centrala bedömarna som undervisade i kurs 3 hade uppskattat både utbildningen och sambedömningen, men de problematiserade bedömningen mer än övriga grupper i sina svar. En kommentar uppmärksammar hur bristen på

likvärdighet i undervisningen mellan olika lärare skulle kunna skapa utmaningar med att uppnå en likvärdig bedömning mellan skolor. En bedömare skriver:

> Något som dök upp under sambedömningen var frågan om disposition. /…/ Jag stöter inte på sådana texter där elever inte vet hur de ska dela in texten i stycken för det har de lärt sig när det är dags att skriva utredande. Jag kände mig därför osäker på några uppsatser gällande dispositionen och de kommentarer som finns i bedömningsanvisningarna. Jag har helt enkelt inte fått in texter med punktform förut. Min spontana känsla var att det inte kan godkännas, men när jag läste bedömningsanvisningen som är vag på den punkten så ändrade jag mig i sista sekund.

En annan bedömare tyckte att:

> Utbildningen var givande, det var bra att själv bedöma först, därefter diskutera med andra, och sist höra hur Skolverket bedömt. Insåg vikten av att verkligen sätta sig in i exempeltexterna och kommentarerna i "röda häftet" innan man tar sig an elevtexterna. Och hur bra sambedömning verkligen är!

8. Intervjuer med betygssättande lärare

I detta kapitel återges uttalande från fyra betygssättande lärare om vilken tillit de hade till bedömningen när större delen av uppsatserna hade bedömts av centrala och externa bedömare. Kapitlet återger först lärarnas uppfattning om hur elevernas resultat överensstämmer med deras egen bild av hur bra texter eleverna brukar skriva. Sedan följer en redogörelse för hur lärarna ser på användbarheten av resultatet för betygssättningen. Kapitlet avslutas med en redogörelse för hur lärarna resonerar generellt om vad som gör central rättning och extern bedömning tillförlitlig och användbar.

Samtliga intervjuade lärare hade deltagit som lokala bedömare i utvärderingen och var legitimerade ämnesbehöriga gymnasielärare med lång erfarenhet av att bedöma nationella prov. Intervjun genomfördes efter att de hade fått tillbaka sina elevers bedömda uppsatser. Eftersom lärarna kom från två skolor intervjuades de parvis med sin kollega. Nedan benämns skolorna som skola 1 och skola 2. Alla fyra lärare undervisade på högskoleförberedande program i svenska kurs 3. De undervisade elever som hade kunskaper som motsvarade samtliga betygsnivåer, men i princip inga elever med F i betyg. Lärarna hade fått ta del av sina elevers resultat genom plattformens digitala rapporter där man även kunde öppna upp och se elevens bedömda uppsats. Det gick inte att se vem som hade bedömt elevens svar och det var därför omöjligt för lärarna att veta om det var en central

bedömare, en extern bedömare eller någon av dem själva[58] som bedömt uppsatsen. För enkelhetens skull benämns bedömningen därför alltid som central bedömning nedan.

8.1 Överensstämmelse med lärarnas egen uppfattning

Alla fyra lärare tyckte att elevernas resultat i stort sett överensstämde med den uppfattning de hade om hur eleverna brukade prestera. En lärare från skola 1 nämnde att framför allt rangordningen av elevernas resultat överensstämde med hens egna underlag. En lärare från skola 2 hade kontrollerat en uppsats skriven av en elev som brukade prestera bättre. Hen fick då bekräftat att eleven hade presterat sämre än vanligt på uppgiften och att den centrala bedömningen stämde. När bedömningen låg i linje med den återkoppling som läraren tidigare hade givit eleven om vad som behövde utvecklas uppfattade också en av lärarna att det var ganska skönt och stärkande att få det bekräftat av en central bedömare.

Alla fyra lärare påpekade dock att i de fall det förekommit oväntade resultat, hade det skapat problem vid återkopplingen till eleverna. En av lärarna hade exempelvis inte varit överens med den centrala bedömaren om att en text skulle ha full poäng:

> När jag tittar på helheten skulle jag inte ha bedömt texten högsta nivån på språket men det har den här personen gjort. Det är svårt att förhålla sig till eftersom det här ska vara ganska styrande för betyget, det kan ju vara så, men hur ska man resonera med eleven att hen inte kanske når ett A i betyg trots resultatet på provet.[59]

Ingen av lärarna uppfattade något generellt mönster i de avvikelser som de hade sett. De ansåg att resultaten möjligen hade varit något lägre än väntat, men de menade att det kunde förklaras med att eleverna inte hade fått lika noggranna förberedelser som inför ett nationellt prov. De sade att de exempelvis inte hade varit lika inlästa på källtexterna.

8.2 Resultatens användbarhet

Lagen säger att det är resultatet från det nationella provet som skall beaktas. Eftersom lärarna endast fått elevernas poäng på en uppgift saknades det ett provresultat att beakta. Hur lärarna skulle använda resultatet på elevtexterna från utvärderingen var alltså inte definierat. Lärarna fick därför frågan om hur de tänkte använda elevernas resultat. De såg alla delprovet som en övning inför de nationella proven som skulle gå senare under våren. Lärarna från skola 1 läste samtliga elevers bedömda texter eftersom de såg detta provtillfälle som en möjlighet att ge formativ återkoppling inför vårens nationella prov. Lärarna på skola 2 tänkte ge en generell formativ återkoppling till hela klassen. Det verkade

[58] Om de inte mindes just den texten från sin egen bedömning.
[59] Eleverna fick aldrig betyg på sina uppsatser, och ingen översättning av poäng till betyg hade existerat under utvärderingen, men lärarna refererade vid ett flertal tillfällen under intervjuerna till betygsbeteckningar när de beskrev uppsatsernas kvalitet.

alltså inte som att lärarna tänkte använda elevernas uppsatser som betygsunderlag som det var sagt i utvärderingen.

