DEL III.
FORMER FÖR BEDÖMNING I SPRÅK

6. Bedömning av muntlig förmåga i engelska – om bedömarvariation och beslutsprocesser ur ett nationellt och europeiskt perspektiv

Linda Borger

Som språklärare står vi inför uppdraget att bedöma komplex språklig förmåga, som till exempel att tala på målspråket. Olika faktorer, både språkliga och innehållsmässiga, vägs in i bedömningen och dessutom ska elevens prestation värderas i relation till kunskapskraven. I detta sammanhang aktualiseras frågor om bedömaröverensstämmelse och likvärdighet. Den studie som presenteras i denna artikel bygger på en licentiatuppsats (Borger, 2014), som undersöker bedömning av muntlig språkfärdighet i ett nationellt prov i engelska på gymnasienivå. Det muntliga provet består av par- eller gruppsamtal, och är ett exempel på ett så kallat *performance-prov*. Typiskt för dessa prov är att de innehåller uppgifter som är utformade för att så långt som möjligt likna autentiska situationer, där eleverna får använda målspråket på ett längre och sammanhängande sätt för att visa sin förmåga (McNamara, 1996). Bedömning av autentiska provuppgifter är komplex och förutsätter tolkning, vilket kan bidra till bedömareffekter, det vill säga variation i betyg som kan kopplas till skillnader mellan bedömarna snarare än elevprestationen. En stor utmaning med denna provtyp är därför att uppnå så goda nivåer som möjligt av bedömarsamstämmighet, medan den stora fördelen är den direkta kopplingen mellan den förmåga som ska bedömas (muntlig produktion och interaktion) och den observerade elevprestationen (i form av ett muntligt samtal).

Hur du refererar till det här kapitlet:
Borger, L. (2021). Bedömning av muntlig förmåga i engelska – om bedömarvariation och beslutsprocesser ur ett nationellt och europeiskt perspektiv. I Bardel, C. et al. (Red.). *Forskarskolan FRAM — lärare forskar i de främmande språkens didaktik* (s. 127–155). Stockholm: Stockholm University Press. DOI: https://doi.org/10.16993/bbg.g.

Bedömareffekter i språk, som beskrivs kortfattat nedan, är ett tämligen väl beforskat område, med uppenbar relevans för den här aktuella studien, som bland annat syftar till att undersöka bedömarvariation och att relatera den till lärares utsagor om sin bedömning. En viss tematik kan skönjas i de studier av bedömareffekter som gjorts. En av de vanligaste är att bedömare är mer eller mindre stränga i jämförelse med andra bedömare (Bachman et al., 1995; Eckes, 2005), vilket på engelska benämns som *severity* och *leniency*; ytterligheterna bland bedömare brukar ibland kallas ”hökar” och ”duvor”. Exempel på andra faktorer som kan påverka bedömningen i autentiska prov är att bedömare tolkar bedömningskriterierna på olika sätt och lägger olika vikt vid dem, vilket kan leda till att de ger olika betyg för samma elevprestation, eller ger samma betyg men av olika skäl (Eckes, 2009; McNamara, 1996; Orr, 2002). Tidigare studier som undersökt bedömares kognitiva processer vid bedömning av muntlig andraspråksfärdighet har visat att bedömare uppmärksammar både aspekter som beskrivs i bedömningskriterierna och aspekter som inte specifikt uttrycks i bedömningskriterierna, till exempel elevens ansträngning, intresse och personliga egenskaper (Ang-Aw & Goh, 2011; Brown, 1995; Meiron, 1998; Orr, 2002). Det framgår även av studier att lingvistiska aspekter, som grammatisk korrekthet och ordförråd, ofta ges större vikt i bedömningen än andra delar av den kommunikativa kompetensen (Magnan, 1988; McNamara, 1990). Delar av dessa olika aspekter av bedömareffekter kan skönjas även i den här aktuella studien.

Parsamtal

Par- och gruppsamtal, som används i de svenska nationella proven i språk och som undersöks i den här aktuella studien, ger i regel möjlighet till en mer naturlig interaktion än intervjuformatet som genomförs med en provdeltagare och en intervjuare/examinator. I parsamtal får provdeltagarna möjlighet att visa upp en större bredd av språkfunktioner och samtalsfärdigheter än i intervjusituationen (Brooks, 2009; ffrench, 1999; Kormos, 1999), där det finns en tydlig hierarkisk struktur mellan

provdeltagare och examinator. Det finns dock svårigheter med par- och gruppsamtal ur bedömningssynpunkt. För det första kan olika bakgrundsvariabler hos provdeltagarna, till exempel personlighet, kön och språklig nivå, påverka samtalspartnerns prestation (Foot, 1999; Nakatsuhara, 2013; Norton, 2005; O'Sullivan, 2002). Forskningsresultaten är inte entydiga beträffande om eller hur dessa så kallade *interlocutor effects* påverkar betyget. Däremot visar ett flertal studier som använder konversationsanalys som metod att bakgrundsvariablerna kan påverka interaktionsmönstret i samtalet (Galaczi, 2014; Lazaraton & Davis, 2008). En ytterligare utmaning vid par- och gruppsamtal är att individuella betyg sätts på en gemensamt skapad prestation, *co-constructed interaction* (Galaczi, 2008), vilket gör att det kan vara svårt att fullt ut skilja prestationerna åt (McNamara, 1997).

Syfte och frågeställningar

Syftet med studien är att undersöka bedömning av muntlig språkfärdighet i det nationella provet i kursen Engelska 6 i gymnasieskolan utifrån två perspektiv. För det första studeras *bedömarvariation* och *bedömarprofiler*; för det andra belyses bedömarnas *beslutsprocesser* genom att identifiera aspekter i elevprestationerna som var framträdande för dem. Ett sekundärt syfte med studien är också att göra en tentativ, empirisk jämförelse av de svenska, nationella kunskapskraven och referensnivåerna i Gemensam europeisk referensram för språk (GERS) (Skolverket, 2009). Följande forskningsfrågor utgjorde utgångspunkt för analysen:

1. Hur ser bedömarvariationen ut i de deltagande lärarnas bedömningar?
2. Vilka aspekter av elevernas muntliga förmåga är framträdande för bedömare när de fattar sina beslut om betyg?
3. Vilken är den möjliga relationen mellan betyg och bedömarnas motivering av dessa betyg?
4. Vilka nivåer i GERS anser externa bedömare att de svenska elevernas prestationer ligger på?

Studiens kontext

De nationella proven i det svenska skolsystemet har som huvudsyfte att vara ett stöd för lärare att göra likvärdiga bedömningar och därmed bidra till en rättvis betygssättning av en elevs kunskaper (Skolverket, 2019).[1] I lärarens arbete ska resultaten från de nationella proven med andra ord inte utgöra det enda underlaget för individuella betyg utan användas som stöd för betygssättning i kombination med kontinuerliga iakttagelser. De nationella proven är därmed inte examina i den traditionella bemärkelsen. I jämförelse med skolsystem med externt bedömda examensprov sätter det svenska systemet således en stor tillit till lärares professionalism vad gäller bedömning.

