UNIVERSITÀ
DI PAVIA

Università degli Studi di Pavia

Dipartimento di Studi Umanistici

Corso di Laurea Magistrale in Linguistica Teorica, Applicata e delle Lingue Moderne

Progettazione e validazione di un modello automatizzato per la valutazione dell'italiano come L2: il sistema ETET.

Relatrice

Prof.ssa Chiara Zanchi

Correlatrice

Prof.ssa Claudia Roberta Combei

Tesi di Laurea Magistrale di

Anna Vignoli

Matricola n° 538505

Anno Accademico 2024/25

INDICE

Introduzione **1**

CAPITOLO 1. RICOGNIZIONE TEORICA **5**

1.1 Fondamenti teorici della Second Language Acquisition 6
1.2 L'interfaccia tra Second Language Acquisition e Language Testing 9
1.3 La competenza linguistico-comunicativa: modelli teorici 13
1.4 La standardizzazione della valutazione linguistica in Europa 21
1.5 Il Quadro Comune Europeo di riferimento per le lingue 23
1.6 Attrattività e pubblico dell'italiano L2 27
1.7 L'evoluzione tecnologica dei test linguistici: dai PPT ai CALT 30
1.8 Verso l'Intelligent CALL: l'integrazione dell'intelligenza artificiale e del trattamento automatico del linguaggio 37
1.9 Intelligent Computer-Assisted Language Learning 46

CAPITOLO 2. IL LANGUAGE TESTING E LE CARATTERISTICHE ESSENZIALI DI UN TEST **55**

2.1 Fondamenti teorici del Language Testing 56
2.2 Le caratteristiche essenziali di un test 60
2.2.1 Validità 63
2.2.2 Affidabilità 74
2.2.3 Autenticità 75
2.2.4 Interattività 76
2.2.5 Impatto 77
2.2.6 Praticabilità 77
2.3 Tipologie di prove e formati di item 79
2.4 Le fasi di sviluppo di un test 83

CAPITOLO 3. METODOLOGIA E PROGETTAZIONE DELLO STRUMENTO DI VALUTAZIONE **88**

3.1 Il caso studio ETET 89
3.2 Architettura e funzionalità di ETET 90
3.3 La costruzione del test 95
3.3.1 Progettazione 95
3.3.2 Operazionalizzazione 103
3.3.2.1 Valutare la grammatica 105
3.3.2.2 Valutare la lettura 107
3.3.2.3 Valutare l'ascolto 112
3.3.2.3.1 Sintesi vocale 117
3.3.2.4 Valutare la scrittura 120
3.3.2.5 Valutare il parlato 123
3.3.2.6 Considerazioni contestuali generali 126

3.3.2.7 Sviluppo dei sistemi di scoring 128
3.3.3 Somministrazione 137

CAPITOLO 4. ANALISI E DISCUSSIONE DEI RISULTATI 141

4.1 Analisi qualitativa della popolazione 142
4.2 Criteri di esclusione e creazione del campione 149
4.3 Costruzione del gold standard 155
4.4 Analisi dell'accordo tra gold standard e modello 160
4.4.1 Disamina dei risultati relativi alla sezione di speaking 161
4.4.2 Disamina dei risultati relativi alla sezione di writing 164
4.5 Analisi dell'accordo tra autovalutazioni e modello 166
4.6 Analisi dell'esperienza utente e dell'usabilità della piattaforma 168

Conclusioni 173

Bibliografia 177

Sitografia 186

Introduzione

La valutazione linguistica ha da sempre svolto una funzione che va oltre la semplice misurazione della competenza, configurandosi come un dispositivo attraverso cui le società stabiliscono, mediante procedure pubbliche e verificabili, chi possiede determinate qualità o abilità e chi ne è privo. In questo senso, i test linguistici non rappresentano semplici strumenti tecnici, ma pratiche istituzionali che incidono concretamente sulle traiettorie individuali dei candidati, influenzando percorsi educativi, opportunità professionali e – in casi estremi – l'accesso alla cittadinanza o al diritto di richiedere asilo.

A fronte della comprovata rilevanza sociale del *testing* e della crescente richiesta di strumenti di certificazione linguistica, il settore del *Language Testing* ha conosciuto negli ultimi decenni un'evoluzione significativa sia sul piano teorico sia su quello tecnologico. Da un lato, lo sviluppo delle teorie dell'acquisizione delle lingue seconde e dei modelli di competenza comunicativa ha favorito un progressivo superamento degli approcci psicometrici-strutturalisti in favore di modelli valutativi più complessi e orientati alla competenza comunicativa. Dall'altro, la diffusione delle tecnologie digitali ha trasformato profondamente le modalità di progettazione e somministrazione dei test, favorendo lo sviluppo di sistemi informatizzati e di nuovi ambiti di ricerca quali il *Computer-Assisted Language Testing* (CALT) e *l'Intelligent Computer-Assisted Language Learning* (ICALL).

Negli ultimi anni, questi processi sono stati ulteriormente accelerati dai progressi nella Linguistica Computazionale e nell'Intelligenza Artificiale: tecnologie come il Trattamento Automatico del Linguaggio naturale (TAL), il riconoscimento automatico del parlato (*Automatic Speech Recognition*; ASR) e i *Large Language Models* (LLM) hanno aperto nuove possibilità per l'analisi automatica delle produzioni linguistiche, rendendo teoricamente realizzabile lo sviluppo di sistemi di valutazione sempre più flessibili e accessibili.

Nonostante i significativi sviluppi tecnologici nel campo del *testing* linguistico, la realizzazione di strumenti pienamente automatizzati continua a rappresentare una sfida complessa. In particolare, la valutazione delle abilità produttive – scrittura e parlato – richiede ancora spesso un intervento umano rilevante, limitando la scalabilità dei sistemi disponibili, soprattutto in contesti caratterizzati da un'elevata domanda.

Il presente lavoro si colloca proprio in risposta a tali esigenze, proponendo la progettazione, l'implementazione e la validazione di un sistema automatizzato per la valutazione della

competenza linguistica di apprendenti di italiano come L2. L'obiettivo principale è sviluppare uno strumento in grado di misurare le abilità produttive attraverso l'impiego di tecnologie di *Automated Essay Scoring* (AES) e *Automated Speaking Assessment* (ASA). Al tempo stesso, il sistema è concepito per rimanere saldamente ancorato ai modelli teorici della *Second Language Acquisition*, fondato sui principi metodologici del *Language Testing* e supportato dalle tecnologie più avanzate della Linguistica Computazionale. In questa prospettiva, la ricerca si colloca all'intersezione di tre ambiti disciplinari distinti ma convergenti, con l'intento di mostrare come l'automazione dei processi valutativi non implichi necessariamente un indebolimento teorico né un impoverimento delle pratiche di valutazione, ma possa invece essere perseguita mantenendo la solidità del costrutto e la coerenza con i framework metodologici consolidati nella letteratura specialistica.

La struttura del lavoro riflette questo triplice ancoraggio disciplinare e, attraverso l'articolazione in quattro capitoli, conduce progressivamente dalla riflessione teorica alla validazione empirica dello strumento sviluppato. Il primo capitolo ricostruisce la cornice concettuale di riferimento, introducendo le principali questioni teoriche rilevanti per uno studio incentrato sull'acquisizione e sulla valutazione delle lingue seconde. In apertura viene affrontato il rapporto tra *Second Language Acquisition* e *Language Testing*, due ambiti di ricerca che, pur condividendo l'interesse per la competenza linguistica degli apprendenti, si sono sviluppati per molto tempo lungo percorsi relativamente indipendenti. Il capitolo mette quindi in evidenza i principali punti di convergenza e le possibili sinergie tra le due discipline. Successivamente vengono posti in rassegna i modelli di competenza linguistico comunicativa che si sono susseguiti nel tempo, con l'obiettivo di definire il "che cosa" un test linguistico dovrebbe misurare. A queste riflessioni si affianca un'analisi del ruolo delle politiche linguistiche europee nella standardizzazione della valutazione, con particolare attenzione alla genesi, alla struttura e ai limiti del Quadro Comune Europeo di Riferimento per le Lingue (QCER). Il capitolo considera inoltre le caratteristiche e le motivazioni del pubblico dell'italiano L2, la cui marcata eterogeneità evidenzia la necessità di strumenti valutativi flessibili, accessibili e potenzialmente distribuibili su larga scala. La seconda parte del capitolo è dedicata all'evoluzione storico-tecnologica del Language Testing. In particolare, viene ricostruito il passaggio dai test *paper-and-pencil*, caratterizzati dalla predominanza di item a risposta chiusa e da una valutazione interamente affidata al giudizio umano, ai più recenti sistemi CALT e ICALL. Questi adottano una concezione multidimensionale e comunicativa della competenza linguistica e mirano a simulare attività linguistiche riconducibili a contesti

d'uso autentici, rendendo possibile, grazie all'integrazione dell'intelligenza artificiale e delle tecnologie di elaborazione automatica del linguaggio, una progressiva riduzione dell'intervento umano nei processi di somministrazione e valutazione. Il capitolo si conclude con una ricognizione delle principali nozioni del Trattamento Automatico del Linguaggio, prerequisito teorico necessario per comprendere le tecnologie impiegate nei sistemi di *Intelligent Computer-Assisted Language Learning* e, più nello specifico, nel modello sviluppato in questo lavoro.

Il secondo capitolo definisce il quadro teorico e metodologico del *Language Testing* che informa la progettazione dello strumento. Dopo aver descritto la funzione sociale storicamente svolta dalla valutazione linguistica e aver esplicitato la sua dimensione inferenziale – ovvero, il fatto che i test non misurano direttamente la competenza sottesa, ma inferiscono capacità a partire da prestazioni osservabili in compiti specifici – il capitolo si concentra sull'analisi delle qualità essenziali di uno strumento valutativo. Vengono analizzati in dettaglio i sei parametri del framework di Bachman e Palmer (1996) – validità di costrutto, affidabilità, autenticità, interattività, impatto e praticabilità – e le cinque dimensioni della validità nel modello socio-cognitivo di Weir (2005): cognitiva, contestuale, di scoring, relativa ai criteri e consequenziale. Il capitolo presenta inoltre le principali tipologie di prove e i formati di item disponibili, dal continuum che va dalle risposte chiuse selezionate alle risposte aperte estese, con le relative implicazioni in termini di autenticità e validità. Si chiude con la descrizione delle tre fasi di sviluppo di un test secondo Bachman e Palmer – progettazione, operazionalizzazione e somministrazione – che costituisce la traccia metodologica seguita nella realizzazione di ETET e fornisce la struttura portante del capitolo successivo.

Il terzo capitolo costituisce il passaggio dalla dimensione concettuale a quella applicativa, descrivendo ETET nella sua architettura tecnica e nelle scelte di progettazione e operazionalizzazione. Dopo un'iniziale descrizione delle componenti di *back-end* e di *back-office* che compongono la piattaforma e delle diverse funzionalità a disposizione dei *test developer*, si mostra come i principi teorici elaborati nei capitoli precedenti si traducano in decisioni operative concrete. La fase di progettazione documenta la definizione dello scopo del test, del dominio d'uso target (TLU), dei candidati, del costrutto, del progetto di validazione dell'utilità e del piano delle risorse a disposizione e dei vincoli. La fase di operazionalizzazione descrive la costruzione delle cinque sezioni del test – grammatica, lettura, ascolto, scrittura e parlato – ciascuna ancorata a un framework cognitivo specifico. In questa sezione, particolare attenzione è riservata allo sviluppo dei sistemi di scoring automatico per le produzioni aperte valutate tramite LLM. Viene descritto il processo iterativo che ha portato alla selezione del

prompt ottimale per il modello realizzato e della pipeline integrata per la sezione di speaking, che combina i punteggi acustici di Azure con la valutazione linguistica del modello sulle trascrizioni. La fase di somministrazione descrive infine le modalità di reclutamento dei partecipanti, del gruppo di controllo e degli annotatori esperti.

Il quarto capitolo presenta i risultati della validazione empirica condotta su un campione eterogeneo di 45 partecipanti (40 partecipanti effettivi e 5 partecipanti madrelingua italiani che hanno costituito il gruppo di controllo). Dopo un'analisi qualitativa della popolazione e la definizione dei criteri di esclusione per la costruzione del campione – necessaria in ragione dell'elevata incidenza di malfunzionamenti tecnici nella sezione di produzione orale, che ha ridotto le risposte orali utilizzabili da 135 a 95 – la validazione si struttura in tre livelli. Il primo riguarda la costruzione del *gold standard* attraverso il confronto *pairwise* tra tre annotatori umani esperti, misurato mediante la kappa di Cohen pesata con pesi quadratici, con selezione della coppia a più alto accordo. Il secondo livello analizza le concordanze dei risultati tra modello e *gold standard* separatamente per le sezioni di produzione orale e scritta, attraverso kappa pesata, *Mean Absolute Error* (MAE) e *bias* medio per determinare la quantità e la distribuzione dello scarto. Il terzo livello esamina la validità relativa ai criteri attraverso il confronto tra i punteggi del sistema e le autovalutazioni fornite dei candidati.

Nel complesso, il presente lavoro intende contribuire al dibattito sulla validità e sull'affidabilità dei sistemi ICALL per la valutazione linguistica delle L2, offrendo un caso di studio empiricamente fondato su una lingua per cui tali strumenti rimangono ancora largamente assenti.

CAPITOLO 1. RICOGNIZIONE TEORICA

Questo capitolo si propone di delineare la cornice teorica entro cui si colloca il presente studio. La costruzione di strumenti per la valutazione linguistica, e in particolare di modelli volti alla misurazione della competenza in una lingua seconda, non può prescindere da una riflessione approfondita sui presupposti teorici che definiscono che cosa debba essere valutato, con quali finalità e secondo quali criteri. Per questa ragione il capitolo intreccia contributi provenienti dalla *Second Language Acquisition*, dal *Language Testing* e dalla Linguistica Computazionale, mettendone in luce differenze e possibili punti di contatto. A partire da tale intersezione, nei primi paragrafi vengono analizzati i principali modelli di competenza linguistico-comunicativa, con particolare attenzione alle implicazioni che essi hanno per la progettazione dei test e per l'interpretazione delle performance degli apprendenti. Successivamente, viene ricostruito il ruolo svolto dalle politiche linguistiche europee nella standardizzazione della valutazione, soffermandosi sulla genesi, sulla struttura e sui limiti del Quadro Comune Europeo di Riferimento per le Lingue. In questa prospettiva, la riflessione si estende alle caratteristiche dei pubblici dell'italiano L2, la cui eterogeneità – tanto in termini di contesti di apprendimento quanto di motivazioni – sottolinea la necessità di strumenti valutativi flessibili, accessibili e distribuibili su larga scala. La seconda parte del capitolo è dedicata all'analisi dell'evoluzione tecnologica nel Language Testing. Viene ricostruito il percorso che ha trasformato la tecnologia da semplice strumento di supporto procedurale a componente integrale dell'intero ciclo di vita del test. In questo contesto, il paradigma *Human-in-the-Loop* emerge come soluzione metodologica fondamentale, capace di integrare le potenzialità dell'automazione con la supervisione umana, garantendo validità e trasparenza nei processi valutativi.

1.1 Fondamenti teorici della Second Language Acquisition

Preliminare a qualsiasi indagine condotta nell'ambito dell'acquisizione delle lingue seconde (*Second Language Acquisition*; SLA) e della valutazione dell'apprendimento linguistico (*Language Testing*; LT) è la definizione di alcune nozioni fondative delle due discipline. Prima fra tutti è la distinzione concettuale tra lingua straniera (LS) e lingua seconda (L2). Si definisce "straniera" la lingua che viene appresa in un contesto in cui essa non è presente se non in una dimensione scolastica e in cui l'input linguistico nella lingua target viene fornito esclusivamente dall'insegnante. Differentemente, la lingua seconda è presente anche in contesto extrascolastico e la maggior parte dell'input linguistico arriva dall'esterno. Il termine "seconda" fa riferimento al fatto che tale lingua viene appresa in una fase successiva rispetto alla lingua madre o prima lingua (L1). È tuttavia consuetudine utilizzare l'espressione L2 per fare riferimento anche a lingue che non vengono apprese cronologicamente come seconde ma anche come terze, quarte e così via (Rod, 2019)[1]. Citando Balboni (2002: 58) «È "straniero" l'inglese studiato in Italia, mentre è "seconda lingua" quando è studiato in Inghilterra».

Il sistema linguistico che viene utilizzato da persone che si apprestano ad apprendere una lingua altra rispetto alla propria L1 viene definito interlingua, concetto introdotto da Larry Selinker in uno dei lavori fondativi della disciplina SLA. Selinker definisce l'interlingua come «un sistema linguistico separato [...] che risulta dai tentativi, da parte di un apprendente, di produrre una norma della lingua d'arrivo» (1972: 14). L'interlingua si delinea pertanto come un sistema linguistico autonomo, differente dalla L1 ma non ancora pienamente identificabile nella L2. Tale sistema è dinamico e provvisorio, ed evolve con il progredire dell'apprendente nella lingua target.

Assieme alla nozione di interlingua, Selinker introduce il concetto di fossilizzazione, ovvero il processo attraverso il quale alcuni tratti dell'interlingua tendono a stabilizzarsi in modo permanente senza più evolvere verso le norme della lingua target. Se nella L1 tutti i parlanti riescono a raggiungere lo stato finale – ovvero a convergere verso un livello di competenza linguistica definita nativa – questa certezza non è invece presente nell'apprendimento e acquisizione della L2. La fossilizzazione può essere favorita da diversi fattori, tra cui l'influenza della L1 (definita *transfer* negativo), la mancanza di *feedback*

[1] In questo lavoro l'espressione L2 verrà utilizzata con valenza di iperonimo rispettivamente rispetto a lingua straniera e lingua seconda, come è consuetudine negli studi in queste discipline.

correttivo o la sufficiente funzionalità comunicativa di forme non *target-like*. Anche per effetto della fossilizzazione gli apprendenti di una L2 dimostreranno livelli differenti di competenza linguistica.

L'altra opera solitamente considerata capostipite della disciplina SLA è il saggio *The Significance of Learners' Errors* di Pit Corder (1967), in cui l'autore ridefinisce il concetto di errore. Corder sostiene che gli errori sistematici nella produzione linguistica degli apprendenti non siano da intendere come sintomatici di un apprendimento difettoso, ma come varchi in cui ricostruire il modo in cui la mente elabora il linguaggio. Attraverso l'analisi di questi errori si può indagare come l'acquisizione procede e determinare alcuni dei processi ad essa sottostanti. Egli propone inoltre una distinzione tra errori propriamente detti, legati a lacune nella competenza, ed errori occasionali, legati a fattori contingenti di performance:

> «Dobbiamo quindi fare una distinzione tra quegli errori che sono il prodotto di tali circostanze casuali e quelli che rivelano la sua conoscenza di base della lingua fino a quel momento, o, come potremmo chiamarla, la sua competenza transitoria. Gli errori di esecuzione saranno tipicamente non sistematici e gli errori di competenza sistematici» (1967: 166; traduzione mia)[2].

Tale distinzione fornisce una base concettuale e diagnostica fondamentale per lo studio della SLA e per la progettazione di strumenti valutativi.

Numerose ricerche hanno poi indagato quali siano le modalità che un apprendente ha a disposizione per "impadronirsi" di una L2. Studio pioneristico in questa direzione è rappresentato dall'opera *Second language acquisition and second language learning* di Stephen Krashen (1981). Nello scritto l'autore teorizza la *Second Language Acquisition Theory* (SLAT) e propone una differenza dicotomica tra acquisizione e apprendimento. Se l'acquisizione è un processo inconscio e involontario, tipico dei contesti immersivi o naturalistici, l'apprendimento – alla base dei modelli di istruzione guidata – è una modalità consapevole che implica un processo razionale. In continuità con la distinzione proposta da Krashen, il neurolinguista Ullman (2001) fornisce con il suo modello Dichiarativo-Procedurale una spiegazione neuro-cognitiva di tale differenza. Ullman sostiene che anche il tipo (e la durata) di esposizione linguistica – ovvero un'esperienza immersiva o un'esperienza guidata – possano influenzare il

[2] «We must therefore make a distinction between those errors which are the product of such chance circumstances and those which reveal his underlying knowledge of the language to date, or, as we may call it his transitional competence. The errors of performance will characteristically be unsystematic and the errors of competence, systematic».

modo in cui l'apprendente si impadronisce della lingua e la relativa dipendenza che ne deriva dai due sistemi di memoria a lungo termine: la memoria dichiarativa e la memoria procedurale. In particolare, l'insegnamento esplicito della grammatica, tipico dei contesti scolastici, tende a favorire l'apprendimento mediato dalla memoria dichiarativa, a scapito di quello procedurale. Al contrario, la mancanza di istruzione formale, come avviene spesso nei contesti immersivi, favorisce l'apprendimento implicito attraverso la memoria procedurale.

Accanto ai processi cognitivi, anche fattori emotivi e motivazionali svolgono un ruolo fondamentale quando si impara una nuova lingua. Ulteriore concetto introdotta da Krashen nella SLAT è la nozione di filtro affettivo, ovvero un meccanismo psicologico che può inficiare la capacità di assimilare un input linguistico. Secondo l'autore, elementi che possono determinare l'innalzamento o l'abbassamento del filtro sono l'ansia, la motivazione e l'autostima. Quando l'apprendente è motivato, sicuro di sé e inserito in un contesto di apprendimento positivo, il filtro affettivo è basso e l'acquisizione linguistica risulta facilitata; al contrario, alti livelli di ansia o scarsa motivazione elevano il filtro, limitando l'efficacia dell'apprendimento. La motivazione risulta in questo senso un fattore determinate per l'acquisizione poiché influisce non solo sulla quantità di input a cui l'apprendente è esposto, ma anche sul grado di coinvolgimento e di persistenza nel processo di apprendimento.

Balboni (2002) distingue tre tipologie differenti di motivazioni che governano l'agire degli individui: il dovere, il bisogno e il piacere. Il primo comporta l'innalzamento del filtro affettivo e non porta all'acquisizione; il secondo è legato alla consapevolezza della difficoltà, il che porta ad un abbassamento del filtro e a generare in questo modo un'acquisizione duratura; l'ultimo, coinvolgendo entrambi gli emisferi del cervello, comporta una situazione di acquisizione definita "potentissima".

Gardner e Lambert (1972) si inseriscono in questa discussione proponendo un'ulteriore distinzione all'interno della motivazione. Secondo gli autori vi è una motivazione strumentale, orientata al solo apprendimento delle abilità linguistiche necessarie per comunicare e una motivazione integrativa, il cui obiettivo è l'inserimento dell'individuo nella cultura che la lingua veicola. Entrambe le motivazioni possono coesistere nei parlanti, seppur con intensità differenti. Tuttavia, quando prevale esclusivamente la motivazione strumentale, si corre il rischio di incappare nella fossilizzazione (Guerini, Dal Negro 2007).

Come si è cercato di delineare fino ad ora, assieme al grado e al tipo di esposizione all'input, anche fattori psicologici e motivazionali concorrono a generare i differenti esiti acquisizionali osservabili negli apprendenti di una L2. Tali fattori non sono direttamente misurabili in una

situazione di test ma, come vedremo nei paragrafi successivi, devono essere tenuti in considerazione nell'interpretazione dei risultati della valutazione, aprendo la riflessione sul rapporto tra processi acquisizionali e pratiche valutative.

1.2 L'interfaccia tra Second Language Acquisition e Language Testing

Come accennato nel paragrafo precedente, le discipline che si occupano di studiare e valutare le lingue seconde sono rispettivamente la *Second Language Acquisition* e il *Language Testing*, entrambe branche della linguistica applicata. Tuttavia, a causa dell'elevato grado di specializzazione richiesto a chi opera in questi ambiti, in unione alla progressiva istituzionalizzazione di conferenze e riviste dedicate esclusivamente all'uno o all'altro campo, le due discipline sono andate incontro a un sorprendente e immotivato isolamento concettuale (Bachman, Cohen 1998). Se l'oggetto in analisi delle due discipline – ovvero la L2 – risulta essere il medesimo, differenti sono le prospettive, le finalità dello studio e le metodologie impiegate. Prima di analizzare nel dettaglio le principali differenze, proponiamo una breve descrizione delle due discipline.

Gli studi sull'acquisizione linguistica affondano le loro radici nella linguistica strutturale e nel comportamentismo, concentrandosi inizialmente in modo esclusivo sull'acquisizione della lingua materna. Sebbene i paradigmi teorici di riferimento siano cambiati nel tempo, seguendo l'evoluzione delle principali correnti linguistiche e psicologiche, è rimasta a lungo invariata l'idea che l'acquisizione linguistica costituisse un fenomeno esclusivo dell'infanzia. Importante è stata in questo senso la scoperta che i bambini apprendessero in maniera spontanea la propria lingua, seguendo ordini e sequenze di sviluppo identificabili come universali nel processo di acquisizione (Brown 1973; Bellugi 1967). È solo con i già citati studi rivoluzionari di Corder e Selinker che si assiste a un cambio di paradigma nella disciplina e ci si inizia a interessare dei processi sottostanti all'acquisizione di individui già pienamente competenti in una L1. Le domande-guida di questa transizione teorica sono le seguenti: l'acquisizione della L2 è paragonabile a quella della L1? Quale è lo stato iniziale della SLA? (VanPatten et al. 2020). La risposta a questi quesiti non è immediata e ad oggi è ancora oggetto di importanti discussioni tra gli studiosi.

La disciplina della valutazione linguistica nasce in ambito anglosassone in risposta alla necessità di standardizzare procedure e strumenti di valutazione e misurazione della competenza linguistica. Tale esigenza è scaturita dalla diffusione su larga scala dell'inglese: inizialmente come lingua del potere coloniale e, in seguito, come lingua dominante nei settori economici, scientifici e tecnologici. Se i numerosi studi e la diffusione sempre più capillare testimoniano la rapida evoluzione che ha coinvolto la disciplina, il dibattitto sulla posizione che questa occupa all'interno delle scienze linguistiche rimane ancora aperto (McNamara 2004). Tuttavia, come evidenziato da Barni:

> «Alla marginalità ancora riservata al LT nel quadro epistemologico delle scienze del linguaggio non corrisponde un'analoga posizione nel sistema sociale che struttura i campi di applicazione di tale disciplina. Ne è testimonianza il ruolo che il LT riveste nel sistema formativo, scolastico e universitario, così come nell'orientamento e nell'inserimento nel mondo del lavoro» (2005: 31).

Le principali differenze delle due discipline vengono schematizzare da Bachman e Cohen nell'opera *Interfaces between second language acquisition and language testing research*, come riportato nella Tabella 1.

#	**Second Language Acquisition**	**Language Testing**
Research perspective	Longitudinal view of inter language development	"Slice of life" view of language ability at a given stage of development, with reference to a given norm or standard of language use.
Focus of research	Antecedents of interlanguage ability: factors and processes that affect or are part of language acquisition, e.g., contextual features, learner characteristics, processes.	Results of language acquisition: components and strategies that are part of language ability, e.g., grammatical competence, pragmatic competence, strategic competence.
Goals of research	To develop and empirically validate a theory of SLA that will (1) describe how SLA takes place and (2) explain why SLA takes place.	(1) To develop and empirically validate a theory of language test performance that will describe and explain variations in language test performance; and (2) to demonstrate the ways in which language test

		performance corresponds to nontest language use.
Research methodology	Variety of research approaches: discourse analysis, case studies, ethnography, experimental, quasi-experimental, ex post facto correlational.	Dominant research approach: ex post facto correlational; increasing use of qualitative analysis of test content and test takers'responses, verbal reports of test taking.

Tabella 1: Second Language Acquisition e Language Testing a confronto (Bachman, Cohen 1998:2-3)

Se la SLA si concentra, da un lato sui processi che soggiacciono all'acquisizione e, dall'altro, analizza l'interlingua con l'obiettivo di scovare sistematicità e prevedere lo sviluppo della competenza linguistica, il LT si focalizza invece sui prodotti dell'acquisizione, ossia su ciò che l'apprendente è effettivamente in grado di fare nella lingua oggetto di valutazione. Allo stesso modo, le metodologie impiegate nelle rispettive analisi sono differenti: sperimentali o semi-sperimentali per la SLA e metodi ex post facto o causali - rappresentativi per il LT. Nello specifico, nel primo caso il ricercatore volontariamente controlla e manipola le condizioni e le variabili che determinano i fenomeni sotto esame, con l'obiettivo di stabilire relazioni causa-effetto (viene quindi posto in esame il processo linguistico); nel secondo, il fenomeno sotto indagine è già avvenuto (e pertanto viene posto in esame il prodotto linguistico). Il compito del ricercatore sarà quello di analizzare i dati, confrontare eventuali campioni cercando di individuare regolarità e pattern ricorrenti al fine di identificare possibili relazioni di causa (Mantovani 1998).

Le differenze appena descritte possono essere attenuate dall'interazione fruttuosa tra le due discipline. Elana Shohamy (1998) individua sei possibili scambi positivi tra i due campi e li suddivide in base alla direzione del contributo: tre riguardano i modi in cui il LT può supportare la SLA e tre in cui la SLA può arricchire il LT. Nello specifico, gli ambiti in cui la valutazione può avvantaggiare la SLA sono identificabili:

1) nella definizione del costrutto di abilità linguistica: uno degli obiettivi principali del LT è la definizione di che cosa si intende per abilità linguistica (come si potrà leggere nel dettaglio nel paragrafo 1.3). Per questa ragione, i modelli sviluppati nel LT possono essere utilizzati per testare la validità dei risultati acquisizionali;
2) nell'applicazione di risultati LT per testare ipotesi SLA: durante i test linguistici i ricercatori riescono a raccogliere numerosi dati sul comportamento linguistico degli

apprendenti. Tali informazioni possono essere utili per generare, verificare confermare ipotesi sull'acquisizione;
3) nel fornire alle ricerche SLA criteri validi per la scelta di task e test: definito da Shohamy come uno dei più importanti contributi che il LT può dare alla SLA, il mondo valutativo può fornire indicazioni pratiche circa la metodologia da utilizzare nell'elicitazione dei dati e nella costruzione di task, contribuendo in questo modo ad aumentare il rigore della ricerca SLA.

D'altra parte, gli ambiti in cui la SLA può giovare al LT sono:

1) l'identificazione delle componenti linguistiche che devono essere elicitate: le scoperte sull'acquisizione possono rappresentare una guida per la definizione di che cosa debba essere valutato e la relativa calibrazione, ovvero se la richiesta linguista sia adatta al livello dell'apprendente;
2) il proporre nuovi task che possono essere utilizzati per l'elicitazione: i numerosi anni di ricerca in ambito SLA hanno permesso di sviluppare vari strumenti per elicitare i dati. Tali strumenti potrebbero ampliare il repertorio di tecniche utilizzate dal LT;
3) l'informare i costruttori di test delle differenze individuali che possono intercorrere nell'acquisizione: gli studi sulla SLA possono evidenziare la presenza di possibili aree problematiche per gli apprendenti. Una tra queste potrebbe essere, per esempio, l'influenza che la L1 esercita sull'acquisizione. Avendo questa informazione, i *tester* potrebbero tenere in considerazione i *background* linguistici dei candidati nella costruzione dei test.

Come sottolineato anche da Ferrari:

> «Chi si occupa di testing può ritrovare nella letteratura acquisizionale informazioni importanti per una migliore definizione degli elementi linguistici testabili nei diversi livelli di proficiency. D'altro canto, chi si occupa di ricerca acquisizionale può imparare dalla letteratura relativa al testing l'importanza di una più rigorosa definizione operativa dei costrutti teorici e della verifica di validità e affidabilità degli strumenti di elicitazione e misurazione» (2019:11-12).

In conclusione, la panoramica finora delineata cerca di mettere in evidenza tanto le specificità delle singole discipline quanto gli elementi di contatto e di complementarità. In questo lavoro, proprio l'integrazione tra la SLA e il LT costituirà il presupposto teorico e metodologico del

progetto. L'obiettivo sarà quello di costruire uno strumento per la valutazione dell'italiano L2 che sia al tempo stesso teoricamente solido – perché ancorato ai principi dello sviluppo acquisizionale – e metodologicamente valido, grazie a criteri e alle norme proprie del Language Testing.
Dopotutto, come affermato da Spolsky, i test linguistici:

> «costringono costantemente questioni pratiche e teoriche a una feconda tensione. Le esigenze del tester mettono regolarmente alla prova il teorico, proprio come le scoperte del teorico tentano ripetutamente il tester» (1995; traduzione mia)[3].

