Samstämmighet i bedömning av muntlig färdighet i andraspråk

Minna Ruohonen

Pro gradu -avhandling i svenska
Jyväskylä universitet
Institutionen för språk- och kommunikationsstudier
Våren 2017

Ett varmt tack för alla informanter för deltagande i min studie. Vidare tackar jag Mikko Kuronen för handledning och Sari Ohranen för allt samarbete i avhandlingens olika faser. Härtill tusen tack för min familj och mina vänner för uppmuntran under studiernas lopp.

JYVÄSKYLÄN YLIOPISTO

Humanistinen tiedekunta	Kieli- ja viestintätieteiden laitos
Tekijä: Minna Ruohonen	
Otsake: Samstämmighet i bedömning av muntlig färdighet i andraspråk	
Ruotsin kieli	Pro gradu -tutkielma
2017	59

Tiivistelmä

Tutkimuksen tarkoituksena oli tutkia suullisen puhetaidon arvioinnin yhteneväisyyttä ruotsin kielen puhekokeessa. Tutkimuksessa informantit arvioivat 10 puheensuoritusta, jotka oli toteutettu Yleisen kielitutkinnon (YKI) puitteissa. Puhetaidon arvioinnin yhteneväisyyden tutkiminen toteutettiin vertaamalla YKI-arvioijien ja informanttien arviointien yhtäläisyyttä. Lisäksi aineistolähtöisen sisällönanalyysin avulla määritin, miten informantit soveltavat puhetaidon kriteerejä, joita eivät entuudestaan tunne. Kirjallisista arviointiperusteluista muodostettiin teemakategoriat arvioinneille, minkä lisäksi numeerisista arvioinneista muodostettiin numeeriset vertailutaulukot.

Tutkimuksen perusteella voidaan todeta, että 61 % informanttien arvioista erosi YKI-arvioinneista. Kriteerejä kuitenkin hyödynnettiin arviointien tukena. Variaatio näkyy etenkin kirjallisesti kriteerien käytössä, mutta myös numeerisissa arvioinneissa. Arvioinneista kävi ilmi, että arvioinnissa oli käytetty myös annettujen kriteerien ulkopuolisia kriteerejä, mikä ei tue yhteneväistä arviointia. Lisäksi arvioinneissa ilmeni vertailua toisiin puhesuorituksiin, minkä ei katsota olevan kriteeriperustaisen arvioinnin kannalta korrektia.

Tutkimus osoittaakin, että useamman arvioijan arvio samasta puhesuorituksesta turvaa arvion yhdenmukaisuuden arvioinnista toiseen. Tämä on tärkeää etenkin, kun arviointi vaikuttaa esimerkiksi tulevaisuuden ura- tai koulutuspolkuun. Lisäksi tutkimusta voidaan hyödyntää esimerkiksi, kun Suomen ylioppilastutkintoon kaavaillaan ruotsin kielen suullista koetta.

Asiasanat – språkbedömning, muntlig färdighet, samstämmighet i bedömning, haloeffekt, interbedömarreliabilitet, Allmänna språkexamina, YKI-testet, YKI-kriterierna, Centret för tillämpad språkforskning, SOLKI

Säilytyspaikka – Jyx

Muita tietoja –

INNEHÅLL

1. INLEDNING

Det sägs att bedömning av muntlig färdighet är ett av de mest utmanande delområdena att bedöma eftersom det finns variation i det talade språket. Det finns olika åsikter om vad som tänks vara viktigt i talat språk, t.ex. uttal, grammatiska brister eller strukturer som används. Det kan påverka betygsättningen vilket indirekt kan inverka på framtidsplanerna, till exempel inträdet till utbildning eller karriärplaner.

Språkbedömning som tema har väckt min uppmärksamhet eftersom bedömning ska vara en stor del av mitt yrke. Som lärarpraktikant fick jag information om bedömningens validitet särskilt i klassrumsbedömning. Det framkom dock inga nya möjligheter att utvärdera muntliga kunskaper. Mest övas bedömning av skriftlig kunskap t.ex. genom att utvärdera uppsatser. Det väckte mitt intresse för att fördjupa mina kunskaper i bedömning av muntlig färdighet i svenska som andraspråk. Det är målet att skapa en helhetsbild av språkbedömningen. Vidare vill jag öka mina kunskaper om standardiserade tester som är allt vanligare och sparar tid vid bedömning av språkkunskaper.

Bedömning av muntlig färdighet medför flera utmaningar. Det finns problem t.ex. med reliabilitet och samstämmighet i bedömning eftersom utvärdering av muntlig färdighet innehåller subjektivitet när en människa gör bedömningen. Detta fenomen är vanligt även om det erbjuds konsekventa ramar i bedömningen. Jag anser att dessa utmaningar måste man vara medveten om och i framtiden är det ett stort behov att ge omfattande utbildning om bedömning av muntlig färdighet om den börjar utvärderas i studentexamen i Finland. Det betyder att kanske i framtiden måste också lärarutbildningen innehålla undervisning i bedömning av muntlig färdighet i svenska som andraspråk.

Särskilt är jag intresserad av att studera bedömning av muntlig färdighet som är ett aktuellt tema nuförtiden. Det är till exempel Studentexamensnämnden i Finland som försöker skapa ett taltest i svenska språket som skulle vara en del av svenska provet i studentexamen ungefär år 2020. I den nuvarande examen testas bara skriftlig förmåga. Nuförtiden är Allmänna språkexamina (YKI-testet) en av de få finska språkexamina som testar muntlig färdighet vid sidan av skriftlig kunskap bland annat i svenska som andraspråk. Därför använder jag som en del av min studie

detaljerade YKI-kriterierna för muntlig färdighet som mina informanter bekantar sig med och får bedöma muntliga prestationer som jag har hämtat från YKI-korpusen.

1.3. Avgränsningar

Språkbedömning är en mångsidig term eftersom språket består av olika delområden, bl.a. tal, skrift, förståelse av skrift och talat språk. Därför avgränsar jag att redogöra för teoretiska utgångspunkter i språkbedömningen i den här avhandlingen men syftet särskilt är att behandla bedömning av muntlig färdighet. Med muntlig färdighet menar jag i den här studien den produktiva förmågan som består t.ex. av talets flyt, flexibilitet, koherens, uttrycksförmågans exakthet och omfattning, behärskning av uttalet och exakthet i strukturer.

1.4. Disposition

Kapitel 2 består av teoretiska utgångspunkter för bedömningen. Först redovisas trender i utveckling av språkbedömningen. Vidare presenteras aspekter som påverkar bedömningens validitet. Därefter presenteras Allmänna språkexamina som är ett exempel på examen som använder analytisk och kriteriebaserad bedömning. De muntliga prestationerna som jag använder i min studie kommer från YKI-korpusen och därför är det relevant att förstå testets syfte, hur resultaten används och hur bedömningen är organiserad.

I avsnitt 3 presenteras metod och material. Vidare presenteras datainsamlingsprocessen. Avsnitt 4 består av resultaten i min studie. Jag svarar på forskningsfrågorna med hjälp av studies resultat. I avsnitt 5 diskuteras resultaten i förhållande till tidigare studier. Avsnitt 6 avslutar avhandlingen

2. BAKGRUND

I det här avsnittet redogörs för utgångspunkter för denna avhandling. Först presenteras aspekter i språkbedömning i allmänhet (se avsnitt 2.1). Det redovisas också möjliga orsaker till variation i bedömning av språkkunskaper i ett språk (avsnitt 2.2). Därefter presenteras Allmänna språkexamina (YKI-testet) i avsnitt 2.3 och tidigare studier i avsnitt 2.4.

2.1. Språkbedömning

I det här avsnittet presenteras utvecklingstendenser, metoder och bedömningstyper i språkbedömningen. Vidare redovisas principer och ramar i språkbedömning. Särskilt presenteras faktorer som måste kontrolleras i bedömning av muntlig färdighet så att bedömningen ska vara valid, reliabel och samstämmig.

2.1.1. Utvecklingstendenser i språkbedömning

Bedömning av språkfärdighet är ett brett begrepp. Det finns flera metoder och verktyg för bedömningen och den kan ske i olika kontexter. Vidare finns det flera syner på hur man ska definiera termen ”språkkunskap” vilket påverkar hur språkfärdigheter tolkas (se t.ex. McNamara 1996). Vidare används bedömningar för olika behov. Det kan vara t.ex. att få information om språkkunskaper för att visa kompetens som krävs för en studieplats (se. tex. Huhta & Takala 1999: 193). Alltfler forskare är intresserade av bedömningens validitet. Därför finns det t.ex. teoretiska ramar (se t.ex. Kane 2013: 3) som ger vägledning för en valid testprocess och vägledning för användning och tolkning av testresultaten.

Det finns med andra ord inte ett sätt att bedöma. Såsom Huhta & Takala (1999: 189, 202) konstaterar, är det kontexten som styr bedömningsprocessen. De understryker också att bedömningstyperna inte kan kategoriseras på ett sätt. De påpekar att bedömningsarbetet följer ofta vissa steg men det finns inte bara ett sätt att planera bedömningen. Huhta & Takala (1999: 205) konstaterar ändå att typiskt är att skilja på förståelse och produktiva kunskaper i bedömningen eftersom förståelse kan mätas bara indirekt genom svar.

Språkbedömning som fenomen har väckt uppmärksamhet i många decennier. Det har varit olika trender i språkbedömningen under olika decenniers lopp. Enligt Cohen (2004: 303–309) var 1970–80-talet betydelsefullt för undersökning av muntlig färdighet i ett andraspråk. Då började de första forskarna undersöka muntlig färdighet empiriskt. Cohen (2004: 306) förklarar ändå att fokus har legat på att beskriva talet (*speech act behavior*). Med andra ord påpekar Cohen (ibid.) att det har varit litet intresse för att utveckla verktyg att bedöma språkkunskaper.

Utvecklingen av språkbedömningen har riktat sig mot att bedöma språkkunskap som en helhet. Uppfattningen om vad som är språkkunskap har förändrats enligt Huhta & Takala (1999: 183). På 1960-talet fanns det en trend att dela språket i olika delområden t.ex. tal och skrift. På 1980–1990-talet skiftade trenden och språkkunskapen började ses som en helhet. Idén att integrera olika kunskapsområdena kom från integrativa språktest enligt Spolsky (1990: 74). Han presenterar Carrolls (1961) tanke att olika delar i språktest visar kunskaper i ett språk. Enligt Spolsky (ibid.) redovisade Carrolls (1961) att TOEFL är ett exempel av test som integrerar deltest för att bevisa kompetensen som en helhet i engelska som andraspråk.

Den kommunikativa språkundervisningen har givit sin del i definitionen av termen språkkunnighet enligt Huhta & Takala (1999: 183). Den kommunikativa modellen ser den sociala kontexten och naturlig användning av ett språk vilket har påverkat forskningens riktning. Den styrande trenden är kriteriebaserad språkbedömning som har blivit allmännare enligt Huhta & Takala (1999: 219). Keurulainen (2013: 42, 46) redovisar att meningen med kriteriebaserad bedömning är att spegla testtagarens kunskaper till kriterier som är bestämd i förväg. Då jämförs prestationen inte med ramgruppens insatser som till exempel i normativ bedömning.

2.1.2. Metoder och bedömningstyper i bedömning av muntlig färdighet

Jag koncentrerar mig på att redogöra för verktyg och metoder i utvärdering av muntlig färdighet. För att mäta muntlig färdighet finns flera metoder. Huhta & Hildén (2013: 166) redovisar att en av de första testen att bedöma muntlig färdighet var Oral proficiency interview (OPI). En utbildad bedömare utför intervjun genom vissa frågor för att kartlägga testtagarens nivå. Det finns ändå flera sätt att bedöma muntlig färdighet: intervju, studioprov, samtal med en partner,

kontinuerlig observation, språkportfolio och formativt prov (se Huhta & Hildén 2013: 169). Vidare redovisar Huhta & Takala (1999: 209–210) också rollspel eller problemlösning som en metod. I allmänhet finns det flera hjälpverktyg för att bedöma språkkunskaper. Det används t.ex. ”papper och penna tester”, muntliga och skriftliga tester och nuförtiden också datorstödda tester enligt Huhta & Takala (1999: 209, 211). De påpekar att flera samtidiga metoder är att föredra.

PTE Academic är ett språktest för kunskaper i engelska som använder automatiserad poängsättning i bedömning av muntlig och skriftlig färdighet (se PTE 2014). Testet utförs i alla delområden av språket och datorer poängsätter prestationen. Automatiseringen baserar sig bland annat på algoritmer och prestationer från tiotusentals testtagare. Enligt PTE (2014: 2) erbjuder automatiserad poängsättning objektivitet och att bedömningsresultaten kan generaliseras.

Huhta & Hildén (2013: 170) redovisar Dialang som ett exempel på ett elektroniskt batteri av tester. Dess mål är att erbjuda ett test med låg tröskel för vem som helst att få en uppfattning om sina kunskaper i ett språk. Det är inte ännu vanligt att använda datorstödd språkbedömning. Det finns ändå ett behov av datorstödd bedömning av muntlig färdighet i framtiden t.ex. i finska studentexamen.

Det finns flera utmaningar i bedömning. Huhta & Hildén (2013: 163) presenterar t.ex. metodeffekten då metoden kan påverka resultaten. Vidare finns det utmaning t.ex. med samtalspartner och bedömningsskalan. Det är också bedömaren som kan påverka reliabiliteten av en bedömning. Det beror på hur släpphänt eller erfaren han är, hur han tolkar kriterier och vilken typ av uppfattning han har om språkkunskap. Vidare påpekar Huhta & Takala (1999: 210) att det är dyrt och tidskrävande att testa och bedöma muntlig färdighet. De poängterar ändå att ett muntligt test i en studio kan spelas in och bedömningen kan ske flexibelt.

Det finns flera bedömningstyper. Det beror på kontexten. Ahola (2012: 55–56) redovisar att ett sätt att dela in bedömningen är t.ex. klassrumsbedömning (fi. *jatkuva arviointi*) och testbedömning (fi. *testiarviointi*). Klassrumsbedömningen sker ofta normbaserad och är dynamiskt. Testbedömningen sker kriteriebaserad och visar kunskaper vid det tillfället. Skillnaden i bedömningen är också hur testresultaten tolkas. På samma sätt delar Huhta & Takala (1999: 189–190) in bedömning i undervisning och bedömning efter kunskapsnivån (jfr Ahola 2012). De förklarar

att båda bedömningstyper finns t.ex. i bedömning av finska studentexamen. De påpekar ändå att den största skillnaden mellan dessa två bedömningstyper är att i kunskapsnivåbaserad bedömning är man mest intresserad av språkliga kunskaper. Då är man inte så intresserad av hur kunskaperna har skaffats.

Klassrumsbedömningen kan ske diagnostiskt, formativt eller summativt (se t.ex. Huhta & Takala 1999: 195–199; Ahola 2012: 57). I diagnostisk bedömning (se Jakku-Sihvonen 2013: 20) kartläggs elevens kunskaper vid en viss tidpunkt. Det visar kunskaper som eleven redan har uppnått men visar också brister i kunskaper. I formativ bedömning (se Keurulainen 2013: 38) bedöms eleven under loppet av en studiehelhet. I summativ bedömning (se ibid.) menas bedömning som sker i slutet av studiehelhet. I diagnostisk, formativ och summativ bedömning är bedömaren den undervisande läraren som kan utvärdera språkkunskaper i flera situationer under en lång tidsperiod. Utvärderingen baserar på GERS (2009).

Testbedömning baserar sig bara på kompetensnivåskalan (se t.ex. Ahola 2012: 58). Bedömningen av kunskaperna kan vara holistisk (*holistic*) eller analytisk (*analytic*) (se. t.ex. McNamara 2015: 43–44). McNamara (ibid.) förklarar att en holistisk bedömning ger en helhetsbild av språkprestationen. Analytisk bedömning består i sin tur av separata bedömningar av olika kompetenser i en prestation. Ahola (2012: 58–59) exemplifierar att i YKI-testet används analytisk bedömning. Då är det ett standardiserat test grunden för testbedömning. Som ett exempel på normbaserade test presenterar Ahola (2012: 55–56) finska studentskrivningar och poängterar att den typen av bedömning sker ofta i klassrumskontext. Vidare redovisar hon att YKI-testet i Finland är ett kriteribaserat standardiserat test som ett exempel på testbaserad bedömning.

