

# La valutazione dell'adeguatezza funzionale di produzioni orali e scritte in italiano L2 in tipologie di task differenti

di Ineke Vedder<sup>1</sup>

## Abstract

The paper discusses the role of variation in L2 learning and, more particularly, the influence of diamesic variation and variation in the communicative context, on oral and written performance in Italian L2. In order to do so, the functional adequacy of oral and written L2 production, in terms of successful task completion is investigated, on the basis of data collected by means of three different tasks. For being able to assess functional adequacy, a rating scale was developed, inspired by the conversational maxims of Grice. The rating scale of functional adequacy distinguishes four scale dimensions: content, task requirements, comprehensibility, coherence and cohesion<sup>2</sup>. In the paper the applicability of the rating scale will be discussed, in relation to variation in both task modality (oral versus written) and task type (decision making task, narration task, instruction task). The underlying assumption of the present study is that assessing learner production is not possible without taking into account, next to the linguistic dimension of L2 performance (CAF: complexity, accuracy and fluency<sup>3</sup>), also the functional dimension<sup>4</sup>.

## Introduzione

Questo articolo indaga il rapporto tra variazione e padronanza linguistica in italiano lingua seconda (L2), focalizzandosi sulle differenze linguistiche dovute alla variazione diamesica (legata al canale di comunicazione) e la variazione determinata dalla situazione comunicativa. Concentrandoci sull'adeguatezza delle produzioni linguistiche in contesti d'uso diversi, esamineremo l'adeguatezza funzionale della *performance*, intesa in termini di "felice" esecuzione del compito, attraverso canali di comunicazione (orale e scritto) e tipologie di task (task argomentativo, narrativo, *instruction task*) differenti.

Presenteremo una scala di valutazione dell'adeguatezza funzionale, considerata come componente specifica della pragmatica, e ispirata alle massime conversazionali di Grice<sup>5</sup>. Scomponendo il costrutto dell'adeguatezza funzionale in quattro componenti diverse, distinguiamo le seguenti dimensioni scalari: contenuto, *task requirements*, comprensibilità, coerenza e coesione. La scala di valutazione è stata messa alla prova in due studi successivi<sup>6</sup>, per testarne l'applicabilità, per produzioni orali e scritte di apprendenti e parlanti nativi di italiano e di olandese. Oggetto di questo studio saranno i dati che sono stati raccolti per l'italiano. Confronteremo l'applicabilità della scala

dell'adeguatezza funzionale per la valutazione di produzioni orali e scritte di un gruppo di informanti di italiano L2 e italiano lingua materna (L1). Verificheremo inoltre l'applicabilità della scala per la valutazione di un piccolo corpus di dati scritti di italiano L2 elicitati attraverso un task argomentativo, un task narrativo e un *instruction task*<sup>7</sup>.

L'articolo è strutturato come segue. Descriveremo il quadro teorico dello studio, *Task-Based Language Assessment* (TBLA, PAR. 1). Dopo aver formulato la definizione dell'adeguatezza funzionale adottata nello studio, come costruito interazionale legato allo svolgimento del task, discuteremo i motivi che hanno portato allo sviluppo della scala di valutazione, insieme ai presupposti teorici e ai descrittori dei livelli di competenza (PAR. 2). Nel paragrafo 3 saranno illustrate l'impostazione e la metodologia della ricerca. I risultati ottenuti dalla verifica dell'applicabilità della scala per produzioni orali e scritte in L2 e L1 saranno presentati nel paragrafo 4. Il paragrafo 5 espone i risultati emersi da uno studio pilota per testare l'applicabilità della scala per tipologie di task differenti. Nelle conclusioni discuteremo le implicazioni teoriche e didattiche della ricerca e le possibili prospettive future.

## I

### Valutazione della pragmatica in L2

Negli studi di *Second Language Acquisition* (SLA) la competenza pragmatica in L2 è stata spesso valutata, sulla scia di Austin<sup>8</sup>, Searle<sup>9</sup>, e Brown e Levinson<sup>10</sup>, in termini di acquisizione di singoli atti linguistici, quali richieste, proteste, complimenti, che sono tra gli atti linguistici più investigati<sup>11</sup>. Gli strumenti di elicitazione più usati sono stati il *Discourse Completion Test* (DCT), il *role play* (aperto o a scelta multipla) e vari tipi di questionari. Minore attenzione è stata prestata invece agli aspetti discorsivi della competenza pragmatica, come la gestione dei turni e l'organizzazione delle sequenze conversazionali, la ricezione e la produzione dell'implicatura, la "riparazione conversazionale" e gli effetti perlocutivi degli enunciati del parlante/scrivente sull'interlocutore<sup>12</sup>. Benché siano stati suggeriti degli approcci alternativi<sup>13</sup>, il numero di studi che ha esaminato gli aspetti discorsivi della competenza pragmatica in L2, in monologhi o scambi più lunghi (*extended discourse*), finora è ridotto. Il nostro studio, proponendo una scala di valutazione dell'adeguatezza funzionale del messaggio che il parlante/scrivente A trasmette al destinatario B, si prefigge di colmare, in parte, tale lacuna.

Uno degli ambiti di ricerca che intende esaminare l'uso della lingua in contesti autentici è la cosiddetta *performance assessment*, la valutazione delle abilità comunicative degli apprendenti attraverso l'utilizzo dei task. Una specifica forma di *performance assessment* è la *Task-Based Language Assessment*<sup>14</sup>, che non considera il task semplicemente come strumento per elicitare e valutare particolari strutture linguistiche, ma si propone di valutare gli enunciati di un locutore e un interlocutore – fisicamente o idealmente presente – in base allo svolgimento di uno specifico task linguistico.

