Nordic Journal of Literacy Research

Peer-reviewed article

Samstämmighet i lärares bedömning av nationella prov i läsförståelse

Michael Tengberg[1]*, and Gustaf B. Skar[2]
[1]*Karlstads universitet;* [2]*Norges teknisk-naturvitenskaplige universitet*

Abstrakt
Tillförlitlighet i bedömning är en avgörande komponent i varje testprogram där testtagares resultat bygger på bedömares tolkningar utifrån en bedömningsskala eller en bedömningsguide. Utförliga svar på öppna uppgifter bedöms exempelvis sällan som antingen ”rätt” eller ”fel”. Istället tillämpas skalan eller bedömningsguiden för att fastställa i vilken utsträckning svaret uppvisar den efterfrågade kompetensen. I den här artikeln redovisas resultat från en studie av bedömarreliabilitet på öppna uppgifter i det nationella provets svenska läsförståelsedel i årskurs nio.

För att undersöka i vilken mån provsystemet skapar förutsättningar för god bedömarreliabilitet har sex lärare fått bedöma tre elevers lösningar av 14 öppna uppgifter, totalt 252 bedömningar. Analyserna innefattar konsensusestimat (procentuell samstämmighet och Cohens kappa) och konsistensestimat (ICC). Dessutom har kvalitativa analyser genomförts på uppgiftsnivå för att visa på aspekter i uppgiftskonstruktionen som kan ligga till grund för låg bedömarreliabilitet.

Resultaten från studien visar på moderata nivåer av bedömarreliabilitet, både ifråga om kappavärden (.73) och ICC (.82), vilket motsvarar en variation mellan bedömningarna som får stora konsekvenser för elevernas slutgiltiga provresultat. I artikeln diskuterar vi resultatens implikationer för rättvis bedömning av elevers läsförmåga i Sverige. Vi för också ett resonemang om olika sätt att stärka bedömarreliabiliteten det nationella provet i läsförståelse.

Nyckelord: *bedömning*; *interbedömarreliabilitet*; *läsning*; *nationella prov*; *reliabilitet*

Abstract
Inter-rater reliability is a critical component in any test program where test-takers’ responses are judged by human raters using scales or scoring rubrics. Lengthy responses to open-ended test items are, for instance, rarely judged objectively as either “correct” or “incorrect”. Rather, rubrics are used to determine the extent to which a particular item response displays the expected competence. This paper reports a study of inter-rater reliability in teachers’ assessment of open-ended items in the Swedish national reading test for 9th grade.

In order to explore whether the test design supports reliable assessment, six experienced teachers of Swedish were asked to rate the responses of three students on 14 items, 252 ratings in all. Analyses included consensus estimates (percent agreement and Cohen’s kappa) and consistency estimates (ICC). In addition, qualitative item analyses were performed in order to investigate possible causes of low reliability for specific items.

Findings indicate moderate levels of inter-rater reliability according to both kappa (.73) and ICC (.82) values, equaling a variation of ratings with large consequences for the students’ final results.

*Correspondence to: Michael Tengberg, Institutionen för pedagogiska studier, Karlstads universitet, 65188 Karlstad, Sverige. E-mail: michael.tengberg@kau.se

Citation: Michael Tengberg and Gustaf B. Skar. ”Samstämmighet i lärares bedömning av nationella prov i läsförståelse.” Nordic Journal of Literacy Research, Vol. 2, 2016, pp. 1–18. http://dx.doi.org/10.17585/njlr.v2.230

Implications for equal assessment of students' reading ability in Sweden are discussed, as well as some suggestions for necessary future development of the national reading test.

Keywords: *assessment*; *inter-rater reliability*; *national test*; *reading*; *reliability*

Received: November 2015; Accepted: February 2016; Published: April 2016

Tillförlitlighet i bedömning är en avgörande komponent i alla slags testprogram där testtagares resultat bygger på bedömares tolkningar av elevsvaren utifrån en bedömningsskala eller en bedömningsguide. Exempelvis kan en muntlig redovisning eller ett flera rader långt skriftligt svar på en läsprovsuppgift sällan bedömas objektivt som antingen ”rätt” eller ”fel”. Istället måste bedömare tillämpa bedömningsguiden för att fastställa i vilken utsträckning uppgiftslösningen uppvisar den efterfrågade kompetensen. Detta förutsätter att bedömningarna som förekommer inom testprogrammet är reliabla, dvs. att de är både konsistenta och fria från olika former bedömareffekter (Haladyna & Rodriguez, 2013).

Interbedömarreliabilitet avser graden av samstämmighet mellan två eller flera bedömare som bedömer eller analyserar samma objekt. Låg interbedömarreliabilitet kan bero på slumpmässiga felkällor, som inkonsistens, eller mer systematiska felkällor, bias, såsom stränghet/generositet och haloeffekter (se t.ex. Engelhardt, 2002; Meadows & Billington, 2005; Myford & Wolfe, 2003).

I den här artikeln diskuterar vi samstämmighet i bedömning av elevers läsförmåga. I fokus för undersökningen står frågan om lärares bedömning av elevers svar på läsprovsuppgifter i det nationella ämnesprovet i svenska i årskurs nio. Eftersom detta prov har en betydande inverkan på elevernas avgångsbetyg i ämnet och därmed för deras framtid så bör provet klassificeras som *high stakes* (Bachman, 2005). Det innebär att det blir extra viktigt att bedömningarna är rättvisande i så mening att elevens slutresultat inte är beroende av vem som bedömer.

Studier av interbedömarreliabilitet av läsförmåga är ett begränsat forskningsområde, men några undersökningar har visat att det går att uppnå höga nivåer av samstämmighet i standardiserade provsystem (DeSanti & Sullivan 1984). Däremot rapporterar Illinois State Board of Education (2013) att vid genomförande av SAT-testerna[1] i läsning 2013 låg nivån för exakt överensstämmelse på moderata 65–67 % vid bedömning av ”extended-response items”, där bedömarna bedömde varje elevsvar på en skala från 0–4. Annan forskning har visat att även efter att bedömare, i det här fallet av fria responsuppgifter på romantext, tränats och kalibrerat bedömningsnivåer till en given uppgift och en bedömningsguide så förekommer hos en andel av bedömarna en statistiskt signifikant glidning[2] både i fråga om sättet att använda bedömningsskalan och i fråga om precision (Myford & Wolfe, 2009). Möjligheten till samstämmighet är dock i stor utsträckning beroende av uppgiftskonstruktionen och därför måste varje provsystem slå vakt om att uppgifter som

[1]SAT (Standard Achievement Test).

[2]Den engelska termen är DRIFT (differential rater functioning over time).

kräver subjektiv bedömning håller en hög grad av interbedömarreliabilitet (Bejar, 2012). I system där man använder en utvald panel med bedömare kan detta kontrolleras både genom kontinuerlig kvalificering (träning och urval) av bedömarna och genom successiv anpassning av uppgiftsurvalet. I system som det svenska nationella provsystemet, där alla ämneslärare i landet deltar som bedömare, blir kvalificering av bedömarna en mer omfattande procedur och kravet på uppgifter som kan bedömas likvärdigt blir högre.