En skillnad mellan klassrumsbedömning och testbedömning är att testbedömningen ofta baserar sig på ett strikt testbatteri (se t.ex. taltest i YKI-testet) som utförs i studion eller i en annan kontrollerad testsituation. Bedömningen i klassrummet kan ändå ske dynamiskt under en lång tidsperiod (se t.ex. Huhta & Takala 1999: 189). I standardiserade test kan bedömaren inte vara den undervisande läraren utan bedömaren måste vara någon som inte känner eller har bedömt testtagaren förut (se t.ex. lag om allmänna språkexamina 964/2004). Bedömaren i klassrummet kan tvärtom känna eleven/studenten förut och kan observera muntlig färdighet i flera situationer. Testet kan också lätt delas ut för en massa elever utan att bedömaren måste möta testtagaren ansikte mot ansikte. Testet kan ändå används som en del av klassrumsbedömningen.

Kritik mot testbedömning och klassrumsbedömning är att bedömaren eller testplaneraren är en människa med subjektiva tolkningar (se t.ex. Ahola 2012: 57). Även om det används kriterier som bas i bedömningen finns det studier som visar att bedömaren kan bygga sina bedömningar på subjektiva tolkningar och åsikter. Det här fenomenet talas både i bedömning av förstaspråk och andraspråk (se t.ex. Davies m.fl. 1999; McNamara 1996). Ofta tycks klassrumsbedömningen vara mycket subjektiv (se. t.ex. Ahola 2012: 59). Men som Tarnanen (2007: 14) presenterar kritiseras det kriteriebaserad test som t.ex. YKI-testet om dess autenticitet. Man får inte tala med en kamrat utan talet spelas in i bandet. Det finns intervju, men bara om man vill visa kunskaper i högre nivå (se avsnitt 2.3.1).

Allt som allt kan konstateras att i språkbedömningen måste bedömaren ta ett stort ansvar för validitet och konsekvenser eftersom bedömningen innebär makt (se Huhta & Takala 1999: 180, 222). Det påverkar till exempel konkurrensen av studie- och arbetsplatser som kan ses t.ex. i Finland. Det kan hända att beroende av bedömaren kan det vara att testtagaren får godkänt när en annan bedömare ger icke-godkänt eller sämre betyg för en och samma prestation. Bedömningen kan alltså påverka hur mycket poäng t.ex. en testtagare får i inträdesprovet. Vidare kan kandidater i en arbetsplatsintervju ordnas genom språkkompetens om det inte finns andra faktorer som skiljer kandidaterna åt.

2.1.3. Principer i språkbedömning

I det här avsnittet presenteras principer för samstämmig språkbedömning i språktest och -bedömning. Vidare redovisar jag ramar för språkbedömning. Särskilt koncentrerar jag mig på att förklara GERS-nivåskalans roll i språkbedömning.

Det finns allmänt accepterade principer i språkbedömning. Inom ramen av denna avhandling redogörs för testets validitet, autenticitet och reliabilitet eftersom testet kan vara ofta ett verktyg att ta reda på människans kunskaper i ett språk. Vidare redogörs det för bedömningens etik och interbedömarreliabilitet.

Den första principen är validitet. Validitet i ett test definieras eftersom testet kan vara ett verktyg i språkbedömningen. Davies m.fl. (1999: 221) redovisar att testet är valitt om det testar egenskaper som man har planerat att testa.

Autenticitet är en annan faktor i ett test som definieras för att förstå varför det är bättre att använda autentiskt material i ett test för att få sådana resultat som är valida och visar testtagarens språkkunskaper. Davies m. fl. (1999: 13) förklarar att ett autentiskt test testar precis sådana kunskaper som behövs i livet. Autenticitet kan man nå i ett test genom att använda material som har samlats in från det verkliga livet i stället för att skriva om eller hitta på situationer i testet.

Den tredje komponenten i ett språktest som stödjer att testet fungerar bra är reliabilitet. Davies m.fl. (1999: 168) redovisar att reliabiliteten betyder att testresultaten är samma i olika tid och rum. Med andra ord måste man ha samma resultat från testet även om testaren eller bedömaren byts ut.

I den här avhandlingen är det relevanta att definiera termen interbedömarreliabilitet eftersom jag undersöker likvärdigt bedömning av muntlig färdighet i ett andraspråk och har valt att undersöka hur likvärdigt olika bedömare har bedömt samma talprestationer. På så sätt försöks visa kvalitet och kvantitet av subjektivitet i bedömningen.

Davies m.fl. (1999: 88) definierar att interbedömarreliabilitet betyder konsensusnivån om två eller flera bedömare som har bedömt testtagarens prestation. De redovisar att det här fenomenet oroar ofta i bedömning av kunskaper i skrift och tal eftersom de är subjektivt bedömda. Ofta kan man ha problem i mellannivå om man jämför bedömningar av två eller flera bedömare. Det kan t.ex. hända att en bedömare kan ge vitsordet som i jämförelse med andra bedömare är mer strängt. Därför poängterar Davies m.fl. (ibid.) att i ett test som poängsätts subjektivt skulle vara pålitligare att två eller flera bedömare bedömer prestationen. De exemplifierar åtminstone high stakes-test vara sådant test. Med high stakes-test menas enligt Davies m.fl. (1999: 185) sådant test som influerar testtagarens framtid t.ex. karriärplaner.

Bedömningens etik är relevant att behandla inom ramen av denna avhandling eftersom informanterna i min studie är sådana bedömare som också måste vara medvetna om gemensamma verksamhetsmodeller. Davies m.fl. (1999: 55) förklarar att etiken i språkbedömning betyder att

bedömningen sker på basis av standarder eller regler som man har konsensus. En språktestare utbildas ofta för att ta ansvar för olika test och normer som de använder. Davies m.fl. (ibid) understryker att etisk testande syns på det sättet att man måste värdera t.ex. om något annat bedömningssätt skulle ge likvärdigare resultat för en testtagare.

2.1.4. Ramar i språkbedömning

Språkbedömningen följer vissa ramar. I Europa finns det ett gemensamt styrdokument för språkbedömningen som heter gemensam europeisk referensram för språk (GERS 2009). Vidare styr lagstiftning vissa principer som lärare och myndigheter måste följa i bedömningsarbetet i Finland. Därtill finns det vissa grunder som måste följas beroende på kontexten. Inom ramen av denna avhandling presenteras i den här delen kort GERS. Andra dokument som är relevanta att beakta i den här studien är lag om allmänna språkexamina (964/2004) och grunderna för allmänna språkexamina (Utbildningsstyrelsen 2011) som talas om i avsnitt 2.3.2.

GERS är ett styrdokument som alla medlemsländerna i EU tillämpar i språkbedömningen (se GERS 2009; Ruohonen 2016). Referensramens roll är att ge gemensamma ramar för att bedöma språkfärdighet med objektiva kriterier. Vidare underlättar GERS samverkan mellan de europeiska länderna. GERS innehåller också nivåskalan i språkundervisning som definierar nivåer som beskriver mål och språkfärdigheter som inläraren väntas nå Eftersom nivåskalan är en gemensam ram för alla bedömare möjliggör det att kunskaper i ett språk kan jämförs med olika talare i olika europeiska länder. Det kan alltså påstås att GERS är ett verktyg att förbättra likvärdigheten av olika betyg eftersom bedömningen ska ske genom riktlinjer som man har konsensus med i alla medlemsländer.

2.2.Variation i språkbedömning

Den här översikten sammanfattar i ett nötskal varför det finns variation i bedömning av språkkunskap, särskilt i bedömning av muntlig färdighet. Det finns allmänna principer som förklarar till viss grad varför bedömningen av ett test kan variera. Därför redogörs för olika aspekter i bedömningskedjan som kan orsaka variation. Inom ramen av denna avhandling koncentrerar jag mig på att redogöra för variation i bedömning av muntlig färdighet.

Som bedömning i alla andra delområden i språket, t ex. skrift och förståelse, har bedömning av muntlig färdighet samma grunder som styr bedömningen. Det tycks ändå att bedömning av muntlig färdighet är en av de mest utmanande uppgifterna eftersom det finns variation i uttal och åsikter om målspråkligt tal varierar. Vidare hänvisar Isaacs (2016: 6) till Lundebergs (1929: 195) tanke att muntlig färdighet inte är så mätbar i jämförelse med skriftlig kunskap eftersom utvärdering av muntlig prestation innehåller bland annat mer variation och är tidskrävande. Subjektivitet tänks vara problemet i bedömningen även om bedömningen baserar sig på objektiva kriterier. Det finns också olika sätt att bedöma språkkunskaper. Det används t.ex. intervju (se t.ex. Brown 2004: 255) men också standardiserad test i studion (se t.ex. Allmänna språkexamina i avsnitt 2.3).

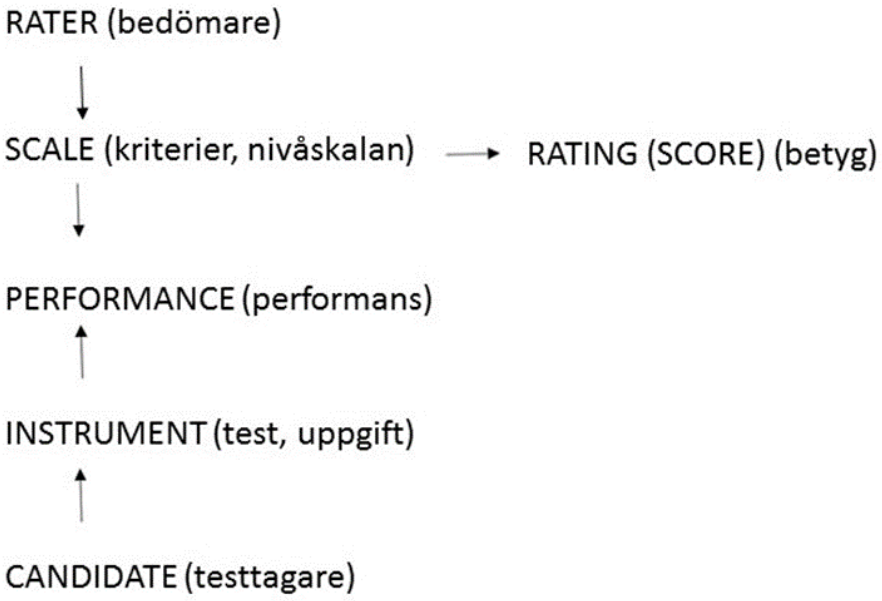
Samstämmighet i bedömning är målet i all bedömning men det finns problematik som påverkar ofta reliabiliteten i språkbedömningen. Utmaningar finns i varje nivå i bedömningskedjan (se figur 1). McNamara (2015: 36–38) påminner att bedömningen alltid innehåller risken att andra faktorer kan påverka bedömningen. McNamara (ibid.) förklarar också att man hellre undviker bedömarcentrisk testbedömning på 1950–1960-talet eftersom man visste att det finns mycket subjektivitet. Men den kommunikativa språkundervisningen pressade att bedömningsprocessen måste bättre förstås.

Som McNamara (2015: 36) redovisar finns det tre aspekter i bedömningsmetoder på att rimligt försäkra pålitlighet vid bedömningen. Den första aspekten är konsensus om gällande konditioner i omständigheterna då prestationen sker. Konsensus kan gälla t.ex. testomständigheter. Den andra aspekten är sammanhängande om olika egenskaper i en prestation. Bedömare måste alltså fastställa t.ex. kriterier för bedömningen. Den tredje aspekten är att bedömare karaktäriserar testtagarens kunskaper med ett betyg eller klass kunskaper att tolka kriterier. Då behöver man deskriptiva kategorier.

Det finns två sätt att bedöma människans kunskaper i språk och det kan ske genom fast respons (*fixed response*) bedömning eller prestationsbedömning (*performance assessment*). Med fast respons test menas att poängsättningen dras direkt från instrumentet. Det betyder att testtagarens svar ger bevis om hans kunskaper (se McNamara 1996: 120–121). I prestationsbedömning utvärderar bedömaren testtagarens prestation vilket indirekt indikerar kunskaper i ett språk. Bedömning baserar sig t.ex. på bedömares observation (se McNamara 1996: 120–121). Det vill

säga att prestationsbedömningen sker genom en bedömare som tolkar från testtagarens prestation de kunskaper som mäts i ett test.

McNamara (1996: 9) redovisar sex komponenter i prestationsbedömningen som visas i figur 1. *Rater* i modellen markerar bedömare som med hjälp av *scale* ger *rating*. *Scale* i modellen beskriver bedömares verktyg som är till exempel kriterier och nivåskalan. Med *performance* menar man prestationen som utvärderas och som indikerar kunskaper i ett språk. *Instrument* i modellen betyder t.ex. test eller uppgift som styr prestationen. *Candidate* betyder kandidaten som tar del i testet. McNamaras (1996: 9) modell för prestationsbedömning (*performance assessment*) fungerar som ram när jag redogör för faktorer i samstämmig bedömning.



Figur 1 Prestationsbedömning efter McNamara (1996: 9) med översättning

Inom ramen av denna avhandling redogörs för faktorer som påverkar bedömningens kvalitet och orsakar variation i bedömningen. Nivåer i bedömningsprocessen efter McNamaras (1996: 9) modell delas in i tre grupper: bedömarnivå (*rater, scale*), testnivå (*instrument, performance*) och testtagarnivå (*candidate*).

2.2.1. Variation i bedömarnivå

I bedömarnivån är det subjektivitet som orsakar problem. Olikvärdig bedömning i bedömarnivå sker ofta genom subjektivitet. Subjektivitet innehåller olika aspekter. Enligt Brown (2004: 255)

kan bedömares fördomar, t.ex. bedömares subjektiva tolkningar eller attityd, påverka bedömningsresultaten. I psykologin talas om haloeffekt. Davies m.fl. (1999: 72) definierar att i haloeffekten en bedömare kan bygga sin bedömning också på t.ex. första intryck eller kritik mot testtagarens egenskaper som bedömare har haft om testtagaren.

Davies m.fl. (1999: 72) omnämner att särskilt när man bedömer ett taltest genom vissa kriterier finns det risken att bedömning av en kategori kan påverka en annan. Det finns t.ex. kriterier som talets flyt eller lämplighet. Då kan bedömaren först bedöma flytet och den bedömningen kan också ha inverkan i bedömningen av lämpligheten. För att kontrollera haloeffekt föreslår Davies m.fl. (ibid.) att bedömningsskalan används på så sätt att varje del i testet bedöms som självständiga delar eller bedömaren får inte veta den första delens bedömning.

I talkontext finns det t.ex. följande problem. Isaacs (2016: 6) redogör för Woods (1927) tanke att variation i bedömning av muntlig färdighet kan bero på att bedömare koncentrerar sig på antingen innehållet av deltagares produktion i ett muntligt test eller bedömare är intresserad av egenskaper i deltagares muntliga produktion. Isaacs (ibid.) påminner ändå att forskare (se t.ex. Bachman & Palmer 1996) ser att dessa faktorer inte får påverka poängsättning i ett test.

McNamara (2015: 37) konstaterar att det finns variation t.ex. i bedömningar mellan bedömare även när vi talar om erfarna bedömare. Alderson (1991: 23) presenterar forskningsresultat som visar att det finns risk att även erfaren bedömare har inte konsensus t.ex. om innehållet eller indikatorer i ett test. När testet planeras är ett språktest ofta en version som baserar sig på en definition av språkkunnighet enligt Huhta & Takala (1999: 182). Om bedömaren inte är konsistent i jämförelse med andra bedömare i sina utvärderingar då är det problem med interbedömarreliabilitet (se Davies m.fl. 1999: 88 och avsnitt 2.1.3). McNamara (2015: 37) påminner att det finns på samma sätt variation mellan bedömare så som mellan kandidater.