Lärarna berättade att de, normalt vid genomförande av nationella prov, vanligtvis använde delprovsresultatet som betygsunderlag, förutom att beakta provresultatet. De motiverade det med att provbetyget inte speglade hela kursen och att det därför var enklare att se på vilken nivå som eleverna presterade inom enskilda delar av ämnesplanen. Lärarna från skola 1 påpekade också att det fanns mer information om elevernas kunskaper i texterna som inte riktigt togs till vara genom de tillgängliga bedömningsaspekterna. Läraren kunde därför även läsa en redan bedömd text för att leta efter något särskilt, exempelvis hur eleven hanterade källmaterialet, lite mer i detalj.

8.3 Vad skapar tillit till en annan lärares bedömning?

På frågan om vilken betydelse för din tillit till resultatet bedömarens kompetens hade, svarade alla fyra lärare att de hade känt en viss osäkerhet eftersom de inte visste vilken erfarenhet den centrala bedömaren hade haft.

De var eniga om att de skulle ha högre tillit till bedömare som hade lång erfarenhet av att bedöma olika typer av elevgrupper, exempelvis såväl elever från högskoleförberedande program som yrkesprogram. De fick också gärna vara utbildade av lärosätet som konstruerar det nationella provet. Lärarna från skola 1 menade till och med att det borde vara ett krav. Efter en sådan utbildning skulle bedömningen kunna sägas utgöra en norm för korrekt bedömning och som minskade behovet av att behöva göra en egen ombedömning[60] av elevtexter med överraskande resultat. De betonade också att en viktig kvalitetsstämpel på bedömningar var att det hade skett en sambedömning där flera bedömare hade läst och diskuterat texten. Även lärarna från skola 2 önskade att texterna skulle vara sambedömda.

Lärarna kommenterade också det egna arbetet med sambedömning i försöket. Lärarna från skola 2 betonade vikten av att få diskutera bedömning med kollegor, så att man inte utvecklar en egen bedömningspraktik på nationella prov och sticker i väg åt något håll. De två lärarna från skola 1 menade att, eftersom de hade bedömt exempeluppsatserna individuellt först hade sambedömningen blivit väldigt effektiv och de hade nått samsyn snabbare än vad de hade förväntat sig. De hade jämfört sina olika poäng de hade satt på aspekterna och hade fokuserat på olikheterna. För den ena läraren hade det blivit tydligt att hen inte varit färdig med den egna bedömningen bara för att poängen var satta. Hen menade att tilliten främst handlade om kännedomen att fler hade tänkt tillsammans. För läraren var det nämligen viktigt att hen hade fått verbalisera det som vid den individuella bedömningen bara varit en känsla:

> Det blev så himla tydligt att vid den första bedömningen när jag hade läst en text var jag absolut inte var färdig med min bedömning. Att

[60] Det fanns ingen möjlighet för lärarna att ändra poängen från en genomförd bedömning i plattformen.

> bara klicka i. Jag litar inte på min egen bedömning utifrån de förutsättningarna.

Alla fyra lärare betonade också att den centrala bedömningen skulle ha underlättats om det funnits fler exempel på bedömda elevexempel i bedömningsanvisningarna. Alla aspekter på de olika nivåerna borde ha funnits representerade, menade de.

Det finns ingen möjlighet för en bedömare att skicka med en motivering eller anteckningar i en bedömd text i Skolverkets provplattform[61]. Därför ställdes en fråga om vilken betydelse det hade att bedömarens resonemang hade saknats i resultatet. Alla fyra lärare menade att det hade haft stor betydelse och de betonade att poängen inte säger allt om en texts kvalitet. Lärarna lyfte behovet av att få en kort förklaring till bedömningen för att själv kunna ta ställning till den och återkoppla till eleven. Men de nämnde också att anteckningar även behövdes för bedömarens egen skull. En lärare påpekade att det var lätt att glömma bort fel som eleven begått i texten om man skulle vänta med att sätta poäng tills hela texten var genomläst. Den bedömande läraren kunde också behöva gå tillbaka till anteckningar i texten för att kontrollera den egna bedömningen. Anteckningarna i texten skulle alltså bli ytterligare en kvalitetsstämpel på arbetet som den centrala bedömaren hade genomfört.

En fråga gällde om lärarna hade haft större tillit till bedömningen av en kollega på skolan än en extern eller central rättares bedömning och i så fall varför. Lärarna menade att det i princip inte hade spelat någon roll, men lärarna konstaterade ändå att skolor hade olika bedömningskulturer som kunde påverka bedömningen. Det kunde exempelvis vara skillnader i bedömningen beroende på om en bedömare arbetade på en skola med många lågpresterande elever. En av lärarna på skola 1 gav exempel från tidigare arbete vid en skola där eleverna hade haft lägre kunskapsnivå än där hon arbetade nu. För att eleverna skulle uppnå ett godkänt betyg hade lärarna behövt lägga ner väldigt mycket arbete på att stötta elever, exempelvis med mycket formativ bedömning. Ibland kunde då läraren, som kände eleven, se till att elevernas resultat precis tippade över till ett E på delprovet. En central bedömare skulle möjligen i de fallen göra en annan bedömning.

[61] Möjligheten att skicka med en motivering till poängsättningen har inte implementerats i plattformen först och främst av etiska skäl. Om centrala bedömare skulle skicka med motiveringar skulle dessa behöva granskas så att de inte innehåller något som skulle kunna uppfattas som kränkande. Det skulle inte vara praktiskt möjligt att granska alla tusentals motiveringar för alla tänkbara reaktioner från lärare och elever.

9. Resultatdiskussion

I detta kapitel diskuteras de olika utvärderingsfrågorna var för sig.

