Nr 37

Samstämmighet i läsbedömning

Statistisk analys vid bedömning av ett nationellt läsförståelseprov

Tobias Dalberg, Martina Zachiu,
Negin Shahsavar, Kristina Eriksson
& Siri Hussenius

2020

Abstract

Tobias Dalberg, Martina Zachiu, Negin Shahsavar, Kristina Eriksson och Siri Hussenius: *Samstämmighet i läsbedömning. Statistisk analys vid bedömning av ett nationellt läsförståelseprov*. Svenska i utveckling nr 37. Uppsala universitet, Uppsala.

Rater agreement in reading assessment. Statistical analysis of rating reading comprehension in Swedish national tests. In order to support consistent and commensurate grading, which has important consequences for individual pupils, tests need to demonstrate high levels of inter-reliability. Rater agreement is an important component of the overall test inter-reliability needed when assessing complex tasks such as written or oral assignments. In this report, we analyse rater agreement of constructed response items in reading comprehension tests assigned to students within the framework of the national test in Swedish 1 at upper secondary school. The purpose of the analysis is to examine agreement levels between different teachers' ratings, and predict expected levels of reliability, given factors such as number of items and raters involved. We observe rater agreement using a range of measures based on different assumptions about whether raters should only be consistent, or whether they should reach consensus. The statistical measures include exact agreement, Cohen's and Fleiss' κ statistics, intraclass correlation coefficient, many-facets Rasch measurement and generalizability coefficients. To aid the interpretability of our generalizations, we also provide probability estimates based on Bayesian inference. We find the reliability of ratings of constructed response items to be well within acceptable range, and on par with levels accepted by large-scale international tests such as PISA. Finally, the results indicate that the overall reliability of the test would benefit more from increasing the number of items rather than increasing the number of raters.

Kph, Trycksaksbolaget, Uppsala 2020

Innehåll

Förord

Rapportserien Svenska i utveckling ges ut vid Institutionen för nordiska språk, Uppsala universitet. I serien redovisas främst studier som utförs med de nationella proven i svenska och svenska som andraspråk som underlag, men även andra arbeten med anknytning till skola och utbildning. Rapporten *Samstämmighet i läsbedömning* är tillkommen som ett led i provgruppens utvecklingsarbete.

I ett läsförståelseprov kan eleven ställas inför uppgifter av flervalstyp eller uppgifter som leder till elevkonstruerade svar. Medan bedömningen av flervalsuppgifter kan antas vara oproblematisk innebär uppgifter med elevkonstruerade svar en större utmaning för bedömaren. I ett provsystem där ett stort antal bedömare behandlar elevsvaren behöver de olika bedömarnas tolkningar vara samstämmiga. Annars skapas ett reliabilitetsproblem, och följden blir ett hot mot möjligheten att göra valida tolkningar av provresultatet.

I denna rapport studeras samstämmigheten i det läsförståelseprov som ingår i det nationella provet för kursen svenska 1 och svenska som andraspråk 1 i gymnasieskolan. En rad statistiska metoder används i analysen, och den sammanfattande slutsats som dras är att samstämmigheten överlag når klart tillfredsställande nivåer, lika höga som vid bedömningen i PISA-undersökningarna. Det kan ses som anmärkningsvärt, eftersom de bedömande lärarna i föreliggande studie inte är särskilt tränade som i PISA-undersökningen, utan slumpvis valda bland lärare som varit villiga att delta.

De statistiska beräkningarna är utförda av Tobias Dalberg som tillsammans med Martina Zachiu har formulerat de delar av texten som handlar om statistisk analys och slutsatser. Negin Shahsavar och Kristina Eriksson initierade forskningsfrågan och har genomfört den praktiska materialinsamlingen samt bidragit med text om provens utformning. Siri Hussenius har granskat rapportens statistiska beräkningar och slutsatser.

Rapporten vänder sig till läsare som är intresserade av bedömning, till exempel forskare, provkonstruktörer och blivande eller verksamma lärare.

Uppsala i augusti 2020

Anne Palmér
universitetslektor och docent,
vetenskaplig ledare för Nationella prov i svenska
och svenska som andraspråk

1 Inledning

På uppdrag av Skolverket utvecklas och konstrueras nationella prov i svenska och svenska som andraspråk vid Institutionen för nordiska språk, Uppsala universitet. Det nationella provet i svenska och svenska som andraspråk i kurs 1 på gymnasiet har funnits sedan läroplanen Lgy11 infördes och består av tre olika delprov som behandlar muntlig framställning (benämnt delprov A), läsning av skönlitteratur och sakprosa (delprov B) respektive skriftlig framställning (delprov C). Proven prövar så långt det är möjligt hur eleven uppnått de mål som respektive ämnesplan tar upp under rubriken ”Ämnets syfte”. Provets innehåll följer ämnesplanernas centrala innehåll medan provets betygsnivåer utgår från kunskapskraven, så som de är formulerade i ämnesplanerna. Provet täcker stora, men inte alla, delar i svenskämnenas ämnesplaner.

De nationella provens huvudsakliga syfte är att fungera som stöd för en likvärdig och rättvis betygssättning. Resultat på prov fungerar som underlag för slutsatser med konsekvenser för individen. Därför är bedömningens kvalitet av stor vikt. Ofta diskuteras samstämmighet och likvärdig bedömning i relation till mer komplexa elevprestationer såsom skrivna texter och muntliga anföranden. Förutsättningarna för en likvärdig och samstämmig bedömning är emellertid en fråga som är central att undersöka även för andra provtyper där en professionell bedömning av elevprestationer sker. Det gäller till exempel även bedömning av elevers läsförståelse, något som uppmärksammats av bland andra Tengberg och Skar (2016) och Tengberg, Roe och Skar (2018).

Läsförståelseprov ingår i samtliga nationella prov i skolämnena svenska och svenska som andraspråk i grundskolan, samt i kurs 1 i gymnasieskolan. Läsförståelseproven är i de båda svenskämnena desamma, med undantag för någon enstaka mer ämnesspecifik uppgift i provet i svenska 1. Resultatet på ett läsförståelseprov är ett av flera underlag från det nationella provet som läraren använder som stöd i sin betygssättning. Det nationella provet i svenska och svenska som andraspråk ger därutöver återkoppling om elevens resultat inom muntliga samt skriftspråkliga färdigheter samt ett sammanfattande så kallat provbetyg som väger ihop den samlade prestationen för de delar som provet prövar inom de båda svenskämnena. Resultatet på ett läsförståelseprov ska alltså tolkas i relation till de övriga resultaten inom det nationella provet när det används som stöd vid betygssättning.

För att slutsatser om en elevs förmåga, i det här fallet läsförståelse, ska kunna anses vara valida behöver olika aspekter i ett så kallat validitetsargument undersökas (jfr Kane et al. 1999). I *Skolverkets systemramverk för nationella prov* (2017:8–9) lyfts det fram ett antal länkar i denna validitetskedja som är centrala i detta avseende. Eftersom läsförståelseprov ofta består av en blandning av flervalsuppgifter och uppgifter där eleven själv formulerar sitt svar utgör bedömning en viktig länk i ett validitetsargument för läsförståelseprov. Skolverket (2017:9) identifierar generellt för samtliga prov med bedömningsinslag ett antal validitetshot, till exempel att bedömningsanvisningen exkluderar aspekter som är relevanta för det som prövas i uppgiften eller att bedömare

fäster oproportionerlig vikt vid vissa sätt att svara. Ett centralt validitetshot för uppgifter med elevproducerade svar är att bedömningen brister i samstämmighet mellan olika bedömare.

Det är, precis som vid bedömning av skrivprov och muntliga prov, inte rimligt att förvänta sig en total samstämmighet i bedömningen av ett läsförståelseprov, eftersom provet innehåller uppgifter som är komplexa, exempelvis skönlitterära tolkningsuppgifter. Sådana uppgifter är också centrala för vad provet prövar.

För att kunna dra valida slutsatser när ett prov används som stöd för betygssättning är det emellertid viktigt att känna till den förväntade tillförlitligheten för den aktuella provtypen, till exempel avseende grad av samstämmighet i bedömningen. Den här rapporten bidrar med kunskap om förutsättningarna för en likvärdig bedömning i ett läsförståelseprov med de förutsättningar som i övrigt gäller för bedömning inom det nationella provsystemet, där bedömningen utförs av verksamma lärare. I denna rapport, som skrivits som en del av pågående valideringsarbete för läsförståelseproven, analyseras samstämmigheten i bedömningar av elevers prestationer på uppgifter med öppna svarsformat producerade inom ramen för ett nationellt läsförståelseprov i kursen svenska 1 på gymnasiet.

En sammanfattande slutsats är att samstämmigheten överlag når klart tillfredsställande nivåer. Det gäller dels sådana nivåer av samstämmighet som finns uttryckta i *Skolverkets systemramverk för nationella prov* (2017), dels nivåer som anges i storskaliga internationella mätningar som PISA-undersökningarna.

1.1 Syfte

Syftet med rapporten är att beskriva och diskutera den nivå av samstämmighet som är förväntad vid lärares bedömning av läsförståelseprov inom det svenska nationella provsystemet. Resonemanget utgår från följande frågor:

- Vilken nivå är förväntad avseende olika vanliga mått på samstämmighet?
- Hur påverkas elevens resultat på provet av om en elevlösning har bedömts av en strängare respektive en mildare bedömare?
- I vilken mån påverkar inslag av flervalsuppgifter, antal uppgifter samt antal bedömare tillförlitligheten i resultatet?

Rapportens resultat beskriver, mer precist, nivåer av samstämmighet som är generaliserbara till situationer där lärare bedömer för dem okända elevers prestationer.

Utöver att skattningen av nivåer för samstämmighet har en direkt praktisk nytta vid tolkning av provresultat för den bedömande läraren, tillhandahåller denna studie viktiga bidrag också på andra sätt. Genom att många olika vanligt förekommande statistiska mått för beskrivning av bedömaröverensstämmelse har undersökts är det möjligt att sätta läsförståelseprovets förutsättningar för bedömning i relation till andra typer av läsförståelseprov och andra provtyper. Dessutom bidrar studien till litteraturen om samstämmighet i bedömning av elevprestationer genom att tillämpa sannolikhetsbaserade modeller för skattningar av måtten.

2 Bakgrund

I detta kapitel beskrivs vad delprov B: läsförståelse i det nationella provet för kursen svenska och svenska som andraspråk 1 prövar samt dess uppgiftstyper och principer för bedömning.

2.1 Delprov B: läsförståelse

Studien som avhandlas i denna rapport undersöker enbart bedömaröverensstämmelse inom det nationella provet, delprov B: läsförståelse, i kursen svenska 1. Men då detta delprov är snarlikt det läsförståelseprov som genomförs inom kursen svenska som andraspråk 1, är beskrivningen i detta bakgrundskapitel överlag giltig för båda dessa läsprov.

I delprov B prövas elevens förmåga att läsa sakprosa och skönlitteratur i löpande och verbal form och att i de aktuella texterna hitta klart uttryckt information, tolka underliggande budskap samt reflektera över innehåll och form. Denna formulering av vad delprovet avser att pröva kallas med ett annat ord för delprovets konstrukt.

Delprovet innebär att eleven löser cirka 27 uppgifter som alla bygger på läsning av texter i ett texthäfte. Texthäftet består av texter av varierande svårighetsgrad, både sakprosa och skönlitteratur, samt alltid någon form av lyrik, vanligen en dikt, som kan visa spännvidden i läsförmåga på olika nivåer. Det sammanlagda ordantalet ligger på omkring 4 500 ord vilket motsvarar ungefär 10 sidor text. Två olika svarsformat förekommer i delprovet: uppgifter med elevkonstruerade svar samt uppgifter av flervalstyp. Omkring hälften av uppgifterna brukar utgöras av flervalsuppgifter, men ett utvecklingsarbete pågår, där provgruppen undersöker möjligheten att utöka andelen flervalsuppgifter i läsförståelseproven.

Utifrån den internationella kunskapsmätningen PISA:s definitioner av tre olika aspekter av läsförståelse (OECD 2009:34–35) har provgruppen formulerat definitioner för tre så kallade läsförståelseprocesser. Viktigt att notera är att delprov B på flera olika sätt skiljer sig i utformning från PISA och att definitionen av de olika processerna inte är direkt överförbar. Inom delprov B benämns processerna:

1. hitta information och dra enkla slutsatser
2. tolka och sammanföra
3. reflektera över och utvärdera innehåll, form och språk.

Vid läsning är dessa processer inte tydligt åtskilda utan samverkar snarare på olika sätt, men en viss process kan antas vara mer framträdande vid lösningen av en specifik uppgift. Varje uppgift i delprov B konstrueras för att motsvara någon av de tre läsförståelseprocesserna. Syftet med kategoriseringen är dels att säkerställa en så hög grad av domäntäckning som möjligt, det vill säga undvika att viss typ

av läsning blir över- eller underrepresenterad i ett visst prov, dels att sammansättningen av uppgiftstyper ska bli så jämförbar som möjligt mellan olika färdiga versioner av delprovet. Läsförståelseprocesserna ger sammantaget information om elevens läsförmåga men har var för sig inte någon direkt koppling till bedömningen av delprovet eller elevernas provbetyg.

Det finns ingen given koppling mellan vilken läsförståelseprocess som dominerar i uppgiften och uppgiftens svarsformat. Inom varje läsförståelseprocess förekommer både uppgifter med elevkonstruerade svar och flervalsuppgifter.

För elever som följer ämnesplanen i svenska omfattar delprovet 120 minuter och för elever som följer ämnesplanen i svenska som andraspråk 180 minuter. Detta är enligt en övervägande andel av de lärare som årligen besvarar provets enkät en rimlig tid för att hinna bearbeta delprovets texter och uppgifter. I enkäten som användes för att följa upp provet vårterminen 2017 angav 89 procent av de svarande lärarna att textmängden var rimlig i förhållande till provtiden.

Eftersom delprov B prövar individuell läsförståelse ska provets texthäfte inte läsas och diskuteras med andra inför genomförandet. Under en förberedelselektion ägnas cirka 20 minuters tid åt att läraren, utifrån ett kopieringsunderlag, informerar eleverna om delprovets upplägg och svarstyper samt om vilka typer av texter som ingår i delprovet. Eleverna får också råd om lämpliga strategier vid genomförandet av delprovet, exempelvis att läsa varje uppgifts instruktion noggrant. För uppgifter som kräver elevernas egna formuleringar indikerar antalet tomma rader vid uppgiften i elevhäftet hur omfattande formuleringen förväntas vara.

2.2 Provformat och bedömning

Ofta görs en skillnad mellan provformat som leder till en direkt bedömning, respektive sådana som prövar en förmåga på ett mer indirekt sätt (jfr Kane et al. 1999). I det nationella provet i svenska 1 och svenska som andraspråk 1 är delprov A och C så kallade performansprov som ger upphov till en mer direkt bedömning. I dessa delprov bedöms elevens förmåga att tala eller skriva genom att eleverna löser uppgifter som leder till just den typ av muntlig eller skriftlig framställning som provet syftar till att uttala sig om. Kopplingen mellan bedömningssituationen och delprovets konstrukt – vad delprovet avser att mäta – blir därmed genomskinlig, vilket kan anses öka provets autenticitet. Men samtidigt som bedömningen blir mer autentisk blir den också mer komplex och provet måste hantera de problem bedömningens komplexitet kan innebära för provets reliabilitet.