1.3 La competenza linguistico-comunicativa: modelli teorici

Il problema principale per chi si appresta a costruire un test di lingua è quale debba essere l'oggetto della valutazione, ovvero che cosa si intende per competenza linguistico-comunicativa. Delineare il costrutto della competenza linguistica non rappresenta una questione esclusivamente teorica, ma ha ricadute dirette sulle modalità di progettazione e valutazione dei test linguistici: ogni modello di competenza implica specifiche scelte valutative, orientando il tipo di prove, i task proposti e i criteri di valutazione adottati.

Il concetto stesso di competenza è stato a lungo al centro del dibattito linguistico. Una delle definizioni che ha esercitato maggiore influenza è quella formulata da Noam Chomsky in *Aspects of the Theory of Syntax*, che definisce la nozione di *competence* (competenza linguistica) ponendola in opposizione a quella di *performance* (esecuzione linguistica):

> «La teoria linguistica si occupa principalmente di un parlante ideale, in una comunità linguistica completamente omogenea, che conosce perfettamente la propria lingua e non è influenzato da condizioni grammaticalmente irrilevanti come limitazioni di memoria, distrazioni, spostamenti di attenzione e interesse ed errori (casuali o caratteristici) nell'applicazione della sua conoscenza della lingua nell'esecuzione effettiva. […] Facciamo quindi una distinzione fondamentale tra competenza

3 «Constantly forces practical and theoretical issues into fruitful tension. The needs of the tester regularly challenge the theorist, just as the findings of the theorist repeatedly tempt the tester».

(la conoscenza della lingua da parte del parlante-ascoltatore) ed esecuzione, l'uso effettivo della lingua in situazioni concrete» (Chomsky 1965: 3-4; traduzione mia)[4].

È possibile quindi definire la competenza linguistica come la conoscenza che un parlante ideale possiede del proprio sistema linguistico, intesa come conoscenza grammaticale; differentemente, l'esecuzione linguistica si caratterizza come una manifestazione temporanea e imperfetta del sistema sottostante, che può esistere solo in situazioni reali e, in quanto tali, condizionate da variabili psicologiche e socioculturali. Nei test linguistici quella che potrà essere valutata sarà esclusivamente l'esecuzione, a meno che non siano previsti task che mirano a riconoscere la grammaticalità o agrammaticalità degli enunciati (Porcelli 1992).

Uno dei primi tentativi di descrivere la competenza linguistica è rappresentato dal modello proposto da Lado e Carrol (1961). I due studiosi proposero una netta differenza concettuale tra le nozioni di abilità e conoscenza, senza però specificarne la reciproca relazione e il legame con il contesto d'uso. Le abilità linguistiche – che da questo momento in poi verranno definite abilità tradizionali – sono: la comprensione scritta (*reading*), la comprensione orale (*listening*), la produzione scritta (*writing*) e la produzione orale (*speaking*). Differentemente, la categoria delle conoscenze linguistiche è costituita dalla grammatica, dal lessico, dalla fonologia e dalla grafematica. Secondo Lado, gli aspetti centrali della competenza sono rappresentati dal sistema grammaticale, dal vocabolario e dalla pronuncia. Questi elementi vengono interpretati come entità indipendenti, separate le une dalle altre e decontestualizzate rispetto alle condizioni reali dell'uso della lingua. Ne deriva un modello fortemente indirizzato alla descrizione strutturale degli elementi linguistici, in cui le abilità e le conoscenze coesistono come entità separate, esulate da qualsiasi riferimento contestuale o concezione comunicativa dell'uso linguistico. Questa teoria si inserisce nella fase del LT definita psicometrico-strutturalista, poiché caratterizzata dall'influenza della linguistica strutturale e della psicometria classica. In questo periodo si afferma l'esigenza di rendere la valutazione più oggettiva, tramite l'ampio utilizzo di *item* a risposta chiusa e domande a scelta multipla e la conseguente esclusione di tutte le prove che implicavano invece un grado di soggettività, come esercizi di scrittura, traduzione e produzione orale. La forma di testing più adottata è quella del *discrete point test* teorizzata da

4 «Linguistic theory is primarily concerned with an ideal speaker-liste, in a completely homogeneous speech community, who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. […] We thus make a fundamental distinction between competence (the speaker–hearer's knowledge of the language) and performance, the actual use of language in concrete situations».

Lado (1961) che implicava uno scorporamento delle diverse componenti linguistiche da valutare mediante, appunto, *item* a scelta multipla o a risposta chiusa. Nello stesso periodo Caroll (1961), in opposizione al framework presentato da Lado, teorizza l'*integrative test*; qui la competenza linguistica non viene misurata attraverso la verifica di singole componenti linguistiche isolate ma la lingua viene considerata nella sua globalità. Tentativi iniziali di superare la frammentazione del testing strutturalista sono rappresentati da strumenti come i cloze test e dal dettato[5].

Negli anni Settanta, il concetto di competenza linguistica viene progressivamente ampliato e arricchito grazie ai contributi di Hymes (1972), Halliday (1976) e Van Dijk (1977). Gli studiosi contestano la visione puramente strutturale proposta da Lado e Carrol e introducono l'importanza del contesto – inteso come contesto situazionale – e del cotesto, ovvero il contesto linguistico. Questa aggiunta ha una portata innovativa, poiché per la prima volta l'attenzione non viene concentrata esclusivamente sul contenuto proposizionale dell'enunciato, ma quest'ultimo viene analizzato anche all'interno di un contesto situazionale comunicativo. In questo rinnovamento, un ruolo centrale è stato rappresentato da Hymes, che per primo conia l'espressione competenza comunicativa, riferendosi alla capacità di:

> «usare un repertorio di atti linguistici, prendere parte a eventi linguistici, comprendere come gli altri li valutano. Questa competenza, inoltre, si integra con attitudini, valori e motivazioni che riguardano la lingua, le sue caratteristiche, i suoi usi, fondendosi con la competenza che i parlanti hanno nell'integrare la lingua ad altri codici comunicativi» (1972: 277-278).

Da questo momento in poi conoscere una lingua non significa più padroneggiare esclusivamente le regole che la governano, ma essere in grado di utilizzarla in modo efficace e appropriato, in contesti diversi e per esigenze comunicative di vario tipo.

Sempre attorno agli inizi degli anni Settanta si colloca il modello di competenza linguistica introdotto da John W. Oller (1979) come risposta alle teorie che separavano la conoscenza della lingua dal suo uso concreto. I *pragmatic tests* proposti dallo studioso avevano l'obiettivo di valutare la capacità dell'apprendente di utilizzare le conoscenze linguistiche in contesti pragmaticamente significativi.

Il legame introdotto da Hymes tra lingua e contesto viene ripreso e sviluppato da Canale e Swain nell'opera *Theoretical bases of comunicative approaches to second language teaching*

[5] Le prove definite "cloze test" richiedono all'utente di completare un testo con le informazioni mancanti.

and testing (1980). Secondo gli studiosi, la competenza linguistica è composta da quattro elementi: la competenza grammaticale, ovvero relativa al sistema formale della lingua; la competenza discorsiva, relativa all'uso in contesto della lingua; la competenza sociolinguistica, da intendersi come capacità di utilizzare la lingua in maniera appropriata rispetto alla situazione comunicativa, all'interlocutore e allo scopo; la competenza strategica, cioè l'essere in grado di utilizzare tecniche di compensazione durante la performance linguistica (per superare eventuali difficoltà durante l'interazione). Il modello introdotto da Canale e Swain contribuisce al passaggio fondamentale da una visione statica della lingua ad una visione dinamica e funzionale, aprendo la strada alla nozione di competenza linguistico-comunicativa.

È grazie al *Communicative Language Ability* (CLA) teorizzato da Lyle Bachman (1990) che si inizia a parlare in maniera sistematica di competenza linguistico-comunicativa; tale modello è ancora oggi considerato una pietra miliare per gli studi sulla valutazione. Nel CLA, Bachman teorizza tre componenti principali dell'abilità linguistico-comunicativa: la competenza linguistica, la competenza strategica e i meccanismi psicofisiologici. La competenza linguistica, come riportato in Figura 1, viene suddivisa in competenza organizzativa e competenza pragmatica. Rientrano all'interno della competenza organizzativa la competenza grammaticale (lessico, morfologia, sintassi, fonetica e fonologia) e la competenza testuale (coesione e organizzazione retorica del discorso). La competenza pragmatica è a sua volta suddivisa in competenza illocutoria (composta dalla funzione ideativa, manipolativa, euristica, immaginativa) e competenza sociolinguistica (sensibilità alle differenze dialettali, al registro, alla naturalezza, uso e interpretazione di riferimenti culturali e figure retoriche).

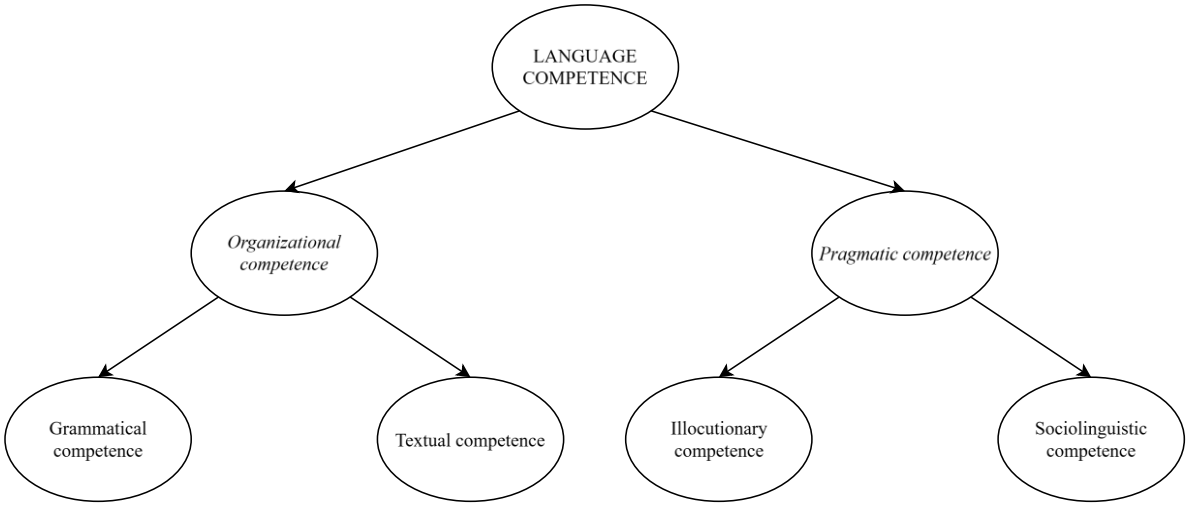


Figura 1:Le componenti della competenza linguistica proposte nel CLA (Bachman1990: 87)

A differenza della modellizzazione proposta da Canale e Swain, Bachman considera la competenza strategica come una componente autonoma, che permette al parlante di pianificare ed eseguire un'interazione comunicativa efficace, selezionando le risorse linguistiche più adeguate al contesto. Questa comprende un fattore di valutazione, uno di pianificazione e uno di esecuzione.

I meccanismi psicofisiologici fanno riferimento ai processi neurologici e psicologici coinvolti nell'atto linguistico. Nello specifico le abilità ricettive (*reading e listening*) prevedono l'utilizzo del canale visivo e uditivo, laddove le abilità produttive (*writing* e *speaking*) impiegano invece componenti neuromuscolari.

Infine, il CLA attribuisce grande rilievo anche alla *knowledge structures*, ovvero alla conoscenza del mondo e alle caratteristiche personali dell'apprendente. La Figura 2 schematizza il modello finora descritto.

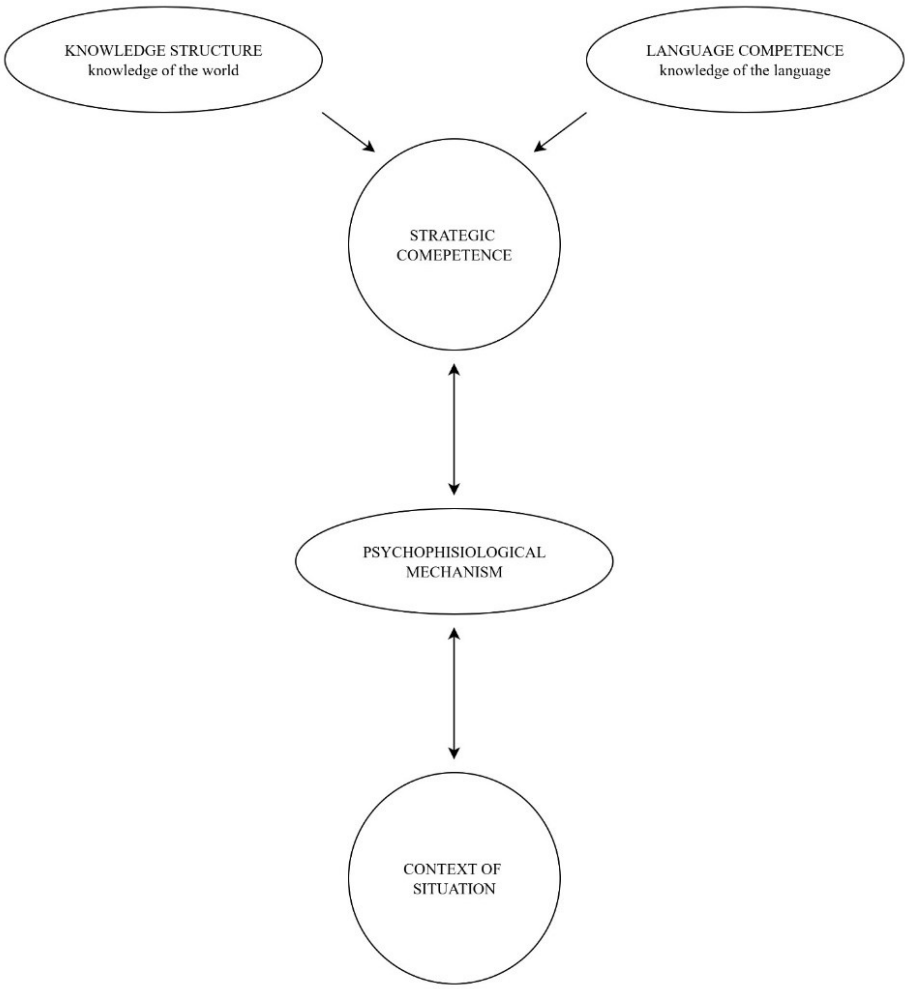


Figura 2: L'abilità linguistico-comunicativa proposta da Bachman (Bachman 1990:84)

Dopo qualche anno, Bachman torna sul modello CLA e, in unione con lo studioso Adrian S. Palmer, ne propone una versione revisionata:

> «A causa della complessità di queste interazioni, crediamo che l'abilità linguistica debba essere considerata all'interno di un quadro interazionale dell'uso della lingua. […] si concentra sulle interazioni tra aree dell'abilità linguistica (conoscenza linguistica e competenza strategica, o strategie metacognitive), conoscenza tematica e schemi affettivi, da un lato, e su come queste interagiscono con le caratteristiche della situazione di uso della lingua, o compito di prova, dall'altro» (Bachman e Palmer, 1996:62; traduzione mia)[6].

Le novità di questo framework sono rappresentate dall'introduzione – in unione alla competenza linguistica e alla competenza strategica – della categoria delle caratteristiche personali, ovvero gli attributi individuali che esulano dalla competenza linguistica ma che possono comunque influenzare una performance. Queste, anche definite *affective schemata* (schemi affettivi), portano l'utente a interpretare e percepire il test e la situazione in essere in funzione di esperienze emotive pregresse. Modifiche terminologiche rispetto al CLA sono la sostituzione delle *knowledge structures* con il *topical knowledge* e *del context of situation* con le *characteristics of the language use situation or test task*. La Figura 3 mostra il modello aggiornato.

[6] «Because of the complexity of these interactions, we believe that language ability must be considered within an interactional framework of language use. […] focuses on the interactions among areas of language ability (language knowledge and strategic competence, or metacognitive strategies), topical knowledge, and affective schemata, on the one hand, and how these interact with the characteristics of the language use situation, or test task, on the other».

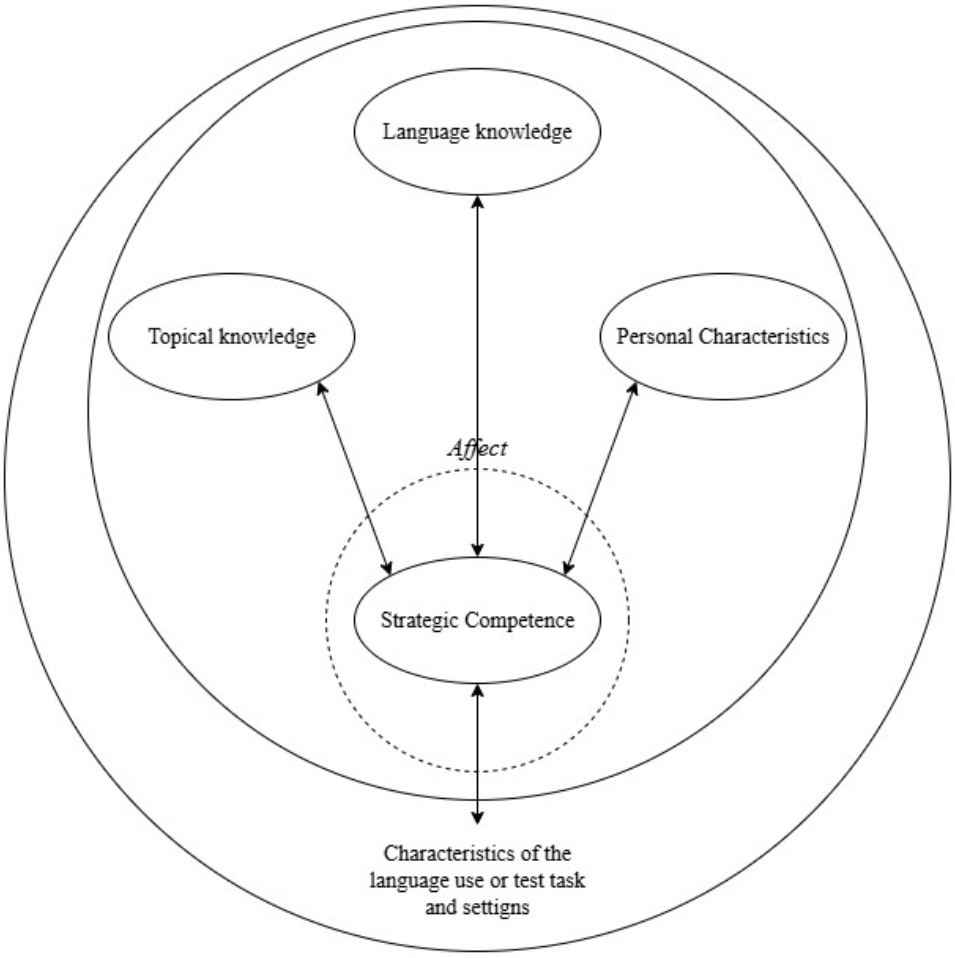


Figura 3: Componenti dell'uso della lingua e della performance (Bachman e Palmer 1996:63)

Attraverso la nuova rappresentazione schematica Bachman e Palmer cercano di esplicitare le principali interazioni che si attivano nell'uso della lingua, indicate attraverso le frecce bidirezionali. Se le caratteristiche generali del compito e della situazione sono raffigurate nel cerchio più esterno, in quello più interno – delimitato da un bordo più marcato – vengono situate le caratteristiche individuali dei soggetti coinvolti nella situazione comunicativa. Ruolo centrale è affidato alla competenza strategica, la quale svolge la funzione di mediazione tra le risorse interne dell'individuo e le richieste poste nella situazione comunicativa o valutativa.

Questo nuovo modello di competenza linguistico-comunicativa permette di delineare le componenti coinvolte nel processo comunicativo e i fattori che possono intercorrere e influenzare la performance dell'esaminato. Tale *framework* si rivela utile tanto nella progettazione delle prove quando nella fase di somministrazione e valutazione. I test che vengono sviluppati a partire da questo modello vengono definiti *communicative language tests* o *performance-based tests*, fondati su compiti che simulano attività linguistiche della vita reale.

Ne fanno parte i *role-play*, le interviste, esercizi di scrittura, di ascolto e di visione di materiale linguistico autentico.

Ulteriore principio fondamentale introdotto dai due studiosi è il concetto di utilità, ovvero: «un test è utile se il suo scopo è chiaro e definito e se costruito in modo da far corrispondere la *performance* del test all'abilità d'uso della lingua del candidato in situazioni non valutative» (Novello 2009: 15). Nel modello proposto da Bachman e Palmer l'utilità può essere scorporata in sei qualità fondamentali e complementari: l'affidabilità, la validità, l'autenticità, l'interattività, l'impatto e la praticabilità. Questi componenti verranno analizzate nel dettagli nel Capitolo 2.

L'ultimo modello di competenza linguistico-comunicativa presentato è quello delineato dal Quadro Comune Europeo di Riferimento per le Lingue (QCER) del Consiglio d'Europa, oggi divenuto paradigma di riferimento in contesto internazionale. Tale modello non propone un nuovo framework teorico ma si inserisce nel solco delle elaborazioni concettuali precedenti, recependone e sistematizzandone i principali contributi. Il QCER, come verrà approfondito più nel dettaglio nei prossimi paragrafi, propone una concezione della competenza comunicativa articolata in competenza linguistica, competenza sociolinguistica e competenza pragmatica, intese come risorse orientate all'azione che l'individuo introduce per agire in contesti sociali. La posizione centrale del QCER nell'ambito della valutazione linguistica è indissolubilmente connessa al contesto istituzionale e alle politiche linguistiche europee che ne hanno orientato la costruzione. Comprendere la genesi del Quadro consente di chiarire le motivazioni che hanno portato alla definizione di criteri comuni per la descrizione e la valutazione delle competenze linguistiche. Il paragrafo seguente ripercorre quindi il panorama delle politiche linguistiche europee che hanno portato all'elaborazione del QCER. Ci avvaliamo del contributo di Vedovelli per ricapitolare gli elementi fin qui delineati:

> «Nella teoria linguistica e nelle metodologie di insegnamento delle lingue si è dunque realizzato il passaggio da un prevalente, se non esclusivo, interesse per la forma (e conseguentemente per la norma) a un prevalente interesse per l'uso della lingua nella sua dimensione socio-pragmatico-comunicativa» (Vedovelli 2003: 28-29).

1.4 La standardizzazione della valutazione linguistica in Europa

Uno dei primi attori nel contesto europeo ad occuparsi della valutazione linguistica è stato il Consiglio d'Europa. Nel secondo dopoguerra infatti, in seguito a rinnovati accordi geopolitici, in Europa sorge una crescente esigenza di cooperazione tra gli stati, espressa anche in termini di mobilità interna e di promozione interculturale. In questo contesto, le lingue mutano progressivamente il loro status, passando dall'essere considerate esclusivamente come oggetto di insegnamento scolastico all'assumere un ruolo politico e sociale centrale; tale mutamento di paradigma concorre a favorire l'integrazione e la coesione tra i cittadini europei.

Primo passo verso la standardizzazione della valutazione in ambito europeo è rappresentato dalla Convenzione culturale europea firmata nel 1954[7]. Da allora il Consiglio d'Europa si è impegnato a sostenere politiche finalizzate tanto all'affermazione del plurilinguismo come valore sociale e culturale, quanto alla creazione di un approccio comune, fondato su presupposti teorici condivisi, in ambito di educazione e valutazione linguistica.

Gli anni Cinquanta e Sessanta vedono la creazione dei primi comitati di esperti e la nascita di numerosi progetti dedicati allo sviluppo delle lingue moderne, con l'obiettivo di promuovere la cooperazione fra sistemi educativi in ambito europeo, favorire il dialogo tra il mondo della ricerca e quello dell'insegnamento e promuovere l'utilizzo delle nuove tecnologie in ambito educativo[8].

A partire dagli anni Settanta – in un'Europa politicamente e socialmente sempre più aperta alla libera circolazione dei propri cittadini – tali proposte trovano una prima sistematizzazione nel modello funzionale-nozionale elaborato da John Trim e Alexander Van Ek (1973). In questo contesto nasce il *Thresold level* (livello soglia) per l'inglese (Van Ek 1975), che rappresenta il primo tentativo di descrivere la competenza linguistica non in termini strutturali, ma, in termini di competenze necessarie per "sopravvivere" in autonomia in un paese in cui la data lingua viene parlata. Nel corso degli anni successivi vari documenti relativi ai livelli soglia delle altre lingue comunitarie iniziano a diffondersi.

[7] La Convenzione culturale europea è operativa dal maggio 1955 e, ad oggi, è sottoscritta da 50 stati (46 stati membri del Consiglio d'Europa e 4 stati che invece non ne fanno parte).
[8] Nel 1957, su iniziativa del governo francese, viene istituito un primo comitato di esperti incaricato di pianificare lo sviluppo dell'insegnamento delle lingue moderne in Europa. Nello stesso anno, la questione viene affrontata anche in un simposio intergovernativo tenutosi a Parigi.

Negli anni Ottanta, grazie al successo e alla diffusione dei livelli soglia con la relativa attenzione alle competenze situazionali e sociali della lingua, si afferma in maniera prorompente l'approccio linguistico-comunicativo. Come osservato da Widdowson:

> «La comunicazione avviene quando si usano frasi per realizzare una serie di atti di natura essenzialmente sociale. Di conseguenza non si comunica componendo delle frasi, ma piuttosto usando le frasi per fare affermazioni di vario tipo, per descrivere, per narrare, per classificare, per fare domande, per fare richieste, per dare ordini. Sapere come mettere insieme delle frasi in modo corretto, è solo una parte di quello che intendiamo per conoscere una lingua e ha di per sé ben poco valore se non supportato dalla consapevolezza di quello che una frase vale come strumento di comunicazione nell'uso quotidiano» (1972: 16).

Tale cambio di prospettiva porta ad una radicale riforma nelle politiche linguistiche: viene avviata un'intensa attività di formazione degli insegnanti attraverso una serie di incontri e seminari a livello internazionale con il fine di attuare anche nella pratica, e non solo nella teoria, l'approccio comunicativo.

È negli anni Novanta che, in concomitanza con i profondi cambiamenti politici e sociali e con il rafforzamento del concetto di cittadinanza Europa, vengono poste le basi per la creazione del QCER. Giuliana Grego Bolli suggerisce che:

> «Il concetto di cittadinanza europea, il conseguente superamento dei confini nazionali, diventa di centrale importanza grazie alle aperture politiche e sociali nate dai radicali cambiamenti politici che hanno investito l'Europa dell'Est e l'Europa Centrale intorno agli anni Novanta. Nuovi paesi divengono membri del Consiglio d'Europa e necessitano, a loro volta, di adeguate politiche linguistiche che rendano effettivamente possibile la comunicazione, l'interscambio e la libera circolazione dei loro cittadini in tutto il territorio europeo» (2010: 7).

In questo clima nel 1991 si tiene il simposio intergovernativo di Rüschlikon dal titolo: *Transparency and coherence in language learning in Europe: objectives, evaluation, certification*; qui vengono stabiliti i principi cardine che dal 1991 al 2001 porteranno alla stesura del documento. Dopo un'iniziale fase di sperimentazione, che ha portato alla pubblicazione di due versioni pilota nel 1996 e nel 1998, nel 2000, nell'anno europeo delle lingue, viene pubblicato il documento ufficiale con il titolo: *Common European Framework of Reference for*

Languages: Learning, Teaching and Assessment (CEFR). In contemporanea vede la luce anche l'European Language Portfolio (ELP).

Nel panorama della valutazione linguistica europea un ruolo fondamentale è ricoperto dall'*Association of Language Testers in Europe* (ALTE)[9]. L'ALTE, fondata nel 1992 su iniziativa delle università di Cambridge e Salamanca, si occupa di generare standard e attuare comparazioni all'interno delle diverse lingue europee seguendo le linee guida proposte dal QCER. La sua attività si colloca in continuità con il lavoro del Consiglio d'Europa e con le proposte del QCER, che l'associazione riprende e sviluppa anche in collaborazione con gli stessi autori del Quadro. Obiettivo principale è favorire il riconoscimento transnazionale delle certificazioni e la loro spendibilità all'interno dello spazio europeo. A questo si affianca la definizione di standard condivisi per tutte le fasi del testing linguistico, dalla progettazione delle prove alla somministrazione, dalla valutazione all'analisi dei risultati, nonché la promozione di progetti comuni e lo scambio di competenze tra gli enti membri. Complessivamente, il lavoro dell'ALTE costituisce un apporto decisivo alla normalizzazione delle pratiche di testing linguistico nel contesto europeo, consolidando la struttura concettuale del QCER e convertendola in metodologie valutative riconosciute su scala internazionale.

1.5 Il Quadro Comune Europeo di riferimento per le lingue

Come accennato nel paragrafo precedente, il QCER nasce in un clima politico di cooperazione e dialogo tra gli stati membri dell'Unione Europea, con l'obiettivo di creare uno strumento tanto teorico quanto pratico, capace di garantire standard di misurazione condivisibili a livello internazionale. L'obiettivo del documento è quello di delineare indicazioni generali per gli ambiti dell'insegnamento e della valutazione, fornendo un repertorio di parametri, categorie e scale a cui gli interessati possono attingere. L'approccio adottato dal quadro è "orientato all'azione", ovvero:

> «considera le persone che usano e apprendono una lingua innanzitutto come "attori sociali", vale a dire come membri di una società che hanno dei compiti (di tipo non solo linguistico) da portare a

[9] https://www.alte.org/

termine in circostanze date, in un ambiente specifico e all'interno di un determinato campo d'azione (2002: 11)».

In questa cornice, la competenza linguistico-comunicativa risulta caratterizzata dalle competenze linguistiche, sociolinguistiche e pragmatiche, ma, va sottolineato:

> «Questa competenza non consiste nella sovrapposizione o nella giustapposizione di competenze distinte, ma è piuttosto una competenza complessa o addirittura composita su cui il parlante può basarsi» (QCER, 2002:205).

Al fine di rendere più chiara la definizione di competenza linguistico-comunicativa nel Quadro vengono proposti dei "livelli comuni di riferimento", raggruppati nelle tre macroaree di competenza elementare, intermedia, avanzata. Ognuna di queste è a sua volta costituita da due elementi, andando a costituire sei livelli complessivi: A1 (contatto), A2 (sopravvivenza), B1 (soglia), B2 (progresso), C1 (efficacia), C2 (padronanza). La Figura 4 rappresenta in maniera schematica i livelli comuni di riferimento.

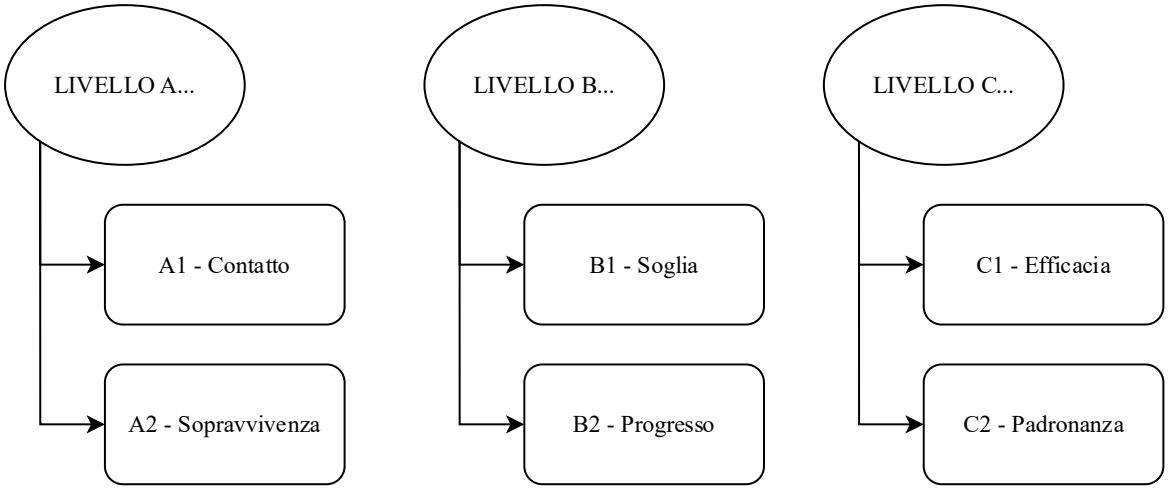


Figura 4: I Livelli comuni di riferimento

Per illustrare i diversi livelli di competenza il documento introduce dei descrittori che fanno riferimento alle attività comunicative, alle strategie e alle competenze linguistico comunicative degli apprendenti; questi vengono presentati attraverso un sistema di griglie. La Tabella 2 mostra quella che viene definita "scala globale". Leggendo la griglia riportata, salterà all'occhio

la formulazione chiara e positiva di tutti i descrittori. È stato infatti scelto di sottolineare quello che l'apprendente è in grado di fare piuttosto che evidenziarne le carenze.