Il riferimento teorico dello studio che presentiamo qui è *Task-Based Language Assessment* (TBLA). La definizione dell'adeguatezza funzionale che adottiamo è quella di costruito interazionale inteso in termini di "felice" esecuzione del task<sup>15</sup>, che coinvolge due partecipanti (A e B). Nella prospettiva teorica della TBLA, l'adeguatezza della *performance* in L2 è giudicata in base allo svolgimento di uno specifico task linguistico (scrivere una mail, rispondere a una telefonata), in un contesto comunicativo ben preciso (una conversazione fra amici, un colloquio di lavoro, una lettera di presentazione), in base a requisiti e criteri derivati dal task stesso<sup>16</sup>.

## 2

### Scala di valutazione dell'adeguatezza funzionale: presupposti, criteri e descrittori

La scala di valutazione dell'adeguatezza funzionale che proponiamo si ispira alla massime conversazionali di Grice relative a quantità, pertinenza, modo e qualità del messaggio trasmesso dal parlante o dallo scrivente A e recepito dall'interlocutore B. La scala è stata sviluppata a partire dai risultati ottenuti da uno studio precedente che aveva l'obiettivo di investigare la relazione tra adeguatezza funzionale e complessità linguistica in L2 e L1 (CAF: complessità, accuratezza, fluenza)<sup>17</sup>. L'assunto primario sottostante lo sviluppo della scala è che non sia possibile valutare la competenza orale e scritta in L2 senza prendere in considerazione, accanto alla dimensione CAF, gli aspetti funzionali del testo<sup>18</sup>. Diversamente dagli indici quantitativi CAF adoperati negli studi SLA per la misurazione della complessità sintattica e lessicale, l'accuratezza e la fluenza delle produzioni linguistiche degli apprendenti<sup>19</sup>, non esistono indici generali per misurare l'adeguatezza funzionale.

I criteri che scaturiscono da questi presupposti sono stati pertanto: (i) scomposizione del costruito dell'adeguatezza funzionale in componenti distinte; (ii) descrittori indipendenti dai descrittori linguistici CAF; (iii) descrittori oggettivi e quantificabili; (iv) applicabilità della scala per valutatori esperti e inesperti; (v) applicabilità sia in L2 che in L1. Scomponendo il costruito dell'adeguatezza funzionale in quattro dimensioni diverse, in base ai risultati emersi da altri studi<sup>20</sup>, abbiamo distinto le seguenti componenti scalari: contenuto; *task requirements*; comprensibilità; coerenza e coesione.

La scala di valutazione, una scala Likert di sei livelli scalari, è così composta dalle seguenti dimensioni (cfr. *Appendice 1*):

- **Contenuto** – Questa dimensione prende in considerazione l'adeguatezza del numero e del tipo di unità informative espresse nel testo del parlante/scrivente A, la loro pertinenza e rilevanza, indipendentemente dai requisiti specifici riguardo all'esecuzione del task.
- **Task requirements** – È stato dato questo nome alla dimensione che considera la misura in cui nel testo si è tenuto conto dei requisiti specifici del task che il parlante/scrivente A deve svolgere, relativi al genere testuale, all'atto linguistico, alla macrostruttura e al registro del testo.

- Comprensibilità – Qui si focalizza il grado di comprensibilità del testo di A e lo sforzo richiesto all’interlocutore B per capire lo scopo e le idee espresse nel testo.
- Coerenza e coesione – Questa dimensione si riferisce alla coerenza e alla coesione del testo del parlante/scrivente A, in termini di presenza o assenza di strategie anaforiche, connettivi, salti logici e argomenti non collegati, ripetizioni ecc.

## 3

**Impostazione e metodologia**

Al fine di testare l’applicabilità della scala dell’adeguatezza funzionale per modalità e tipi di task differenti sono stati condotti diversi studi<sup>21</sup>. Nel paragrafo 4 presenteremo i principali risultati emersi dagli studi condotti da Kuiken e Vedder<sup>22</sup> aventi come obiettivo la verifica dell’applicabilità della scala per produzioni orali e scritte in italiano L2 e L1. Nel paragrafo 5 ci focalizzeremo sui risultati di uno studio pilota che intendeva verificare l’applicabilità della scala per tre tipologie di task differenti (argomentativo, narrativo, *instruction task*<sup>23</sup>).

I dati scritti del corpus di Kuiken e Vedder provengono da 39 apprendenti di madrelingua olandese, con una padronanza linguistica tra i livelli A2 e B2 del QCER e da un gruppo di controllo di 18 parlanti nativi dell’italiano. Ai partecipanti sono stati somministrati due task di scrittura argomentativa. Nel primo compito veniva chiesto di indicare quale organizzazione *no-profit*, da un elenco di tre, dovesse essere finanziata dall’università (Task 1). Nel secondo compito si trattava di scegliere, fra tre articoli di giornale proposti, quello ritenuto più adatto per la pubblicazione in prima pagina (Task 2). Gli studenti dovevano motivare la propria scelta con un testo di circa 150 parole che contenesse almeno tre argomentazioni. Gli informanti del gruppo di controllo, studenti universitari della stessa età, sono stati sottoposti agli stessi test. Per stabilire il livello di padronanza linguistica globale dei partecipanti è stato somministrato un C-test<sup>24</sup>.