Naturligtvis kan man i princip utesluta bedömarvariation som felkälla genom att endast inkludera uppgifter som kan bedömas objektivt som antingen ”rätt” eller ”fel”. Dessa uppgifter får ofta multiple choice-format (MC) och kan rättas antingen maskinellt eller av lärare. Så har man exempelvis utformat det danska avgångsprovet i läsning i årskurs nio (Tengberg, in press) och det har också varit en internationell standard för utformning av stora läsprov (Alderson, 2000; Campbell, 2005). Det finns flera provtekniska fördelar med MC-formatet: fler uppgifter kan ingå i provet eftersom de går fort att besvara och skrivförmågan blir inte en begränsande faktor för uppvisandet av läsförmåga (Campbell, 2005; Roe & Lie, 2009). Men när man bestämt sig för att pröva i vilken mån elever exempelvis kan motivera sina tolkningar, relatera olika texter till varandra eller relatera det lästa till sina tidigare erfarenheter så är det ofta lämpligare att låta eleverna formulera dessa svar på egen hand, dvs. genom constructed response-uppgifter (CR) (Campbell, 2005; Kane, Crooks & Cohen, 1999; Pearson & Hamm, 2005). Det gör att frågan om samstämmighet i bedömningen blir ofrånkomlig.

I det svenska bedömningssystemet ansvarar ämnesläraren både för betygsättning i ämnet och för bedömning av elevernas nationella prov. På många skolor förekommer olika former av gemensam rättning eller utbyten av elevlösningar för att den enskilde läraren inte ska rätta sina egna elever. Detta kan vara ett sätt att undvika den sortens bias som ofta uppstår när bedömaren känner den som ska bedömas (Jacob & Levitt, 2003). Däremot avhjälper inte sådana system en eventuellt bristande samstämmighet mellan hur lärarna värderar elevers svar på olika uppgifter. Hur det är ställt med samstämmigheten i svensklärares bedömning av elevers läsförmåga vet vi emellertid lite om eftersom några sådana studier inte genomförts och eftersom provkonstruktörerna själva i liten utsträckning mätt interbedömarreliabiliteten i det nationella läsprovet. I en undersökning, publicerad av Skolverket (2009), redovisas bedömaröverensstämmelse mellan ordinarie rättande lärare och tre övriga bedömare. Dessa övriga bedömare var en provkonstruktör, en lärare som deltagit i utprövning av provuppgifter samt en ytterligare lärare med skolerfarenhet. Överensstämmelsen mättes bland annat genom betygsättning av 100 elevlösningar på en tiogradig skala (här ingick inte de ordinarie rättande lärarna). Korrelationen uppmättes till mellan .80 och .86, men eftersom samstämmighet gäller det sammanlagda provbetyget kan resultatet dölja en stor variation på enskilda uppgifter. Skolverket noterar att man funnit särskilt låg överensstämmelse för uppgifter ”vilka kräver utförligare svar, så som förklaringar, exempel eller flera aspekter (till exempel både förklaring och motivering)” (Skolverket, 2009, s. 23), även om statistiken inte redovisas. Studien är intressant eftersom den implicerar att det är möjligt att uppnå en generellt hög

överensstämmelse mellan olika bedömare. Däremot säger den inte mycket om samstämmigheten mellan vanliga lärares bedömningar inom ramen för provsystemet. Den visar heller inte vilken konsekvens differensen mellan två olika bedömares bedömningar riskerar att få för den enskilde eleven på det givna provet. I föreliggande studie använder vi ett mindre empiriskt material för att följa upp de här frågorna och undersöker samstämmigheten i svensklärares bedömningar av elevers läsförmåga så som den uttrycks i nationella provuppgifter.

Syfte och frågeställningar

Syftet med studien är att ta reda på i vilken mån det svenska nationella läsprovet möjliggör tillförlitliga bedömningar av niondeklassares läsförmåga. Detta undersöks genom att jämföra bedömningar från flera lärare av samma elevlösningar. Mer specifikt vill vi ta reda på

- i vilken utsträckning lärarna bedömer enskilda constructed response-uppgifter samstämmigt, och
- hur elevernas resultat varierar med vem som bedömer deras provlösningar.

Constructed response-uppgifter i läsförståelsetester

Även om multiple choice-uppgifter är både kostnadseffektiva och lättadministrerade så har constructed response-uppgifter under de senaste decennierna blivit allt vanligare i storskaliga tester av läsförmåga och det är numer brukligt med så kallade mixed-models i vilka de två uppgiftsformaten blandas (Pearson, Calfee, Walker-Webb, & Fleischer, 2002; Solheim & Skaftun, 2009). I både PISA och PIRLS utgörs ca hälften av uppgifterna av CR-formatet (IEA, 2009; OECD, 2013). Vanligast är de för uppgifter som avser att pröva elevernas förmåga att tolka eller att reflektera över språk och innehåll i texter, uppgifter som antas pröva djupare förståelse.

Användandet av CR-uppgifter medför dels att uppgifter måste bedömas av någon med ämneskompetens, dels att svar inte helt och hållet kan förutses och att somliga svar måste graderas kvalitativt som mer eller mindre korrekta eller fullständiga. För detta ändamål används bedömningsguider med noggranna anvisningar om hur olika elevsvar ska bedömas. Att formulera bedömningsguider innebär emellertid en avvägning mellan att å ena sidan begränsa tolkningsutrymmet i uppgiften i syfte att undvika irrelevant bedömaravvikelse och å andra sidan lämna tillräckligt stort utrymme för att alla rimliga och oförutsedda, men likväl goda elevtolkningar ska kunna ge poäng. Detta är alltså en fråga om semantisk öppenhet som dels måste förhandlas för varje enskild uppgift, dels måste förankras i kursplanens kunskapskrav liksom i relevant teori om läsförståelse och i etablerade ämnestraditioner (Solheim & Skaftun, 2009).

I många läsprov, inklusive det svenska nationella provet, bedöms CR-uppgifter på en skala i två eller flera steg, dvs. att en elev kan få två eller flera poäng för en uppgift

som kräver ett mer utförligt svar. Detta möjliggör att även icke fullständiga svar kan ge poäng, men skapar samtidigt problem i förhållande till övriga uppgifter i provet. Uppgifter som genererar fler poäng implicerar att de kräver mer kvalificerade läsningar, men ofta är det snarare en additiv logik som bygger upp dessa uppgifter, dvs. att uppgifterna består av olika delar där en elev exempelvis ska lämna *två* referenser från texten, inte *en*, för att få full poäng. Full poäng ges alltså när svaret är fullständigt snarare än mer kvalificerat i ett textförståelseperspektiv. Solheim & Skaftun (2009) kallar detta fenomen för *semantisk atomism* och har påvisat problemet i PIRLS-provet. Tengberg (2014) har visat att samma problem finns i de svenska nationella läsproven.

Att mäta bedömarreliabilitet

Låg bedömarreliabilitet kan ha olika orsaker. En orsak är att bedömarna gjort olika tolkningar av bedömningsanvisningarna, dvs. kriterierna. Andra vanliga källor till irrelevant bedömarvariation är grad av stränghet, haloeffekt, centraltendens och omfångsrestriktion (Engelhard, 1994; Van Moere, 2014). Stränghetsvariation innebär att bedömare systematiskt skiljer sig i fråga om andelen höga respektive låga omdömen. Haloeffekter innebär att bedömningen av en aspekt av ett svar inverkar på bedömningen av andra, konceptuellt åtskilda aspekter. Centraltendens betyder att bedömaren främst använder mittdelen av bedömningsskalan. Omfångsrestriktion betyder att en bedömare använder en begränsad del av skalan, och därmed inte diskriminerar tillräckligt mellan testtagare (se också Myford & Wolfe, 2003).