McNamara (2015: 38) redovisar att det finns möjlighet att bedömningar av en bedömare inte heller är konsistent. Enligt Davies m.fl. (1999: 91) är det då problem med intrabedömarreliabilitet. Det betyder att bedömares bedömning av samma prestation varierar. Med andra ord håller alla bedömningar en konsistent linje som inte är beroende av tid och plats om man har en bra intrabedömarreliabilitet.

2.2.2. Variation i testnivå

I testnivån kontrolleras resultatens pålitlighet i ett test genom en försiktig planering. Innan testet används finns det t.ex. regler och instruktioner hur testet ska planeras. Om språktestandet finns litteratur som fokuserar på planering och bedömning av ett språktest (se. t.ex. Bachman & Palmer 1996, McNamara 2015, Spolsky 1989). Som redan sagt måste testet vara valitt (se avsnitt 2.1.3). Även om testet är välplanerat finns det ändå risk att testet t.ex. inte mäter det som det är planerat för.

Bachman & Palmer (2004: 21) redovisar att strukturvaliditet måste kontrolleras. Med struktur validitet menar Bachman & Palmer (ibid.) att det finns till viss mån möjlighet att tolka testresultat genom indikatorer som tycks visa kunskaper i ett visst kunskapsområde. Med andra ord måste indikatorer vara sådana som visar reella kunskaper t.ex. i tal. Om struktur validitet inte är pålitlig är generalisering av testresultaten osäker enligt Bachman & Palmer (ibid.). Alderson (1991: 23) påminner att testet måste säkerställa på förhand så att testet fungerar i själva testsituationen.

I testnivån är vi medvetna också om testmetodeffekt. Davies m.fl. (1999: 203–204) redogör för att en metod kan påverka insikten som man får om testtagarens kunskaper. Vidare kan andra faktorer inverka på testtagarens prestation. Enligt Davies m.fl. (ibid.) kan det vara fysisk testsituation, provets tidpunkt eller hurdana svar man hoppas få. Bland annat kan testtyp styra hur kandidaten svarar på testet eller det finns vissa förväntningar om testtagarens kunskaper innan man tar testet.

2.2.3. Variation i testtagarnivå

I testtagarnivå finns det några variabler som kan påverka bedömningen. Cohen (2004: 320) konstaterar att sådana variabler som kan påverka prestationen är till exempel testtagarens socioekonomiska bakgrund eller om testtagaren är introvert eller extrovert. Vidare finns det möjlighet att variation i resultaten av olika kandidater i ett samma test kan komma t.ex. från skillnaden mellan utbildning eller erfarenhet om språket. Ahola m.fl. (2016) redogör för resultaten som visar att t.ex. vuxna som hade invandrarbakgrund tog del i finska provet och resultaten mellan

testtagarna varierade beroende av grundutbildning. Men samma studie också visar att vistelsetiden i Finland var inte en säker markör att behärska ett språk utan man alltid behöver kontakter att utveckla sina kunskaper i ett språk. Men dessa faktorer inte indikerar att testet inte är valitt.

2.2.4. Sammanfattning

Som McNamara (2015: 44), Alderson (1991: 21) och Davies m.fl. (1999: 161) konstaterar är en konstant utbildning av bedömare ett sätt att förbättra bedömningens likvärdighet. I Finland utbildas t.ex. grundskolelärare genom ”veso-dagar” (se t.ex. Korkeakivi 2014: 27) där man får utbildning om bland annat bedömning. Vidare utbildar t.ex. Centret för tillämpad språkforskning (SOLKI) YKI-bedömarna för bedömningsarbetet (se avsnitt 2.3).

För att dela ut information om språkbedömning och språktest så att testbedömare skulle ha de bästa färdigheter att tolka testresultaten för att bedöma språkfärdigheter finns det till exempel internationella organisationer som ALTE (*Association of language testers in Europe*) och EALTA (*European association for language testing assessment*) i Europa som delar ut information om riktlinjer i bedömningen. I ALTEs webbsida redovisas att den sätter bland annat standarder för språkbedömning. EALTAs webbsidor presenterar att den delar ut test- och bedömningspraktiker i Europa och informerar om de teoretiska principerna i språkbedömningen och -testandet.

Som sagt är det också viktigt att kontrollera testplaneringen för att nå valida resultat. Dessutom måste man ta hänsyn till att det finns variabler också i testtagaren som kan påverka resultaten och i slutet bedömningen av språkkunskaper. I testkontext kommer det att vara utmaningar. Brown (2004: 306) påminner att talhandling studerades på 1980-talet och det fanns på 1990-talet en trend att undersöka hur beskrivs beteendet i talsituationer i stället för att utveckla verktyg för bedömning av muntlig färdighet.

Man behöver ändå en människa som bedömare i testbedömningen även om det finns risken för subjektivitet i bedömningen. Alderson (1991: 22–24) redovisar att bedömaren ofta gör beslut om testets innehåll så att det kommer att visa kandidatens aktuella språkkunskaper.

2.3. Allmänna språkexamina

I det här avsnittet redogörs utgångspunkter för Allmänna språkexamina (YKI-testet). Enligt Ahola (2012: 9) är YKI-testet en standardiserad examen i Finland och SOLKI är ansvarig för anordnandet av den. Enligt SOLKI (2017a) är examens syfte att testa och bedöma vuxnas språkfärdigheter. Examen består av fyra delprov som mäter produktiva och kommunikativa färdigheter. Delproven är textförståelse, skriftlig framställning, talförståelse och muntlig framställning (se SOLKI 2011a). Provet är frivilligt och det kan tas i nio språk, bl.a. svenska, engelska och finska. Det finns tre nivåer (grundnivå, mellannivå och högsta nivå) och deltagaren bestämmer själv vilken nivå hen tar provet. Nivåskalan är sexgradig, bedömningen är kriteriebaserad och kriterierna motsvarar GERS-nivåskalan. Enligt Ahola & Hirvelä (2016) är YKI-testet ett high stakes-test som oftast avläggs när man söker medborgarskap eller ska visa sina språkliga kunskaper till arbetsgivare.

2.3.1. Testet för muntlig framställning i YKI-testet

Enligt SOLKI (2011b) varierar testet för muntlig framställning beroende på grundnivå, mellannivå och högre nivå. I allmänhet tar provet 20 minuter och innehåller 3 olika uppgifter. I högre nivå utförs också en intervju. I allmänhet är situationerna vardagliga och teman i dessa talsituationer är bland annat livsmiljön, fritiden och arbetslivet. Oftast är situationer informella och det viktigaste målet är att testtagaren kan argumentera och variera sitt tal (se t.ex. Utbildningsstyrelsen 2011: 9).

I alla nivåer spelas muntliga prestationen in i språkstudion för att bedöma dem senare. I bedömningen kontrolleras vissa kunskaper. Till exempel i mellannivå bedöms t.ex. kunskaper som förståelig kommunikation enligt instruktioner och kunnighet att kommunicera i vardagliga situationer som situationen kräver. Vidare bedöms hur talaren uttrycker sin åsikt och motiverar den.

Enligt SOLKI (2011b) finns det tre uppdrag i testet för muntlig framställning. Jag avgränsar att redogöra för testtyperna i mellannivå. Testtyperna är följande: 1. samtalsuppgift (*keskustelutehtävä*), 2. situationell uppdrag (*tilannetehtävä*) och 3. berättande uppgift (*kertomistehtävä*). Uppgift 1 innehåller 3–6 anförande som spelas på bandet och kandidaten reagerar på dem. I uppgift 2 reagerar kandidaten till 6–8 situationer snabbt enligt tipsar som kandidaten får

på förhand. I den tredje uppgiften förbereder kandidaten sig 1–2 minuter att tala om ett valt ämne för 1–2 minuter.

Som Ahola m.fl. (2015) redovisar är testövningar i YKI-testet standardiserade och kalibreras. Övningar kan används alltså på nytt och det är ett sätt att försäkra konsekvens i bedömningar mellan olika testtillfälle. Det är alltså trenden att övningar inte är generellt tillgängliga (se t.ex. SOLKI:s webbsidor). I den här studien kan jag därför inte definiera uppdraget i testövningar som testdeltagarna har haft. Det finns ändå information om utformning av testet för muntlig framställning (se Ahola, 2012: 132–165)

2.3.2. Bedömning av muntlig färdighet i YKI-testet

Bedömarna i YKI-testet är utbildad av SOLKI. Enligt lag om allmänna språkexamina (964/2004) får YKI-testet ”bedömas endast av personer som har sådan behörighet som föreskrivs genom förordning av statsrådet”. Utbildningsstyrelsen är ansvarig för YKI-testet och i samband med kommissionen för Allmänna språkexamina utvecklar examina. Vidare styr det register över bedömare. På det sättet försäkras enhetlig nivå på bedömningen.

Ramar som styr bedömning av talprestationer i YKI-testet är t.ex. GERS (2009). YKI-kriterierna motsvarar GERS-nivåskalan (se avsnitt 2.3.2). Vidare ger lag om allmänna språkexamina (964/2004) riktlinjer bland annat för undervisning och bedömning i YKI-testet. Vidare redovisar grunderna för allmänna språkexamina t.ex. examens innehåll och syftet med examen (se Utbildningsstyrelsen 2011: 9).

2.3.3. Allmänna kriterier i YKI-testet

De muntliga prestationerna som jag har valt i min studie ligger på mellannivå (nivå 3 och 4). Nivåerna motsvarar kunskapsnivåer B1 och B2 i GERS-nivåskalan (2009). Det finns allmänna och nivåfast (grund, mellan och högre) kriterier för språkkunskaper. De allmänna kriterierna är skildringar för att beskriva språkkunskaperna som en helhet (se t.ex. Utbildningsstyrelsen 2011). I den här avhandlingen fokus är på kriterier i mellannivå (nivå 3 och 4). Nedan redogörs för allmänna YKI-kriterierna för nivå 4:

> *Förstår tal i normalt tempo om allmänna ämnen, men vissa detaljer kan bereda svårigheter. Snabbt talspråk och dialektalt språk kan vara svårt att förstå. Förstår*

> *utan svårigheter texter som handlar om allmänna företeelser, även om vissa nyansskillnader i texten kan gå förlorade. Reder sig rätt bra i olika såväl officiella som inofficiella talsituationer. Kan skriva både private och delvis officiella texter och framställa sina tankar så att de bildar ett logiskt sammanhang.*
>
> -Utbildningsstyrelsen 2011

Därefter redogörs för allmänna YKI-kriterierna för nivå 3:

> *3: Förstår längre talsekvenser och det centrala innehållet i många TV-och radioprogram om ämnet är bekant och taltempot normalt. Förstår vanliga texter om vardagliga företeelser, men krävande texter om ämnen som är främmande kan vålla svårigheter. Reder sig i vanliga praktiska talsituationer och kan skriva enkla sammanhängande texter om vanliga företeelser, även om brister i grammatik och ordförråd ibland kan vålla förståelseproblem.*
>
> -Utbildningsstyrelsen 2011

2.3.4. Kriterier för muntlig färdighet i YKI-testet

Jag koncentrerar mig på att redogöra för bedömningskriterier för muntlig färdighet eftersom informanterna i min studie använder dem med hjälp av bedömningen. I YKI-testet används följande kriterier i bedömning av muntlig färdighet: talets flyt, flexibilitet, koherens, uttrycksförmågans exakthet och omfattning, behärskning av uttalet och exakthet i strukturer. YKI-kriterierna är mer detaljerade beskrivna i bilagan (se bilaga 1).

Eftersom bedömning av muntlig färdighet består av en helhet, inte bara ett kriterium (t.ex. koherens) redovisas också allmänna YKI-kriterierna för tal i kunskapsnivåer 3 och 4. I nivå 3 klarar testtagaren sig i de vanligaste praktiska talsituationerna och tar initiativ i dagliga språkanvändningssituationer. Tal kan vara långsamt men det förekommer inte onaturliga pauser speciellt ofta. Talaren blir förstått trots att hen har transfer från sitt modersmål eller andraspråk i strukturer. Uttalet kan vara icke-målspråkligt (SOLKI 2017b). I nivå 4 klarar testtagaren sig ganska bra också i främmande kommunikationssituationer. Hen kan avskilja formell och informell talform i sitt tal i viss mån. Talaren kan presentera och motivera sina åsikter förståeligt.

Vidare kan man berätta och beskriva vad hen har sett, hört och upplevt. Talaren måste sällan använda omskrivningar i vardagliga talsituationer på grund av hens brister i språkkunskaper (SOLKI 2017b).

Det är viktigt att notera att testtagaren kan själv bestämma att vilken testnivå hen tar provet (se SOLKI 2017a). Det betyder att det finns en möjlighet att testtagaren ligger på någon annan nivå en den som hen har tagit testet. Till exempel kan testtagaren ligga på högre nivå än hen har tänkt. Men i situationen då testtagaren visar att sina kunskaper skulle passa till högre nivå, får hen ändå vitsord i testnivån hen har anmält sig. Det finns också möjlighet att testtagaren kan visa sämre kunskaper och fyller inte kraven i testet på den nivå som hen har anmält sig. Enligt SOLKI (ibid.) kommenterar bedömaren att testet indikerar att testtagaren hör till lägre nivå än vad hen har anmält sig (t.ex. >3 eller <3).

2.4. Tidigare studier

Det finns litteratur som redogör mycket för teoretiska ramar för språktestandet (se t.ex. Bachman & Palmer 1996, McNamara 2015 och Luoma 2001). Tidigare studier om likvärdig bedömning av språkkunskaper finns t.ex. i Finland och Sverige. Mest har intresset varit att undersöka bedömning och dess reliabilitet. Det finns sammanlagt mycket få studier om muntlig färdighet. Min studie är därför en viktig pusselbit i kunskapsbygget.

I Finland har Juutilainen (2011) och Lindroos (2010) avlagt avhandlingar där de båda har studerat bedömningens likvärdighet bland olika bedömare. Juutilainen (2011: 4, 5) har undersökt jämförbarhet av utvärderingar av fyra informanter. Syftet med hennes studie var att ta reda på varför finns det variation i bedömningen även om bedömningskriterierna var läroplanens nivåskala för språkkunskaper. Enligt Juutilainen (2011: 18) var informanterna fyra lärare och de fick bedöma sex elevtexter på svenska som var skrivna av finska högstadieelever.

Som resultat redovisar Juutilainen (2011: 41) att alla informanterna hade samma nivåskala men de använde också kriterier som inte tillhör nivåskalan. Vidare hade informanterna använt likartade motiveringar i bedömningen även om det var oenighet i elevens språkliga nivå. Som metod använde Juutilainen (2011: 2) en halvstrukturerad temaintervju och kvalitativ analys av intervjumaterialet.

Lindroos (2010: 7) har också utfört en studie om liknande tema men på gymnasie- och högskolenivån. Hon studerade utvärdering av muntlig prestation. Materialet bestod av bedömningar av muntlig färdighet i svenska som andraspråk i A-svenska i gymnasiet av fyra universitetslärare och utvärderingar av muntlig färdighet i svenska på högskolenivå av tre universitetslärare. Syftet med studien var att ta reda på om sju bedömares subjektiva syn på kriteriebaserad bedömning av språkkunskaper i svenska. Hon analyserar bedömarnas reflektioner i bedömning av muntlig kunskap i svenska.

Som metod använde Lindroos (2010: 11) en intervju och analyserade materialet med hjälp av meningskoncentreringsmetod. Som huvudresultat presenterar Lindroos (2010: 67) att ”attityd till bedömning av de kommunikativa aspekterna av muntlig språkfärdighet” skiljde sig mellan bedömarna. Lindroos (2010: 71, 72) påpekar att det ändå inte finns drastiska skillnader i bedömningen mellan två grupper av informanterna även om några informanter hade gett högre vitsord. Den här variationen beror på hur mycket bedömaren har betonat förut den kommunikativa aspekten i bedömning (Lindroos. 2010: 71).

I Sverige har man undersökt bland annat hur läraren bedömer andraspråktexter (se Fransson 2010). Vidare har Nordström (2005) studerat likvärdig bedömning i elevtexter i åk 9 i sin uppsats. Det finns också en studie där Grube (2012) studerade språkbedömning bland elever med annat modersmål och kartläggande av rutiner och granskning av språkliga bedömningsverktyg.