Skolinspektionen samlar varje år in nationella prov som genomförts av elever runt om i Sverige för ombedömning (se till exempel Skolinspektionen, 2017). Resultaten visar att samstämmigheten mellan ursprungsbedömarens och ombedömarens bedömning, den så kallade interbedömarreliabiliteten, inte håller tillräckligt hög nivå för de delar av proven som innehåller längre skrivuppgifter, till exempel uppsatserna i svenska och delvis engelska. Dock har den metod som Skolinspektionen använt kritiserats av till exempel Gustafsson & Erickson (2013) och andra studier har visat på mera positiva resultat i engelska (Erickson, 2009). Skolinspektionens omrättningar har lett till en debatt om bedömning av nationella prov och bedömarvariabilitetens betydelse för en rättvis och likvärdig bedömning. De muntliga delarna av de nationella proven har emellertid inte undersökts, eftersom dessa inte behöver dokumenteras och det följaktligen inte finns tillgängligt underlag att samla in. Detta visar på ytterligare behov av undersökningar av bedömareffekter i muntliga prov.

Gemensam europeisk referensram för språk

Arbetet med att sammanlänka de svenska kursplanerna i främmande språk och språknivåerna i Gemensam europeisk referensram

[1] En ny regel som anger att nationella provresultat särskilt ska beaktas vid betygssättning infördes i skollagen 2018. Vad ”särskilt beaktas” innebär är dock inte närmare preciserat eller kvantifierat i lagtexten.

för språk (GERS) (Skolverket, 2009) har framförallt gjorts genom textuella jämförelser, och det finns få empiriska undersökningar att tillgå. Därför är ett sekundärt syfte med denna studie att tentativt jämföra de svenska kunskapskraven i kursen Engelska 6 med GERS referensnivåer. GERS, som gavs ut av Europarådet 2001 (Council of Europe, 2001), har som huvudsyfte att ge en gemensam riktlinje för lärande, undervisning och bedömning av andra- och främmandespråk. I GERS beskrivs olika nivåer av språkfärdighet, så kallade gemensamma referensnivåer (från A1 till C2). Kurserna i engelska, moderna språk och svenska som andraspråk i det svenska skolsystemet är indelade i steg, vilka i sin tur är direkt relaterade till språknivåerna i GERS (Skolverket, 2011a). Ett godkänt resultat i gymnasiekursen Engelska 6, till exempel, ska motsvara den lägre nivån för B2 i GERS.

Studien

Metod och material

Studien har en så kallad *mixed-methods design*, vilket innebär att olika typer av data samlades in och analyserades med varierande metod av såväl kvalitativt som kvantitativt slag för att belysa forskningsfrågorna ur olika perspektiv och därmed få en större bredd. Materialet i studien består av sex ljudinspelade samtal från utprövningen av ett nationellt prov i kursen Engelska 6 på gymnasienivå, vilket alltså motsvarar tolv individuella elevprestationer. Samtalen hade valts ut för att spegla olika färdighetsnivåer. Provet syftar till att pröva muntlig produktion och interaktion i enlighet med ämnesplanen, då eleverna ska uttrycka och utveckla ett innehåll både på egen hand och i samspel med partnern.[2] Två olika bedömargrupper deltog i studien. Den första gruppen bestod av 17 gymnasielärare i engelska som bedömde elevprestationerna enligt de nationella kunskapskraven. Den andra gruppen bestod av 14 externa europeiska bedömare, som använde referensnivåerna i GERS i sin bedömning, de flesta av dessa med gedigen erfarenhet av undervisning i engelska. De senares

[2] Exempel på olika typer av uppgifter som kan förekomma i nationella proven i Engelska 6 återfinns på https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-6-gymnasiet/exempel-pa-uppgiftstyper-for-engelska-6 (hämtat 2021-02-10).

medverkan avsåg att möjliggöra en tentativ jämförelse mellan de svenska kunskapskraven och nivåerna i GERS. Data bestod dels av betyg som bedömarna satte individuellt på de tolv elevprestationerna, dels av skriftliga kommentarer som de skrev för att motivera och förklara vilka aspekter de framför allt tog hänsyn till när de satte betygen.

Deltagarna valdes ut genom en kombination av bekvämlighets- och målinriktat urval. De 17 svenska gymnasielärarna arbetade på olika gymnasieskolor, både kommunala och fristående skolor, i och runt två svenska städer. En förfrågan om att delta i studien skickades ut till rektorer på skolor i dessa två områden. Av de medverkande lärarna var fyra män och 13 kvinnor. Undervisningserfarenheten varierade från 1 till 29 år med ett medelvärde på 12 år. De europeiska deltagarna valdes ut genom ett målinriktat urval av bedömare som hade erfarenhet av GERS-bedömning. De kom från Finland (n = 7) och Spanien (n = 7) och arbetade på skolor/universitet och/eller statliga myndigheter. I dessa länder används skalor baserade på referensnivåerna i GERS i betydligt större utsträckning än i svenska sammanhang. Bedömarna använde som nämnts ovan två olika skalor. De svenska lärarna bedömde elevprestationerna enligt en tiogradig skala (se tabell 4 nedan) baserad på betygsstegen F-A, och som avsåg steg 6/kurs 6 i den svenska gymnasieskolan. De europeiska bedömarna använde en niogradig skala baserad på referensnivåer i GERS, från A1-C2, inklusive de så kallade plus-nivåerna (se tabell 5 nedan). Eftersom bedömarna använde olika skalor var syftet inte att jämföra enskilda deltagares bedömningar. För de svenska bedömarna var fokus på att undersöka bedömarvariation. Syftet med att inkludera europeiska bedömare i studien var att studera vilka nivåer i GERS som de svenska elevernas prestationer ansågs motsvara. Det som kunde jämföras var de två bedömargruppernas rangordning av elevprestationerna, eftersom denna inte är beroende av skalorna. Dessutom skrev både de svenska och de externa europeiska bedömarna skriftliga kommentarer till sina bedömningar, och dessa analyserades på samma sätt för att undersöka vilka aspekter av elevernas muntliga kommunikativa kompetens som var framträdande vid bedömningen.

Tabell 4. Tiogradig skala använd av de svenska bedömarna baserad på de svenska betygsstegen.

F–	F+	E–	E+	D–	D+	C–	C+	B	A
1	2	3	4	5	6	7	8	9	10

Tabell 5. Niogradig GERS-baserad skala använd av de externa europeiska bedömarna.

A1	A2	A2+	B1	B1+	B2	B2+	C1	C2
1	2	3	4	5	6	7	8	9

Data samlades in under endagsseminarier som var upplagda på samma sätt men som ägde rum vid olika tillfällen för de olika bedömargrupperna. Tillfället inleddes med en gemensam introduktion och ett kort övningstillfälle. Bedömarna fick sedan lyssna individuellt i hörlurar på de sex samtalen och sätta betyg på elevprestationerna samt skriva kommentarer som motiverade betyget. De hade ca 30 min till sitt förfogande per samtal. Bedömarna hade tillgång till bedömningskriterierna. För de svenska bedömarna innebar detta de nationella kunskapskraven för muntlig produktion och interaktion i kursen Engelska 6 (Skolverket, 2011b) och de analytiska bedömningsfaktorer som är avsedda att vara ett stöd för helhetsbedömningen (se bilaga 1). De externa europeiska bedömarna använde GERS-skalor för muntlig produktion och interaktion från *Manual for Relating Language Examinations to the Common European Framework of Reference for Languages* (Council of Europe, 2009, s. 184–186). Eftersom data var av två olika typer, dels betyg, dels skriftliga kommentarer till betygen, delades analysen in i två delar, en kvantitativ och en kvalitativ.

Analys av data

I de kvantitativa analyserna av betygen användes beskrivande statistik och korrelationsanalys.