9.1 Skillnad på reliabilitet mellan central rättning, extern bedömning och bedömning av elevens undervisande lärare

Skolverkets utvärdering ger inte något stöd för att reliabiliteten av uppsatstexter blir högre om de bedöms av andra bedömare än elevernas egna lärare. Reliabiliteten ökade inte ens när bedömarna var legitimerade, behöriga, hade genomgått en särskild bedömarutbildning för den specifika uppgiften i provet eller när de sambedömde i par.

Central bedömning av standardiserade prov är i regel standard runtom i världen, men eftersom få länder verkar kombinera centraliserad och decentraliserad bedömning av samma prov är det är svårt att hitta liknande jämförelser. Resultaten i den här utvärderingen stämmer dock överens med studier genomförda i Bhutan där man i stället gick från central till decentraliserad bedömning av nationella prov. I dessa studier förändrades reliabiliteten obetydligt i matematik[62] respektive engelska[63].

9.1.1 Inga effekter av sambedömning

Man skulle kunna förvänta sig att reliabiliteten skulle öka med sambedömning. Bedömare som avviker, brukar ofta modereras mot övriga bedömare som de sambedömer med för den aktuella bedömningsuppgiften.[64] Någon långvarig effekt av sambedömning, d.v.s. att lärare som sambedömer får en mer likvärdig betygssättning har däremot varit svår att påvisa empiriskt.[65]

Lärarna uttryckte en tydligt positiv attityd och förtroende för sambedömning i såväl enkäten som intervjuerna. I utvärderingen syns en något högre korrelation på sambedömningen i gymnasieskolan, men sambedömarna i grundskolans två prov avviker inte positivt. Att sambedömarna ligger högre än övriga bedömargrupper i gymnasiet ska man därför vara försiktig att dra några slutsatser av. Ytterligheter, i form av de strängaste respektive mest generösa bedömningarna, borde ha försvunnit, men skillnaden mellan bedömarparen ser, i stort, inte ut att ha minskat jämfört med de individuella bedömarna. Möjligtvis

[62] Jurmi, C. (2003). *A Comparison of Examination Scores Under Decentralized and Centralized Systems of Marking in Class VI Mathematics in Bhutan*. National Library of Canada= Bibliothèque nationale du Canada, Ottawa.
[63] Dolkar, D. (2009). *Studying school-based summative assessments in high-stakes examinations in Bhutan: A question of trust?* (Master's thesis, University of Twente).
[64] Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education*, *41*(4), 638–653.
[65] Ibid.

skulle återkommande parbyten och ombedömningar ha fått bedömarna att närma sig varandra.

Intressant är också vad som kommer fram i enkätsvaren om sambedömning. Trots att bedömningsutfallen mellan lärare skiljde sig åt ansåg majoriteten av de lokala och externa bedömarna att de redan var överens och inte ändrade sin bedömning i någon större omfattning vid sambedömningen. Om sambedömningen inte medförde att bedömare ändrade sina poäng så är det inte underligt att avvikelsen mellan bedömarna bibehållits. Endast några av de centrala bedömarna menade att sambedömningen medförde en omvärdering och att de gick vidare från sambedömningen med ny kunskap. Lärarnas upplevelse av att inte ha förändrat sin syn på texters kvalitet utifrån sambedömningen, stämmer överens med svårigheterna att påvisa någon positiv långvarig effekt av sambedömning i forskning.[66] En förklaring till att de lokala bedömarna inte ändrade sig kan vara att de lokala bedömarna var kollegor och hade kanske redan utvecklat en samsyn på skolan. De externa bedömarna hade däremot inte träffat varandra tidigare och borde därför ha kunnat lära sig något nytt av varandra.

9.1.2 Inga effekter av bedömarutbildning

De centrala bedömarna i utvärderingen uppvisade inte högre reliabilitet än de externa eller lokala bedömarna trots att de genomgått en utbildning. Att uppnå en hög reliabilitet på längre fritextsvar med högre komplexitet, som uppsatsskrivning, har visat sig vara utmanande även internationellt.[67] Det finns dock exempel på lyckade projekt där man genom träning och bevakning av expertbedömare kunnat höja såväl konsensus som konsistens bland oerfarna bedömare.[68]

Det är inte förvånande att bedömarutbildning, i sig, inte skapade högre reliabilitet för bedömning av uppsatser. Bedömarträning behöver inte betyda att bedömarna blir mer överens om en uppsats poäng, men bedömare som avviker mycket från de andra kan komma närmare de andra bedömarna.[69] De uteblivna effekterna av bedömarutbildningen kan bero på att skrivuppgifterna i svenska och engelska var så öppet utformade att det blir svårt att träna bedömare mot mer enhetlig bedömning. Eftersom eleverna fick frihet att utforma sin text med stor variation behövde bedömningsanvisningar och bedömarutbildningar bli så generella att de hanterade den innehållsliga bredden av elevtexter. Denna generalisering behövde göras på bekostnad av tydligheten, vilket riskerade att olika bedömare tillämpade bedömningsinstruktionerna på olika sätt. Bedömare behövde, trots omfattande utbildning, till syvende och sist, ändå värdera många unika fall och kunde komma att poängsätta dessa olika. Liknande resonemang fördes vid en studie bland

[66] Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education*, *41*(4), 638–653
[67] Brown, G. T., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing writing*, *9*(2), 105–121.
[68] Ibid.
[69] Weigle, S. C. (1994). Effects of training on raters of ESL compositions. Language Testing, 11(2), pp. 197–223

bedömare i Australien[70] där små effekter på reliabilitet, speciellt bland erfarna bedömare, kunde uppmätas även efter mer omfattande bedömarutbildning. Det bör dock sägas att det inte är en given lösning för ökad reliabilitet att i större omfattning styra upp elevernas skrivande mot givna mallar för att det skall bli lättare att bedöma dem eftersom mycket av kurs- och ämnesplanernas syften skulle gå förlorade i de nationella proven då.