Vidare undersökte Stolt (2016: 215) bedömare som fick utvärdera elevernas prestationer i studentexamen i modersmål och litteratur i svenska. Syftet med undersökningen var att studera ”värderande uttryck och en institutionell interaktion i bedömarkommentarer.” Stolt studerade tendenser genom bedömarkommentarer. Som resultat fick Stolt (2016: 225) att det är ”vanligare att goda uppsatser läses och kommenteras av fler än en bedömare”.

Inom ramen av projektet Inlärningsgångar i andraspråket (Topling) har det kommit ut flera studier som handlar om språkbedömning. Syftet med Topling -projektet (Topling. 2012) var ”att undersöka inlärning av skriftlig färdighet i finska, engelska och svenska som andraspråk i det finska utbildningssystemet, genom att jämföra tvärsnittsdata med longitudinella data. Även vuxna inlärare studeras.”

Vidare undersökte Toropainen m.fl. (2012) hur bedömare använde GERS-nivåskalan i bedömning av skriftlig prestation i svenska som förstaspråk och svenska som andraspråk. Syftet var att ta reda på vilka aspekter bedömare koncentrerade sig i bedömningen och hur uppdraget påverkade bedömningsprocessen.

3. SYFTE, MATERIAL OCH METOD

I det här avsnittet presenteras utgångspunkter för datainsamling och analysprocess. Först redovisas uppsatsens syfte och hypoteser. Vidare redogörs för informanterna och genomförandet av materialinsamlingen. Därefter presenteras insamlat material. Dessutom finns det information om metod som används i datainsamling och metod i analysprocess.

3.1. Uppsatsens syfte och hypoteser

De teoretiska utgångspunkter som presenteras i avsnitt 2 bygger en ram för min analys. Vidare återspeglas resultaten till tanken av Davies m.fl. (1999: 88). De påpekar att samstämmighet är högre i änderna av nivåskalan och ofta finns variation i mitten av skalan. Det kan ha konsekvenserna. Det kan påverka t.ex. att man blir godkänt i ett test. För att nå renhårighet rekommenderar Davies m.fl. (ibid.) att två eller flera bedömare utvärderar prestationer när det är frågan om high-stakes test som bedöms subjektivt.

Centrala begrepp som jag använder är interbedömarreliabilitet och muntlig färdighet. Med interbedömarreliabilitet menas i denna avhandling reliabiliteten mellan flera bedömare. Med muntlig färdighet menas den produktiva förmågan som består t.ex. av talets flyt, flexibilitet, koherens, uttrycksförmågans exakthet och omfattning, behärskning av uttalet och exakthet i strukturer.

Min studie skiljer sig från tidigare studier för jag använder prestationer från YKI-korpusen. Min synvinkel är high stakes-test som inte har på det här sättet studerat. Vidare ska min studie fylla information om interbedömarreliabiliteten i utvärderingar av en samma prestation.

Syftet med föreliggande studie är att undersöka samstämmighet i bedömning av muntlig färdighet i svenska som andraspråk. Jag söker svar på följande forskningsfrågor:

1. Motsvarar informanternas bedömningar YKI-bedömarnas bedömningar?
2. Hur tillämpar informanterna YKI-kriterierna?
3. Hurdan variation finns det i informanternas bedömningar?

Dessa frågor är intressanta för att se hur YKI-kriterierna tillämpas av språkproffs som inte har fått YKI-utbildning men har insikt i frågan genom annan bakgrund. Vidare visar frågorna om bedömningen är samstämmig. Det kan vara t.ex. att en annan bedömares linje är striktare i jämförelse med de andra bedömarna.

Att studera tillämpning av kriterierna och samstämmighet i bedömning är intressant eftersom det ändå finns risk att bedömningarna varierar även om det finns gemensamma riktlinjer för att bedöma muntlig färdighet. Det finns alltid en risk att reliabiliteten mellan bedömare kan variera eftersom bedömningar utförs av människor med sina egna tolkningssätt och erfarenheter. Det betyder att det finns subjektivitet i bedömningar, särskilt i bedömning av muntlig färdighet. Det kan hända att testtagare får godkänt i ett test samtidigt en annan bedömare ger icke-godkänt eller sämre betyg på samma prestation. Bedömningen kan på det sättet påverka framtidsplaner t.ex. karriär eller utbildning särskilt när vi pratar om high stakes-test.

Hypoteser som kompletterar min studie finns tre. För det första förväntas att informanterna kan tillämpa kriterierna mycket väl om de har erfarenhet att bedöma muntliga kunskaper i svenska språket. För det andra förväntas att informanternas sätt att bedöma varierar. De kan t.ex. betona kriterier på olika sätt i sina bedömningar. Den tredje hypotesen är att informanterna använder kriterier utanför givna kriterier så som t.ex. Juutilainen (2011) visade i sin studie.

3.2. Informanterna

Jag skickade förfrågan till ca 40 lärare i Finland var av 8 deltog i studien. Kraven för att ta del i studien var att informanten måste ha erfarenhet att undervisa och bedöma vuxnas muntliga färdigheter. Minimikraven var att läraren har undervisat på andra stadiet. Informanterna jobbar i olika delar i Finland. Informanterna bedömer muntliga prestationer med hjälp av YKI-kriterierna i min studie. Informanterna fick slumpmässigt valda pseudonymer för att försäkra anonymitet. Jag presenterar kort profiler för alla informanter. Profilen innehåller följande information: A. ålder, B. kön, C. arbetsår som lärare, D. läroämnen, E. undervisningsstadie som du har gjort bedömningsarbete. Informationen insamlades med formulär (se bilaga 2). Profilerna för informanterna finns nedan:

Maria, 47 år, kvinna, 23 år som lärare, svenska, andra stadiet

Helena, 55 år, kvinna, 29 år som lärare, svenska, alla stadier, för närvarande andra stadiet

Harri, 60 år, man, 31 år som lärare, engelska och svenska, andra stadiet

Anna, 44 år, kvinna, 17 år som lärare, svenska och engelska, andra och tredje stadiet, för närvarande andra stadiet

Liisa, 53, 20+ år som lärare, svenska och franska, grundskola och andrastadiet, för närvarande andra stadiet

Julia, 28 år, 1 år som lärare, engelska och svenska, tredje stadiet

Mia, 27 år, 2 1/2 år som lärare, svenska, tredje stadiet

Eeva, 26 år, 2 år som lärare, svenska, tyska och engelska, alla stadier, för närvarande grundskola

Enligt en elektronisk förfrågning svarade alla att de hade utvärderat muntlig färdighet förut. Alla medger också att de kände till GERS. I formulären frågades också om informanten kände till YKI-testet förut. Det var bara Liisa som hade gjort YKI-bedömningar förut. Andra deltagare hade antingen hört om YKI-testet eller inte alls kände till YKI-testet.

3.3. Genomförandet av materialinsamlingen

I detta avsnitt redogörs för genomförandet av materialinsamlingen. Först redovisas materialpaketet som skickades per e-post till alla informanter. Vidare presenteras muntliga prestationer som fungerar som bas för informanternas bedömningar.

3.3.1. Materialpaketet för informanterna

I början av datainsamlingen bad jag informanterna att fylla i ett formulär med de första uppläsningarna. Jag frågade t.ex. om de har erfarenhet av språkbedömning, har de undervisat studerande i olika åldrar, hur mycket de känner den europeiska referensramen eller YKI-kriterierna (se bilaga 2). Efter att informanten hade sagt ja till att medverka i studien, skickades instruktionerna och allt material per e-post. Materialpaketet innehöll informationsbrevet om undersökningen och privatutdelade länken till alla 10 audioklipp som jag lade till SoundCloud appen från YKI-korpusen. Vidare innehöll e-posten YKI-kriterierna för kunskapsnivåer 1 till 6 (se bilaga 1). Informanterna fick fylla i svarsformulären och formulären med de första upplysningar (se bilaga 2) och skicka de per epost till mig.

3.3.2. Muntliga prestationer från YKI-korpusen

Med hjälp av materialinsamling används 10 muntliga prestationer som samlades in från YKI-korpusen. Korpusen är en databank som innehåller testprestationer som kommer från YKI-testet. Korpusen innehåller bland annat testdeltagarnas prestationer (bl.a. uppsatser och muntliga prestationer), bakgrundsinformation och bedömningsresultat. Muntliga prestationer välde jag från mellannivå (nivå 3 och 4). Jag bestämde att testdeltagarna får ha bara finska som modersmål. Informanterna får alltså koncentrera sig bara på att bedöma muntliga prestationer där testdeltagare inte har t.ex. olika accent från olika språk. På det sättet försöker jag avgränsa variation mellan bedömningarna.

I YKI-korpusen kan man hitta prestationer med deltagare ID efter att man har ansökt ID för att använda YKI-korpusen. Bakgrunden mellan testtagaren varierade bland annat i kön, åldern och varför de tar testet. Fem av testtagaren var kvinnor och fem var män. Åldern i prestationer var mellan 26–55 år. Fem tog YKI-testet för att visa sina språkkunskaper till den nuvarande arbetsgivaren, tre för att få feedback om sina språkkunskaper och två för att söka arbete. De 10 prestationer som jag använde var från mellannivå (nivå 3 och 4). Prestationerna är bedömt av YKI-bedömarna som har fått en särskild utbildning för uppgiften.

3.4. Insamlat material

Jag samlade in materialet för undersökningen 13 januari–20 februari 2017. Materialet består av informanternas bedömningar som insamlades med ett svarsformulär (se bilaga 3). Informanterna motiverar sina bedömningar med hjälp av YKI-kriterierna (se bilaga 1). Därefter samlade jag in med ett formulär (se bilaga 2) följande upplysningar: ålder, kön, undervisningsstadier som du har gjort bedömning, undervisningsstadie du för närvarande undervisar, arbetsår, läroämnen du undervisar.

Efter att informanterna bedömde muntliga prestationer bad jag informanterna att ge anonymt feedback med ett elektroniskt formulär. Det frågas t.ex. hurdant det var att bedöma muntliga prestationer. Meningen med förfrågan var att få information om det finns externa faktorer som kan påverka variation i bedömningar, t.ex. bråttom. Förfrågan är en kombination av Likert-skala och öppna frågor. Efter analysen skickade jag feedback till informanterna för deras utvärderingar i april 2017.

3.5. Analysmetod

Eftersom insamlat data består av informanternas skriftliga svar utförs analysen kvalitativt. Som metod används kvalitativ innehållsanalys. Analysen sker induktivt genom att materialet kategoriseras genom skriftliga kommentarer som informanterna har givit med numerisk bedömning.

Syftet med den induktiva analysen är att sammanfatta en teoretisk helhet genom att analysera materialet (se t.ex. Tuomi & Sarajärvi 2009: 95). Tidigare teorier eller information påverkar inte analysen. Enligt Tuomi och Sarajärvi (2009: 108) är syftet i innehållsanalysen att beskriva fenomenet genom att sammanfatta information utan att det förloras. I denna avhandling följs Miles och Hubermans modell (1994) som skiljer analysprocessen till 1) *Reducering*, 2) *Kategorisering* och 3) *Abstrahering* av materialet.

Med *reducering* försöks hitta svar på forskningsfrågorna i studien. Målet är att ta bort onödig information och filtrera uttryck som har likheter. Dessa uttryck bygger sedan *kategorier*. Vidare bygger dessa kategorier delkategorier och till sist en gemensam huvudkategori. Bestämda kategorierna fungerar som teoretiska begrepp som hjälper att svara på forskningsfrågorna i studien

i slutet. Det kallas *abstrahering* (se Tuomi & Sarajärvi 2009: 101, 108, 109). I min studie bygger kategorier på informanternas bedömningar. Kategorier formas enligt informanternas beskrivningar om muntlig färdighet.

I denna studie kvantifieras materialet också efter kategorisering. Jag beräknar frekvenser hur många gånger informanterna omnämner samma beskrivningar. Kvantifiering är inte nödvändigt men den tycks vara ett effektivt sätt att ge nytt perspektiv till tolkning av materialet (se Tuomi & Sarajärvi 2002: 119).

4. ANALYS OCH RESULTAT

I kapitel 4 introduceras resultaten för denna studie. Resultaten presenteras kvantitativt och kvalitativt. Kvantitativt redovisas bedömningsresultaten genom tabeller som jag har utfört i Excel. Med hjälp av innehållsanalys kategoriseras informanternas kommentarer som informanter gav i samband med poängsättning. Avsnitt 4.3, 4.4 och 4.5 baserar sig på forskningsfrågorna i denna studie. I avsnitt 4.6 sammanfattas de faktorer som har påverkat bedömningen enligt informanterna.

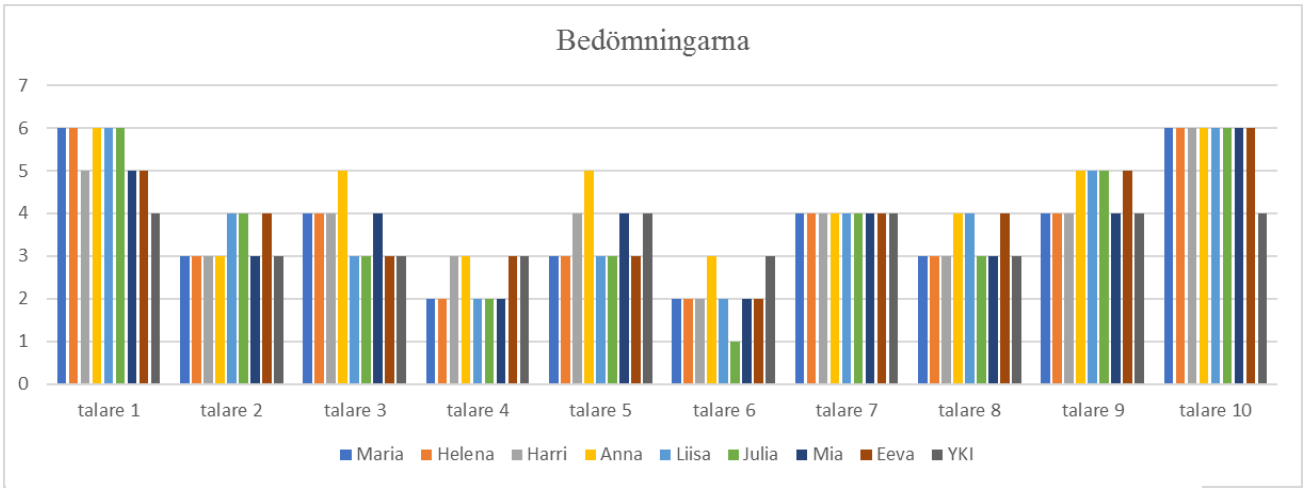
4.3. Motsvarar informanternas bedömningar YKI-bedömarnas bedömningar?

För att jämföra YKI-bedömarnas utvärderingar med informanternas bedömningar finns det några faktorer bakom YKI-testets bedömningar. Enligt Ohranen (e-post, 15.12.2016) har YKI-bedömarna dubbelbedömt talare 4 och 7 så vitsordet i dessa prestationer består av medelvärdet av två bedömningar. Det är ändå så att båda YKI-bedömarna har givit nivå 3 för talare 4. Med talare 7 finns det oenighet med bedömning mellan YKI-bedömarna. Det är variation mellan 4 och 3, men den slutliga bedömningen är nivå 4. Vidare är det talare 10 som YKI-bedömarna har poängsatt till nivå 4. Enligt Ohranen (ibid.) tänkte YKI-bedömarna att talaren fick en stark 4. Det betyder att hans kunskaper ändå skulle nå den högre nivån (nivå 5 eller 6) i bedömningsskalan.