Innehållet i bedömarnas skriftliga kommentarer analyserades genom ett slags innehållsanalys som kallas *verbal protocol*

analysis (VPA) (Green, 1998). VPA är en metod för att samla in och analysera data om kognitiva processer. Bedömarna ombads att verbalisera sina grunder för betyget i en sammanfattande skriftlig kommentar.[3] Dessa kommentarer delades i analysen in i mindre enheter, så kallade segment, bestående av fraser eller meningar som beskriver en övergripande aspekt som bedömarna tog hänsyn till. Ett exempel på indelning i segment, markerade med snedstreck, ges nedan. Exemplet kommer från en av de svenska bedömarna.

> *General communication skills are good/*
> *she has fluency/*
> *and structure./*
> *She listens to what the male is saying and as the conversation develops she acknowledges his thoughts and even adds her own opinion to the subject at hand. She even puts the question back to him for further discussion./*

I nästa steg utvecklads ett kodningsschema, enligt vilket segmenten kodades. Kodningskategorierna grundar sig både på teori och empiri (Galaczi, 2013), och består av huvud- och underkategorier (se kodningsschema och exempel i bilaga 2). De teorigrundade kodningskategorierna skapades utifrån beskrivningen av kommunikativ kompetens i GERS (Council of Europe, 2001), vilken både de svenska och europeiska bedömarnas betygskriterier baseras på. I GERS beskrivs kommunikativ kompetens som bestående av tre huvudkomponenter: 1) lingvistisk kompetens, 2) pragmatisk kompetens och 3) sociolingvistisk kompetens. Utöver dessa delar beskrivs kommunikativa strategier i GERS, vilka innefattar både interaktions- och produktionsstrategier. De kriteriegrundade kodningskategorierna bestod alltså av *Accuracy* och *Range*, som tillhör den lingvistiska kompetensen, *Coherence* och *Fluency*, som tillhör den pragmatiska kompetensen, och *Sociolinguistic appropriateness*, som tillhör den sociolingvistiska kompetensen. Kategorierna

[3] De svenska bedömarna fick välja om de ville skriva på svenska eller engelska, och de europeiska bedömarna skrev på engelska.

Interaction och *Production strategies* beskriver kommunikativa strategier. I de skriftliga kommentarerna fanns också uttalanden som inte direkt kunde härledas till bedömningskriterierna. Dessa var kategorierna *Intelligibility*, som handlade om hur väl det gick att förstå eleverna från bedömarens perspektiv, *Task Realisation*, som bestod av kommentarer om hur väl eleverna lyckades lösa uppgiften och *Other* med uttalanden som inte platsade in i de övriga kategorierna. Slutligen bildades en kategori kallad *Comparisons* med kommentarer som innehöll olika slags jämförelser av elevernas prestationer. Det fanns även kommentarer som omfattade bedömarnas tankar och funderingar kring olika aspekter och dessa kodades under kategorin *Rater reflections*. Utöver kodningskategorierna som beskrivits ovan gjordes även en kodning utifrån värderingen i kommentaren, det vill säga om den var positiv, negativ eller blandad.

Kvalitativ innehållsanalys bygger på kodarens tolkning av materialet, vilket kan leda till skillnader mellan kodare. Därför kodades 10 % av materialet, som sammanlagt bestod av 372 set av bedömarkommentarer (31 bedömares kommentarer till 12 elevprestationer), av ytterligare en person med lång erfarenhet inom området. Bedömaröverensstämmelsen visade sig ligga på en rimlig nivå (ca. 85 % på huvudkategorinivå). I de fall där det fanns oenighet i kodningen diskuterades orsakerna till detta noggrant och förändringar gjordes i kodningsschemat där det var lämpligt, enligt en modell som beskrivs bland annat av Galaczi (2013).

Resultat

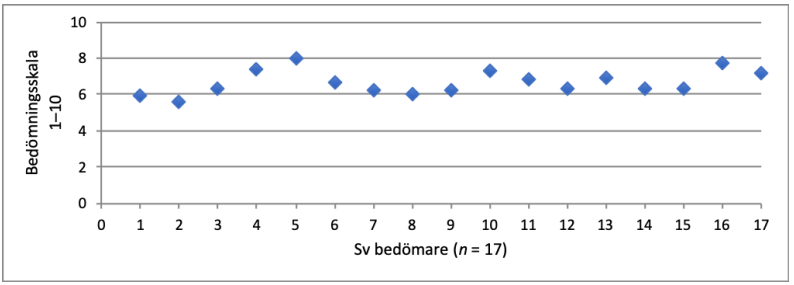
Bedömarvariation

För att besvara den första forskningsfrågan om variabilitet för de svenska lärarnas bedömningar gjordes statistiska analyser av betygen. Först presenteras deskriptiv statistik för de 17 svenska lärarnas bedömningar av de 12 elevprestationerna. Varje elev har en kod, där siffran står för elevparets nummer (1–6) och bokstaven står för kön (F = flicka, P = pojke).

Tabell 6. Deskriptiv statistik: betyg per elev (N = 12) för de svenska bedömarna (n = 17).

Elev	Medel	Median	Typvärde	SD	Variationsvidd
1F	5,9	6,0	6	1,5	(4–8)
1P	7,4	7,0	9	1,5	(5–9)
2F	9,1	9,0	9	0,8	(7–10)
2P	8,0	8,0	8	1,5	(4,5–10)
3F	4,9	5,0	3	1,7	(3–8)
3P	6,4	7,0	7	1,5	(3–9)
4F	3,4	3,0	3	1,0	(1–5)
4P	2,9	3,0	3	1,0	(1–5)
5F	9,4	9,0	9	0,5	(9–10)
5P	7,1	7,0	7	1,1	(5–9)
6F	8,2	9,0	9	1,1	(6–10)
6P	7,3	7,0	7	1,3	(4–10)

Som framgår i tabell 6 varierar medelvärde, median och typvärde mellan 3 (motsvarande E) och 9 (motsvarande B) på den tiogradiga skalan, vilket tyder på att de tolv elevprestationerna representerar en bredd av språkliga nivåer. Vidare syns en tydlig variabilitet i bedömningarna när man ser på variationsvidden. Det finns flera exempel på elevprestationer där lärarnas bedömningar går isär. Elevprestation 3F, till exempel, har en variationsvidd från E till C+ (3–8), och hög standardavvikelse, 1,7, vilket innebär stor spridning i lärarnas bedömningar. Även 3P och 6P är exempel på elevprestationer där bedömningarna går isär. För de två elevprestationerna som har fått högst betyg av bedömarna, 2F och 5F, är variabiliteten dock mindre framträdande. 2F har en variationsvidd på 7–10, och relativt låg standardavvikelse, 0,8. 5F har en variationsvidd på 9–10 och låg standardavvikelse, 0,5. Likaså har de två elevprestationerna som bedömts som de svagaste, 4F och 4M, en något lägre standardavvikelse, vilket tyder på att lärarna var mer överens om dessa betyg. Det bör dock betonas att variationsvidd är ett problematiskt mått eftersom det styrs av den högsta och den lägsta bedömningen och därmed är känsligt för så kallade *outliers*. Om man tittar på fördelningen av betyg för varje elevprestation (se Borger, 2014) framkommer att de flesta bedömarna var tämligen överens och i synnerhet



Figur 2. Medelvärden för de 17 svenska bedömarna. Anpassad från figur 8 i Borger (2014, s. 74).

var skillnaden stor mellan enstaka mycket stränga respektive milda bedömningar.

