Pre-print: Ferrari S., Nuzzo E. (2010), *La valutazione delle competenze orali in italiano L2. Una verifica sperimentale dei criteri suggeriti dal Quadro Comune Europeo* (con E. Nuzzo), in E. Lugarini (a cura di), «Valutare le competenze linguistiche, Atti del XV Convegno nazionale GISCEL, Milano, 6-8 marzo 2008», Franco Angeli, Milano, pp. 279-293

La valutazione delle competenze orali in italiano L2: una verifica sperimentale dell'affidabilità dei criteri suggeriti dal QCER

di *Stefania Ferrari, Elena Nuzzo**

> *It is the duty of research to question current wisdom, problematize accepted 'certainties', to investigate the problems and seek solutions.* C. J. Alderson (2005)

0. Introduzione

Il *Quadro Comune Europeo di Riferimento per le lingue* (d'ora in avanti QCER, Consiglio d'Europa, 2001) è il risultato del lavoro trentennale del Consiglio d'Europa nell'elaborazione di scale e descrittori dedicati alle competenze e agli usi linguistici. Oggi è considerato il punto di riferimento ufficiale per l'insegnamento e la valutazione degli apprendimenti linguistici. Il QCER è un documento complesso e ambizioso che si offre come strumento di lavoro a tutti coloro che si occupano di apprendimento linguistico: docenti e studenti naturalmente, ma anche autori di programmi e curricoli, responsabili delle politiche linguistiche, esponenti del mondo dell'editoria e delle certificazioni. Gli autori del QCER descrivono così il documento e le sue finalità:

> Il *Quadro Comune Europeo di Riferimento per le lingue* fornisce una base comune in tutta l'Europa per l'elaborazione di programmi, linee guida curricolari, esami, libri di testo per le lingue moderne ecc. Descrive in modo esaustivo ciò che chi studia una lingua deve imparare per usarla per comunicare e indica quali conoscenze e abilità deve sviluppare per agire in modo efficace. La descrizione riguarda anche il contesto culturale nel quale la lingua si situa. Inoltre il *Quadro di Riferimento* definisce i livelli di competenza che permettono di misurare il progresso dell'apprendimento ad ogni stadio del percorso, nella prospettiva dell'educazione permanente (QCER: 1).

* Università di Verona. Il lavoro nasce da una stretta collaborazione tra le due autrici. Stefania Ferrari è responsabile della stesura dei paragrafi 0, 1 e 3.1, Elena Nuzzo è responsabile della stesura dei paragrafi 2, 3.2 e 4.

Il QCER pone particolare attenzione ai temi della valutazione e della certificazione ed esplicita chiaramente le sue possibili funzioni in questi due ambiti.

Rispetto alla valutazione della competenza linguistica le possibilità d'uso sono tre:

1. per specificare il contenuto di test ed esami;
2. per definire i criteri in base ai quali un obiettivo di apprendimento si può considerare raggiunto, con riferimento sia alla valutazione di una particolare prestazione orale o scritta, sia alla valutazione continua dell'insegnante, dei pari e all'autovalutazione;
3. per descrivere i livelli di competenza rilevati in test ed esami esistenti, in modo da fare confronti tra sistemi di certificazione diversi (QCER: 24).

Sin dalla sua pubblicazione il QCER ha stimolato l'interesse di chi si occupa di valutazione, e la sua applicazione è stata oggetto di sperimentazione in diversi contesti di insegnamento e valutazione a livello europeo (cfr. "Progetto Lingue Moderne"; "The Dutch CEF Construct Project Dialang", in Alderson 2002, 2005; www.coe.int/portfolio) e in ambito italiano (cfr. "Progetto Lingue 2000"; Grego Bolli 2006; www.coe.int/portfolio). Questi studi, che pure riconoscono le potenzialità e il valore del documento, mettono in luce alcuni dei limiti nella sua applicazione al testing, evidenziando la necessità di indicazioni più dettagliate ai fini dell'implementazione.

In Italia la valutazione costituisce un problema spinoso per gli operatori della scuola, che si trovano spesso nell'urgenza di certificare le competenze degli allievi. A costoro il QCER offre la possibilità di disporre di un linguaggio comune per definire le competenze linguistiche di studenti italiani e stranieri.

Proprio nel contesto specifico della scuola si colloca la nostra ricerca, che si pone l'obiettivo di verificare se il QCER possa essere, come crediamo, un utile punto di riferimento per gli insegnanti nella valutazione delle competenze orali e in che modo si possano valorizzare al meglio le sue potenzialità. Per fare questo abbiamo elaborato un protocollo di ricerca che ci permettesse di indagare sperimentalmente le possibilità di utilizzo del QCER per la valutazione delle competenze orali in italiano L2 e, in particolare, di verificare l'affidabilità del documento così come formulato nella sua attuale versione.