Tabell 1 Bedömningarna av 8 informanter och YKI-bedömarna

Bedömningarna enligt kriterier för muntlig färdighet (nivå 1-6)										
	prestation									
informant	talare 1	talare 2	talare 3	talare 4	talare 5	talare 6	talare 7	talare 8	talare 9	talare 10
Maria	6	3	4	2	3	2	4	3	4	6
Helena	6	3	4	2	3	2	4	3	4	6
Harri	5	3	4	3	4	2	4	3	4	6
Anna	6	3	5	3	5	3	4	4	5	6
Liisa	6	4	3	2	3	2	4	4	5	6
Julia	6	4	3	2	3	1	4	3	5	6
Mia	5	3	4	2	4	2	4	3	4	6
Eeva	5	4	3	3	3	2	4	4	5	6
YKI	4	3	3	3	4	3	4	3	4	4

I tabell 1 och figur 2 förevisas informanternas och YKI-bedömarnas utvärderingar av de 10 talprestationer. Informanternas bedömningar varierar men bedömningarna varierar också i jämförelse med YKI-bedömarna.



Figur 2 Bedömningarna enligt kriterier för muntlig färdighet (se bilaga 1)

Figur 2 visar att det finns en prestation som alla informanterna är eniga med YKI-bedömarna (se talare 7). Vidare finns det två prestationer där ingen informant är enig med YKI-bedömarna (se talare 1 och 10). Informanternas bedömningar syftar till högre nivå för talare 1 och 10. Med talare 1 och 10 finns det en möjlighet att YKI-bedömarna skulle ha givit vitsordet på högre nivå, men testtagarna har själv valt att ta del i testet på mellannivå. Vidare finns det bara en informant som är enig med YKI-bedömarna med talare 6. Andra informanterna har poängsatt prestationen till lägre nivå. Med andra prestationer finns det 3–4 informanter som är eniga med YKI-bedömarna och andra har poängsatt prestationer till högre nivå (se talare 2, 3, 4, 5, 8 och 9).

Tabell 3 visar interbedömarreliabiliteten av informanterna och YKI-bedömarna. Totalsumman av alla bedömningar är ett relationstal som används när interbedömarreliabiliteten behandlas. Tumregeln är att summan av varje bedömares bedömningar jämförs med andra bedömare. Ju större summan desto mildare har bedömaren bedömt i allmänhet. I tabell 3 ser vi att Anna är den som har varit mild i sina bedömningar i jämförelse med andra bedömare. Mildare bedömning syns också i bedömningar av talare 3, 4 och 6. Anna har varit det enda av informanterna som har givit högre vitsord än andra informanter. Men Maria, Helena, Julia och Mia har varit striktaste i jämförelse med alla informanter. Summan av YKI-testets bedömningar kan inte

kompareras eftersom det inte har varit möjlighet för YKI-bedömarna att ge högre eller lägre vitsord än 3 eller 4.

Tabell 3 Interbedömarreliabilitet

Bedömningarna, interbedömarreliabilitet									
	informant								
prestation	Maria	Helena	Harri	Anna	Liisa	Julia	Mia	Eeva	YKI
talare 1	6	6	5	6	6	6	5	5	4
talare 2	3	3	3	3	4	4	3	4	3
talare 3	4	4	4	5	3	3	4	3	3
talare 4	2	2	3	3	2	2	2	3	3
talare 5	3	3	4	5	3	3	4	3	4
talare 6	2	2	2	3	2	1	2	2	3
talare 7	4	4	4	4	4	4	4	4	4
talare 8	3	3	3	4	4	3	3	4	3
talare 9	4	4	4	5	5	5	4	5	4
talare 10	6	6	6	6	6	6	6	6	4
Total	37	37	38	44	39	37	37	39	35

Alla tio talare har deltagit i testet på mellannivå (3 och 4), men mina informanter visste inte det. Informanterna fick alltså fri händer att utvärdera prestationer så att de skulle ge vitsord i grundnivå (1 och 2) och högre nivå (5 och 6). I tabell 4 visas informanternas bedömningar som baserar sig på informanternas bedömningar men meningen är att visa variation mellan olika nivåstegen.

Tabell 4 Bedömningarna och gränsen att bli godkänt, h=högre nivå, g=grundnivå

Bedömningarna som siktar mot grundnivå (1-2) och högre nivå (5-6)										
	prestation									
informant	talare 1	talare 2	talare 3	talare 4	talare 5	talare 6	talare 7	talare 8	talare 9	talare 10
Maria	h			g		g				h
Helena	h			g		g				h
Harri	h					g				h
Anna	h		h		h				h	h
Liisa	h			g		g			h	h
Julia	h			g		g			h	h
Mia	h			g		g				h
Eeva	h					g			h	h
YKI	4	3	3	3	4	3	4	3	4	4

Tabell 4 visar att informanterna skulle sätta talare 1 och 10 på högre nivå. Vidare ser vi att Anna skulle sätta talare 3 och 5 på högre nivå vilket avviker från andra bedömningar. Det är också

talare 4 som fem informanter skulle sätta på högre nivå. Vidare sätter fyra informanter talare 9 till högre nivå.

Det är talare 6 som är intressant om vi tänker på gränsen att bli godkänt i testet. Det är bara Anna vars bedömning skulle ge talare 6 vitsordet att bli godkänt i mellannivå och sammanhänger med YKI-bedömarna. Det är då sju informanter som tycker att talare 6 hör till lägre nivå än vad testet mäter. Det betyder att bara en av informanterna ock YKI-bedömarna skulle ge godkänt i testet samt sju informanter skulle inte ge godkänt. Det är också talare 4 vars kunskaper enligt Maria, Helena, Liisa, Julia och Mia hör till grundnivå vilket betyder att talare 4 inte skulle bli godkänt i mellannivå.

Om alla 80 bedömningar relateras till 31 bedömningar som är sammanhållande med YKI-bedömarna är det 39 % av alla bedömningar som samstämmer med YKI-bedömarna. Det betyder att 61 % av bedömningarna motsvarar inte YKI-bedömarnas bedömningar. Talare 1 och 10 finns inte med jämförelsen eftersom det skulle förvränga jämförelsen.

Tabell 5 visar medelvärdet av informanternas bedömningar jämfört med YKI-bedömarnas bedömningar. Tabellen visar att spridning och informanterna har varit strikta på talare 4, 5 och 6 i jämförelse med YKI-bedömarna. Enligt medelvärde i jämförelse med YKI-bedömarna kan påstås att 50 % av informanternas bedömningar motsvarar YKI-bedömarnas bedömningar (se talare 2, 5, 7, 8 ja 9). Det är ändå inte så klart om man jämför studiens alla bedömningar till YKI:s bedömningar.

Tabell 5 Medelvärdet av informanternas bedömningar jämfört med bedömningar i YKI

	mv	YKI
talare 1	5,63	4
talare 2	3,38	3
talare 3	3,75	3
talare 4	2,38	3
talare 5	3,50	4
talare 6	2,00	3
talare 7	4,00	4
talare 8	3,38	3
talare 9	4,50	4
talare 10	6,00	4

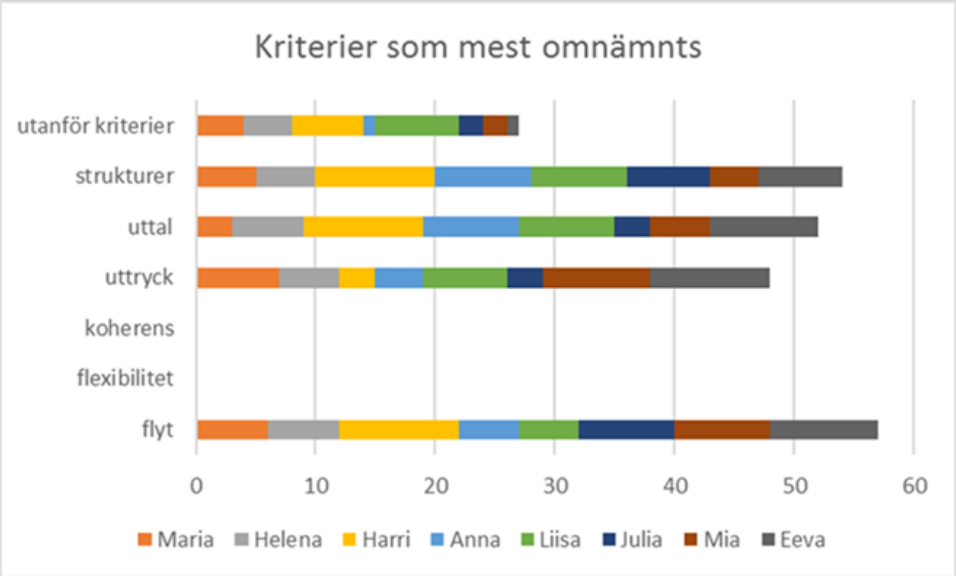
Allt som allt visar dessa resultaten att det finns risken att bedömares utvärdering kan variera från andra bedömare. Om vi jämför medelvärdet, interbedömarreliabilitet och variationen som finns i omdömen och relateras alla bedömningar (80) till sammanhållande bedömningar med YKI-bedömarna kan påstås att informanternas bedömningar inte motsvarar alla prestationer med YKI-bedömarnas bedömningar. Det kan i värsta fall betyda att testtagaren råkar ha en strikt bedömare och blir inte godkänt även om andra testtagare i med samma kunskaper blir godkänt. Vidare kan det hända, som med talare 6, att testtagaren blir godkänt eller får högre betyg än de som bedöms av andra bedömarna. Jag återkommer i diskussionen till möjliga orsaker till skillnaderna och vad som kan sägas om deras betydelse.

4.4. Hur tillämpar informanterna YKI-kriterierna?

I detta avsnitt redogör jag för hur informanterna tillämpar allmänna språkexamens kriterier. Motiveringar i bedömningar kategoriserades med induktiv innehållsanalys till 7 kategorier som följer delkriterium i YKI-kriterierna: talets flyt, flexibilitet, uttryck, uttal, strukturer och motiveringar utanför kriterier.

Det kom fram att i skriftliga motiveringar finns det kriterier som inte alls omnämns och kriterier som kommenteras mest i skriftliga kommentarer. YKI-kriterierna som inte tas upp i kommentarer är koherens och flexibilitet. Sammanlagt betonar informanterna flera gånger flyt och strukturer, se figur 4.

Tabeller 6–13 visar att det finns variation mellan informanterna t.ex. hurdana kriterier de omnämner. Vidare betonar informanterna olika kriterier. Det är t.ex. Harri som kommenterar alltid uttal, flyt och strukturer i prestationer. Vidare kommenterar Eeva nästan alltid flyt, uttryck, uttal och strukturer. Maria och Mia kommenterar mest flyt och uttryck när Helena kommenterar mest flyt och uttal. Anna har kommenterat mest uttal och strukturer. Liisa har också betonat mest uttal och strukturer men också uttryck. Julia kommenterar mest uttryck och flyt i sina bedömningar (se tabeller 6–13).



Figur 4 Kriterier som mest omnämnts i skriftliga motiveringar

Tabell 6 Kategorisering av Marias skriftliga motiveringar

				Maria				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x			x				
Talare 2				x			x	
Talare 3						x		
Talare 4	x			x	x	x		
Talare 5							x	
Talare 6	x							
Talare 7	x			x	x	x		
Talare 8	x			x		x		
Talare 9				x	x		x	
Talare 10	x			x		x	x	

Tabell 7 Kategorisering av Helenas skriftliga motiveringar

				Helena				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x			x			x	
Talare 2	x				x	x		
Talare 3				x		x		
Talare 4				x	x	x		
Talare 5					x	x	x	
Talare 6	x							
Talare 7	x				x	x	x	
Talare 8	x				x			
Talare 9	x			x	x			
Talare 10				x			x	

Tabell 8 Kategorisering av Harris skriftliga motiveringar

				Harri				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x			x	x	x	x	
Talare 2	x				x	x	x	
Talare 3	x				x	x	x	
Talare 4	x				x	x		
Talare 5	x				x	x		
Talare 6	x				x	x	x	
Talare 7	x				x	x		
Talare 8	x			x	x	x		
Talare 9	x			x	x	x	x	
Talare 10	x				x	x	x	

Tabell 9 Kategorisering av Annas skriftliga motiveringar

				Anna				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x			x				
Talare 2					x	x		
Talare 3	x				x	x		
Talare 4					x	x		
Talare 5	x				x	x		
Talare 6	x				x	x		
Talare 7				x	x	x		
Talare 8				x	x	x		
Talare 9	x				x	x		
Talare 10				x			x	

Tabell 10 Kategorisering av Liisas skriftliga motiveringar

				Liisa				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1				x	x	x		
Talare 2	x			x	x	x		
Talare 3	x			x	x	x	x	
Talare 4				x		x	x	
Talare 5				x	x	x	x	
Talare 6	x				x	x	x	
Talare 7	x				x	x	x	
Talare 8	x			x	x	x		
Talare 9				x			x	
Talare 10					x		x	

Tabell 11 Kategorisering av Julias skriftliga motiveringar

				Julia				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x				x		x	
Talare 2				x		x		
Talare 3	x				x	x		
Talare 4						x		
Talare 5	x					x		
Talare 6	x			x				
Talare 7	x					x		
Talare 8	x					x		
Talare 9	x				x	x		
Talare 10	x			x			x	

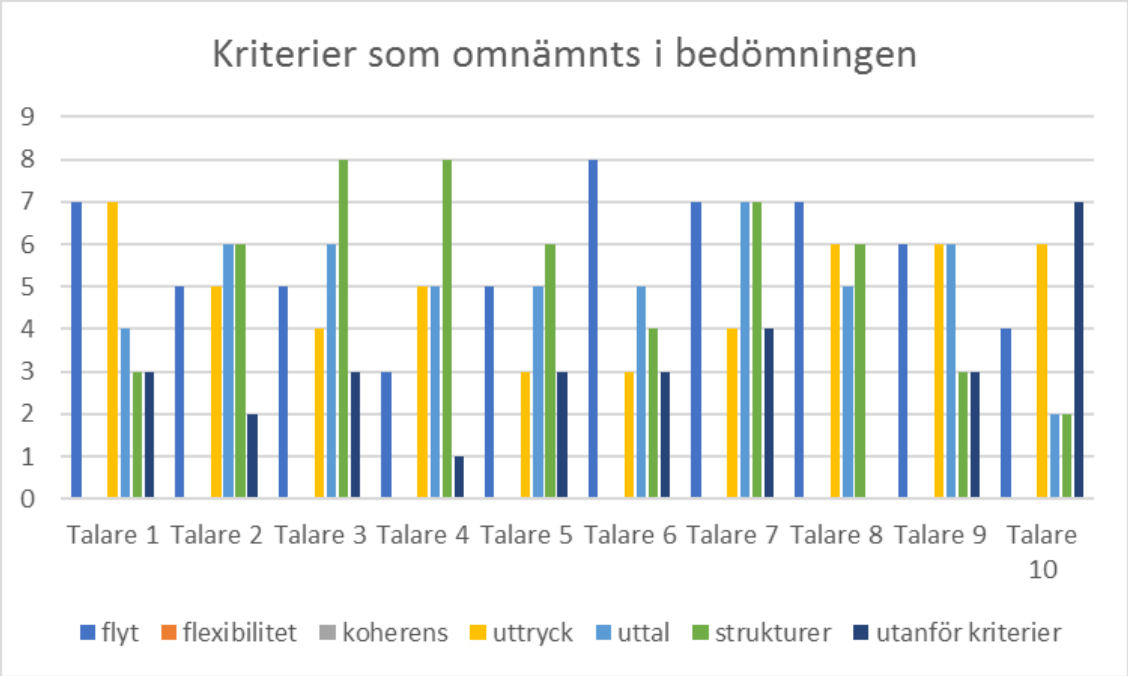
Tabell 12 Kategorisering av Mias skriftliga motiveringar

				Mia				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x			x				
Talare 2	x			x	x	x		
Talare 3				x	x	x	x	
Talare 4	x			x		x		
Talare 5	x			x				
Talare 6	x			x	x			
Talare 7	x			x	x		x	
Talare 8	x			x				
Talare 9	x				x	x		
Talare 10				x				

Tabell 13 Kategorisering av Eevas skriftliga motiveringar

				Eeva				
	flyt	flexibilitet	koherens	uttryck	uttal	strukturer	utanför kriterier	
Talare 1	x			x	x	x		
Talare 2	x			x	x			
Talare 3	x			x	x	x		
Talare 4				x	x	x		
Talare 5	x			x	x	x		
Talare 6	x			x	x	x		
Talare 7	x			x	x	x		
Talare 8	x			x	x	x		
Talare 9	x			x	x			
Talare 10	x			x			x	

Figur 3 visar att kriterier som informanterna har omnämnt i bedömning av samma prestation varierar. Figuren visar till exempel att med talare 6 kommenteras mest prestationens flyt men med talare 3 och 4 kommenteras mest talets strukturer i bedömningar. Det är alltså variation att nämna samma kriterier till alla prestationer. Det kan förstås bero på att talarna är olika.