Bedömarreliabilitet kan undersökas på olika sätt. Stemler och Tsai (2008) nämner tre typer av undersökningar som är vanliga. Den första typen används för att mäta om bedömare rangordnar elevgensvar likadant. Måtten på detta kallas för *konsistensestimat* och inkluderar Pearsons *r*, Spearmans rho (ρ), Cronbachs alpha (α) och Intra Class Correlation (ICC). Den andra typen av undersökning mäter i vilken utsträckning bedömare når samma slutsats om elevers absoluta förmåga, ofta uttryckt som betygssteg eller verbala deskriptorer av förmåga. Måtten på detta kallas följdriktigt för *konsensusestimat* och inkluderar procentuell fullständig och angränsande samstämmighet samt Cohens viktade och oviktade kappa (κ). Den tredje typen av undersökning är främst aktuell i provutvecklingssammanhang och kallas av Stemler och Tsai för ”mätestimat” (*measurement estimates*). Vanliga mått här baseras på Many-Facet Rasch Measurement, faktoranalyser och Generalizability Theory. I regel används denna typ av undersökningar när en provutvecklare (eller -användare) vill inkorporera all tillgänglig information och också studera eventuella interaktioner mellan bedömare, uppgifter och testtagare. Vi kommer inte att tala mer om mätestimat i denna artikel.

Eftersom konstistens- och konsensusmått bidrar med olika typer av information är det lämpligt att genomföra båda typer av mätningar (Stemler, 2004). Det är till exempel fullt möjligt att erhålla höga kappavärden, utan att motsvarande höga konstistensmått uppnås. Detta beror på de matematiska antaganden som ligger bakom en del konsistensmått. På motsvarande sätt är det möjligt att nå höga värden

vid konstistensundersökningar, utan att för den skull erhålla höga konsensusmått. Det beror på att bedömare systematiskt kan skilja sig åt i absolut stränghet, men ändå rangordna elevgensvar likadant.

Med tiden har det etablerats så kallade benchmarks, eller riktvärden, för vad som räknas som acceptabla nivåer av samstämmighet. För konsistensindikatorerna, dvs. korrelationskoefficienter (som antar ett värde i intervallet -1.0 till 1.0), anges vanligen att dessa bör överstiga .70, eftersom ".70 represents a rock-bottom minimum of acceptable agreement among raters" (McNamara, 2000, s. 58). Vilka nivåer som är eftersträvansvärda är emellertid starkt kontextavhängigt; när provresultatet får stora konsekvenser är behovet av att minimera bedömareffekter stort och det är inte ovanligt att ha ett korrelationsvärde om .90 som krav.

För konsensusindikatorerna finns flera benchmarks (Gwet, 2014). Vanligen anges att värden omkring .60–.80 är acceptabla och att värden från .80 och uppåt är goda. Även här styr konsekvenserna av provanvändningen krav på miniminivåer.

Tidigare forskning om samstämmighet relaterat till CR-uppgifter tar sig ofta ett av två uttryck (jfr Meadows & Billington, 2005): endera rapporteras i nivåer av samstämmighet som är möjliga att uppnå, givet en viss utformning av kriterier, en viss bedömarträning och så vidare (t.ex. Congdon & McQueen, 2000; Lim, 2011; Moser, Sudweeks, Morrison, & Wilcox, 2014), eller så fokuseras metoder för att identifiera och eller åtgärda bristande samstämmighet (t.ex. Attali, 2014; Güler, 2014).[3] Av förklarliga skäl betyder det att vi från tidigare forskning i bästa fall kan få intressanta referensvärden (i tillägg till etablerade riktvärden) och metodiska tillnärmningar.

Material och metod

Urval och datainsamling

För att besvara de tre forskningsfrågorna i studien har vi utgått från det delprov i läsförståelse som genomfördes i mars 2015, utformat av Gruppen för nationella prov i svenska och svenska som andraspråk vid Uppsala universitet.[4] Provet består av 7 texter med 20 tillhörande uppgifter och bedömningsanvisningar. Av de 20 uppgifterna är 14 st CR-uppgifter, varav 4 kortsvarsuppgifter som ger 0 eller 2 poäng samt 10 uppgifter som kräver längre svar och som kan ge 0-2-4 poäng (8 st), 0-2-4-6 poäng (1 st) eller 0-2-4-6-8 poäng (1 st).

Sex lärare vid tre olika skolor har bedömt tre anonymiserade elevlösningar, vilket ger ett sammanlagt dataset om 252 bedömningar. En av lärarna kallas ursprungsbedömare (Bedömare 1) och var den lärare som på skolan hade i formell uppgift att bedöma elevernas lösningar (ej anonymiserat).

[3]I den här artikeln diskuterar vi inte datorbaserad bedömning av elevgensvar. Fältet *automated essay scoring* (AES) är emellertid stort och intresserade läsare kan exempelvis konsultera ett specialnummer om AES i *Assessing Writing* (Elliot & Williamson, 2013).

[4]Hela provmaterialet omfattas av en sekretess som sträcker sig till 2021-06-30. Av det skälet kan vi inte visa vare sig texter eller uppgifter utan endast beskriva dessa såsom typer eller kategorier.

Urvalet av lärare och elevlösningar är med andra ord litet och fyller ingen representativ funktion i förhållande till populationen svensklärare. I så motto skulle studien kunna betraktas som en fallstudie av bedömarreliabilitet. Men med det sagt är det som undersöks i studien heller inte svensklärares generella förmåga att bedöma elevers provlösningar samstämmigt, utan hur väl den specifika provkonstruktionen fungerar för att producera likvärdiga bedömningar av elevers läsförmåga. För det syftet är också ett mindre dataset tillfyllest så länge det är stort nog för att kunna generera statistiskt signifikanta mått på samstämmigheten.

Lärarna som bedömt elevlösningarna är erfarna, behöriga, legitimerade och kvalificerade svensklärare (se tabell 1). Vid tiden för datainsamlingen hade de nyligen genomfört bedömning av delprovet i läsförståelse på sina respektive skolor och ombads att inom ramen för studien rätta och bedöma ytterligare tre elevlösningar. De tre elevlösningarna anonymiserades och sändes via mejl till lärarna, som bedömde, scannade in sammanställningen av sin poängsättning och mejlade tillbaka.

Statistiska analysmetoder

För att undersöka samstämmigheten mellan de sex bedömarna använder vi oss av följande tre mått: procentuell samstämmighet, Cohens viktade kappa, κ_v (Cohen, 1968) och Intraclass Correlation Coefficient, ICC (McGraw & Wong, 1996). De två första måtten tillhör gruppen konsensusmått och ICC är ett konsistensmått. Vi kommer att utföra beräkningar på bedömningar av CR-uppgifterna.

Procentuell samstämmighet och kappamåtten estimeras för ett par i taget. Med sex bedömare får vi 15 parkombinationer. Procentuell samstämmighet räknas ut genom att dividera antalet exakt samstämmiga bedömningar av uppgiftslösningar med det totala antal bedömningar. Detta görs manuellt för var och en av de 15 parkombinationerna.

Med Cohens kappa kan vi kontrollera den samstämmighet som vi kan anta att två bedömare uppnår enbart genom slumpen. Med viktad kappa tas också hänsyn till omfattningen av diskrepansen mellan bedömare; ju fler skalsteg två bedömare

Tabell 1. Uppgifter om lärarna och deras kvalifikationer i yrket

	Kön	Ålder	Yrkeserf.	Relevant formell utb.[a]	Leg.	Förstelärare	Övriga ämnesrelaterade uppdrag
Bedömare 1	K	41	18 år	X	X	X	Läs- och skrivutv.
Bedömare 2	K	59	20 år	X	X		Ämnesföreträdare
Bedömare 3	K	53	29 år	X	X	X	Ämnesföreträdare
Bedömare 4	K	59	13 år	X	X		
Bedömare 5	K	46	20 år	X	X	X	Processledare
Bedömare 6	K	44	15 år	X	X		

[a]Med relevant formell utbildning menas examen från ålders- och ämnesadekvat lärarutbildning. Fyra av lärarna har ämneslärarexamen för åk 4–9 i bl.a. ämnet svenska. Två av lärarna har ämneslärarexamen för högstadiet respektive för högstadiet och gymnasiet i bl.a. svenska.

befinner sig från varandra, desto lägre kappavärde. Även viktad kappa beräknas manuellt för vart och ett av de 15 paren.