I valutatori erano quattro studenti universitari, parlanti nativi dell’italiano, per indirizzo di studio e età comparabili agli informanti coinvolti nello studio. I valutatori potevano essere considerati “ingenui” e “inesperti”, in quanto non avevano precedenti esperienze di valutazione di testi. La decisione di coinvolgere nello studio valutatori “inesperti” era stata presa a seguito di un’osservazione emersa da uno studio precedente<sup>25</sup>: per la valutazione dell’adeguatezza funzionale è importante affiancare ai valutatori “esperti” (insegnanti di lingua) valutatori “inesperti”, in quanto i primi possono dimostrarsi più clementi o, al contrario, più severi nei loro giudizi rispetto a parlanti nativi “ingenui”. Sono state organizzate due sedute di *training* (sei ore in totale), per spiegare ai valutatori le finalità della ricerca e per farli familiarizzare con l’uso della scala. Dopo la fase della valutazione è stata organizzata una discussione retrospettiva di gruppo con i valutatori per discutere il lavoro svolto e le riflessioni dei valutatori sull’utilizzo delle scale.

In uno studio successivo, questa volta focalizzato interamente sulla L2, sono stati raccolti, per la verifica della scala per produzioni orali, nuovi dati provenienti da un altro gruppo di 26 apprendenti di italiano L2 di madrelingua olandese, con un livello di padronanza linguistica tra A2 e C1. Gli apprendenti sono stati sottoposti agli stessi test, benché in versione orale, che già erano stati somministrati in versione scritta al gruppo dei primi 39 informanti. Come nel primo studio, i quattro valutatori, studenti universitari e parlanti nativi dell'italiano, hanno partecipato a due sedute di *training* di sei ore totali. L'organizzazione del *training* dei "nuovi" valutatori è stata identica a quella avvenuta per la valutazione dei dati scritti. Le produzioni orali sono state giudicate in base alla scala usata per la valutazione dei testi scritti, ad eccezione di alcuni adattamenti minori nei descrittori, come la sostituzione della parola "testo" al posto di "produzione". La valutazione delle produzioni orali è avvenuta sulla base dell'ascolto delle registrazioni audio, non sulla base della loro trascrizione.

Riportiamo di seguito un esempio della produzione orale di uno dei partecipanti (livello di competenza A2), che spiega i motivi per i quali i soldi dovrebbero essere dati all'associazione ASPH (Associazione sportiva portatori di handicap) che promuove attività sportive per bambini handicappati (1).

(1) *Volo andare i soldi al organizzazione ASPA; è un'organizzazione che supporta l'associazioni sportive di bambini con handicap fisici e mentali. Ehm primo è importante per tutti i bambini a fare esercizi sportiva. Secondo, tutti, tutti i bambini eeb deservono li stessi opportunity e terzo fare esercizi sportiva è divertente.*

Nello studio pilota di Del Bono<sup>26</sup> erano coinvolti 15 studenti universitari di madrelingua olandese, con un livello di competenza linguistica in italiano tra A2 e B1, a cui sono stati somministrati tre task (argomentativo, narrativo, *instruction task*). Nel task argomentativo, a partire da tre proposte di alloggio per un anno di studio all'estero (*residence* per studenti internazionali, appartamento in condivisione, studio) gli studenti dovevano scrivere un'email al Direttore degli Studenti Internazionali della facoltà ospitante e indicare i motivi per i quali preferivano un determinato alloggio e non avevano scelto gli altri due. Il task narrativo chiedeva agli informanti di raccontare un particolare aneddoto accaduto durante un viaggio di studio. Il testo doveva contenere informazioni specifiche quali la destinazione del viaggio, quando era accaduto e se avevano imparato qualcosa dell'esperienza vissuta. Nell'*instruction task* gli apprendenti dovevano scrivere un biglietto di istruzioni da lasciare a una coppia che aveva preso in affitto una casa per una breve vacanza. Nel biglietto dovevano inserire delle informazioni specifiche, come indicare dove si trovavano certi oggetti di cui gli ospiti potevano aver bisogno, fornire le istruzioni relative ad alcuni apparecchi elettronici e spiegare come lasciare la casa alla partenza, come nel testo riportato di seguito, scritto da un apprendente con un livello di competenza linguistica tra A2 e B1 (2).

(2) *Gentili ragazzi (ciao Maria e Max!)*

*Benvenuto a casamia. Nella mia casa si trovano diverse cose che loro possono usare. È per mezzo d'usare tutti gli accomdazioni per esempio quando ti/si piace usare il mio netflix account e il xbox non c'è problema. Ho una casa piccola, dunque ciò di cui possono aver bisogno la loro permanenza è molto facile. Non ho apparecchi elettronici, quelli sono complicati per usare. Vorrei ritrovare casamia in il stesso modo che ho lasciato.*

I testi raccolti in Olanda sono stati successivamente valutati in Italia da cinque valutatori non esperti, studenti universitari e parlanti nativi dell'italiano della stessa età e stesso indirizzo di studio dei partecipanti. Il *training* dei valutatori, per farli familiarizzare con le scale e far convergere i loro giudizi, si è svolto in due fasi separate con un primo *training* di gruppo e un secondo *training* individuale. I testi da valutare sono stati suddivisi per tipologia di task; il *training* di gruppo comprendeva però soltanto il task narrativo. I valutatori hanno avuto un mese di tempo per la valutazione, in seguito è stata organizzata una discussione di gruppo retrospettiva sull'utilizzo delle scale e le eventuali difficoltà incontrate dai valutatori.