Figur 3 Kriterier som omnämnts i bedömningen

Informanterna har använt kriterier utanför givna YKI-kriterierna. Informanterna har motiverat bedömningar genom att ge svar som hör till följande kategorier: tolkning, jämförelse till andra prestationer, observationer och subjektiv bedömning. Med tolkning menas att informanten har konkluderat efter att de har analyserat och givit vitsordet till prestationer. Jämförelse till andra prestationer sammanfattar de gånger som bedömare har jämfört föregående prestationen till prestationen i fråga. Observationer sammanfattar gånger då informanterna har kommenterat egenskap i talarens muntliga färdighet som inte direkt hör till delkriterier. Med subjektiv bedömning menas gånger då bedömaren har använt komparativ och utvärderat prestationens kvalitet med fraser som är subjektivt laddad. I tabell 14 redovisas kommentarer.

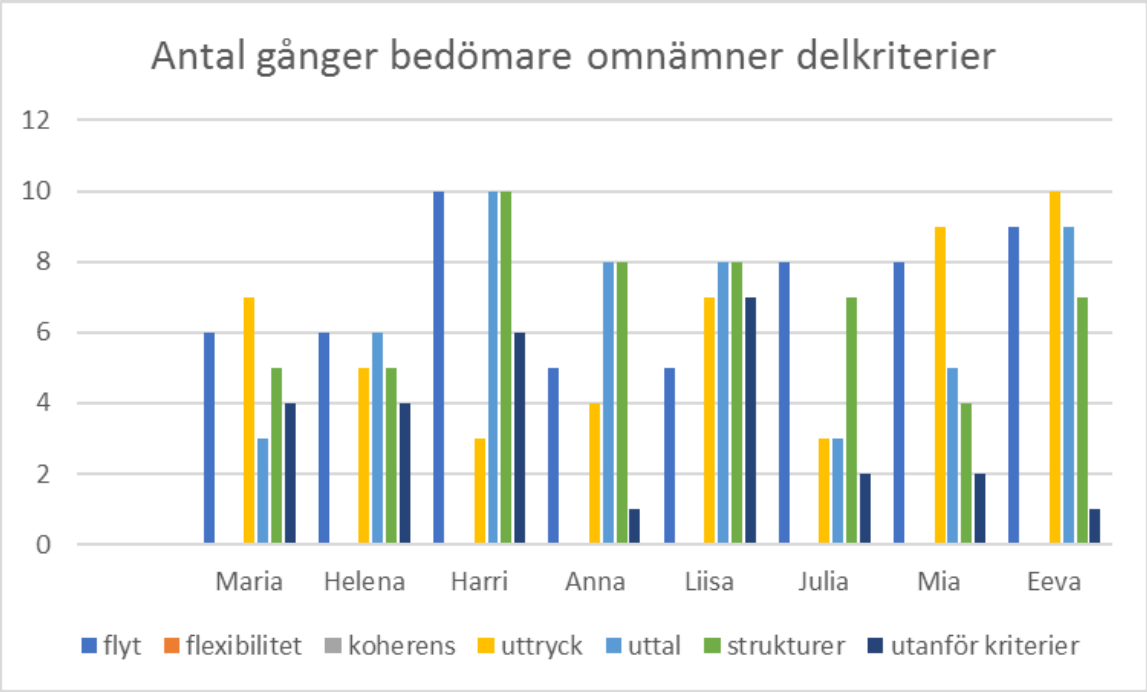
Tabell 14 Kriterier utanför YKI-kriterierna (översättningarna är författarens)

Utanför kriterier	
Tolkningar	**Observation**
Helena, talare 1: ”Använder inte svenska varje dag.” Julia, talare 1: ”Efter vad jag har hört tror jag inte att personen skulle ha svårigheter att klara sig av i inga situationer på svenska.” Liisa, talare 4: ”Jag tror att samtal om ett abstrakt ämne skulle orsaka problem.” Helena, talare 7: ”…har lärt sig svenskan kanske i ett finlandssvenskt område.” Mia, talare 7: ”Det verkar som om talaren har använt svenska i vardagen.” Helena, talare 10: ”Svenska som modersmål.” Harri, talare 10: ”Låter autentiskt.” Anna, talare 10: ”Talarens modersmål.” Liisa, talare 10: ”Använder svenska i vardagen.” Julia, talare 10: ”Tycks inte avvika från en infödd talare.” Eeva, talare 10: ”Jag får ett intryck att personen talar svenska som modersmål...; ”…om det var mer tid skulle talaren kanske haft flera nya synvinklar i sitt tal…; ”Talaren visar också humor i sitt tal…”	Harri, talare 1: ”Det finns inte främmande inflytande i uttal.” Harri, talare 2: ”Inflytande från andra språk kan man höra i ordförrådet.” Harri, talare 3: ”Ingen märkvärdig främmande inflytande i tal.” Maria, talare 5: ”Lite inflytande från engelska i tal.” Helena, talare 5: ”Engelska (i tal).” Harri, talare 5: ”Anlitar även på engelska (i tal).” Harri, talare 6: ”Främmande inflytande i uttal.” Liisa, talare 6: ”Engelskan (i tal).” Julia, talare 6: ”Engelskan kommer fram (i tal).” Harri, talare 7: ”Främmande inflytande i uttal.” Harri, talare 9: ”Lite främmande inflytande i uttal” Liisa, talare 9: ”Behöver inte använda strategier för att undvika vissa strukturer.”
Jämförelse med andra prestationer	**Subjektiv bedömning**
Liisa, talare 3: ”…i jämförelse med föregående talare.” Mia, talare 3: ”…i jämförelse med föregående talare.” Liisa, talare 7: ”…i jämförelse med föregående talare.” Maria, Talare 10: ”Den bästa av alla prestationerna.”	Maria, talare 2: ”Tillräckligt brett ordförråd.” Maria, talare 5: ”Blir bättre i slutet.” Helena, talare 5: ”Blir bättre i slutet.” Maria, talare 9: ”Blir bättre i slutet.”

Sammanfattningsvis visar resultaten att alla informanterna har tillämpat delkriterierna som finns i YKI-kriterierna men användning av kriterierna varierar. Vidare finns det kriterier i motiveringar utanför YKI-kriterierna. Det finns t.ex. jämförelse med andra prestationer vilket tycks inte vara korrekt i kriteriebaserad bedömning.

4.5. Hurdan variation finns det i informanternas bedömningar?

Variationen syns i numeriskt omdöme (tabell 1) som redovisas i del 4.3. Vidare syns variation i skriftliga motiveringar som informanterna ger efter varje numerisk bedömning. I avsnitt 4.4 visar tabeller 6 till 13 linjen som bedömaren har haft när man jämför bedömning av alla prestationer. Vidare presenterar figur 3 vilka kriterier bedömaren har kommenterat i sina motiveringar när man undersöker varje talare. Figur 5 består av antal gånger bedömaren har omnämnt ett visst delkriterium eller kriterium utanför givna kriterier.



Figur 5 Antal gånger som bedömare omnämner kriteriet

Figur 5 visar att över i hälften av prestationer har informanterna kommenterat flyt i prestationerna (se Maria, Helena, Harri, Liisa, Mia och Eeva). Vidare kommenteras uttryck över i hälften av prestationerna (se Maria, Liisa, Mia och Eeva). På samma sätt kommenteras uttal också i prestationer (se Helena, Harri, Anna, Liisa och Eeva). Strukturer kommenteras över i hälften av prestationerna (se Harri, Anna, Liisa, Julia och Eeva). Alla informanter har omnämnt kriterier utanför givna kriterier, men det är två informanter (Harri och Liisa) som har omnämnt mest kriterier utanför givna YKI-kriterierna.

Som figur 5 visar, kan en bedömare betona olika kriterier än annan bedömare. Betoningen av samma kriterier beror ändå på prestationen. Jag återkommer i diskussionen till möjliga orsaker till skillnaderna och vad som kan sägas om deras betydelse.

4.6. Hur upplevde informanterna bedömningen?

I detta avsnitt presenteras hur informanterna upplevde bedömning av muntlig färdighet. Efter att informanterna hade utvärderat prestationer svarade de på en elektronisk enkät som var en kombination av Likert-skalan 1–5 och öppna frågor. Vidare fick de komplettera deras svar i öppna fälten.

Största delen av informanterna tyckte att utvärdering av muntlig färdighet inte var utmanande, se tabell 15. Alla gav ändå motiveringar för vad som gjorde bedömningen utmanande. En svarade att hen inte kände till kriterier förut. En informant rapporterade att det var utmanande att ge ett tydligt svar på frågorna i formuläret. Vidare svarade tre informanter att det var en utmaning att sätta prestationen bara i en nivå eftersom kriterierna för nivå 3 och 4 var liknande. En informant svarade att om man inte ser talsituationen är det svårt att tolka muntlig färdighet i en naturlig talsituation. Vidare var det svårt att veta vilket delområde i kriterierna som man ska betona (kommunikativ förmåga vs. strukturer) eller i vilken mån kan man kompensera brister i grammatiken med ett brett ordförråd.

Tabell 15 Svårighetsgraden att bedöma muntlig färdighet

	1	2	3	4	5		totalt	medeltal
utmanande	1	1	3	3	0	icke utmanande	8	3

Informanterna fick värdera påverkan av yttre faktorer i bedömningen. Resultaten visar att största delen inser att det inte var yttre faktorer som påverkade bedömningen, se tabell 16.

Tabell 16 Utvärdering om yttre faktorer påverkade bedömningen

	1	2	3	4	5		totalt	medeltal
håller inte med alls	3	3	1	1	0	håller helt med	8	2

Två informanter angav att tidspress påverkade bedömningen. Vidare nämnde en att inga yttre faktorer påverkade bedömningen. En informant svarade ändå att hon jämförde nivåer med kunskaper som hennes elever skulle ha vilket hon trodde att kan påverka sin bedömning. En informant svarade inte på denna punkt.

Informanterna approximerade också om det var utmanande att tillämpa YKI-kriterierna. I tabell 17 står att informanterna är neutrala med sina svar. Det finns inte variation i änderna av skalan utan informanterna har antingen svarat att det lite är utmanande att tillämpa YKI-kriterierna eller det är inte så värst utmanande.

Tabell 17 Svårighetsgraden att tillämpa YKI-kriterierna.

	1	2	3	4	5		totalt	medeltal
utmanande	0	2	3	3	0	icke utmanande	8	3,13

Sex informanter motiverade att tillämpning av kriterierna är icke-utmanande eftersom de var tydliga och hänger ihop med GERS-nivåskalan. Två informanter nämnde att tillämpa kriterierna var utmanande eftersom de inte kände till kriterierna förut. Utmanande för en informant var ändå att skilja nivå 3 och 4 från varandra för det finns så mycket likheter mellan dessa nivåer.

Tabell 18 Tid som informanter använde i bedömningen

	Antal
30 min	0
1 h	3
1 h 30 min	3
2 h	1
mer än 2 h	1

Informanterna approximerade tidsanvändningen av bedömningen. I tabell 18 står det att alla använde minst en timme vilket betyder att informanterna har koncentrerat sig på bedömningen. De omnämnde flera faktorer som påverkade tidsanvändningen av bedömningen. Fem av informanterna nämnde att de lyssnade talprestationer flera gånger. Två av informanterna nämnde att de jämförde en kandidat med en annan kandidat vilket de tyckte att tog tid. I allmänhet tyckte

informanterna att det tog tid att sätta sig in i materialpaketet och särskilt YKI-kriterierna och talprestationer.

Sammanfattningsvis kan påstås att informanterna hade vissa svårigheter att bedöma prestationerna, och det var speciellt utmanande att välja mellan nivå 3 och 4 eftersom det finns likheter mellan kriterierna. Det kom också fram att de kände att det tog tid att sätta sig till kriterierna men det var hjälp att kriterierna baserar på GERS-nivåskalan för språkkunskaper.

5. DISKUSSION

I detta avsnitt presenteras de centralaste resultaten och intressanta fynd som reflekteras i förhållande till den teoretiska ramen. Vidare diskuterar jag orsaker till resultaten och vilka betydelser resultaten kan ha.

Resultaten stämmer delvis med mina hypoteser. Om alla 80 bedömningar relateras till 31 bedömningar som är sammanhållande med YKI-bedömarna är det 39 % av alla bedömningar som samstämmer med YKI-bedömarna. Det betyder att 61 % av bedömningarna inte motsvarar YKI-bedömarnas bedömningar. Det betyder att samstämmig bedömning är låg. Om medelvärdet av informanternas bedömningar jämförs med YKI-bedömarnas bedömningar motsvarar informanternas bedömningar 50 % YKI-bedömarnas utvärderingar. Inom ramen av denna avhandling är det relevant att visa att enstaka lärarens utvärderingar visar mycket variation. Men resultaten bevisar också att medelvärdet av alla bedömningar motsvarar best bedömningen som YKI-bedömarna har givit.

Variationen som finns i bedömningar syns i numeriska omdömen och skriftliga motiveringar men mest i skriftliga motiveringar. Det finns variation också mellan informanterna. Genom att analysera informanternas slutliga svar i den elektroniska förfrågningen fick jag analysera orsaken till variationen i informanternas bedömningar. Enligt informanterna kan bedömningar variera eftersom det var utmanande att bestämma vilket delkriterium man ska mest betona. Vidare svarade informanterna att det var svårt att bedöma kunskapen eftersom situationen inte var en naturlig talsituation som till exempel i observation. Vidare svarade informanterna att det var utmanande att bygga upp motiveringar på tydligt sätt. Informanterna svarade också att det var svårast att sätta talprestation mellan nivå 3 och 4 för det finns likheter mellan kriterier i båda nivåer. Som Davies m. fl. (1999: 88) påpekar är det samstämmighet högre i änderna av nivåskalan. Ofta finns variation i mitten av skalan. Det syns också i mina resultat.

Skriftliga motiveringar i bedömningen visar att YKI-kriterierna omnämns men de beror på bedömaren vilka kriterier hen betonar. Variationen visar att det inte finns systematik att omnämna samma kriterier till alla prestationer. Det var också två delkriterier, flyt och strukturer, som betonas mest och två delkriterier, koherens och flexibilitet, som omnämns inte alls i motivering-

arna. Vidare har informanterna använt kriterier utanför givna kriterier vilket betyder att subjektivitet kan påverka bedömning av muntlig färdighet. Vidare finns det jämförelse med prestationer i skriftliga motiveringar vilket inte tycks höra till kriteriebaserad bedömning. Det här fenomenet kan kanske kontrolleras genom utbildning som nämndes också i teoridelen (se avsnitt 2.2.4).

I allmänhet visar resultaten att språkproffs tillämpade kriterierna även om de inte hade haft YKI-utbildning. Resultaten visar ändå att dubbelbedömningen är relevant i high stakes-test som Davies m.fl. (1999: 88) påstår, eftersom det finns variation i bedömningar. Vidare använde informanterna kriterier utanför givna kriterier i min studie. Intressanta fynd var att även om informanterna har motiverat deras bedömningar genom att nämna olika kriterier i skriftliga motiveringar skulle deras utvärderingar ändå samstämma numeriskt. Det kan betyda att de har processat bedömningen genom flera kriterier men helhetsbilden har bestämt vitsordet.