ICC är en grupp modeller som kan skatta konsistensen eller konsistens *och* konsensus i en grupp av bedömare. I det här sammanhanget är vi intresserade av den första typen av estimat. Vi kommer nedan att använda oss av modellen *ICC two-way random, consistency, single measure*, också kallad ICC(C,1) (McGraw & Wong, 1996). Den ger oss ett värde på reliabilitetskoefficienten under antagandet att en elevlösning bara blir rättad och bedömd av en lärare. ICC kräver en i sammanhanget mer komplicerad uträkning, vilken vi genomför i SPSS 23.0 (IBM, 2015).

För att vidare analysera skillnader mellan bedömarna kommer vi att beskriva bedömarnas relativa stränghet med hjälp av medelvärdesanalyser. Vi kommer också att göra en närmare kvalitativ analys av några exempel på uppgifter i provet där variationen mellan bedömarna leder till stora poängskillnader för eleverna. Avslutningsvis redovisar vi också elevernas sammanlagda provpoäng för de sex olika bedömarna.

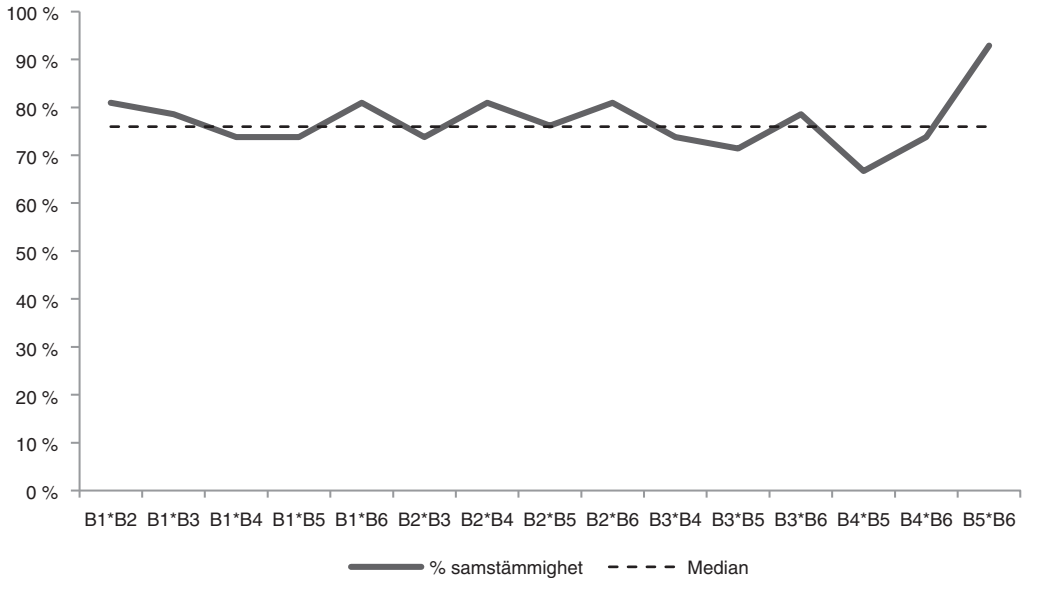
Resultat

Procentuell samstämmighet

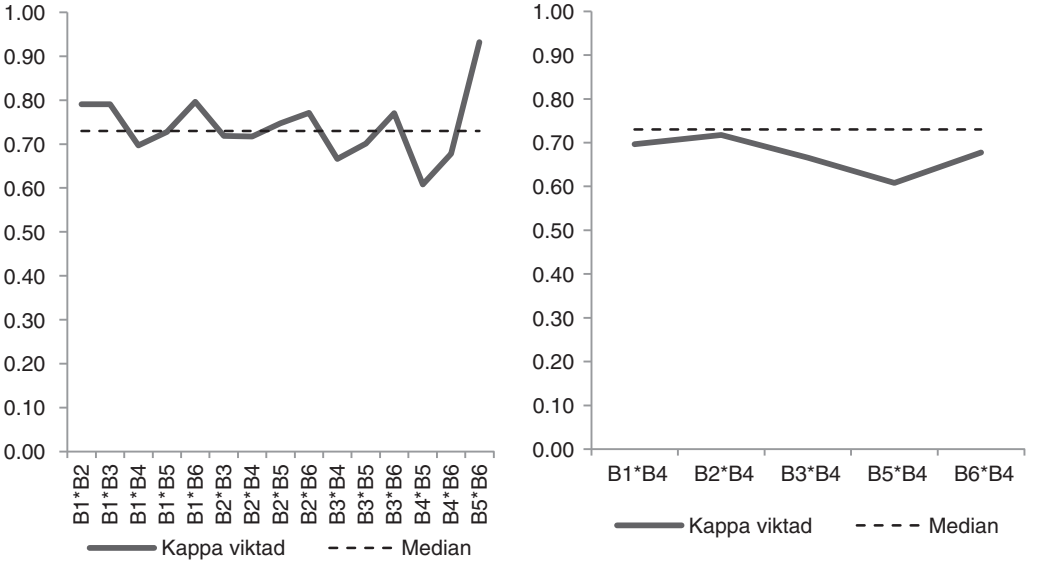
Inledningsvis studerar vi procentuell samstämmighet mellan bedömarna. Medianvärdet för samtliga parkombinationer är 76 %. Det minsta värdet, 67 %, uppmäter vi mellan bedömare 4 och bedömare 5. Det högsta värdet, 93 %, finner vi mellan bedömare 5 och bedömare 6. Skillnaden mellan högsta och lägsta värdet är därmed 26 procentenheter. Vi har alltså att göra med en parkombination (B4–B5) där man bedömer uppgifterna olika i fler än tre fall av tio, samtidigt som resten av parkombinationerna resulterar i en samstämmighet som åtminstone överstiger 70 %. I figur 1 åskådliggörs resultaten för vart och ett av paren (se appendix för resultat i tabellformat).

Cohens kappa

Mot den procentuella överensstämmelsen kan vi nu ställa kappavärdet, som alltså både kontrollerar den samstämmighet som uppstår till följd av slumpen och tar hänsyn till hur stort avståendet mellan två bedömare är. Det genomsnittliga kappavärdet är .73. Detta värde är, som förväntat, lägre än procentuella samstämmigheten och understiger med god marginal det eftersträvansvärda kappavärdet på .80. Det minsta värdet som vi uppmäter är mellan B4 och B5, med kappavärde på .61. Det högsta finner vi mellan B5 och B6, med kappavärde på .93. Tre av värdena i serien understiger .70, vilket gäller för B3–B4, B4–B5 och B4–B6. Resultaten indikerar att B4:s bedömning avviker i högre grad från de övriga bedömarnas och att B4 därigenom bidrar negativt till samstämmigheten. I inget av fallen då B4 ingår i ett par överstiger samstämmigheten genomsnittet för hela gruppen. Emellertid nås inte ett tillfredsställande genomsnitt även om B4 exkluderas. Det nya genomsnittet blir då .77, vilket även det understiger eftersträvansvärda nivåer. Resultaten för hela gruppen och för kombinationer i vilka B4 ingår redovisas i figur 2.



Figur 1. Procentuell samstämmighet för samtliga parkombinationer. Bedömarna är numrerade från 1–6, vilket innebär att B1 avser bedömare 1 osv. Den streckade linjen anger medianvärdet (76 %).