Nella TAB. 1 sono schematizzati i dati relativi agli studi di Kuiken e Vedder e a quello di Del Bono.

TABELLA 1  
Lingua, task, valutatori e modalità

Studio	Lingua				Task	Valutatori		Modalità	
	Lingua target	L2	L1	Livello L2	Tipologia	Non esperti	Scritto	Orale	
Kuiken, Vedder (2014, 2017)	italiano (L2/L1)	N=39	N=18	A2-B2	ARG	4	X		
Kuiken, Vedder (2018)	italiano (L2)	N=26	-	A2-C1	ARG	4		X	
Del Bono (2017)	italiano (L2)	N=15	-	A2-B2	ARG NAR INSTRUC	5	X		

#### 4

#### Applicabilità della scala: produzioni scritte e orali

In questo paragrafo confrontiamo i dati ottenuti dalla verifica dell'applicabilità della scala di valutazione per le produzioni scritte con quella per le produzioni orali<sup>27</sup>. Le domande di ricerca che ci siamo posti riguardavano nello specifico: (i) l'affidabilità

dei giudizi dei valutatori tra di loro (*intraclass correlations*) per ciascuna delle quattro dimensioni della scala (contenuto, *task requirements*, comprensibilità, coerenza e coesione); (ii) le correlazioni tra i giudizi dei valutatori sulle quattro dimensioni dell'adeguatezza funzionale precedentemente distinte; (iii) le correlazioni tra i giudizi dei valutatori sull'adeguatezza funzionale dei due task prodotti dai partecipanti. Riassumiamo qui brevemente i principali risultati delle analisi statistiche, per ulteriori dettagli rimandiamo a Kuiken e Vedder<sup>28</sup>.

Per misurare il grado di accordo tra i valutatori nelle produzioni scritte come in quelle orali sono stati calcolati i coefficienti di correlazione intraclasse, per ciascuna delle quattro dimensioni scalari. I risultati evidenziano che nel complesso i valutatori concordano nell'applicazione della scala, in quanto tutte le correlazioni sono significative per le due modalità. Per i testi scritti i valori variano da accettabili (.725 per *task requirements*) a eccellenti (.901 per comprensibilità). Per le produzioni orali le correlazioni risultano un po' più alte: da .893 (comprensibilità) a .929 (coerenza/coesione).

Il secondo obiettivo era quello di stabilire le correlazioni tra i giudizi dei valutatori per le quattro dimensioni dell'adeguatezza funzionale. Per l'esame del rapporto tra questi giudizi sono stati calcolati i coefficienti di correlazione di Pearson. I risultati dell'analisi statistica dimostrano che tutte le correlazioni sono significative, nelle produzioni scritte come in quelle orali. La relazione tra comprensibilità e coerenza/coesione appare particolarmente significativa nella modalità scritta (.938,  $p < .01$ ), mentre per il task orale risulta piuttosto alta la correlazione tra *task requirements* e contenuto (.957,  $p < .01$ ).

Al fine di poter stabilire le correlazioni nella modalità scritta e in quella orale tra i giudizi dei valutatori sull'adeguatezza funzionale per i due task prodotti dai gruppi dei partecipanti<sup>29</sup>, sono state confrontate le medie dei giudizi assegnati dai valutatori, per ciascuna delle quattro dimensioni scalari. In base ai coefficienti di correlazione di Pearson, possiamo concludere che le correlazioni tra i giudizi dei due task risultano significative in tutti i casi, il che indica che i due task vengono giudicati in modo simile dai valutatori. Nell'insieme, per la modalità scritta le correlazioni risultano lievemente più alte rispetto alle produzioni orali. Sono più basse le correlazioni riscontrate tra i giudizi assegnati per *task requirements*, sia nella modalità scritta (.455,  $p < .001$ ) che in quella orale (.679,  $p < .01$ ). Risultano più alte, per i task scritti come per quelli orali, le correlazioni tra i giudizi di coerenza e coesione (scritto: .802;  $p < .001$ ; orale .830,  $p < .01$ ) e comprensibilità (scritto: .766,  $p < .001$ ; orale: .832,  $p < .001$ ).

Dall'intervista retrospettiva è emerso che i valutatori avevano avuto maggiori difficoltà nella valutazione dell'adeguatezza delle produzioni orali rispetto a quella dei testi scritti, dato che soprattutto la comprensibilità risulta influenzata da fattori quali prosodia, esitazioni, riformulazioni e ripetizioni. Ciò vale anche per la valutazione di coerenza/coesione, anch'essa apparsa più difficile nella modalità orale.



A questo proposito è comunque interessante osservare che nell'insieme il grado di accordo tra i valutatori risulta più alto per i task orali rispetto ai task scritti.

Infine è emerso che è anche possibile l'utilizzo della scala in L1, benché ai valutatori risultasse più difficile giudicare i testi degli studenti nativi, valutati tutti come appartenenti ai livelli più alti della scala, rispetto alle produzioni degli studenti non nativi<sup>30</sup>.

## 5

### Applicabilità della scala per tipologie di task differenti

In questo paragrafo saranno presentati i principali risultati ottenuti nello studio pilota di Del Bono. Lo studio, seguendo i suggerimenti di Kuiken e Vedder<sup>31</sup> di mettere alla prova l'applicabilità della scala anche per altre tipologie di task, aveva l'obiettivo di verificare la validità dello strumento di valutazione per tre generi testuali (task argomentativo, task narrativo, *instruction task*). Per un resoconto più dettagliato e per le analisi statistiche rinviando al testo di Del Bono; dato il piccolo campione (15 informanti) i risultati vanno comunque interpretati con una certa cautela.