9.1.3 Skillnader i effekter av avvikelser på aspekt- respektive helhetsnivå

Reliabiliteten var något högre för proven i engelska, som använde en holistisk rapportering av poäng, än för proven i svenska som använde en rapportering genom analytiska aspekter. Även om motiveringen till analytisk bedömning ofta är att uppnå högre reliabilitet så ger holistisk bedömning i flera fall högre reliabilitet.[71] Det märks speciellt vid konsistensmåtten eftersom bedömarna lättare kan komma överens om hur uppsatserna skall rangordnas holistiskt. Värderingar av enskilda aspekter riskerar dock att stöka om i rangordningen mellan bedömare i aspektbedömning. Holistisk bedömning ger dock inte lika stor fördel i konsensusestimat eftersom de inte behöver blir lättare att komma överens om vilken poäng uppsatsen skall få holistiskt jämfört med analytiskt. Man kan argumentera för att de små avvikelserna på aspekterna, åtminstone de som var av slumpmässig karaktär, till viss del tog ut varandra när bedömning aggregerades upp på en högre nivå. Det är viktigt att understryka, att det dock alltid kommer finnas enskilda elever som drabbas på så vis att de slumpmässiga bedömningsavvikelserna leder till ett mer än genomsnittligt ”fel” om de har otur – eller tur om bedömningarna faller väl ut för dem. Skulle bedömningsavvikelserna i stället bero på systematiska avvikelser, så kallade bedömareffekter, så blir saken än mer intrikat och det är inte självklart hur det skulle slå mot olika elever och deras resultat i termer av likvärdighet.

9.2 Effekt på provbetyget

I sammanhanget ska det betonas att endast bedömningen av elevers uppsatsskrivning har utvärderats. Bedömning av uppsatser som uppgiftsformat är den bedömning som är mest komplex att göra och som därmed i regel ger lägst reliabilitet.[72] Endast bedömningen av muntlig produktion, som dessutom är beroende av vad en observerande bedömare uppmärksammar under elevens framställning, brukar ha lägre reliabilitet. Man kan således argumentera för att det som utvärderats är ett av de två delprov som ger sämst reliabilitet och att den här utvärderingens resultat därmed kan symbolisera en lägsta nivå för vad man skulle

[70] Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49–58.
[71] Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing writing*, *12*(2), 86–107.
[72] Brown, G. T., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing writing*, *9*(2), 105–121.

kunna åstadkomma i termer av ökad reliabilitet genom införande av central rättning eller extern bedömning.

Uppsatsskrivningen har olika betydelse för provresultatet i de tre proven som användes i utvärderingen. Eftersom provet för svenska och svenska som andraspråk för kurs 3 endast består av en del för muntlig och en för skriftlig framställning blir det provbetyget ytterst känsligt för bristande reliabilitet. Effekten på ökad likvärdighet av central rättning kommer begränsas på olika sätt beroende på om prov är sammansatta av uppgiftsformat som i termer av bedömningskomplexitet är mer eller mindre komplexa.

9.3 Bedömarnas uppfattning av bedömningen

Upplevelsen av bedömning i provplattformen skiljde sig en del beroende på i vilket ämne och i vilken årskurs/kurs som man bedömde elevtexter i. De som bedömde svenska åk 6 hade varit genomgående positiva till bedömningssituationen. De hade inte haft problem att komma överens vid sambedömningen och de var nöjda med att de fått möjlighet att delta i försöket. De som bedömde i engelska hade uppskattat att bedömningen varit enkel och gått snabbt att genomföra. De hade inte heller sett några större problem med att läsa texterna digitalt. Däremot tyckte flera av bedömarna i engelska att det hade varit svårt att bedöma med poäng i stället för med delprovsbetyg. Till skillnad från bedömningsanvisningen i svenska, fanns ingen generell beskrivning av poängnivåerna för engelska utan dessa exemplifierades genom bedömda och kommenterade uppsatser. Poängen kan därför ha varit mer abstrakta för bedömarna i engelska än för de som bedömde uppsatser i svenska. Bedömarna i kurs 3 såg de största problemen med digital bedömning eftersom det inte gick att skriva ut elevtexterna eller skriva anteckningar i dem. De hade också de längsta och mest avancerade uppsatserna vilket troligtvis medförde störst utmaningar med att värdera kvaliteten utifrån de olika bedömningsaspekterna.

De externa och de lokala bedömarna hade haft lättare att komma överens med kollegan vid sambedömningen än de centrala bedömarna. Centrala bedömare hade också ändrat sina poäng i högre grad än de övriga grupperna. En klar majoritet av de centrala bedömarna menade att deras arbete med bedömningen hade påverkat deras uppfattning om elevtexters kvalitet, vilket inte var lika tydligt bland externa och lokala bedömare. Att de centrala bedömarna i högre grad ansåg att deras bedömningskompetens utvecklats av sambedömningen kan möjligen förklaras av att de hade gått en utbildning som de hade upplevt som positiv. Den gemensamma utbildningen kan ha öppnat för en bredare och djupare diskussion om de olika elevtexterna under sambedömningen. Detta kan ha inneburit en mer omfattande förhandling i sambedömningen som hade resulterat i en tydligare känsla av utveckling av synen på kvalitet. Det syns dock inte i måtten för konsensus att bedömarna som grupp närmat sig varandra i sambedömningen.

Bedömarna i gymnasieskolan var i högre grad missnöjda med möjligheterna till bedömning i provtjänsten än grundskolans bedömare. Dessa olika uppfattningar

kan troligen förklaras med skillnader i de olika konstrukten, det vill säga kunskapen som skulle mätas, som låg till grund för uppgiftskonstruktionen. Bedömarna i svenska åk 6 hade visserligen flest bedömningsaspekter att ta ställning till, men texterna var mindre komplexa än motsvarande på gymnasieskolan. Proven i svenska kurs 3 var därmed troligen de mest utmanande att bedöma. De var långa med ett stort utredande innehåll och många språkliga element att ta ställning till och värdera för en bedömare. Enkätsvaren indikerar att dessa uppsatser ställde högre krav på den digitala bedömningen och det har möjligtvis blivit svårare att upprätthålla en hög reliabilitet.