La prima parte di questo contributo sarà dedicata a una breve presentazione del QCER e del suo ruolo nella scuola, cui seguirà l'illustrazione della metodologia utilizzata e del tipo di dati raccolti. I paragrafi successivi presenteranno l'analisi dei dati e i risultati. Infine si passerà alle conclusioni e alle possibili applicazioni pratiche dei risultati ottenuti.

Pre-print: Ferrari S., Nuzzo E. (2010), *La valutazione delle competenze orali in italiano L2. Una verifica sperimentale dei criteri suggeriti dal Quadro Comune Europeo* (con E. Nuzzo), in E. Lugarini (a cura di), «Valutare le competenze linguistiche, Atti del XV Convegno nazionale GISCEL, Milano, 6-8 marzo 2008», Franco Angeli, Milano, pp. 279-293

1. Il QCER per la valutazione

L'approccio di fondo che caratterizza il QCER[1] è l'orientamento all'azione. La competenza linguistica viene quindi formulata attraverso una serie di scale costituite da descrittori in cui si illustra che cosa l'apprendente è in grado di fare nello svolgimento di un compito comunicativo, in un determinato contesto, utilizzando un certo repertorio di risorse linguistiche. Il documento permette di descrivere la competenza dell'apprendente secondo due dimensioni: una verticale, data dalle scale globali, e una orizzontale, costituita dai parametri per le attività comunicative e la competenza linguistico-comunicativa (QCER, capp. 4 e 5).

Il QCER propone una tripartizione dei livelli che risultano così suddivisi in A (livello elementare), B (livello intermedio) e C (livello avanzato). A sua volta ciascuno dei tre livelli è suddiviso in due ulteriori sotto-livelli, dando origine così a una scala a 6 bande, potenzialmente ancora suddivisibile secondo un sistema ad albero.

Per riuscire a realizzare appieno le sue funzioni, in particolare nell'ambito della valutazione, il QCER deve soddisfare alcuni criteri: deve cioè essere esaustivo[2], trasparente e coerente. Ecco come vengono definiti tali criteri nel QCER:

- esaustivo: dovrebbe cercare di specificare nel modo più ampio possibile la gamma di conoscenze, capacità e usi linguistici (QCER: 8)
- trasparente: l'informazione deve essere formulata in modo chiaro ed esplicito, in modo da essere accessibile e facilmente comprensibile alle persone interessate (QCER: 8)
- coerente**:** la descrizione deve essere scevra da contraddizioni interne (QCER: 9)

Al tempo stesso, essendo un documento non prescrittivo, ma di riferimento, il QCER è multifunzionale, flessibile, aperto, dinamico, amichevole, e soprattutto non dogmatico (QCER: 9). Ecco perché gli esempi sono semplicemente indicativi e il lettore è invitato a usare scale e descrittori in modo critico (QCER: xiv, xv).

[1] Per una presentazione del QCER in relazione all'italiano cfr. Vedovelli (2002).

[2] Nella versione inglese del QCER si usa l'aggettivo *comprehensive*: "It describes in a comprehensive way what language learners have to learn to do in order use a language" (CEFR, p.1). A questo proposito Figueras *et al*. (2005: 264) sottolineano che "The CEFR has aimed to be *comprehensive*, but this has to be understood in a very practical way, as not being synonimous with *exhaustive*". Ci sembra opportuno tenere conto di questa precisazione anche a proposito dell'italiano *esaustivo*, forse più cautamente sostituibile con *ampio*.

Il QCER è un sistema descrittivo astratto, e non vi è corrispondenza immediata tra i descrittori disponibili e i bisogni concreti di chi deve effettuare una valutazione delle competenze sulla base di alcune prestazioni. Il problema principale risiede nel fatto che le descrizioni presenti nelle scale non coincidono sempre con comportamenti direttamente osservabili. Questo richiede interpretazione da parte di chi utilizza il QCER. L'unico modo per accertarsi che i descrittori portino a interpretazioni valide è esaminare l'allineamento dei giudizi espressi da diversi valutatori, riportando l'attenzione alla domanda fondamentale per chi si occupa di testing linguistico, ossia: *come posso avere la certezza che il mio B1 (o A2, o C1 ecc.) corrisponde al B1 (o A2, o C1 ecc.) di un altro valutatore?*

Tra gli autori che l'hanno esaminato criticamente, Weir (2005: 281-282) sostiene che il QCER nella sua forma attuale non è sufficientemente comprensivo, coerente e trasparente per un uso efficace nel testing linguistico. L'autore identifica in particolare quattro aree problematiche:

1. le scale sono collocate entro una gamma di variabili contestuali e di condizioni di uso linguistico incompleta e non ben applicata;
2. si tiene poco conto della natura dei processi cognitivi coinvolti nei diversi livelli di abilità;
3. le attività sono raramente messe in relazione con la qualità con cui ci si aspetta che vengano svolte;
4. la formulazione di alcuni dei descrittori non è sufficientemente coerente e trasparente per lo sviluppo di test.