Replicando in gran parte la metodologia degli studi di Kuiken e Vedder condotti su task argomentativi nella modalità scritta e orale discussi in precedenza, la ricerca esplorativa di Del Bono si proponeva di (i) indagare per le tre tipologie di task l'affidabilità dei giudizi dei valutatori tra di loro per ciascuna delle quattro dimensioni della scala; (ii) stabilire le correlazioni tra i giudizi dei valutatori su ciascuna dimensione; (iii) verificare le correlazioni tra i giudizi assegnati ai tre task prodotti dai partecipanti.

Dall'analisi dei dati è emerso che nell'insieme i coefficienti di correlazione intraclassa (*interrater agreement*), che danno informazioni sul grado di accordo tra i valutatori, risultano molto più bassi di quelli riscontrati da Kuiken e Vedder. I punteggi di correlazione stabiliti per contenuto, *task requirements* e coerenza/coesione non superano in nessun caso la soglia del grado di accordo dello 0.70; risultano più alte nei tre task solo le correlazioni tra i valutatori per la comprensibilità. Il testo con le correlazioni più basse è quello argomentativo. L'analisi dell'alfa di Cronbach<sup>32</sup>, ovvero la consistenza delle valutazioni di ciascun valutatore (*interrater reliability*), dimostra invece delle correlazioni piuttosto alte e significative per le tre tipologie di task (tra .72.8 e .952). Le più alte risultano essere le correlazioni per la comprensibilità nel task narrativo e l'*instruction task*, e quelle per *task requirements* nel task argomentativo.

In merito alle correlazioni tra le quattro dimensioni dell'adeguatezza funzionale, stabilita con il calcolo del coefficiente di correlazione di Pearson, è emerso, similmente ai risultati di Kuiken e Vedder per il confronto della modalità scritta e quella orale, che nei tre task le correlazioni tra le quattro dimensioni sono significative e possono dunque essere considerate componenti dello stesso costrutto. Dall'analisi è apparso anche che le dimensioni con maggiori correlazioni sono contenuto e *task require-*



*ments*. Al contrario, la comprensibilità è risultata essere la dimensione con le correlazioni più basse con *task requirements* e contenuto, confermando che un testo può essere comprensibile indipendentemente dai contenuti e il rispetto dei criteri specifici del task, e viceversa.

Le correlazioni tra i giudizi dei valutatori nei tre testi scritti dal gruppo di partecipanti risultano statisticamente significative, con delle variazioni a seconda del tipo di task. Le più alte sono le correlazioni nel task narrativo, mentre risultano più basse quelle riscontrate nel task argomentativo. L'analisi dei dati ha pertanto confermato la possibilità di applicare la scala alle tre tipologie di task, anche se le correlazioni divergono. Ciò potrebbe significare che la scala è in qualche modo sensibile al tipo di testo che va a valutare. I valutatori potrebbero essere condizionati dal tipo di task che valutano e quindi, a seconda del testo valutato, la loro interpretazione della scala potrebbe essere diversa. Questo ultimo risultato potrebbe comunque essere determinato dal fatto che il *training* di gruppo (un unico incontro) dei valutatori dello studio di Del Bono era avvenuto in base a un unico tipo di task (task narrativo).

## Conclusioni

Abbiamo visto che in base ai risultati ottenuti da Kuiken e Vedder è in effetti possibile misurare l'adeguatezza funzionale in modo affidabile su una scala suddivisa in quattro dimensioni, nelle produzioni scritte come in quelle orali, di studenti con livelli di competenza linguistica differenti. In tutti i casi i quattro valutatori concordano nei loro giudizi, con valori di affidabilità tra accettabili, buoni e eccellenti. La verifica dell'applicabilità della scala dimostra così che l'adeguatezza funzionale, come componente fondamentale della competenza in L2, può essere valutata oggettivamente, indipendentemente dagli indici CAF. Nonostante le correlazioni significative riscontrate tra le quattro dimensioni della scala che dimostrano che esse fanno parte dello stesso costrutto, è importante tenerle distinte nella valutazione, in quanto concettualmente differenti. La ricerca mostra inoltre che anche valutatori non esperti riescono a lavorare con le scale dopo due sessioni di *training* di gruppo di sei ore in totale.

I dati emersi dallo studio esplorativo di Del Bono condotto in base a un piccolo campione di informanti confermano in parte i risultati di Kuiken e Vedder. Le differenze riscontrate rispetto alle osservazioni di Kuiken e Vedder, che probabilmente sono da attribuire alle differenze nell'organizzazione del *training*, sottolineano l'importanza del *training* di gruppo – piuttosto che quello individuale – e della discussione tra i valutatori per acquisire familiarità con l'uso delle scale e l'interpretazione dei descrittori confrontandosi tra loro. I primi risultati ottenuti da Del Bono dimostrano inoltre la necessità di indagare ulteriormente l'impatto della tipologia di task sull'adeguatezza funzionale, per capire meglio i meccanismi in gioco nello sviluppo e nella valutazione dell'adeguatezza funzionale in L2.

## Appendice 1

### Scala di valutazione dell'adeguatezza funzionale

Contenuto: Il numero delle unità informative espresse nel testo è adeguato e rilevante?