Delprov B: läsförståelse ger i stället upphov till en mer indirekt bedömning. En analys av ämnesplanen i svenska visar att flera av de utvecklingsmål och kunskapskrav som berör läsning kopplar samman läsande och skrivande genom att de beskriver läsaktiviteter som mynnar ut i skriftlig produktion. Delprovet i läsförståelse är i stället inriktat mot att pröva elevens förutsättningar att utföra de

beskrivna aktiviteterna snarare än aktiviteterna i sig. Exempelvis mäter inte delprovet huruvida eleverna kan skriva ”egna texter som **lyfter fram huvudtanken** i det lästa” (ur kunskapskraven för betyg C i kursen svenska 1), utan ger i stället indikationer om elevens förmåga att urskilja huvudtanken i en läst text.

I ett prov med mer av indirekt bedömning blir kopplingen till den förmåga som prövas alltså inte lika omedelbart synlig, men uppgifterna kan i stället tillåtas att bli mer slutna och därmed mer lättbedömda. En indirekt bedömning förutsätter ofta ett stort antal uppgifter (på engelska *items*) som kan bedömas på ett mer objektivt sätt än vad som är fallet för prov med direkt bedömning. Ett prov som liksom delprov B består av många men mindre omfattande uppgifter som sammantaget ger ett resultat på provet kan benämnas som ett itemprov.

För att kunna pröva läsförmåga är det avgörande att det provmaterial som skolorna erbjuds är så genomarbetat att eleverna förstår vad uppgiften efterfrågar och att lärarna förstår bedömningsanvisningarna och kan bedöma i enlighet med dem. Dessutom bör provet avspegla vad som avses med god läsförmåga i förhållande till provets konstrukt. Det bör till exempel råda någorlunda konsensus kring att poängsättning och svårighetsnivåer på uppgifter är ett utslag av elevens läsförmåga och inte någonting annat, eventuellt godtyckligt. Under konstruktionen av nya provversioner genomförs utprövningar i flera omgångar, materialet granskas av referensgrupper och experter i olika skeden och olika statistiska beräkningar görs för att i så hög grad som det är möjligt säkerställa att allt provmaterial håller tillräckligt hög kvalitet. Med kvalitet avses här huruvida en uppgift mäter det den är avsedd att mäta, har ett förväntat utfall som går att ringa in i ett bedömningsunderlag, håller en rimlig svårighetsgrad med tanke på provets mottagare etc.

Bedömningsanvisningar konstrueras parallellt med att nya uppgifter tas fram. I samband med utprövningar dokumenterar provgruppen vilka olika typer av elevsvar som uppgifter med elevkonstruerade svar ger upphov till samt hur väl de olika typsvaren går att bedöma utifrån den tänkta bedömningsanvisningen. Särskild vikt läggs vid att undersöka sådana svar som initialt inte bedöms som fullgoda lösningar. Här är det centralt att ta reda på om svaren kan antas vara fullt rimliga tolkningar av frågeformuleringen i uppgiften eller om de faktiskt ger uttryck för en sämre läsförståelse. Sådana analyser ger sammantaget indikationer om och hur uppgiftsformuleringar och bedömningsunderlag bör omformuleras.

Vid utprövning får även deltagande lärare ta del av det preliminära bedömningsunderlaget och ombeds kommentera eventuella brister eller otydligheter i både uppgifter och bedömningsunderlag. Detta är ett steg i att stärka konsistent bedömning då uppgifter som inte fungerar tillfredsställande lättare kan upptäckas och sorteras bort under konstruktionsprocessen. Av samma anledning granskas materialet kontinuerligt av flera olika medarbetare inom provgruppen samt av speciellt inbjudna forskare och erfarna provkonstruktörer från exempelvis högskoleprovet och den norska motsvarigheten till de nationella svenska läsförståelseproven.

Kvalitativa skattningar av vilken interbedömarreliabilitet en uppgifts bedömningsanvisning ger upphov till vägs in i det slutgiltiga urvalet av uppgifter till ett färdigt delprov.

Något som är speciellt med de svenska nationella proven jämfört med exempelvis PISA- och PIRLS-proven är att det inte finns någon centralt organiserad eller extern bedömning, utan att elevernas lösningar i normalfallet bedöms av den egna läraren. För bedömning av delprov B i svenska 1 och svenska som andraspråk 1 finns ett häfte med anvisningar. Där listas inledningsvis följande övergripande bedömningsprinciper som gäller för delprov B.

Instruktioner för bedömning av delprov B: läsförståelse:

- I uppgifter där eleven själv formulerar sitt svar ska eleven visa att hen förstått syftet med uppgiften och kan lösa den på ett rimligt sätt i förhållande till texten. Lösningar som visar att eleven missförstått uppgiften ska inte godkännas. Detta gäller också lösningar som är alltför vaga eller oklara.
- Eleven behöver inte ordagrant formulera sig som i bedömningsmallen. Det viktiga är att lösningen innehåller information motsvarande den aktuella bedömningsanvisningen.
- Elevlösningar som innehåller ytterligare information än det som efterfrågas i uppgiften får poäng så länge eleven tydligt uttrycker det som ska framgå enligt bedömningsmallen. Vid motsägande information eller uppenbara garderingar ges däremot inga poäng.
- I vissa uppgifter är det möjligt att citera ett textavsnitt i stället för att formulera sig med egna ord. För att nå poäng ska citatet vara väl avgränsat och tydligt fungera som svar på uppgiften.
- I enstaka fall förekommer det att goda elevlösningar inte stämmer överens med bedömningsanvisningarna. Om ett sådant avvikande svar ger uttryck för att eleven har förstått texten på ett riktigt sätt i förhållande till uppgiften kan läraren välja att godkänna det.

Därefter följer en bedömningsmall med uppgiftsspecifika bedömningsanvisningar. För flervalsuppgifter anges vad som är det korrekta alternativet. Uppgifter med elevkonstruerade svar har alltid en generell bedömningsanvisning som konkretiseras genom en uppställning av exempellösningar och hur dessa har bedömts.

Vid användning av bedömningsanvisningarna ligger fokus på att identifiera tecken på visad kunskap, så kallad positiv bedömning, snarare än att utgå från en tänkt fullständig och korrekt lösning och identifiera brister. Bedömningen ska vidare utgå från den lösning en elev faktiskt har producerat i en viss uppgift, inte från tolkningar av vad eleven kan ha försökt uttrycka eller hur eleven brukar prestera i andra sammanhang.

Ingen specificering görs av vilka uppgifter som ska ha lösts för en viss betygsnivå. Delprovet är alltså inbördes helt och hållet kompensatoriskt. Alla godtagbart lösta uppgifter ger 2 poäng i det aktuella provet, och om lösningen inte är godtagbar får den 0 poäng. Det är inte möjligt att ge elevlösningen 1 poäng. Uppgifterna är därmed dikotoma, lösningarna är antingen godtagbara eller ej godtagbara. Ingen viktning av uppgifter eller poäng görs. En uppgift är endast värd fler poäng om den innebär att eleven löser fler uppgiftsled, som exempelvis i uppgifter som av

pedagogiska skäl delas upp i ett a- respektive b-led. När alla uppgifter är bedömda summeras poängen till en totalpoängsumma. Utifrån en tabell där de fastställda poänggränserna redovisas översätts totalpoängen sedan till ett delprovsbetyg. Dessa gränser tas fram med hjälp av så kallad kravgränssättning inför publiceringen av varje enskild provversion. De tre olika delprovsresultaten sammanställs slutligen till ett sammanvägt provbetyg som även det tas fram med hjälp av en tabell i bedömningsanvisningarna.

2.3 Provets syfte

Det nationella provet i svenska 1 har som huvudsakligt syfte att stödja en likvärdig och rättvis betygssättning. Provet kan också bidra till att stärka skolornas kvalitetsarbete genom analys av provresultaten i relation till uppnådda kunskapskrav på skolnivå, på huvudmannanivå och på nationell nivå. Det nationella provet ska användas för att bedöma elevernas kunskaper i förhållande till kunskapskraven.

Resultatet på det nationella provet ska särskilt beaktas vid betygssättningen. Att resultatet särskilt ska beaktas innebär att resultatet har en större betydelse i samband med lärarens allsidiga utvärdering av elevernas kunskaper vid betygssättningen än resultatet på andra enskilda underlag, och att lärare inte kan bortse från resultatet om det inte finns särskilda skäl för det. Formuleringen ”särskilt beakta” betyder däremot inte att provresultatet helt ska styra betyget, utan att provet ska utgöra ett stöd vid betygssättningen. Resultatet på ett nationellt prov kan alltså inte ensamt utgöra underlag för betygssättningen och läraren kan aldrig enbart motivera ett betyg utifrån ett nationellt provbetyg. Viktigt att notera i sammanhanget är att det är elevens sammanvägda provbetyg som ska beaktas, inte resultatet på de enskilda delproven. Utifrån syftet att stödja en likvärdig och rättvis bedömning är det uppenbart att alla delar av provet måste uppnå god reliabilitet, till exempel avseende samstämmighet i bedömningen.

2.4 Provets målgrupp

Sedan den 1 januari 2018 är provet i svenska 1 och svenska som andraspråk 1 obligatoriskt på de program där kursen är den högsta avslutande kursen i ämnet men valfritt på de program där eleverna även läser svenska 2 och 3. Det är rektor som avgör om provet ska användas i de fall där det är frivilligt. Om rektorn beslutar att provet ska användas innebär det att alla elever som läser kursen ska genomföra provet och att provet ska hanteras på samma sätt som ett obligatoriskt prov. Inom vuxenutbildningen är provet fortsatt obligatoriskt.

3 Teori och metod: Ansatser för att analysera samstämmighet

I rapporten används olika metoder för att mäta samstämmigheten. Dessa olika metoder är i sin tur förknippade med olika antaganden om vilken typ av samstämmighet man kan förvänta sig under optimala förutsättningar. Det svenska nationella provsystemet förutsätter i hög grad den så kallade standardansatsen, benämnd så av Thomas Eckes (2015:42–43). Standardansatsen utgår från antagandet att testtagaren har en sann prestationsnivå. Under optimala förutsättningar antar man alltså att det finns en ”sann” prestation som är möjlig för bedömaren att nå fram till i sin bedömning. Ett alternativ till standardansatsen är den begreppspsykometriska ansatsen – mätansatsen – som till exempel genom statistisk modellering kan skatta elevens prestation med hänsyn tagen till olika faktorer, till exempel den aktuella bedömarens stränghet. En sådan typ av modellering är inte en integrerad del av provsystemet, men kan användas i utvärderande syfte för att ta reda på vilken roll skillnader i bedömarens stränghet spelar för skattningen av elevens resultat. I den här rapporten utvärderas samstämmigheten därför utifrån båda dessa ansatser. Inom facklitteraturen skiljer man ofta mellan *interbedömarreliabilitet* som står för samstämmighet mellan bedömare och *intrabedömarreliabilitet* som avser varje enskild bedömares interna konsistens, det vill säga överensstämmelsen mellan en och samma persons bedömningar. I denna rapport syftar samstämmighet på *inter*bedömarreliabilitet.

För att mäta samstämmigheten mellan olika bedömare använder man sig inom standardansatsen av två typer av skattningar: konsensus- respektive konsistensskattningar, vilka också beräknats i föreliggande rapport. Inom mätansatsen har för denna undersökning så kallad mångfasetterad Rasch-modellering samt så kallad generaliserbarhetsteori använts för att besvara frågan om skillnader i resultat mellan stränga och milda bedömare, samt hur denna bedömarvariation i stränghet inverkar på elevens slutliga resultat. I detta avsnitt ges en översikt över dessa olika statistiska mått och de antaganden dessa bygger på. Allra sist i avsnittet presenteras den metod för statistisk inferens som används för att generalisera slutsatser utifrån studien.

3.1 Standardansatsen

3.1.1 Konsensusskattningar: procentuell överensstämmelse och Cohens kappa

Konsensusskattningar mäter i vilken utsträckning bedömarna är helt överens i sina bedömningar, och därför talar man ibland om absolut eller exakt överensstämmelse när man använder sig av konsensusmått. Om bedömningen handlar om

att poängsätta uppgifter bör bedömarna med andra ord nå fram till samma poäng. Vanliga tekniker för konsensusskattningar är exakt överensstämmelse respektive Cohens κ (utläses Cohens kappa), uppkallat efter Jacob Cohen (1960;1968). Den exakta överensstämmelsen kallas ibland procentuell överensstämmelse. För att beskriva hur beräkningen går till kan vi föreställa oss två bedömare som var för sig har bedömt samma 10 uppgifter. Om 3 av 10 gemensamma uppgifter har tilldelats exakt samma poäng av två bedömare blir den exakta överensstämmelsen mellan dessa två bedömare 30 procent. I föreliggande rapport, där sammanfattande mått är av större intresse än överensstämmelse mellan specifika par av bedömare, redovisas medelvärdet och medianen av alla parvisa beräkningar av exakt överensstämmelse tillsammans med spannet från det lägsta till det högsta värdet.

Cohens κ-koefficient mäter överensstämmelsen mellan två bedömare med hänsyn tagen till att de då och då – beroende på antalet poängnivåer eller skalsteg – förväntas sätta samma poäng även om de inte är särskilt överens. Den ovannämnda exakta överensstämmelsen tar ingen hänsyn till sådana skillnader i antal poängnivåer, och blir därför svår att jämföra från ett sammanhang till ett annat. Cohens κ-koefficient underlättar dock jämförelser av skattningar genomförda för bedömningar utförda med olika bedömningsskalor. Antagandet om att bedömningen skulle kunna ha utförts helt slumpmässigt vid bedömningen av uppgifter med mycket tydligt formulerade bedömningsgrunder, kan emellertid i vissa fall vara orealistiskt, vilket medför att måttet kan tendera att bli konservativt (jfr Uebersax 1987).

3.1.2 Konsistensskattningar: korrelationsmått

Konsistensmåtten mäter i vilken utsträckning bedömarna är överens om rangordningen av prestationer. Det innebär att bedömarna skulle kunna skilja sig åt vad gäller stränghet men ändå vara överens om själva rangordningen av prestationerna. Därför talar man ibland om relativ överensstämmelse när man använder sig av konsistensmått såsom korrelationskoefficienter av olika slag. Konsistensmått kan fungera som ett bra komplement till konsensusmått för att tolka på vilket sätt bedömningar avviker ifrån varandra, när konsensus kring den exakta bedömningen är låg.

I tablå 1a nedan återges ett exempel där två lärares totala bedömning av samtliga uppgifter på ett läsprov avviker från varandra. Det är emellertid inte rimligt att dra slutsatsen att bedömningen i det här fallet skulle vara problematisk, vilket en konsensusskattning skulle indikera. Avvikelsen är nämligen inte alls osystematisk. Båda lärarna har rangordnat de båda elevlösningarna på samma sätt, och det är inte sannolikt att en bedömare hamnar på exakt samma poängsumma som en annan, när skalan består av många skalsteg. Om konsistensen, det vill säga överensstämmelsen i rangordning, är god så är det därför ändå rimligt att sluta sig till att bedömningen har skett systematiskt och är av god kvalitet.