I jämförelse med mina hypoteser finns det ingen indikation på att erfarenhet påverkar tillämpningen av YKI-kriterierna eftersom det finns samstämmighet mellan erfarna bedömare (se t.ex. Maria och Helena) och icke-erfarna bedömare (se t.ex. Eeva och Mia). Men resultaten stämmer med hypotesen att bedömare kan betona kriterier på olika sätt i sina motiveringar. Vidare stämmer resultaten med Juutilainen (2011) att informanterna kan använda kriterier utanför givna kriterier. I min studie finns det t.ex. subjektivt laddad bedömning eller jämförelse med andra prestationer.

Nyttan som man kan ha av resultaten är att man får en inblick i vad som man måste ta hänsyn till när man i framtiden ska forma en muntlig test och bedömning av detta i studentexamen i Finland. Det är t.ex. dubbelbedömning av prestationerna och kontrollering av interbedömarreliabiliteten som är viktiga om man tänker studentexamen eftersom den påverkar t.ex. karriärplaner i framtiden.

Jag kan inte dra för direkta slutsatser eftersom det empiriska underlaget är relativt litet. Det kan också påverka resultaten att informanterna inte visste testnivån som testtagaren hade valt. Vidare visste mina informanter inte hurdan uppgift som en helhet testtagarna har gjort. YKI-bedömarna har haft denna information när de har bedömt prestationer.

Å ena sidan kan informanternas bedömningar visa den verkliga nivån av talarens kunskaper eftersom det finns en risk att talaren inte tar testet i passande nivå utan i nivån som han klarar utmärkt eller klarar inte av. Vidare kan det påverka bedömningens validitet om bedömaren redan vet att prestationen har tags på en viss nivå. YKI-bedömarna har haft förkunskaper om nivån som testtagaren har tagit testet vilket kan påverka YKI-bedömarnas utvärdering omedvetet.

Metoden fungerade bra för den här studien eftersom det gav mig svar för mina forskningsfrågor. Det finns risken att i kvalitativa analysen av materialet är litet och kvantifieringen ger ingen ytterligare information (se Tuomi & Sarajärvi 2002: 119). I min studie gav kvantifiering nytt perspektiv till analysen och klargjorde analysen av resultaten eftersom det var 10 bedömningar och 8 informanter med i studien.

Validitet i studien har försäkrats genom att analysera materialet induktivt. I analysprocessen har det varit målet att information inte förloras. Det är ändå klart att jag har byggt upp temakategorierna vilket kan påverka resultaten. Reliabilitet i studien försäkrades genom en strikt analys genom att följa innehållsanalysen stegvis. Det betyder att det förminskas risken att resultaten skulle variera i framtiden.

En fördel med min studie är att jag har autentiska och färdiga muntliga prestationer hämtade från YKI-korpusen som informanterna får bedöma. En annan fördel är också att informanternas svar analyserades effektivt eftersom svaren var skriftliga i stället för att spela in svaren. Ytterligare efter att jag hade insamlat materialet (m.a.o. informanternas bedömningar) hade jag alltid tillgång till materialet och det ger mig möjlighet att analysera och skriva avhandlingen när som helst.

En nackdel i den här typen av studien är att informanterna motiverar deras utvärderingar på olika sätt vilket orsakar variation i svaren. Informanterna måste motivera varje bedömning och jag var medveten om att det finns en möjlighet att variation i svaren t.ex. bero på hur de tolkar kriterier eller hurdana egenskaper i tal de uppskattar.

6. AVSLUTNING

Som en helhet svarade min studie på mina forskningsfrågor. Det finns ändå några utmaningar i genomförandet av studien. En utmaning var att locka informanternas intresse att ta del i min studie. Därför lovade jag skicka feedback om deras utvärderingar per epost efter att studien var färdig. Vidare var det en utmaning att distribuera muntliga prestationerna till informanterna utan att de laddar ner prestationerna vilket enligt SOLKI:s regler var förbjudet. Detta problem löste jag med SoundCloud -program som var ett effektivt verktyg att utdela prestationerna men samtidigt säkerställde jag att informanterna inte kan ladda ned prestationerna.

Viktigaste fyndet i min studie är att informanternas bedömningar inte helt motsvarade YKI-bedömarnas bedömningar. Det kan bero på att utbildning och förkunskaper som YKI-bedömarna har fått tidigare kan påverka bedömningen i YKI-testet. Det kan ha något med haloeffekten att göra. Det kan också påverka att jag skulle ha givit tydligare instruktioner till mina informanter om nivån som informanterna hade tagit testet. Det visar betydelsen av utbildning före bedömningsarbetet utförs. Men intressant var att informanternas svar visade till viss mån samstämmighet och var samstämmigare mellan varandra än med YKI-bedömarna. Studien visar också att lärare i undervisningsfältet kan tillämpa givna YKI-kriterierna för muntlig färdighet även om de inte känner dessa kriterier förut. Det visar att YKI-kriterierna för muntlig färdighet möjligen skulle kunna imiteras i bedömning av muntligt prov i studentexamen.

Min studie är en relevant pusselbit för att visa variation i bedömning av muntlig färdighet. Det kan förväntas att i framtiden finns det ännu mer intresse att försäkra rättssäker bedömning av muntlig färdighet eftersom trenden har blivit att bedöma muntliga kunskaper vid sidan av skriftliga kunskaper. Min studie är därför en relevant pusselbit till studier inom språkbedömning. Att vara medveten av faktorer som kan påverka sin bedömning är ett steg mot samstämmig bedömning som kan ses i resultaten av denna studie.

I framtiden kräver närmare granskning om en taltest i studiomiljön imiterar en äkta talsituation och visar kandidatens kunskaper i naturliga talsituationer. Jag funderar också om pragmatiska kunskaper kan mätas i en studiomiljö. Vidare tänkte jag över om det finns risken att systematiken och datordriven kontroll av muntliga prestationer förlorar information om testtagarens reella kunskaper i ett språk. I framtiden finns det ännu möjlighet att datorstödd bedömning tar

fotfäste i bedömning av muntlig färdighet som man redan gör t.ex. i PTE test. Dessa är heta frågor t.ex. i planering av ett muntligt prov i studentexamen i Finland under de kommande åren.

LITTERATUR

Ahola, S., & Hirvelä, T. 2016. *Mikä merkitys osallistujan taustatekijöillä on menestymiseen kielitestissä?* Kieli, koulutus ja yhteiskunta. https://jyx.jyu.fi/dspace/handle/123456789/51355. (Hämtad 7.2.2017)

Ahola, S; Hirvelä, T; Härmälä, M; Lammervo, T; Neittaanmäki, R & Tossavainen, H. 2015. *Esimerkkejä YKIn ydinalueita koskevista tutkimus- ja kehittämiskohteista.* https://www.jyu.fi/hytk/fi/laitokset/solki/tutkimus/hankkeet/yki/tutkimusjakehittaminen

Ahola, S. 2012. *Yleisten kielitutkintojen laatijoiden käsityksiä kielestä ja tehtävien laadinnasta.* Jyväskylä: Centralen för tillämpad språkforskning. Licentiatexamen. Jyväskylä universitet. https://jyx.jyu.fi/dspace/bitstream/handle/123456789/40753/URN%3aNBN%3afi%3ajyu-201301241110.pdf?sequence=1 (Hämtad 8.2.2017)

Alderson, J. 1991. *Language Testing in the 1990s: How Far Have We Come? How Much Further Have We to Go?* Institute of education sciences. https://eric.ed.gov/?id=ED365145. (Hämtad 8.2.2017)

ALTE=Association of language testers in Europe. http://www.alte.org/. (Hämtad 8.2.2017)

Bachman, L & Palmer, A. 1996. *Language testing in practice: Designing and developing useful language tests*. Oxford: Oxford University Press.

Brown, A. 2004. *Discourse analysis and the oral interview: Competence or performance*? I: Boxer, D; Cohen, A. Studying speaking to inform second language learning. (red.) Clevedon: Multilingual matters ltd. 253–301.

SOLKI 2017a=Centret för tillämpad språkforskning. 2017. *Kielitutkintojen arviointi ja todistus*. https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista/tietoakielitutkinnoista/arviointi. (Hämtad 9.2.2017)

SOLKI 2017b=Centret för tillämpad språkforskning. 2017. *Puhumisen arviointikriteerit*. https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista/tietoakielitutkinnoista/puhumisen_arviointikriteerit.pdf. (Hämtad 9.2.2017)

SOLKI 2016=Centret för tillämpad språkforskning. 2016. *Kielitutkintojen rakenne ja sisältö*. https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista/tietoakielitutkinnoista/rakenne. (Hämtad 9.2.2017).

SOLKI 2011a=Centret för tillämpad språkforskning. 2011. *Osakokeet ja aihealueet.* https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista/osat_aihealueet. (Hämtad 9.2.2017)

SOLKI 2011b=Centret för tillämpad språkforskning. 2011. *Puhuminen.* https://www.jyu.fi/hytk/fi/laitokset/solki/yki/yleista/osat_aihealueet/puhuminen. (9.2.2017)

Cohen, A. D. 2004. *Assessing speech acts in a second language*. I: Boxer, D; Cohen, A. Studying speaking to inform second language learning. (red.) Clevedon: Multilingual matters ltd. 302–324

Davies, A; Brown, A; Elder, C; Hill, K; Lumley, T & McNamara, T. 1999. *Studies in language testing-Dictionary of language testing*. Cambridge: University of Cambridge Local Examinations Syndicate.

EALTA=European association for language testing and assessment. http://www.ealta.eu.org/ (Hämtad 8.2.2017)

Lag om allmänna språkexamina 964/2004. http://www.finlex.fi/sv/laki/ajantasa/2004/20040964#P7. (Hämtad 30.12.2016)

Fransson, I-L. 2010. *Lärares bedömning av andraspråkstexter – uppfattningar om språkbehärskning, nivå och betyg*. Uppsats. Linnéuniversitet. http://www.diva-portal.org/smash/get/diva2:329137/FULLTEXT01.pdf. (Hämtad 10.2.2017)

GERS=Gemensam europeisk referensram för språk. Skolverket 2009. *Gemensam europeisk*

referensram för språk: lärande, undervisning och bedömning. (Hämtad 9.2.2017)

Grube, S. 2012. *Språkbedömning i samband med pedagogiska utredningar av elever med annat modersmål än svenska inför mottagande i särskolan- en kartläggning av rutiner och granskning av språkliga bedömningsverktyg.* Uppsats. Stockholms universitet. http://www.andrasprak.su.se/polopoly_fs/1.177233.1400051321!/menu/standard/file/Susanne_Grube_kandidatuppsats.pdf. (Hämtad 10.2.2017)

Huhta, A. & Takala, S. 1999. *Kielitaidon arviointi*. I: Kielenoppimisen kysymyksiä, Sajavaara, K.& Marsh-Piirainen, A. (red.) Jyväskylä: Jyväskylän yliopistopaino

Huhta A. & Hildén, R. 2013. *Kielitaidon arvioinnin metodologisia vaihtoehtoja.* I: Räisänen, A. (red.) Oppimisen arvioinnin kontekstit ja käytännöt. Utbildningsstyrelsen. https://karvi.fi/app/uploads/2013/09/OPH_R0313.pdf. (Hämtad 2.3.2017)

Isaacs, T. 2016. *Assessing speaking.* I: D. Tsagari & J. Banerjee (red.): *Handbook of second*

language assessment. Berlin: DeGruyter Mouton. 131–146. http://nebula.wsimg.com/cff6bea3be4b70ca68aa14a3f5e185b0?AcessKeyId=EF0C57FC220BDEC3825D&disposition=0&alloworigin=1. (Hämtad 28.10.2016)

Jakku-Sihvonen, R. 2013. *Oppimistulosten arviointijärjestelmistä ja niiden kehittämishaasteista.* I: Räisänen, A. (red.) Oppimisen arvioinnin kontekstit ja käytännöt. Utbildningsstyrelsen. https://karvi.fi/app/uploads/2013/09/OPH_R0313.pdf. (Hämtad 3.3.2017)

Juutilainen, M. 2011. *Att bedöma skriftliga färdigheter i svenska med hjälp av en nivåskala- hur skillnader mellan bedömare uppstår.* Kandidatavhandling. Jyväskylä universitet.

Kane, M. 2013. *The Argument-Based Approach to Validation*. Education Testing Services. School Psychology Review, upplaga 42, nr. 4, s.1–11

Keurulainen, H. 2013. *Pelisääntöjä arviointipäätösten tekemistä varten.* I: Räisänen, A. (red.) Oppimisen arvioinnin kontekstit ja käytännöt. Utbildningsstyrelsen. Juvenes Print-Suomen yliopistopaino Oy.

Korkeakivi, R. 2014. *Mihin opettaja tarvitsee veso-päiviä? Miksei veso-päivinä saa lomauttaa*? Opettaja 43/2014. s. 27

Lindroos, S. 2010. *Bedömningen av muntliga språkkunskaper i svenska: en analys av bedömarnas reflektioner kring elevbedömningens reliabilitet*. Pro gradu. Helsingfors universitet.

McNamara, T. 2015. *Language testing*. Oxford: Oxford University Press

McNamara, T. 1996. *Measuring second language performance.* New York: Addison Wesley Longman limited

Nordström, J. 2005. *Olika lärare, olika betyg-om (o)likvärdig bedömning av elevtexter i år 9.* Växjö universitet. http://www.diva-portal.org/smash/get/diva2:206766/FULLTEXT01.pdf. (Hämtad 10.2.2017)

Ohranen, S. Projektforkare i Centret för tillämpad språkforskning. (Fråga om muntliga prestationer i YKI-korpusen, epostkommunikation) 15.12.2016

PTE= Pearson test of english academic. 2014. *Pearson test of english academic: automated scoring*. http://pearsonpte.com/wp-content/uploads/2015/05/7.-PTEA_Automated_Scoring.pdf. (Hämtad 3.3.2017)

Ruohonen, M. 2016. *Innehållsanalys av grunderna för läroplanen 2004 och 2014-Vad som förnyas i grunderna för läroplanen av den medellånga lärokursen i svenska*? Kandidatavhandling. Jyväskylä universitet. https://jyx.jyu.fi/dspace/bitstream/handle/123456789/49387/URN-NBN-fi-jyu-201604212277.pdf?sequence=4 (Hämtad 19.4.2017)

Spolsky, B. 1990. *Conditions for second language learning- introduction to a general theory*. Oxford University Press

Stolt, S. 2016. *Kollegial bedömning i summativa bedömarkommentarer på studentexamensuppsatser i modersmål och litteratur*. I: Huhta, A; Hildén, R. (red.) Kielitaidon arviointitutkimus 2000-luvun Suomessa. AFinLA-e. Soveltavan kielitieteen tutkimuksia 2016/no:9. file:///C:/Users/mintt/AppData/Local/Temp/4267-228-PB.pdf. (Hämtad 2.3.2017)

Tarnanen, M. 2007. *Testiaineistosta kielenoppijakorpukseksi*. I: Salo, O.-P., T. Nikula & P. Kalaja (red.) Kieli oppimisessa – Language in Learning. AFinLAs årbok 2007. Finländska föreningen för tillämpad språkvetenskaps publikationer nr. 65. Jyväskylä. s. 197–213

Topling. 2012. *Inlärningsgångar i andraspråket.* https://www.jyu.fi/hytk/fi/laitokset/kivi/tutkimus/hankkeet/paattyneet-tutkimushankkeet-kansio/topling/se. (Hämtad 2.3.2017)

Toropainen, O; Härmälä, M & Lahtinen, S. 2012. *Kaksi asteikkoa, kaksi eri tilannetta: äidinkielellä ja vieraalla kielellä kirjoitettujen tekstien kriteeripohjaisen arvioinnin haasteita.* I: Meriläinen, L; Kolehmainen, L; Nieminen, T. (red.) AFinLA-e. Soveltavan kielitieteen tutkimuksia 2012/ no: 4. 60-79. file:///C:/Users/mintt/AppData/Local/Temp/7038-1-16632-1-10-20121025.pdf. (Hämtad 2.3.2017)

Tuomi, J & Sarajärvi, A. 2002. *Laadullinen tutkimus ja sisällönanalyysi*. Helsingfors: Tammi

Tuomi, J. & Sarajärvi, A. 2009. *Laadullinen tutkimus ja sisällönanalyysi*. Helsingfors: Tammi.