C2	È in grado di comprendere senza sforzo praticamente tutto ciò che ascolta o legge. Sa riassumere informazioni tratte da diverse fonti, orali e scritte, ristrutturando in un testo coerente le argomentazioni e le parti informative. Si esprime spontaneamente, in modo molto scorrevole e preciso e rende distintamente sottili sfumature di significato anche in situazioni piuttosto complesse.
C1	È in grado di comprendere un'ampia gamma di testi complessi e piuttosto lunghi e ne sa ricavare anche il significato implicito. Si esprime in modo scorrevole e spontaneo, senza un eccessivo sforzo per cercare le parole. Usa la lingua in modo flessibile ed efficace per scopi sociali, accademici e professionali. Sa produrre testi chiari, ben strutturati e articolati su argomenti complessi, mostrando di saper controllare le strutture discorsive, i connettivi e i meccanismi di coesione.
B2	È in grado di comprendere le idee fondamentali di testi complessi su argomenti sia concreti sia astratti, comprese le discussioni tecniche nel proprio settore di specializzazione. È in grado di interagire con relativa scioltezza e spontaneità, tanto che l'interazione con un parlante nativo si sviluppa senza eccessiva fatica e tensione. Sa produrre testi chiari e articolati su un'ampia gamma di argomenti e esprimere un'opinione su un argomento d'attualità, esponendo i pro e i contro delle diverse opzioni.
B1	È in grado di comprendere i punti essenziali di messaggi chiari in lingua standard su argomenti familiari che affronta normalmente al lavoro, a scuola, nel tempo libero ecc. Se la cava in molte situazioni che si possono presentare viaggiando in una regione dove si parla la lingua in questione. Sa produrre testi semplici e coerenti su argomenti che gli siano familiari o siano di suo interesse. È in grado di descrivere esperienze e avvenimenti, sogni, speranze, ambizioni, di esporre brevemente ragioni e dare spiegazioni su opinioni e progetti.
A2	Riesce a comprendere frasi isolate ed espressioni di uso frequente relative ad ambiti di immediata rilevanza (ad es. informazioni di base sulla persona e sulla famiglia, acquisti, geografia locale, lavoro). Riesce a comunicare in attività semplici e di routine che richiedono solo uno scambio di informazioni semplice e diretto su argomenti familiari e abituali. Riesce a descrivere in termini semplici aspetti del proprio vissuto e del proprio ambiente ed elementi che si riferiscono a bisogni immediati.
A1	Riesce a comprendere e utilizzare espressioni familiari di uso quotidiano e formule molto comuni per soddisfare bisogni di tipo concreto. Sa presentare sé stesso/a e altri ed è in grado di porre domande su dati personali e rispondere a domande analoghe (il luogo dove abita, le persone che conosce, le cose che possiede). E'in grado di interagire in modo semplice purché l'interlocutore parli lentamente e chiaramente e sia disposto a collaborare.

Tabella 2: Scala globale, Quadro comune europeo di riferimento

Nonostante il ruolo centrale rivestito dal QCER nel panorama europeo, negli ultimi anni gli studiosi ne hanno messo in luce anche alcuni limiti e criticità. Il Quadro nasce per avere una strutturazione generale e non prescrittiva, volta a fornire linee guida per ambiti diversi

all'interno dello studio delle lingue. Proprio questa natura non vincolante, se da un lato ne ha agevolato la diffusione in ambiti e contesti variegati, dall'altro fa emergere alcune carenze strutturali. Nello specifico:

> «la natura generale del documento stesso che, non essendo rivolto a una lingua specifica, richiede necessariamente un'interpretazione, un adattamento e un'integrazione in relazione al contesto della lingua e cultura di riferimento. Una certa flessibilità è di conseguenza necessaria per permettere di descrivere le varianti socioculturali e linguistiche delle diverse lingue» (Spinelli, Parizzi: 2010).

Proprio per queste ragioni i singoli stati hanno sentito la necessità di creare degli strumenti che aiutassero a trasporre le indicazioni del QCER nei contesti specifici delle singole lingue. L'obiettivo di questa nuova generazione di strumenti – definiti Descrizioni per le lingue dei Livelli di Riferimento del Quadro (DLR) – è stato quello di:

> «tradurre, in una particolare lingua, i descrittori del Quadro (che definiscono le competenze degli utenti/apprendenti a un dato livello e per una data abilità) attraverso il materiale linguistico, a essa relativo, necessario per mettere in atto tali competenze. Si "interpretano" i descrittori del Quadro al fine di definire del materiale linguistico "corrispondente" che consenta di produrre dei testi rilevanti per quel livello. Per questa ragione, le DLR si inseriscono nella tradizione dei Livelli soglia, che sono stati a loro volta concepiti per descrivere le diverse lingue, senza essere tuttavia ancorati a un documento di riferimento condiviso» (prefazione XIII).

Sul sito del consiglio d'Europa è possibile leggere che questi livelli di riferimento non sono stati elaborati dal consiglio stesso bensì da team nazionali che hanno utilizzato approcci e metodologie differenti, in alcune occasioni appoggiati da esperti nominati dal consiglio[10]. Sono attualmente disponibili DLR per la lingua ceca, croata, francese, italiana, inglese, portoghese, spagnola e tedesca.

Nonostante le criticità sottolineate, l'apporto del QCER rimane innegabile; riducendo l'arbitrarietà nella definizione della competenza linguistico-comunicativa e nei relativi livelli, ha portato ad una standardizzazione – pur lasciando spazio allo sviluppo di strumenti di supporto – necessaria per equiparare la didattica e la valutazione nel contesto Europeo.

[10] https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions

1.6 Attrattività e pubblico dell'italiano L2

Il campione preso in analisi in questo lavoro è costituito da apprendenti che utilizzano l'italiano come L2; risulta pertanto interessante cercare di delineare quali siano le caratteristiche degli utenti che si accingono ad apprendere la lingua italiana e le motivazioni che li guidano. Secondo le statistiche riportate da Ethnologue, la lingua italiana risulta parlata da circa 3.287.300 parlanti non nativi[11]. Questo numero riflette in gran parte i grandi flussi di emigrazione che hanno caratterizzato l'Italia nel corso della sua storia.

Nello specifico, De Mauro (1991) identifica due ondate migratorie principali: la grande ondata postunitaria, che ha avuto luogo dalla seconda metà dell'Ottocento fino allo scoppio della Grande guerra, e l'ondata avvenuta dopo la Seconda guerra mondiale. Vedovelli (2021) amplia la prospettiva di De Mauro riportandola fino ai giorni nostri:

> «processi paralleli a quelli vissuti entro i confini nazionali hanno investito e visto protagonisti milioni di italiani che a più ondate hanno lasciato il Paese per fare fortuna, nella Merica, o in Australia, in Asia quanto in Africa, prima perlopiù come poveri analfabeti e dialettofoni, spesso esuli politici, di recente anche come cervelli in fuga, laureati e italofoni. E cambiamenti linguistici profondi, anche laceranti, hanno attraversato le famiglie italiane emigrate, le diverse generazioni, fino a quelle dei loro discendenti giovani e giovanissimi per i quali l'italiano è forse ormai solo una lingua straniera, un luogo dove mito e pregiudizi alimentano un immaginario che si intuisce – il più delle volte, solo si intuisce – capace di generare valori» (2021: 23-24).

Nell'ottica dello "spazio linguistico italiano globale", come viene definito da Vedovelli, il Mistero degli Affari Esteri gestisce un complesso ecosistema istituzionale articolato su scala mondiale: dagli Istituti Italiani di Cultura (88) che costituiscono i presìdi culturali ufficiali, a un variegato sistema educativo che comprende istituti statali, scuole paritarie e sezioni italiane integrate in contesti scolastici europei e internazionali (complessivamente oltre 130 strutture), fino alla rete associativa della Società Dante Alighieri con i suoi 422 comitati distribuiti globalmente (Vignoli et al. 2025).

Opera fondamentale per indagare i parlanti di italiano non nativi presenti fuori dal contesto italiano è *Italiano 2000. I pubblici e le motivazioni dell'italiano diffuso tra stranieri* (2002) di Tullio De Mauro. Nell'opera De Mauro osserva come la diffusione dell'italiano

[11] http://ethnologue.com/language/ita/

all'estero sia dovuta principalmente alla sua percezione come lingua di cultura, il cui interesse è sostenuto da motivazioni simboliche, identitarie, affettive, culturali ed in misura più limitata, strumentali. Anche secondo la definizione proposta da L. J. Calvet (2002) l'italiano si configura come una lingua di cultura, ovvero non riveste una posizione dominante in termini socioeconomici o come lingua veicolare nel sistema mondiale delle lingue (definito dall'autore "mercato"), tuttavia, detiene una rilevante capacità attrattiva grazie al capitale culturale e simbolico accumulato nel corso della storia. L'italiano si delinea principalmente come una lingua studiata per scelta e non per necessità.

Come accennato in precedenza, il primo e più consistente gruppo di parlanti di italiano L2 delineato da De Mauro è rappresentato dai discendenti degli emigrati italiani, per i quali, l'apprendimento della lingua svolge una funzione identitaria e risponde all'esigenza di recuperare le proprie radici e di affermare la propria appartenenza culturale. Questi parlanti vengono denominati *Heritage speakers*, ovvero coloro che parlano una lingua ereditata dai propri familiari, diversa dalla lingua maggioritaria del paese in cui il soggetto vive (Benmamoun 2021).

Seconda categoria è costituita dagli studenti appartenenti ad università straniere che si avvicinando alla lingua per ricercare un legame con la tradizione intellettuale che ad essa soggiace. Motori attrattivi principali in questo contesto sono lo studio della letteratura, della produzione artistica e musicale che, secondo De Mauro, «attivano processi di attrazione che mitizzano tutto il "sistema Italia"» (2002: 242). In questi contesti la padronanza della lingua originale costituisce un valore aggiunto rispetto a forme di fruizione mediata. In contesto estero, tuttavia, l'italiano non costituisce la prima lingua scelta dagli studenti ma si colloca comunque all'interno delle cinque lingue che vengono maggiormente selezionate (assieme all'inglese, francese, spagnolo, tedesco).

Un ulteriore pubblico è composto dai professionisti dei settori specifici del *made in Italy*, quali per esempio la gastronomia, i motori e la moda. Anche in questo caso l'approccio alla lingua è dettato principalmente per questioni simboliche e positive impersonificate dall'Italia.

Accanto a queste categorie definite tradizionali, nella sua indagine, De Mauro identifica nuove categorie attrattive quali: il turismo, alcuni aspetti della società e cultura moderna italiana come il cinema e il cantautorato, la partecipazione a programmi di mobilità internazionale, la presenza obbligatoria dell'italiano nel curriculum scolastico, particolari interessi lavorativi e infine, motivazioni personali come l'utilizzo della lingua per interagire con il partner.

In studi più recenti rispetto all'indagine condotta da De Mauro sono stati presi in considerazione ulteriori contesti di uso dell'italiano come L2; in particolare, è stato analizzato il caso del Business Process Outsourcing e dei call center delocalizzati (Wang et al 2013; Bourding et al 2023; Combei 2023). Gli studiosi evidenziano come in queste realtà la lingua non rappresenti un semplice strumento di comunicazione, ma costituisca una risorsa produttiva centrale.

In questi ambiti, il più delle volte, i parlanti L2 si interfacciano con un pubblico di parlanti nativi ed è stato mostrato come la percezione dell'accento agisca come un marcatore sociale, influenzando negativamente il riconoscimento e la legittimazione da parte dell'utente. È stato osservato inoltre come l'accento dell'operatore possa influire sulla fiducia del cliente, sulla disponibilità alla cooperazione, sulla valutazione della competenza professionale e sulla soddisfazione del servizio. In questo quadro, la lingua diventa parte integrante del valore vendibile del servizio, contribuendo in maniera significativa alla costruzione della qualità percepita e alla competitività dell'offerta nei contesti di *outsourcing*.

Parallelamente alle dinamiche di diffusione internazionale della lingua si delinea un pubblico cospicuo di parlanti che hanno appreso l'italiano direttamente in Italia. In questo caso, il pubblico principale è rappresentato da persone immigrate che necessitano di imparare la lingua per inserirsi nel tessuto sociale e lavorativo del paese. Per questa categoria di persone l'italiano riveste una funzione strumentale e funzionale, diventando allo stesso tempo un veicolo e dispositivo di inclusione. Tale dimensione funzionale è stata progressivamente rafforzata da un processo di istituzionalizzazione della competenza linguistica, che ha condotto alla formalizzazione di requisiti linguistici in ambito giuridico, formativo e lavorativo. Un caso emblematico è costituito dall'inserimento dell'articolo 9.1 nella normativa n. 91/1992, che vincola l'acquisizione della nazionalità italiana al raggiungimento di una padronanza linguistica pari almeno al livello B1 stabilito dal QCER. La verifica di tale competenza linguistica avviene attraverso il superamento di esami di italiano L2 riconosciuti a livello internazionale, rilasciati da organismi certificatori accreditati dal Ministero degli Affari Esteri e della Cooperazione Internazionale. I principali enti certificatori sono l'Università per Stranieri di Perugia, che rilascia le certificazioni CELI (Certificato di Conoscenza della Lingua Italiana)[12], l'Università per Stranieri di Siena, responsabile delle certificazioni CILS (Certificazione di Italiano come Lingua Straniera)[13], l'Università degli Studi Roma Tre, che gestisce la certificazione CERT.IT

[12] https://cvcl.unistrapg.it/
[13] https://www.unistrasi.it/

(Certificazione dell'italiano come lingua straniera) [14], e la Società Dante Alighieri, che rilascia la certificazione PLIDA (Progetto Lingua Italiana Dante Alighieri)[15].

Analogamente, in ambito universitario, agli studenti provenienti da paesi non appartenenti all'Unione Europea è generalmente richiesto il possesso di una competenza linguistica di livello B2 per l'accesso ai corsi erogati in lingua italiana.

Il quadro delineato mostra un'eterogeneità nei pubblici e nelle motivazioni che inducono gli apprendenti ad acquisire la lingua italiana come L2. Allo stesso tempo, tale varietà di situazioni linguistiche mette in evidenza come la competenza in italiano assuma per ogni individuo valori e funzioni differenti. In questo contesto, e vista la crescente richiesta di certificazioni e riconoscimenti in ambito istituzionale, scolastico e lavorativo, emerge la necessità di affiancare alla valutazione tradizionale strumenti che siano in grado di fornire una valutazione solida dal punto di vista teorico ma allo stesso tempo rapida, accessibile e distribuibile su larga scala. È in risposta a queste esigenze che si colloca il presente lavoro, proponendo uno strumento pensato per svolgere una funzione di supporto nei contesti ad alta richiesta o come complemento ai percorsi didattici e certificativi. Prima di illustrarne le caratteristiche e il funzionamento risulta necessario inquadrare le tecnologie che vengono impiegate tramite una breve ricognizione storico-tecnologica.

1.7 L'evoluzione tecnologica dei test linguistici: dai PPT ai CALT

L'evoluzione dei test linguistici e delle pratiche valutative nelle L2 ha proceduto in parallelo con l'avanzamento tecnologico, riflettendo un'interazione sinergica tra innovazioni digitali e metodologie di valutazione. Se per anni lo stato dell'arte è stato rappresentato dai test *paper and pencil* (PPT), l'introduzione dell'informatica in questo campo ha portato all'emergere del *Computer Assissted Language Testing* (CALT) del *Computer-Assisted Language Assesment* (CALA) e del *Computer Assisted Language Learning* (CALL), in cui la performance linguistica viene elicitata e valutata con l'ausilio di un computer (Noijons 1994)[16]. Recentemente, con l'avvento dell'intelligenza artificiale (IA), si è assistito alla creazione di un nuovo campo all'interno del CALL, ovvero l'*Intelligent Computer Assisted Language Learning* (ICALL),

[14] https://certificazioneitaliano.uniroma3.it/
[15] https://plida.dante.global/it
[16] In questo lavoro i termini CALT, CALA e CALL verranno utilizzati in maniera interscambiabile.

che applica concetti, tecniche, algoritmi e tecnologie dell'AI al CALL (Heift 2021). Prima di discutere nello specifico le peculiarità di quest'ultima disciplina, si propone una breve ricostruzione delle evoluzioni tecnologiche che si sono susseguite nel Language Testing e che hanno posto le basi per gli strumenti odierni.

Come descritto in precedenza, la metodologia di verifica utilizzata agli albori del LT si fondava sull'utilizzo di prove oggettive altamente controllate che venivano veicolate tramite esami cartacei definiti test *paper and pencil*. Questa modalità rimase la più utilizzata fino agli anni Settanta data l'elevata standardizzazione e l'affidabilità della misurazione. Tuttavia, tale procedura implicava costi e tempistiche di correzione estremamente elevati, soprattutto con l'aumento del numero di candidati nei sistemi educativi di massa del dopoguerra. Per far fronte a questa esigenza, nel 1935 negli Stati Uniti venne sviluppato il modello IBM 805, una macchina elettromeccanica per la tabulazione e il punteggio automatico basata sull'uso di schede perforate[17]. Questo dispositivo, inizialmente concepito per applicazioni commerciali e censimenti, venne rapidamente adottato in ambito educativo per la correzione automatizzata degli esami. I test che venivano distribuiti insieme alla macchina dell'IBM erano costituiti esclusivamente da domande a risposta multipla, un formato che si prestava alla lettura meccanica delle schede. Tale implementazione diede un forte impulso all'utilizzo di esami fondati su questa tipologia di domande a causa della loro semplicità nella correzione; ancora oggi, benché criticata per i suoi limiti nella valutazione di competenze comunicative complesse, rappresenta una delle categorie di task più utilizzati in numerosi contesti valutativi (Fulcher 2000).

A partire dagli anni Sessanta e Settanta, con la diffusione dei primi computer nelle università e nei centri di ricerca, si assiste alla diffusione dei *computer-administered conventional tests*. Questi erano la semplice trasposizione al computer degli esami cartacei e ne mantenevano di fatto il formato e la struttura (Lynch 2022). L'unica differenza sostanziale risiedeva nella modalità di somministrazione che avveniva in formato digitale e la conseguente possibilità del calcolo automatico dei punteggi.

Da un punto di vista psicometrico, queste prime formalizzazioni di test si fondavano sulla *Classical Test Theory* (CTT), secondo la quale, la misurazione della competenza avveniva attraverso la trasformazione del numero di risposte corrette fornite dall'utente in un punteggio globale, ottenuto dalla somma del punteggio osservato e dall'errore di misurazione. Questo

[17] Per maggiori informazioni sul funzionamento del IBM 805 consultare: https://www.ibm.com/history/805-scoring-test

implicava che l'intero test venisse utilizzato come unità di analisi (Gulliksen 1950). Tale misurazione, sebbene efficace per test standardizzati e altamente controllati, risulterà progressivamente meno adatta a forme di valutazione più complesse e dinamiche come quelle che si svilupperanno in seguito.

In queste due prime fasi che hanno caratterizzato il LT la tecnologia si inserisce dapprima come semplice strumento per velocizzare la correzione di domande a risposta predefinita e, in una fase successiva, come medium per supportare il test. Sebbene queste innovazioni abbiano rappresentato un progresso in termini di efficienza di somministrazione, gestione dei dati e velocità di restituzione dei risultati, la semplice digitalizzazione degli esami non comportò un immediato ripensamento della progettazione dei test, dei formati di item o delle procedure di scoring utilizzate. In questa fase, il computer svolgeva principalmente il ruolo di un semplice supporto tecnologico, ospitando pratiche valutative originariamente pensate per il formato cartaceo.

È soltanto a partire dalla metà degli anni Ottanta, con lo sviluppo dei microprocessori e la commercializzazione dei personal computer, che si assiste ad una svolta epocale in ambito LT. Lo sviluppo dei microchip rese possibile la costruzione di computer sempre più piccoli, economici e accessibili, favorendone una diffusione su larga scala. In questo contesto si assiste alla nascita delle discipline del *Computer Assissted Language Testing, Computer Assissted Language Assessment, Computer Assissted Language Learning* e del *Computer applications in Second Language Acquisition* (CASLA) (Chapelle 2001).

Il continuo aumento della capacità computazionale, lo sviluppo di dispositivi hardware sempre più sofisticati (come schede audio, CD-ROM e monitor a colori ad alta risoluzione), la diffusione di sistemi operativi user-friendly e il conseguente miglioramento delle interfacce grafiche, ampliarono significativamente le possibilità di progettazione dei test durante gli anni Novanta. Per la prima volta divenne possibile integrare nei test elementi multimediali quali registrazioni audio, immagini e, successivamente, brevi sequenze video, consentendo la creazione di compiti valutativi più autentici e comunicativamente orientati (Chapelle, Douglas 2006).

Nel corso dei primi anni Duemila la diffusione di Internet e delle tecnologie web-based portarono ad un'ulteriore e radicale trasformazione delle pratiche valutative. I test iniziarono a essere somministrati online attraverso browser, permettendo una maggiore scalabilità, l'accesso remoto da qualsiasi località geografica, la gestione centralizzata e in tempo reale dei dati, oltre a nuove forme di interazione.

Alla luce delle repentine evoluzioni tecnologiche è facile intuire come il *Computer Assissted Language Testing* non si presenti come un campo statico ma come un settore in continua trasformazione. Proprio in virtù di questa complessità intrinseca, Ruslan Suvorov e Volker Hegelheimer (2014) propongono di scorporare e analizzare il CALT attraverso un framework articolato in nove attributi (vedi Tabella 3), ciascuno dei quali composto da specifiche categorie descrittive. Se le prime cinque classi, insieme alla categoria interattiva dell'ultimo attributo, sono tipiche dei sistemi CALT, i restanti quattro attributi sono condivisi da tutte le tipologie di test poiché fanno riferimento a dimensioni valutative e concettuali indipendenti dalla tecnologia utilizzata.

#	Attribute	Categories
1	Directionality	Linear, adaptive, and semi-adaptive testing
2	Delivery format	Computer-based and Web-based testing
3	Media density	Single medium and multimedia
4	Target skill	Single language skill and integrated skills
5	Scoring mechanism	Human-based, exact answer matching, and analysis-based scoring
6	Stakes	Low stakes, medium stakes, and high stakes
7	Purpose	Curriculum-related (achievement, admission, diagnosis, placement, progress) and non-curriculum-related (proficiency and screening)
8	Response type	Selected response and constructed response
9	Task type	Selective (e.g., multiple choice), productive (e.g., short answer, cloze task, written and oral narratives), and interactive (e.g., matching, drag and drop)

Tabella 3: Framework per la descrizione dei Computer Assissted Language Test (Suvorov & Hegelheimer 2014: 2)

Il primo attributo è inerente alla direzionalità, in base alla quale i test possono essere: lineari, adattivi (definiti Computer Adaptive Test, CAT) o semi-adattivi. Negli esami lineari gli utenti svolgono tutti lo stesso test, composto dai medesimi quesiti ed articolato in un tempo predefinito. Gli esami adattivi invece, sono caratterizzati dalla possibilità di adattarsi alla competenza mostrata dagli utenti, proponendo compiti adeguati alle capacità dei candidati. Questo è possibile grazie all'applicazione dell'*Item Response Theory* (IRT), una teoria psicometrica proposta nei lavori pioneristici di Lord (1952) e Rasch (1960).

Se, come descritto in precedenza, la Classical Test Theory utilizza come unità di analisi il test intero, l'IRT viene definita un'«*itemized theory*» (Baker, 1985), ovvero, l'attenzione è

focalizzata sui singoli item e sulla relazione probabilistica tra la presentazione di un item e il livello di abilità latente del candidato. Gli item possono essere caratterizzati da tre attributi: la capacità di discriminare tra due livelli o sottolivelli di competenza contigui, la difficoltà intrinseca e lo *pseudo guessing*, ovvero la possibilità che un utente con bassa competenza indovini la risposta (Baker 2001). La probabilità che l'utente risponda in maniera corretta all'item può essere modellata matematicamente come funzione dei parametri dell'item e del livello di competenza del candidato. Questa relazione è espressa attraverso l'*Item Characteristic Curves* (ICC). Esistono diversi modelli IRT a seconda del numero di parametri che vengono presi in considerazione nella costruzione degli item. Il funzionamento dell'algoritmo di selezione degli item viene descritto in Figura 5. Il test inizia con la somministrazione di un item, tendenzialmente di livello intermedio di difficoltà. In base alla risposta fornita dall'utente il sistema stima il livello di abilità latente del candidato e seleziona dalla banca dati l'item successivo più coerente con tale stima. Se la risposta fornita dall'utente è corretta il sistema selezionerà una domanda più difficile, mentre se la risposta è errata verrà scelto un item più semplice. Questa operazione iterativa continuerà fino a che non verrà raggiunto il criterio di arresto prestabilito.

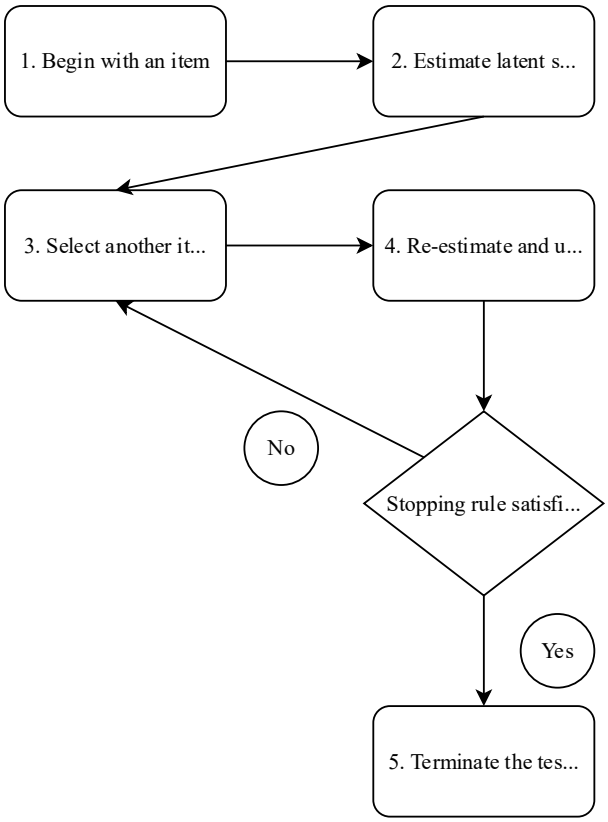


Figura 5:Il diagramma di flusso dei sistemi CAT (Sari et al. 2016: 390)

L'utilizzo dei test CAT permette di diminuire significativamente il numero di item necessari per ottenere una valutazione precisa della competenza della persona testata, riducendo i tempi di somministrazione e l'affaticamento del candidato. Allo stesso tempo, può risultare migliorata anche l'esperienza di test, grazie alla presentazione di item né troppo facili né troppo difficili. Infine, i modelli IRT permettono di risparmiare tempo e risorse per quanto riguarda la somministrazione e la valutazione (Sari et al 2016).

Tuttavia, l'implementazione di esami CAT presenta sfide notevoli. In primo luogo, è necessario costruire banche date di item calibrati estremamente ampie, il che richiede tempo e denaro. Qualora il pool di item non fosse sufficientemente esteso, e questi non coprissero in maniera adeguata l'intero spettro dei livelli di abilità dei candidati, potrebbe avvenire una sovraesposizione degli item (Fulcher 2000). Inoltre, maggiore è il numero di parametri utilizzati dall'algoritmo, più grande è la dimensione del campione necessaria per il pre-test e la calibrazione. Infine, durante gli esami CAT, il candidato non ha la possibilità di rivedere le risposte date in precedenza e gli item non possono essere saltati.

I test semi-adattivi invece, costituiscono una soluzione intermedia rispetto ai test lineari e ad i test adattivi. L'adattamento in questo caso non avviene item per item come nei CAT ma attraverso intere sezioni di domande.

Il secondo attributo che si può analizzare in relazione ai CALT riguarda il formato di erogazione, ovvero la modalità attraverso la quale il test viene somministrato. Gli esami CALT si suddividono in test basati sul computer (CBT) e test basati sul web (WBT). I primi vengono generalmente distribuiti offline tramite postazioni dedicate in ambienti controllati, mentre i secondi sono fruibili online tramite i *browsers*. Ad oggi i WBT rappresentano la categoria più diffusa grazie alle potenzialità legate all'eliminazione di vincoli logistici, temporali e geografici. Questa tipologica offre infatti un'elevata flessibilità, accessibilità e scalabilità.

Non mancano tuttavia le criticità come la mancanza di sicurezza in contesti non supervisionati, dove è difficile garantire l'identità del candidato o impedire aiuti esterni. Inoltre, il malfunzionamento del server può impedire l'accesso al test o interrompere una sessione in corso. In ultimo, la scarsa dimestichezza con il browser o il computer può introdurre varianza irrilevante per il costrutto, influenzando negativamente il punteggio dei candidati meno esperti (Roever 2001).

Terzo parametro da tenere in considerazione è la densità dei media, ovvero la presenza di elementi multimediali nei test come audio, video e immagini che possono avvicinare i compiti linguistici a situazioni comunicative reali. Chapelle & Douglas (2006) argomentano

che in questo modo si possa aumentare l'autenticità situazionale, ovvero la corrispondenza tra il compito di valutazione e le caratteristiche del contesto reale (quali ambientazione, partecipanti, contenuti, tono e genere), e l'autenticità interattiva, cioè il livello di integrazione tra le conoscenze linguistiche del candidato e le richieste comunicative del compito proposto.

Il quarto parametro che possiamo descrivere riguarda la tipologia di competenza che può essere indagata. Negli esami è possibile testare una singola abilità fondamentale (ad esempio la lettura, l'ascolto, la scrittura e il parlato), più abilità in successione o un insieme di competenze integrate, come ascolto e scrittura o ascolto e parlato. La valutazione delle competenze integrate rispecchia la natura complessa dell'utilizzo reale della lingua. Tuttavia, i compiti basati sulle competenze integrate richiedono un notevole sforzo computazionale e di progettazione.

Il quinto attributo è costituito dalla tipologia di valutazione utilizzata nell'assegnazione del punteggio, il quale può essere affidato esclusivamente a valutatori umani, a sistemi automatizzati o ad una combinazione dei due approcci. La valutazione automatica può essere scomposta in due categorie differenti: l'abbinamento di risposte preimpostate attuato da semplici algoritmi e l'utilizzo di sistemi che effettuano un'analisi linguistica più complessa. Quest'ultima tipologia verrà descritta in maniera più approfondita nei prossimi paragrafi e nel Capitolo 3.

Ulteriori attributi riguardano la posta in gioco (*stakes*), che può essere bassa, media o alta a seconda delle conseguenze che i risultati del test comportano per i candidati (da un semplice *feedback* formativo a decisioni di ammissione universitaria o certificazione professionale), e lo scopo del test, che può essere correlato al curriculum (ad esempio per un posizionamento nei corsi o per monitorare i progressi personali) oppure non correlato al curriculum, come nei test utilizzati per scopi certificativi, professionali o di selezione istituzionale.