Figur 2. I grafen till vänster återges kappavärde för samtliga parkombinationer. Bedömarna är numrerade från 1–6, vilket innebär att B1 avser bedömare 1 osv. Den streckade linjen anger medianvärdet (.73). I grafen till höger återges kappavärde för parkombinationer i vilka B4 ingår. Även här anger den streckade linjen medianen.

Intra Class Correlation (ICC) och medelvärdesanalys

Konsistensundersökningarna indikerar att bedömarna rör sig i samma riktning, dvs. att de rangordnar elevsvaren på liknande sätt. När vi undersöker alla CR-items får vi ett ICC-värde på .815, vilket är i underkant av .90. Konfidensintervallet (se tabell 2) är emellertid tämligen brett och det går inte att utesluta att det egentliga värdet ligger både under eller över .815.

När vi delar upp item efter skallängd kan vi få indikationer på vilken typ av item som leder till störst konsistens. Resultaten visar högre konsistens för item med 2 skalsteg (.732) än för item med 3 skalsteg (.697). Högst konsistens har emellertid en item med 5 skalsteg (.842), men här bidrar ett litet urval till stor osäkerhet. Med ett konfidensintervall på 95 % i beaktande ligger det egentliga värdet någonstans mellan .457 och .995. I tabell 2 och i figur 3 återges resultaten.

Tabell 2. Konsistensundersökning

	KI[a] 95 % nedre	ICC(C.1)[b]	KI 95 % övre
Alla items (n = 42)	.736	.815**	.883
2 skalsteg (n = 12)	.530	.732**	.895
3 skalsteg (n = 24)	.550	.697**	.880
5 skalsteg (n = 3)	.457	.842**	.995

Kommentar. Av mättekniska skäl saknas underlag för att estimera ICC för item med fyra skalsteg. [a]KI = konfidensintervall. [b]ICC(C,1) avser ICC, two-way random, consistency, single measure. ** = $p < .01$

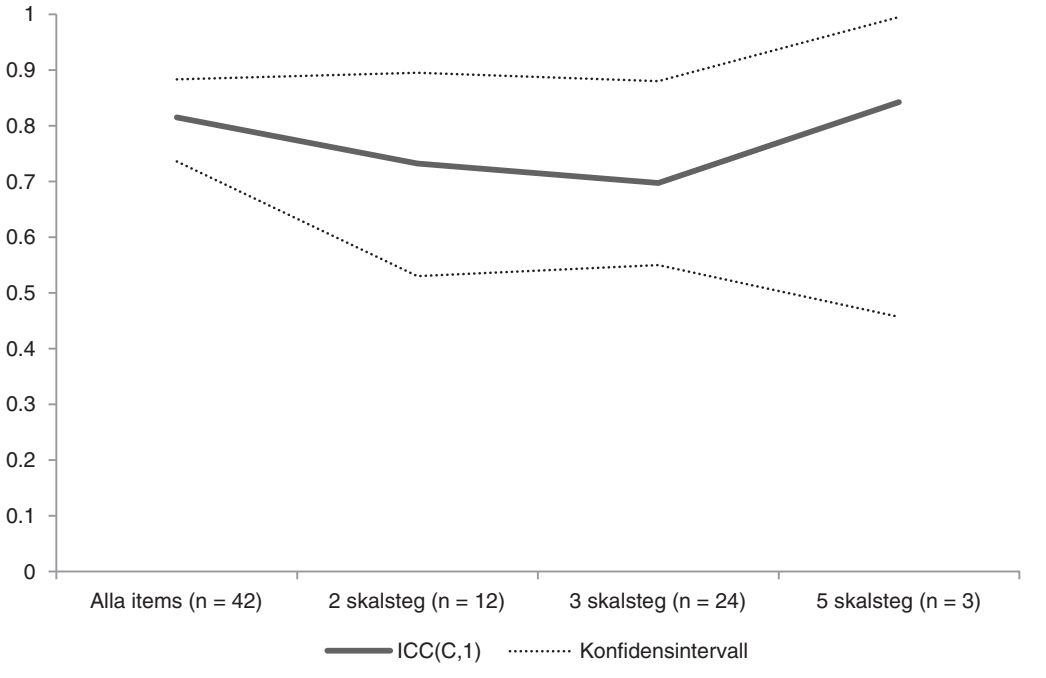
Sammantaget ser vi att några särskilt höga nivåer av samstämmighet inte uppnås vare sig vi uttrycker det med procentuell samstämmighet, kappavärden eller med ICC-värden. Bedömarna som ingår i den här studien är alltså inte utbytbara med varandra, utan tycks tolka elevsvaren på olika sätt och dessutom vara olika stränga. För att närmare undersöka just stränghet har vi utfört en enkel medelvärdesanalys på alla CR-items. Resultatet av denna analys återges i tabell 3.

Tabell 3. Medelvärdesanalys

	Bedöm. 1	Bedöm. 2	Bedöm. 3	Bedöm. 4	Bedöm. 5	Bedöm. 6
Medelvärde	2.86	2.71	2.62	2.62	2.48	2.52
Standardavvikelse	1.93	1.92	1.90	1.85	1.97	1.98

Kommentar. Medelvärdesanalys baserad på CR-uppgifter (n = 42). Skillnaderna mellan de olika bedömarparen är icke-signifikanta.

Medelvärdesanalysen visar att bedömarna är olika stränga, även om skillnaderna inte är statistiskt signifikanta. Vi kan bland annat notera att bedömare 1, den ursprungliga bedömaren, är den mest generösa med ett medelvärde på 2,86 poäng per uppgift. Bedömare 5 är strängast med ett medelvärde på 2,48 poäng per uppgift. Konsensusundersökningen visar också att just B1 och B5 är oeniga i drygt 25 % av



Figur 3. ICC-värden och konfidensintervall för alla items, och items med 2, 3 respektive 5 skalsteg.

fallen. Även om skillnaderna inte är statistiskt signifikanta hade det för en elev ändock varit betydligt mer gynnsamt att bedömas av B1 än B5. Vilka konsekvenser det får för eleverna och för bedömningen av deras sammanlagda provprestation återkommer vi till i avsnittet om konsekvenser för eleven nedan.

Sammantaget är bedömarna i den här studien som grupp betraktat mindre konsistenta än vad som är eftersträvansvärt. Förvisso tycks par som B5–B6 (93 %, $\kappa = .93$) ha tämligen likartade uppfattningar om hur elevsvaren skall bedömas, men andra par, som t.ex. B3–B4 (74 %, $\kappa = .67$), ligger betydligt längre ifrån varandra. Variationen i den här relativt lilla gruppen är därmed omfattande.

Uppgiftsanalys

För att närmare undersöka möjliga orsaker till denna variation har vi närstuderat bedömningen av några av provets uppgifter. Eftersom antalet bedömare är för få för att använda Kappa och ICC på enskilda uppgifter har vi istället analyserat uppgifterna med hjälp av deskriptiv statistik.

Vi ska se på två exempel på uppgifter med betydande variation i poängsättningen. Det ena gäller uppgift 6 som ges till en narrativ text i provhäftet. Huvudpersonen i berättelsen ska lösa en skrivuppgift utifrån två givna kriterier. Det ena kriteriet rör i vilken mån innehållet i uppgiften angår honom, det andra gäller i vilken mån han har egen erfarenhet av ämnet ifråga. I formulering som ges till eleverna, som löser det nationella provet, uppges att huvudpersonen i berättelsen löser skrivuppgiften i enlighet med det ena kriteriet men inte det andra. Uppgiften för eleverna som

Tabell 4. Bedömning av uppgift 6

	Bedöm. 1	Bedöm. 2	Bedöm. 3	Bedöm. 4	Bedöm. 5	Bedöm. 6
Elev 1	0	0	0	0	0	0
Elev 2	4	4	4	0	2	2
Elev 3	2	2	2	2	2	2

genomför provet blir att förklara hur. I tabell 4 redovisas hur de tre eleverna bedömts av de sex bedömarna.