Tablå 1a. Totalpoäng på läsförståelseprovet i dess helhet.

	Lärare X	Lärare Y
Totalpoäng elev 1	24 poäng	28 poäng
Totalpoäng elev 2	20 poäng	26 poäng

Tablå 1b. Poängsumma för enskild uppgift.

	Lärare X	Lärare Y
Elevlösning 1	2 poäng	0 poäng
Elevlösning 2	0 poäng	2 poäng

För läsförståelseprov där varje uppgift vanligen bedöms som antingen poänggivande eller icke-poänggivande, det vill säga är dikotom, är det däremot inte relevant att undersöka rangordningen av prestationer på uppgiftsnivå. Den enda möjliga skillnaden i rangordning för en enskild sådan dikotom uppgift illustreras av tablå 1b ovan. I praktiken säger ett konsistensmått exakt samma sak som konsensusmåtten för dessa typer av uppgifter, eftersom bedömningarna när de skiljer sig åt också alltid per definition skiljer sig i rangordning.

För att skatta konsistens i bedömning kan flera olika mått användas. Ofta används olika typer av enkla korrelationsmått såsom Kendalls rangkorrelation och Pearsons korrelation. När det gäller bedömning av enskilda uppgifter med binära utfall – rätt eller fel – ger konsensusmått som tidigare nämnda Cohens κ och konsistensmått som Pearsons korrelation näst intill identiska resultat, medan Pearsons korrelation och Kendalls rangkorrelation ger helt identiska resultat. Det är först när de enskilda uppgifternas binära utfall sammanförs till en totalpoängsumma som konsensus- och konsistensskattningar kan leda till motstridiga slutsatser (Eckes 2015:45). En konsekvent mildare bedömares och en konsekvent strängare bedömares totalpoängsummor kan till exempel generera identiska rangordningar av prestationer utan att någonsin generera exakt samma poängsumma till samma prestation. Med konsistensmått skulle vi utifrån detta scenario dra slutsatsen om att det råder god samstämmighet, medan konsensusmåtten skulle föranleda omedelbara åtgärder för att förbättra samstämmigheten.

Utifrån konsistensmått kan alltså samvariation mellan två bedömare skattas. Men måtten fungerar inte lika bra som beskrivningar när många bedömare har bedömt samma uppgifter. Ett mått som då är möjligt att använda är inomklasskorrelation som presenteras i nästa avsnitt.

3.1.3 Inomklasskorrelation

En vanlig uppsättning modeller för att skatta bedömaröverensstämmelse när det finns ett antal bedömare som alla bedömt samtliga uppgifter och man vill genomföra en skattning på ett mer sammanfattande sätt än de tidigare nämnda måtten medger, är inomklasskorrelationen (ofta förkortad ICC). Den utgår från de två aspekter som varierar i en bedömningssituation av det här slaget:

(1) Eleverna varierar i läsförmåga, det vill säga det provet syftar till att testa.

(2) Bedömarna varierar i sin bedömning av samma elevprestationer, den variation som kan ses som felkällan i bedömningen.

Vid bedömning av elevers lösningar av uppgifter är det till exempel önskvärt att variation i bedömningen beror på att lösningarna i sig varierar i kvalitet och inte på att de som bedömer texterna varierar i stränghet eller pålitlighet. Inomklasskorrelationen kan alltså något förenklat sägas beskriva hur stor andel av variationen som består av det vi vill mäta, nämligen elevernas variation i läsförmåga, jämfört med hur stor variation som beror på felkällan, nämligen bedömarnas variation.

Det finns flera olika reliabilitetsmått som kan användas för att skatta inomklasskorrelation. Ett sådant relativt välkänt konsistensmått är Cronbachs α (utläses Cronbachs alfa) (jfr McGraw & Wong 1996). Ofta används just Cronbachs α för att skatta hur väl uppgifterna mäter samma sak, men det kan användas också för att skatta hur väl bedömningarna hänger samman. Som alltid vid beräkning av statistiska mått är det viktigt att den modell som används stämmer in på undersökningsdesignen och dataurvalet. Det finns en rad olika modeller som kan specificeras vid beräkning av inomklasskorrelation, och exempelvis McGraw & Wong (1996) redovisar olika ICC-modeller. Valet av modell skiljer sig till exempel beroende på om syftet är att utvärdera konsensus eller konsistens enligt resonemanget ovan. I detta fall har den modell som utvärderar konsistens använts.

I vårt fall bör dessutom inomklasskorrelationen beräknas med den variant som kallas *two-way random single measure*. Denna modell används nämligen när varje text bedöms av varje bedömare, och dessa bedömare anses vara representativa för en större population av liknande bedömare, samt när lösningarna kan antas utgöra ett representativt stickprov (Koo & Li 2016). Reliabiliteten bestäms utifrån de enskilda bedömningarna (*single measure*) snarare än medelvärdet av flera bedömningar (*average*).

3.1.4 Möjligheten att generalisera resultatet av studien

Centralt för studier som använder sig av statistisk metod är möjligheten att dra slutsatser som är generellt giltiga. Det är helt enkelt inte särskilt intressant att beskriva hur just dessa nio lärare bedömde dessa enstaka elevlösningar, med hjälp av de ovan nämnda statistiska måtten. Önskvärt är snarare att utifrån en studie av

det här slaget kunna uttala sig om vad som generellt är att vänta sig när lärare i Sverige bedömer elever som gjort det nationella läsförståelseprovet i kurs 1. De lärare och elever som deltagit i studien kan sägas utgöra ett stickprov av populationen lärare och elever i Sverige.

När ett stickprov från en population tas kommer både de bedömare och elevlösningar som utgör urvalet i studien att uppvisa vissa egenheter. Detta gör att det inte är rimligt att vänta sig att värdet i stickprovet skulle vara exakt detsamma som i populationen. Med statistisk metod är det möjligt att skatta den osäkerhet som uppstår på grund av sådana slumpmässiga egenheter i stickprovet. Förutsättningen för att kunna skatta osäkerheten är att det urval som använts är slumpmässigt i relation till den population man vill uttala sig om. I den här studien har både bedömare och elevlösningar behandlats som någorlunda slumpmässigt utvalda, men med vissa begränsningar.

Vanligen används signifikanstestning för att avgöra om det är rimligt att dra slutsatsen att det värde som uppvisas i stickprovet går att generalisera till populationen. En ofta förekommande invändning mot signifikanstestning är att utfallet är beroende av stickprovets storlek. Stora stickprov kan ge ett signifikant resultat trots att bedömaröverensstämmelsen är relativt liten och tvärtom kan små stickprov ge icke-signifikanta resultat trots att bedömaröverensstämmelsen är hög. Det går inte heller utifrån denna metod att säga hur troligt det är att det uppmätta värdet även skulle påträffas om vi fick chansen att fortsätta samla in data och mäta en allt större del av den tänkta målpopulationen. Om man vill kunna uttala sig om sannolikheten för att samstämmigheten befinner sig på en viss nivå, eller inom ett visst spann, får man vända sig till den statistiska grenen sannolikhetsbaserad inferens (vanligen benämnd bayesiansk statistik).

Inom sannolikhetsbaserad inferens skattas sannolikheter för hur ett visst mått uppträder i populationen, till exempel utifrån en tidigare undersökning. Detta görs för att på ett säkrare sätt närma sig sannolikheten för ett visst utfall i den population man vill uttala sig om. Mot bakgrund av att det finns få tidigare undersökningar att rätta sig efter har vi i föreliggande studie emellertid utgått från antagandet att den kunskap vi för närvarande kan ha om sannolikheter för mått på samstämmighet finns i de data vi har samlat in. Det innebär att vi i beräkningarna använder oss av en likformig förhandsfördelning där alla möjliga värden är lika sannolika tills vi låter modellen konfrontera våra data. Den sannolikhetsbaserade metoden för inferens resulterar i ett intervall för de värden av samstämmighet som är att vänta i populationen.

3.2 Mätansatser

De två varianter av mätskattningar, den mångfasetterade Rasch-analysen respektive generaliserbarhetsteorin, som används i föreliggande rapport bär flera gemensamma drag, varav det viktigaste är att ett resultat på ett prov betraktas som resultatet

av en uppsättning samverkande faktorer eller, som de ofta benämns, fasetter. De båda mätskattningarna är emellertid intressanta var för sig, eftersom de besvarar delvis olika frågor. Den ena varianten av mätskattning, den mångfasetterade Rasch-analysen, hör hemma inom den så kallade item-response-teorin (IRT), som liksom namnet signalerar tar den enskilda uppgiften som utgångspunkt för analysen vid identifikation av bedömares stränghet, uppgifters svårighetsgrad och elevernas läsförmåga. Generaliserbarhetsteorin söker till skillnad ifrån IRT-modellerna att förklara totalpoängsresultatet och vilka beståndsdelar variationen i ett totalpoängsresultat består av. Eftersom rapportens syfte är att undersöka samstämmighet begränsas framställningen till mätskattningar som rör just bedömares samstämmighet. I detta avsnitt presenteras de båda mätansatserna var för sig.

3.2.1 Mångfasetterad Rasch-analys

Den mångfasetterade Rasch-analysen är som ovan nämnts en IRT-modell. Det betyder att den utgår från den enskilda uppgiften och mönster i datasetet. Varje uppgift ses som en indikator på den förståelse som testas. Utifrån svarsmönster och mönster för bedömning kan sedan sådant som uppgifters svårighetsgrad, testtagarens förmåga och bedömarens stränghet skattas. Kännetecknande för item-response-teorin är att den modell som används för skattning av de värden som man är intresserad av, till exempel bedömarens stränghet, specificeras på förhand rent matematiskt. Den matematiska definitionen, formeln om man så vill, uttrycker sambandet mellan sannolikheten för eleven att svara rätt på en uppgift och ett antal okända parametrar. De okända så kallade parametrarna i formeln är i denna undersökning elevens läsförmåga, uppgiftens svårighetsgrad samt just bedömarens stränghet.[1] Eftersom det är flera obekanta parametrar som ska skattas, går det inte att på ett enkelt sätt matematiskt räkna ut ett värde. I stället skattas eller estimeras värdet på de obekanta parametrarna, genom ett datorprogram. Här har *Facets* använts. Programmet genomför ett mycket stort antal sökningar som testar olika värden på parametrarna och försöker hitta de värden som bäst stämmer in på datasetet. Efter ett tag konvergerar skattningen och datorn hittar de värden som bäst motsvarar den givna formeln och det dataset som analyseras. Det finns flera olika sådana estimeringsmetoder, det vill säga sätt för datorn att söka efter de bästa värdena, och den som använts här är *Joint Maximum Likelihood Estimation*.

Till skillnad från andra statistiska modeller, där man tänker sig att det är de empiriskt observerade datapunkterna som representerar de sanna värdena, vänder item-response-teorin på resonemanget. Utgångspunkten är helt enkelt att det är den på förhand definierade modellen som på ett sant sätt representerar hur den

[1] Den modell som här använts formuleras matematiskt som följer: $P = \frac{e^{T-S-B}}{1+e^{T-S-B}}$, där T står för testtagarens förmåga, S för uppgiftens svårighetsgrad och B för bedömarens stränghet. P står för sannolikheten att klara uppgiften. Av formeln följer att när exponenten, dvs. T – S – B är noll, är sannolikheten att klara en uppgift 0,5 (eftersom $e^0 = 1$). (jfr Linacre 1989:1)

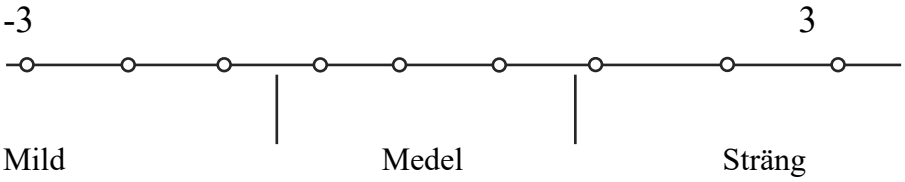
förmåga som mäts beter sig, vilket i det här fallet är läsförståelse. Om enskilda observationer inte stämmer med modellen, så bör man vidta åtgärder för att komma tillrätta med dessa avvikelser (Linacre 1989:43). Även om modellens syfte i första hand är preskriptivt, det vill säga att identifiera vilka datapunkter som inte stämmer med modellen i syfte att åtgärda dessa, till exempel genom att sortera bort uppgifter och träna bedömare, kan den också ses som deskriptiv vid utvärdering av det slag som presenteras här.

Fokus i föreliggande undersökning är just skattningen av bedömarparametern i modellen, det vill säga hur sträng eller mild bedömaren tycks vara. Att en bedömare är sträng eller mild är enligt modellen förväntat, när en persons förmåga ska skattas. Estimeringsprocessen hittar då ett värde för personens förmåga som korrigerar för att bedömaren kan ha varit alltför sträng eller alltför mild. I det här fallet, när syftet snarare är att deskriptivt belysa skillnader i bedömning än att skatta testtagarens förmåga, är skillnader i bedömares stränghet av vikt för att få syn på hur bedömningen varierar på poängskalan.

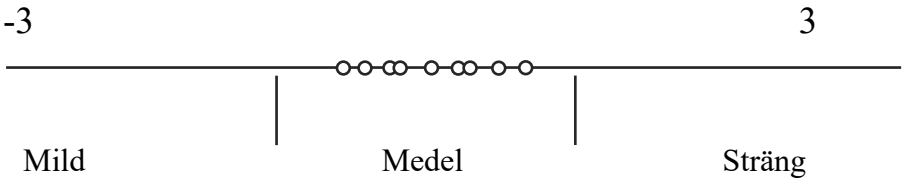
En fördel med att anpassa en mångfasetterad Rasch-modell till data över bedömningar av provuppgifter är att det går att identifiera särskilt stora avvikelser mellan modellens förutsägelser och observerade data, så kallad modellanpassning. Vissa elevsvar på specifika uppgifter kan dessutom visa sig vara särskilt svåra att bedöma samstämmigt och uppvisar då en stor avvikelse från modellens förutsägelser. De bedömare som uppvisar potentiellt säregna bedömningar i relation till modellen är särskilt viktiga att få syn på.

För Rasch-modeller utvärderas vanligen överensstämmelse med modellen genom så kallade inlier-sensitive fit-mått. Måttet visar hur stor avvikelsen är mellan den enskilda observationen och själva modellens förutsägelser. Vid hög överensstämmelse – att en bedömares rangordning av uppgifter och prestationer motsvarar den som beräknats av modellen – hamnar inlier-sensitive fit-måttet nära 1. Om en bedömare avviker uppåt (>1,3) finns det en risk att den har ett säreget sätt att bedöma uppgifterna, i förhållande till övriga bedömare som den jämförs med. Om bedömaren i stället avviker nedåt (<0,75) finns en risk att den är försiktig i sina bedömningar och mestadels håller sig till mitten av skalan, eller att man för respektive elev sätter samma poäng på alla uppgifter trots att andra bedömare betraktar prestationerna på dessa uppgifter som varierande. Värden över 1,3 och under 0,7 brukar bedömas som problematiska. Samma gränsvärden kan också tillämpas på modellanpassningen för uppgiften.