Infine, dal punto di vista della tipologia di risposta e di attività, i test possono prevedere domande chiuse o predefinite, aperte o produttive e interattive (per esempio task di trascinamento). Nel prossimo capitolo, incentrato in parte sul processo di realizzazione di un test, si parlerà più nel dettaglio di queste specifiche.

Gli attributi fin qui descritti non operano in modo isolato ma sono strettamente interconnessi fra loro, evidenziando come nel campo del CALT e del Language testing in generale l'innovazione tecnologica, la validità teorica del costrutto e l'affidabilità delle misurazioni siano dimensioni interdipendenti. La complessità intrinseca che caratterizza questo campo e il relativo dinamismo sono efficacemente descritti da Hubbard (2009):

> «L'apprendimento delle lingue assistito tramite computer (CALL) è un campo di ricerca e pratica al tempo stesso entusiasmante e frustrante. È entusiasmante perché è complesso, dinamico e in rapida evoluzione, ed è frustrante per le stesse ragioni[18]».

Questa caratterizzazione assume rilevanza ancora maggiore con la recente integrazione dell'intelligenza artificiale nel campo del Language Testing. Se l'evoluzione fin qui descritta ha sfruttato le potenzialità computazionali per automatizzare e arricchire i test l'linguistici, è con l'avvento dell'IA che si apre una frontiera valutativa radicalmente nuova.

1.8 Verso l'Intelligent CALL: l'integrazione dell'intelligenza artificiale e del trattamento automatico del linguaggio

L'introduzione dell'Intelligenza Artificiale nel Language Testing si concretizza principalmente attraverso lo sviluppo del Natural Language Processing (NLP), in italiano Trattamento Automatico del Linguaggio (TAL), un ambito dell'IA che si occupa della modellizzazione e dell'analisi computazionale del linguaggio naturale. La comprensione delle fondamenta tecnologiche del TAL non costituisce un semplice approfondimento tecnico, ma rappresenta un presupposto metodologico essenziale per contestualizzare criticamente le applicazioni ICALL che verranno descritte in seguito. Questa conoscenza fornisce inoltre le basi concettuali necessarie per la comprensione e l'analisi dello strumento che si trova alla base di questo lavoro e che verrà descritto nel Capitolo 3.

Di seguito – affidandoci alla schematizzazione proposta da Hadi et al. (2025) – verrà proposta una ricognizione relativa all'evoluzione interna al TAL, partendo dai primi sistemi simbolici degli anni Sessanta fino ad arrivare agli attuali Large Language Models (LLM). La Figura 6 fornisce una visione d'insieme di questa progressione e guiderà nell'articolazione del paragrafo.

[18] «Computer Assisted Language Learning (CALL) is both exciting and frustrating as a field of research and practice. It is exciting because it is complex, dynamic and quickly changing – and it is frustrating for the same reasons» (Hubbard 2009)

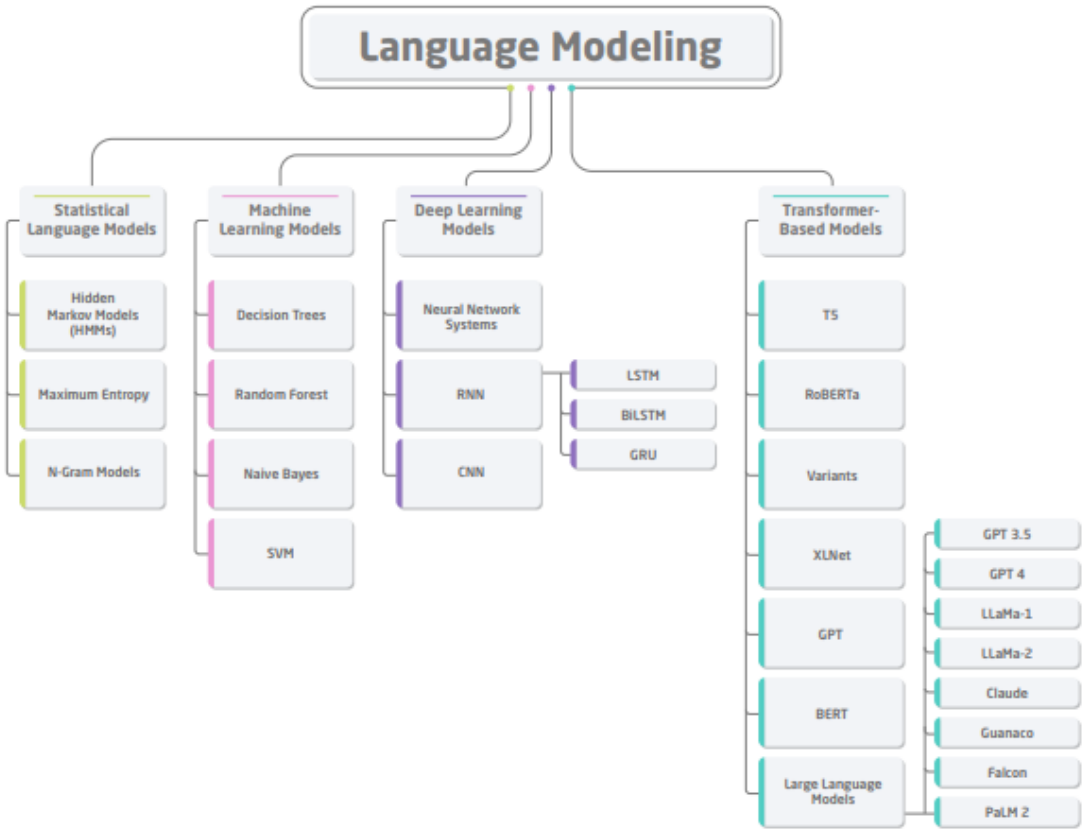


Figura 6: Tassonomia dei modelli di linguaggio computazionali. La progressione evidenzia quattro fasi evolutive: Statistical Language Models (anni '80-'90), Machine Learning Models (anni '90-2010), Deep Learning Models (2010-2017), e Transformer-Based Models (Usman Hadi et al. 2025: 3)

I primi sistemi di TAL affondavano le proprie radici nella tradizione metodologica razionalista e simbolica all'ora dominante tanto in ambito IA quanto nella linguistica generativa. Il paradigma che ne derivava era un approccio epistemologico fondato su metodi logico-deduttivi che intendevano il linguaggio come composto da un insieme di regole che potevano essere interpretate dal computer (Lenci et al 2005). Questa prima fase, definita simbolica o *rule based*, dava origine a modelli costruiti a partire da assiomi codificati manualmente da esperti. Il funzionamento di questi sistemi si basava su grammatiche formali in cui ogni fenomeno linguistico doveva essere esplicitamente descritto attraverso regole logiche. Nonostante l'apporto innovativo per l'epoca, questi sistemi si rivelarono presto inclini a rigidità e a scarsa robustezza. Ogni regola grammaticale doveva essere esplicitamente codificata da professionisti, rendendo i sistemi estremamente fragili di fronte a fenomeni di variazione linguistica, errori ortografici, costruzioni non standard o input devianti. Queste regole, inoltre, dovevano essere costantemente aggiornate e riscritte, il che rendeva il processo estremamente laborioso e costoso. In ultimo, i sistemi non erano in grado di gestire l'ambiguità

linguistica: fenomeni come la polisemia lessicale, l'ambiguità strutturale, i riferimenti anaforici e le interpretazioni pragmatiche rappresentavano elementi di disturbo per il sistema.

A partire dagli anni Ottanta e Novanta si assiste a un progressivo abbandono dell'approccio simbolico in favore di una metodologia basata su modelli statistici e probabilistici. Questa transizione fu scaturita dall'avanzamento della potenza computazionale e dalla disponibilità sempre maggiore di dati linguistici digitalizzati sotto foma di *corpora*. La tradizione di riferimento era quella empirista e *data-driven*, che «si basava sull'uso di metodi statistici per definire modelli probabilistici del linguaggio in termini di generalizzazioni induttive di dati linguistici estratti automaticamente da corpora» (Lenci 2013: 920). Anziché codificare manualmente regole linguistiche, i nuovi sistemi apprendevano pattern e regolarità direttamente dai dati attraverso tecniche di apprendimento automatico (Machine Learning; ML). L'apprendimento automatico è il processo attraverso il quale un modello computazionale impara a eseguire un compito linguistico sulla base di esempi, senza che le regole vengano programmate esplicitamente. I principali paradigmi di ML applicati al TAL sono tre: l'apprendimento automatico supervisionato, non supervisionato e autonomo (*self-supervised*). Nei paradigmi di apprendimento supervisionato, il training del modello avviene attraverso l'esposizione a coppie input-output esplicite: ciascun esempio nei dati di addestramento è accompagnato dall'etichetta corretta (ad esempio frasi annotate con la relativa categoria grammaticale), permettendo all'algoritmo di apprendere la funzione di mappatura che associa caratteristiche dell'input alle categorie o valori target. Una volta addestrato, il modello è in grado di produrre predizioni corrette su input mai incontrati in precedenza. Nell'apprendimento non supervisionato, al contrario, il sistema opera su dati non annotati, individuando autonomamente pattern, raggruppamenti e regolarità intrinseche alla struttura informativa dei dati stessi, senza disporre delle istanze predefinite. L'assenza della fase di annotazione rende l'apprendimento non supervisionato meno dispendioso in termini di tempo e risorse economiche, ma questa efficienza si traduce generalmente in performance inferiori rispetto ai modelli supervisionati (Ježek, Sprugnoli 2023). Il *self-supervised learnin*g si trova in una situazione intermedia tra gli approcci appena descritti. Il sistema genera autonomamente le proprie etichette di supervisione direttamente dai dati non annotati, sfruttando la struttura intrinseca dell'informazione per creare task di previsione (ad esempio la predizione di parole mancanti o future). Tale approccio si è rivelato particolarmente efficace nel trattamento del linguaggio naturale, poiché consente di sfruttare grandi quantità di dati grezzi senza ricorrere a costose annotazioni manuali.

È in questo contesto che emerge il concetto di modello linguistico (Language Model; LM), ovvero un modello computazionale progettato per stimare la probabilità che una determinata parola – o più precisamente token[19] – o una sequenza di token compaia dato il contesto linguistico precedente (Jurafsky, Martin 2026, cap 3). In termini formali, un LM apprende una distribuzione di probabilità sulle sequenze linguistiche. Come anticipato sopra i primi modelli linguistici si fondavano su approcci statistici e probabilisti, tra cui gli *n-gram* e le catene Markoviane. Essendo estremamente complicato dal punto di vista computazionale stimare la probabilità di occorrenza di una parola data la sua intera storia precedente, questi modelli stimavano la probabilità considerando esclusivamente un numero limitato di parole precedenti, definito dalla dimensione dell'*n-gram* (ad esempio bigrammi o trigrammi). Sebbene efficaci per compiti semplici, tali modelli risultavano incapaci di catturare dipendenze sintattiche e semantiche a lungo raggio, rendendoli inadatti a modellare la complessità del linguaggio naturale (Manning, Schutze 1999).

È con l'affermazione del Machine Learning che si assiste alla progressiva introduzione del concetto di apprendimento basato su vettori di caratteristiche (*feature vectors*). Questo paradigma consentiva di trasformare unità linguistiche discrete in *feature* numeriche utilizzabili dagli algoritmi di apprendimento. Su tali rappresentazioni operavano modelli statistici e discriminativi come le *Support Vector Machines* (SVM) che miravano a stimare la probabilità condizionata modellando direttamente il confine tra le classi, e i *Naive Bayes*, modelli generativi che tentavano invece di valutare la probabilità congiunta distinguendo le feature che componevano gli insiemi creando direttamente delle classi. I vettori che venivano realizzati con queste tecniche presentavano limiti strutturali che ne riducevano drasticamente l'efficacia. In primo luogo, venivano trattati come entità discrete e atomiche la cui dimensionalità coincideva con la grandezza del vocabolario stesso. Ciò comportava la creazione di vettori estremamente grandi e quasi interamente composti da zeri (definiti vettori sparsi) che implicavano un notevole spreco di risorse. Secondariamente, non veniva codificata alcuna relazione semantica o sintattica tra le parole; tutti i vettori risultavano equidistanti nello spazio geometrico, indipendentemente dal loro significato.

Per far fronte a queste difficoltà si attuò un cambio di paradigma: si passò da una rappresentazione discreta del significato a una rappresentazione distribuita, attraverso l'introduzione dei *Word Embeddings*. La concettualizzazione di questa nuova tipologia di

[19] «I token sono le unità minime che compongono un corpus, includono le occorrenze delle forme delle parole, la punteggiatura, i numeri, i simboli e le sigle» (Ježek & Sprugnoli 2023).

vettori poneva le proprie radici nell'ipotesi distribuzionale, formulata intorno alla metà degli anni Novanta dal linguista americano Z. Harris (1954) e dal lessicografo britannico J.R. Firth (1957). L'assunto alla base dell'ipotesi è che l'insieme dei contesti in cui una parola ricorre – cioè le parole con cui co-occorre, ovvero la sua distribuzione – sia indicativa del significato della parola stessa e che quindi «esiste una correlazione tra similarità distributiva e similarità di significato, che ci consente di utilizzare la prima per stimare la seconda[20]» (Sahlgren 2008; traduzione mia). La rappresentazione del significato delle entità linguistiche si fonda sull'utilizzo di una metafora spaziale, secondo la quale, il significato di una parola è un luogo in uno spazio geometrico – con un numero variabile di dimensioni – e la similarità semantica tra parole può essere interpretata come vicinanza in questo spazio. Due vettori geometricamente vicini saranno indicativi di due parole con significato simile. La similarità tra vettori viene calcolata attraverso il coseno e il prodotto scalare normalizzato. Gli *embeddings* si classificano quindi come vettori corti (poiché più efficienti dal punto di vista computazionale), densi (in opposizione ai precedenti vettori sparsi) e informativi (in quanto sensibili al contesto di occorrenza. Questa tipologia specifica di *embeddings* viene definita *embedding* contestuale).

L'inizio del XXI secolo segna una svolta metodologica fondamentale nel TAL attraverso l'utilizzo degli *embeddings* e l'affermazione del paradigma connessionista. Questo si fondava sull'utilizzo delle reti neurali artificiali, ovvero sistemi computazionali costituiti da insiemi di unità elementari interconnesse, dette nodi[21]. Ogni nodo riceve segnali in input, li elabora attraverso una funzione matematica e trasmette il risultato ai nodi successivi attraverso connessioni ponderate tramite pesi. A differenza degli approcci algoritmici tradizionali basati su un processamento centralizzato, queste architetture distribuiscono il calcolo attraverso molteplici unità che collaborano tramite le loro connessioni.

A seconda dell'architettura e di come avviene il calcolo le reti neurali si suddividono in diverse famiglie. Le tre principali nel contesto TAL sono: le reti neurali *feedforward*, le reti neurali ricorrenti (Recurrent Neural Networks; RNN) e i *transformers*.

La struttura base delle reti *feedforward* (o *feed-forward neural networks*) – schematizzata in Figura 7 – si articola gerarchicamente in tre strati funzionali: un primo livello riceve i dati da processare (*input layer*), uno o più livelli intermedi effettuano le trasformazioni

[20] «There is a correlation between distributional similarity and meaning similarity, which allows us to utilize the former in order to estimate the latter».
[21] Il concetto di rete neurale si ispira al funzionamento neuronale del cervello.

computazionali attraverso cui il modello apprende rappresentazioni progressive dei dati (*hidden layers*), e un livello finale produce il risultato dell'elaborazione (*output layer*).

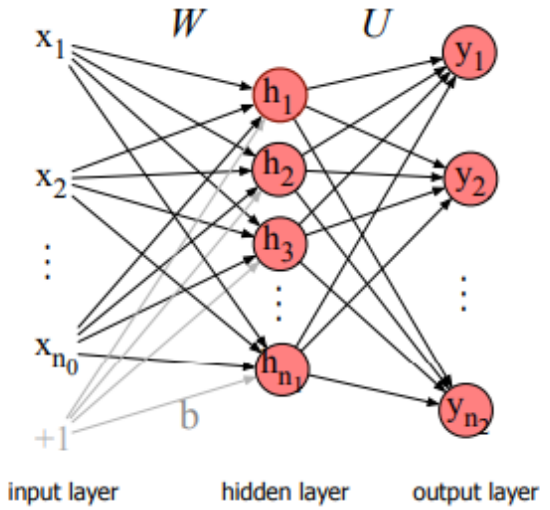


Figura 7: Modello dell'architettura feedforward (Jurafsky, Martin 2026 capitolo 6 pag 8)

In queste architetture il calcolo si propaga esclusivamente in maniera unidirezionale, dallo strato di input verso quello di output, senza retroazioni. Queste sono tipicamente *fully connected* (completamente connesse), dove ogni unità di uno strato trasmette il proprio output a tutte le unità dello strato successivo. Le reti *feedforward* sono efficaci per compiti di classificazione o regressione ma, poiché prive di meccanismi di memoria, risultano non particolarmente adatte per la gestione del linguaggio naturale, che è intrinsecamente sequenziale. Durante la fase di addestramento, la rete apprende attraverso un processo iterativo: i dati di training vengono presentati alla rete, che produce un output; l'errore tra l'output prodotto e quello desiderato viene calcolato e "propagato all'indietro" attraverso la rete tramite l'algoritmo di *backpropagation*, aggiornando i pesi delle connessioni in modo da ridurre progressivamente l'errore. Questo processo, ripetuto su milioni di esempi, permette alla rete di apprendere autonomamente le rappresentazioni ottimali per il compito assegnato (Rumelhart et al., 1986). È in questo contesto che si inizia a parlare di *Deep learning*, ovvero l'impiego di reti neurali profonde, definite in questo modo per la presenza di numerosi strati nascosti, che permettono al modello di apprendere rappresentazioni gerarchiche e astratte dei dati linguistici.

Le *Recurrent Neural Networks* (RNN) introducono una caratteristica fondamentale per il trattamento del linguaggio: la capacità di memoria temporale. Il meccanismo distintivo consiste nel fatto che l'output prodotto in ciascun nodo viene duplicato e reintrodotto nella rete

come input aggiuntivo per il passo temporale successivo. In questo modo, quando la rete processa un nuovo elemento della sequenza, considera simultaneamente l'input corrente e lo stato interno che codifica informazioni sugli elementi precedentemente elaborati. Questa architettura ricorrente permette alla rete di mantenere una "memoria" della storia processuale: ogni computazione dipende non solo dall'input presente ma anche dagli stati precedenti, consentendo di catturare dipendenze e pattern che si sviluppano lungo la sequenza temporale (Elman 1990). Grazie a questi stati interni ricorrenti, le RNN possono modellare relazioni tra elementi distanti nella sequenza, rendendo possibile l'elaborazione di linguaggio naturale dove il significato spesso dipende dal contesto precedente. La Figura 7 mostra il funzionamento di una rete RNN.

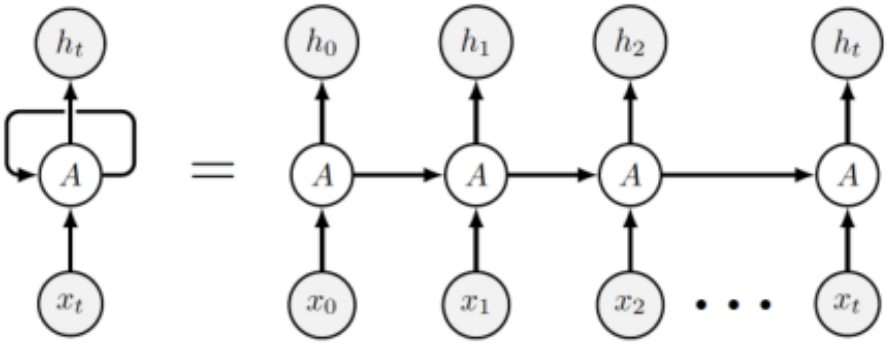


Figura 8: Modello dell'architettura RNN (Ježek, Sprugnoli 2023: 101)

Varianti più sofisticate come le *Long Short-Term Memory* (LSTM) e le *Gated Recurrent Units* (GRU) furono sviluppate per migliorare la capacità di memoria delle RNN attraverso meccanismi di cancelli – detti *gates* – che regolano selettivamente quali informazioni mantenere o dimenticare (Hochreiter & Schmidhuber 1997). Nonostante l'introduzione della dimensione temporale nella modellizzazione del linguaggio, le reti neurali ricorrenti presentano limiti strutturali rilevanti. La loro natura sequenziale limita la parallelizzazione del calcolo e ne riduce la scalabilità, rendendo il conteggio su dataset ampi estremamente dispendioso in termini di tempo e costi.

Tali criticità hanno motivato lo sviluppo di tecnologie alternative, culminate nell'introduzione dei Transformer, teorizzati da Vaswani et al. nel celebre articolo *Attention Is All You Need* (2017). Questo lavoro ha trasformato radicalmente il campo del TAL e costituisce la base architettonica su cui sono costruiti tutti i moderni modelli del linguaggio. L'architettura Transformer (Figura 9) si compone di due macro-componenti: *l'encoder*, cioè la componente

di codifica, ed il *decoder*, ovvero la componente di decodifica. L'encoder è composto da una serie di strati che elaborano la sequenza di input e ne costruiscono una rappresentazione contestualizzata. Ogni parola viene trasformata in un token che codifica il suo significato lessicale, il ruolo sintattico e le sue relazioni con tutte le altre parole della sequenza. Il decoder invece, a sua volta composto da una serie di strati, utilizza le rappresentazioni prodotte dall'encoder per generare la sequenza di output.

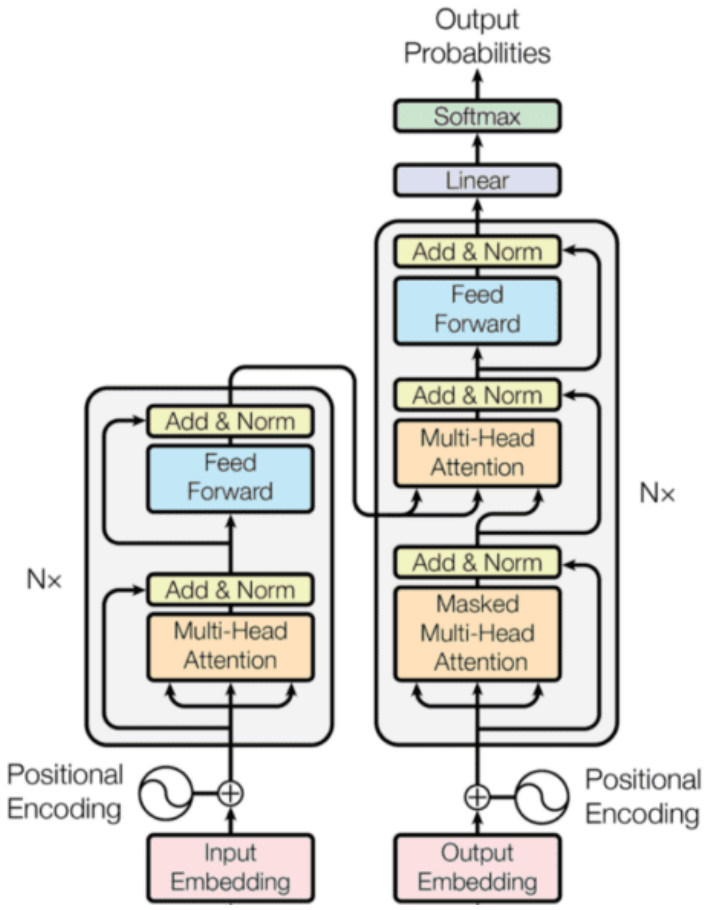


Figura 9: Architettura del modello Transformer (Vaswani et al 2017:3)

L'innovazione fondamentale che viene aggiunta nei transformer è il meccanismo dell'attenzione (definito *self-attention*), ovvero, un'operazione matematica che assegna un peso ad ogni token nella sequenza, basandosi sulle relazioni semantiche tra i token stessi. Ciò gli permette di ponderare dinamicamente l'importanza relativa dei diversi elementi della sequenza durante l'elaborazione di ciascun token. Tale procedura si basa sul principio secondo il quale, in una frase, non tutte le parole concorrono in egual misura a definire il significato delle altre. Per aumentare le capacità del modello, nei Transformer viene introdotta la *multi-head attention*,

che replica il meccanismo di attenzione più volte in parallelo. Specifico della struttura del decoder è il meccanismo di *self-attention* mascherata, che impedisce l'accesso ai token futuri durante la generazione dell'output, e un meccanismo di attenzione incrociata (*encoder–decoder attention*), che permette al decoder di integrare le rappresentazioni prodotte dall'encoder.

Originalmente l'architettura transformer comprendeva sia la componente di encoder che quella di decoder; tuttavia, si sono successivamente affermate tre categorie distinte di modelli (Minaee et al. 2025): *l'encoder-only*, come per esempio BERT (Bidirectional Encoder Representations from Transformers; Devlin et al 2019) specializzato in compiti di comprensione; il *decoder-only*, come la famiglia GPT (Generative Pre-trained Transformer; Radford et al 2018; Radford et al 2019; Brown et al 2020), abili in compiti di generazione; e chi mantiene l'architettura *encoder-decoder*, come T5 (Text-to-Text Transfer Transformer; Raffel et al. 2019), distintiva per i compiti di trasformazione *sequence-to-sequence*.

Le innovazioni tecnologiche che sono state inserite nelle architetture *transformer* hanno consentito di addestrare modelli di linguaggio di dimensioni senza precedenti, addestrati su una quantità di dati mai vista prima attraverso pratiche di *self-supervising*. Questo ha portato all'emergere dei *Large Language Models* (LLM) che hanno rivoluzionato il campo dell'elaborazione del linguaggio naturale. A differenza dei sistemi precedenti, progettati e addestrati per compiti specifici (come la traduzione, la classificazione o la generazione), gli LLM acquisiscono competenze linguistiche generaliste che possono essere applicate a molteplici attività di TAL senza necessità di riaddestramenti costosi (Jurafsky, Martin 2026 cap.7). Questo è reso possibile dall'introduzione delle rivoluzionarie procedure di *pre-train* e *fine-tune* (rappresentate in Figura 10).

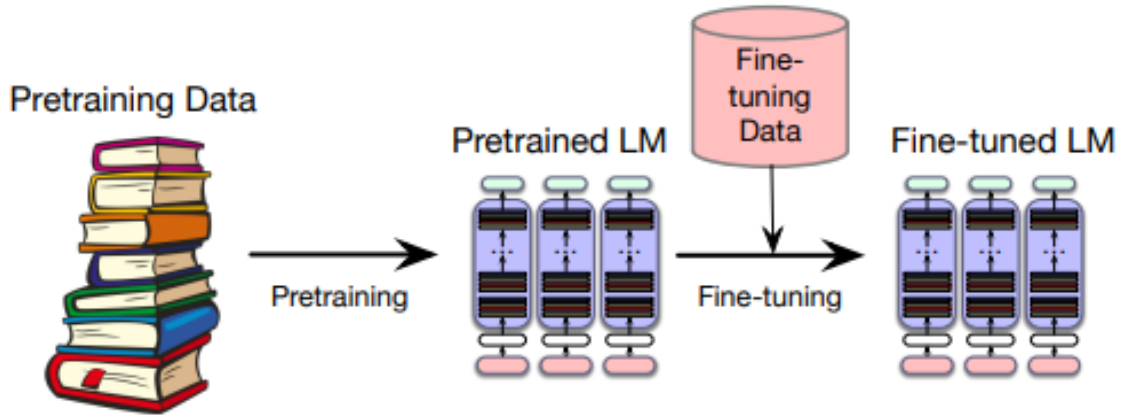


Figura 10: Paradigma pre-train - fine-tune (Jurafsky e Martin cap 7 pag 19)

Grazie paradigma di *pre-training* il modello apprende rappresentazioni linguistiche generali attraverso task *self-supervised* su corpora di vastissime dimensioni; una volta pre-addestrato il modello è in grado di specializzarsi su compiti specifici tramite un ulteriore addestramento supervisionato su dataset annotati più piccoli e task-specifici. Tale procedura permette un trasferimento di conoscenza – definito *transfer learning* – che consente di ridurre in maniera esponenziale i costi e le tempistiche di addestramento. Il paradigma *pre-train - fine-tune* ha rivoluzionato il campo del trattamento automatico del linguaggio e rappresenta oggi lo stato dell'arte della disciplina.

Il percorso evolutivo delineato in questo paragrafo non evidenzia esclusivamente un avanzamento sul piano tecnico, ma segnala un mutamento profondo di natura epistemologica nell'ambito del trattamento automatico del linguaggio. Il passaggio da sistemi rigidi basati su regole a favore di modelli capaci di rappresentare il significato in modo distribuito e di modellare relazioni contestuali estese ha contribuito a ridurre la distanza tra le formalizzazioni computazionali e la dinamicità propria del linguaggio naturale. Queste caratteristiche risultano particolarmente significative nel contesto del Language Testing, dove la misurazione della competenza linguistica richiede strumenti in grado di affrontare la variabilità intrinseca nell'uso autentico della lingua. L'impiego di tecnologie basate sul TAL si configura dunque non come una semplice innovazione tecnica, ma come un cambiamento metodologico che apre nuove possibilità per la progettazione, l'analisi e la validazione delle prove linguistiche.

1.9 Intelligent Computer-Assisted Language Learning

L'integrazione sistematica di concetti, tecniche, algoritmi e tecnologie proprie dell'intelligenza artificiale nel campo della valutazione linguistica ha dato origine all'*Intelligent Computer Assisted Language Learning.* Se nei sistemi CALT descritti nei paragrafi precedenti la tecnologia veniva utilizzata essenzialmente per automatizzare, standardizzare, arricchire multimedialmente e – sebbene in maniera superficiale – valutare i test linguistici, è con l'avvento dell'ICALL che quest'ultima interviene in maniera trasversale e sistemica lungo l'intero ciclo di vita del test. L'introduzione dell'IA, degli LLM e l'impiego di tecnologie vocali ha portato ad un cambio di paradigma nella concezione stessa dell'*assessment*.

L'interesse crescente per l'utilizzo di queste nuove tecnologie nel mondo della valutazione è testimoniato dalla nascita e dal consolidamento di organizzazioni dedicate. Una

delle più rilevanti in ambito internazionale è il Computer Assisted Language Instruction Consortium (CALICO)[22], fondato nel 1983 e significativamente rivitalizzato negli ultimi anni. L'obiettivo di CALICO è quello di favorire la collaborazione tra esperti appartenenti ai settori della linguistica, della didattica delle lingue e dell'informatica al fine di favorire la ricerca e lo sviluppo di strumenti didattici e valutativi in ambito L2 basati sull'IA e sul TAL.

Oltre al punto del processo di creazione del test in cui la tecnologia si inserisce, nell'ICALL emerge una differenza fondamentale che riguarda il ruolo epistemologico e funzionale assegnato al computer all'interno del processo valutativo. In questa direzione – dapprima teorizzata nei confronti del CALT e successivamente applicata all'ICALL – Taylor 1980 propone una distinzione concettuale tra l'utilizzo del *computer as a tutor* (computer come un tutor) e *computer as a tool* (computer come uno strumento). Nel primo caso il sistema informatico svolge la funzione attiva di guida nell'apprendimento, emulando le funzioni tradizionalmente eseguite dall'insegnante. Il computer, sulla base del livello dell'utente, propone dei materiali e delle attività e monitora le sue risposte, decidendo sulla base di queste quale compito proporre in seguito e fornendo suggerimenti durante l'intero processo. Al termine del percorso, il sistema valuta il livello di competenza mostrato dall'utente e fornisce esercizi mirati per l'automiglioramento e un *feedback* correttivo personalizzato. Una delle realizzazioni più concrete del *computer as a tutor* per l'apprendimento linguistico sono i sistemi di Tutoring intelligente (Intelligent Tutoring Systems, ITS), software educativi che combinando la conoscenza dell'allievo (definita modello dello studente), la conoscenza del dominio (detta modello esperto) e la conoscenza delle strategie di insegnamento (denominato modello pedagogico) forniscono percorsi di apprendimento e valutazione individualizzati (Katinskaia 2025). Proprio questa personalizzazione del processo di apprendimento e accompagnamento costante rappresenta il vantaggio principale del modello *computer as a tutor*. Tale vantaggio permette di creare percorsi didattici individuali per ogni studente, procedura che risulterebbe estremamente complessa e dispendiosa in termini di tempo in un ambiente didattico tradizionale. Non mancano tuttavia delle criticità: i sistemi incorporano necessariamente delle gerarchie di priorità didattiche e modelli valutativi che potrebbero non essere condivisi da tutti i docenti o appropriati a tutti i contesti educativi. La metafora del tutor rischia inoltre di creare aspettative eccessive riguardo le capacità di comprensione del sistema, mascherando i limiti intrinseci di queste tecnologie.