In questo contributo ci concentriamo sugli ultimi due aspetti, che nella nostra esperienza sono quelli che sembrano creare le maggiori difficoltà ai valutatori. Infatti, chi utilizza le griglie avverte la necessità di sapere quale sia il livello di qualità della prestazione che ci si aspetta dagli apprendenti nello svolgere un determinato compito a un determinato livello. Occorrerebbe cioè che fossero chiaramente esplicitati i criteri di valutazione di *come* un apprendente sa fare qualcosa. Tali criteri dovrebbero essere strettamente connessi con i parametri contestuali e teorici che influenzano l'abilità che si sta misurando. Questo punto si intreccia con quello della trasparenza dei termini, come rivela il seguente esempio, che peraltro rappresenta un caso tutt'altro che isolato all'interno del QCER. Il descrittore relativo a *Coerenza e coesione* per il livello A2+ è così formulato:

- [l'apprendente] *è in grado di collegare frasi semplici usando i connettivi più usuali per raccontare una storia o descrivere qualcosa, realizzando un semplice elenco di punti*.

Quello del livello B1 recita invece:

- *è in grado di collegare una serie di elementi relativamente brevi e semplici in una sequenza lineare per punti*.

È evidente che un valutatore rischia di trovarsi in seria difficoltà nel momento in cui deve determinare se l'apprendente che sta valutando dal punto di vista della coerenza e coesione si trova a livello A2+ o a livello B1: come definire le *frasi semplici*? E come distinguerle dagli *elementi relativamente brevi e semplici*? E quali sono i *connettivi più usuali*? Curiosamente, il descrittore del livello precedente risulta invece più preciso nelle indicazioni, giacché specifica a quali indicatori linguistici fa riferimento: *È in grado di collegare gruppi di parole con connettivi semplici quali "e", "ma" e "perché"*. Perché tale esemplificazione è presente solo per questo livello e non per i due successivi?

2. Metodologia

Come accennato nell'introduzione, questo lavoro parte dall'idea che le potenzialità del QCER nell'ambito della scuola siano numerose e vadano opportunamente sviluppate. Il QCER offre infatti la possibilità di disporre di un linguaggio comune per individuare obiettivi validi e significativi per l'apprendimento linguistico (L2, LS e, perché no, anche L1[3]), e consente quindi di elaborare criteri condivisi per la rappresentazione precisa delle capacità linguistiche e del loro sviluppo nel tempo. Abbiamo deciso di verificare sperimentalmente tali potenzialità, o almeno una parte di esse, cercando di rispondere alla domanda:

1. Il QCER è uno strumento efficace per la valutazione delle competenze orali in contesto scolastico?

E, in particolare:

1a) Una volta adattate e utilizzate in modo critico, le griglie del QCER permettono di ottenere giudizi affidabili sul livello di competenza nel parlato?
1b. I descrittori, quando vengono utilizzati nella valutazione, risultano effettivamente esaustivi, trasparenti e coerenti?

[3] Cfr. le esperienze nazionali del Progetto Poseidon e del Progetto PON, e varie esperienze locali presso singoli istituti scolastici.

Dato il nostro interesse specifico per l'utilizzo del QCER nel contesto scolastico, e quindi da parte di insegnanti con diversi livelli di confidenza con lo strumento e di esperienza nella valutazione, ci domandiamo inoltre:

2. Come vengono utilizzati i descrittori da diversi tipi di insegnanti? Che differenze si possono osservare tra valutatori esperti e non esperti?

Per rispondere alle domande di ricerca è stato ideato un protocollo di indagine che permettesse una validazione empirica dell'affidabilità delle valutazioni espresse da due gruppi di valutatori, esperti e non, su un corpus di 9 videoregistrazioni di produzioni orali in italiano L1 ed L2.

L'indagine ha coinvolto 10 insegnanti di italiano L2[4] suddivisi in due gruppi: 5 insegnanti esperti nella valutazione e 5 insegnanti non esperti nella valutazione. Gli insegnanti esperti sono tutti docenti di un Centro Territoriale Permanente, formati come esaminatori e regolarmente coinvolti come valutatori della prova orale negli esami di certificazione CELI dell'Università per Stranieri di Perugia. Gli insegnanti non esperti sono docenti con diverse esperienze nel mondo della scuola, tutti impegnati come volontari in scuole di italiano L2 per adulti.

Gli strumenti utilizzati per l'indagine sono due: un test diagnostico e alcune griglie contenenti una selezione di descrittori.