Utbildningsstyrelsen. 2011. *Grunderna för allmänna språkexamina.* Tammerfors. Juvenes Prints-Tampereen yliopistopaino Oy. http://oph.fi/download/141378_valmisykiperusteetruo2011.pdf. (Hämtad 9.2.2017)

Sekundära:

Lundeberg, O. 1929. *Recent developments in audition-speech tests*. The Modern Language Journal 14(3). 193–202.

Carroll, J. B. 1961. Fundamental considerations in testing for English language proficiency of foreign students. Testing. Washington DC: Center for Applied Linguistics.

Wood, B. 1927. *New York experiments with new-type modern language tests*. New York: MacMillan.

7. BILAGA

Bilaga 1: Kriterier för muntlig färdighet (YKI-kriterierna, SOLKI 2016)

Yleiset kielitutkinnot

Puhumisen arviointikriteerit

Kaikkia tehtäviä koskevat kriteerit (Huom. Suorituksen arvio muotoutuu kokonaisuudesta, ei yhden yksittäisen kriteerin mukaan.)

TASO 6 Yleiskriteerit	Sujuvuus	Joustavuus	Koherenssi/ sidosteisuus	Ilmaisun tarkkuus/ laajuus/ idiomaattisuus	Ääntäminen/ fonologinen hallinta	Rakenteiden tarkkuus
Puhuu erittäin sujuvasti ja puheessa esiintyy vain satunnaisesti eikohdekielisiä piirteitä, kuten vierasta korostusta. Pystyy ilmaisemaan täsmällisesti hienojakin merkitysvivahteita, ja käyttää myös idiomaattisia ilmauksia monipuolisesti ja tarkoituksenmukaisesti. Pystyy kuvailemaan monimutkaistakin aihetta sekä liittämään kuvaukseen alateemoja, kehittelemään erilaisia näkökulmia ja päättämään esityksen sopivalla tavalla.	Pystyy ilmaisemaan itseään sujuvasti, luontevasti ja eräröimättä pidemmässäkin puhetuotoksessa. Pysähtyy vain joskus pohtimaan oikeaa sanaa ilmaistakseen ajatuksensa tai löytääkseen sopivan esimerkin tai selityksen.	Pystyy joustavasti muotoilemaan ajatuksiaan uudelleen käyttäen erilaisia kielellisiä muotoja - karsiakseen epäselvyyksiä - osoittaakseen painotuksia ja - mukauttaakseen puheensa vastaanottajan ja tilanteen mukaan.	Pystyy tuottamaan sisällöltään ja muodoltaan johdonmukaista puhetta käyttäen monipuolisesti ja tarkoituksenmukaisesti erilaisia keskustelun jäsentämistapoja ja koherenssikeinoja.	Hallitsee erittäin laajan sanaston ja käyttää johdonmukaisesti ja asianmukaisesti idiomaattisia ja sivumerkitykset esiin tuovia ilmauksia (ironia, leikinlasku). Pystyy ilmaisemaan hienoimpiakin merkitysvivahteita täsmällisesti. Käyttää tarkasti ja varioivasti erilaisia tarkennuskeinoja (esim. astetta ilmaisevia adverbejä).	Kuten taso 5: Osaa varioida intonaatiota ja asettaa lausepainon oikein ilmaistakseen hienojakin merkitysvivahteiden eroja.	Hallitsee komplek-sisetkin rakenteet johdonmukaisesti. Puheessa esiintyy vain joitakin satunnaisia lipsahduksia.

TASO 5 > 6

- Ilmaisu on tarkempaa ja varioivampaa.
- Puhe on idiomaattisempaa.
- Sisällön abstraktiotaso on korkea.

TASO 5 Yleiskriteerit	Sujuvuus	Joustavuus	Koherenssi/ sidosteisuus	Ilmaisun tarkkuus/ laajuus/ idiomaattisuus	Ääntäminen/ fonologinen hallinta	Rakenteiden tarkkuus
Puhuu sujuvasti tarvitsematta useinkaan selvästi hakea ilmauksia. Puheenvuorot ovat luontevia, yhtenäisiä ja sopivan pituisia. Pystyy esittämään selkeän, yksityiskohtaisen kuvauksen monimutkaisestakin aiheesta. Osaa käyttää idiomaattisia ilmauksia ja arkielämän sanontoja ja pystyy ilmaisemaan sävyeroja kohtalaisen hyvin.	Pystyy ilmaisemaan itseään jokseenkin luontevasti, sujuvasti ja spontaanisti. Vain käsitteellisesti vaikeissa aiheissa voi esiintyä kielellistä epäröintiä.	Kuten taso 4: Pystyy tavanomaisissa ja vähän vieraammissakin tilanteissa sovittamaan sanottavansa tilanteen ja vastaanottajan mukaan.	Pystyy tuottamaan selkeää ja loogisesti etenevää puhetta ja osoittaa hallitsevansa hyvin keskustelun jäsentämistapoja ja koherenssikeinoja.	Hallitsee yleisiä aihepiirejä koskevan sanaston hyvin. Löytää vaivatta kiertoilmaisuja ja käyttää idiomaattisia ilmauksia melko luontevasti. Joutuu vain harvoin turvautumaan ilmausten hakemiseen ja välttämisstrategioihin. Sanastolliset lipsahdukset ovat vähäisiä. Pystyy ilmaisemaan täsmällisesti varmuutta/epävarmuutta, luottamusta/epäilyä, todennäköisyyttä.	Osaa varioida intonaatiota ja asettaa lausepainon oikein ilmaistakseen hienojakin merkitysvivahteiden eroja.	Hallitsee lähes kaikki rakenteet virheettömästi. Puheessa esiintyy joitakin epäluontevia muotoja.

TASO 4 Yleiskriteerit	Sujuvuus	Joustavuus	Koherenssi/ sidosteisuus	Ilmaisun tarkkuus/ laajuus/ idiomaattisuus	Ääntäminen/ fonologinen hallinta	Rakenteiden tarkkuus
Selviää melko hyvin vieraammissakin viestintätilanteissa. Erottaa puheessaan muodollisen ja epämuodollisen kielimuodon ainakin jossain määrin. Pystyy esittämään ja perustelemaan mielipiteensä ymmärrettävästi. Pystyy kertomaan ja kuvailemaan näkemäänsä, kuulemaansa ja kokemaansa. Joutuu vain harvoin käyttämään kiertoilmauksia arkielämän puhetilanteissa kielitaidon puutteellisuuden vuoksi.	Puhuu suhteellisen sujuvasti. Pitkät tauot ovat harvinaisia vaikka epäröintiä voi esiintyä rakenteita ja ilmauksia etsiessä.	Pystyy tavanomaisissa ja vähän vieraammissakin tilanteissa sovittamaan sanottavansa tilanteen ja vastaanottajan mukaan.	Pystyy käyttämään erilaisia kielellisiä keinoja osoittaakseen ajatusten väliset suhteet ja saadakseen aikaan koherentin tuotoksen. Pitkissä puheosuuksissa voi esiintyä jonkin verran koherenssin puutetta.	Pystyy ilmaisemaan sanottavansa tarkasti ja melko yksityiskohtaisesti. Hallitsee useimpia yleisiä aihepiirejä koskevan sanaston. Pystyy välttämään liiallista toistoa. Sanastossa voi olla epätarkkuutta, mutta se ei estä ymmärtämistä.	On omaksunut selkeän ääntämyksen ja intonaation, vaikka ääntämisessä on puutteita.	Hallitsee rakenteet suhteellisen hyvin. Rakenteissa esiintyy vähäisiä puutteita, jotka eivät kuitenkaan aiheuta väärinkäsityksiä.

TASO 3 > 4

- Ilmaisu on varioivampaa
- Sisällön abstraktiotaso on korkeampi
- Puhuja tekee eron muodollisen ja epämuodollisen rekisterin välillä

TASO 3 Yleiskriteerit	Sujuvuus	Joustavuus	Koherenssi/ sidosteisuus	Ilmaisun tarkkuus/ laajuus/ idiomaattisuus	Ääntäminen/ fonologinen hallinta	Rakenteiden tarkkuus
Selviää tavallisimmissa käytännön puhetilanteissa ja pystyy olemaan aloitteellinen jokapäiväisissä kielenkäyttötilanteissa. Puhe voi olla melko hidasta, mutta epäluontevia katkoja ei esiinny kovin paljon. Tulee ymmärretyksi siitä huolimatta, että siirtää äidinkielen tai muiden kielten rakenteita ja sanastoa kohdekieleen, ja ääntäminen saattaa olla selvästi ei-kohdekielenomaista.	Pystyy ilmaisemaan itseään ymmärrettävästi ja suhteellisen helposti ilman apua. Tauot, jotka johtuvat puheen muotoiluun liittyvistä ongelmista, ovat yleisiä etenkin pidemmissä yhtäjaksoisissa tuotoksissa.	Selviytyy arkipäivän viestintätilanteista hyvin, mutta muodollisen ja epämuodollisen rekisterin vaihtelu on puutteellista.	Pystyy yhdistämään ilmauksia yhtenäiseksi ja johdonmukaiseksi puheeksi, vaikka sidoskeinojen käyttö voi olla puutteellista ja toistavaa.	Pystyy ilmaisemaan sanottavansa ymmärrettävästi ja melko tarkasti. Virheitä esiintyy kompleksisia ajatuksia ilmaistaessa tai silloin, kun kyseessä on vähemmän tuttu tilanne tai aihe. Hallitsee riittävän laajan sanaston tullakseen toimeen jokapäiväisen elämän tilanteissa (esim. perhe, työ, harrastukset).	Ääntäminen on selvästi ymmärrettävää, vaikka siinä on puutteita.	Rakenteiden käyttö on kohdekielenomaista tutuissa tilanteissa. Vaikka vaativammissa tilanteissa rakenteiden käyttö on epäluontevaa, viesti tulee kuitenkin ymmärretyksi.

TASO 2 Yleiskriteerit	Sujuvuus	Joustavuus	Koherenssi/ sidosteisuus	Ilmaisun tarkkuus/ laajuus/ idiomaattisuus	Ääntäminen/ fonologinen hallinta	Rakenteiden tarkkuus
Selviää vain rutiininomaisissa puhetilanteissa, jotka vaativat yksinkertaista tiedonvaihtoa. Kielitaidon vähäisyys rajaa paljolti sitä, mitä asioita puhuja pystyy käsittelemään. Viestin perillemeno edellyttää, että puhekumppani on valmis auttamaan puhujaa sanottavansa muotoilemisessa. Ääntäminen voi olla hyvin ei-kohdekielenomaista, mikä vaatii kuulijalta paljon ja vaikeuttaa viestin perillemenoa.	Pystyy tekemään itsensä ymmärretyksi lyhyissä aihepiiriltään tutuissa puhetilanteissa, vaikkakin tauot ja epäröinnit ovat hyvin tavallisia.	Pystyy käyttämään fraaseina oppimiaan yksinkertaisia/tavallisimpia ilmauksia arkipäiväisissä tilanteissa.	Pystyy yhdistämään tavanomaisimmilla sidesanoilla (kuten "ja", "mutta" tai "koska") yksinkertaisia lauseita.	Pystyy ilmaisemaan asiat yksinkertaisesti tutuissa rutiininomaisissa viestintätilanteissa. Hallitsee suppeaa, konkreettisiin ja jokapäiväisiin tarpeisiin liittyvää sanastoa.	Ääntäminen on yleensä niin selkeää, että puhe on ymmärrettävää selvästä vierasperäisestä korostuksesta huolimatta. Keskustelukumppanit joutuvat pyytämään toistoa aika ajoin.	Käyttää joitakin yksinkertaisia rakenteita oikein, mutta tekee silti systemaattisesti virheitä perusrakenteissa. Viesti tulee kuitenkin yleensä selväksi.

TASO 1 > 2

- Puhe on yhtäjaksoisempaa.
- Puheessa on enemmän sisältöä.
- Puhuja yrittää aktiivisemmin tuottaa puhetta.

TASO 1 Yleiskriteerit	Sujuvuus	Joustavuus	Koherenssi/ sidosteisuus	Ilmaisun tarkkuus/ laajuus/ idiomaattisuus	Ääntäminen/ fonologinen hallinta	Rakenteiden tarkkuus
Selviää kaikkein yksinkertaisimmissa puhetilanteissa, mutta joutuu käyttämään runsaasti ei-kielellisiä keinoja tullakseen ymmärretyksi. Viestintä on hidasta ja hyvin katkonaista. Pystyy käyttämään yksinkertaisia kohteliaisuusmuotoja ja kysymään ja vastaamaan yksinkertaisiin kysymyksiin, jotka käsittelevät välittömiä, jokapäiväisiä tarpeita.	Käyttää vain hyvin lyhyitä ja tavallisimpia fraaseina oppimiaan ilmauksia. Pysähtyy usein etsimään ilmausta, ääntämään vähemmän tuttuja sanoja ja paikkailemaan kommunikaatiota.	Pystyy reagoimaan yksinkertaisiin kysymyksiin ja repliikkeihin.	Pystyy liittämään sanoja tai sanaryhmiä toisiinsa käyttäen muutamia sidesanoja, kuten "ja" tai "sitten".	Hallitsee yksittäisiä irrallisia sanoja ja ilmauksia, jotka liittyvät jokapäiväisiin yksinkertaisiin tilanteisiin.	Puhujan vähäisen tuotoksen ymmärtäminen vaatii puheeseen totuttelemista ja ponnistelemista.	Hallitsee vain muutamia yksinkertaisia rakenteita ja fraaseina oppimiaan lauseita.

ALLE 1	
	• **Puhe koostuu yksittäisistä irrallisista sanoista.** • **Puhuja pystyy sanomaan oman nimensä ja 12 asiaa itsestään (esim. missä asuu ja mitä tekee työkseen).** • **Puhetta on mahdoton ymmärtää.**

Bilaga 2: Formulär med de första upplysningarna

Esitietolomake

Luettuasi tutkimuksen infokirjeen, täytä huolellisesti tämä esitietolomake. Lomakkeen täyttäminen toimii suostumuksena tutkimukseen osallistumiselle

Ohjeet lomakkeen täyttämiseen:
Vastaa jokaiseen annettuun kohtaan huolellisesti. Kohdissa vastaus annetaan sanallisesti, kohdissa 7-9 arvioidaan osaamista numeerisesti asteikolla 1-5 (1=en lainkaan, 5=erittäin hyvä) ja vastausta voi täydentää perustelu kohdassa.

Tiedot	Vastauskenttä:	
Ikä		
Sukupuoli		
S-posti osoite (johon haluat palautteen arvioinnista):		
Opintoasteet, joilla olet tehnyt arviointityötä (kirjaa kaikki)	Perusaste (peruskoulu, muu)	
	Toinen aste (lukio, ammattikoulu, muu)	
	Kolmas aste (yliopisto, amk, muu)	
Opintoaste, jolla opetat tällä hetkellä		
Työvuodet opettajana		
Opetettavat oppiaineet		

Täytä alla olevat tiedot mahdollisimman huolellisesti.

Merkitse kirjaimella ”X” siihen kohtaan, miten koet seuraavat väittämät. Perustele.

Puheen suoritusten arviointikokemus (1=ei lainkaan, 5=erittäin hyvä):

		3.	4.	5.

Perustelut:

Eurooppalaisen viitekehyksen tuntemus: (1=ei lainkaan, 5=erittäin hyvä):

1.	2.	3.	4.	5.

Perustelut:

Yleisen kielitutkinnon (SOLKI) tuntemus: (1=ei lainkaan, 5=erittäin hyvä):

1.	2.	3.	4.	5.

Perustelut:

Bilaga 3: Svarsformulär

Vastauslomake

Kuuntele puheensuoritukset ja täytä taulukko siten, että annat kullekin suoritukselle taitotasoarvion (taso 1-6, ks. LIITE 4). Perustele arviosi omin sanoin (voit käyttää arviointikriteereitä apuna, ks. LIITE 4).

Suori-tus	Tasoarvio	Perustelut
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		