[22] https://calico.org/

Per quando riguarda invece il paradigma del *computer as a tool*, il sistema informatico riveste un ruolo ausiliare e di supporto alla didattica tradizionale. Il pieno controllo sul processo di insegnamento resta nelle mani del docente che sfrutta le potenzialità della tecnologia per migliorare e facilitare l'acquisizione. Il paradigma *tool* mantiene pertanto l'*agency* umana al centro del processo valutativo e decisionale.

La distinzione concettuale dell'utilizzo del computer come tutor e del computer come strumento non dovrebbe essere intesa come una dicotomia rigida ma piuttosto come un continuum entro il quale le due modalità convergono. Le applicazioni ICALL ad oggi più promettenti tendono infatti a coniugare pratiche automatizzate con la supervisione umana, in particolare negli ambiti valutativi ad alto impatto.

Come accennato in precedenza, nei sistemi ICALL la tecnologia, e in particolare l'intelligenza artificiale, rivestono un ruolo capillare in ogni punto del processo di creazione e valutazione di un test. Vediamo di seguito gli utilizzi più comuni che ne vengono fatti.

In primis, durante la fase di creazione e progettazione dei test, l'IA trova applicazione nei processi di *Automatic Item Generation* (AIG), permettendo di generare automaticamente le domande che compongono i test. Queste si possono differenziare a seconda del metodo utilizzato per la generazione che può fare affidamento su template, regole o metodi statistici (Kurdi et al., 2020). Se i metodi basati su template e regole richiedono comunque l'elaborazione di quest'ultime da parte dell'esperto umano, quelli che utilizzano i metodi statistici deducono implicitamente le regole per la generazione delle domande. Numerosi studi hanno analizzato nel corso del tempo la possibilità di realizzare item a risposta chiusa come le domande a risposta multipla (Donati et al 2024). È solo a partire dagli ultimi anni che l'IA viene utilizzata anche per creare domande più complesse – come esercizi di comprensione interattiva – per la generazione di testi calibrati su specifici livelli di difficoltà o scritti con un linguaggio specialistico, per la creazione di immagini specifiche o per la realizzazione di distrattori. Queste implementazioni hanno ridotto notevolmente i tempi e i costi di progettazione degli item e dei test stessi, portando ad un progressivo ampliamento delle banche di item disponibili.

In secondo luogo, durante la fase di somministrazione, i sistemi ICALL possono sfruttare i già citati sistemi adattivi, che aiutano a rendere personalizzata l'esperienza di test. Attraverso l'analisi in tempo reale di variabili come i tempi di risposta, la frequenza e la tipologia degli errori, l'IA rende possibile il monitoraggio in tempo reale delle performance del candidato. Ne consegue la possibilità di eliminare i sentimenti di frustrazione derivanti dal confronto con compiti troppo complessi e, allo stesso tempo, la mancanza di stimoli causata da

task eccessivamente semplici (Baker & Siemens 2014; Chen et al. 2021). L'esperienza che ne deriva favorisce un approccio educativo su misura, che consente agli studenti di apprendere e avanzare al proprio ritmo.

Nei sistemi ICALL l'intelligenza artificiale riveste un ruolo centrale soprattutto nell'ambito della correzione e valutazione automatizzata delle prestazioni, in particolare attraverso l'utilizzo di tecnologie di *Automated Essay Scoring* (AES) (Yavuz et al 2025) e di *Automated Speaking Assessment* (ASA) (Zechner, Evanini 2019). I primi modelli di AES, sviluppati tra la fine degli anni Novanta e l'inizio degli anni Duemila, si fondavano sull'identificazione di *features* linguistiche ritenute indicative per garantire la bontà di un testo e sullo sviluppo di algoritmi per l'estrazione automatica di tali caratteristiche. Queste riguardavano l'utilizzo di metriche informative per quanto riguarda il numero di parole, la lunghezza media delle frasi, la ricchezza lessicale, la complessità sintattica, la densità di connettivi e marcatori e così via. I sistemi AES più recenti sfruttano invece architetture di deep learning che apprendono automaticamente rappresentazioni distribuite dei testi senza necessità di feature engineering esplicito. Questi sistemi rappresentano un cambio paradigmatico: anziché identificare a priori quali caratteristiche testuali siano rilevanti, l'algoritmo apprende implicitamente pattern complessi direttamente dai dati. Nello specifico, vengono sfruttate le potenzialità offerte dai *transformer* ai quali vengono forniti i testi da valutare e informazioni specifiche – sotto forma di *prompt* – che riguardano i criteri di scoring e le scale valutative. Il modello genera una valutazione sotto forma di punteggio, accompagnata da una giustificazione testuale. Numerosi studi hanno descritto i vantaggi che si possono trarre dall'utilizzo di queste tecnologie: in primo luogo è possibile ottenere un'efficienza elevata data dall'immediatezza della correzione, si può inoltre riscontrare un'ampia concordanza con i giudizi di valutatori umani esperti (Bridgeman et al 2012), aumentare l'affidabilità e la standardizzazione delle valutazioni ed eliminare le variabili soggettive dovute ai valutatori umani. Non mancano tuttavia criticità legate a questi strumenti. I sistemi AES faticano a cogliere la profondità concettuale e dimostrano una comprensione semantica e pragmatica limitata. Sono inoltre vulnerabili a strategie di manipolazione: è infatti possibile ingannare l'algoritmo introducendo nel testo elementi lessicali ricercati ed elaborate costruzioni sintattiche, anche se privi di reale contenuto informativo. Particolarmente rilevanti sono le questioni legate alla possibile presenza di bias derivanti dai dati su cui i modelli vengono addestrati, che possono portare alla penalizzazione di studenti non madrelingua o di stili di scrittura non conformi ai modelli dominanti. Infine, dal

punto di vista pedagogico, il *feedback* fornito potrebbe rivelarsi non rilevante o utile per l'apprendimento e poco personalizzato.

Le tecnologie di *Automated Speaking Assessment* adottano invece architetture modulari che integrano componenti specializzate: è presente un modulo di *Automatic Speech Recognition* (ASR), un modulo di analisi acustica e prosodica, un modulo di analisi linguistica e un modulo legato alla valutazione. Dopo aver effettuato una trascrizione del segnale acustico in oggetto l'audio viene analizzato per estrarre feature prosodiche e di pronuncia che non sono catturate dal testo trascritto, per esempio la velocità di eloquio, la frequenza e la durata delle pause, il ritmo e così via. Successivamente, il testo trascritto dall'ASR viene processato per estrarre feature linguistiche come nei modelli di AES ed infine viene fornita una valutazione. I modelli ASA si rivelano particolarmente utili nei contesti di apprendimento e valutazione ad alta richiesta. Consentono infatti di ottenere una valutazione rapida e scalabile ma allo stesso tempo coerente e standardizzata, riducendo i costi e i tempi di correzione. Infine, la grande novità introdotta da queste tecnologie risiede nella possibilità di svolgere l'esame completamente da remoto, non necessitando della presenza dell'esaminatore umano come valutatore. Pur offrendo numerosi vantaggi i modelli ASA nascondono criticità. Prima fra tutte, i principali modelli di ASR sono addestrati prevalentemente su parlanti nativi e mostrano risultati meno eccellenti sul parlato L2. È possibile, inoltre, che a causa della presenza di *bias* derivanti dai dati si verifichino discriminazioni sulla base della L1, del genere e dell'accento del parlante, dell'utilizzo di una varietà non standard o per l'appartenenza a categorie sottorappresentate nei dati di training (Periyasamy et al. 2025). Infine, la qualità della registrazione incide drasticamente sulle performance di ASR. Rumore ambientale, riverbero, caratteristiche del microfono e la compressione dell'audio possono comportare un'errata trascrizione con un conseguente errore nella valutazione. Un'analisi più approfondita sul funzionamento, sui punti di forza e sui limiti dei sistemi AES e ASA sarà effettuata nel Capitolo 3.

Un ulteriore aspetto innovativo introdotta dall'AI nei modelli ICALL riguarda l'impiego di agenti conversazionali e chatbot. Questi, attraverso la possibilità costante di intavolare conversazioni e di ricevere *feedback* correttivi, offrono un'opportunità di pratica simile a quella immersiva. L'accesso massiccio ad un input simile a quello presente in un ambiente di apprendimento reale favorisce il miglioramento della competenza linguistica (Dokukina, Gumanova 2020; Huang et al. 2023). Numerosi studiosi hanno inoltre evidenziato come l'interazione con i sistemi di dialogo possa contribuire notevolmente ad abbassare il filtro affettivo descritto da Krashen (Al-Obaydi et al. 2023). Le dinamiche di gamification e role-play

riducono l'ansia che insorgerebbe nel rapporto con il valutatore umano e rendono l'apprendente più propenso a esporsi linguisticamente, eliminando il timore del giudizio dell'insegnante (Dhimolea et al. 2022).

Infine, in particolare nei contesti di somministrazione dei test a distanza, l'IA svolge un ruolo importante inerentemente alla sicurezza e all'integrità dei test. Sono stati infatti sviluppati algoritmi che permettono di attuare il monitoraggio biometrico dei candidati tramite il riconoscimento facciale, l'analisi dei pattern di digitazione e l'eyes track (Hylton et al. 2016; Suvorov 2024).

Nonostante il crescente livello di autonomia dei sistemi ICALL basati sull'intelligenza artificiale, un numero sempre maggiori di studi sottolinea l'importanza di mantenere attiva la presenza dell'intervento umano nei processi automatizzati (Wu et al 2022; Mosqueira-Rey et al. 2022). Quello che viene proposto è un paradigma definito *Human In-The-Loop* (HITL), in cui l'intervento umano è integrato in modo esplicito e continuo all'interno del ciclo di funzionamento del sistema. L'essere umano è coinvolto nella fase di raccolta e annotazione dei dati, nella definizione dei criteri di valutazione, nell'addestramento e calibrazione dei modelli ed infine nella validazione dei risultati prodotti. Questa procedura consente di monitorare costantemente le prestazioni, correggere eventuali errori, attenuare la presenza di bias e migliorare l'affidabilità generale.

In ambito ICALL la presenza umana è ravvisabile in diversi punti del processo di vita del test. Nella fase di progettazione dell'esame e di generazione degli item il contributo umano risulta essenziale per garantire la validità del costrutto[23], l'adeguatezza dei task ai livelli di competenza e la pertinenza rispetto agli obiettivi dell'apprendimento (Attali et al. 2022). Nei processi di correzione automatica e scoring il coinvolgimento umano è essenziale per supervisionare e garantire l'affidabilità e la stabilità delle valutazioni. Questo si dimostra particolarmente utile nel caso di esami ad alta posta in gioco o in situazioni di valutazioni ambigue. Per quanto riguarda la supervisione e interpretazione del feedback generato automaticamente l'intervento umano risulta fondamentale per garantire che il riscontro sia pedagogicamente appropriato, comprensibile e utile dal punto di vista formativo.

L'analisi condotta in questo capito ha inteso delineare una duplice traiettoria evolutiva: da un lato i mutamenti intervenuti nella concettualizzazione della competenza linguistica, dall'altro i progressi registrati in ambito tecnologico. Parallelamente, si è cercato di mostrare

[23] McNamara (2000: 13) definisce il costrutto come: « those aspects of knowledge or skill possessed by the candidate which are being measured».

come queste due dimensioni siano profondamente interconnesse e come il progredire dell'una abbia costantemente influenzato le riflessioni teoriche e le scelte metodologiche relative all'altra.

A partire dalle divergenze che riguardano le discipline della Second Language Acquisition e del Language Testing si è cercato di evidenziare come una maggiore sinergia tra i due ambiti possa risultare reciprocamente vantaggiosa, contribuendo al rafforzamento tanto dei modelli teorici quanto degli strumenti operativi. Gli studiosi di Language Testing possono trovare nella ricerca sull'acquisizione linguistica un contributo fondamentale per chiarire quali aspetti della lingua siano effettivamente valutabili e come questi si distribuiscano lungo i diversi livelli di competenza. Allo stesso tempo, la letteratura sul testing offre alla ricerca sull'acquisizione un quadro metodologico rigoroso, sottolineando la necessità di definizioni operative chiare e di strumenti di elicitazione e misurazione affidabili. La successiva disamina delle teorie sulla competenza linguistico-comunicativa ha messo in luce come il concetto stesso di competenza, così come le modalità attraverso cui essa viene elicitata e valutata, abbia subito nel tempo trasformazioni significative, in linea con i paradigmi dominanti nella linguistica dei diversi periodi. È stato inoltre evidenziato il ruolo centrale svolto dalle istituzioni europee e, in particolare, dal QCER, sia nella definizione della competenza linguistico-comunicativa sia nella standardizzazione di processi condivisi. Infine, la disamina dell'eterogeneità dei pubblici e delle motivazioni che spingono gli utenti ad apprendere l'italiano come lingua seconda ha evidenziato come questa competenza assuma per ciascuno valori e funzioni differenti a seconda dei contesti e degli obiettivi degli apprendenti.

La seconda traiettoria evolutiva ha riguardato gli sviluppi tecnologici e il loro impatto sulle pratiche valutative nel Language Testing. In una fase iniziale, dominata dai test *paper and pencil* e successivamente dai *computer-administered conventional tests*, la tecnologia viene impiegata prevalentemente a fini di efficientamento procedurale. La digitalizzazione delle prove cartacee ha consentito una somministrazione più rapida e una correzione automatizzata, ma non ha comportato una ristrutturazione sostanziale dei formati di test né dei modelli teorici sottostanti, promuovendo una visione composizionale della competenza basata su unità discrete. Con l'avvento dei personal computer e del *Computer-Assisted Language Testing*, l'integrazione di elementi multimediali ha consentito un primo superamento della frammentazione delle abilità linguistiche e ha favorito una concezione della competenza più comunicativamente orientata e una maggiore attenzione all'autenticità dei compiti. La successiva diffusione delle tecnologie *web-based* ha ulteriormente ampliato queste possibilità,

eliminando i vincoli spazio-temporali, garantendo maggiore flessibilità e scalabilità. Un ulteriore cambiamento paradigmatico è stato introdotto con lo sviluppo del Trattamento Automatico del Linguaggio e delle evoluzioni interne a questa disciplina. Il passaggio dai sistemi simbolici basati su regole, espressione di una concezione atomistica delle unità linguistiche, ai modelli statistici, fino al paradigma connessionista delle reti neurali e alle architetture transformer, ha determinato una trasformazione profonda nella modellizzazione del linguaggio. In questo quadro – attraverso l'utilizzo di una metafora spaziale – il significato linguistico viene rappresentato come una posizione in uno spazio geometrico multidimensionale, in cui la similarità può essere interpretata come vicinanza o lontananza in questo spazio. Tali innovazioni hanno aperto le porte *all'Intelligent Computer-Assisted Language Learning*, in cui la tecnologia interviene in maniera sistematica e trasversale in tutte le fasi del processo di creazione del test. È possibile ravvisarla nelle fasi di generazione degli item, nella selezione di quest'ultimi attraverso i sistemi di somministrazione adattiva, nella simulazione di contesti immersivi tramite *chatbot* e *Intelligent Tutoring Systems*, nella correzione automatizzata delle prove e nella creazione di feedback personalizzati. Attraverso queste innovazioni l'*assessment* linguistico può così indagare la competenza in tutte le sue sfumature, avvicinandosi in modo inedito alla complessità dell'uso linguistico autentico. È importante sottolineare come in questo processo la supervisione umana rimanga centrale per garantire validità, equità, trasparenza e affidabilità.

L'analisi delineata in questo capitolo ha portato alla luce il fatto che i limiti più significativi dell'ICALL non siano di natura computazionale, bensì epistemologica. Essi riguardano la nostra comprensione ancora parziale e dibattuta di che cosa costituisca la competenza linguistica, dei meccanismi profondi che guidano l'acquisizione linguistica, di quali dimensioni della performance linguistica siano rilevanti in contesti d'uso autentici, e di come queste dimensioni possano essere validamente ed equamente misurate. Come Alderson osservava lucidamente già nel 1990, in una riflessione che mantiene piena attualità nonostante i progressi tecnologici degli ultimi tre decenni:

> «Le possibilità di diagnosi e correzione sollevano un problema importante per i linguisti applicati e gli insegnanti di lingue. Il limite allo sviluppo di tali test non risiede nella capacità dell'hardware o nella complessità del compito di programmazione, ma nella nostra inadeguata comprensione della natura dell'apprendimento e dell'uso linguistico … la sfida del [ICALT] è più per il linguista e per il linguista applicato nel fornire un input appropriato sulla natura delle

routine di ramificazione e sui suggerimenti, gli indizi e il feedback che potrebbero aiutare gli studenti, che per il programmatore di computer nel produrre un software adeguato»[24].

[24] Possibilities for diagnosis and remediation raise an important problem for applied linguists and language teachers. The limitation on the development of such tests is not the capacity of the hardware, or the complexity of the programming task, but our inadequate understanding of the nature of language learning and of language use … the challenge of [CALT] is more to the linguist and applied linguist to provide appropriate input on the nature of branching routines, and on the hints, clues and feedback that would help learners, than to the computer programmer to produce adequate software. (Alderson1990: 25)

CAPITOLO 2. IL LANGUAGE TESTING E LE CARATTERISTICHE ESSENZIALI DI UN TEST

Il presente capitolo definisce il quadro teorico e metodologico che precede la somministrazione pilota del test volto alla valutazione della competenza linguistico-comunicativa di apprendenti di italiano L2. In primo luogo, verrà esplicitato il ruolo sociale storicamente svolto dalla valutazione, intesa come dispositivo attraverso cui le società verificano e legittimano il possesso di determinate qualità o competenze. In questa prospettiva, si evidenzierà come, in specifici contesti educativi e istituzionali, il Language Testing abbia assunto – e continui in parte ad assumere – una funzione di *gatekeeping* sociale, contribuendo alla costruzione di gerarchie che incidono non solo sulle rappresentazioni della lingua e della competenza, ma anche sui percorsi e sulle opportunità degli individui. Proprio in ragione di tale dimensione sociale e delle sue implicazioni etiche, i *test developer* sono chiamati ad assumersi la responsabilità di progettare strumenti teoricamente fondati e metodologicamente solidi. A tal fine, verrà chiarito come esami e prove di valutazione non costituiscano misure dirette della competenza sottesa, bensì inferenze basate su prestazioni osservabili in compiti specifici; la loro attendibilità e legittimità dipende pertanto dalla qualità dello strumento attraverso cui tali prestazioni vengono elicitate e interpretate. La qualità di un test verrà quindi ricondotta al concetto di utilità complessiva, intesa come costrutto multidimensionale derivante dal bilanciamento dei parametri di validità di costrutto: affidabilità, autenticità, interattività, impatto e praticabilità. L'analisi di queste dimensioni consentirà di esplicitare i criteri attraverso cui uno strumento valutativo può essere giudicato adeguato rispetto agli scopi per cui è stato progettato e alle decisioni che supporta. Su tali basi teoriche, il capitolo descriverà successivamente le principali fasi di costruzione di un test: dalla definizione del costrutto alla stesura delle specifiche (progettazione), dalla scrittura e revisione degli item alla sperimentazione pilota e all'analisi dei *feedback* (operazionalizzazione), fino alle procedure di somministrazione e raccolta dati (somministrazione).

2.1 Fondamenti teorici del Language Testing

Nell'opera *Language Testing: The Social Dimension* (2006) Tim McNamara e Carsten Roever analizzano la funzione sociale della valutazione, evidenziando come questa abbia da sempre rivestito un ruolo fondamentale nelle comunità, rivelandosi una costante antropologica per stabilire, attraverso procedure pubbliche e verificabili, chi possiede determinate qualità o competenze. Nello specifico, i test linguistici costituiscono strumenti per la raccolta di informazioni finalizzate a prendere decisioni che producono conseguenze concrete – talvolta determinanti – per apprendenti, programmi educativi, istituzioni e società nel loro complesso. Essi fungono da dispositivi di regolazione dell'accesso a opportunità educative, professionali e sociali, determinando chi può iscriversi a una determinata università, chi può ottenere la cittadinanza, chi è qualificato per una posizione lavorativa. In questo senso, i test operano come meccanismi di *gatekeeping* sociale, esercitando un potere che va ben oltre la mera misurazione di competenze individuali: essi contribuiscono a definire standard linguistici, influenzano curricula e pratiche didattiche, producono gerarchie tra varietà linguistiche e, in ultima analisi, tra parlanti (Bachman, Palmer 2010).

Durante la storia, i test di lingua sono stati utilizzati come strumenti per attuare una sorta di controllo delle frontiere per resistere al movimento dei popoli. McNamara (2004) propone un excursus di questo fenomeno: partendo dal caso dello *Shibboleth* nella Bibbia (Giudici 12: 4-6), in cui i soldati sconfitti che tentavano di farsi passare per appartenenti al gruppo etnico vincitore venivano identificati ed eliminati sulla base di una minima differenza fonematica, ossia l'incapacità di pronunciare correttamente il fonema /ʃ/ presente nella parola "shibboleth"; passando per il caso della Germania riunificata degli anni Novanta, in cui i test linguistici venivano utilizzati per moderare le richieste di riconoscimento avanzate da gruppi minoritari provenienti dall'ex blocco orientale. In questo contesto il test, sotto forma di colloquio orale, mirava a identificare tratti linguistici non standard, considerati indicativi dell'appartenenza al gruppo minoritario. Tuttavia, questo criterio penalizzò coloro i quali avevano maggiormente assimilato tratti linguistici caratteristici dei paesi di residenza, comportando per queste persone la negazione al diritto di accesso; in ultimo, alcuni paesi tra cui l'Australia, utilizzano i test linguistici per operare una selezione tra i richiedenti asilo: appellandosi al diritto internazionale – per il quale uno stato non è obbligato a concedere lo status di rifugiato a una persona che non provenga direttamente dal paese d'origine ma da un paese terzo considerato sicuro – le autorità australiane ricorrono a interviste orali nella lingua madre dei richiedenti asilo, che vengono

registrate e successivamente sottoposte a un'analisi linguistica specialistica. L'obiettivo è individuare tratti derivanti dal contatto linguistico che possano indicare una permanenza in un paese di transito sicuro piuttosto che nella zona originaria di persecuzione. Anche in questo caso, la presenza di tali tratti implicherebbe il mancato diritto allo status di rifugiato. Questi esempi evidenziano come i test linguistici possano operare non solo come strumenti neutrali di misurazione della competenza, ma anche come tecnologie di potere che determinano chi possa attraversare confini, accedere a diritti, rivendicare appartenenze. In questi casi estremi i test linguistici si configurano come pratiche di *linguistic profiling* con conseguenze che possono essere drammatiche per gli individui coinvolti.

In netta contrapposizione agli usi storici e contemporanei del Language Testing (LT) come strumento di controllo ed esclusione si colloca la visione promossa dalle politiche linguistiche europee e, in particolare, dal Quadro Comune Europeo di Riferimento per le Lingue. Come discusso nel capitolo precedente (cfr. paragrafi 1.4 e 1.5), la genesi stessa del documento è legata alla volontà di facilitare la mobilità delle persone all'interno dello spazio europeo, favorendo e incentivando l'incontro interculturale e interlinguistico, definendo criteri per la comparabilità e il riconoscimento delle competenze linguistiche.

Sebbene rispetto al passato il mancato superamento di una situazione di test possa comportare conseguenze meno immediate e nefaste – tranne, forse, per le persone richiedenti asilo – si tratta comunque di situazioni ad alto rischio per chi li sostiene, con il potere di definire la propria identità. Come afferma Ferrari:

> «Proprio per la dimensione sociale che il testing assume, la valutazione è luogo di possibili conflitti politici o scontro di valori; attraverso la validazione, chi si occupa di testing ha l'obbligo di considerare le possibili conseguenze di un test ed evitare quelle negative. Di qui l'attenzione nel testing per definizione del costrutto e validazione» (Ferrari 2019: 95).

Questa consapevolezza delle implicazioni sociali e politiche del *testing* linguistico impone alla disciplina una responsabilità che va oltre la competenza tecnica nella costruzione degli strumenti. L'obiettivo del LT contemporaneo è quindi quello di definire teorie e strumenti che si fondino sulla conciliazione delle richieste derivanti dal ruolo sociale svolto dal settore con esigenze epistemologiche e metodologiche che ne garantiscano la validità, l'affidabilità, l'equità e la trasparenza.

È proprio per rispondere a questa duplice esigenza che il LT ha sviluppato *framework* teorici articolati che chiariscono la natura inferenziale del testing e i criteri attraverso cui valutare la legittimità delle interpretazioni prodotte. Comprendere come i test linguistici operino in quanto strumenti inferenziali costituisce il prerequisito per qualsiasi riflessione sulla loro validità. Riprendendo la definizione fornita da Green: «Un test è un evento appositamente predisposto per sollecitare una performance (solitamente entro un lasso di tempo prestabilito) allo scopo di formulare giudizi sulle conoscenze, le competenze o le abilità di una persona» (2020: 6; traduzione mia) [25]. L'obiettivo del test, quindi, non è semplicemente quello di registrare le risposte fornite dall'utente ma di utilizzarle per formulare giudizi sulla capacità dell'apprendente di usare la lingua in situazioni reali, al di fuori del contesto artificiale del test. Si tratta, in altre parole, di un processo inferenziale: dai comportamenti osservabili si risale a capacità sottostanti non direttamente analizzabili. Questa dinamica introduce una distinzione fondamentale tra ciò che si può effettivamente misurare, ovvero le risposte a determinati item o la produzione di specifici testi, e ciò che attraverso questi vogliamo conoscere, ovvero la competenza linguistico comunicativa dell'individuo. In questo binomio, il test costituisce lo strumento attraverso il quale raccogliamo evidenze, il criterio rappresenta invece l'oggetto ultimo della nostra indagine (McNamara 2004). Figura 11 schematizza la relazione tra test, criterio e costrutto.

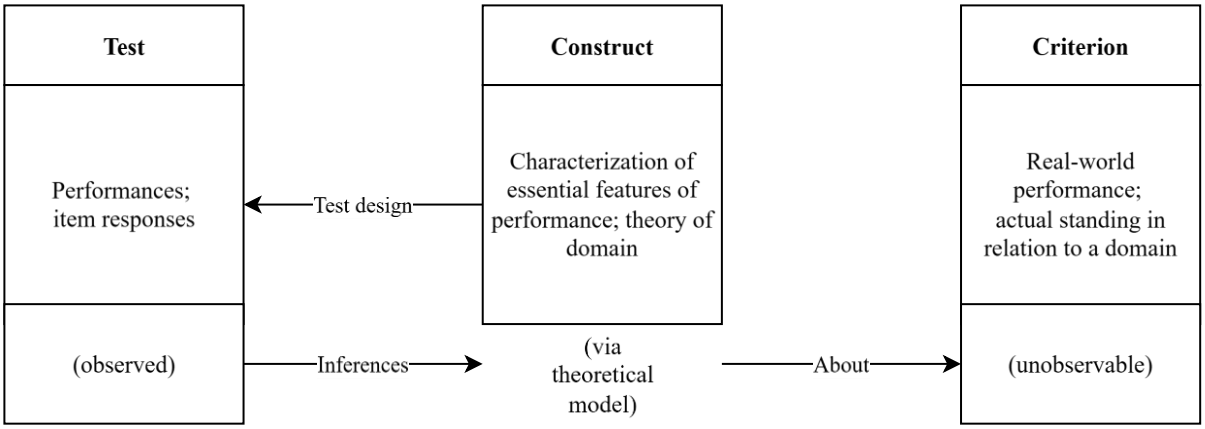


Figura 11: Test, costrutto e criterio (McNamara 2004:765)

Il problema epistemologico centrale del LT risiede nel fatto che il criterio non sia direttamente accessibile poiché quello che è osservabile è una sua manifestazione parziale e

[25] «A test is an event that is especially set up to elicit a performance (usually within a predetermined time frame) for the purpose of making judgements about a person's knowledge, skills or abilities».

situata (si veda la distinzione chomskiana tra *competence* e *performance* discussa nel paragrafo 1.3). Questa impossibilità di accesso diretto alla competenza richiama il paradosso dell'osservatore formulato da Labov (1972); come il ricercatore che, per il solo fatto di osservare, modifica il fenomeno studiato così lo sviluppatore di test con l'atto stesso di testare crea una situazione artificiale che inevitabilmente distorce, almeno in parte, ciò che vuole indagare. Per superare – o per lo meno gestire consapevolmente – questa difficoltà, i *test developer* elaborano modelli teorici della competenza linguistica che specificano quali dimensioni debbano essere oggetto di valutazione e come esse possano manifestarsi in comportamenti osservabili. Questi modelli, che nella letteratura specialistica prendono il nome di costrutti, fungono da mediatori concettuali tra la *performance* osservata nel test e la competenza generale che si intende inferire. È importante sottolineare che lo stesso costrutto non è semplicemente una descrizione neutra della realtà, ma una costruzione teorica che riflette specifiche assunzioni sul significato di competenza linguistica e su come questa possa essere rilevata. Proprio a causa della natura teorica ma al tempo stesso arbitraria dei costrutti, essi non possono essere accettati acriticamente, bensì, le scelte effettuate, devono essere argomentate scientificamente al fine di garantirne la validità generale.

Per sintetizzare quanto descritto finora, il processo di progettazione di un test linguistico implica una serie di passaggi metodologicamente ordinati. In primo luogo, occorre circoscrivere quali elementi della competenza linguistica debbano essere oggetto di indagine e attraverso quali manifestazioni essi possano emergere. Questo implica fare riferimento ad un modello teorico della competenza linguistica comunicativa – il costrutto – che fornisca una descrizione articolata di cosa significhi padroneggiare una lingua e che guidi la costruzione di strumenti valutativi. Con il costrutto chiaramente definito, il progettista seleziona i tipi di task più adatti a far emergere le competenze *target*. Seguono le procedure di *scoring* e la conversione delle valutazioni qualitative in dati quantitativi. Per concludere, è fondamentale che il tester sia in grado di stabilire una relazione coerente tra la *performance* osservata, l'abilità linguistica che si intende misurare e le modalità di impiego di tale abilità in contesti d'uso reali. La sequenza metodologica qui descritta non rappresenta solo un quadro teorico astratto, ma costituisce la traccia effettivamente seguita nella progettazione del test oggetto di indagine in questo lavoro. Le sezioni successive di questo capitolo documenteranno come ciascuna fase sia stata concretamente realizzata, rendendo esplicite le scelte operate e le loro giustificazioni teoriche ed empiriche.

È tuttavia opportuno sottolineare che elementi di soggettività e incertezza sono intrinseci nelle procedure di misurazione. Una serie di variabili può infatti influire sulla validità delle inferenze formulate a partire da un test: i task selezionati, il valutatore e le modalità di giudizio utilizzate, la scelta del framework di competenza linguistica, le caratteristiche proprie del candidato e infine le possibili interazioni tra le variabili stesse. Alan Davies nell'opera *Principles of language testing* spiega che:

> «Ogni misurazione linguistica comporta incertezza. Le caratteristiche gemelle di variabilità ed errore sono endemiche a tutti i tentativi di studiare e misurare l'apprendimento linguistico: variabilità dovuta all'imprecisione linguistica, errore dovuto a errori di misurazione» (1990: 179; traduzione mia)[26].

Nonostante questa condizione, lo studioso sostiene che sia necessario operazionalizzare l'incertezza ovvero, ridurre l'incertezza attraverso la definizione esplicita di criteri che guidino sia la concettualizzazione teorica della competenza sia la progettazione concreta degli strumenti. È in risposta a questa esigenza che negli anni sono stati sviluppati framework rigorosi e sistematici per valutare la qualità complessiva degli strumenti valutativi.