9.4 Lärarnas tillit

De fyra intervjuade lärarna tyckte att de bedömda elevtexterna som de fick tillbaka i stort sett hade stämt med deras egna betygsunderlag. Men de påpekade att alla avvikelser från det egna underlaget skulle medföra problem, bland annat när resultatet skulle återkopplas till eleven. De hade inte sett några särskilda mönster i avvikelserna, förutom att resultaten ibland blivit lite lägre, vilket de förklarade med att förberedelserna inte genomfördes lika noggrant som vid det ordinarie provtillfället. Lärarna var eniga om att en försäkran om den centrala bedömarens kompetens skulle ha stor betydelse för deras tillit till resultaten. De menade också att de skulle ha högre tillit om elevtexterna var sambedömda av flera bedömare. Däremot skulle det inte spela någon större roll om bedömaren kom från den egna skolan eller om hen kom utifrån. Med andra ord kan man säga att lärarna inte uttryckte några större bekymmer av att bedömningen gick dem ur händerna så länge bedömningen hade säkrat god kvalitet genom utbildning och sambedömning.

Lärarna utryckte alla att bristen på motivering till bedömningen och uteblivna kommentarer i elevtexterna från bedömaren varit en stor brist. De behövde en kort återkoppling från bedömaren för att kunna återkoppla bedömningen till eleverna och de tyckte att deras egen erfarenhet av att bedöma nationella prov visade att det var viktigt att även som bedömare kunna skriva anteckningar när de läste och bedömde texterna. Detta behov ställs mot behovet av att förenkla och effektivisera bedömningen. Genom att ersätta dokumentation av kommentarer på papper med knapptryckningar går den digitala bedömningen snabbare och blir enklare, men information går även förlorad i effektiviseringen.

Ingen av de fyra intervjuade lärarna verkade ha något större problem med att lämna ifrån sig bedömningen till bedömare de inte kände, så länge de fick någon slags garanti för att bedömarna är kompetenta och har erfarenheter av uppsatser från såväl högpresterande som lågpresterande elever. De såg också sambedömning som en kvalitetsstämpel, även om de inte själva hade ändrat något vid sambedömningen. Lärarna framförde flera gånger tankar om att olika lärare bedömer uppsatserna utifrån varierande referenser och de var medvetna om att bedömningarna av en och samma text kunde skilja sig åt exempelvis beroende på om en lärare arbetade på en skola med hög- respektive lågpresterande elever.

Därför såg de behov av att kunna göra en värdering av bedömarnas kvalitet och för det krävdes information om hur bedömaren hade resonerat. I dag finns inte någon möjlighet att en bedömare skickar med en motivering till sin bedömning i elevernas resultatrapport. Utmaningarna med medskickade motiveringar är inte bara tekniska. Förutom att en utlovad effektivisering som följd av digitalisering av proven skulle minska skulle sådana kommentarer även behöva granskas så att de inte uttrycker något som skulle kunna väcka anstöt hos betygssättande lärare eller elev, vilket skulle medföra en omfattande arbetsinsats i en centraliserad bedömningsorganisation.

De intervjuade lärarna hade svårt att se några mönster i de avvikelser som de såg mellan det egna betygsunderlaget och resultatet på den centralt bedömda uppsatsen. Möjligen såg de en tendens att elever som behövde mycket stöd i undervisningen skulle få en strängare bedömning. I det sammanhanget är det viktigt att utifrån ett likvärdighetsperspektiv även förklara hur olika former av bedömning eventuellt påverkar olika elevgrupper. Skulle det till exempel vara så att uppsatser av lägre kvalitet i större utsträckning skulle bedömas strängare när central rättning införs?

10. Slutsatser

I det här avsnittet redovisas implikationer om vilka konsekvenser ett införande av central rättning eller extern bedömning skulle kunna få utifrån resultat om uppnådd reliabilitet och bedömarnas upplevelser.

10.1 Inget stöd för att enbart central rättning skulle öka likvärdigheten

En viktig slutsats är att utvärderingen inte gav stöd för att bedömningen kommer att bli mer reliabel enbart genom att förflytta bedömning av de provdelar som prövar skriftlig produktion från skolans lärare till centralt utbildade och organiserade bedömare. Det är dock inte möjligt att dra slutsatser om effekterna på likvärdigheten i provbetygen, baserat på utvärderingens resultat, då den empiri som använts är mer lämpad för ett kvalitativt angreppssätt. Det går med andra ord inte att dra några generaliserbara slutsatser om populationen bedömande lärare i relation till populationen elevlösningar.[73]

[73]Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication methods and measures*, *5*(2), 93–112.

10.2 Likvärdighet över tid

En annan slutsats är att utvärderingen inte gav något stöd för att reliabiliteten kommer att förbättras på kort tid. De centrala bedömarna upplevde i något högre grad att de närmade sig varandra i sambedömningen än de externa och lokala bedömarna, men reliabilitet mellan paren i hela bedömarpopulationen ökade dock inte jämfört med bedömarnas individuella bedömning. Det är dock möjligt att flera på varandra följande år av centraliserad bedömning med organiserad sambedömning ger en mer likvärdig bedömning.[74] De centrala bedömarna blir kanske med tiden, som kollektiv, mer samspelta med varandra genom att varje år sambedöma i olika team, med bedömare från olika delar av landet, som leds av en erfaren huvudbedömare. Det är ett sådant upplägg som Skolverket föreslår i 2022 års regeringsredovisning.[75] Vidare vore det viktigt ur ett likvärdighetsperspektiv att studera hur man i ett system med central rättning på bästa sätt skulle kunna komma till rätta med fler typer av bedömareffekter än vad som undersöktes i utvärderingen. I utvärderingen studerades inte hur väl bedömare gör samma bedömningar upprepat över tid, så kallad, *intra*bedömarreliabilitet. Andra effekter som skulle behöva studeras är till exempel om centrala bedömare, över tid, blir strängare, så kallad *rater drift*.