Bedömarvariationen kan också illustreras med figur 2 där en jämförelse av medelvärdena för de 17 svenska lärarnas bedömningar ges. Figuren visar att det finns tydliga bedömarprofiler bland lärarna med skillnader i stränghet. Lärare 2 är till exempel den strängaste bedömaren med ett medel på 5,6, medan lärare 5 är den mildaste med ett medel på 8,0. Att det finns bedömarprofiler med olika grad av stränghet, så kallad *rater severity*, är ett välkänt fenomen inom forskning om performance-prov. Hur man ska minimera sådana bedömareffekter är därför en viktig fråga i ett nationellt provsystem där denna typ av bedömning ingår.

För att ytterligare belysa samvariationen i bedömningarna utfördes parvisa korrelationer mellan de 17 svenska lärarnas bedömningar med hjälp av statistikprogrammet SPSS. Två olika korrelationskoefficienter användes: Spearman's rho och Kendall's tau-b. Båda koefficienterna används för icke-parametriska data och bygger på rangordning, men Kendall's tau-b är ett mer robust mått som ger lägre värden än Spearman's rho och mer korrekta p-värden vid små urval.[4] Resultaten visar att korrelationskoefficienterna mellan de parvisa bedömningarna, mätt med Spearman's rho, varierar mellan .59 och .95 ($p < .05$). Det fanns ett fåtal icke-signifikanta korrelationer från .39 till .56. Kendall's tau-b koefficienterna var som väntat något lägre och varierar mellan .47 och .89 ($p < .05$). Det fanns ett fåtal icke-signifikanta korrelationer även här (från .30 till .44). För att få en mer generell bild räknades

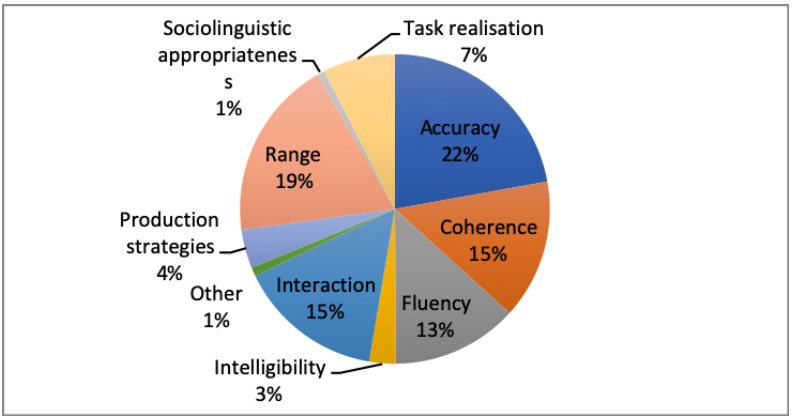
[4] Information hämtad från http://www.statisticssolutions.com/, *Kendall's Tau and Spearman's Rank Correlation Coefficient*.

medianen för de parvisa korrelationerna ut och den låg på .77 för Spearman's rho och .66 för Kendall's tau-b, vilket kan ses som relativt god samstämmighet och visar att bedömarna rangordnar elevprestationerna på ett liknande sätt. Om man kvadrerar korrelationskoefficienten får man ut andelen förklarad varians, det vill säga hur stor del av variationen i betyg som kan förutspås från en bedömare till en annan (se Borgström & Ledin, 2014). Räknat med medianen av Spearman's rho-koefficienterna kan alltså 59 % av bedömarbeteendet förutspås. Detta resultat visar att det finns utrymme för förbättring, men med tanke på att bedömarna inte genomgått någon specifik bedömarträning, vilket är vanligt i liknande forskningsstudier, är det ändå acceptabelt. Slutligen beräknades Cronbach's alpha, som mäter den interna konsistensen för de svenska lärarnas bedömningar. Den var mycket hög, .98, vilket pekar på att bedömarna uppfattar det kunskapsområde som ska bedömas, nämligen muntlig produktion och interaktion, på ett samstämmigt sätt.

Resultaten visade också att de europeiska bedömarna i genomsnitt bedömde elevprestationerna på den nivå i GERS som provet avser mäta. Medelvärdena för de europeiska bedömarna låg mellan B1+ och C1 för alla elevprestationer utom två. De två elevprestationer som bedömdes ligga under provets minimumnivå av de europeiska bedömarna hade även bedömts som icke godkända av några av de svenska bedömarna. Slutligen jämfördes rangordningen av elevprestationer mellan den svenska och europeiska gruppen och resultaten visar på stora likheter (se Borger, 2014).

Bedömningsaspekter

Bedömarnas kommentarer kodades enligt kodningsschemat som återfinns i bilaga 2 och som beskrivits ovan. I figur 3 och tabell 7 presenteras resultaten från sammanräkningen av de kodade kommentarerna. Figur 3 ger en översiktsbild av fördelningen av kodningskategorierna, medan tabell 7 visar frekvens och procent av kodade bedömningsaspekter. Exempel på kommentarer ges i redovisningen nedan för att illustrera. Det är markerat om kommentaren är positiv, negativ eller blandad och om det är en svensk (Sv) eller extern europeisk bedömare (GERS).



Figur 3. Fördelning i procent av kodade bedömningsaspekter för det totala antalet deltagande bedömare (N = 31).

Tabell 7. Frekvens och procent av kodade bedömningsaspekter

	Acc*	Coh	Flu	Inte	Inter	Other	Strat	Range	Soc.li	Task	Totalt
Sv (n = 17)											
Frekv.	385	261	157	39	219	18	78	289	18	130	1594
%	24 %	16 %	10 %	3 %	14 %	1 %	5 %	18 %	1 %	8 %	100 %
GERS (n = 14)											
Frekv.	159	97	166	29	154	5	21	174	1	55	861
%	19 %	11 %	19 %	3 %	18 %	1 %	3 %	20 %	0 %	6 %	100 %
Totalt (N = 31)											
Frekv.	544	358	323	68	373	23	99	463	19	185	2455
%	22 %	15 %	13 %	3 %	15 %	1 %	4 %	19 %	1 %	7 %	100 %

**Kategorier i följande ordning: Accuracy, Coherence, Fluency, Intelligibility, Interaction, Other, Production strategies, Range, Sociolinguistic appropriateness, Task realisation*
Anpassad från tabell 8 i Borger (2014, s. 82)

Som syns i figur 3 hänvisade bedömarna till en stor bredd av bedömningsaspekter när de motiverade sina betyg. I tabell 7 framgår även att det fanns en viss skillnad i fördelning av de olika

bedömningsaspekterna mellan de svenska och europeiska bedömarna. Till exempel kommenterade GERS-bedömarna i högre grad elevernas muntliga flyt än vad de svenska lärarna gjorde. Av utrymmesskäl kan dessa skillnader dock inte diskuteras djupare i denna text men går att läsa om i Borger (2014). Ett huvudresultat är att kategorierna *Accuracy* och *Range*, dvs. aspekter av lingvistisk kompetens, utgjorde den största andelen av de kodade bedömarkommentarerna med 22 % respektive 19 % av totalen. De lingvistiska aspekterna som bedömarna uppmärksammade i elevprestationerna handlade om språklig säkerhet i form av grammatisk (Exempel 1), fonologisk (Exempel 2) och lexikal korrekthet. I kategorin *Range* fanns kommentarer som beskrev språkets omfång och bredd vad gäller ord, uttryck och språkstrukturer (Exempel 3), men också vad gäller förmågan att uttrycka synpunkter (Exempel 4).