1. *Per niente*: Il numero di idee è insufficiente, non è affatto adeguato, e non sono collegate le une alle altre.
2. *Appena*: Il numero di idee è appena adeguato, le idee sono poco consistenti.
3. *Parzialmente*: Il numero di idee è parzialmente adeguato anche se non sono molto consistenti.
4. *Sufficientemente*: Il numero di idee è sufficientemente adeguato e le idee sono sufficientemente consistenti.
5. *Largamente*: Il numero di idee è largamente adeguato e appaiono molto consistenti e coerenti le une con le altre.
6. *Assolutamente*: Il numero di idee è assolutamente adeguato e appaiono molto consistenti e coerenti le une con le altre.

*Task requirements*: Sono soddisfatti i requisiti e i criteri specifici del task (ad es. genere, atto linguistico, registro)?

- 1 *Per niente*: Nessuno dei requisiti e dei criteri specifici del task è stato soddisfatto.
- 2 *Appena*: Pochi requisiti e criteri specifici del task (meno della metà) sono stati soddisfatti.
- 3 *Parzialmente*: Circa la metà dei registri e dei criteri specifici del task sono stati soddisfatti.
- 4 *Sufficientemente*: Più della metà dei requisiti e dei criteri specifici del task sono stati soddisfatti.
- 5 *Largamente*: Quasi tutti i requisiti e i criteri specifici del task sono stati soddisfatti.
- 6 *Assolutamente*: Tutti i requisiti e i criteri specifici del task sono stati soddisfatti.

Comprensibilità: Quanto sforzo è richiesto per capire lo scopo del testo e le idee?

- 1 *Per niente*: Il testo non è affatto comprensibile. Le idee e lo scopo sono espressi in modo oscuro e il lettore, anche sforzandosi, non riesce a capire.
- 2 *Appena*: Il testo è appena comprensibile, i suoi scopi non sono chiari e il lettore deve sforzarsi molto per capire le idee dell'autore. Il lettore deve cercare di indovinare la maggior parte delle idee e degli scopi del testo.
- 3 *Parzialmente*: Il testo è parzialmente comprensibile, ma alcune frasi non si capiscono bene a una prima lettura. Un'ulteriore rilettura è utile per chiarire gli scopi del testo e le idee espresse, ma rimangono alcuni dubbi.
- 4 *Sufficientemente*: Il testo è sufficientemente comprensibile, solo certe frasi sono poco chiare ma si possono capire senza grandi sforzi con una rilettura.
- 5 *Largamente*: Il testo è largamente comprensibile e si legge agevolmente, non ci sono problemi di comprensibilità.
- 6 *Assolutamente*: Il testo è assolutamente comprensibile e si legge molto agevolmente, le idee e gli scopi sono espressi con chiarezza.

Coerenza e coesione: Il testo è coerente e coeso (ad es. ci sono connettivi e strategie di coesione)?

- 1 *Per niente*: Ci sono frequenti salti logici e argomenti non collegati. L'autore non usa riferimenti anaforici (pronomi, frasi con soggetto sottinteso chiaramente interpretabile). Il testo non è affatto coeso. I connettivi sono praticamente assenti e le idee non sono collegate tra loro.
- 2 *Appena*: L'autore spesso non collega gli argomenti tra loro; se c'è coerenza, questa è espressa per lo più da ripetizioni. Pochi riferimenti anaforici. Alcuni salti logici. Il testo è poco coeso. Vengono usati pochi connettivi, che non collegano bene le idee.
- 3 *Parzialmente*: Ci sono frequenti salti di argomento e/o ripetizioni. Più di due frasi di seguito esprimono esplicitamente lo stesso soggetto, anche quando questo sarebbe chiaro. Vengono usati alcuni riferimenti anaforici. Possono esservi alcune interruzioni della coerenza. Il testo è abbastanza coeso. Vengono usati alcuni connettivi, ma sono per lo più delle congiunzioni.
- 4 *Sufficientemente*: Ci sono frequenti salti di argomento e/o ripetizioni. Più di due frasi di seguito esprimono esplicitamente lo stesso soggetto, anche quando questo sarebbe chiaro. Vengono usati alcuni riferimenti anaforici. Possono esservi alcune interruzioni della coerenza. Il testo è abbastanza coeso. Vengono usati alcuni connettivi, ma sono per lo più delle congiunzioni.
- 5 *Largamente*: Quando si introduce un nuovo argomento, di solito ciò avviene mediante l'uso di connettivi o espressioni di collegamento esplicite. Le ripetizioni sono molto rare. Si trovano numerosi riferimenti anaforici. Nessuna interruzione della coerenza. Il testo è molto coeso e le idee sono ben collegate tra loro mediante connettivi verbali o avverbiali.
- 6 *Assolutamente*: L'autore produce un'ottima coerenza integrando le nuove idee nel testo con connettivi o espressioni di collegamento esplicite. I connettivi anaforici sono usati regolarmente. Qualche raro caso di argomenti non collegati e nessuna interruzione della coerenza. La struttura del testo è estremamente coesa, grazie a un abile uso dei connettivi (in particolare verbali, avverbiali e formule di collegamento), usati spesso per descrivere le relazioni tra le idee.

## Note

1. Un particolare ringraziamento va a Paolo Orrù (Università di Debrecen), Chiara Sale (Università di Cagliari) e Saskia Nijman (Universitair Medisch Centrum Groningen) per il loro aiuto. Ringraziamo inoltre Gabriele Pallotti (Università di Modena e Reggio Emilia), Laura Sanna (Università di Torino), Anna Vittoria Ottaviani (Università di Trento), Francesca Musella (Università di Napoli L'Orientale), Sible Andringa, Raffaello Ibba, Irene Mistro, Flora Prota, Luisa Seguin (Università di Amsterdam).