2.2 Le caratteristiche essenziali di un test

Gli studiosi Bachman e Palmer (1996) ritengono che la componente più importante da tenere in considerazione nello sviluppo di un test di lingua sia l'uso a cui esso è destinato; pertanto, la qualità fondamentale di un test è la sua utilità, ovvero il suo essere adatto agli obiettivi della valutazione. A seconda delle modalità di valutazione effettuate, in letteratura si distinguono tre principali tipologie di test: gli *achievement* o *attainment* test, i *proficiency* test e i *diagnosis* o *placement* test; questi si differenziano per i contenuti oggetto di valutazione e per il momento e le modalità di somministrazione (McNamara 2004).

I test di *achievement* o *attainment* sono associati ad un percorso formativo, si trovano principalmente all'interno di contesti didattici istituzionalizzati e misurano il grado di

[26] «All language measurement involves uncertainty. The twin features of variability and error are endemic to all attempts to study and measure language learning: variability because of linguistic imprecision, error because measurement failure».

apprendimento raggiunto in relazione a un curriculum o programma specifico. Sono strettamente ancorati ai contenuti insegnati nel corso e il loro obiettivo è misurare quanto l'apprendente ha appreso ciò che è stato insegnato. I test di *proficiency*, d'altra parte, non fanno riferimento a curricula ma mirano a valutare la competenza linguistica complessiva indipendentemente dal percorso di apprendimento seguito. Riguardano l'adeguatezza o meno della competenza linguistica di una persona a soddisfare un bisogno o uno standard predeterminato. I test di *proficiency* sono tipicamente ancorati a *framework* esterni – come il QCER in ambito europeo o l'American Council on the Teaching of Foreign Languages (ACTFL) in contesto americano – e intendono fornire una misura generalizzabile della competenza in domini d'uso reali. Infine, i test di *diagnosis* o *placement* servono a collocare gli apprendenti nel livello o corso più appropriato all'interno di un programma didattico strutturato. Sono interessati ad indagare quello che il candidato non sa piuttosto che quello che sa, con il fine di attenuare le lacune rilevate.

Per lungo tempo la letteratura sul *Language Testing* ha trattato le diverse qualità dei test come caratteristiche sostanzialmente indipendenti, ciascuna da massimizzare separatamente. Questa prospettiva ha generato posizioni estreme e problematiche, tra cui l'idea che determinati parametri siano intrinsecamente in conflitto; si sosteneva, per esempio, che aumentare l'affidabilità di un test significasse necessariamente sacrificarne la validità, o che compiti autentici non potessero essere valutati in modo affidabile. Sebbene esistano effettivamente tensioni tra le diverse qualità – ad esempio, task molto strutturati tendono a essere più facilmente valutabili ma meno rappresentativi dell'uso comunicativo reale – presentare tali tensioni come incompatibilità assolute è fuorviante e metodologicamente paralizzante.

In risposta a questa visione disgregante delle qualità dei test, Bachman e Palmer (1996) propongono un approccio che vede il costrutto dell'utilità composto da sei parametri posti in una relazione di complementarità; tale approccio è ritenuto tutt'oggi fondamento di qualsiasi riflessione relativa al *test design*. Piuttosto che cercare di massimizzare ciascuna qualità isolatamente, chi progetta un test deve individuare un equilibrio appropriato tra di esse, equilibrio che varierà necessariamente in funzione del contesto valutativo. Un test su larga scala destinato a prendere decisioni ad alto impatto richiederà livelli elevati di affidabilità e validità, a scapito, eventualmente, di una certa autenticità dei task; diversamente, un test somministrato in contesto didattico potrà privilegiare l'autenticità e l'interattività, accettando margini maggiori di variabilità nella valutazione.

Bachman e Palmer propongono di concettualizzare l'utilità complessiva di un test come una funzione di sei qualità essenziali – ovvero l'affidabilità, la validità di costrutto, l'autenticità, l'interattività, l'impatto e la praticabilità – tutte necessarie ma variamente bilanciate a seconda del contesto. La Figura 12 modellizza tale concettualizzazione.

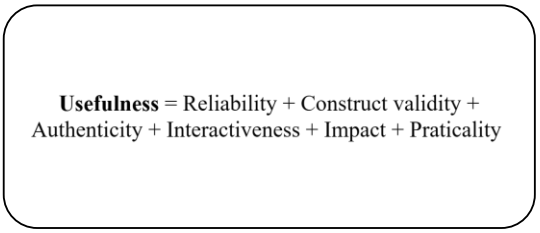


Figura 12: L'utilità e i suoi parametri costitutivi (Bachman e Palmer 1966:18)

Questa formulazione ha diverse implicazioni importanti. In primo luogo, sottolinea che nessuna qualità può essere completamente sacrificata: un test privo di affidabilità, per quanto autentico, non fornisce informazioni utilizzabili; un test privo di validità, per quanto affidabile, misura semplicemente qualcosa di irrilevante in modo coerente. In secondo luogo, riconosce che la valutazione dell'utilità è inevitabilmente contestuale e situata: non esiste un equilibrio universalmente ottimale, ma questo è sempre orientato allo scopo. In terzo luogo, ammette che determinare l'equilibrio appropriato richiede giudizi di valore da parte di chi sviluppa il test: si tratta di decisioni che, pur dovendo essere corroborate da evidenze empiriche e principi teorici, comportano inevitabilmente scelte interpretative attuate dagli sviluppatori.

Il *framework* proposto da Bachman e Palmer fornisce una cornice generale per valutare l'utilità complessiva di un test, trattando la validità di costrutto come una delle sei qualità che, insieme, determinano se uno strumento valutativo sia appropriato agli scopi che si prefigge. Tuttavia, per comprendere analiticamente cosa renda un test valido – quali evidenze specifiche siano necessarie per sostenere le interpretazioni dei punteggi – è utile fare ricorso a un ulteriore approccio complementare proposto da Cyril J. Weir (2005). Lo studioso, adottando una prospettiva socio-cognitiva, scompone la validità di costrutto di Bachman e Palmer in cinque dimensioni specifiche. Questi due approcci non si escludono a vicenda ma si integrano: il framework di Bachman e Palmer costruisce la cornice generale per pensare l'utilità; quello di Weir fornisce gli strumenti analitici per validare sistematicamente le interpretazioni dei punteggi.

2.2.1 Validità

La validità costituisce probabilmente la qualità più cruciale e al contempo più complessa di un test linguistico. In termini generali, un test è valido quando effettivamente misura ciò che si prefigge di misurare. Il processo attraverso cui si costruisce l'argomento a favore della legittimità delle interpretazioni proposte è un processo articolato che richiede la raccolta di evidenze di natura diversa.

Il framework più influente per organizzare tale processo è quello proposto da Weir (2005), che distingue diverse dimensioni della validità: la validità teorica (definita nelle opere successive dell'autore validità cognitiva, termine che verrà utilizzato d'ora in avanti anche in questo lavoro), la validità contestuale, la validità di scoring, la validità relativa ai criteri e la validità consequenziale. Ciascuna di queste richiede tipi specifici di evidenze e interviene in fasi differenti del ciclo di vita del test. L'autore sostiene che:

> «La validità è multiforme e sono necessari diversi tipi di prove per supportare qualsiasi affermazione sulla validità dei punteggi di un test. Queste non sono alternative, ma aspetti complementari di una base probatoria per l'interpretazione del test. Nessuna validità può essere considerata superiore a un'altra. La carenza di una di esse solleva dubbi sulla fondatezza di qualsiasi interpretazione dei punteggi dei test» (2005: 13; traduzione mia)[27].

È possibile innanzitutto portare una prima distinzione che vede una validità a priori, che riguarda la progettazione del test ed è relativa alla fase precedente alla somministrazione e una validità a posteriori, che riguarda tutto ciò che concerne il post somministrazione. Rientrano nella prima categoria la validità cognitiva e la validità di contesto; fanno parte della seconda, la validità di scoring, la validità relativa ai criteri e la validità consequenziale. Weir sostiene che, per l'elaborazione di un test metodologicamente appropriato, occorra tenere in considerazione le caratteristiche fisiche, psicologiche e esperienziali del candidato – definite da O'Sullivan (2000) *test taker characteristics* e illustrate in Tabella 4; i processi cognitivi richiesti per l'esecuzione del compito (validità cognitiva); le caratteristiche del compito e le condizioni in cui il test è somministrato (validità di contesto); le procedure e i criteri di valutazione (validità

[27] «Validity is multifaceted and different types of evidence are needed to support any claims for the validity of scores on a test. These are not alternatives but complementary aspects of an evidential basis for test interpretation. No single validity can be considered superior to another. Deficit in any one raises questions as to the well-foundedness of any interpretation of test scores».

di scoring); la correlazione con criteri esterni che possiedono una comprovata esperienza di misurazione (validità legata ai criteri).

Physical / Physiological	Psychological	Experiential
Short-term ailments (Toothache, cold, etc.)	Personality	Education preparedness
Longer-term disabilities	Memory	Examination experience
Speaking, hearing, vision (e.g., dyslexia)	Cognitive style	Communication experience
Age	Affective schemata	TL country residence
Sex	Concentration	
	Motivation	
	Emotional state	

Tabella 4: Caratteristiche dell'utente che si sottopone al test

Infine, vanno valutate le conseguenze che il test può avere sugli *stakeholder* – ovvero gli attori che in qualche misura prendono parte al processo di testing come i candidati, gli insegnanti, i genitori, il governo, gli enti ufficiale e il mercato in generale – e il grado di coerenza con gli scopi pedagogici e didattici del sistema educativo, e con i valori condivisi dalla società in cui il test ha luogo (validità consequenziale). La Figura 13 modellizza il framework appena descritto:

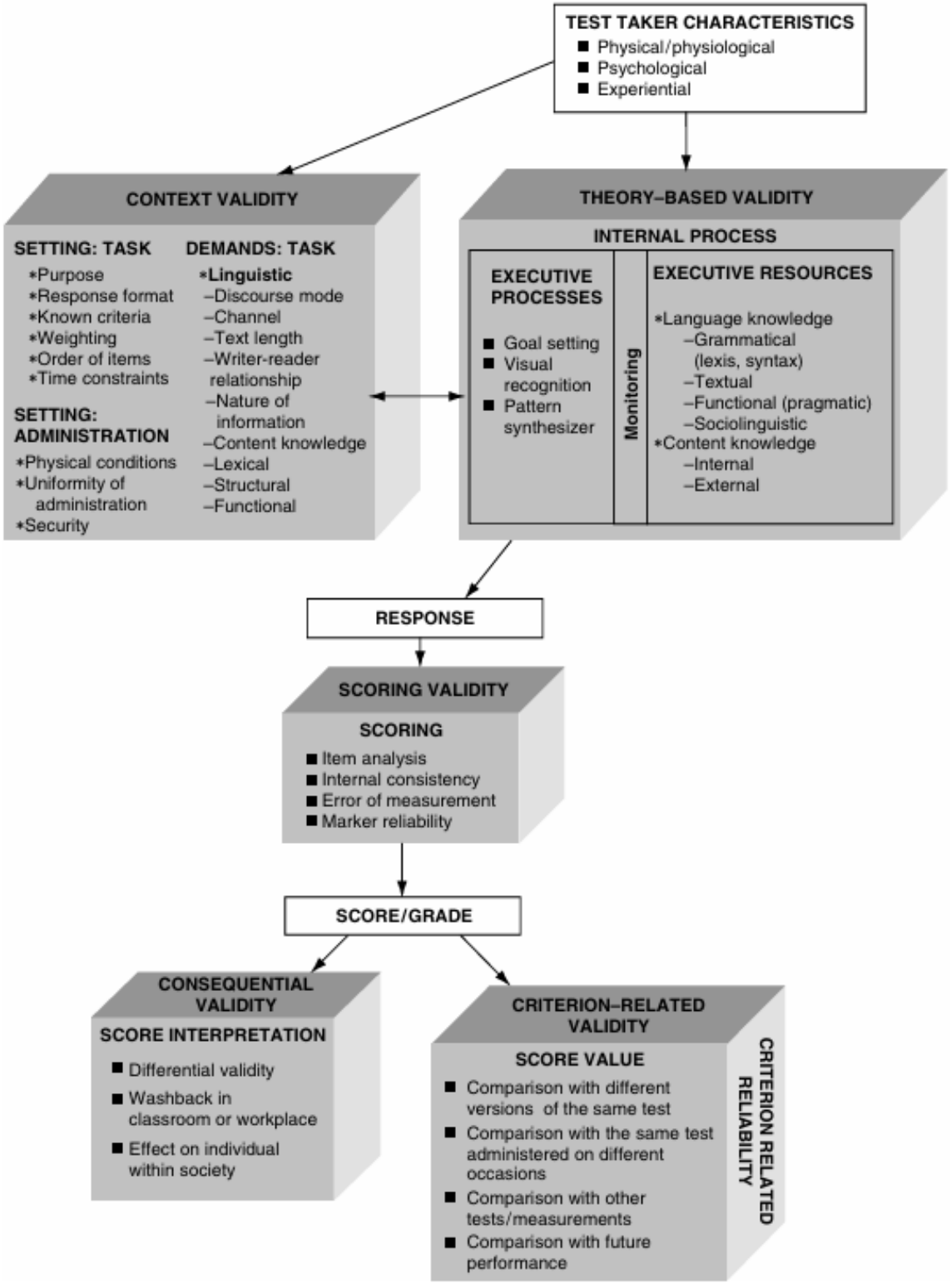


Figura 13: Framework socio-cognitivo per la validazione dei test. Nello specifico questo modello si riferisce all'abilità di lettura (Weir 2005: 44).

In primo luogo, per validità cognitiva si intende:

> «la misura in cui i compiti impiegati riescono a suscitare nei candidati un insieme di processi simili a quelli utilizzati da un ascoltatore esperto in un evento di ascolto nel mondo reale. Una seconda considerazione riguarda la precisione con cui i processi rilevanti vengono classificati nei vari livelli della suite in termini di richieste cognitive che impongono al candidato» (Field 2013:77; traduzione mia)[28].

Tradizionalmente, per validare le ipotesi cognitive venivano utilizzate procedure *post hoc*: attraverso la somministrazione del test si raccoglievano i dati necessari e su questi venivano applicate analisi statistiche per verificare se i punteggi si distribuissero secondo la struttura teorica attesa; in altre parole, si tentava di risalire ai processi cognitivi attraverso i *pattern* prodotti dai risultati. Cyril J. Weir si dimostra critico verso questa metodologia, poiché tali *pattern* potrebbero essere artefatti del formato del test, della selezione degli item, o di assunzioni teoriche nella progettazione dello strumento piuttosto che riflettere genuinamente la struttura della competenza linguistica. In risposta a queste criticità egli propone di integrare un approccio a priori che parta da evidenze empiriche indipendenti su come effettivamente funzioni l'elaborazione linguistica. Come lui stesso afferma:

> «Quanto più siamo in grado di descrivere in modo completo il costrutto che stiamo cercando di misurare nella fase a priori, tanto più significative potrebbero essere le procedure statistiche che contribuiscono alla convalida del costrutto, che possono essere successivamente applicate ai risultati del test. I dati statistici di per sé non generano etichette concettuali. Non possiamo mai sfuggire alla necessità di definire cosa viene misurato» (Weir 2005:18; traduzione mia)[29].

In altre parole, prima di costruire un test, occorre definire il modello cognitivo della competenza target, fondato sulla ricerca empirica in psicologia cognitiva e psicolinguistica sull'uso reale del linguaggio. Non è quindi sufficiente che un test dimostri che un candidato possiede un certo

[28] «By "cognitive validity" is to be understood the extent to which the tasks employed succeed in eliciting from candidates a set of processes which resemble those employed by a proficient listener in a real- world listening event. A second consideration is how finely the relevant processes are graded across the levels of the suite in terms of the cognitive demands that they impose upon the candidate».

[29] «The more fully we are able to describe the construct we are attempting to measure at the a priori stage the more meaningful might be the statistical procedures contributing to construct validation that can subsequently be applied to the results of the test. Statistical data do not in themselves generate conceptual labels. We can never escape from the need to define what is being measured».

livello di conoscenza linguistica astratta, deve anche fornire evidenze che il candidato sia capace di comportamenti comunicativi che soddisfano le richieste cognitive del contesto d'uso target.

Field (2013) articola tre questioni principali che la validazione cognitiva deve affrontare:

1. Similarità di elaborazione: I processi cognitivi attivati durante il test sono sufficientemente simili a quelli che verrebbero impiegati nel contesto comunicativo target oppure i candidati adottano strategie aggiuntive o alternative determinate da caratteristiche artificiali del test (quali procedure, modalità di somministrazione o tipologia degli item)? Questo criterio corrisponde alla preoccupazione di Messick (1989) per la *construct-irrelevant variance*, ovvero fattori estranei al costrutto che influenzano i punteggi.
2. Completezza: Gli item del test sollecitano un sottoinsieme limitato dei processi cognitivi che un utente competente impiegherebbe in contesti naturali, oppure coprono una gamma sufficientemente ampia di tali processi da poter essere considerati rappresentativi del comportamento comunicativo reale? Questo criterio riflette la preoccupazione di Messick per la *construct underrepresentation*, il rischio che il test catturi solo una porzione ristretta del costrutto teorico. La completezza può dipendere dal metodo di testing adottato (alcuni formati privilegiano certi processi a scapito di altri), ma anche dalla diversità degli item all'interno del test. Gli item operano prevalentemente a un singolo livello di elaborazione (ad esempio, tutti richiedono un riconoscimento lessicale ma non l'integrazione discorsiva) o a livelli molteplici?
3. Calibrazione: In test che valutano prestazioni secondo scale di livello (come quelle del QCER), le richieste cognitive imposte ai candidati a ciascun livello sono appropriatamente calibrate rispetto alle competenze caratteristiche di ciascun livello?

È inoltre importante chiarire che l'approccio cognitivo alla validazione non assume che tutti gli utenti di una lingua attuino i medesimi comportamenti. Esistono infatti variazioni individuali legate all'età, al *background* linguistico e alle strategie apprese che concorrono ad influenzare il modo in cui diverse persone affrontano gli stessi compiti comunicativi. Gli apprendenti di L2, in particolare, adottano spesso percorsi meno diretti ed efficienti rispetto ai parlanti nativi per compensare lacune nel repertorio lessicale o grammaticale. Tuttavia, come argomenta Field (2013), esistono ragioni solide per ritenere che alla base delle competenze linguistiche operino

routine di elaborazione consolidate e condivise che possono essere identificate studiando utenti esperti.

Esistono infatti due argomentazioni a favore di questa teoria: l'argomento universale e l'argomento legato all'esperienza. Il primo sostiene che tutti i cervelli umani condividono un'architettura fondamentale. A un certo livello di generalità, è ragionevole ipotizzare che questi abbiano in comune anche routine di elaborazione largamente simili, che riflettono i punti di forza e i limiti dell'apparato cognitivo umano e i meccanismi attraverso cui l'informazione viene elaborata. Queste routine influenzano non solo le forme che il linguaggio assume, ma anche i modi in cui viene processato durante la produzione e la comprensione. Il secondo vede invece la differenza fondamentale tra un parlante adulto di L1 e un apprendente di L2 nell'esperienza accumulata. L'adulto ha avuto infatti anni per sviluppare le routine di elaborazione più rapide ed efficaci e lo ha fatto senza l'interferenza di routine profondamente radicate associate a un'altra lingua. Comprendere come operano questi utenti esperti fornisce un modello verso cui orientare lo sviluppo degli apprendenti.

Il quadro di riferimento proposto da Weir pone quindi l'attenzione sulla dimensione cognitiva del candidato, sui processi mentali coinvolti nella dimostrazione delle competenze possedute o acquisite e sull'aspetto sociale dell'uso della lingua piuttosto che sul mero aspetto linguistico.

La validità di contesto concerne invece il grado in cui i compiti proposti nel test costituiscono un campione rappresentativo dell'universo più ampio di situazioni comunicative che il test pretende di misurare. Essa riguarda non solo le caratteristiche linguistiche dei *task* ma anche le condizioni in cui tali *task* vengono eseguiti: i vincoli temporali, le modalità di presentazione dell'input, i formati di risposta richiesti, le istruzioni fornite, le dinamiche interazionali previste. Come osserva Weir:

> «Se i compiti del test riflettono compiti di vita reale in termini di condizioni e operazioni contestualmente appropriate, è più facile affermare cosa uno studente può fare attraverso la lingua... a meno che non si adottino misure per identificare e incorporare tali caratteristiche, sembrerebbe imprudente fare affermazioni sulla capacità di un candidato di funzionare in condizioni normali nella sua futura situazione target» (1993: 28-9; traduzione mia).[30].

[30] «If the test tasks reflect real-life tasks in terms of important contextually appropriate conditions and operations it is easier to state what a student can do through the medium of English....unless steps are taken to identify and incorporate such features it would seem imprudent to make statements about a candidate's ability to function in normal conditions in his or her future target situation».

Un concetto chiave per operazionalizzare la validità di contesto è quello di dominio d'uso della lingua target, *Target Language Use domain* (TLU), definito da Bachman e Palmer come: «un insieme specifico di compiti di uso della lingua che il candidato probabilmente incontrerà al di fuori del test stesso, e a cui vogliamo che le nostre inferenze sulla capacità linguistica siano generalizzate» (1996: 44; traduzione mia)[31]. Il test dovrebbe quindi costituire un campione rappresentativo del TLU domain: non può ovviamente includere tutte le possibili situazioni, ma deve catturarne la varietà essenziale in modo che le prestazioni osservate nel test permettano di generare inferenze ragionevoli sulle prestazioni in situazioni non testate ma appartenenti allo stesso dominio.

Weir (2005) articola la validità di contesto distinguendo tra caratteristiche legate all'impostazione del compito e caratteristiche legate alle richieste del compito. Le prime riguardano il fatto di fornire consegne inequivocabili, chiarire le finalità comunicative dei compiti, diversificare i formati di risposta, ponderare il peso degli esercizi, ordinare gli item secondo criteri predefiniti, calibrare efficientemente i vincoli temporali; le seconde concernono il diversificare i generi testuali e il registro dei testi, la lunghezza, la natura delle informazioni fornite o richieste, le caratteristiche linguistiche. Entrambe le dimensioni devono essere attentamente calibrate per garantire che i *task* del test sollecitino elaborazioni appropriate e rappresentative.

L'autore sottolinea inoltre che la validità di contesto non opera in isolamento ma intrattiene relazioni simbiotiche con la validità cognitiva e la validità di *scoring*. Le caratteristiche contestuali dei *task* determinano infatti quali processi cognitivi vengono attivati: un compito di lettura con vincoli temporali molto stretti sollecita strategie di *skimming* selettivo[32]; lo stesso testo con tempo illimitato permette invece un'elaborazione intensiva e globale. Analogamente, i criteri di valutazione – se noti ai candidati – influenzano le strategie metacognitive adottate: se l'accuratezza formale non viene valutata, i candidati non dedicheranno risorse cognitive al monitoraggio di ortografia e grammatica.

Raggiungere piena validità di contesto è innegabilmente problematico data la difficoltà di caratterizzare con precisione la competenza linguistica e le minacce alla validità introdotte dall'artificialità inevitabile di ogni situazione di test. Tuttavia, come sottolinea Weir, queste

[31] «A set of specific language use tasks that. the test taker is likely to encounter outside of the test itself, and to which we want our inferences about language ability 'to generalize».
[32] Una discussione approfondita di queste pratiche si leggerà nel paragrafo 3.3.2.2.

difficoltà non esonerano gli sviluppatori di test dal cercare di rendere gli esami il più possibile pertinenti in termini di contesto.

D'altra parte, la validità di *scoring* riguarda la misura in cui i risultati di un test sono stabili, coerenti e privi di distorsioni sistematiche. Essa mira a valutare quanto i punteggi siano esenti da errori di misurazione e, di conseguenza, quanto sia legittimo fare affidamento su di essi per prendere decisioni sui candidati. In questa prospettiva, la qualità del processo di assegnazione dei punteggi costituisce una componente essenziale della validità complessiva del test: se i punteggi non riflettono in modo affidabile le prestazioni osservate, ogni interpretazione della competenza sottesa risulta compromessa (Weir 2005).

Nel dibattito metodologico sul *testing*, l'affidabilità è stata a lungo contrapposta ad altre dimensioni della validità, come se si trattasse di qualità in tensione tra loro. Da un lato, vi era chi sosteneva che senza affidabilità non potesse esistere in generale la validità, poiché punteggi incoerenti non possono supportare inferenze credibili; dall'altro, vi era chi osservava che un'enfasi eccessiva sulla coerenza numerica rischiasse di restringere il costrutto, privilegiando aspetti facilmente misurabili a scapito di dimensioni più complesse ma teoricamente rilevanti. Nella prospettiva contemporanea, tuttavia, l'affidabilità è concepita come una forma di evidenza di validità piuttosto che come proprietà separata: la validità di valutazione integra quindi tutte le dimensioni del processo di assegnazione e interpretazione dei punteggi che incidono sulla loro utilizzabilità per decisioni conseguenti.

Dal punto di vista psicometrico, la validità di *scoring* è strettamente connessa al problema dell'errore di misurazione. Come già accennato nel primo capitolo, nella Teoria Classica dei Test, il punteggio osservato ottenuto da un candidato è il risultato di una componente vera, che riflette l'abilità sottesa, e da una componente di errore, dovuta a fattori casuali o irrilevanti rispetto al costrutto. L'obiettivo della costruzione del test e delle procedure di scoring è quindi quello di massimizzare la varianza attribuibile alla competenza e minimizzare la varianza d'errore.

Le fonti di errore possono derivare da diversi fattori. Come già citato, alcuni sono legati ai candidati, come la stanchezza, l'ansia o predisposizioni personali, altri alla costruzione del test – potrebbero essere causati dall'ambiguità degli item, da un campionamento inadeguato del contenuto o da problemi tecnici generali – ed infine, altri ancora potrebbero essere legati alle procedure di valutazione derivanti da criteri poco chiari, da incoerenza tra valutatori o da errori di trascrizione degli input.

Nei test a risposta chiusa, l'analisi psicometrica degli item consente di identificare problemi attraverso indicatori quali la facilità e la discriminazione, che permettono di valutare rispettivamente il livello di difficoltà e la capacità dell'item di distinguere tra candidati di diversa abilità (Baker 1985). L'eventuale presenza di errore sistematico costituisce una minaccia particolarmente grave, poiché introduce distorsioni stabili nei punteggi. Sarà pertanto necessaria una fase di *pre-testing* atta a verificare e minimizzare la loro eventuale presenza, al fine di ridurre il più possibile la varianza. Per compiti a risposta aperta che implicano un giudizio umano, l'aspetto cruciale sarà rappresentato dall'affidabilità *inter-rater*, ossia, il grado di accordo tra valutatori diversi nell'applicazione degli stessi criteri alle stesse prestazioni. Punteggi differenti saranno indicativi del fatto che la variabilità deriva dal valutatore piuttosto che dalle prestazioni del candidato. La validità di *scoring* richiede pertanto criteri espliciti e operazionalizzati, formazione e calibrazione dei valutatori e procedure di controllo della coerenza nel tempo.

Quando la validità di punteggio viene applicata a sistemi di valutazione automatizzati, essa acquista delle caratteristiche peculiari che richiedono considerazioni specifiche. In questo contesto, l'affidabilità riguarda non solo la coerenza interna del sistema, ma anche la corrispondenza tra le valutazioni prodotte in maniera meccanica e quelle fornite dagli esperti umani. Risulta quindi necessario verificare empiricamente che il sistema automatizzato generi giudizi stabili nel tempo e che non introduca distorsioni sistematiche.

Un'ulteriore dimensione della validità di punteggio riguarda l'interpretabilità dei risultati attraverso una comparazione sistematica tramite una scala di prestazione predefinita. I valori numerici prodotti da un test non hanno di per sé un significato intrinseco; essi acquisiscono rilevanza solo quando sono collegati a livelli di competenza attraverso procedure e scale standardizzate, come per esempio il QCER.

In sintesi, la validità di punteggio costituisce una condizione necessaria ma non sufficiente per la validità complessiva di uno strumento. Da un lato, punteggi incoerenti o distorti compromettono qualsiasi inferenza sulla competenza del candidato, dall'altro, la sola coerenza matematica non ne garantisce la validità. L'obiettivo metodologico consiste quindi nel perseguire simultaneamente qualità tecnica del processo di *scoring* e fedeltà alla definizione teorica della competenza linguistico-comunicativa.

Mentre la validità teorica e la validità di contesto richiedono una comprensione approfondita di che cosa il *test* misuri – quale costrutto, attraverso quali operazioni cognitive, in quali condizioni – la validità relativa ai criteri adotta una prospettiva diversa e

complementare. Essa concerne la misura in cui i punteggi del test correlano con misure esterne di prestazione che possiedono proprietà consolidate e riconosciute (Messick 1989). Si tratta di un approccio prevalentemente quantitativo e a posteriori: richiede evidenze empiriche di relazioni sistematiche tra i risultati del test e altri indicatori rilevanti della competenza o della performance.

Tuttavia, come sottolineano diversi autori (Bachman 1990; Oller 1979), l'applicazione di questo tipo di validazione presenta difficoltà metodologiche non banali. Il problema centrale risiede nell'identificare misure di criterio sufficientemente valide con cui correlare i punteggi del test. Esiste infatti un rischio circolare: non si può affermare che un test abbia validità relativa ai criteri semplicemente perché correla altamente con un altro test, se quest'ultimo non è esso stesso una misura valida del costrutto in questione. Nonostante queste cautele metodologiche, la validità relativa ai criteri fornisce evidenze importanti, soprattutto quando le misure di criterio sono scelte con cura e le correlazioni vengono interpretate nel contesto di altre evidenze di validità.

La validità relativa ai criteri si articola tradizionalmente in due sottotipi, distinti in base alla dimensione temporale della relazione tra test e criterio: la validità concorrente, che esamina la correlazione tra i punteggi del test e misure di criterio raccolte simultaneamente o in tempi molto ravvicinati, e la validità predittiva, che analizza la capacità dei punteggi del test di predire performance future in contesti rilevanti. Una metrica tipicamente utilizzata nelle misure di criterio è la correlazione con i risultati di test precedenti che analizzano costrutti simili. In contesto didattico è possibile invece correlare i punteggi con le valutazioni fornite da insegnanti che conoscono approfonditamente i candidati e possono giudicare le loro competenze sulla base di osservazioni prolungate. È inoltre possibile effettuare un confronto tramite le autovalutazioni prodotte dagli utenti sulla propria competenza.

Con lo sviluppo e la diffusione di *framework* di riferimento come il QCER, la questione della comparabilità ha assunto rilevanza centrale poiché è possibile affidarsi nel confronto a livelli descrittivi ampiamente riconosciuti a livello internazionale. Nella misura in cui ci si può fidare dell'affidabilità delle misure di criterio scelte, correlazioni positive e significative forniscono evidenza che il test misura costrutti sovrapponibili. Tuttavia, correlazioni basse non implicano automaticamente mancanza di validità: potrebbero semplicemente riflettere che il *test* e il criterio catturano aspetti diversi della competenza linguistica.

Stabilire validità predittiva presenta invece difficoltà pratiche considerevoli poiché richiede studi longitudinali che seguano i candidati nel tempo, spesso per mesi o anni. Così

facendo la *performance* futura potrebbe essere influenzata da numerose variabili difficili da controllare come l'esposizione all'input della lingua *target*, eventuali opportunità di pratica e fattori personali e contestuali. Può inoltre succedere che i candidati non riescano a essere rintracciati o che non possano completare i percorsi previsti, riducendo la potenza statistica degli studi. Per queste ragioni, gli studi predittivi rimangono relativamente rari. Date le evidenti criticità, in questo lavoro la validità predittiva non verrà analizzata.