(1) *Grammar seems to be an improvement area; ing-form/ no ing-form is mixed at will, subject-verb agreement is a problem area. (negativ) /Sv*
(2) *Some mispronunciations, especially /z/. Other examples are "age", "students", "essay"… (negativ) /GERS*
(3) *Examples of idiomacy and variation to vocabulary. Not advanced, but extensive and varied. (blandad) /Sv*
(4) *She finds some problems to describe her point of view, but she ends up finding the way to do it without help. (blandad) /GERS*

De pragmatiska aspekterna av kommunikativ kompetens, *Coherence* och *Fluency*, utgjorde den nästa största gruppen av kommentarer med 15 % respektive 13 % av totalen. Kommentarer i kategorin *Coherence* beskriver elevens förmåga att formulera sig sammanhängande och strukturerat (Exempel 5) och med anpassning till mottagare och situation (Exempel 6). Även elevens förmåga att uttrycka sig fylligt och varierat och utveckla ett innehåll på ett tydligt sätt (Exempel 7) ingick i denna kategori. *Fluency* handlar om elevens förmåga att uttrycka sig med flyt och ledighet i språket, vilket också inbegriper kommentarer om tvekan och pauser (Exempel 8).

(5) *The parts about personal brands and smart phones were very tricky to understand. The content is not coherent. (negativ) /Sv*
(6) *Formal, well-adapted level of English mostly but also some (VERY) informal expressions (sucks, kind of) too. (blandad) /Sv*
(7) *She makes several good observations and uses examples to develop her thoughts, which moves the topics along. (positiv) /Sv*
(8) *She speaks with several pauses and hesitation, which impairs the understanding. (negativ) /Sv*

Den tredje största gruppen av bedömarkommentarer beskriver kommunikativa strategier, både interaktionsstrategier, som var den största kategorin av de två med 15 %, och produktionsstrategier, som utgjorde 4 % av det kodade materialet. Interaktionsstrategierna handlar om elevens förmåga att samarbeta, visa engagemang och bidra till den gemensamma prestationen (*co-constructed performance*) på olika sätt, till exempel genom att lyssna aktivt, ställa frågor, ge bekräftelse och bygga vidare på partnerns inlägg (Exempel 9). En annan aspekt av interaktionsstrategierna handlade om elevens förmåga att delta i turtagningen (Exempel 10). Det fanns också kommentarer om elevernas ”roller” i samtalet, speciellt i de fall då interaktionsmönstret tenderade att bli asymmetriskt, det vill säga att en av provdeltagarna var mer dominant och den andra mer passiv (Galaczi, 2008) (Exempel 11 och 12). Bedömarna var inte helt överens om hur asymmetrisk interaktion skulle påverka betyget, vilket även framkommit i tidigare studier (May, 2009, 2011). En del bedömare ansåg att den elev som tog mindre plats hjälptes av den mer pratsamme eleven medan andra tyckte att asymmetrin påverkade den passiva eleven negativt.

(9) *She moves the discussion along using questions and she also adds constructive comments and valid points when responding to her partner. (positiv) /Sv*
(10) *Takes the initiative to start and to continue conversation. (positiv) /GERS*

(11) She is helpful, a bit bossy though, dominates in interaction, explains on behalf of her counterpart, overpowers rather than collaborates with him to achieve a conversation. (blandad) /GERS

(12) During the rest – he is repeatedly interrupted by the female student, who speaks too much. It is hard to hear his full range, since he does not "fight" her verbally, he lets her takes over. (negativ). /Sv

I kategorin *Production strategies* fanns det kommentarer om elevens användning av två sorters strategier – strategier för att rätta sitt eget språk och strategier för att lösa språkliga problem genom till exempel omformulering, förklaring och förtydliganden (Exempel 13).

(13) He tries to work out any problems that may arise in the conversation, he struggles with explaining how some students might feel when they are not receiving top grades in school and he finally manages to work it out in the end. (positive) /Sv

Den sista kategorin som baseras på beskrivningen av kommunikativ kompetens i GERS är *Sociolinguistic appropriateness*. I GERS inbegriper denna komponent framför allt sociokulturella aspekter av språkanvändning, såsom "språkliga markörer i sociala relationer, artighetskonventioner, folkliga uttryck, skillnader i stilnivå, dialekt och accent" (Skolverket, 2009, s. 115). Det verkar som att det finns begränsade möjligheter för eleverna att visa upp sociokulturella aspekter av den språkliga förmågan i detta prov, och bara 1 % av kommentarerna kodades i denna kategori (Exempel 14).

(14) Använder ordet "crap" vilket inte hör hemma i sammanhanget – han ber dock om ursäkt för detta, så han är medveten om det. (blandad) /Sv

De kodningskategorier som inte explicit gick att relatera till bedömningskriterierna var *Intelligibility*, *Task realisation* och

Other (se exempel i kodningsschemat i bilaga 2). Dessa tre kategorier utgjorde sammanlagt en relativt liten del av totalen (11 %). Kommentarer kodade som *Task realisation* utgjorde den största gruppen av de tre och innehöll kommentarer om hur eleverna lyckades följa instruktionerna, framförallt hur väl de lyckades sammanfatta den korta texten som de hade läst igenom som förberedelse. Kommentarerna i kategorin *Intelligibility* handlade om hur svårt eller lätt det var för bedömaren att förstå elevens muntliga framställning. I många fall fanns det en överlapp mellan kategorin *Intelligibility* och kategorierna som handlade om fonologisk korrekthet och flyt, eftersom svårigheter med att förstå elevens uttalanden ofta grundade sig i problem med uttal eller brist på flyt och ledighet. Kategorin *Other* (1 %) var mycket liten och bestod av kommentarer som inte platsade i de övriga kategorierna.

Bland de kodade kommentarerna fanns även olika slags relativa jämförelser av elevernas prestationer, vilket även har uppmärksammats i tidigare studier om bedömning av parsamtal (Ang-Aw & Goh, 2011; May, 2011; Orr, 2002). Bedömarna gjorde jämförelser som handlade om likheter (Exempel 15), skillnader (Exempel 16) och om provdeltagarnas språkliga nivå (Exempel 17). I vissa jämförelser var elevernas individuella prestation inte helt urskiljbar då kommentaren handlade om paret som 'helhet' (Exempel 18) (se också May, 2011), vilket förstärker det faktum att samtalet är *co-constructed*.

(15) Both speakers gave the same impression: not terribly talkative, a bit shy perhaps, but positive towards each other and the texts. They didn't have much to say about the topics. / GERS

(16) Vocabulary not quite as comprehensive as the male speaker's, also simpler. /GERS

(17) She seems to me to be at about the same levels as her interlocutor. /GERS

(18) The speakers help each other well here, they give and take, ask for clarifications, examples (positive) /Sv

Den sista kodningskategorin, *Rater reflections*, omfattade bedömarnas tankar och funderingar om olika aspekter, till exempel

hur styrkor och svagheter i prestationen vägs samman till ett helhetsbetyg (Exempel 19) och hur ett visst beteende kan tolkas (Exempel 20). Det fanns även reflektioner kring hur betyget påverkas – i både positiv och negativ riktning – av provdeltagarnas roller i samtalet (Exempel 21 och 22), vilket belyser frågan om ”interlocutor effects” i parsamtal.