2. F. Kuiken, I. Vedder, *Rating Written Performance: What Do Raters Do and Why?*, in "Language Testing", 31, 3, 2014, pp. 329-48; F. Kuiken, I. Vedder, *Functional Adequacy in L2 Writing. Towards a New Rating Scale*, in "Language Testing", 34, 3, 2017, pp. 321-36; F. Kuiken, I. Vedder, *Assessing Functional Adequacy of L2 Performance in a Task-Based Approach*, in N. Taguchi, Y. Kim (eds.), *Task-Based Approaches to Teaching and Assessing Pragmatics*, Benjamins, Amsterdam 2018, pp. 275-96.

3. A. Housen, F. Kuiken, I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Proficiency in SLA*, Benjamins, Amsterdam 2012.

4. G. Pallotti, *CAF: Defining, Refining and Differentiating Constructs*, in "Applied Linguistics", 30/4, 2009, pp. 590-601; A. Révész, M. Ekiert, E. N. Torgersen, *The Effects of Complexity, Accuracy and Fluency on Communicative Adequacy in Oral Task Performance*, in "Applied Linguistics", 6, 1, 2016, pp. 828-48.

5. H. P. Grice, *Logic and Conversation*, in P. Cole, J. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York 1975, pp. 41-58.

6. Kuiken, Vedder, *Functional Adequacy in L2 Writing*, cit.; Idd., *Assessing Functional Adequacy of L2 Performance in a Task-Based Approach*, cit.; I. Vedder, *Il ruolo dell'adeguatezza funzionale nelle produzioni scritte in lingua seconda: Proposta per una scala di valutazione*, in E. Santoro, I. Vedder (a cura di), *Pragmatica e interculturalità in italiano lingua seconda*, Franco Cesati Editore, Firenze 2016, pp. 79-91.
7. F. Del Bono, *Aspetti pragmatici nella valutazione di testi scritti: Uno studio sull'adeguatezza funzionale in italiano L2*. Tesi di Laurea Magistrale, Università degli Studi Roma Tre, Roma 2017.
8. J. L. Austin, *How to Do Things with Words*, Clarendon Press, Oxford 1962.
9. J. R. Searle, *Speech Acts: An Essay in the Philosophy of Language*, Cambridge University Press, Cambridge 1969; J. R. Searle, *Indirect Speech Acts*, in P. Cole, J. L. Morgan (eds.), *Syntax and Semantics 3: Speech Acts*, Academic Press, New York 1975, pp. 59-82.
10. P. Brown, S. C. Levinson, *Politeness. Some Universals in Language Usage*, Cambridge University Press, Cambridge 1987.
11. A. Barron, *Acquisition in Interlanguage Pragmatics*, Benjamins, Amsterdam 2003; K. C. Grabowski, *Assessing Pragmatic Competence*, in D. Tsagari, J. Banerjee (eds.), *Handbook of Second Language Assessment*, Palgrave, New York 2016, pp. 149-71; A. Martínez-Flor, E. Usó Juan, *Speech Act Performance: Theoretical, Empirical and Methodological Issues*, Benjamins, Amsterdam 2010.
12. C. Roever, *Testing ESL Pragmatics: Development and Validation of a Web-Based Assessment Battery*, Peter Lang, Berlin 2005; C. Roever, *Testing of Second Language Pragmatics: Past and Future*, in "Language Testing", 28, 4, 2011, pp. 463-81; C. Roever, *What learners get for free: Learning of pragmatic formulae in ESL and EFL environments*, in "The ELT Journal", 66, 1, 2012, pp. 10-21.
13. J. D. Brown, T. Hudson, J. Norris, W. J. Bonk, *An Investigation of Second Language Task-Based Performance Assessment*, in "SLTCC Technical Report" 24, Second Language Teaching & Curriculum Centre, University of Hawai'i at Manoa, Honolulu 2002; S. J. Ross, G. Kasper (eds.), *Assessing second language pragmatics*, Palgrave Macmillan, Basingstoke 2013.
14. Brown et al., *An Investigation of Second Language Task-Based Performance Assessment*, cit.; J. M. Norris (ed.), *Special Issue: Task-Based Language Assessment*, in "Language Testing", 19, 4, 2002.
15. Searle, *Speech Acts*, cit.
16. Brown et al., *An Investigation of Second Language Task-Based Performance Assessment*, cit.; M. González-Lloret, *A Practical Guide to Integrating Technology into Task-Based Language Teaching*, Georgetown University Press, Washington 2016; T. McNamara, "Instruction" in *Second Language Performance Assessment: Whose Performance?*, in "Applied Linguistics" 18, 4, pp. 446-66; Norris *Special Issue: Task-Based Language Assessment*, cit.; J. M. Norris, *Task-Based Teaching and Testing*, in M. H. Long, C. J. Doughty (eds.), *The Handbook of Language Teaching*, Wiley-Blackwell, Malden (MA)-Oxford 2009, pp. 578-94; Pallotti, *CAF: Defining, Refining and Differentiating Constructs*, cit.; A. Shehadeh, *Task-Based Language Assessment: Components, developments, and implementation*, in C. Coombe, P. Davidson, B. O'Sullivan, S. Stoyonoff (eds.), *The Cambridge Guide to Second Language Assessment*, Cambridge University Press, Cambridge 2012, pp. 156-63.
17. F. Kuiken, I. Vedder, R. Gilabert, *Communicative Adequacy and Linguistic Complexity in L2 Writing*, in I. Bartning, M. Martin, I. Vedder (eds.), *Communicative Proficiency and Linguistic Development: Intersections between SLA and Language Testing Research*, Eurosla Monographs Series 1, 2010, pp. 81-100.
18. Pallotti, *CAF: Defining, Refining and Differentiating Constructs*, cit.; Révész, Ekiert, Torgersen, *The Effects of Complexity, Accuracy and Fluency on Communicative Adequacy*, cit.
19. J. M. Norris, L. Ortega, *Towards an organic approach to investigating CAF in instructed SLA: The Case of Complexity*, in "Applied Linguistics", 30, 4, 2009, pp. 555-79; B. Bulté, A. Housen, *Defining and Operationalising L2 Complexity*, in A. Housen, F. Kuiken, I. Vedder (eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*, Benjamins, Amsterdam 2012, pp. 121-42.
20. B. Bridgeman, D. Powers, E. Stone, P. Mollan, *TOEFL iBT speaking test scores as indicators of oral Communicative Language Proficiency*, in "Language Testing", 29, 1, pp. 91-108; N. H. De Jong, M. P. Steinel, A. F. Florijn, R. Schoonen, J. H. Hulstijn, *The Effect of Task Complexity on Functional Adequacy, Fluency and Lexical Diversity in Speaking Performances of Native and Non-Native Speakers*, in Housen, Kuiken, Vedder (eds.), *Dimensions of L2 Performance and Proficiency. Complexity, Accuracy and Fluency in SLA*, cit., pp. 121-42; N. H. De Jong, M. P. Steinel, A. F. Florijn, R. Schoonen, J. H. Hulstijn, *Facets of Speaking Proficiency*, in "Studies in Second Language Acquisition", 34, 1, 2012, pp. 5-34; U. Knoch, *Diagnostic Assessment of Writing: A Comparison of Two Rating Scales*, in "Language Testing", 26, 2, 2009, 275-304.
21. D. Cortés Velásquez, E. Nuzzo, *Assessing Task Performance in the Academic Context: Functional Adequacy Scales for Italian L1 University Students*, 7<sup>th</sup> TBLT Conference Tasks in Context, University of Barcelona,