10.3 Risk för sämre provresultat

En risk som lyfts i de intervjuade lärarnas utsagor är att centrala och externa bedömare skulle kunna vara strängare, i förhållande till lokala bedömare. Lokala bedömare skulle kunna tänkas vara mer generösa om det gällde ordinarie obligatoriska nationella prov vars resultat särskilt skulle beaktas i betygssättningen. Studier visar ofta att expertbedömare tenderar att vara strängare än elevernas egna lärare vid ombedömningar.[76] Det skulle i så fall kunna innebära att andelen elever som får F i provbetyg riskerar att öka när central rättning införs.

Så länge olika uppgifter i en elevs prov bedöms av olika bedömare så borde effekten av stränga bedömare tas ut av andra mer generösa bedömare för flertalet elever. En enskild elev kan dock råka ut för bara bedömare som är strängare eller generösare än genomsnittet.

10.4 Större skillnad mellan betyg och provbetyg

En fjärde slutsats skulle kunna vara att skillnaden mellan betyg och provbetyg skulle kunna öka. En lärare som ställer låga krav för de olika betygsstegen är även generös i sin bedömning av nationella prov. Om central rättning fungerar som önskat, kommer centralt bedömda svar i högre grad spegla den genomsnittliga lärarens krav. Eftersom poängen från den centralt bedömda delen av provet endast

[74] Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors?. *Language Testing*, *37*(3), 311–332.
[75] Skolverket (2022) Redovisning av uppdrag om att införa central rättning av nationella prov U2021/03346.
[76] Kuhlemeier, H., & Kremers, E. (2013). De praktijk van de eerste en tweede correctie. Samenvatting van onderzoek naar het functioneren van het CSE. In *Technical Report*. Arnhem: Cito.

utgör en liten del av provresultatet behöver inte effekten av en ökad reliabilitet bli så stor på elevpopulationens provbetyg. En enskild elevs poäng kan dock hamna ovanför eller under en kravgräns för ett provbetyg. Tidigare kunde då elevens lärare justera poängen så att eleven hamnade över kravgränsen. Då provets övriga delar i digitala prov delvis automaträttas, innebär däremot central rättning av skriftlig produktion att skolan får relativt små möjligheter att justera elevers resultat mot önskade provbetyg.[77] Skolan och den betygssättande läraren kommer därför behöva acceptera den eventuella skillnad som uppstår mellan provresultatet och den bedömda nivån på lärarens eget betygsunderlag. Det finns dock ingen gräns för hur mycket elevers betyg får avvika från de nationella provens betyg. När underlaget från ett nationellt prov avviker från lärarens bild av elevens kunskapsnivå, avgörs elevens betyg av hur läraren värderar giltigheten av sitt eget underlag i relation till det nationella provresultatet. Det är inte alls säkert att en lärare omprövar giltigheten i sitt eget underlag, utan snarare ogiltigförklarar provresultatet. Det är därför möjligt att vi kommer se ökade skillnader mellan betyg och provbetyg efter ett införande av central rättning.

10.5 Ytterligare systemåtgärder är nödvändiga för att betygssättningen ska bli mer likvärdig

Skolverket har tidigare uppmärksammat de likvärdighetsproblem och elevers rättssäkerhet som dagens prov- och betygssystem ger upphov till[78] och anser att en reform som avser att centralisera bedömningen av de nationella proven inte är tillräcklig för att komma till rätta med de problem som finns i systemet[79].

Oavsett hur reliabel Skolverket kan få den centrala rättningen att bli, så kan inte central rättning i sig anses vara en åtgärd för att öka likvärdigheten i betygssättning. Så länge det är upp till varje lärare att själv bestämma hur det nationella provresultatet ska användas vid betygssättningen går det inte att hävda att betygssättningen kommer att bli mer likvärdig i landet bara för att bedömningen av svaren blir mindre beroende av vilken lärare eleven har. Skolverket har lämnat förslag på hur regeringen skulle kunna öka likvärdigheten i betygssättningen med olika modeller för hur nationella provresultat kan styra betygssättningen.[80] Dessa modeller förutsätter central rättning, just för att förhindra skolors möjlighet att påverka provresultaten. Skolverket uppmanar regeringen ännu en gång att se över Skolverkets förslag och inte nöja sig med ett införande av central rättning som åtgärd för att åstadkomma en mer rättvis och likvärdig betygssättning.

---

[77] Diamond, R., and P. Persson. 2016. "The Long-Term Consequences of Teacher Discretion in Grading of High-Stakes Tests". (No. w22207). National Bureau of Economic Research.

[78] Skolverket (2020a). Grundskolors systematiska avvikelser i betygssättning och elevers gymnasieresultat. Skolverkets rapport 2020:7

[79] Skolverkets (2020b) *Likvärdiga betyg och meritvärden.* Rapport 2020:7

[80] Ibid.

Litteraturförteckning

Barkaoui, K. (2007). Rating scale impact on EFL essay marking: A mixed-method study. *Assessing writing*, *12*(2), 86–107.

Barrett, P. (2001). Assessing the reliability of rating data. Retrieved from http://www.pbarrett.net/presentations/rater.pdf

Barrett, S. (2001). The impact of training on rater variability. *International Education Journal*, *2*(1), 49–58.