(19) *Interaction is high, range not so much, but fluent speaking and ok grammar takes this one to B2. /GERS*
(20) *She does not contribute much to the conversation. She keeps asking “What do you think?” as she struggles to find something to say. We cannot be sure if it’s for lack of ideas or lack of language, but I’m inclined to think it’s the latter as they’re discussing a subject which should be quite relevant to their generation and interests. Still, it’s only my perception… /GERS*
(21) *Jag tror att hon bidrar till att hennes partner får ett högre betyg än vad han har presterat tidigare för hon anpassar sitt språk och ställer bra frågor. /Sv*
(22) *I felt that this speaker was somewhat disadvantaged due to a domineering partner. I would have liked to hear more. /Sv*

Relationen mellan kommentarer och betyg

Bedömarvariationen studerades slutligen genom en tentativ jämförelse mellan bedömarnas kommentarer och betyg. Bland annat jämfördes två sorters elevprestationer: två elevprestationer (3P och 6P) med relativt hög standardavvikelse och därmed högre grad av bedömarvariation, och en elevprestation (5F) med lägre standardavvikelse och därmed högre grad av bedömaröverensstämmelse. Resultaten (se bilaga 11 i Borger, 2014) visade att bedömare som satte lågt respektive högt betyg på samma elevprestation antingen uppmärksammade samma aspekter men värderade dem olika (alltså som positivt eller negativt), eller uppmärksammade delvis olika aspekter. I fallet med elevprestationen där bedömarna i högre grad var överens om betyget visade analysen att bedömarna generellt uppmärksammade samma aspekter och kommenterade dem på ett likartat sätt.

Diskussion och didaktiska implikationer

I studien som presenteras i denna artikel undersöktes bedömarvariation och bedömares beslutsprocesser i ett muntligt nationellt prov i engelska. Den deskriptiva statistiken och korrelationsanalysen för de 17 svenska engelsklärarnas bedömningar visar på relativt god samstämmighet, men med klart utrymme för förbättring. Bedömarprofiler med olika grad av stränghet identifierades, vilket är vanligt förekommande vid performance-prov. Resultaten från de kvantitativa analyserna sammanfaller väl med tidigare studier kring bedömning av muntliga prov i främmande/andraspråk, där ofta ett större antal bedömare ingår och mer sofistikerade statistiska analyser används (Bachman et al., 1995; Bonk & Ockey, 2003; Brown, 1995; Eckes, 2005; Lynch & McNamara, 1998; Yan, 2014).

En åtgärd för att öka interbedömarreliabilitet och minska bedömareffekter är att involvera flera bedömare i bedömningsprocessen (Henning, 1996), till exempel genom att arbeta med med- och sambedömning,[5] då elevprestationer diskuteras i förhållande till kunskapskrav, elevexempel och övrigt bedömningsmaterial. Denna metod rekommenderas även av Skolverket som ett sätt att öka likvärdigheten. Enkäter som genomförs med lärare som bedömer de nationella proven i engelska i gymnasieskolan visar att med- och sambedömning används i större utsträckning vid uppsatsprovet än vid de muntliga delproven.[6] Resultaten av denna studie pekar alltså på att med- och sambedömning är viktigt även i samband med det muntliga provet, där detta troligtvis är mer tidskrävande och kräver extra resurser. Om med- och sambedömning genomförs på ett systematiskt och organiserat sätt kan det också ses som en form av *bedömarträning* (Lumley & McNamara, 1995; Weigle, 1994). Målet med bedömarträning

[5] Med- och sambedömning har något olika innebörd. Medbedömning innebär att två eller fler lärare oberoende av varandra gör en bedömning (och sedan diskuterar den), medan ”sambedömning innebär att lärare samarbetar om bedömning eller betygssättning, till exempel genom att bedöma elevers prestationer tillsammans eller genom att diskutera bedömningen” (Skolverket, 2013, s. 11).

[6] Projektet Nationella Prov i Främmande Språk vid Göteborgs universitet redovisar resultat från lärarenkäterna på sin hemsida: https://www.gu.se/nationella-prov-frammande-sprak (hämtat 2021-02-10)

är att bedömarna ska tolka elevprestationer, kunskapskrav och bedömningsanvisningar på ett så likartat och enhetligt sätt som möjligt. Forskning om performance-prov visar att bedömarträning kan ha en positiv effekt på interbedömarreliabilitet, framför allt genom att bedömare som hör till ytterligheterna vad gäller stränghet/mildhet minskar (Davis, 2016). Bedömarträning har även visat sig ha en bra effekt på intra-bedömarreliabiliteten, det vill säga att bedömare blir bättre på att bedöma elevprestationer och använda bedömningskriterierna på ett konsekvent sätt.

I denna studie undersöktes även vilka aspekter som är framträdande för bedömare när de fattar beslut om betyg. Innehållsanalysen av bedömarnas kommentarer visade att de tog hänsyn till en stor bredd av den kommunikativa språkkompetensen, vilket är i linje med beskrivningen i ämnesplanen. De lingvistiska och pragmatiska aspekterna samt elevernas interaktionsstrategier var mest framträdande. Däremot kommenterade bedömarna elevernas sociolingvistiska kompetens i mindre utsträckning, vilket kan ha att göra med att det aktuella provet inte specifikt möjliggör att visa upp sociokulturella aspekter av språkanvändning, såsom språkliga markörer i sociala relationer och artighetskonventioner. Liksom i tidigare studier utgjorde de lingvistiska aspekterna av elevernas kommunikativa förmåga den största gruppen av kommentarerna. Detta betyder dock inte nödvändigtvis att bedömarna lägger större vikt vid de lingvistiska delarna jämfört med övriga delar. Forskning pekar på att de lingvistiska aspekterna uppmärksammas i högre grad av bedömare eftersom de är lättare att ”kvantifiera” än andra aspekter av muntlig språkfärdighet, till exempel struktur och sammanhang (Fulcher, 2003). Bedömare bör alltså vara medvetna om detta och sträva efter att ta hänsyn till en så stor bredd av elevernas muntliga förmåga som möjligt i sin helhetsbedömning.

Resultaten visade också att bedömarna använde bedömningskriterierna i en mycket stor utsträckning. De kommentarer som inte hänvisar direkt till kriterierna utgjorde ca. 11 % av materialet. Dessa icke-kriterierelaterade aspekter var dock relevanta för bedömningen av provet även om de inte direkt beskrivs i kriterierna. Flera tidigare studier av bedömares beslutsprocesser i främmande/andraspråks muntliga prov visar att bedömare har en

tendens att väga in icke-kriterierelaterade aspekter i bedömningen, som ansträngning och intresse (se t.ex. Orr, 2002), vilket alltså inte var fallet i någon större utsträckning i denna studie.

Det fanns även många exempel i bedömarkommentarerna på hur eleverna använde interaktionsstrategier för att utveckla samtalet, till exempel genom turtagning, att lyssna aktivt och bygga vidare på partnerns samtalsämnen. Par- och gruppsamtal ger möjligheter för provdeltagare att visa upp sin *interaktionskompetens* (Kramsch, 2006), vilken kan ses som ytterligare en del av den kommunikativa kompetensen. Det finns dock svårigheter med att bedöma interaktionskompetens eftersom den är kontextberoende och skapas av provdeltagarna tillsammans i samtalet. Bedömarna kommenterade till exempel elevernas roll i samtalet i de fall då interaktionen var asymmetrisk. Att asymmetrisk interaktion är speciellt svår ur ett bedömningsperspektiv har även visats i tidigare studier (se till exempel May, 2011) och är alltså en fråga som kräver speciell uppmärksamhet, inte minst i bedömningsanvisningarna. En fördel i den svenska kontexten är att eleverna har fler chanser än det nationella provet att visa upp sin interaktionsförmåga, med olika samtalspartners, vilket ökar möjligheten att ge en så rättvisande bild som möjligt av elevens muntliga kommunikativa kompetens.