Barcelona 2017, April 19-21; S. Faone, F. Pagliara, *How to Assess L2 Information-Gap Tasks through Functional Adequacy Rating Scales*, 7<sup>th</sup> TBLT Conference Tasks in Context, University of Barcelona, Barcelona 2017, April 19-21; A. Herraiz Martínez, *On the Use of Second Language Pragmatic Markers: A Study Abroad Perspective*, Congresso AITLA 2018 Lingua in Contesto: La prospettiva pragmatica. Università Roma Tre, Roma 2018, 22-24 febbraio; F. Pagliara, *Valutare l'adeguatezza funzionale in produzioni scritte di studenti Marco Polo*, Convegno Dieci anni di didattica dell'italiano a studenti cinesi. Risultati, esperimenti, proposte, Università per Stranieri di Siena, Siena 2017, 6-7 ottobre

22. Kuiken, Vedder, *Functional Adequacy in L2 Writing*, cit.; Idd., *Assessing Functional Adequacy of L2 Performance in a Task-Based Approach*, cit.; Vedder, *Il ruolo dell'adeguatezza funzionale nelle produzioni scritte in lingua seconda*, cit.

23. F. Del Bono, *Aspetti pragmatici nella valutazione di testi scritti: Uno studio sull'adeguatezza funzionale in italiano L2*. Tesi di Laurea Magistrale, Università degli Studi Roma Tre, Roma 2017.

24. C. Klein Braley, *C-Tests in the Context of Reduced Redundancy Testing: An Appraisal*, in "Language Testing", 14, 1, pp. 47-84; U. Raatz, C. Klein-Braley, *Introduction to language testing and to C-tests*, in J. A. Coleman, R. Grotjahn, U. Raatz (eds.), *University Language Testing and the C-test*. AKS-Verlag, Bochum 2002, pp. 75-91.

25. Kuiken, Vedder, Gilabert, *Communicative Adequacy and Linguistic Complexity in L2 Writing*, cit.

26. F. Del Bono, *Aspetti pragmatici nella valutazione di testi scritti: Uno studio sull'adeguatezza funzionale in italiano L2*. Tesi di Laurea Magistrale, Università degli Studi Roma Tre, Roma 2017.

27. Kuiken, Vedder, *Functional Adequacy in L2 Writing*, cit.; Idd., *Assessing Functional Adequacy of L2 Performance in a Task-Based Approach*, cit.; Vedder, *Il ruolo dell'adeguatezza funzionale nelle produzioni scritte in lingua seconda*, cit.

28. Kuiken, Vedder, *Assessing Functional Adequacy of L2 Performance in a Task-Based Approach*, cit.

29. Ricordiamo che i dati orali che sono stati raccolti in un secondo momento, provengono da un altro gruppo di apprendenti di italiano L2, e sono stati valutati da altri valutatori.

30. Per l'utilizzo delle scale in italiano L1, si veda Cortés Velásquez, Nuzzo, *Assessing Task Performance in the Academic Context*, cit.

31. Kuiken, Vedder, *Functional Adequacy in L2 Writing*, cit.; Idd., *Assessing Functional Adequacy of L2 Performance in a Task-Based Approach*, cit.; Vedder, *Il ruolo dell'adeguatezza funzionale nelle produzioni scritte in lingua seconda*, cit.

32. L'alfa di Cronbach, a volte chiamata anche coefficiente di Cronbach, fu ideata nel 1951 dal pedagogista statunitense Lee Cronbach.