Le dimensioni della validità finora discusse si concentrano primariamente su questioni tecniche e psicometriche: cosa misura il test, come lo misura, quanto affidabilmente e come i punteggi si relazionano ad altri framework e misure. La validità consequenziale introduce una prospettiva radicalmente diversa, spostando l'attenzione dalle proprietà intrinseche dello strumento alle sue conseguenze – intenzionali e non – sugli individui, sulle istituzioni e sulla società. Messick argomenta che:

> «La questione è se le potenziali e reali conseguenze sociali dell'interpretazione e dell'uso dei test non solo supportino gli scopi previsti, ma siano anche coerenti con altri valori sociali. Poiché i valori attribuiti ai risultati desiderati e inattesi delle interpretazioni e dell'uso dei test derivano e contribuiscono al significato dei punteggi dei test, anche la valutazione delle conseguenze sociali dei test è considerata un aspetto della validità di costrutto... Per una visione completamente unificata della validità, si deve anche riconoscere che l'appropriatezza, la significatività e l'utilità delle inferenze basate sui punteggi dipendono anche dalle conseguenze sociali dei test. Pertanto, i valori e le conseguenze sociali non possono essere ignorati nelle considerazioni sulla validità» (1989: 18; traduzione mia)[33].

La letteratura sul LT distingue tipicamente tra due livelli di conseguenze, sebbene i termini utilizzati non siano sempre uniformi: l'impatto e il *washback*. Queste specifiche verranno trattate nel sottoparagrafo 2.2.5, essendo questi concetti equivalenti e trattati in entrambe le teorie.

[33] «The questions are whether the potential and actual social consequences of test interpretation and use are not only supportive of the intended testing purposes, but at the same time are consistent with other social values. Because the values served in the intended and unintended outcomes of test interpretations and test use both derive from and contribute to the meaning of test scores, the appraisal of social consequences of testing is also seen to be subsumed as an aspect of construct validity....For a fully unified view of validity, it must also be recognized that the appropriateness, meaning fulness, and usefulness of score based inferences depend as well on the social consequences of the testing. Therefore social values and social consequences cannot be ignored in considerations of validity».

È importante notare che i *framework* di Bachman e Palmer (1996) e di Weir (2005) non sono in contraddizione ma complementari. Bachman e Palmer propongono un modello delle qualità che contribuiscono all'utilità complessiva del *test*, trattando la validità di costrutto come concetto unitario. Weir, adottando una prospettiva socio-cognitiva, scompone analiticamente questo concetto unitario in dimensioni specifiche che richiedono tipi diversi di evidenza propri e intervengono in fasi diverse del ciclo di vita del test.

2.2.2 Affidabilità

Il concetto di affidabilità promosso da Bachman e Palmer (1996) riprende la teorizzazione proposta da Weir (2005) in relazione alla validità di *scoring* e indica la coerenza nella misurazione dei risultati. Un test affidabile fornisce punteggi stabili per lo stesso individuo in somministrazioni successive, a parità di competenza effettiva. In altre parole, se la competenza linguistica del candidato non è cambiata, il test dovrebbe assegnare punteggi molto simili in occasioni diverse.

Esistono diversi approcci tradizionali per stimare l'affidabilità di un test, ciascuno dei quali fornisce informazioni su aspetti diversi della consistenza: il metodo *test-retest*, il metodo delle forme parallele e il metodo del *semitest*. Il primo consiste nel somministrare lo stesso identico test due volte allo stesso gruppo di candidati a distanza di un breve lasso temporale e calcolare la correlazione tra i punteggi ottenuti nelle due occasioni. Un coefficiente di correlazione elevato indica stabilità temporale: il test produce risultati simili per gli stessi individui in momenti diversi. Tuttavia, il metodo *test-retest* presenta limiti metodologici significativi che ne condizionano l'utilizzo nel *Language Testing* contemporaneo. La *performance* nella seconda somministrazione può essere influenzata dall'esperienza acquisita nella prima; inoltre, se l'intervallo tra le somministrazioni fosse molto breve, i candidati potrebbero ricordare le risposte specifiche fornite nella prima occasione. D'altra parte, con intervalli lunghi, la competenza linguistica dei candidati potrebbe modificarsi a causa di esperienze intercorrenti, rendendo difficile distinguere instabilità del test da cambiamenti reali di competenza.

L'approccio delle forme parallele prevede la creazione di versioni multiple di test costruite per essere interscambiabili e utilizzabili in somministrazioni diverse. Due forme sono considerate parallele quando misurano le stesse competenze e sottocompetenze linguistiche, quando utilizzano la stessa distribuzione di formati di item (stessa proporzione di scelta multipla, vero o falso o produzioni) e prevedono le stesse condizioni di somministrazione,

vincoli temporali, criteri di valutazione. La correlazione tra i punteggi cerca di individuare quanto i risultati dipendano dagli item specifici selezionati piuttosto che dalla competenza sottostante misurata. Tuttavia, anche questa metodologia non è priva di criticità. In primo luogo, garantire che due forme siano genuinamente parallele è tecnicamente impegnativo e costoso: richiede *piloting* estensivo, analisi statistiche sofisticate, revisioni iterative. Inoltre, anche con cura estrema nella costruzione, è difficile essere certi che le forme siano perfettamente equivalenti. Piccole differenze nella formulazione degli item, nella familiarità degli argomenti, nella complessità sintattica possono introdurre differenze sistematiche di difficoltà.

In ultimo, il metodo del *semitest* o *split-half* rappresenta un compromesso che cerca di superare alcuni problemi dei metodi precedenti. Consiste nel somministrare il test una sola volta, ma poi dividere gli item in due metà e trattarle come se fossero forme parallele separate. La divisione tipica è quella pari-dispari: gli item con numero pari formano una metà, quelli con numero dispari l'altra. Si calcolano i punteggi separati per le due metà e si confrontano. Tuttavia, questa metodologia presuppone equivalenza delle metà, ovvero è efficace solo se le due metà sono effettivamente parallele in difficoltà e contenuto.

Come più volte specificato in questo lavoro, raggiungere l'affidabilità assoluta nei test linguistici si rivela un'impresa impossibile. I punteggi sono sempre soggetti a variabilità dovuta a fattori che vanno oltre il test stesso. L'obiettivo non deve quindi essere quello di eliminare completamente la variabilità ma di ridurla al minimo.

2.2.3 Autenticità

Anche il concetto di autenticità riprende la già citata validità contestuale. L'autenticità misura il grado di somiglianza tra i compiti proposti nel test e le situazioni comunicative che gli esaminandi affronteranno effettivamente utilizzando la lingua *target* al di fuori del contesto valutativo.

Essa riveste un ruolo cruciale per almeno due ragioni fondamentali. In primo luogo, l'autenticità costituisce un criterio decisivo per valutare la generalizzabilità delle interpretazioni: quanto più i compiti del test somigliano a situazioni d'uso reale della lingua, tanto più legittimo diventa estendere le conclusioni tratte dalla performance osservata durante il test alla competenza complessiva del candidato in contesti comunicativi autentici. In secondo luogo, essa influenza profondamente la qualità della performance stessa. Quando i candidati riconoscono una corrispondenza tra i compiti valutativi e situazioni concrete che incontrano o potrebbero incontrare nella vita reale, percepiscono la rilevanza e l'utilità della valutazione.

Questa percezione tende a generare un atteggiamento più positivo verso il test, riducendo resistenze psicologiche e permettendo ai candidati di mobilitare pienamente le proprie risorse linguistiche, manifestando così le loro effettive potenzialità comunicative (Bachman e Palmer 1996).

2.2.4 Interattività

L'interattività designa la misura e la modalità con cui le caratteristiche individuali del candidato vengono mobilitate e coinvolte attivamente durante lo svolgimento dei compiti del test. Secondo Bachman e Palmer, le dimensioni individuali più rilevanti che entrano in gioco durante una prestazione linguistica includono la competenza linguistica, la conoscenza de mondo, le caratteristiche personali e gli schemi affettivi (vedi Figura 14).

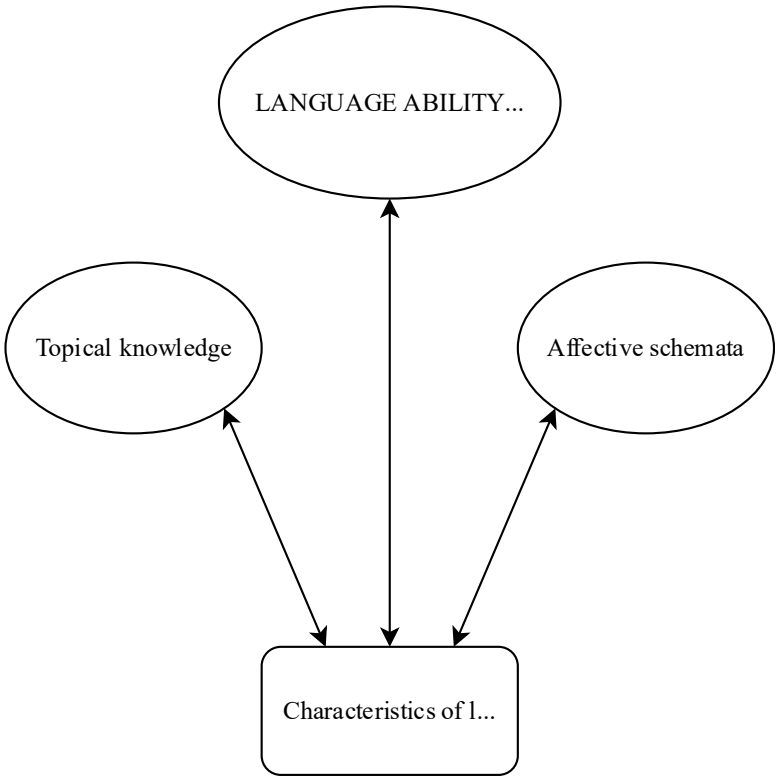


Figura 14: L'interattività (Bachman e Palmer 1996: 26)

L'interattività risiede nell'interazione dinamica tra queste risorse individuali e le richieste poste dai compiti del test. Un compito altamente interattivo è quello che richiede al candidato di mobilitare simultaneamente e in modo integrato molteplici dimensioni delle proprie risorse cognitive, linguistiche e affettive, piuttosto che attivare meccanicamente una sola competenza isolata.

È fondamentale riconoscere che l'interattività, analogamente all'autenticità, costituisce una qualità relativa piuttosto che assoluta: un test non è interattivo o non interattivo in senso binario, ma può essere più o meno interattivo.

2.2.5 Impatto

La nozione di impatto è equivalente al concetto di validità consequenziale proposta da Weir (1995). Essa si riferisce alle influenze che i test esercitano nella collettività. Tali effetti possono manifestarsi a livello macro, incidendo su contesti sociali più ampi – quali sistemi educativi, politiche linguistiche nazionali, percezioni pubbliche della competenza linguistica e accesso a opportunità educative e professionali – oppure a livello micro, producendo conseguenze per il singolo individuo. Quest'ultimo concetto viene definito *washback* e si riferisce principalmente alle influenze del test sui contesti immediati dell'insegnamento e dell'apprendimento: cosa accade in aula, come insegnanti progettano le loro lezioni, come studenti orientano i loro sforzi di studio, quali materiali didattici vengono prodotti e utilizzati. È importante sottolineare che sia l'impatto sia il *washback* sono concetti neutrali che possono tuttavia manifestarsi in forme positive o negative

Liz Hamp-Lyons (1997) sostiene che gli strumenti di valutazione dovrebbero essere esaminati non solo dalla prospettiva di chi li progetta ma anche da quella degli altri *stakeholder*, in quanto differenti attori sociali sperimentano conseguenze diverse derivanti dall'esistenza e dall'uso di un test. A tal fine, la studiosa individua cinque principali categorie di *stakeholder*: gli apprendenti, gli insegnanti, i genitori, le istituzioni governative e gli organismi ufficiali, nonché il mercato in senso ampio.

2.2.6 Praticabilità

La praticabilità rappresenta la sesta e ultima qualità che contribuisce all'utilità complessiva di un test secondo il *framework* proposto da Bachman e Palmer (1996). Essa possiede una natura distintiva rispetto alle altre cinque qualità: mentre gli altri parametri riguardano primariamente la qualità interpretativa dei punteggi e le conseguenze dell'uso del test, la praticabilità concerne aspetti più strettamente operativi e logistici – come il test viene concretamente progettato, sviluppato, somministrato e corretto. Bachman e Palmer definiscono la praticabilità come «la relazione tra le risorse che saranno richieste nella progettazione, nello sviluppo e nell'uso del test e le risorse che saranno effettivamente disponibili per queste attività» (1996:36; traduzione

mia)[34]. Questa definizione sottolinea come la praticabilità non sia una proprietà intrinseca del test in astratto, ma che emerga dal rapporto tra le risorse richieste e le risorse disponibili.

Come sottolineano gli autori, la praticabilità non può essere determinata in modo universale ed entra spesso in tensione con altre qualità dell'utilità: compiti autentici e interattivi tendono a essere più costosi da sviluppare e somministrare rispetto a formati standardizzati semplici; garantire alta affidabilità attraverso doppia correzione, procedure elaborate di formazione dei valutatori, analisi psicometriche sofisticate richiede risorse considerevoli; fornire un *feedback* dettagliato e personalizzato che produca *washback* formativo può rivelarsi dispendioso in termini di tempo e denaro. Ciò che rimane fondamentale è che questi compromessi siano consapevoli, espliciti e giustificati in relazione al contesto specifico, e che non compromettano le qualità minime indispensabili.

L'introduzione delle tecnologie digitali sta trasformando radicalmente i parametri della praticabilità: i test computerizzati riducono drasticamente i costi e i tempi di somministrazione; le pratiche di scoring automatico rendono possibili valutazioni su larga scala di produzioni aperte che prima richiedevano correzione umana intensiva; le *web-app* eliminano vincoli spaziali e temporali, permettendo somministrazioni asincrone e distribuite. Tuttavia, queste tecnologie introducono anche nuove richieste di risorse in termini di competenze tecniche, infrastrutture digitali, connettività affidabile e sicurezza informatica.

La Figura 15 riassume e modellizza tutti i parametri che, opportunamente bilanciati, concorrono alla realizzazione di un test utile.

[34] «The relationship between the resources that will be required in the design, development, and use of the test and the resources that will be available for these activities».

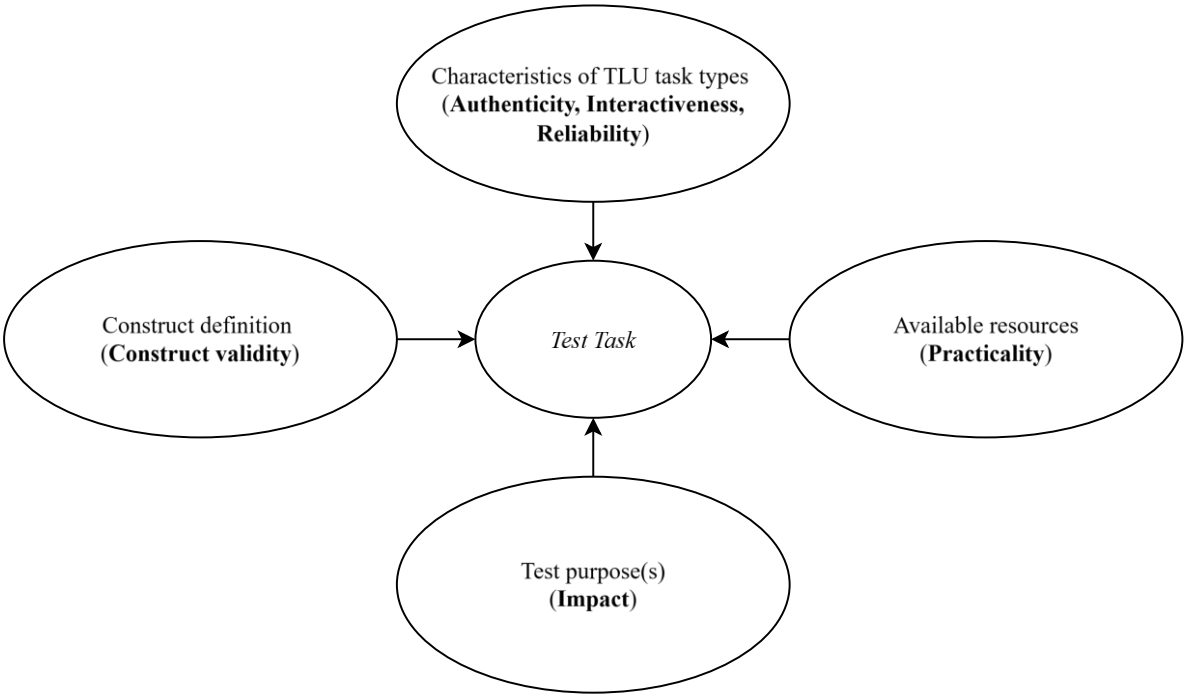


Figura 15: Componenti che intercorrono nella realizzazione del test (Bachman, Palmer 1996: 172)

2.3 Tipologie di prove e formati di item

Se il costrutto del test costituisce il "che cosa" debba essere valutato, le tecniche costituiscono il "come" questo avvenga (Bachman 1990). Le modalità attraverso cui si elicitano le prestazioni linguistiche dei candidati – i formati di *item*, i tipi di compito, le tecniche di risposta – non sono scelte neutrali o puramente tecniche, ma influenzano significativamente sia i processi cognitivi attivati durante l'esecuzione sia i punteggi finali ottenuti. Alderson, Clapham e Wall (1995) definiscono questo fenomeno "effetto metodo" (*method effect*): i candidati possono performare diversamente in funzione del formato utilizzato, indipendentemente dalla loro competenza effettiva nella dimensione linguistica oggetto di valutazione. Per far fronte a questa possibilità Sprugnoli sottolinea:

> «La conoscenza delle tecniche consente di determinare il tipo di test più adeguato, valido e affidabile, per verificare una determinata abilità. Consente, inoltre, di trovare un equilibrio tra dimensione produttiva e recettiva, tra difficoltà e facilità di elaborazione, somministrazione, correzione. La variazione dei tipi di test, inoltre, fa sì che non siano penalizzate o avvantaggiate particolari categorie di apprendenti» (2005: 165).

Risulta quindi opportuno variare le tecniche utilizzate nel test per non favorire alcuni candidati rispetto ad altri e scegliere dei formati adatti a verificare l'abilità in oggetto.

Indipendentemente dal formato specifico adottato, esistono insidie comuni che chi progetta test deve tenere in considerazione e cercare di evitare. In primo luogo, vi è la possibilità di testare qualcosa di diverso da ciò che ci si era prefissati. Per esempio, le conoscenze enciclopediche o culturali vengono spesso valutate inavvertitamente al posto della comprensione linguistica; una domanda che richiede argomenti specifici o che presuppone conoscenze specialistiche può penalizzare candidati competenti nella lingua ma privi di quelle informazioni extralinguistiche. Secondariamente, alcune costruzioni possono richiedere il ricorso a capacità generali, finendo così per testare l'abilità di *problem solving* o particolari tipi di intelligenza piuttosto che la competenza linguistica sottesa.

Le metodologie attualmente utilizzate nella valutazione linguistica includono un'ampia gamma di formati, che possono essere collocati lungo un *continuum* che si muove da formati altamente strutturati con *scoring* oggettivo (denominati *selected-response items*), passando per formati a risposta costruita limitata (detti *constructed-response items*) fino ad arrivare a formati aperti con scoring soggettivo che richiede un giudizio esperto (chiamati *extended constructed-response*) (Brown, Hudson 2002). La Tabella 5 schematizza queste differenze.

Response type	**Item format**
Selected response	Scelta multipla, vero o falso, matching, test a completamento, riordino, abbinamento
Constructed response	Cloze, gap filling
Extended constructed-response	Scrittura guidata, composizioni, riassunti, role play, conversazioni guidate

Tabella 5:Formati di item e tipologie di risposte

I formati a risposta selezionata richiedono ai candidati di scegliere la risposta corretta tra un insieme predefinito di opzioni, senza produrre alcun input linguistico autonomamente. Questa caratteristica li rende particolarmente appropriati per valutare abilità ricettive – comprensione scritta e comprensione orale – nonché conoscenze discrete come competenza morfosinttattica, lessicale o fonemica (Brown, Hudson 2002). I vantaggi generali di queste metodologie risiedono nella rapidità di somministrazione e correzione, nella possibilità di

automatizzare completamente lo *scoring* e nell'oggettività della valutazione. Tuttavia, gli aspetti negativi sono altrettanto significativi: creare *item* ben formulati con distrattori plausibili risulta tecnicamente impegnativo, i candidati possono rispondere correttamente "per caso" introducendo varianza d'errore e l'autenticità linguistica di questa tipologia di *task* risulta inevitabilmente limitata.

All'interno di questa categoria il formato di scelta binaria richiede ai candidati di giudicare la verità o correttezza di un'affermazione selezionando tra due sole alternative. Sebbene offra una focalizzazione chiara sul costrutto testato, il formato presenta una probabilità di indovinare del 50% che può essere attenuata solo aumentando significativamente il numero di item. Inoltre, è stato osservato che chi progetta test tende a costruire item ingannevoli basati su ambiguità, compromettendone potenzialmente la validità (Brown, Hudson 2002).

La metodologia di abbinamento richiede di collegare elementi di una lista con elementi di un'altra, tipicamente termini con definizioni o frasi con contesti appropriati. Questo formato offre efficienza spaziale (poiché con un unico esercizio possono essere testati molteplici quesiti) e bassa probabilità di *guessing*. Per garantirne l'efficacia occorre delineare liste brevi con dieci o dodici elementi, garantirne omogeneità di contenuto, includere più opzioni che prompt ed esplicitare chiaramente se le opzioni possono essere riutilizzate (Popham 1981).

La scelta multipla rappresenta il formato più diffuso per la tipologia a risposta predefinita. Tale metodologia è costituita da un quesito seguito tipicamente da tre, quattro o cinque opzioni di risposta, tra le quali una sola risulta corretta. I punti di forza del formato sono molteplici: la riduzione del *guessing* casuale rispetto a formati binari (con probabilità che scendono al 33%, 25% o 20% a seconda del numero di opzioni utilizzate)[35], la versatilità che rende questo formato applicabile a molteplici dimensioni linguistiche e la particolare efficacia nel testare abilità ricettive e conoscenze discrete. I limiti, tuttavia, sono altrettanto significativi. Questo formato non riflette l'uso produttivo della lingua e può risultare inappropriato per valutare competenze comunicative complesse o abilità produttive come scrittura e parlato. La costruzione di *item* a scelta multipla validi è inoltre tecnicamente complessa e richiede tempo considerevole. Per massimizzare l'efficacia di questo formato, diverse linee guida progettuali devono essere rispettate: è fondamentale garantire che esista una sola alternativa corretta, evitando ambiguità che comprometterebbe la validità dell'*item*; tutte le opzioni devono essere coerenti con il *prompt* e la lunghezza delle diverse alternative dovrebbe essere simile per evitare

[35] Una trattazione più dettagliata di questa tematica sarà riportata nel paragrafo 3.3.2.1.

di introdurre indizi involontari; la risposta corretta non dovrebbe essere identificabile senza riferimento al testo o al contesto, basandosi esclusivamente su conoscenze generali del mondo (Alderson et al. 1995).

I formati a risposta costruita limitata richiedono produzioni brevi e controllate – singole parole o brevi frasi – rappresentando una via intermedia tra l'assenza totale di produzione e la produzione estesa. Tale metodologia permette la valutazione integrata di abilità ricettive e produttive, offrendo flessibilità e autenticità moderate che li rendono più rappresentativi di alcuni usi linguistici rispetto ai formati puramente selettivi. È inoltre possibile ridurre quasi del tutto la possibilità di indovinare la risposta attraverso la richiesta di una produzione attiva da parte dell'utente. Il grande limite di questo formato è rappresentato tuttavia dalla variabilità delle risposte possibili per ciascun quesito. Esistono difatti molteplici risposte che potrebbero essere considerate corrette nel contesto preso in esame, costituendo un problema notevole per le pratiche di valutazione.

La tipologia principale di questa categoria è rappresentata dal completamento o *gap-filling*. Questo formato fornisce un contesto linguistico – una frase o un testo più ampio – al quale vengono rimosse delle sezioni che devono essere successivamente completate dai candidati. Per realizzare domande efficaci occorre fornire un contesto sufficiente ampio al fine di ridurre l'ambiguità, mantenere lunghezza uniforme degli spazi, evitare spazi consecutivi che creerebbero interdipendenza, preferire parole funzionali che ammettono meno variabilità e predeterminare criteri chiari per l'accettabilità delle risposte alternative.

L'ultimo formato è rappresentato dalle risposte libere che richiedono ai candidati produzioni linguistiche autonome e complete – testi articolati, interazioni orali prolungate, performance comunicative complesse – rappresentando l'estremo del continuum orientato verso l'autenticità e la produzione libera. I punti di forza di questa metodologia riguardano principalmente la validità; questi, costituiscono difatti l'unico formato appropriato per testare abilità produttive nella loro complessità, permettendo di osservare l'interazione tra abilità multiple e offrendo autenticità elevata e validità di contesto superiore. D'altra parte, uno dei limiti più rilevanti è costituito dalla soggettività dello scoring insito in questi *task*, che introduce variabilità nella valutazione e che necessita di un controllo sistematico. Inoltre, la complessità logistica, il tempo e i costi richiesti per la somministrazione e valutazione di questa metodologia risultano considerevoli.

All'interno di questa categoria la scrittura guidata elicita una produzione scritta attraverso l'ausilio di supporti strutturati quali grafici, tabelle o prompt dettagliati che

specificano con precisione il contesto comunicativo, lo scopo del testo, i destinatari previsti e il registro appropriato.

Le composizioni, invece, richiedono produzioni scritte più autonome su argomenti prestabiliti, riducendo il livello di strutturazione e supporto rispetto alla scrittura guidata. I punti di forza di questo formato includono principalmente la familiarità dovuta al fatto che il formato sia ampiamente utilizzato in contesti didattici. Le composizioni permettono inoltre la valutazione di capacità discorsive, organizzative e retoriche complesse che difficilmente emergono in formati più vincolati.

I riassunti e i *task* integrati richiedono ai candidati di leggere o ascoltare input linguistici e successivamente produrre testi che sintetizzino, analizzino o elaborino le informazioni ricevute. Questi formati integrano esplicitamente abilità ricettive e produttive in un unico compito complesso.

Infine, le performance orali includono formati quali giochi di ruolo (*role-play)*, conversazioni guidate e *information gap*[36], che richiedono interazione orale, simulazioni di situazioni comunicative e scambi dialogici tra candidati, tra candidato ed esaminatore o tra candidato e sistema. Questi formati permettono di valutare la competenza interazionale, catturando aspetti della comunicazione orale che metodologie meno interattive non potrebbero rilevare, quali la gestione dei turni, la negoziazione del significato e l'adattamento diafasico.

2.4 Le fasi di sviluppo di un test

Le diverse dimensioni della validità, le qualità dell'utilità e le considerazioni sui formati di task discusse nelle sezioni precedenti non rappresentano semplicemente un inventario di principi astratti, ma costituiscono un insieme di criteri operativi che devono informare sistematicamente ogni fase della costruzione dello strumento valutativo. Bachman e Palmer concettualizzano lo sviluppo di un test come un percorso articolato in tre fasi principali – progettazione, operazionalizzazione e somministrazione – che, pur seguendo una progressione tendenzialmente lineare, presenta carattere profondamente ciclico:

[36] Esercizi in cui due o più partecipanti possiedono informazioni differenti e devono comunicare per completare un compito comune.

> «Lo sviluppo del test è l'intero processo di creazione e utilizzo di un test. Il processo è organizzato in tre fasi: progettazione, operazionalizzazione e amministrazione. Sebbene lo sviluppo del test sia generalmente lineare, con lo sviluppo che procede da una fase all'altra, il processo è anche iterativo, in cui le decisioni prese e le attività completate in qualsiasi fase possono indurci a riconsiderare e rivedere le decisioni e a ripetere le attività svolte in un'altra fase» (1996: 93; traduzione mia)[37].

La complessità e l'investimento richiesto da questo processo variano in funzione degli scopi e delle conseguenze dell'uso del test. Per valutazioni formative a basso rischio le procedure possono essere relativamente rapide e informali, mentre in test ad alto impatto destinati a decisioni importanti richiedono rigore estremo, il coinvolgimento di *team* multidisciplinari e fasi di *piloting* estensivo.

Pur riconoscendo la natura ciclica che investe la procedura di sviluppo del test, è necessario organizzare concettualmente lo sviluppo in tre macrofasi che forniscono una struttura utile per monitorare l'utilità durante tutto il processo. La Figura 16 schematizza i principali passaggi che intercorrono nello sviluppo di un test.

[37] «Test development is the entire process of creating and using a test. The process is organized into three stages: design, operationalization, and administration. While test development is generally linear, with development progressing from one stage to the next, the process is also an iterative one, in which the decisions that are made and activities that are completed at any stage may lead us to reconsider and revise decisions and repeat activities that have been performed at another stage».

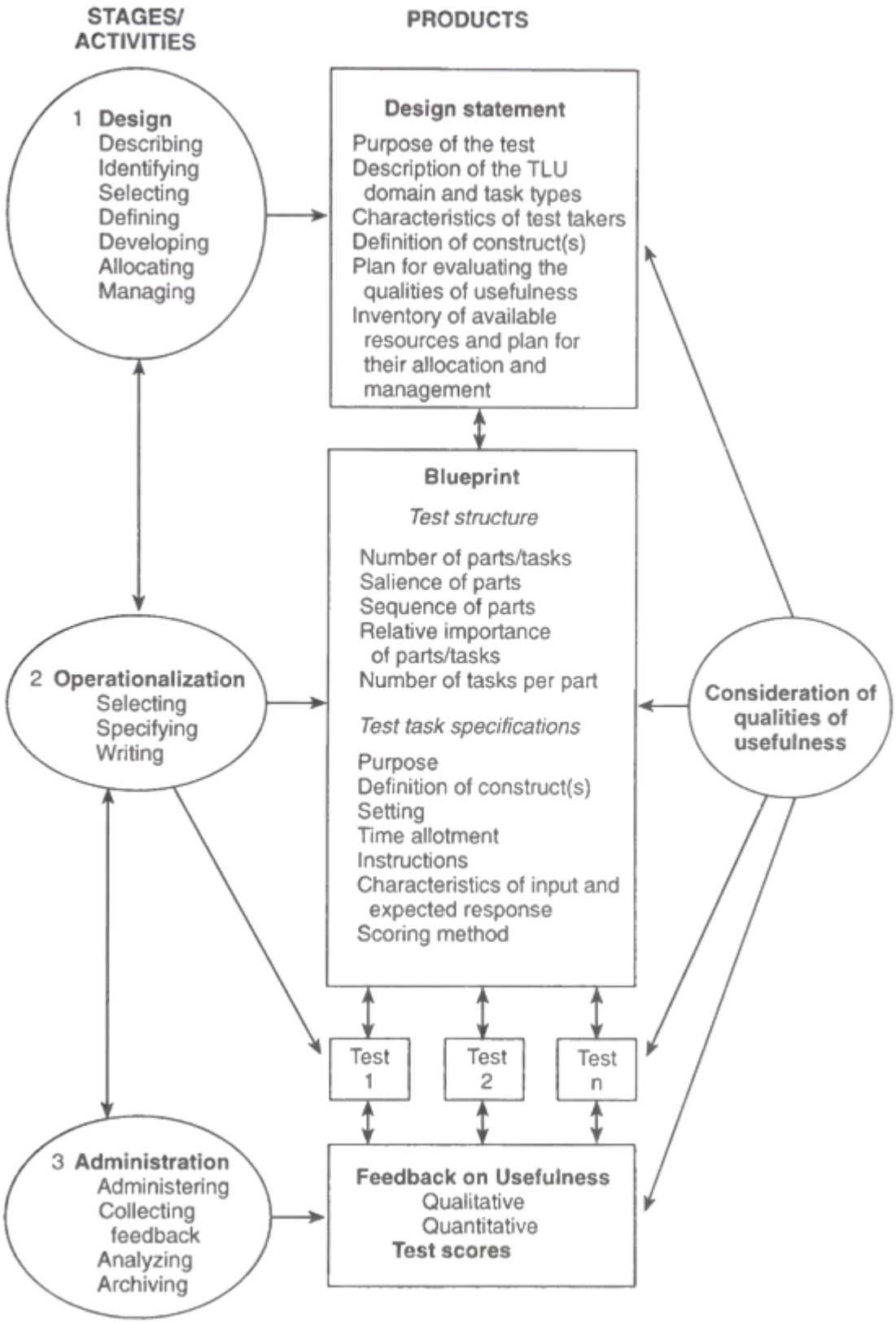


Figura 16:Procedure di sviluppo dei test (Bachman e Palmer 1996: 87)

La fase di progettazione mira a descrivere nel dettaglio le componenti che garantiranno che le prestazioni elicitate corrispondano all'uso linguistico nei contesti *target* e che i punteggi ottenuti siano utili per gli scopi previsti. Questo avviene attraverso una dichiarazione di progettazione (il *design statement*) che definisce lo scopo e gli usi specifici del *test*, indicando chiaramente quali inferenze sull'abilità linguistica o sulla capacità d'uso della lingua si intendono trarre sulla base dei risultati. In secondo luogo, viene definito l'insieme di situazioni comunicative *target* (il dominio TLU) a cui si intendono generalizzare le inferenze sui candidati. Come discusso nella sezione relativa alla validità di contesto e all'autenticità, la qualità di questa analisi determinerà direttamente quanto i compiti del *test* rispecchieranno usi linguistici reali e quindi quanto saranno generalizzabili i risultati. Successivamente, la descrizione dei candidati rende esplicita la natura della popolazione *target* in termini di caratteristiche demografiche, *background* linguistico ed educativo, esperienze pregresse, bisogni e obiettivi. La seguente definizione del costrutto in esame esplicita in termini teorici precisi la natura dell'abilità da misurare, specificando quali dimensioni della competenza linguistico-comunicativa sono oggetto di interesse e quali processi cognitivi sono attesi nell'esecuzione dei compiti. È inoltre prevista un'iniziale valutazione dell'equilibrio appropriato tra le sei qualità di utilità – validità, affidabilità, autenticità, interattività, impatto, praticabilità – nel contesto specifico e la definizione di livelli minimi accettabili per ciascuna qualità. Il piano specifica inoltre quali tipi di *feedback* – sia quantitativi (punteggi, statistiche degli *item*) sia qualitativi (osservazioni, autovalutazioni verbali dei candidati) – verranno raccolti durante il *piloting* e l'uso operativo, e quali procedure verranno utilizzate per analizzare tali informazioni. In ultimo, l'esplicitazione delle risorse umane, materiali, temporali ed economiche necessarie al fine di verificare la praticabilità effettiva del test progettato.