Il test diagnostico è costituito da tre attività comunicative. La prima è un'intervista informale che segue una traccia tematica anche se con una certa flessibilità. La seconda attività è il racconto di un breve estratto dal film *Modern Times* di Charlie Chaplin. La terza è la telefonata di servizio effettuata per raccogliere informazioni allo scopo di selezionare il miglior cellulare da un catalogo e di organizzare una gita di classe. Il test è stato somministrato a tre studentesse, due apprendenti e una parlante nativa. Dal corpus[5] contenente le videoregistrazioni delle studentesse impegnate nelle attività sono stati selezionati nove estratti da sottoporre al giudizio dei valutatori.

[4] Cogliamo l'occasione per ringraziare gli insegnanti del CTP di Modena e i volontari della scuola di Canegrate e della scuola Babele di Legnano (MI) che hanno partecipato alla sperimentazione. Senza il loro prezioso contributo questo lavoro non sarebbe stato possibile.

[5] Si tratta di un corpus più ampio costituito dalle produzioni di otto studentesse sottoposte a una batteria di task comunicativi (tra cui i tre selezionati per la presente sperimentazione). Il corpus è realizzato nell'ambito dei due progetti di ricerca: COFIN 2003, co-finanziato dal MIUR e dall'Università di Verona (coordinatore locale C. Bettoni), e COFIN 2006, co-finanziato dal MIUR e dall'Università di Modena e Reggio Emilia (coordinatore locale G. Pallotti).

	Lingua 1	**Data di nascita**	**Luogo di nascita**	**Scolarità in Italia**	**Scolarità nel paese d'origine**
Apprendenti					
Shirley	inglese	1986	Nigeria	6 anni (3 sup.)	7 anni
Pandita	punjabi	1987	Punjab, India	4 anni (3 sup.)	8 anni
Parlante nativa					
Elisa	italiano	1989	Italia	10 anni (1 sup.)	//

Ai valutatori sono state fornite tre tabelle contenenti griglie del QCER, selezionate in base alla loro pertinenza rispetto ai compiti inclusi nel test diagnostico e alle competenze nella lingua parlata.

La tabella 1 conteneva i descrittori generali relativi alle seguenti competenze: *Produzione orale generale* (QCER: 73), *Comprensione orale generale* (p. 83) e *Interazione orale generale* (p. 93).

La tabella 2 riportava i descrittori relativi agli *Aspetti qualitativi dell'uso della lingua parlata*: *Repertorio linguistico generale* (p. 135), *Correttezza grammaticale* (p. 140), *Fluenza nel parlato* (p. 158), *Coerenza e coesione* (p. 154), *Prendere la parola* (p. 158).

La tabella 3 riportava i descrittori relativi alle singole attività comunicative: *Monologo articolato* per il racconto di film (p. 74), *Transazioni per ottenere beni e servizi* per le telefonate di servizio (p. 99), *Intervistare ed essere intervistati* per l'intervista (p. 101).

Ciascun gruppo di insegnanti si è reso disponibile per una sessione di valutazione della durata di 3 ore circa. Agli insegnanti è stato richiesto di valutare le 9 videoregistrazioni riferendosi alle griglie predisposte. L'attribuzione del livello avveniva subito dopo la visione di ciascuna produzione ed era suddivisa in due momenti: una prima valutazione individuale e una seconda valutazione condivisa da tutti i valutatori. Gli incontri sono stati registrati e le discussioni trascritte.

Ai fini dell'analisi sono state considerate le valutazioni individuali, le valutazioni condivise e le trascrizioni delle discussioni nei gruppi.

3. Analisi e risultati

3.1 Analisi quantitativa

Con l'analisi quantitativa abbiamo osservato l'affidabilità dei giudizi misurando il grado di accordo tra i valutatori tramite due procedure statistiche (cfr. Bachman, 1990)[6]. La prima consiste nel calcolo del coefficiente di correlazione, per il quale abbiamo usato lo *Spearman rank-order* (*Rho*)[7] e, rifacendoci alla letteratura, abbiamo considerato affidabili i giudizi quando il valore del coefficiente era pari o superiore a 0,80 (Alderson *et al.*, 1995). Questa operazione permette di confrontare l'ordine in cui i diversi valutatori distribuiscono gli apprendenti su una scala, ma non confronta direttamente le valutazioni. Ciò significa che sarebbe possibile, per due valutatori, ordinare gli studenti nello stesso modo sulla scala, dando però voti sistematicamente più alti l'uno e sistematicamente più bassi l'altro. Una seconda procedura statistica, ossia l'analisi della frequenza e della distribuzione delle valutazioni, ci ha consentito di registrare tali possibili differenze di severità. I casi di astensione dal giudizio sono stati considerati a parte.