Ett sekundärt syfte med studien var att jämföra de nationella kunskapskraven och referensnivåerna i GERS genom att låta de deltagande europeiska bedömarna bedöma elevprestationerna från det nationella provet. Resultaten visade att de externa bedömarna i genomsnitt bedömde elevprestationerna på den nivå i GERS som provet avser mäta, dvs. B2.1. Det bör dock påpekas att detta är en liten empirisk jämförelse och att resultaten därmed är högst tentativa, men med tanke på att det finns få empiriska studier som sammanlänkar de svenska kursplanerna i främmande språk och GERS språknivåer är detta ett litet bidrag.

Slutligen visade analyserna av relationen mellan kommentarer och betyg två situationer som belyser bedömarvariation och bedömareffekter ytterligare. Dels kan bedömare uppmärksamma samma aspekter av en elevprestation men värdera dem olika, dels kan de uppmärksamma delvis olika aspekter i samma elevprestationer, och alltså basera sitt beslut på olika grunder. Ett sätt att förstärka

en gemensam syn på kvaliteter i elevprestationer och därmed bidra till att nå högre likvärdighet i bedömningen av performance-prov är som tidigare nämnts med- och sambedömning. I ett skolsystem där lärares bedömningar utgör en central del finns även ett stort behov av kompetensutveckling för att öka bedömarkompetensen – både för verksamma lärare och på lärarutbildningen.

Bilaga 1

Bedömningsfaktorer till den muntliga delen av det nationella provet för kurs Engelska 6 i gymnasieskolan*

Innehåll

- tydlighet
- fyllighet och variation
  - olika exempel och perspektiv
- sammanhang och struktur
- anpassning till syfte, mottagare, situation och genre

Språk och uttrycksförmåga

- kommunikativa strategier
  - för att utveckla och föra samtalet vidare
  - för att lösa språkliga problem genom t.ex. omformuleringar, förklaringar och förtydliganden
- flyt och ledighet
- omfång, variation, komplexitet, tydlighet och säkerhet
  - vokabulär, fraseologi och idiomatik
  - uttal och intonation
  - grammatiska strukturer
- anpassning till syfte, mottagare, situation och genre

*Faktorerna bygger på den kommunikativa språksyn som ligger till grund för ämnesplanen i engelska och moderna språk och ska ses som olika aspekter av kvaliteter i talat språk. De är avsedda att utgöra ett stöd i analysen vid en helhetsbedömning https://www.gu.se/nationella-prov-frammande-sprak/prov-och-bedomningsstod-i-engelska/engelska-6-gymnasiet/nationellt-prov-i-engelska-6 (hämtat 2021-02-10)

Bilaga 2

Kodningsschema med exempel

Huvudkategorier	Underkategorier	Exempel
Accuracy	Grammatical accuracy	Hon gör en del grammatiska misstag (that has learn to handle, homeworks) (Sv)
	Phonological control	Intonation very affected by Swedish, but speaks with ease and clarity (Sv)
	Vocabulary control	Somewhat unidiomatic at times, e.g. on your spare time, homeworks (Sv)
Coherence	Coherence and cohesion	can use a limited number of cohesive devices (GERS)
	Flexibility to circumstances	Formal, well-adapted level of English mostly but also some (VERY) informal expressions (sucks, kind of) too (Sv)
	Topic development	she finds it difficult to develop her arguments and opinions (GERS)
Fluency	Fluency mentioned in general	She manages to put her message across all along, though she's clearly finding it hard to show consistent fluency (GERS)
	Hesitation and pauses	Some hesitations in his discourse (GERS)
	Speed of delivery fast or slow	He struggles to get a word in and when he does, he is too slow for his partner so she jumps in (GERS)
Intelligibility*		Difficult to understand at times – sounds tend to become muddled, inaccuracies (nothing (to?) fat lose), pronunciation errors (e.g. [jast] i st f [dzast]), struggles to form utterances. (GERS)

Kodningsschema med exempel *(Fortsatt)*

Huvudkategorier	Underkategorier	Exempel
Interaction	Co-operating strategies	Han bekräftar det partnern säger och ställer frågor vilket driver samtalet framåt (Sv)
	Turn-taking strategies	She takes her turn when appropriate (GERS)
	Has a passive role in discussion	Let's his partner take command too often and is not as involved in the discussions as he maybe could be (Sv)
	Manages discussion (usually pos.)	Actually, she controls discussion, asks the questions, asks him for clarification of what he says (GERS)
	Dominates discussion (usually neg.)	Eleven har dock en tendens att ta över samtalet och släpper inte in sin partner i samtalet (Sv)
Other*		Seems to be enjoying the conversation (Sv)
Production strategies	Monitoring and repair	but is generally good at self-correction (GERS)
	Compensating	och han använder strategier när han inte hittar orden, han förklarar t.ex. vad han menar (Sv)
Range	General linguistic range	Examples of variation in structure and vocab, but rather ordinary (Sv)
	Vocabulary range	Her vocabulary is quite varied and expressive (Sv)
	Ability to express viewpoints	Han motiverar också det han säger och ger exempel som belyser hans ståndpunkt (Sv)
Sociolinguistic appropriateness		Använder ordet "crap" vilket inte hör hemma i sammanhanget – han ber dock om ursäkt för detta, så han är medveten om det (Sv)

Huvudkategorier	Underkategorier	Exempel
Task realisation*	Completing and understanding task requirements	Does not fully get the statements in the instructions (Sv)
	Brief or extended contributions	...and his answers are short (GERS)
	Overall comments	A very competent speaker (Sv)
Rater reflections*	Rater reflection general	This conversation runs quite smoothly all the way (Sv)
	Rater reflection about rating decision	It's mostly because of the interaction that I want to award B2 (GERS)
	Matching of candidates – how candidates perform in relation to one another	I felt that this speaker was somewhat disadvantaged due to a domineering partner (Sv)

**Empirigrundade kategorier*

Referenser

Ang-Aw, H. T., & Goh, C. C. M. (2011). Understanding discrepancies in rater judgement on national-level oral examination tasks. *RELC Journal, 42*, 31–51. https://doi.org/10.1177/0033688210390226

Bachman, L. F., Lynch, B. K., & Mason, M. (1995). Investigating variability in tasks and rater judgements in a performance test of foreign language speaking. *Language Testing, 12*, 238–257. https://doi.org/10.1177/026553229501200206

Bonk, W. J., & Ockey, G. J. (2003). A many-facet Rasch analysis of the second language group oral discussion task. *Language Testing, 20*, 89–110. https://doi.org/10.1191/0265532203lt245oa

Borger, L. (2014). *Looking beyond scores. A study of rater orientations and ratings of speaking* [Licentiatuppsats, Göteborgs universitet]. http://hdl.handle.net/2077/38158

Borgström E., & Ledin, P. (2014). Bedömarvariation: Balansen mellan teknisk och hermeneutisk rationalitet vid bedömning av skrivprov. *Språk & Stil, 24,* 133–165. http://urn.kb.se/resolve?urn=urn:nbn:se:oru:diva-42199

Brooks, L. (2009). Interacting in pairs in a test of oral proficiency: Co-constructing a better performance. *Language Testing, 26,* 341–366. https://doi.org/10.1177/0265532209104666

Brown, A. (1995). The effect of rater variables in the development of an occupation-specific language performance test. *Language Testing, 12,* 1–15. https://doi.org/10.1177/026553229501200101

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR).* Cambridge University Press.