Som framgår av tabellen är de sex bedömarna helt överens om bedömningen av två av elevsvaren medan bedömningarna för det tredje (Elev 2) går helt isär. För denna elev (2) används alla tre bedömningsalternativen på skalan. För att få full poäng, ska det enligt bedömningsanvisningen framgå att huvudpersonen i berättelsen skriver om det som angår honom, men inte utifrån egen erfarenhet. Elev 2 besvarar uppgiften på följande vis.

> Det angick verkligen huvudpersonen eftersom han/hon alltid velat ha en och kunde tänka sig in i det, men eftersom det inte hände så har han inte upplevt det.

Genom svaret visar eleven att han har förstått premisserina både i texten och i uppgiften. Att ämnet angår huvudpersonen och att han därmed följer det första kriteriet redovisas (han hade alltid velat ha en [hund]), liksom att det andra kriteriet inte följs (han hade inte upplevt det) [haft en hund]. Samtidigt är elevens text vag i flera avseenden. Det framgår inte *vad* huvudpersonen alltid velat ha, vilket gör precisionen i förhållande till texten svag. Att skriva att ”det inte hände” kan på liknande vis uppfattas som en vag förklaring till det sätt på vilket huvudpersonen inte följer det andra kriteriet i skoluppgiften. En rimlig förklaring till variansen i bedömningarna här är alltså att vagheten i elevsvaret gjort att bedömningsanvisningarna tycks ge utrymme för olika tolkningar av dess kvalitet. Inte minst när uppgiften frågar om flera saker samtidigt och därmed kräver mer komplexa bedömningsanvisningar blir detta en naturlig konsekvens. Åtminstone så länge bedömarna inte först fått möjlighet att utförligt diskutera kriterierna och hur de ska tillämpas.

I uppgift 12 finns en liknande problematik. Uppgiften ges till ett reportage i texthäftet där olika personer uttalar sig om ett givet ämne. I uppgiften nämns två av dessa personer och att deras åsikter går isär på en punkt. ”Vilken? Hur motiverar de

Tabell 5. Bedömning av uppgift 12

	Bedöm. 1	Bedöm. 2	Bedöm. 3	Bedöm. 4	Bedöm. 5	Bedöm. 6
Elev 1	6	4	4	4	4	4
Elev 2	8	8	8	4	8	8
Elev 3	0	4	0	0	2	0

sina åsikter?" För att få full poäng på uppgiften (8 poäng) ska svaret innehålla båda personernas respektive åsikt och motivering. Om motivering eller åsikt saknas i elevsvaret dras poäng av; skalan över möjliga bedömningar är 0-2-4-6-8. I tabell 5 redovisas lärarnas bedömningar av de tre elevsvaren.

Vi ser att variationen mellan bedömarna får stora konsekvenser för de enskilda eleverna. För två av eleverna kan det skilja upp till fyra poäng på en enda uppgift beroende på vem av lärarna som gjort bedömningen. Samtidigt är ICC för just den här uppgiften .84, även om konfidensintervallet är brett (se tabell 2). Detta understryker vikten av att genomföra olika typer av samstämmighetsundersökningar. Lärarna har i detta fall rangordnat elevsvaren på ett liknande vis, men skalans längd innebär att även små bedömarvariationer får stora konsekvenser. Vad detta innebär för det sammanlagda provresultatet redovisas i nästa avsnitt.

Konsekvenser för eleven

Som statistiken ovan visar är samstämmigheten mellan de sex bedömarna enligt vedertagna tumregler bristande. Den befintliga variationen mellan lärarnas bedömningar får också konsekvenser i praktiken för eleverna. Grad av stränghet i bedömning är en välkänd variabel på vilken variationen mellan olika bedömare kan vara stor. Skillnaden mellan bedömarna i vårt urval kan beskrivas på olika vis. Dels kan vi titta på variationen i stränghet över hela datasetet, vilket redovisades ovan i tabell 3. Dels kan vi undersöka utfallet i form elevernas sammanlagda poäng på provet för respektive bedömare. Av jämförelsen framgår att urspungsbedömaren (Bedömare 1) är klart mer generös i sin bedömning än de övriga fem, som i sin tur ligger relativt närmre varandra ifråga om genomsnittlig stränghet. För eleverna får det här betydelse för deras slutgiltiga provpoäng (se tabell 6 nedan).

Tabell 6. Elevernas sammanlagda resultat på provet för respektive bedömare

	Bedöm. 1	Bedöm. 2	Bedöm. 3	Bedöm. 4	Bedöm. 5	Bedöm. 6	*Max diff*
Elev 1	48	46	46	46	40	42	*8*
Elev 2	58	50	48	46	48	48	*12*
Elev 3	38	40	38	40	38	38	*2*

Maxpoäng: 66 p. Betygsgränser på provet: F (0–30p), E (32–38p), D (40–44p), C (46–54p), B (56–58p), A (60–66p).

I de samlade poängsummorna ovan ingår även poäng för MC-uppgifterna. Även bland dessa finns faktiskt en liten variation genom att några rena felrättningar gjorts. Som framgår av tabellen kan poängsumman för en enskild elev skilja så smycket som 12 poäng beroende på vem av lärarna som rättar provet. För eleven i det här urvalet (Elev 2) ger det en skillnad på ett betygsteg, men som framgår av tabellen skulle en så stor poängskillnad för en annan elev kunna innebära en skillnad med två hela betygssteg. Vilken enskild lärare som bedömer elevens prestationer på det nationella läsprovet har alltså avgörande betydelse för hans eller hennes provbetyg, och därmed rimligen också betydelse för ämnesbetyget i svenska.

Diskussion

Sammanfattning av resultaten

Samstämmighet i bedömning är en förutsättning för tillförlitligheten i ett testinstrument. Kort sagt har vi inte någon större nytta av ett testinstrument som inte kan upprätthålla höga nivåer av samstämmighet (Haladyna & Rodriguez, 2013). Resultaten från den här studien indikerar att samstämmigheten ifråga om bedömning av elevers läsförmåga på det nationella provet inte helt uppfyller gängse förväntningar på interbedömarreliabilitet när det gäller tester av stor betydelse för den enskilde testtagaren (McNamara, 2000). Detta redovisas i artikeln på flera vis. För det första är den genomsnittliga överensstämmelsen i gruppen låg. Det statistiska underlaget för studien är visserligen begränsat och man bör iaktta försiktighet ifråga om generaliseringar. Men om liknande resultat skulle bekräftas även på större statistiska underlag och om vi tar interbedömarreliabilitet som utgångspunkt för att mäta tillförlitligheten i läsprovet såsom en del av det nationella provsystemet så kan vi konstatera att endast sju av tio bedömningar är tillförlitliga. För det andra är variationen i överensstämmelse mellan olika bedömarpar stor. Det betyder å ena sidan att det bevisligen går att uppnå högre nivåer av samstämmighet ifråga om hur uppgifter på läsprovet ska rättas. Å andra sidan betyder det att vi har en andel svensklärare vars bedömningar avviker påtagligt från det som kan antas vara den gängse normen för bedömning av elevers läsförmåga. För det tredje får variationen i lärarnas bedömning betydande konsekvenser för den enskilde elevens slutgiltiga provresultat, för det sammanlagda provbetyget och potentiellt även för slutbetyget i svenskämnet.