I risultati del calcolo di *Rho* ottenuti su tutte le valutazioni evidenziano come gli insegnanti siano nel complesso poco concordi nell'applicazione della scala. Sono comunque evidenti differenze tra i due gruppi (cfr. Tabella 2). Gli insegnanti esperti dimostrano infatti maggior accordo rispetto ai colleghi non esperti, ottenendo un valori pari a 0,80 contro 0,70. Questo risultato è in linea con ciò che viene da più parti evidenziato nella letteratura sul testing: la formazione aumenta la coerenza e l'affidabilità delle scelte dei valutatori (Lunz *et al.*, 1990) oltre che ridurre le differenze di severità tra esaminatori (Wigglesworth, 1993; Weigle, 1994). Lumley e Mcnamara (1995) mostrano però che i vantaggi della formazione non sono di lunga durata, pertanto è necessario un momento di formazione prima di ogni sessione di valutazione.

Interessanti i risultati ottenuti nelle valutazioni condivise, dove per i due gruppi *Rho* è pari a 0,81: la discussione nei gruppi per il raggiungimento di un accordo sembra quindi portare gli insegnanti ad allinearsi maggiormente nelle loro scelte. Questo dato sostiene l'utilità della pratica diffusa nelle certificazioni di impiegare due valutatori per ciascuna valutazione.

[6] Ringraziamo Claudio Sartini per il supporto tecnico nell'utilizzo degli strumenti statistici.

[7] Lo *Spearman rank-order* è un indice calcolato sulla base della comparazione delle correlazioni ottenute tra coppie di variabili

I risultati dell'analisi effettuata variabile per variabile confermano il quadro appena descritto. Sebbene i giudizi su alcuni descrittori risultino meno affidabili di quelli espressi per altri, gli insegnanti esperti e le valutazioni condivise ottengono i risultati migliori. È interessante notare come i descrittori relativi alla correttezza grammaticale abbiano ottenuto il maggior accordo, mentre altri - per esempio quelli relativi alla produzione orale generale o alla fluenza nel parlato - hanno ottenuto risultati nettamente meno soddisfacenti.

Tab. 2. Affidabilità dei giudizi. Spearman rank-order (Rho)

	Tutti gli insegnanti	**Insegnanti esperti**	**Insegnanti non esperti**	**Giudizi condivisi**
	Da R1 a R10	R1, R2, R3, R4, R5	R6, R7, R8, R9, R10	R17, R18
Tutte le variabili	0,73	0,80	0,70	0,81
Produzione orale generale	0,59	0,74	0,63	0,86
Comprensione orale generale	0,58	0,89	0,64	-
Interazione orale generale	0,80	0,87	0,68	0,78
Repertorio linguistico generale	0,73	0,80	0,61	0,96
Correttezza grammaticale	0,90	0,88	0,81	0,87
Fluenza nel parlato	0,68	0,87	0,55	0,83
Coerenza e coesione	0,83	0,89	0,83	0,79
Prendere la parola	0,73	0,76	0,84	0,95
Task	0,78	0,82	0,71	0,56

Nonostante il risultato per i 10 insegnanti non sia ottimale, i risultati dell'analisi statistica della distribuzione delle valutazioni permettono alcune osservazioni in positivo. Su 11 livelli disponibili[8] per ciascuna competenza, la gamma delle scelte effettuate oscilla tra 2 e 4, con una preferenza per 2 valutazioni diverse. Le diverse valutazioni inoltre si riferiscono nella maggior

[8] Calcolando livelli e sotto-livelli, dal momento che in molti casi i livelli delle fasce A e B sono ulteriormente suddivisi (per esempio, A1, A1+, A2, A2+)

parte dei casi a livelli contigui sulla scala. Questo suggerisce come il QCER possa essere un buon punto di riferimento, anche se sembra non riuscire a guidare in modo chiaro e preciso le scelte degli insegnanti nella distinzione tra livelli contigui, in particolare per la fascia C. Difficoltà per i livelli più alti sono state evidenziate anche in altre ricerche: Kaftandjieva e Takala (2002: 113), per esempio, sottolineano come il livello più alto (C2) non sia sempre distinto da quello inferiore (C1); Cassandro e Maggini (2004: 71), conducendo un'esercitazione con un gruppo di insegnanti, osservano come non manchino divergenze vistose nell'attribuzione dei livelli, specialmente per quanto riguarda la parte alta del *continuum.*

Nel corpus di valutazioni individuali, gli insegnanti hanno raggiunto un accordo del 100% in un numero assai limitato di casi. Su 81 valutazioni espresse da ciascun insegnante per ciascun video è stato raggiunto un accordo totale in 7 casi (8,6%), 5 dei quali tra gli esperti e 2 tra i non esperti. L'accordo totale è stato raggiunto quasi esclusivamente nelle valutazioni relative alle produzioni di Pandita, l'apprendente meno avanzata.