Bloxham, S., Hughes, C., & Adie, L. (2016). What's the point of moderation? A discussion of the purposes achieved through contemporary moderation practices. *Assessment & Evaluation in Higher Education*, *41*(4), 638–653.

Brown, G. T., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing writing*, *9*(2)

Diamond, R., & Persson, P. (2016). *The long-term consequences of teacher discretion in grading of high-stakes tests* (No. w22207). National Bureau of Economic Research.

Casabianca, J. M., & Wolfe, E. W. (2017). The impact of design decisions on measurement accuracy demonstrated using the hierarchical rater model. *Psychological Test and Assessment Modeling*, *59*(4)

Dolkar, D. (2009). *Studying school-based summative assessments in high-stakes examinations in Bhutan: A question of trust?* (Master's thesis, University of Twente).

Gustafsson, J. E., & Erickson, G. (2013). To trust or not to trust? – teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability*, *25*

Jurmi, C. (2003). *A Comparison of Examination Scores Under Decentralized and Centralized Systems of Marking in Class VI Mathematics in Bhutan*. National Library of Canada Bibliothèque nationale du Canada, Ottawa

Krippendorff, K. (2011). Agreement and information in the reliability of coding. *Communication methods and measures*, *5*(2), 93–112.

Kuhlemeier, H., & Kremers, E. (2013). De praktijk van de eerste en tweede correctie. Samenvatting van onderzoek naar het functioneren van het CSE. In *Technical Report*. Arnhem: Cito.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia medica*, *22*(3)

Riksdagen. Förordning (2017:1106) om en försöksverksamhet med datorbaserade nationella prov, extern bedömning och central rättning.

Riksdagen. Skolförordning (2011:185).

Şahan, Ö., & Razı, S. (2020). Do experience and text quality matter for raters' decision-making behaviors?. *Language Testing*, *37*(3), 311–332.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, *86*(2)

Skolinspektionen: Ombedömning av nationella prov 2019 Diarienummer 2019:503

Skolverkets allmänna råd (SKOLFS 2022:417) om betyg och prövning

Skolverket (2019a). *Grundskolors systematiska avvikelser i betygssättning och elevers gymnasieresultat*. Skolverkets rapport 2020:7

Skolverket (2019b). *Analyser av likvärdig betygssättning mellan elevgrupper och skolor.* Rapport 475.

Skolverket (2020a). *Analyser av likvärdig betygssättning i gymnasieskolan. Jämförelser mellan kursbetyg och kursprov*. Rapport 2020:3.

Skolverkets (2020b) *Likvärdiga betyg och meritvärden.* Rapport 2020:7

Skolverket: Redovisning av uppdrag om att införa central rättning av nationella prov U2021/03346

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. Practical Assessment, Research, and Evaluation, 9(1)

Stemler, S.E.; Tsai, J (2008). Best Practices in Interrater Reliability: Three Common Approaches. In Best Practices in Quantitative Methods; Osborne, J.W., Ed.; Sage: Los Angeles, CA, USA

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing*, *11*(2)

Material

Frisläppta nationella prov som ingick i utvärderingen

Ämnesprov 2017/2018. Svenska och svenska som andraspråk, årskurs 6, delprov C 1. Den magiska silverringen.

Ämnesprov 2015/2016. English, årskurs 9, delprov C. My media.

Kursprov vårterminen 2015. Svenska och svenska som andraspråk, 3, delprov A. Det var en gång, uppgift 1 Törnrosa lever.

Bilagor

Bilaga 1. Enkätfrågor

1. Hur lång erfarenhet har du av att bedöma nationella prov?
   - 10 år eller mer
   - 5 - 9 år
   - 3 - 5 år
   - 1 - 2 år
   - Jag är nybörjare

2. Jag bedömde elevtexter i
   - svenska åk 6
   - engelska åk 9
   - svenska kurs 3

3. Jag var
   - lokal bedömare
   - extern bedömare
   - central bedömare

4. Att bedöma uppsatser digitalt jämfört med på papper gick
   - mycket snabbare
   - snabbare
   - långsammare
   - mycket långsammare

5. Bedöma genom att klicka på knappar jämfört med att dokumentera på papper var
   - mycket enklare
   - enklare
   - svårare
   - mycket svårare

6. Att läsa texterna på datorn gjorde bedömningen
   - mycket enklare
   - enklare
   - svårare
   - mycket svårare
   - Kommentar:

7. Att bedöma med poäng i stället för betygsbeteckningar var
    - mycket enklare
    - enklare
    - svårare
    - mycket svårare

8. Bedömningsexemplen utnyttjade jag
    - genom att läsa dem grundligt innan jag började bedöma
    - genom att använda dem som referens när jag blev osäker
    - genom att använda dem både inför och under bedömningen
    - inte alls eftersom jag inte behövde dem
    - inte alls, eftersom jag inte hittade dem

9. Instruktionsrutan för bedömningsaspekterna klickade jag fram? (svenska)
    - hela tiden
    - mycket i början, sedan någon gång
    - bara i början
    - inte alls eftersom jag inte behövde dem
    - inte alls, eftersom jag inte hittade dem
    - inte alls eftersom jag bedömde engelska

10. Jag ändrade mina poäng under sambedömningen
    - i de flesta fall
    - i mer än hälften
    - i färre än hälften
    - i inget eller i något enstaka fall

11. När vi skulle sätta en gemensam poäng under sambedömningen
    - var det inga problem alls att komma överens
    - krävdes det en omfattande diskussion
    - kom vi inte överens, men jag gav mig
    - kom vi inte överens, men jag lyckades övertyga den andra bedömaren
    - kompromissade vi om poängen

12. Jag tror att sambedömningen har påverkat min uppfattning om elevtexters kvalitet
    - i hög grad
    - i ganska hög grad
    - i mindre grad
    - inte alls

13. Utifrån din erfarenhet, vad kändes positivt med bedömning i provplattformen jämfört med att bedöma på papper?

14. Utifrån din erfarenhet, vad kändes negativt med bedömning i provplattformen jämfört med att bedöma på papper?

15. Vi tar tacksamt emot övriga synpunkter

Frågor som bara ställdes till de centrala bedömarna som genomgick en utbildning i bedömning.