Som sammanfattande mått för hur bedömarna varierar i stränghet används ofta det så kallade separationsindexet (G) alternativt separationskvoten (H) (Wright & Masters 1982:105–106). Separationskvoten visar hur många kategorier som bedömarna kan delas in i (Eckes 2015:62–63; Wright & Masters 1982:92, 105–106). Ju mer bedömningar varierar, desto fler kategorier är det möjligt att urskilja. Detta illustreras i de fiktiva figurerna 1a och 1b nedan, där varje bedömares ”strängheten” representeras av en punkt på logitskalan.



Figur 1a. Variation i stränghet – högt värde på separationsindex.



Figur 1b. Variation i stränghet – lågt värde på separationsindex.

Ett högt separationsindex visar hur många kategorier det är möjligt att dela in bedömarna i, med hänsyn tagen till eventuella mätfel. Ett högt värde för separationsindexet är problematiskt, eftersom bedömarvariationen helst ska vara så liten som möjligt. Vid ett slumpmässigt urval av bedömare används separationsindexet om man inte har skäl att anta att extrema bedömningar i sig skulle kunna vara sanna. Om man vill ta större hänsyn till mer extrema bedömningar, eftersom man har skäl att anta att de kan vara lika sanna, används separationskvoten. Ju lägre värde dessa mått uppvisar, desto mer likartat har bedömningen genomförts. Helst ska det inte vara möjligt att dela in bedömarna i olika kategorier, som i figur 1a, utan bedömarna bör göra en ungefärlig likadan bedömning, i enlighet med figur 1b.

Separationsreliabiliteten uttrycker med vilken säkerhet bedömarna skiljer sig från varandra. Den talar alltså om hur tillförlitligt separationsindexet och separationskvoten delar in gruppen bedömare i olika kategorier, vilket är anledningen till att man strävar efter en låg separationsreliabilitet i detta sammanhang. Om bedömarna på ett tillförlitligt sätt är nära nog utbytbara i termer av stränghet närmar sig separationsreliabiliteten 0,0 och separationsindex 1,0.

3.2.2 Generaliserbarhetsteori

En teori som brukar föras till den så kallade mätansatsen är den så kallade generaliserbarhetsteorin. Generaliserbarhetsteoretiska statistiska modeller utgör i någon mening den klassiska testteorins motsvarighet till den moderna testteorins mångfasetterade

Rasch-analys (Brennan 2001:2; Eckes 2015:169). Generaliserbarhetsteorin och item-response-teorin skiljer sig emellertid på flera sätt från varandra vilket gör det motiverat att undersöka samstämmigheten i bedömning utifrån båda perspektiven. Till skillnad från item-response-teorin, som utgår från egenskaper på uppgiftsnivå vid skattning av de olika fasetterna, står totalpoängsummor på provlösningen som helhet i fokus för generaliserbarhetsteorin. Utgångspunkten är att det är totalsumman i ett prov som någorlunda speglar elevens kompetens, men att det kan finnas felkällor, till exempel variation i bedömningen.

Syftet med en generaliserbarhetsstudie är ofta utforskande. Det är helt enkelt viktigt att förstå hur olika kända fasetter påverkar den totalpoängsumma eleven får. En sådan studie benämns ofta G-study (generalizability-study) (Brennan 2001). När elever som genomför ett läsförståelseprov får olika slutresultat i form av en poängsumma är det till exempel eftersträvansvärt att så stor portion som möjligt av deras olika slutresultat förklaras av att de varierar i läsförståelse snarare än att de reagerar egenartat på specifika uppgifter eller bedöms olika beroende på vem det är som bedömer. Den typ av analyser som är vanligt förekommande i generaliserbarhetsteoretiska studier fungerar väl för att analysera på förhand kända källor till variationer i prestationernas resultat. En generaliserbarhetsstudie kan också fungera som underlag för beslut, till exempel huruvida ett test kan behöva designas om. En sådan studie kallas D-study (decision-study), här beslutsstudie (Brennan 2001:8–9).

Centralt inom generaliserbarhetsteorin är att värdera hur ett provresultat kan generaliseras. Den lärare som använder ett specifikt läsförståelseprov som stöd i sin betygssättning på kursen är egentligen inte särskilt intresserad av vad eleven exakt svarade på de uppgifter som ingick i just det prov som eleven genomfört. Läraren vill snarare kunna dra en slutsats om nivån på elevens läsförmåga generellt, oavsett vilket prov som genomförts och vem som har bedömt själva provlösningen. Inom generaliserbarhetsteorin söker man uttrycka med vilken reliabilitet det är rimligt att generalisera ett resultat på ett prov, givet kunskapen om vissa felkällor.

Utifrån generaliserbarhetsteorin betraktar vi alltså ett observerat resultat från en elevprestation som ett stickprov från ett universum av möjliga elevprestationer och möjliga mått på dessa prestationer. Vi kan föreställa oss att eleverna på ett rent principiellt plan skulle kunna utsättas för en enorm mängd uppgifter, vilka i sin tur bedöms av en ofantlig mängd bedömare (Bachman 2004:178). Medelvärdet av alla möjliga totalpoängsummor på ett prov och av bedömningarna skulle då betraktas som elevens universumpoäng, eller dennas sanna prestationsnivå med avseende på det konstrukt eller den egenskap som vi är intresserade av att observera och mäta. Eftersom det är omöjligt att observera alla möjliga uppgifter och bedömningar av dessa uppgifter får vi hålla till godo med ett stickprov och bestämma med vilken precisionsgrad detta stickprov kan generaliseras till en sann prestationsnivå.

Generaliserbarhetsteoretiska modeller används för att analysera stickprovets generaliserbarhet och hur stort inflytandet från olika felkällor är. I jämförelse med måttet för inomklasskorrelation inom standardansatsen kan fler felkällor tas med i beräkningen. I denna undersökning kan själva urvalet av texter med provuppgifter

som ingick i denna provversion ses som ett någorlunda representativt stickprov av alla de läsförståelseuppgifter som hade varit möjliga i provet. Bedömarna kan ses som ett stickprov av populationen bedömare och elevlösningarna som någorlunda representativa för populationen elever.

Centralt är alltså utgångspunkten att varje elevs totalpoäng kommer att avvika från den genomsnittliga poängen för samtliga provlösningar. Den avvikelsen i relation till medelvärdet kan med statistisk terminologi kallas varians (ett mått som också har vissa formella egenskaper).[2] För ett provresultat vill vi att avvikelserna i relation till medelvärdet är höga för elever med höga poäng respektive låga poäng. Framför allt är det viktigt att den avvikelsen, variansen, ska förklaras av att eleverna har olika god läsförståelse. Det generaliserbarhetsteorin är intresserad av är att få fram vilket slags ”brus” det kan finnas i tolkningen av elevens resultat för att så säkert som möjligt kunna uttala sig om elevens förmåga. I den här undersökningen om samstämmighet i bedömning är det viktigt att förstå hur mycket varians bedömarna ger upphov till, vid tolkningen av elevernas resultat.

Formellt handlar generaliserbarhetsteorin om att beskriva hur stor portion av den totala variansen som kan härledas till den totala variansens olika komponenter. Med hjälp av generaliserbarhetsteoretiska modeller går det alltså att besvara frågor om hur stor del av resultatvariationerna som beror på 1) skillnader i elevprestationer, 2) skillnader i bedömarnas stränghet, och dessutom 3) skillnader i uppgifternas egenskaper. I föreliggande studie tillämpas en design där vi antar att elevernas resultat beror på fasetterna uppgifter och bedömare. Med generaliserbarhetsteoretisk vokabulär benämns denna design tvåfasettdesign. Inom generaliserbarhetsteori är det i detta fall endast uppgifter och bedömare som benämns fasetter. De komponenter som skapar varians i en tvåfasettdesign är emellertid fler.

De grundläggande komponenterna är elevens förmåga, bedömarens stränghet och uppgiftens egenskaper, precis som ovan nämnts. Dessutom förekommer kombinationer av dessa komponenter som skapar unik varians, till exempel mellan elever och bedömare (elever*bedömare). Bedömare kan reagera specifikt på vissa typer av elevprestationer. Det skulle kunna vara så att vissa lärare tenderar att reagera på vissa drag, såsom stavfel, eller något annat utmärkande för vissa elevers lösningar, trots att detta inte ska vara utslagsgivande för bedömningen. På samma sätt kan elever reagera säreget beroende på uppgiftens specifika egenskaper (elever*uppgifter). Ett exempel skulle kunna vara om en viss elev inte har fått undervisning om inslag som är centrala i kursen, såsom centrala motiv, berättarteknik och vanliga stildrag. Då klarar inte denna elev just denna uppgiftstyp, vilket skapar varians i totalresultatet. Bedömare kan också reagera på ett individuellt sätt i relation till vilken uppgift det är de bedömer. Det skapar varianskomponenten (uppgifter*bedömare). Övrig varians sammanfattas i det som benämns residual, ett mer obestämt mätfel. (Verhelst 2004:2; Brennan 2001:22–23)

[2] Var (X) $=(\bar{X}-\mu)^2$, där μ står för avvikelsen i relation till medelvärdet.

Med hjälp av varianskomponenterna kan därefter ett reliabilitetsmått – generaliserbarhetskoefficienten – beräknas. Poängen med reliabilitetsmått överlag kan metaforiskt beskrivas som att man försöker skatta den så kallade signalen relativt bruset i det aktuella testet. Ett läsförståelseprov syftar till att signalera elevens läsförmåga. Men det förekommer alltid brus som stör tolkningen. Reliabilitetsmåttet kan ses som en indikator på andelarna brus respektive signal, det vill säga hur stor andel varians som är signal – och hur stor andel varians som stör. Om hälften är brus och hälften är signal, det vill säga vid reliabilitet på 0,5 så ger testet inte något underlag för tolkning av signalen. Ju mindre brus, desto högre värde, och provresultatet blir mer tillförlitligt att generalisera.

Generaliserbarhetskoefficienten (g-koefficienten) kan kvalitativt alltså tolkas på samma sätt som många andra reliabilitetsmått inom klassisk testteori, såsom Cronbachs α (Brennan 2001:35). Man behöver dock ta hänsyn till att ju fler fasetter man väljer att inkludera i en viss undersökningsdesign, desto mer varians kommer man också att upptäcka. I en studie av det här slaget är skillnaden mellan mer traditionella reliabilitetsmått för itemprov såsom Cronbachs α, som endast utgår ifrån uppgiftsfasetten, att g-koefficienten dessutom väger in bedömarfasetten i den samlade reliabilitetsskattningen. Det för med sig att det är möjligt att skatta reliabiliteten för poängsummor från ett mer generellt så kallat universum, än vad Cronbachs α som rent internkonsistensmått gör. Bilden blir mer nyanserad, genom att studien ger bättre underlag kring hur det aktuella provet faktiskt fungerar när också bedömningsfasetten läggs till. Men det medför samtidigt att g-koefficienten per definition kommer att bli lägre än Cronbachs α, eftersom den tagit hänsyn till ännu mer möjligt brus.

I en beslutsstudie kan därutöver andra scenarier testas, till exempel för att besvara frågor om hur stor reliabilitet det är möjligt att förvänta sig att uppnå givet att provet administreras på olika sätt, till exempel genom att använda flera bedömare för att bedöma varje enskild elevprestation. En grundlig introduktion till generaliserbarhetsteori står att finna i Brennan (2001).

3.3 Allmänna riktvärden och tidigare studier av samstämmighet vid bedömning av läsprov

När rapportens indikatorer på samstämmighet ska värderas används riktvärden från tre olika sammanhang, vilka redovisas i detta avsnitt. För det första finns det riktvärden som ofta fungerar som standardreferenser, i regel hämtade från handböcker och annan forskningslitteratur. De värden som anges brukar vara resultatet av ackumulerad erfarenhet och utredningar av rent statistiska egenskaper hos måtten i fråga. För det andra finns det stipulerade riktvärden som förvisso bygger på standardreferenser eller erfarenheter, men med skillnaden att de kan vara direkt styrande för vilka beslut som bör fattas givet ett specifikt resultat. Ett tredje slag av riktvärden står att finna i undersökningar av bedömaröverensstämmelse i liknande

provtyper. Här finns det också ett par studier genomförda med utgångspunkt i svenska nationella prov, vars uppmätta värden är av särskilt stort intresse eftersom de mäter samstämmighet inom ramen för samma nationella provsystem.

3.3.1 Standardreferenser

En flitigt citerad skala för värdering av kappa-koefficienter såsom Cohens κ är den som föreslagits av J. Richard Landis och Gary G. Koch (1977:165). Ambitionen har från författarnas sida varit att skapa en enhetlig nomenklatur för att med ord beskriva statistiska utfall, snarare än att ge statistiskt motiverade gränser. I själva verket är gränsdragningarna enligt författarna helt godtyckligt dragna. Skalan föreslår ändå följande spann med tillhörande tolkningar:

- < 0,00 dålig överensstämmelse
- 0,0–0,20 liten överensstämmelse
- 0,20–0,40 skälig/lovande överensstämmelse
- 0,40–0,60 rimlig/rätt bra överensstämmelse
- 0,60–0,80 påtaglig överensstämmelse
- 0,80–1,00 nästan perfekt överensstämmelse

En något modifierad variant av denna nivåtolkning har, i relation till Landis och Koch (1977), föreslagits av Joseph L. Fleiss (2003:604). Fleiss gradering är mer grovhuggen:

- < 0,40 dålig överensstämmelse
- 0,40–0,75 medelgod/god överensstämmelse
- 0,75–1,00 utomordentlig överensstämmelse

Fleiss (1971) har dessutom utvecklat Cohens κ för att kunna ge en sammanfattande skattning av tre eller fler bedömare som har bedömt samma prestationer. Denna koefficient bär Fleiss namn (Fleiss κ, utläses Fleiss kappa) och tolkas enligt samma skalor som Cohens κ.

När det gäller reliabilitetsmått som inomklasskorrelation går det att göra värderingar både utifrån allmänna skalor för reliabilitetsmått och utifrån mer specifika skalor för inomklasskorrelation, eftersom nivåerna är ungefär desamma. Robert F. DeVellis (1991:85) menar att reliabilitetskoefficienter rent generellt under 0,6 är att betrakta som oacceptabla, medan 0,7–0,8 är respektabla och värden över 0,9 så höga att man kan överväga att korta ner testlängden om testet är av sådan art. Det gäller alltså också den ovan nämnda generaliserbarhetskoefficienten. Mer sällan föreslås specifika skalor för inomklasskorrelation, och Leslie G. Portney och Mary Watkins hör till undantagen. Enligt Portney och Watkins (2015:595) är 0,75 den lägsta godtagbara nivån av inomklasskorrelation, medan mått på 0,9 eller högre är eftersträvansvärda när ambitionen är att dra valida slutsatser med stora konsekvenser i enskilda fall utifrån det mätinstrument som används.