I risultati dell'analisi della distribuzione delle astensioni permette alcune osservazioni ulteriori. Innanzitutto si confermano le differenze tra esperti e non esperti. Nelle valutazioni individuali gli esperti si astengono dal valutare in modo significativamente maggiore rispetto ai non esperti (18,27% dei casi contro 0,74%). Probabilmente gli insegnanti non esperti, consapevoli della scarsa confidenza con il QCER, hanno limitato al massimo le critiche verso le griglie, cercando piuttosto di svolgere al meglio il compito assegnato e arrivare comunque a una valutazione; al contrario gli insegnanti esperti, più sicuri nel compito, si sono sentiti liberi di esternare le loro difficoltà attribuendo la scelta a lacune nei descrittori o nel tipo di produzione da valutare.

Le astensioni espresse dagli insegnanti esperti coinvolgono tutti e tre i tipi di attività comunicative, ma si riferiscono in particolare all'apprendente più avanzata, Shirley, e all'italiana Elisa, e ad alcuni descrittori più che ad altri: *Comprensione orale generale* (26 occorrenze), *Prendere la parola* (17), *Interazione orale generale* (9).

Gli insegnanti non esperti si sono astenuti dal valutare in 3 casi, tutti relativi allo stesso video, il racconto del film realizzato da Shirley. La distribuzione delle astensioni nei giudizi condivisi supporta quanto osservato nei due gruppi: gli esperti si astengono in 17 casi (20,99%), i non esperti in 1 caso solo (1,23%). La distribuzione per i non esperti è speculare a quella illustrata sopra.

3.2 Analisi qualitativa

L'analisi quantitativa rivela nel complesso un grado di affidabilità piuttosto basso, dal momento che i giudizi dei valutatori risultano spesso non allineati. La scarsa affidabilità va probabilmente in parte attribuita alla mancanza di confidenza con lo strumento, come dimostra il fatto che il disaccordo emerge molto più evidente all'interno del gruppo dei non esperti. A questo proposito, vale la pena di sottolineare un'altra differenza interessante tra il gruppo degli esperti e quello dei non esperti: mentre nel primo il tentativo di raggiungere un giudizio condiviso ha generato quasi sempre lunghe e accese discussioni, perché nessuno intendeva "cedere" e voleva piuttosto convincere gli altri delle proprie ragioni, nel gruppo dei non esperti si avvertiva una maggiore disponibilità a rinunciare alla propria decisione in favore di una "media" tra le valutazioni che potesse costituire il giudizio collettivo. Al di là delle differenze caratteriali tra i membri dei due gruppi, si può forse intravedere in questa diversità di comportamenti la maggiore o minore sicurezza nell'uso del QCER determinata dalla maggiore o minore consuetudine all'utilizzo dello strumento.

Anche se la confidenza con il QCER appare come un elemento centrale, non si possono escludere fattori interni al QCER per la scarsa affidabilità dei giudizi, dal momento che anche tra i valutatori esperti si riscontrano problemi di allineamento. Per comprendere quali siano questi fattori interni analizziamo le discussioni emerse nella fase di attribuzione del giudizio condiviso.

Tra i numerosi commenti emersi nelle discussioni, abbiamo selezionato quelli che mettono in luce la mancata aderenza del QCER ai criteri cui il documento mira a conformarsi. Riportiamo qui di seguito alcune citazioni, rappresentative delle principali categorie di problemi rivelate dall'analisi.

1. *Criterio di esaustività*

Dalle osservazioni emergono essenzialmente due elementi di contrasto con il criterio della esaustività. Da un lato non tutti i descrittori risultano avere lo stesso livello di dettaglio, dall'altro per alcuni livelli ci sono indicazioni su quello di cui l'apprendente sa parlare, per altri no. Ne risulta che, per esempio, il lettore ha la sensazione di non poter collocare a livello C2 un parlante che racconta la trama di un film, poiché questo compito compare solo a livello B1+. Ecco i commenti di due valutatori esperti che illustrano quanto detto:

> (1) *Esperto*: C'è qualche problema nei descrittori, perché il B2 a volte nel dettaglio risulta più alto di un C1. Il descrittore B2 è ricchissimo, è molto più dettagliato. Per esempio il descrittore "Prendere la parola", è ricchissimo. Quindi, mi rispondeva meglio di quell'altro.

(2) *Esperto*: Il problema è che fino al B1 ci occupiamo di viaggi, sport, ecc. Poi non ce ne occupiamo più. Non sappiamo più se questo C1 lo sa fare o non lo sa fare. Si parla di argomenti altissimi, argomentazioni su linguaggi specifici ed è il problema di questo framework che ormai è fuori uso nel senso che fino al B1 parli di viaggi, poi dal B2 si parla di argomenti anche complessi. Quindi sparisce tutto l'altro repertorio che invece è tanto importante. Fino al B1 ci occupiamo di cose pratiche, di routine, poi dopo…

2. *Criterio di trasparenza*

I due punti critici che ricorrono con maggiore frequenza nelle osservazioni dei valutatori sono l'uso di termini vaghi, che lettori diversi possono interpretare diversamente, e l'uso di parole di significato analogo che non si capisce se siano utilizzate come sinonimi oppure con l'intenzione di suggerire sfumature di significato differenti (per osservazioni analoghe, cfr. Alderson *et al.*, 2004: 8-11). Osserviamo alcuni commenti dei valutatori:

(3)
Esperto 4: All'inizio usa anche delle forme improprie.
Esperto 3: C1 mantiene costantemente un livello elevato di correttezza grammaticale; gli errori sono rari e poco evidenti.
Esperto 2: Sono *evidenti* perché non mette l'articolo, omissione dell'articolo. La concordanza poi non è rispettata, io ho visto molti errori.
Esperto 5: No, la concordanza mi sembrava abbastanza…

(4) *Non esperto*: Mi sfugge la differenza tra "formare un elenco" (A2) e "strutturandola in una sequenza lineare di punti" (B1).

3. Criterio di coerenza

Per quanto riguarda la coerenza, dalle osservazioni dei valutatori emerge soprattutto il fatto che le stesse descrizioni compaiono in descrittori di livelli diversi (cfr. Alderson *et al.* 2004: 10):

(5) *Esperto*: Questi + non sono ben formulati, spesso fanno un riassunto del precedente e aggiungono un elemento che potrebbe anche far pensare ad un livello più basso. Possono confondere le idee. Per esempio, questa ragazza nell'attacco, quando deve chiedere il costo del telefonino, non avvia nemmeno il discorso, si precipita immediatamente nel discorso, non riesce a reggere. Non sa partire, non sa attirare l'attenzione.

Tra i punti critici segnalati dai valutatori quello che sembra suscitare le difficoltà maggiori nell'utilizzo del QCER per la valutazione è la frequente mancanza di corrispondenza tra il tipo di task svolto dall'apprendente e la qualità della prestazione. In pratica, i valutatori hanno la sensazione che l'apprendente si trovi a un certo livello, ma trovano che il compito non rientri tra quelli indicati nel descrittore di quel livello, pertanto tendono ad attribuire alla prestazione un livello diverso da quello che secondo la loro intuizione sarebbe il più adatto. Pare quindi che il descrittore non dia agli insegnanti le parole per tradurre la loro intuizione in valutazione. Il problema emerge soprattutto ai livelli più alti, dove nei descrittori si citano solo - e non sempre - compiti di un certo tipo, come parlare di argomenti astratti. Accade così che un valutatore ritenga per esempio di poter considerare la fluenza dell'apprendente in una telefonata tale da essere collocata a livello C1, ma senta di non poter assegnare questo livello perché nel descrittore del C1 "non si parla di telefonate ma di argomenti astratti". Si può invece ragionevolmente ipotizzare che un apprendente dimostri una fluenza nel parlato collocabile a livello C1 anche nel gestire una telefonata, solo che la qualità della prestazione sarà in questo caso superiore rispetto a quella di un parlante che svolge il medesimo compito a livello B1. Sarebbe dunque essenziale trovare un riferimento preciso anche alla qualità della prestazione che ci si deve attendere a ogni livello (Weir 2005: 285). Vediamo alcuni commenti dei valutatori:

(6)
Esperto 4: È stata molto pronta a rispondere alla telefonata.
Esperto 3: Ma nel C stiamo parlando di argomenti alti, non di interazione in una telefonata. Io ho messo B2.
Esperto 4: È il compito che la limita, non valutabile.

(7)
Esperto 4: Il contenuto di cui si parla è molto condizionante. Come faccio a giudicare tutti gli elementi se l'interlocutore si limita ad un argomento di tipo familiare? Io non ho strumenti di valutare più di così.
Esperto 5: Il compito era facile, ma lei è stata brava.

4. Conclusioni

Alla luce dei risultati ottenuti, possiamo confermare che il QCER è uno strumento efficace per la valutazione delle competenze orali. Infatti, sebbene l'affidabilità dei giudizi non risulti sempre buona e i commenti dei valutatori rivelino vari nodi di criticità nell'uso delle griglie, l'analisi statistica mostra un quadro d'insieme tutt'altro che sconfortante, suggerendo la possibilità di operare interventi in grado di sviluppare le potenzialità dello strumento.

Occorrerebbe in primo luogo rielaborare i descrittori a partire da una teoria dell'apprendimento linguistico, ossia dalla descrizione di come i processi cognitivi evolvono da un livello di competenza al successivo (Weir, 2005: 295). In secondo luogo, sarebbe essenziale avere a disposizione un'ampia rassegna di esempi concreti di prestazioni ai vari livelli (cfr. il DVD prodotto in questa prospettiva dal Consiglio d'Europa), nonché un repertorio di strutture linguistiche e di forme d'uso elaborato a partire dall'analisi di *corpora* di italiano parlato e scritto. Come suggerisce lo stesso Weir (2005: 283) il QCER non è pensato per rispondere alle esigenze specifiche della valutazione e pertanto «it will require considerable, long-term research, much reflective test development by providers, and prolonged critical interaction between stakeholders in the field to address these deficiencies».