Bedömarutbildningen gjorde mig till

- en betydligt säkrare bedömare
- en säkrare bedömare
- en lika säker bedömare som innan
- en osäkrare bedömare

Att bygga utbildningen på bedömning av och diskussion om elevexempel

- fungerade mycket bra
- fungerade bra
- fungerade ganska dåligt
- fungerade inte alls

Att bygga utbildningen med elevtexter utifrån aktuellt prov är

- viktigt
- mindre viktigt
- saknar betydelse

Bilaga 2 Intervjufrågor

Överensstämde den bedömning du fick tillbaka med din uppfattning om elevernas kunskaper?

Om inte, vilka avvikelser såg du? Uppfattade du något generellt mönster?

Hur tänker du använda resultatet?

Vilken betydelse för din tillit till resultatet har bedömarens kompetens?

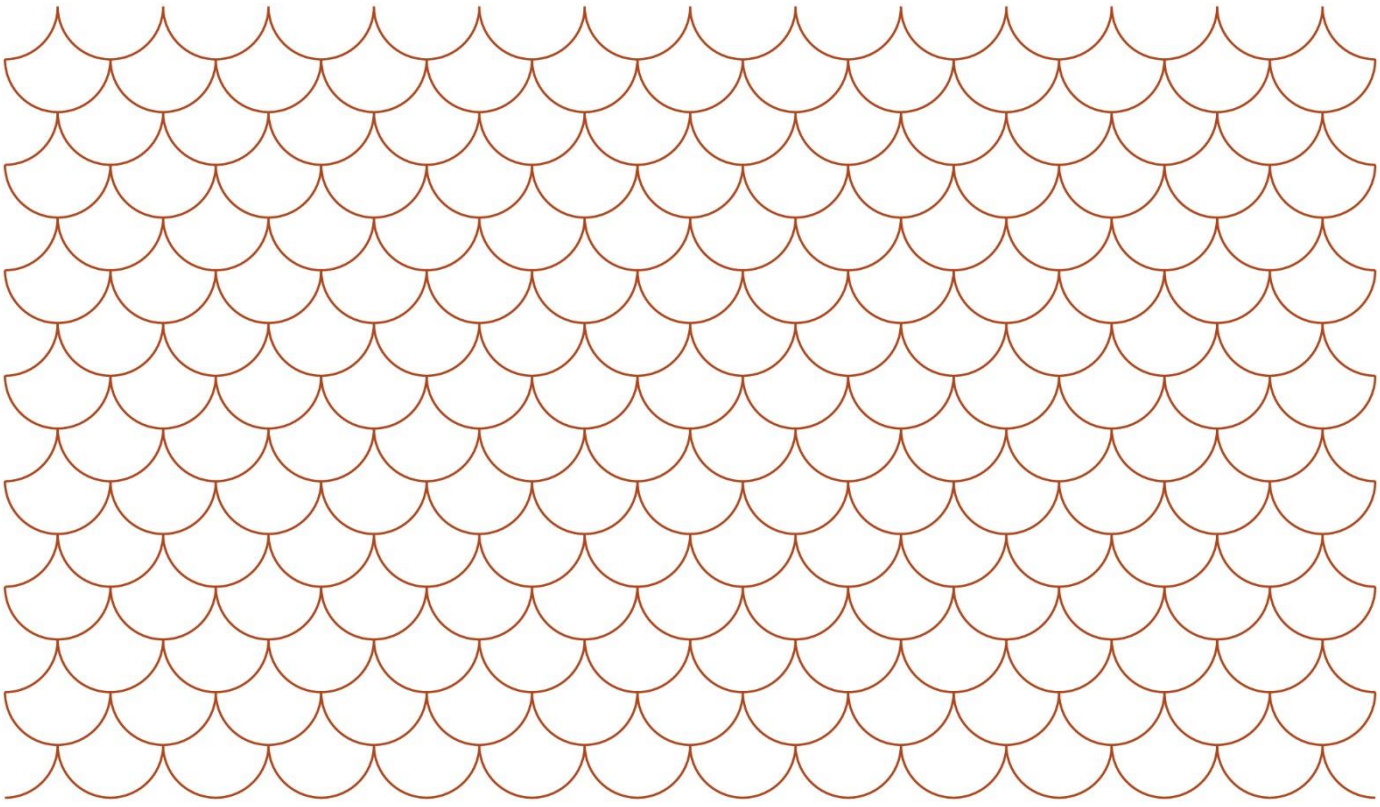
Vilken betydelse har det att bedömarens resonemang saknas i resultatet?

Har du större tillit din en kollega på skolans bedömning än en extern eller central rättares bedömning. Varför i så fall.

Om eleven får ett lägre provbetyg på grund av uppsats delen skulle de påverka din betygssättning på ett annat sätt än hur du skulle resonera i dag.

Skolverket har fått i uppdrag av regeringen att utvärdera hur central rättning och extern bedömning kan påverka likvärdigheten i bedömningen av de nationella proven inom ramen för försöksverksamheten.

Skolverkets utvärdering ger inget stöd för att bedömningen av uppsatser i nationella prov blir mer likvärdig av att bedömaren är någon annan än elevens betygsättande lärare, det vill säga extern bedömning. Utvärderingen ger inte heller något stöd för att bedömningen blir mer likvärdig om bedömningen genomförs av legitimerade, behöriga lärare som genomgått en extra bedömarutbildning av Skolverket, det vill säga central rättning.



Skolverket

www.skolverket.se