3.3.2 Stipulerade riktvärden

Stipulerade riktvärden föreslås i regel av organisationer och myndigheter med ansvar för prov med särskilt höga insatser, som de nationella proven i Sverige. Skolverket (2017:21) har till exempel angivit nivån 0,6 som ett minimikrav för performansuppgifter, alldeles oavsett vilken typ av mått det gäller. Detta minimikrav gäller prestationer som skapar komplexa bedömningssituationer. Uppgifter i läsförståelseprov är emellertid inte lika komplexa som prov i muntlig och skriftlig framställning. Samstämmigheten vid bedömningar av elevproducerade svar i ett läsförståelseprov har därför förutsättningar att nå högre nivåer, men detta regleras inte på ett uttalat sätt i *Skolverkets systemramverk för nationella prov* (2017).

När PISA genomför sina reliabilitetsundersökningar förväntar man sig till exempel att bedömarna inom respektive deltagande land ska ha en genomsnittlig överensstämmelse på 92 procent och som lägst 85 procent för enstaka uppgifter (OECD 2015:257). Liknande riktvärden används även inom PIRLS, där man under konstruktionsfasen brukar resa varningsflagg för uppgifter som har en bedömaröverensstämmelse som är lägre än 85 procent (Martin et al. 2017:10.6). Värt att tillägga är att förutsättningarna för bedömning inom de internationella kunskapsmätningarna skiljer sig avsevärt från genomförandet i det svenska skolsystemet. I PISA och PIRLS sker all bedömning inom ett land av en liten grupp bedömare som specialtränats för ändamålet. I det nationella provsystemet i Sverige sker bedömningen av elevers provlösningar på varje skola av undervisande lärare. Sådana skillnader i förutsättningar gör det väntat att samstämmigheten i bedömningen är olikartad inom de olika provtyperna.

3.3.3 Uppmätta värden för bedömarsamstämmighet i tidigare studier

När det gäller uppmätta värden av samstämmighet i bedömning av kortare svar på prov i läsförståelse är antalet publicerade undersökningar förhållandevis få. I en undersökning av samstämmighet i bedömning av prov utformade som lucktexter, det vill säga texter där testtagaren ska fylla i de ord som utelämnats, förefaller samstämmigheten vara hög (de Santi & Sullivan 1984). Inomklasskorrelationen uppmättes till 0,89–0,99 i genomsnitt vid bedömningar av själva förståelsen och språkanvändningen. Dock var det klart lägre samstämmighet vid bedömningar av svarens grammatiska standard, med genomsnittliga inomklasskorrelationer mellan 0,56 och 0,86.

Vid en undersökning av samstämmighet i bedömning av elevsvar i de norska läsförståelseproven i årskurs 8 har Tengberg et al. (2018) jämfört samstämmigheten i en grupp bestående av tjugo lärare med samstämmigheten i en grupp bestående av sju provkonstruktörer. Alla bedömare gjorde bedömningar av 23 elevers svar på 11 uppgifter. I lärargruppen varierade konsensusmåttet Cohens κ från 0,51 till 0,89 med ett medianvärde på 0,74, medan motsvarande värden hos provkonstruktörerna var 0,77–0,92 med ett medianvärde på 0,89. Måttet på samstämmighet i grupperna

som helhet, Fleiss κ, uppmättes till 0,72 i lärargruppen och 0,87 i provkonstruktörsgruppen. Med hjälp av en mångfasetterad Rasch-analys, som bland annat kan användas för att analysera varierande grader av stränghet bland bedömare, visade det sig att det fanns klart större variation i stränghet inom lärargruppen än bland provkonstruktörerna. Detta torde också vara en bakomliggande förklaring till de lägre κ-koefficienterna i lärargruppen.

Inom ramen för det svenska nationella provsystemet finns det två publicerade studier av samstämmighet i bedömning av prov med kortare elevproducerade svar. Den ena har utförts av Anna Lind Pantzare (2015) som undersöker bedömningar av det nationella provet i matematik för gymnasieelever. I studien har fem lärare, utvalda från en grupp av femton frivilliga lärare, bedömt 99 elevlösningar på ett prov där 23 av 24 frågor innebar någon form av elevproducerade svar. Den procentuella överensstämmelsen mellan bedömarna var som lägst 86 procent och som högst 94 procent, medan κ-koefficienten varierade mellan 0,78 och 0,91. Konsistensmåttet Spearmans rangkorrelation uppmättes till mellan 0,88 och 0,93, medan det sammanfattande måttet på konsistensen som helhet i bedömargruppen, Cronbachs α, uppgick till 0,98. Samstämmigheten i bedömningar av detta nationella prov i matematik tycks alltså befinna sig på samma nivå som eftersträvas inom PISA-ramverket.

Den andra studien använder läsförståelsedelen från det nationella provet i svenska för elever i årskurs 9. Mikael Tengberg och Gustaf B Skar (2016) samlade in bedömningar från sex lärare fördelade på tre skolor som bedömde tre elevlösningar. En av de i studien ingående lärarna hade redan som formell uppgift att bedöma de elevlösningar som användes. Provet som bedömdes innehöll 14 uppgifter som krävde elevproducerade svar, varav 10 skulle bedömas enligt en skala med tre eller fler poängnivåer. Den procentuella överensstämmelsen låg som lägst på 67 procent och som högst på 93 procent, med ett medianvärde på 76 procent. Cohens κ-koefficient visar på en liknande spännvidd med värden från 0,61 till 0,93 och ett medianvärde på 0,73. Slutligen mättes inomklasskorrelationen till 0,82.

Sammanfattningsvis kan vi konstatera att det är önskvärt att de beräknade koefficienterna från mätningar av samstämmighet i bedömningar når de övre nivåerna av riktvärden som anges i forskningshandböcker. Man bör emellertid ha i åtanke att handböckernas riktvärden ofta är generellt formulerade utan hänsyn tagen till specifika syften med det mätinstrument som ska användas eller utvärderas. Vid sammanhang där bedömningar av enskilda elevers prestationer ska användas för att fatta beslut på individnivå är en hög samstämmighet viktigare än om alla elevers resultat ska sammanfattas på nationell nivå. Därför är det viktigt att utvärderingen av samstämmighet i bedömningar av elevprestationer görs med hänsyn tagen till stipulerade riktvärden, som när Skolverket föreslår 0,6 som ett minimikrav för performansuppgifter. Tidigare studier av samstämmighet i bedömningar av elevprestationer under likartade förutsättningar visar att det förvisso är möjligt att nå höga riktvärden men att det är troligt att vi också kommer att observera förhållandevis låga uppmätta värden på samstämmighet.

3.4 Urval och datainsamling

Urvalet av elevtexter och insamlingen av bedömningar har gjorts av medarbetare inom Gruppen för nationella prov i svenska och svenska som andraspråk vid Institutionen för nordiska språk vid Uppsala universitet. I följande avsnitt beskrivs urvalsprocesserna bakom de elevlösningar och de lärare som bedömde elevlösningarna samt en statistisk motivering till urvalsstorlekarna.

3.4.1 Insamling av elevlösningar

Insamlingen av elevlösningar gjordes under våren 2017 i samband med att det nationella provet i fråga genomfördes på skolor runt om i landet. Provet som ligger till grund för studien bestod av 18 uppgifter där eleverna själva formulerar en lösning samt 7 flervalsuppgifter.

Urvalet av elevlösningar gjordes slumpmässigt från prov genomförda i två mindre klasser på yrkesförberedande program och en klass på ett studieförberedande program på två olika skolor. Insamlingen bestod av sammanlagt 63 elevlösningar (34 från yrkesförberedande program och 29 från studieförberedande program). Elevlösningarna kopierades och avidentifierades på plats. På den ena skolan gjordes detta av undervisande lärare och på den andra skolan av en provkonstruktör.

En provkonstruktör sorterade sedan bort de elevlösningar där en eller flera uppgifter lämnats obesvarade eftersom beräkningar av samstämmighet i bedömningen av sådana lösningar skulle kunna leda till missvisande höga skattningar. Sammanlagt 10 elevlösningar – 5 från varje programinriktning – valdes ut slumpmässigt. Provkonstruktören gjorde sedan en bedömning av dessa 10 elevlösningar. Enligt denna bedömning skulle fördelningen av betyg innebära att de utvalda lösningarna täcker hela betygsskalan (tabell 1).

Tabell 1. Urval av elevlösningar. Provkonstruktörens bedömning.

Betyg	F	E	D	C	B	A
Antal lösningar	1	2	2	2	2	1

3.4.2 Insamling av bedömningar

Urvalet av bedömare gjordes enligt följande steg. Först kontaktades var femte lärare som besvarat den lärarenkät som följde på det nationella provet vårterminen 2017. I brevet som gick ut till totalt 65 lärare beskrevs syftet med studien tillsammans med uppgifter om ersättning, och lärarna fick frågan om de var villiga att delta under dessa förutsättningar. Sammanlagt 18 av de 65 lärarna svarade att de var

villiga att delta. Bland de 18 lärarna valdes varannan lärare ut enligt bokstavsordning, och urvalsalgoritmen fortsatte genom listan tills 10 lärare valts ut. Det slutgiltiga antalet bedömare i studien uppgår till 9, eftersom en av bedömarna valde att avstå under själva bedömningsfasen. I strikt statistisk mening rör det sig med andra ord om ett slumpmässigt stickprov som på grund av det frivilliga elementet kan vara skevt i förhållande till den ursprungliga målpopulationen, som var gymnasielärare i svenska. Säkra slutsatser om nivån av samstämmighet är därför begränsade till gymnasielärare som kan tänka sig delta i bedömningsstudier. Hur stor andel denna grupp utgör i populationen gymnasielärare i svenska som helhet är svårt att säga, även med vetskapen om att lite mer än en fjärdedel av alla tillfrågade lärare svarade att de ville delta.

Kopior av alla tio elevlösningarna skickades tillsammans med ett följebrev ut till de utvalda bedömarna. I följebrevet fanns en beskrivning av utskicket och instruktioner om att utifrån ordinarie bedömningsanvisningar bedöma uppgifterna där eleverna själva formulerat sina svar. Till skillnad från bedömningar av ordinarie prov, där det står lärarna fritt att ta hjälp av sina kollegor, uppmanades bedömarna att i detta fall göra bedömningarna enskilt.

4 Resultat

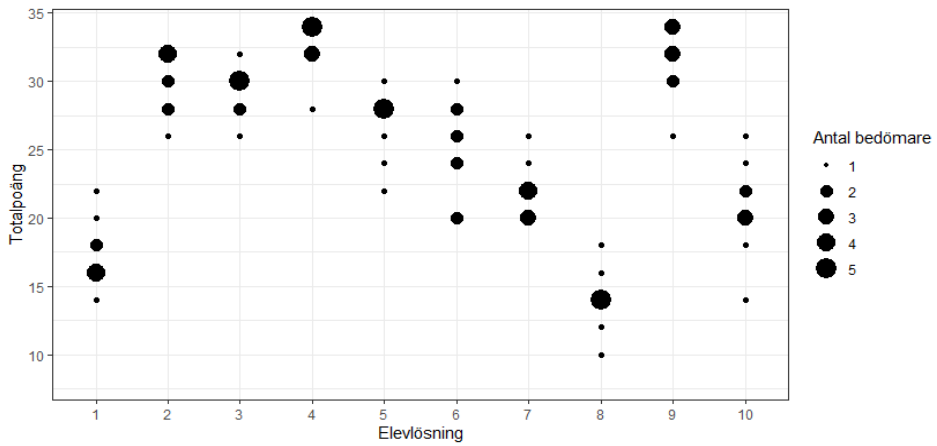
I detta kapitel beskrivs materialet i korthet och därefter rapporteras resultatet av de statistiska beräkningarna. Resultaten av konsensusskattningar, inomklass-korrelation, mångfasetterad Rasch-analys samt generaliserbarhetsanalys redovisas också här.

4.1 Deskriptivt om materialet

För att mäta samstämmigheten i bedömningarna används i den här rapporten endast de 18 uppgifter som kräver elevers egna formuleringar. Det förefaller nämligen rimligt att utgå från att samstämmigheten är total när det gäller flervalsfrågor.

Den bedömande läraren tar ställning till om elevens lösning ska bedömas som ej godtagbar och alltså ge 0 poäng, eller om lösningen ska bedömas som godtagbar, det vill säga ge 2 poäng. I figur 2 nedan visas ett diagram över hur totalpoängsummorna för de 18 uppgifterna kan variera för olika elevlösningar. På y-axeln återges totalpoängsumman och på x-axeln elevlösningens nummer. Varje punkt i diagrammet motsvarar en bedömningspunkt. Om två bedömare kommit fram till exakt samma poängsumma motsvaras dessa av en och samma punkt. Den genomsnittliga totalpoängsumman i materialet med hänsyn tagen till alla bedömares skattningar och alla elevers totala poäng är 25,0.

Det finns emellertid en viss spridning i bedömningar för varje elevlösning, vilket framgår av diagrammet. Men spridningen är systematisk. Den elevlösning där bedömningarna varierar minst är elevlösning 4. Här har de flesta bedömningarna landat på 32 eller 34 poäng och medelvärdet av bedömningarna är 32,6. Variansen i bedömningen är således liten, vilket är önskvärt. I elevlösning 1 sträcker sig den samlade bedömningen i form av en totalpoängsumma från 14 poäng som lägst till 22 poäng som högst. Variansen är alltså något större för denna elevlösning.



Figur 2. Variation i totalpoäng och bedömning.

4.2 Konsensusskattningar

Den exakta samstämmigheten var 89,8 procent i genomsnitt med medianen 91,4 för samtliga parvisa kombinationer av bedömare. Spridningen mellan det högsta (93,9 procent) och det lägsta (81,1 procent) värdet för olika bedömarpar var 12,8 procentenheter, vilket kan betraktas som en relativt liten spridning. Överensstämmelsen är med andra ord nästan lika stor som man eftersträvar i PISA-undersökningarna (OECD 2015:257). Samstämmigheten tycks vidare vara ungefär lika stor som i Lind Pantzares (2015) undersökning av nationella prov i matematik, där den exakta samstämmigheten varierade från 86 till 94 procent. Jämfört med Tengbergs och Skars studie (2016:8) är den exakta samstämmigheten större och spridningen mindre i föreliggande undersökning. I deras studie var medianvärdet 76 procent och spridningen 26 procentenheter.

Eftersom uppgifterna poängsätts enligt en skala med två steg är det fullt möjligt att två bedömare av ren slump sätter samma poäng – eftersom det bara finns två steg att välja mellan. För att ta hänsyn till sådana slumpmässiga faktorer som beror på skalans utformning används som ovan nämnts Cohens κ-koefficient för att beräkna konsensus mellan samtliga parvisa kombinationer av bedömare. Cohens κ var 0,76 i genomsnitt[3] med en median på 0,79. Enligt de riktvärden som finns tyder detta resultat på utomordentlig samstämmighet (Fleiss 2003:604) och kategoriseras enligt Landis och Kochs (1977:165) som påtaglig och nära nästan perfekt samstämmighet. I Tengbergs och Skars studie för ett läsförståelseprov i årskurs 9 uppmättes Cohens κ till 0,73 i genomsnitt och i motsvarande studie för det norska nationella läsförståelseprovet var medianen 0,74 (Tengberg et al. 2018). Skillnaden mellan kappa-värdena i Tengbergs och Skars studie och föreliggande studie är mindre än när vi jämförde den procentuella överensstämmelsen, vilket illustrerar hur κ-koefficienten kan vara användbar för att jämföra värden mellan undersökningar där antalet poängnivåer skiljer sig åt.