Il confronto tra le valutazioni degli esperti e quelle dei non esperti ha mostrato come una buona conoscenza del QCER sia un prerequisito essenziale per il raggiungimento di un maggior grado di affidabilità nei giudizi. Emerge pertanto la necessità di una formazione specifica degli insegnanti nell'uso di questo strumento. Tale formazione dovrebbe prevedere sia una fase iniziale di familiarizzazione con il QCER sia dei momenti di esercitazione e di confronto tra valutatori (*benchmarking*), con l'obiettivo di uniformare il più possibile le interpretazioni dei descrittori e delle prestazioni. Procedure di questo tipo sono ampiamente diffuse tra chi si occupa di valutazione delle competenze linguistiche a scopo di certificazione e non costituiscono momenti isolati, bensì pratiche abituali. Le ricerche nel settore suggeriscono inoltre l'opportunità di ripetere le sessioni di confronto e standardizzazione dei giudizi prima di intraprendere ogni processo di valutazione (Lumley & McNamara, 1995). Nella scuola, ente certificatore per eccellenza, purtroppo non sono generalmente diffusi i momenti dedicati alla condivisione e standardizzazione di criteri comuni. Il nostro studio evidenzia come procedure di questo tipo potrebbero migliorare di molto l'affidabilità delle valutazioni scolastiche e costituire pertanto una buona pratica che auspichiamo trovi maggiore diffusione all'interno della scuola.

Pre-print: Ferrari S., Nuzzo E. (2010), *La valutazione delle competenze orali in italiano L2. Una verifica sperimentale dei criteri suggeriti dal Quadro Comune Europeo* (con E. Nuzzo), in E. Lugarini (a cura di), «Valutare le competenze linguistiche, Atti del XV Convegno nazionale GISCEL, Milano, 6-8 marzo 2008», Franco Angeli, Milano, pp. 279-293

Riferimenti bibliografici

Alderson J.C. (ed.) (2002), *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Case studies*, Council of Europe, Strasbourg.

Alderson J.C. (2005). "Editorial". *Language Testing*, 2005: 22.

Alderson J.C., Clapham, C. & Wall, D. (1995), *Language test construction and evaluation*, Cambridge University Press, Cambridge.

Alderson J.C., Figueras N., Kuijper H., Nold G., Takala S. & Tardieu C. (2004), *The development of specifications for item development and classification within the Common European Framework of Reference for Languages: learning, teaching, assessment: reading and listening: final report of the Dutch Construct Project*. Available on request from the Project Coordinator, c.alderson@lancaster.ac.uk.

Bachman L.F. (1990), *Fundamental considerations in language testing*, Oxford University Press, Oxford.

Cassandro M. & Maggini M. (2004), "Osservazione e valutazione di apprendenti stranieri in relazione ai livelli comuni di riferimento del Quadro comune europeo, in E. Jafrancesco (a c. di), *Le tendenze innovative del quadro comune europeo di riferimento per le lingue ed il portfolio. Atti del XII convegno nazionale ILSA*, Edilingua, Roma: 64-107.

Consiglio d'Europa (2001), *Modern languages: learning, teaching, assessment. A Common European Framework of Reference*, CUP, Cambridge. (Trad. it. *Quadro Comune Europeo di Riferimento per le lingue*, La Nuova Italia-Oxford, Firenze).

Figueras, N., North B., Takala S., Verhelst N. & Van Avermaet P. (2005), "Relating examinations to the Common European Framework: a manual", *Language Testing*, 22,3: 261-79.

Grego Bolli G. (2006), *Progetti europei. Nuove prospettive sulla scia del Quadro Comune Europeo di Riferimento*, http://www.cvcl.it/canale.asp?id=198

Kaftandjieva F. & Takala S. (2002), "Council of Europe scales of language proficiency: a validation study, in Alderson: 106-29.

Language Testing (2005), *Special issue on the Common European Framework of Reference*: 22, 3.

Lumley T. & McNamara T.F. (1995), "Rater characteristics and rater bias: implications for training", *Language Testing*, 12,1: 54-71.

Lunz M.E., Wright B.D. & Linacre J.M. (1990), Measuring the impact of judge severity on examination scores, *Applied Measurement in Education* 3: 331-345.

Vedovelli M. (2002), *Guida all'italiano per stranieri: la prospettiva del quadro comune europeo*, Carocci, Roma.

Weigle S.C. (1994), "Effects of training on raters of ESL compositions", *Language Testing*, 11: 197-223.

Weir C.J. (2005), "Limitations of the Common European Framework for developing comparable examinations and tests", *Language Testing*: 22.

Wigglesworth G. (1993), "Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction" *Language Testing*, 10, 3: 305-335.