Council of Europe. (2009). *Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) – A manual.* https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a2d

Davis, L. (2016). The influence of training and experience on rater performance in scoring spoken language. *Language Testing, 33,* 117–135. https://doi.org/10.1177/0265532215582282

Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly, 2,* 197–221. https://doi.org/10.1207/s15434311laq0203_2

Eckes, T. (2009). On common ground? How raters perceive scoring criteria in oral proficiency testing. I A. Brown & K. Hill (Red.), *Tasks and Criteria in Performance Assessment: Proceedings of the 28th Language Testing Research Colloquium* (s. 43–73). Peter Lang.

Erickson, G. (2009). *Nationella prov i engelska – en studie av bedömarsamstämmighet*. Göteborgs universitet. https://www.gu.se/nationella-prov-frammande-sprak/rapporter-och-skrifter#Studie-av-bed%C3%B6marsamst%C3%A4mmigheti-engelska-%C3%A5k-9

ffrench, A. (1999). *Study of qualitative differences between CPE individual and paired test formats* (Internal UCLES EFL report). University of Cambridge Local Examinations Syndicate.

Foot, M. C. (1999). Relaxing in pairs. *ELT Journal, 53*, 36–41. https://doi.org/10.1093/elt/53.1.36

Fulcher, G. (2003). *Testing second language speaking*. Longman.

Galaczi, E. D. (2008). Peer–peer interaction in a speaking test: The case of the *First Certificate in English* Examination. *Language Assessment Quarterly, 5*, 89–119. https://doi.org/10.1080/15434300801934702

Galaczi, E. D. (2013). Content analysis. I A. J. Kunnan (Red.), *The Companion to Language Assessment* (Vol. 3, s. 1325–1339). Wiley-Black.

Galaczi, E. D. (2014). Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics, 35*, 553–574. https://doi.org/10.1093/applin/amt017

Green, A. (1998). *Verbal protocol analysis in language testing research: A handbook*. Cambridge University Press.

Gustafsson, J.-E., & Erickson, G. (2013). To trust or not to trust?—teacher marking versus external marking of national tests. *Educational Assessment, Evaluation and Accountability, 25*, 69–87. https://doi.org/10.1007/s11092-013-9158-x

Henning, G. (1996). Accounting for nonsystematic error in performance ratings. *Language Testing, 13*, 53–61. https://doi.org/10.1177/026553229601300104

Kormos, J. (1999). Simulating conversations in oral-proficiency assessment: A conversation analysis of role plays and non-scripted interviews in language exams. *Language Testing, 16*, 163–188. https://doi.org/10.1177/026553229901600203

Kramsch, C. (2006). From communicative competence to symbolic competence. *The Modern Language Journal, 90*, 249–252. https://doi.org/10.1111/j.1540-4781.2006.00395_3.x

Lazaraton, A., & Davis, L. (2008). A microanalytic perspective on discourse, proficiency, and identity in paired oral assessment.

Language Assessment Quarterly, 5, 313–335. https://doi.org/10.1080/15434300802457513

Lumley, T., & McNamara, T. F. (1995). Rater characteristics and rater bias: Implications for training. *Language Testing, 12*, 54–71. https://doi.org/10.1177/026553229501200104

Lynch, B. K., & McNamara, T. F. (1998). Using G-theory and Many-facet Rasch measurement in the development of performance assessments of the ESL speaking skills of immigrants. *Language Testing, 15*, 158–180. https://doi.org/10.1177/026553229801500202

Magnan, S. (1988). Grammar and the ACTFL oral proficiency interview: Discussion and data. *The Modern Language Journal, 72*, 266–276. https://doi.org/10.1111/j.1540-4781.1988.tb04187.x

May, L. (2009). Co-constructed interaction in a paired speaking test: The rater's perspective. *Language Testing, 26*, 397–421. https://doi.org/10.1177/0265532209104668

May, L. (2011). Interactional competence in a paired speaking test: Features salient to raters. *Language Assessment Quarterly, 8*, 127–145. https://doi.org/10.1080/15434303.2011.565845

McNamara, T. F. (1990). Item Response Theory and the validation of an ESP test for health professionals. *Language Testing, 7*, 52–76. https://doi.org/10.1177/026553229000700105

McNamara, T. F. (1996). *Measuring Second Language Performance*. Longman.

McNamara, T. F. (1997). 'Interaction' in second language performance assessment: Whose performance? *Applied Linguistics, 18*, 446–466. https://doi.org/10.1093/applin/18.4.446

Meiron, B. E. (1998). *Rating oral proficiency tests: A triangulated study of rater thought processes*. (Unpublished master's thesis). California State University Los Angeles, Los Angeles, CA.

Nakatsuhara, F. (2013). *The co-construction of conversation in group oral tests*. Peter Lang.

Norton, J. (2005). The paired format in the Cambridge Speaking Tests. *ELT Journal, 59*, 287–297. https://doi.org/10.1093/elt/cci057

Orr, M. (2002). The FCE Speaking test: Using rater reports to help interpret test scores. *System, 30*, 143–154. https://doi.org/10.1016/S0346-251X(02)00002-7

O'Sullivan, B. (2002). Learner acquaintanceship and oral proficiency test pair-task performance. *Language Testing, 19*, 277–295. https://doi.org/10.1191/0265532202lt205oa

Skolinspektionen. (2017). *Bedömningsprocessernas betydelse för likvärdigheten: Ombedömning av nationella prov 2016. Redovisning av regeringsuppdrag*. Dnr U2014/7535/GV. https://skolinspektionen.se/globalassets/02-beslut-rapporter-stat/granskningsrapporter/regeringsrapporter/redovisning-av-regeringsuppdrag/2017/ombedomning_nationellaprov_omg8_slutgiltig.pdf

Skolverket. (2009). *Gemensam europeisk referensram för språk: Lärande, undervisning och bedömning*. https://www.skolverket.se/publikationsserier/ovrigt-material/2009/gemensam-europeisk-referensram-for-sprak-larande-undervisning-och-bedomning?id=2144

Skolverket. (2011a). *Skolverkets kommentarmaterial till ämnesplanen i engelska*. https://www.skolverket.se/download/18.6011fe501629fd150a28916/1536831518394/Kommentarmaterial_gymnasieskolan_engelska.pdf

Skolverket. (2011b). *Ämne – Engelska*. https://www.skolverket.se/undervisning/gymnasieskolan/laroplan-program-och-amnen-i-gymnasieskolan/gymnasieprogrammen/amne?url=1530314731%2Fsyllabuscw%2Fjsp%2Fsubject.htm%3FsubjectCode%3DENG%26courseCode%3DENGENG05%26lang%3Dsv%26tos%3Dgy%26p%3Dp&sv.url=12.5dfee44715d35a5cdfa92a3#anchor_ENGENG05

Skolverket. (2013). *Sambedömning i skolan – exempel och forskning*. https://www.skolverket.se/publikationsserier/stodmaterial/2014/sambedomning-i-skolan

Skolverket. (2019). *Nationella prov*. Hämtad 2019-03-07 från https://www.skolverket.se/a-o/landningssidor-a-o/nationella-prov

Weigle, S. C. (1994). Effects of training on raters of ESL compositions. *Language Testing, 11*, 197–223. https://doi.org/10.1177/026553229401100206

Yan, X. (2014). An examination of rater performance on a local oral English proficiency test: A mixed-methods approach. *Language Testing, 31*, 501–527. https://doi.org/10.1177/0265532214536171