För att använda studien som underlag för slutsatser om bedömaröverensstämmelse för ett läsförståelseprov i kurs 1 i populationen lärare, används som tidigare nämnts sannolikhetsbaserad inferens. Givet data från bedömningsstudien, och den modell för fyra bedömare som beskrivs i Broemeling (2012:187–190), är det 95 procents sannolikhet att κ-koefficienten för fyra pseudo-slumpmässigt dragna bedömare, beräknad enligt Fleiss metod, befinner sig inom spannet 0,71–0,84 med medelvärdet 0,78 och medianen 0,78. Detta är alltså vår statistiskt sett bästa gissning om nivån på samstämmigheten i den oerhörda mängd liknande studier som på ett teoretiskt plan skulle kunna genomföras.

Sammanfattningsvis kan vi konstatera att samstämmigheten i bedömningarna i denna studie nära nog når samma nivåer som eftersträvas i storskaliga internationella mätningar, vilket är anmärkningsvärt med tanke på de skilda förutsättningarna för bedömningens genomförande. Samstämmigheten är vidare jämförbar eller i vissa fall högre än de nivåer som uppmätts för andra nationella prov i svenska och matematik.

[3] Fleiss κ, som alltså beräknas för alla bedömare, uppmättes till 0,76.

4.3 Inomklasskorrelation

Konsistensskattningarna beräknas för bedömningar som resulterar i en totalpoäng på provet. Bedömningar av enskilda uppgifter i detta prov är emellertid inte lika intressanta att undersöka med hjälp av konsistensmått, i det här fallet den ICC-modell som har använts. För konsistensmått resulterar skillnader i bedömning automatiskt i motsatt rangordning för dikotoma uppgifter. Det förväntade genomsnittliga värdet är detsamma som för Cohens κ, vilket redovisats ovan, nämligen 0,76.

I det urval som undersökts framgår att inomklasskorrelationen uppgår till 0,92 (konfidensintervall 0,84–0,98) vilket enligt riktlinjerna redovisade ovan är uppe på nivåer av det som benämns ”nästan perfekt överensstämmelse”. Detta värde motsvarar för övrigt generaliserbarhetskoefficienten för en bedömare, om den beräknas för den sammanlagda poängsumman.

Värdena är signifikant skilda från det kritiskt låga värdet 0,6 som alltså är den nedre gränsen Skolverket fastställt för samstämmighet i bedömningar av performansuppgifter. Sannolikheten är med andra ord minimal att vi hade fått dessa värden om den egentliga inomklasskorrelationen faktiskt ligger omkring 0,6.

4.4 Mångfasetterad Rasch-analys

Den mångfasetterade Rasch-analysen bidrar till att skatta bedömarnas stränghet. I tabell 2 nedan redovisas stränghets- och separationsskattningar för lärarnas bedömningar. Det övergripande resultatet är att bedömningarna överlag är samstämmiga trots att det finns vissa mindre skillnader i stränghet mellan bedömarna. Spannet mellan den i genomsnitt mildaste och strängaste bedömaren sträcker sig från 0,59 till 0,76 råpoäng (spännvidden 0–1). Även om detta förefaller vara små skillnader går det emellertid att med ett separationsindex på 2,40 urskilja två skikt av bedömare, baserat på variationen i stränghet och mätt enligt den logitskala som används i den mångfasetterade Rasch-analysen. Dessa skillnader är ungefär lika stora som de skillnader Tengberg et al. (2018) uppmätte bland 20 lärare som bedömde norska läsförståelseprov.

För att få en uppfattning om huruvida ett separationsindex på 2,4 är högt eller lågt kan vi jämföra med bedömningar av längre texter. I en studie av bedömningar av skrivproven för svenska 1 var separationsindex 1,87 för enskilda bedömningar (Dalberg 2019). Det är väntat med tanke på att antalet skalsteg är färre, sex betygssteg, än antalet möjliga poängsummor i detta prov, 27. Ju fler steg desto mer nyanserat kommer bedömare kunna urskiljas. I en sammanställning av resultaten från totalt 12 studier visar Eckes (2015:180–182) att separationsindex i regel indikerar ett flertal distinkta klasser och att separationsreliabiliteten är stor. Faktum är att av de 12 studierna Eckes redovisar är det ingen som har ett separationsindex under 2,43, och alla utom tre av studierna har ett index på 4 eller högre (ett separationsindex på 11,45 som högsta notering).

Tabell 2. Stränghets- och separationsskattningar av lärarnas bedömningar.

Mått	Värden
Logit min (max råpoäng)	-0,48 (0,76)
Logit max (min råpoäng)	0,76 (0,59)
Logit medelvärde (råpoäng)	0,00 (0,69)
Logit standardavvikelse (råpoäng)	0,36 (0,05)
Separationsreliabilitet R	0,71
Separationsindex H (Separationskvot G)	2,40 (1,55)

Skillnaderna i stränghet kan också omvandlas till summor av de 18 bedömda uppgifterna. När provet genomförs ger varje uppgift 0 eller 2 poäng. Den genomsnittliga poängsumman för samtliga bedömares bedömningar är 25 poäng (se avsnitt 4.1). Att jämföra enskilda bedömares genomsnittliga poängsumma för de tio bedömda eleverna med genomsnittet för samtliga bedömningar ger ett mått som kan vara lättare att tolka, då det uttrycks i råpoäng.

Med hjälp av den mångfasetterade Rasch-analysen kan summorna för enskilda bedömares genomsnittliga poäng skattas. Dessa spänner enligt analysen då i genomsnitt, för alla 10 elevlösningar, från 21,24 till 27,36 poäng (differens 6,12). Den strängaste läraren ger därmed en genomsnittlig totalpoäng på samtliga elevers lösningar på 21,2 medan den mildaste bedömaren ger något fler poäng än det genomsnittliga antalet poäng i materialet, det vill säga 27,4. Det går också att tolka som att den strängare bedömaren avviker mer från det samlade medelvärdet än den mildare bedömaren. Det skulle till exempel kunna innebära att en elev som blir bedömd av den strängaste bedömaren riskerar att få 6 poäng mindre på uppgifter som kräver elevproducerade svar än om samma elev blir bedömd av den mildaste bedömaren. I en studie av norska prov (Tengberg et al. 2018) bedömdes en skillnad mellan den strängaste och mildaste bedömaren på 1,87 råpoäng, motsvarande 17 procent av provets maxpoäng, som påtaglig. I föreliggande studie motsvarar 6 poäng 11 procent av provets maxpoäng, vilket är väntat med tanke på att provet innehåller ett antal mer komplexa tolkningsuppgifter.

Den stora samstämmigheten mellan bedömarna uttrycks bland annat i att modellen är välanpassad till alla bedömare. Detta framgår av att måttet inlier-sensitive fit befinner sig inom spannet 0,95–1,14 för alla bedömare, det vill säga med god marginal inom det spann som brukar anges som riktvärde (0,75–1,3). Med andra ord är det ingen av bedömarna som gjort varken säregna eller försiktiga bedömningar. Eftersom vi vet att uppgifterna varierar i svårighetsgrad och endast rymmer två skalsteg är det inte särskilt troligt att observera någon ”försiktig” bedömare.

4.5 Studier utifrån generaliserbarhetsteorin

Syftet med generaliserbarhetsteorin är att identifiera hur stor andel av variationen, i totalpoäng, den så kallade variansen, som har att göra med på förhand kända felkällor, så kallade varianskomponenter. Det är också möjligt att genomföra undersökningar, som fungerar som underlag för beslut, så kallade beslutsstudier. Inledningsvis i detta avsnitt redovisas en generaliserbarhetsstudie. Därefter redovisas en beslutsstudie.

4.5.1 Generaliserbarhetsstudie

I den tvåfasettdesign som denna studie har, kan följande komponenter urskiljas:

(1) Elever
(2) Bedömare
(3) Uppgifter
(4) Elever*bedömare
(5) Elever*uppgifter
(6) Uppgifter*bedömare
Residual (övrig varians)

I tabell 3 nedan beskrivs hur stor andel respektive komponent står för av den totala variansen i elevresultaten. Skattningarna har gjorts dels utan flervalsuppgifter, dels med flervalsuppgifter. Den del av variansen som hänförs till eleverna speglar skillnader i elevernas faktiska prestationer inom den egenskap som testas – i det här fallet läsförståelse. Från ett psykometriskt perspektiv vill man att mätinstrumentet ska maximera variansen hos denna komponent, men det är också väntat att det i itemprov är så att det finns en interaktion mellan enskilda elever och uppgiftens egenskaper. Den största komponenten är elever och uppgifter (elever*uppgifter), med 49,4 procent, vilket framstår som naturligt i jämförelser med andra prov. I en undersökning av PISA-uppgifter (OECD 2005:213) hamnar denna varianskomponent på motsvarande procentandel. Denna komponent utgör en ännu större andel – 61,2 procent – när flervalsuppgifterna inkluderas i beräkningarna.

Tabell 3. Varianskomponenter i andelar av den totala variansen.

Varianskomponent	Andelar (%), 18 uppgifter (utan flervalsuppgifter)	Andelar (%), 25 uppgifter (med flervalsuppgifter)
Elever	11,2	10,0
Bedömare	0,9	0,4
Uppgifter	16,1	13,1
Elever*Bedömare	0,1	0,0
Elever*Uppgifter	49,4	61,2
Uppgifter*Bedömare	3,6	2,6
Residual	18,8	12,7

Bedömarkomponenten indikerar framför allt variationer i stränghet mellan bedömarna. Om man utgår från antagandet att det finns en korrekt bedömning av elevprestationer bör bedömarkomponenten stå för en blygsam andel av variansen. I det här fallet utgör denna komponent 0,9 procent, vilket alltså antyder att bedömarna är nära nog jämbördiga i sina bedömningar. När flervalsuppgifterna inkluderas i beräkningarna sjunker denna komponent till 0,4 procent.

Uppgiftskomponenten fångar upp variationer i uppgifternas svårighetsgrad. Eftersom läsförståelseproven i regel består av uppgifter som varierar i svårighetsgrad vet vi på förhand att denna komponent kommer att stå för en påtaglig andel av variansen, vilket den också gör med 16,1 procent när undersökningen begränsas till bedömda uppgifter, och 13,1 procent när flervalsuppgifterna inkluderas.

Komponenten elever och bedömare (elever*bedömare) tar fasta på hur konsekventa bedömarna varit i bedömningarna av respektive elev. För att minimera denna komponent bör bedömarna vara överens om rangordningen av eleverna. Denna komponent står för bara 0,1 procent av variansen, vilket indikerar att bedömarna på det stora hela är överens om rangordningen av eleverna.

Bedömarnas variation i bedömningarna av olika uppgifter (bedömare*uppgifter) står för 3,6 procent av variansen. Det innebär att bedömarnas stränghet inte varierar nämnvärt från en uppgift till en annan. Om denna komponent hade visat sig stå för en avsevärt större andel av den totala variansen hade det alltså berott på att några bedömare varit strängare än andra vid bedömningen av vissa uppgifter, och mildare än andra bedömare vid bedömningen av andra uppgifter.

Slutligen samlas i komponenten benämnd residual i tabell 3 den skattning av variansen som beror på en kombination av icke observerade mätfel och interaktionseffekter som inte går att skatta separat från dessa mätfel. Här står den för 18,8 procent

av variansen. Det är inte ovanligt att denna komponent står för en påtaglig del av den totala variansen.

Resultaten från variansanalysen kan sammanfattningsvis tolkas som att endast en försumbar del av variansen för elevens totalpoängsummor beror på variationer mellan bedömare. Det är helt enkelt svårt att observera några påtagliga skillnader i bedömarnas stränghet eller ens några inkonsekventa bedömningar av enskilda elevers prestationer.

Utifrån de skattade varianskomponenterna kan man beräkna ett reliabilitetsmått – den så kallade generaliserbarhetskoefficienten – som beskriver hur stor andel av den observerade poängens varians som utgörs av universumpoängens varians. Denna koefficient har i föreliggande studie skattats till 0,80 när flervalsuppgifterna exkluderas respektive 0,81 när flervalsuppgifterna inkluderas, vilket enligt konventionella måttstockar betraktas som en respektabel nivå.

4.5.2 Beslutsstudie

En styrka hos generaliserbarhetsteoretiska ansatser är att det finns beräkningsmodeller för att med hjälp av varianskomponenterna skatta generaliserbarhetskoefficienter för olika antal uppgifter och bedömare. Det blir med andra ord möjligt att göra skattningar av hur mycket tillförlitligare ett beslut om en elevs prestationsnivå baserat på ett kunskapstest skulle bli om det utökades med ytterligare bedömare eller uppgifter. Det går därför att genomföra en beslutsstudie, där syftet alltså är att närmare utforska hur reliabiliteten hos elevernas resultat förändras om vi skulle förändra förutsättningarna i form av antalet bedömare eller antalet uppgifter.

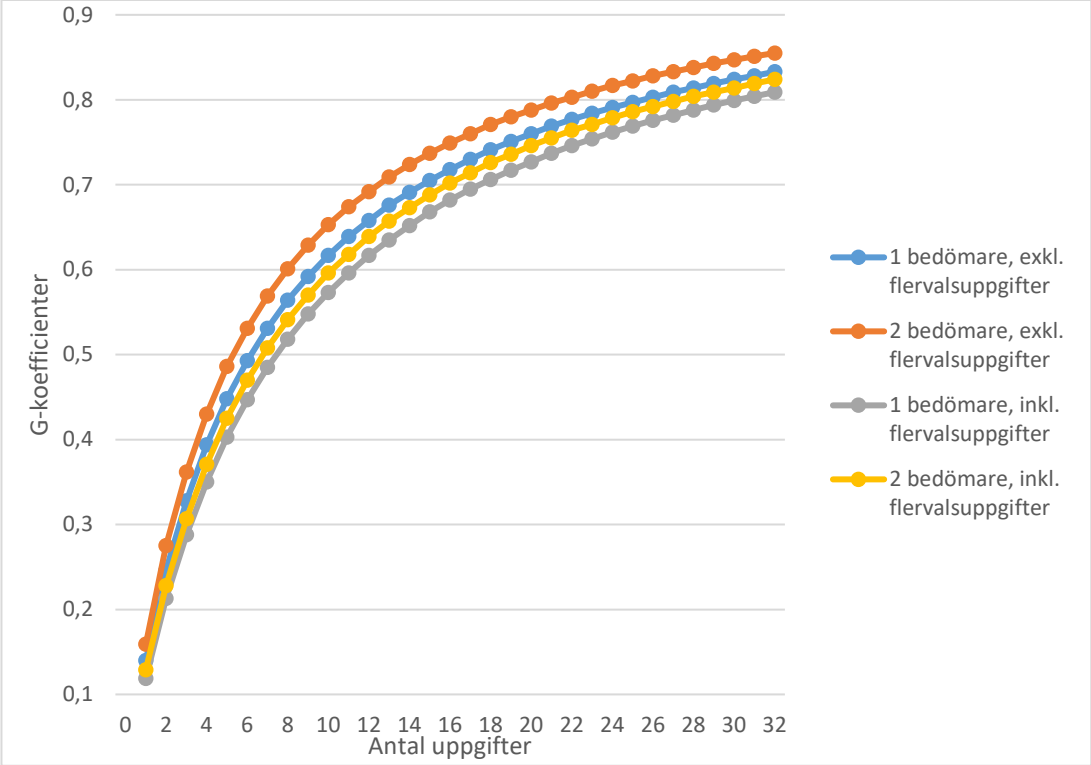
Förhållandet mellan generaliserbarhetskoefficienten och antalet uppgifter, bedömare samt om uppgifterna inkluderar flervalsuppgifter eller ej illustreras i figur 3a och 3b, där x-axeln visar antalet uppgifter och y-axeln nivån på generaliserbarhetskoefficienten. Kurvorna illustrerar en specifik kombination av antalet bedömare och ingående uppgiftstyper i beräkningarna. Generaliserbarhetskoefficienten – och därmed också reliabiliteten – blir större ju fler bedömare som bedömer en elevs lösning på en uppgift. Koefficienten blir också större ju fler uppgifter som bedöms per elev. Reliabiliteten hos elevernas resultat kan alltså höjas både genom att öka antalet uppgifter och genom att öka antalet bedömare. Det analyserade läsförståelseprovet innehåller 18 uppgifter som kräver bedömning av elevers egna formuleringar. Om en enda bedömare används för att bedöma alla elevers lösningar förväntas generaliserbarhetskoefficienten bli 0,74. För att nå koefficienter över 0,80 skulle man antingen behöva utöka provet till 26 uppgifter (se blå linje i figur 3a nedan) eller använda två bedömare som bedömer 22 uppgifter (röd linje i figur 3a).