Användbarheten i CR-uppgifter

Vad säger då resultaten om konstruktionen av det svenska nationella läsprovet i årskurs nio? Är det rimligt att behålla uppgifter på vilka variationen mellan lärarnas bedömningar av elevers lösningar är så stor? En vanlig reaktion vid påvisande av bristande interbedömarreliabilitet är att kraven på objektivitet höjs, antingen genom att uppgifter som fordrar subjektiv bedömning tas bort till förmån för uppgifter som möjliggör objektiv bedömning, eller också genom att lärares bedömningskompetens underkänns, följt av krav på externa censorer (jfr t.ex. Skolinspektionen, 2011). I många läsprov används mycket riktigt en högre andel MC-frågor som ett sätt att undanröja problemet med potentiellt bristande interbedömarreliabilitet (Campbell, 2005; Roe & Lie, 2009; Solheim & Skaftun, 2009). I de svenska proven under de senaste åren har vi också sett en gradvis ökande andel MC-frågor i läsprovet i åk 9 (Tengberg, in press), vilket möjligen åstadkommer en skenbar objektivitet. Vill man få en reell effekt ifråga om höjd reliabilitet krävs emellertid en betydligt högre andel MC-frågor än vad som hittills varit fallet i de svenska proven. Samtidigt riskerar en sådan förändring undergräva den ekologiska validiteten i provet, dvs. det att förhållanden som råder under testet (som exempelvis rör testets metoder och material) i så hög utsträckning som möjligt återspeglar de naturliga förhållanden som testet avser mäta (Schmuckler, 2001). MC-frågor (åtminstone i den traditionella

form som används i skandinaviska läsprov med ett korrekt svar och tre distraktorer) för med sig åtminstone tre för sammanhanget väsentliga inskränkningar av vad för slags kunskap som blir möjlig att pröva. Den första handlar om att det endast går att formulera frågor som har ett givet korrekt svar och där övriga svar är mer eller mindre objektivt felaktiga. Det innebär att en rad olika frågeställningar, som vi inom ämnessammanhanget normalt anser vara både intressanta och rationella när det gäller att pröva elevernas läsförmåga, måste utelämnas (Nordenfors, 2008; Pearson & Hamm, 2005; Tengberg, 2014). Den andra inskränkningen gäller det faktum att vi ibland vill kunna skilja mellan mer och mindre kvalificerade lösningar på en uppgift, vilket inte är möjligt i MC-uppgifter men som möjliggörs av CR-uppgifter givet att testkonstruktionen innehåller noggranna beskrivningar av vad som utmärker olika kvalitetsnivåer av elevsvar. Den tredje inskränkningen handlar om att elever i MC-uppgifter endast har att ta ställning till texttolkningar som någon annan föreslagit för dem istället för att själva ställas inför uppgiften att formulera dessa tolkningar, något som vi inom ramen för undervisningen i ämnet oftast tar för självklart att elever ska lära sig att göra.

Samtidigt måste alltså CR-uppgifterna i det nationella läsprovet konstrueras på ett sätt så att de borgar för högre nivåer av samstämmighet mellan bedömarna. Det är förvisso ett problem att även om många läsprov kompletterar MC-uppgifter med CR-uppgifter så saknar vi än så länge specifika benchmarks för inter-bedömarreliabilitet av de senare. Ett syfte med den här studien kan därför sägas vara att den lämnar referensvärden som kommande undersökningar kan relatera till.

Hur bör man hantera bedömareffekter?

I *Standards for Educational and Psychological Testing* konstateras att reliabilitet är en nödvändig, men inte tillräcklig förutsättning för validitet (AERA, APA, & NCEM, 2014). I den här studien har vi alltså kunnat visa varför reliabla mått är en förutsättning: samma elever får olika resultat, beroende på vem som gör bedömningen. Detta betyder att provresultatet, och i förlängningen betyget, inte har någon entydig innebörd. Därför kan de inte behandlas som entydiga uttryck för elevernas läsförmåga.

Vad kan då förklara resultaten i studien? Vi kan misstänka att informanterna i denna studie, liksom är vanligt bland bedömare, engagerar sig i den komplexa relationen mellan bedömare, elevrespons och bedömningsanvisningar på kvalitativt olika sätt (Bejar, 2012). Bedömare tolkar både anvisningar och elevresponser olika, men det kan också finnas andra förklaringar. Till exempel har innovativ forskning på fältet *rater cognition* kunnat belägga att den av bedömaren uppfattade betydelsen av ett givet kriterium inverkar på stränghetsgraden. Ju viktigare kriterium, desto strängare bedömning (Eckes, 2012). Oavsett vad som har orsakat variationen bland bedömarna i denna studie är vi i behov av åtgärder som kan säkra reliabiliteten framgent.

Utöver alternativet att helt utesluta uppgifter som kräver subjektiv bedömning är utbudet av kvalitetssäkrande metoder strängt taget begränsat till två huvudkategorier: bedömarträning och multipla bedömare. Bedömarträning innebär att

fokusera på bedömares kompetens och i största möjliga mån få bedömare att tolka såväl bedömningsanvisningar som elevresponser på ett likartat sätt. Detta har visat sig vara en effektiv åtgärd för att höja intrabedömarreliabiliteten, men i mindre grad interbedömarreliabiliteten (Eckes, 2011). Bedömarträning tycks med andra ord göra enskilda bedömare mer konsistenta med sig själva, utan att konsistensen och samstämmigheten *mellan* bedömarna för den skull ökar.

Att införa ett system med multipla bedömare innebär en möjlighet att låta bedömareffekter ta ut varandra. En ordning med multipel bedömning medför dock nya ekonomiska och tekniska förutsättningar. Att låta två eller tre personer bedöma varje elevrespons innebär en ökning av antalet arbetstimmar förbundet med bedömning, vilket i sin tur leder till ökade kostnader för vikarier eller övertidsersättning. I teknisk bemärkelse bygger ett system med multipla bedömare på att varje enskild bedömare är konsistent, annars tar inte bedömareffekterna ut varandra. Av den orsaken bör man även i ett sådant system vinnlägga sig om rigorös bedömarträning.

Att bedömarna i den här studien inte uppnår tillfredsställande grad av samstämmighet är en stark signal om behovet av mera deskriptiv och experimentell forskning relaterat till det nationella läsprovet. I det kommande är det angeläget att söka svar på om bedömarsamstämmigheten på ett större urval bedömare är densamma eller skiljer sig från den som är uppmätts här. Det vore också betydelsefullt att undersöka betingelser för bedömning (bedömningsanvisningar, procedurer m.m.), både deskriptivt och som del av studier som prövar olika tillvägagångssätt.

Att bedömarsamstämmigheten i svenska nationella läsprov motsvarar den nivå som forskningen klassificerar som ”rock-bottom minimum” för vad som kan anses acceptabelt kan inte vara i linje vare sig med provkonstruktörens eller med Skolverkets ambitioner. Hittills har dock inga motsvarande undersökningar genomförts och vi menar att det är upp till provkonstruktören och Skolverket att demonstrera att reliabilitet och validitet föreligger i det nationella provsystemet.

Referenser

Alderson, J. C. (2000). *Assessing reading*. Cambridge: Cambridge University Press.

American Educational Research Association, American Psychological Association & National Council on Educational Measurement (Red.). (2014). *Standards for Educational and Psychological Testing*. Washington: American Educational Research Association.

Attali, Y. (2014). A ranking method for evaluating constructed responses. *Educational and Psychological Measurement*, *74*(5), 795–808.

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, *2*(1), 1–34.

Bejar, I. (2012). Rater cognition: Implications for validity. *Educational Measurement: Issues and Practice*, *31*(3), 2–9.

Campbell, J. R. (2005). Single instrument, multiple measures: Considering the use of multiple item formats to assess reading comprehension. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 347–368). Mahwah, New Jersey: Lawrence Erlbaum Ass.