När vi inkluderar provets flervalsuppgifter skulle det krävas 31 uppgifter för en bedömare (grå linje i figur 3b) eller 28 uppgifter för två bedömare (gul linje i figur 3b) för att nå koefficienter över 0,80. Eftersom ett prov med flervalsuppgifter enligt

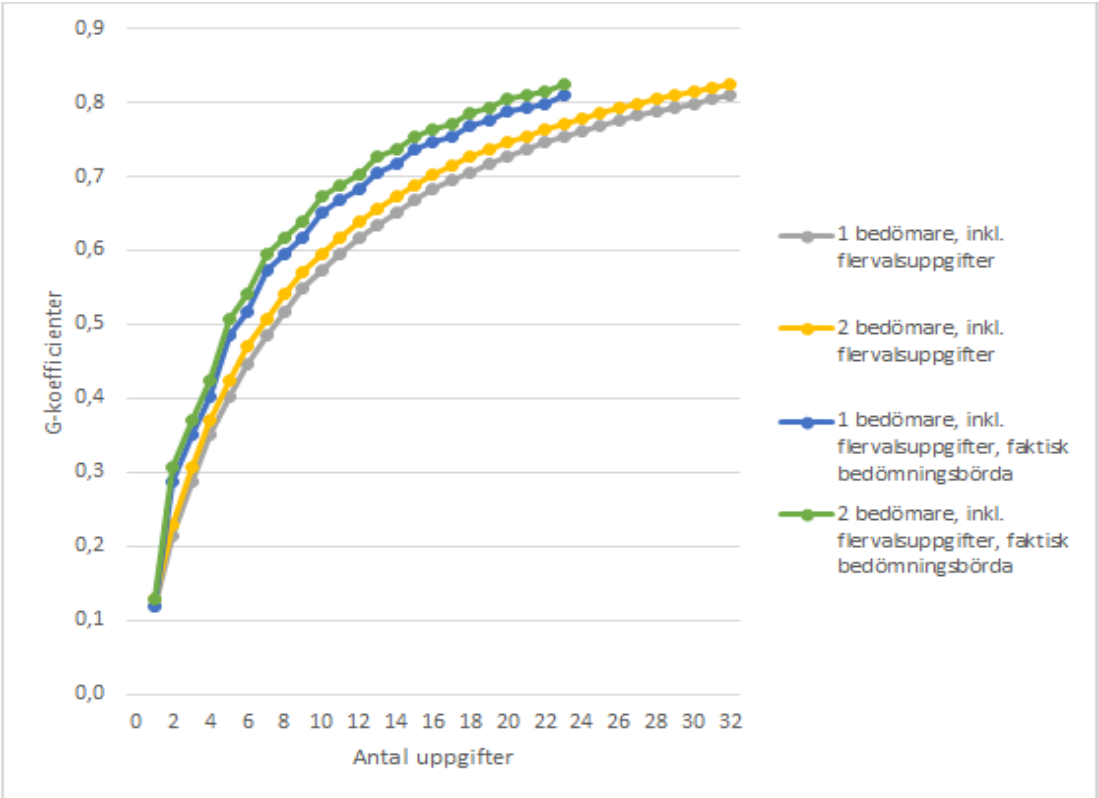
nu rådande balans mellan uppgiftstyper – 18 uppgifter med elevkonstruerade svar och 7 flervalsuppgifter – innehåller omkring 72 procent uppgifter som kräver bedömning blir själva bedömningsbördan något mindre än det totala antalet uppgifter kan ge sken av. En enskild bedömare skulle med andra ord behöva bedöma omkring 22 elevkonstruerade svar tillsammans med 9 flervalsuppgifter, medan två bedömare skulle behöva bedöma 20 elevsvar och 8 flervalsuppgifter för att nå koefficienter över 0,80.

Det torde vid ett försök att höja reliabiliteten vara mer kostnadseffektivt att utöka antalet uppgifter (företrädesvis av flervalstyp) i stället för antalet bedömare. Även om det vid varje steg i figur 3a och 3b ger en större ökning av generaliserbarhetskoefficienten att gå från en till två bedömare än att lägga till ytterligare en uppgift innebär utökandet av antalet uppgifter att det krävs färre bedömningar totalt sett. Därutöver kan man konstatera, något förenklat, att det går åt lika många bedömningar totalt sett att bedöma ett prov med dubbelt så många provuppgifter som att vid ett givet antal uppgifter gå från en till två bedömare. Det kan helt enkelt vara praktiskt mer genomförbart att utöka antalet provuppgifter än att utöka antalet bedömare.

Utifrån analyserna med generaliserbarhetsteoretiska verktyg kan man konstatera att variationer i bedömarnas bedömningar står för en relativt liten andel av den totala variansen jämfört med hur mycket skillnader i elevprestationer tycks bidra till densamma. Annorlunda uttryckt kan man tolka de statistiska resultaten som att elevernas provresultat snarare beror på elevprestationernas variation i kvalitet än på bedömarnas variation i stränghet. Det ger vidare en högre grad av reliabilitet att utöka antalet provuppgifter, även om det också skulle gå att höja reliabiliteten med ytterligare bedömare. Det torde dock vara mer kostnadseffektivt för det fullskaliga nationella provgenomförandet att utöka antalet uppgifter än att utöka antalet bedömare.



Figur 3a. Generaliserbarhetskoefficienten som en funktion av antalet uppgifter och antalet bedömare.



Figur 3b. Generaliserbarhetskoefficienten som en funktion av antalet uppgifter och antalet bedömare.

5 Slutsatser och diskussion

De olika statistiska mått som har använts i denna rapport vilar på skilda antaganden och ger underlag för olika typer av slutsatser. Tillsammans ger de en samlad bild av tillförlitligheten i en grupp bedömares bedömningar av elevprestationer på uppgifter med öppna svarsformat producerade inom ramen för nationella prov i svenska för kurs 1 på gymnasiet. Bedömargruppen utgjordes av ett slumpmässigt urval av gymnasielärare som var villiga att delta i studien.

Konsensusmått på bedömaröverensstämmelse utgår från att bedömarna kan och bör nå fram till exakt samma bedömning. När det gäller bedömningar av elevformulerade lösningar av enskilda uppgifter är det rimligt att förvänta sig att konsensus råder mellan bedömarnas bedömningar. I föreliggande rapport visade sig den exakta samstämmigheten vara 89,8 procent i genomsnitt, vilket kan betraktas som en god samstämmighet. Vidare uppmättes Cohens κ-koefficient till 0,76 i genomsnitt, vilket utifrån gängse riktlinjer brukar ses som utomordentlig respektive påtaglig överensstämmelse. Konsensusmåtten tycks sammanfattningsvis visa att samstämmigheten i bedömningarna i denna studie nära nog når samma nivåer som eftersträvas i storskaliga internationella mätningar, och den är vidare jämförbar eller i vissa fall högre än de nivåer som uppmätts för andra nationella prov i svenska och matematik. Inomklasskorrelation beräknades för totalpoängsumman av de bedömda uppgifterna. Den uppmätta nivån 0,92 indikerar nästan perfekt överensstämmelse.

Utifrån den bayesianska metoden att dra slutsatser om bedömaröverensstämmelse för populationen, det vill säga alla lärare som genomför provet, kan konstateras att samstämmigheten i bedömningar av elevsvar på det specifika läsförståelseprovet med 95 procents sannolikhet befinner sig mellan 0,71 och 0,84 för Cohens κ och mellan 0,72 och 0,95 för inomklasskorrelationen. Den bedömaröverensstämmelse som skulle kunna bli utfallet vid det faktiska provgenomförandet enligt dessa skattningar befinner sig, i relation till riktlinjerna, alltså inom spannet för påtaglig överensstämmelse.

Inom ramen för en mångfasetterad Rasch-analys gjordes en närmare analys av skillnaderna i stränghet mellan bedömarna liksom av deras benägenhet att göra säregna bedömningar. Skillnaderna i stränghet mellan bedömarna förefaller något mindre än resultaten från tidigare studier av liknande prov, och påfallande små i jämförelse med resultat från studier som behandlat bedömningar av komplexa uppgifter. Den genomsnittliga bedömningen skilde 6 poäng mellan den mildaste och den strängaste bedömaren. Separationsindex var 2,40 vilket visar på att skillnaden mellan den mildaste och den strängaste bedömaren är relativt liten. Det faktum att den mångfasetterade Rasch-modellen är statistiskt välanpassad till bedömarnas bedömningar visar dessutom att de inte gjort vare sig säregna eller oförutsägbara bedömningar.

Även analyserna utifrån den generaliserbarhetsteoretiska ansatsen visar på mycket små variationer mellan bedömarna, både vad gäller stränghet och rangordning av elevlösningarna. Generaliserbarhetskoefficienten, det vill säga provets totala reliabilitet utifrån den här studien, skattades till 0,79. En slutsats som kan dras utifrån dessa analyser är emellertid att ett utökat antal provuppgifter torde höja reliabiliteten i elevernas resultat. Ett utökat antal provuppgifter vore dessutom att föredra framför att öka antalet bedömare, eftersom det troligen slukar mer resurser att kräva minst två bedömare per elevlösning än att en bedömare bedömer fler uppgifter. I kommande utvecklingsarbete planerar provgruppen för en utökning av antalet uppgifter, främst av flervalstyp. Detta medför att generaliserbarhetskomponenten sannolikt kommer att kunna nå över 0,8, med viss marginal.

Liksom vid andra studier av den här typen behöver studiens begränsningar beaktas vid tolkning av resultatet. När det aktuella provet genomfördes var det obligatoriskt på samtliga gymnasieprogram. Sedan 2018 är provet endast obligatoriskt på de yrkesförberedande programmen. Det är en omständighet som tyvärr försvårar tolkningen av resultatet. Undersökningen är gjord utifrån elevlösningar i ämnet svenska och kan visa sig få ett annat utfall om den skulle genomföras utifrån elevlösningar i ämnet svenska som andraspråk. Mer generella omständigheter som ofta gäller för denna typ av undersökningar rör frivilligheten i deltagande, som kan ha fått konsekvenser för hur representativt urvalet kommit att bli i relation till populationen lärare. Det är visserligen möjligt att bedömningarna i den grupp av lärare som inte anmält intresse skulle likna de som samlats in från lärarna som efter anmält intresse valdes ut, men vi saknar helt enkelt bevis för att uttala oss om eventuella likheter eller skillnader.

Resultatens räckvidd hade dock fortfarande varit begränsad i andra avseenden. Det saknas nämligen fortfarande studier av svenska nationella prov som fullt ut kan beskriva skillnader mellan att göra bedömningar av okända elevers lösningar som därtill saknar konsekvens för dessa okända elever (såsom i denna studie) och att göra bedömningar av elevlösningar från den egna klassen där det slutgiltiga omdömet får konsekvenser för eleverna. Detta slags hinder för generaliseringar av resultaten till faktiska förhållanden för genomförandet av nationella prov låter sig inte undanröjas genom mer representativa urval av bedömare, utan vi får nöja oss med approximationer.

Dessa begränsningar av resultatens räckvidd delar rapporten med de flesta undersökningar av samstämmighet. Däremot ger försöket att dra sannolikhetsbaserade slutsatser ett unikt bidrag till studier av samstämmighet i bedömningar av elevprestationer. Detta statistiska ramverk ger nämligen konkreta verktyg för att inte bara göra utsagor om hur stor samstämmigheten kan vara i andra, ännu ogenomförda studier med samma prov och målpopulation, utan också med vilken statistisk säkerhet dessa utsagor görs. Inom klassisk frekventistisk statistik, där man endast kan uttala sig om sannolikheten för att en specifik nivå av samstämmighet har uppnåtts, givet att den faktiska nivån är 0 (eller annan specificerad nollhypotes), är det alltså inte möjligt att beskriva hur säker man är på att samstämmigheten befinner sig på en viss nivå.

Avslutningsvis kan ett par slutsatser dras utifrån studien med avseende på hur förutsättningar för reliabla bedömningar kan optimeras. Det svar som rapporten kan ge, och som diskussionen ovan antytt, är att bedömningar av elevproducerade svar på uppgifter med öppna svarsformat kan göras tillförlitligt av en bedömare. Vi vet från lärarenkäter att sambedömning inte är särskilt utbrett i bedömning av läsförståelseprov. Som också har nämnts tidigare i rapporten skulle det dessutom med största sannolikhet vara mer kostnadseffektivt att utöka antalet uppgifter än att rekommendera sambedömning i syfte att rent allmänt höja reliabiliteten i läsförståelseprovet. Sedan återstår förstås frågan om i vilken utsträckning formuleringar i bedömningsanvisningarna liksom variationer i hur strikt olika lärare följer dessa påverkar graden av samstämmighet. Detta ligger emellertid utanför rapportens undersökningar.

Sammanfattningsvis pekar rapportens resultat på att samstämmigheten överlag når klart tillfredsställande nivåer. Det gäller dels sådana nivåer av samstämmighet som finns uttryckta i *Skolverkets systemramverk för nationella prov* (2017), dels nivåer som anges i storskaliga internationella mätningar som PISA-undersökningarna. Skillnaderna mellan milda och stränga bedömare är relativt små och det är möjligt att med ett ökat inslag av flervalsuppgifter uppnå mycket god sammantagen reliabilitet utifrån generaliserbarhetskoefficienten. Att bedömningarna för ett läsförståelseprov där slumpmässigt utvalda bedömare som inte tränats särskilt på uppgiften tycks nå nästan lika hög genomsnittlig överensstämmelse som de tränade bedömarpanelerna i PISA är anmärkningsvärt. En sista slutsats som är viktig att lyfta fram utifrån den här studien är lärares professionalitet i att tolka de givna anvisningarna till den här typen av läsförståelseprov. Studien pekar på att lärare därmed också kan förvänta sig relativt hög tillförlitlighet i bedömarhänseende vid tolkning av enskilda elevers provresultat på läsförståelseprov inom det nationella provsystemet i Sverige.

Referenser

Litteratur

Bachman, Lyle F., 2004: *Statistical analyses for language assessment*. Cambridge: Cambridge University Press.

Brennan, Robert L., 2001: *Generalizability theory*. New York: Springer.

Broemeling, Lyle D., 2009: *Bayesian methods for measures of agreement*. Boca Raton: CRC Press.

Broemeling, Lyle D., 2012: *Advanced Bayesian methods for medical test accuracy*. Boca Raton: CRC Press.

Cohen, Jacob, 1960: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:1. S. 37–46.

Cohen, Jacob, 1968: Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70:4. S. 213–220.

Dalberg, Tobias, 2019: *Samstämmighet i skrivbedömning. Statistisk analys vid bedömning av två nationella skrivprov*. Svenska i utveckling 36. Uppsala: Uppsala universitet.

de Santi, R. J., & Sullivan, V. G., 1984: Inter-rater reliability of the Cloze Reading Inventory as a qualitative measure of reading comprehension. *Reading Psychology*, 5(3–4), 203–208.

DeVellis, Robert F., 1991: *Applied social research methods series, Vol. 26. Scale development: Theory and applications*. Thousand Oaks, CA, US: Sage Publications, Inc.

Eckes, Thomas, 2015: *Introduction to many-facet Rasch measurement: analyzing and evaluating rater-mediated assessments*. Andra reviderade och uppdaterade upplagan. Frankfurt am Main: Peter Lang.

Fleiss, Joseph L., 1971: Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76:5. S. 378–382.

Fleiss, Joseph L., 2003: *Statistical methods for rates and proportions*. Tredje upplagan. Hoboken, N.J.: Wiley-Interscience.

Kane, Michael, Crooks, Terence & Cohen, Allan, 1999: Validating Measures of Performance. *Educational Measurement: Issues and Practice 18*(2), S. 5–17. https://doi.org/10.1111/j.1745-3992.1999.tb00010.x

Koo, Terry K, & Li, Mae Y., 2016: A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of chiropractic medicine*, 15:2. S. 155–63.

Lakes, K. & Hoyt, W., 2009: Applications of Generalizability Theory to Clinical Child and Adolescent Psychology Research. *Journal of Clinical Child & Adolescent Psychology*, Vol. 38, nr 1, 144–165.

Landis, J. Richard, & Koch, Gary G., 1977: The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33:1. S. 159–174.

Linacre, John M., 1989: *Many-Facet Rasch Measurement*. Chicago: MESA Press.

Lind Pantzare, Anna, 2015: Interrater reliability in large-scale assessments – Can teachers score national tests reliably without external controls?, *Practical Assessment, Research, and Evaluation*: Vol. 20, Article 9.
Martin, M. O., Mullis, I. V. S., & Hooper, M. (red.), 2017: Methods and Procedures in PIRLS 2016.
McGraw, Kenneth O. & Wong, S. P., 1996: Forming Inferences About Some Intraclass Correlation Coefficients. *Psychological Methods*, 1:1. S. 30–46.
OECD 2005: *PISA 2003. Technical Report*. OECD Publishing.
OECD 2009: *PISA 2009. Assessment Framework. Key competencies in reading, mathematics and science*. OECD Publishing.
OECD 2015: *PISA 2015. Technical Report, PISA*. OECD Publishing.
Portney, Leslie Gross. & Watkins, Mary P., 2015: *Foundations of clinical research: applications to practice*. Tredje reviderade upplagan. Philadelphia, PA.: F.A. Davis Company.
Skolverket, 2017: *Skolverkets systemramverk för nationella prov*. Stockholm: Skolverket.
Tengberg, Michael, & Skar, Gustaf B., 2016: Samstämmighet i lärares bedömning av nationella prov i läsförståelse. *Nordic Journal of Literacy Research*, 2(1).
Tengberg, Michael, Roe, Astrid, Skar, Gustaf B., 2018: Interrater reliability of constructed response items in standardized tests of reading. *Nordic Studies in Education*. 38: 2. S. 118–137.
Uebersax, John S., 1987: Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101(1), 140–146. https://doi.org/ 10.1037/0033-2909.101.1.140
Verhelst, Norman D., 2004: Section E: Generalizability Theory. I: *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment*, red. av Sauli Takala. Strasbourg, France: Council of Europe/Language Policy Division.
Wright, Benjamin D., & Masters, Geofferey N., 1982: Rating scale analysis. Chicago: Mesa Press.

Programvara

Gamer, Matthias, Lemon, Jim & Singh, Ian Fellows Puspendra, 2019: *irr: Various Coefficients of Interrater Reliability and Agreement*. R package version 0.84.1. https://CRAN.R-project.org/package=irr
Linacre, John M., 2018: *Facets computer program for many-facet Rasch measurement*, version 3.80.4. Beaverton, Oregon: Winsteps.com
Lunn, D.; Spiegelhalter, D.; Thomas, A.; Best, N., 2009: The BUGS project: Evolution, critique and future directions. *Statistics in Medicine*. 28 (25): 3049–3067.
Mizumoto, Atsushi, 2015: Langtest (Version 1.0) *Generalizability Theory*. Hämtad från http://langtest.jp [2018-06-26]

Tabellbilagor

Tabell 4. Skattningar av överensstämmelse med bayesiansk metod: central- och spridningsmått.

Mått	Medelvärde	Standardavvikelse	Median	Bayesianskt 95 % konfidensintervall
Absolut överensstämmelse	0,833	0,028	0,835	0,776–0,884
Slumpmässig överensstämmelse	0,243	0,039	0,240	0,175–0,326
Fleiss κ	0,780	0,035	0,781	0,707–0,845

Fotnot: Alla värden hänvisar till posteriorifördelningen. Den är i sin tur är baserad på de 94 000 simuleringar som återstår efter att vi bortsett från de 1 000 första simuleringarna. Eftersom antalet simuleringar är relativt stort blir beräknings-precisionen hög, vilket indikeras av att Monte Carlo-mätfelen är långt mindre än 0,001.

Tabell 5. Konsensusestimat. Exakt (procentuell) samstämmighet och Cohens κ för alla parvisa jämförelser av bedömarnas bedömningar.

Bedömarpar	Antal	Exakt (%)	Cohens κ
1/2	180	81,1	0,59
1/3	180	86,1	0,70
1/4	180	85,0	0,68
1/5	180	83,3	0,64
1/6	180	81,1	0,58
1/7	180	85,6	0,69
1/8	180	81,7	0,60
1/9	180	88,9	0,77
2/3	180	91,7	0,80
2/4	180	93,9	0,85
2/5	180	92,2	0,81
2/6	180	92,2	0,80
2/7	180	90,0	0,76
2/8	180	91,7	0,79
2/9	180	87,8	0,72
3/4	180	93,3	0,84
3/5	180	91,7	0,80
3/6	180	89,4	0,74
3/7	180	92,8	0,83
3/8	180	92,2	0,81
3/9	180	91,7	0,81
4/5	180	92,8	0,82
4/6	180	93,9	0,85

4/7	180	90,6	0,78
4/8	180	92,2	0,81
4/9	180	92,8	0,84
5/6	180	92,2	0,80
5/7	180	93,3	0,84
5/8	180	90,6	0,76
5/9	180	91,1	0,80
6/7	180	90,0	0,75
6/8	180	92,8	0,81
6/9	180	87,8	0,72
7/8	180	91,7	0,80
7/9	180	90,0	0,78
8/9	180	88,3	0,73

Rapportserien Svenska i utveckling

1. Margareta Andersson, 1995: Prov i modersmålet. En översikt över centralt utarbetade prov i Sverige och nio andra länder. (= FUMS Rapport nr 176.) 40 s.
2. Catharina Nyström, 1996: Skrivandet, kursplanerna och läromedlen. En studie av gymnasieskolans läromedel i svenska 1970–1995. (= FUMS Rapport nr 178.) 52 s.
3. Birgitta Garme, 1996: Att pröva skrivförmågan med nationella prov. En presentation av proven i svenska för skolår 2 och 5. (= FUMS Rapport nr 180.) 55 s.
4. Birgitta Garme & Anne Palmér, 1996: Samspel och Kommunikation – om utvecklandet av nationella kursprov i svenska för gymnasieskolan. (= FUMS Rapport nr 182.) 73 s.
5. Orla Vigsø, 1996: Valgplakaten som kommunikation og marketing. (= FUMS Rapport nr 183.) 56 s.
6. Eva Östlund-Stjärnegårdh, 1997: Skriva debattartiklar i skolan – går det? Om texttypen debattartikel i nationella prov. (= FUMS Rapport nr 186.) 40 s.
7. Hanna Sofia Öberg, 1997: Referensbindning i elevuppsatser. En preliminär modell och en analys i två delar. (= FUMS Rapport nr 187.) 109 s.
8. Björn Melander, 1998: Inskickat och registrerat – sammanställning av uppgifter rörande de nationella proven i svenska vår- och höstterminen 1996 samt vårterminen 1997. (= FUMS Rapport nr 188.) 53 s.
9. Annika Persson, 1998: Texten och inspirationskällan. En studie av förlagans betydelse för elevers fria textproduktion i skolår 5. (= FUMS Rapport nr 191.) 42 s.
10. Anne Palmér, 1999: Tankar om tal – lärares och elevers syn på muntlig framställning i undervisning och bedömning. (= FUMS Rapport nr 193.) 57 s.
11. Helena Olevard, 1999: ”Tonårsliv”. En pilotstudie av 60 elevtexter från standardproven för skolår 9 åren 1987 och 1996. (= FUMS Rapport nr 194.) 28 s.
12. Eva Östlund-Stjärnegårdh, 1999: Principen och praktiken. En enkätundersökning av lärares syn på bedömning av gymnasieelevers texter. (= FUMS Rapport nr 195.) 40 s.
13. Catharina Nyström & Maria Ohlsson (red.), 1999: Svenska på prov. Arton artiklar om språk, litteratur, didaktik och prov. (= FUMS Rapport nr 196.) 150 s.
14. Catharina Nyström, 2000: Ledfamiljer och referentrelationer. En modell för analys av referensbindning tillämpad på gymnasisttexter. (= FUMS Rapport nr 197.) 59 s.
15. Kerstin Lagrell, 2000: Växa i skrivandet. Om förhållningssättet till de yngre skolbarnens skrivutveckling. (= FUMS Rapport nr 199.) 70 s.
16. Maria Ohlsson, 2001: Säker stil eller rätt i sak? En studie i gymnasisters analyser av en enkät. (= FUMS Rapport nr 204.) 45 s.
17. Anne Palmér, 2002: Röster i samspel. Analys av gymnasieelevers radioprogram. (= FUMS Rapport nr 205.) 59 s.
18. Helena Andersson, 2002: Svenska som första- och andraspråk – en jämförande studie av texter från skolår 9. (= FUMS Rapport nr 208.) 97 s.
19. Catharina Nyström, 2003: Argumentera! En presentation av gymnasisters texter i databasen ARGUS. (= FUMS Rapport nr 209.) 61 s.
20. Katharina Hallencreutz, 2003: Särskrivningar och andra skrivningar i elevspråk. (= FUMS Rapport nr 210.) 101 s.
21. Maria Eklund Heinonen, 2005: Godkänd eller underkänd? Hur processbarhetsteorin kan tillämpas vid muntliga språktester av andraspråksinlärare. (= FUMS Rapport nr 215.) 59 s.

22. Inger Gröning, 2006: Interaktion och lärande i flerspråkiga klasser. (= FUMS Rapport nr 218.) 89 s.
23. Eva Östlund-Stjärnegårdh, 2006: Att förmedla egna och andras tankar. Om gymnasisters källhantering i det nationella provets skrivuppgift. (= FUMS Rapport nr 219.) 52 s.
24. Karin Wesslén, 2008: Processkrivande – en etablerad metod på gymnasiet? (= FUMS Rapport nr 225.) 55 s.
25. Ciolek Laerum, Beatrice, 2009: Elever skriver och lärare bedömer – en studie av elevtexter i åk 9. (= FUMS Rapport nr 226.) 60 s.
26. Palmér, Anne, 2010: Att bedöma det muntliga. Utvärdering av ett delprov i gymnasieskolans nationella kursprov, Svenska B. (= FUMS Rapport nr 227.) 61 s.
27. Nyström Höög, Catharina, 2010: Mot ökad diskursivitet? Skrivutveckling speglad i provtexter från årskurs 5 och årskurs 9. (= FUMS Rapport nr 228.) 62 s.
28. Hagberg-Persson, Barbro, Berg, Elisabeth & Lagrell, Kerstin, 2010: Ämnesprov i svenska och svenska som andraspråk för årskurs 3 – en utprövningsomgång. (= FUMS Rapport nr 229.) 89 s.
29. Nordberg, Olle, 2013: Att finnas till som läsare – skönlitterär läsning i ett elevperspektiv. Didaktiska tillämpningar av en empirisk studie baserad på elevers egna texter om sin läsning. 91 s.
30. Hagberg-Persson, Barbro & Wiberg, Cecilia, 2013: Elever visar vad de kan. Två studier kring ämnesprovet i svenska och svenska som andraspråk för årskurs 3. 61 s.
31. Palmér, Anne, 2013: Nationella skrivprov baserade på två olika läroplaner. Genre, kommunikationssituation och skrivdidaktiska diskurser. 72 s.
32. Schüssler, Eija L., 2014: Visuell kommunikation – en modell för bedömning av gymnasieelevers bruk av bildspel vid muntliga anföranden. 46 s.
33. Hagberg-Persson, Barbro, 2015: Nationellt prov i årskurs 3 – en redovisning av ämnesprovet i svenska och svenska som andraspråk. 53 s.
34. Mark, Malin & Palmér, Anne, 2017: En utvecklande tolkningsgemenskap? Matrisanvändning, interaktion och kontext i bedömningssamtal om ett nationellt prov i muntlig framställning. 91 s.
35. Nordberg, Olle, 2019: Berättelser som förändrar. Utvärdering och didaktisk diskussion kring ett nationellt läsprojekt för ungdomar. 161 s.
36. Dalberg, Tobias, 2019: Samstämmighet i skrivbedömning. Statistisk analys vid bedömning av två nationella skrivprov. 47 s.
37. Dalberg, Tobias, Zachiu, Martina, Shahsavar, Negin, Eriksson, Kristina & Hussenius, Siri, 2020: Samstämmighet i läsbedömning. Statistisk analys vid bedömning av ett nationellt läsförståelseprov. 47 s.

Beställning

För beställning av rapporter i serien Svenska i utveckling, kontakta någon av provgruppens administratörer. Kontaktuppgifter finns på följande webbplats: **natprov.nordiska.uu.se**. Där finns även vissa av rapporterna nätpublicerade.