Cohen, J. (1968). Weighted Kappa. *Psychological Bulletin*, *70*(4), 213–220.

Congdon, P. J. & McQueen, J. (2000). The Stability of Rater Severity in Large-Scale Assessment Programs. *Journal of Educational Measurement*, *37*(2), 163–178.

DeSanti, R. J. & Sullivan, V. G. (1984). Inter-rater reliability of the cloze reading inventory as a qualitative measure of reading comprehension. *Reading Psychology: An International Journal*, *5*, 203–208.

Eckes, T. (2011). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments*. Frankfurt am Main: Peter Lang.

Eckes, T. (2012). Operational rater types in writing assessment: Linking rater cognition to rater behavior. *Language Assessment Quarterly: An International Journal*, *9*(3), 270–292.

Elliot, N. & Williamson, D. M. (2013). Assessing Writing special issue: Assessing writing with automated scoring systems. *Assessing Writing*, *18*(1), 1–6.

Engelhard, G. (1994). Examining rater errors in the assessment of written composition with a many-faceted Rasch-model. *Journal of Educational Measurement*, *31*(2), 93–112.

Engelhard, G. (2002). Monitoring raters in performance assessments. In G. Tindal & T. M. Haladyna (Eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation* (pp. 261–288). Mahwah, NJ: Lawrence Erlbaum Associates.

Gwet, K. L. (2014). *Handbook of inter-rater reliability.* Gaithersburg, MD: Advanced Analytics, LLC.

Güler, N. (2014). Analysis of Open-Ended Statistics Questions with Many Facet Rasch Model. *Eurasian Journal of Educational Research*, *55*, 73–90.

Haladyna, T. M. & Rodriguez, M. C. (2013). *Developing and validating test items*. New York: Routledge.

IBM. (2015). *SPSS Statistics* (version 23) [Computer Software]. New York: IBM.

IEA (2009). *PIRLS 2011 Assessment framework*. Chestnut Hill, MA: Boston College.

Illinois State Board of Education (2013). *Illinois standards achievement test 2013. Technical Manual*. Springfield, IL: Illinois State Board of Education, Division of Assessment.

Jacob, B. A. & Levitt, S. D. (2003). Rotten apples: An investigation of the prevalence and predictors of teacher cheating. *The Quarterly Journal of Economics*, *3*, 843–877.

Kane, M., Crooks, T. & Cohen, A. (1999). Validating measures of performance. *Educational Measurement: Issues and Practice*, *18*(2), 5–17.

Lim, G. S. (2011). The development and maintenance of rating quality in performance writing assessment: A longitudinal study of new and experienced raters. *Language Testing*, *28*(4), 543–560.

McGraw, K. O. & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46.

McNamara, T. F. (2000). *Language testing*. Oxford: Oxford University Press.

Meadows, M. & Billington, L. (2005). *A review of the literature on marking reliability.* London: National Assessment Agency.

Moser, G. P., Sudweeks, R. R., Morrison, T. G. & Wilcox, B. (2014). Reliability of ratings of children's expressive reading. *Reading Psychology*, *35*(1), 58–79.

Myford, C. M. & Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Measurement: Part I. *Journal of Applied Measurement*, *4*(4), 386–422.

Myford, C. M. & Wolfe, E. W. (2009). Monitoring rater performance over time: A framework for detecting differential accuracy and differential scale category use. *Journal of Educational Measurement*, *46*(4), 371–389.

Nordenfors, M. (2008). Kan en skrattande ödla vara nyttig? *Pedagogiska magasinet nr 1/08*, 68–70.

OECD (2013). *PISA 2012. Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* OECD Publishing. http://dx.doi.org/10.1787/9789264190511-en. Hämtad 2013-12-02.

Pearson, P., Calfee, R., Walker Webb, P. & Fleischer, S. (2002). *The role of performance-based assessments in large scale accountability systems: Lessons learned from the inside*. Washington DC: Council of Chief State School Officers.

Pearson, P. D. & Hamm, D. N. (2005). The assessment of reading comprehension: A review of practices: Past, present, and future. In S. G. Paris & S. A. Stahl (Eds.), *Children's reading comprehension and assessment* (pp. 13–70). Mahwah, NJ: Law. Erlbaum Ass.

Roe, A. & Lie, S. (2011). Nasjonale leseprøver i et didaktisk og testteoretisk perspektiv. In S. Dobson, A. B. Eggen & K. Smith (Eds.), *Vurdering, prinsipper og praksis: Nye perspektiver på elev- og læringsvurdering* (s. 145–165). Oslo: Gyldendal.

Schmuckler, M. A. (2001). What is ecological validity? A dimensional analysis. *INFANCY*, *2*(4), 419–436.

Skolverket (2009). *Bedömaröverensstämmelse vid bedömning av nationella prov.* Dnr 2008:286. Stockholm: Skolverket.

Solheim, O. J. & Skaftun, A. (2009). The problem of semantic openness and constructed response. *Assessment in Education: Principles, Policy & Practice*, *16*(2), 149–164.

Skolinspektionen (2011). *Lika eller olika: Omrättning av nationella prov i grundskolan och gymnasieskolan*. Stockholm: Skolinspektionen.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation*, *9*(4). Retrieved from http://pareonline.net/getvn.asp?v=9&n=4

Stemler, S. E. & Tsai, J. (2008). Best practices in interrater reliability. Three common approaches. In J. Osborne (Ed.), *Best practices in quantitative methods* (pp. 29–50). Thousand Oaks, CA: SAGE Publications, Inc.

Tengberg, M. (in press). National reading tests in Denmark, Norway, and Sweden. *A comparison of construct definitions, cognitive targets, and response formats. Language Testing.* http://ltj.sagepub.com/content/early/2015/10/28/0265532215609392.full.pdf+html

Tengberg, M. (2014). Konstruktion och bedömning av förmågan att läsa (och förstå) skönlitterär text. *Två nedslag i det nationella provets läsförståelsedel, åk 9. Educare*, *1*, 79–97.

van den Bergh, H. (1990). On the construct validity of multiple-choice items for reading comprehension. *Applied Psychological Measurement*, *14*(1), 1–12.

Van Moere, A. (2014). Raters and ratings. In A. J. Kunnan (Ed.), *The Companion to Language Assessment* (pp. 1358–1374). Chichester, West Sussex: Wiley-Blackwell.

Appendix

Tabell A. Procentandel samstämmighet och kappavärden för respektive parkombination

	Agree %	Viktad kappa	Standard error	KI 95 nedre	KI 95 % övre
B1*B2	81 %	.79	.07	.64	.94
B1*B3	79 %	.79	.06	.67	.91
B1*B4	74 %	.70	.09	.53	.86
B1*B5	74 %	.73	.08	.57	.89
B1*B6	81 %	.80	.07	.65	.94
B2*B3	74 %	.72	.08	.56	.88
B2*B4	81 %	.72	.09	.53	.90
B2*B5	76 %	.75	.08	.58	.91
B2*B6	81 %	.77	.08	.61	.94
B3*B4	74 %	.67	.09	.49	.84
B3*B5	71 %	.70	.08	.54	.86
B3*B6	79 %	.77	.07	.63	.92
B4*B5	67 %	.61	.10	.42	.80
B4*B6	74 %	.68	.09	.50	.86
B5*B6	93 %	.93	.04	.85	1.00
Median	*76 %*	*.73*	*.08*	*.57*	*.90*
Min	*67 %*	*.61*	*.04*	*.42*	*.80*
Max	*93 %*	*.93*	*.10*	*.85*	*1.00*
Variationsvidd	*26 %*	*.32*	*.06*	*.43*	*.20*