La fase di operazionalizzazione traduce i principi astratti della progettazione in specifiche concrete e materiali effettivi. Partendo dai *framework* teorici definiti nella fase di progettazione vengono modellizzate le operazioni cognitive richieste attraverso la realizzazione degli *item*, la definizione dell'*output* atteso e la formalizzazione delle condizioni di esecuzione. Attraverso la realizzazione di un *blueprint* viene descritta la struttura organizzativa complessiva del *test*: viene formalizzato il numero di sezioni che lo comporranno, come verranno presentati e distribuiti i *task*, le tempistiche concesse e le procedure di bilanciamento attuate. In ultimo vengono specificati i metodi di *scoring*, definiti i criteri di valutazione e le procedure per ottenerli, specificando chi valuterà, con quale formazione e attraverso quali protocolli.

La fase di somministrazione fa riferimento all'utilizzo effettivo del *test* da parte dei candidati. Questa può essere preceduta da una o più fasi di *piloting* operativo che permettono di raccogliere informazioni al fine di migliorare il *test*: tale meccanismo consente di individuare eventuali istruzioni ambigue, *task* inadeguati o vincoli temporali inappropriati. Le revisioni che ne conseguono possono essere più o meno radicali. Una volta ottemperati questi passaggi è possibile distribuire il *test*. La ricerca sistemica di evidenze circa l'affidabilità del *test* creato dovrebbe continuare anche dopo la diffusione dello strumento mediante l'analisi descrittiva dei punteggi, il monitoraggio degli *item*, la stima dell'affidabilità e la raccolta di *feedback* qualitativo.

È fondamentale sottolineare che lo sviluppo del test non termina con la prima somministrazione operativa. Come più volte esplicitato in questo lavoro, i test esistono in ecosistemi dinamici dove la popolazione in indagine, i contesti e le teorie evolvono costantemente. Un *test* ben progettato richiede manutenzione continua attraverso monitoraggio dell'utilità, aggiornamento periodico e revisione basata sulle evidenze accumulate. Le specifiche non dovrebbero essere fissate rigidamente ma rappresentare documenti in evoluzione, soggetti a revisioni quando nuove evidenze o cambiamenti contestuali lo richiedano.

CAPITOLO 3. METODOLOGIA E PROGETTAZIONE DELLO STRUMENTO DI VALUTAZIONE[38]

Il presente capitolo costituisce il passaggio dalla dimensione teorica a quella applicativa della ricerca, traducendo i *framework* elaborati nei Capitoli precedenti in scelte operative concrete. Se il Capitolo 1 ha definito il contesto teorico entro cui si colloca lo sviluppo dei sistemi ICALL e il quadro generale della valutazione delle lingue seconde, il Capitolo 2 ha approfondito i fondamenti teorici e metodologici del *Language Testing*, illustrando le qualità essenziali che caratterizzano uno strumento valutativo efficace e le principali fasi che ne guidano il processo di sviluppo. Alla luce di tali premesse, questo capitolo si propone di mostrare come tali principi vengano tradotti nella progettazione e nell'implementazione di un sistema di valutazione concreto. In questa prospettiva, dopo aver illustrato le motivazioni che hanno condotto alla realizzazione dello strumento, verrà presentato il caso studio oggetto del presente lavoro, ovvero il sistema ETET. Seguirà una breve descrizione dell'architettura della *web-app*, funzionale a comprendere il funzionamento generale della piattaforma, prima di analizzare in modo più approfondito il processo che ha guidato la costruzione del test e le principali scelte metodologiche e tecnologiche adottate. In particolare, nella sezione dedicata alla progettazione verranno illustrate le decisioni relative allo scopo dello strumento, alla definizione del dominio d'uso target e al piano di validazione della sua utilità. Successivamente, nella sezione di operazionalizzazione verrà descritta la costruzione concreta delle cinque sezioni che compongono il test – grammatica, lettura, ascolto, scrittura e parlato – con riferimento ai rispettivi framework cognitivi di base. Parallelamente, verranno presentate le soluzioni tecniche adottate per l'automazione del processo valutativo, fondate sull'integrazione di tecnologie di riconoscimento automatico del parlato e di *Large Language Models*. Infine, la sezione dedicata alla somministrazione illustrerà le modalità di reclutamento dei partecipanti e le procedure adottate per la raccolta dei dati, ponendo le basi metodologiche per l'analisi e la validazione empirica dei risultati presentate nel Capitolo 4.

[38] Alcuni degli argomenti trattati in questo capitolo sono stati esposti per la prima volta nel paper presentato alla conferenza di Linguistica Computazionale. Per maggiori informazioni visionare: Vignoli A., Combei C.R., Zappulla F., (2025), Verso la valutazione automatizzata dell'italiano L2: ETET tra LLM e tecnologie vocali, Proceedings fo the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), Cagliari, Italy, CEUR Proceedings & ACL Anthology, 282-291.

3.1 Il caso studio ETET

Lo studio realizzato in questo lavoro è stato reso possibile grazie alla collaborazione con ETET, Easy Talk Easy Test, una *start-up* italiana specializzata nello sviluppo di strumenti di valutazione linguistica automatizzata per le lingue seconde. L'azienda ha messo a disposizione la propria piattaforma tecnologica e le competenze del *team* di sviluppo, permettendo di tradurre i principi teorici discussi in questo lavoro in uno strumento operativo concreto, specializzato nella valutazione dell'italiano come L2.

Come anticipato nel Capitolo 1, questo studio prende le mosse dalla costatazione della necessità di strumenti che permettano di attuare una valutazione delle competenze linguistiche in maniera rapida e scalabile ma al contempo sicura e teoricamente valida. Negli ultimi anni si è infatti assistito a un incremento significativo nella domanda di strumenti valutativi per l'taliano L2, incentivata da interventi normativi che hanno reso la certificazione linguistica un requisito obbligatorio in specifici contesti amministrativi, accademici e lavorativi, generando una pressione crescente sugli organi certificatori tradizionali.

Alla luce di tali criticità, lo sviluppo di strumenti di valutazione automatizzati e scalabili, che non richiedano la presenza di un valutatore umano e che si fondino su tecnologie di intelligenza artificiale, si configura come una risposta strategica per democratizzare l'accesso alla valutazione linguistica. Tale approccio consente, al contempo, di preservare il rigore metodologico e di garantire l'allineamento agli standard internazionali di riferimento. L'automazione attraverso tecnologie ASA e AES permette di abbattere drasticamente i costi marginali per candidato, eliminare vincoli spaziotemporali attraverso somministrazione *web-based*, e fornire risultati semi-istantanei.

Per l'italiano L2, tuttavia, al momento della progettazione del presente lavoro non erano disponibili strumenti completamente automatizzati, facilmente accessibili e altamente personalizzabili, in grado di coprire l'intero spettro delle competenze linguistiche. La presente ricerca nasce dunque dall'esigenza di colmare tale lacuna attraverso lo sviluppo e la validazione di uno strumento di valutazione interamente automatizzato per l'italiano L2, progettato in conformità ai principi metodologici del *Language Testing* contemporaneo e validato empiricamente mediante un confronto sistematico sia con i giudizi di valutatori umani esperti sia attraverso evidenze riconducibili al più ampio quadro teorico della validità.

L'obiettivo dello strumento non è sostituire i sistemi di *assesment* consolidati – che rimangono imprescindibili nei contesti certificativi ad alto impatto – bensì integrare

l'ecosistema valutativo esistente, offrendo uno strumento pensato per essere utilizzato in situazioni in cui la valutazione tradizionale risulti impraticabile o difficilmente accessibile.
L'auspicio è che questo lavoro contribuisca tanto all'avanzamento scientifico nel campo del *Language Testing* automatizzato per l'italiano L2 – un ambito ancora relativamente poco esplorato – quanto alla disponibilità pratica di strumenti valutativi accessibili che possano supportare concretamente apprendenti, insegnanti, istituzioni educative e contesti professionali nel valutare e sviluppare competenze comunicative in italiano.

3.2 Architettura e funzionalità di ETET

La piattaforma ETET costituisce un sistema *web-based* che integra tecnologie avanzate di trattamento automatico del linguaggio, *machine learning* e tecnologie vocali per consentire la somministrazione scalabile di test linguistici e la valutazione automatizzata di produzioni sia chiuse che aperte. Il sistema si compone di diversi elementi: un'interfaccia utente tramite cui i candidati visualizzano e svolgono i test, un *back-end* per gestire lo *scoring* e la logica applicativa ed *un back-office* amministrativo per la costruzione dei test e il monitoraggio dei risultati.

Per quanto riguarda le funzionalità di valutazione automatizzata, la piattaforma integra tecnologie di *Automated Essay Scoring* per la valutazione di produzioni scritte e di *Automated Speaking Assessment* per la valutazione di produzioni orali, due strumentazioni consolidate in questo ambito applicativo. Il sistema di riconoscimento vocale automatico è implementato attraverso Azure AI di proprietà Microsoft, che utilizza il modello Whisper sviluppato da OpenAI. Gli input vocali vengono acquisiti direttamente tramite browser in formato audio compresso (.ogg) o non compresso (.wav), registrati in canale mono senza normalizzazione preventiva.

La valutazione di tutte le produzioni aperte – scritte e orali – è affidata a GPT-4o, LLM anch'esso sviluppato da OpenAI e disponibile tramite API. Per ogni tipologia di test e per ciascuna domanda di produzione vengono sviluppati *prompt* personalizzati che permettono al modello di assegnare i punteggi secondo criteri predefiniti, allineati ai livelli proposti dal QCER. La valutazione avviene in modalità asincrona; mentre il candidato completa il test, le risposte vengono elaborate in *background* dal sistema e i risultati vengono restituiti solo al

termine della somministrazione, garantendo che l'utente non riceva *feedback* parziali durante l'esecuzione che potrebbero influenzare le risposte successive.

Tutti i dati relativi agli esami – produzioni, punteggi e metadati – sono archiviati in un database ospitato in *cloud* da Microsoft Azure e gestiti conformemente al Regolamento Generale della Protezione dei Dati (GDPR).

La piattaforma ETET (cfr. Figura 17) è stata concepita per offrire flessibilità operativa agli amministratori dei test al fine di permettere la costruzione di configurazioni valutative personalizzate in base agli obiettivi specifici del contesto d'uso. Tale scelta progettuale riflette l'adesione al principio secondo cui non esiste uno strumento di valutazione universalmente ottimale per tutti i contesti; al contrario, l'efficacia e l'utilità di un test risultano sempre subordinate agli scopi specifici per cui esso viene impiegato.

La piattaforma permette di testare tutte e quattro le abilità fondamentali: la lettura, l'ascolto, la scrittura e il parlato. I test possono essere costruiti per riportare una valutazione complessiva della competenza oppure possono riportare una valutazione selettiva di una o più abilità di particolare interesse per il contesto specifico. Tale valutazione avviene attraverso l'utilizzo di item a risposta chiusa che permettono uno scoring automatico oggettivo immediato, e item a risposta aperta valutati automaticamente tramite LLM secondo criteri analitici multiparametrici.

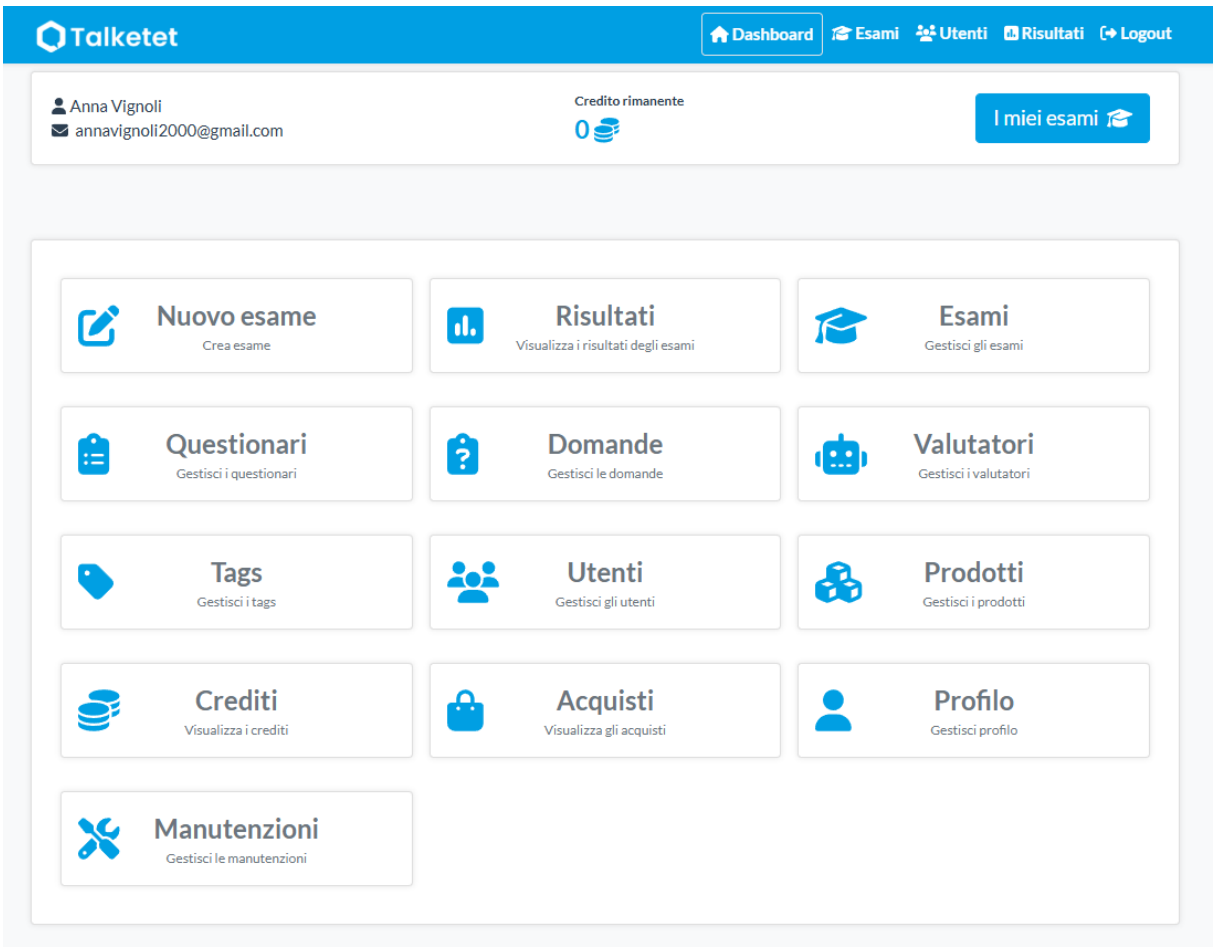


Figura 17: Schermata di back-office disponibile agli sviluppatori di test

La costruzione di una configurazione valutativa avviene attraverso un processo che segue una logica articolata in fasi sequenziali: una volta definite le abilità linguistiche che devono essere incluse nel test si procede alla costruzione delle domande e dei questionari, ovvero un insieme strutturato di *item* che valutano quella specifica competenza. L'*item bank* è composta da domande organizzate per lingua, abilità, livello QCER e oggetto epistemico testato o processo psicolinguistico sotteso, consentendo al *test designer* di prendere decisioni informate quando compone il questionario. Questi ultimi costituiscono l'unità organizzativa fondamentale: ciascuna abilità corrisponde a un questionario distinto, e l'insieme dei questionari costituirà la forma complessiva del test.

Il processo di costruzioni di un *item* inizia con la scelta della tipologia di *task* e della lingua oggetto di valutazione; segue la denominazione della domanda, indicizzata secondo i parametri sopra citati. Per esempio, una domanda di italiano livello B2 volta a indagare l'abilità di lettura e, nello specifico, il processo di *skimming* sarà nominata "IT-R-B2-2: comprensione per il succo del discorso, lettura globale rapida [Rita Levi-Montalcini 4]", includendo un numero

progressivo e, tra parentesi, un'indicazione sintetica dell'argomento utile allo sviluppatore. Successivamente vengono scritte le consegne della domanda, viene selezionato il *tag* relativo all'abilità linguistica che sarà utile al sistema nella valutazione finale del punteggio suddiviso per abilità, avviene l'assegnazione del minutaggio e, qualora necessario, vengono allegati file audio o immagini. Nel caso in cui la domanda rientrasse nella categoria delle composizioni libere, occorrerà selezionare il valutatore più idoneo (cfr. paragrafo 3.3.2.7) e nel caso delle domande di produzione orale, selezionare la casella per permettere la registrazione di un file audio. La Figura 18 mostra la schermata con i parametri necessari per costruire le domande.

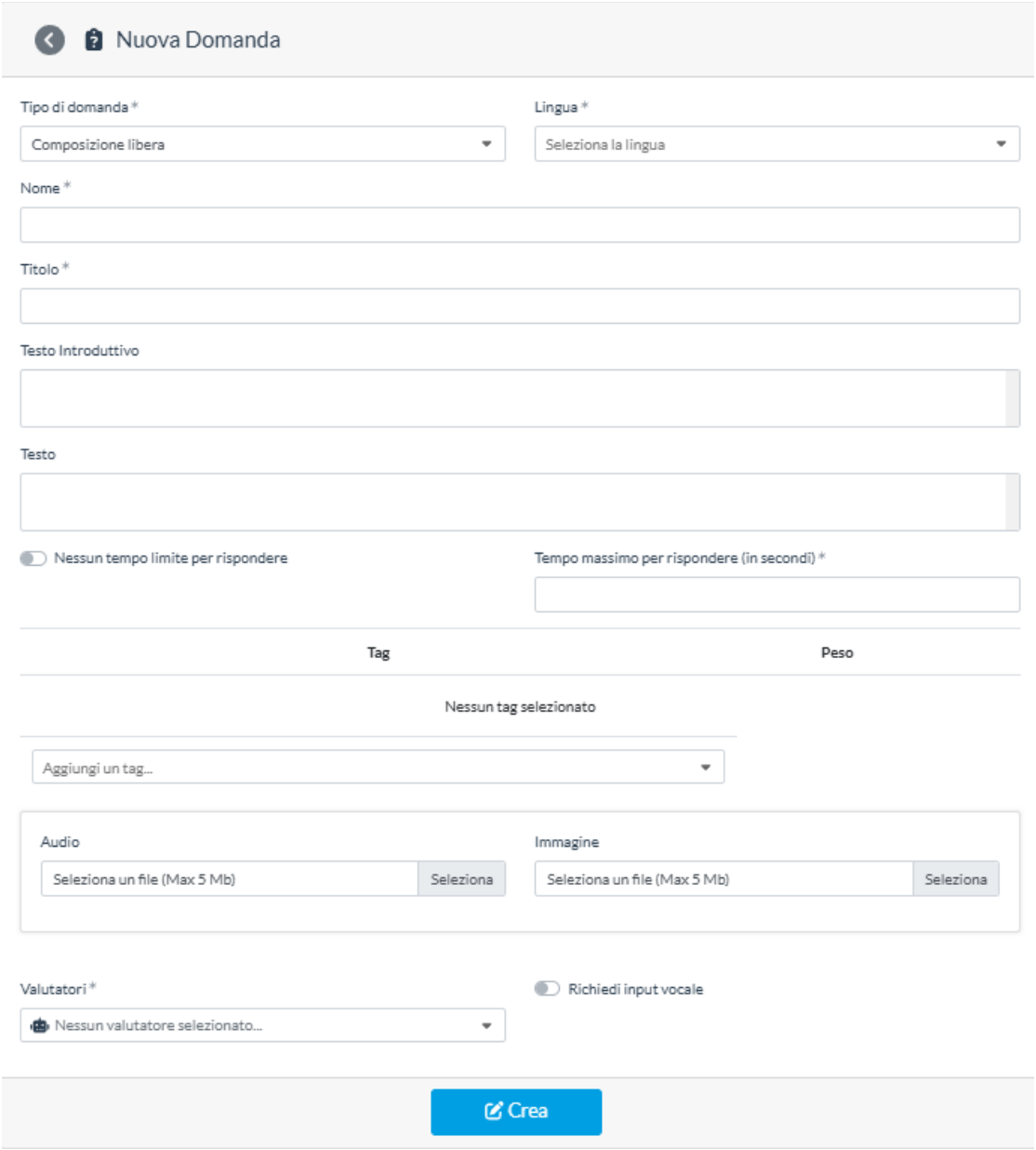
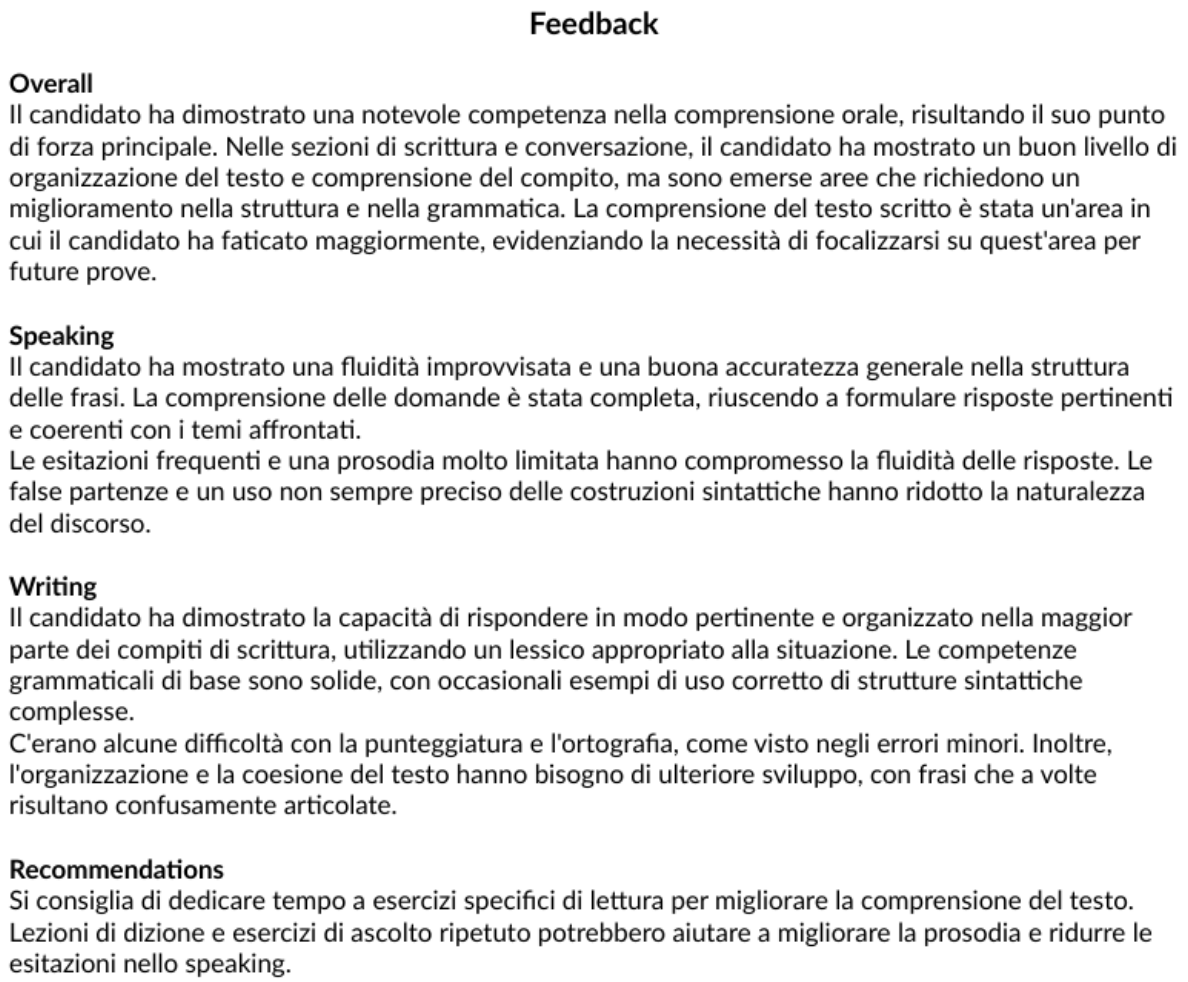


Figura 18: Schermata per la costruzione delle domande

L’output finale generato dal sistema consiste in punteggi su scala 0 -100 per ciascuna abilità testata e per il test complessivo, con una mappatura automatica sui livelli QCER sia globali sia disaggregati per singola competenza, fornendo un profilo diagnostico articolato della competenza del candidato.

Al termine della prova la piattaforma integra un sistema di feedback automatico personalizzato basato su tecnologie generative. Analizzando i risultati quantitativi conseguiti e le produzioni aperte del candidato, il sistema genera un report che identifica punti di forza della competenza linguistica, evidenzia aree di criticità emerse durante il test e fornisce raccomandazioni didattiche mirate per il miglioramento delle abilità che hanno ottenuto i risultati più bassi. Il feedback costituisce un valore aggiunto rispetto ai soli punteggi numerici, traducendo dati quantitativi in interpretazioni qualitative che orientano percorsi di apprendimento individualizzati. La Figura 19 mostra un esempio di feedback automatico personalizzato.



Feedback

Overall
Il candidato ha dimostrato una notevole competenza nella comprensione orale, risultando il suo punto di forza principale. Nelle sezioni di scrittura e conversazione, il candidato ha mostrato un buon livello di organizzazione del testo e comprensione del compito, ma sono emerse aree che richiedono un miglioramento nella struttura e nella grammatica. La comprensione del testo scritto è stata un'area in cui il candidato ha faticato maggiormente, evidenziando la necessità di focalizzarsi su quest'area per future prove.

Speaking
Il candidato ha mostrato una fluidità improvvisata e una buona accuratezza generale nella struttura delle frasi. La comprensione delle domande è stata completa, riuscendo a formulare risposte pertinenti e coerenti con i temi affrontati.
Le esitazioni frequenti e una prosodia molto limitata hanno compromesso la fluidità delle risposte. Le false partenze e un uso non sempre preciso delle costruzioni sintattiche hanno ridotto la naturalezza del discorso.

Writing
Il candidato ha dimostrato la capacità di rispondere in modo pertinente e organizzato nella maggior parte dei compiti di scrittura, utilizzando un lessico appropriato alla situazione. Le competenze grammaticali di base sono solide, con occasionali esempi di uso corretto di strutture sintattiche complesse.
C'erano alcune difficoltà con la punteggiatura e l'ortografia, come visto negli errori minori. Inoltre, l'organizzazione e la coesione del testo hanno bisogno di ulteriore sviluppo, con frasi che a volte risultano confusamente articolate.

Recommendations
Si consiglia di dedicare tempo a esercizi specifici di lettura per migliorare la comprensione del testo. Lezioni di dizione e esercizi di ascolto ripetuto potrebbero aiutare a migliorare la prosodia e ridurre le esitazioni nello speaking.

Figura 19: Esempio di feedback personalizzato disponibile per l’utente dopo il completamento della prova

L'architettura tecnologica e le sue funzionalità hanno costituito il punto di partenza operativo per lo sviluppo del test oggetto della presente ricerca. Le sezioni successive documentano come, partendo da queste capacità tecniche, sia stato seguito sistematicamente il processo tripartito di progettazione, operazionalizzazione e validazione identificato da Bachman e Palmer (1996), traducendo i principi teorici descritti in scelte concrete relative alla definizione del costrutto, alla selezione e costruzione dei task, alle procedure di *scoring* e ai metodi di validazione.

3.3 La costruzione del test

Viene di seguito proposta una schematizzazione delle scelte teoriche e delle metodologie pratiche che ricalca la tripartizione dello sviluppo dei test proposta da Bachman e Palmer (1996) discussa nel capitolo precedente. Si ritiene che tale scelta possa agevolare nella comprensione dei numerosi passaggi attuati.

3.3.1 Progettazione

La fase di progettazione ha seguito in maniera rigorosa il *framework* proposto da Bachman e Palmer (1996), articolandosi nelle sei componenti fondamentali della dichiarazione di progettazione. Ciascuna componente è stata sviluppata con l'obiettivo di garantire che il test risultante raggiungesse un equilibrio adeguato tra le diverse qualità dell'utilità in relazione agli scopi specifici della valutazione.

Il test è stato concepito per fornire una panoramica generalizzabile della *proficiency* del candidato, con l'obiettivo di formulare inferenze valide sul suo livello di competenza linguistico-comunicativa. Le inferenze che lo strumento intende supportare riguardano la capacità di utilizzare l'italiano in situazioni comunicative quotidiane che richiedono l'integrazione di abilità ricettive e produttive, sia orali sia scritte. Più specificamente, i punteggi ottenuti dovrebbero consentire di determinare il livello del QCER al quale il candidato può essere collocato in termini di competenza complessiva; individuare quali abilità linguistiche risultino più o meno sviluppate, fornendo così un profilo articolato piuttosto che un punteggio unitario indifferenziato; e infine, identificare eventuali aree di debolezza che richiedano interventi didattici mirati, orientando la progettazione di percorsi formativi personalizzati.

Il dominio d'uso della lingua *target* a cui il test intende generalizzare le proprie inferenze è costituito da situazioni comunicative quotidiane che un apprendente di italiano L2 potrebbe

incontrare tipicamente in contesti di vita reale. Questa scelta riflette l'obiettivo di costruire uno strumento di valutazione generale, applicabile a popolazioni eterogenee di apprendenti con bisogni comunicativi diversificati. L'analisi dei domini TLU ha identificato categorie di situazioni comunicative rilevanti, organizzate secondo le quattro abilità linguistiche fondamentali e la loro integrazione in contesti d'uso autentici. Per la comprensione scritta, le situazioni includono la lettura di comunicazioni pratiche come annunci pubblici, istruzioni operative e messaggi informativi che regolano attività ordinarie (vedi Figura 20); sono inoltre compresi testi informativi di interesse generale quali articoli di giornale, descrizioni relative a eventi e luoghi; infine, sono stati inseriti anche testi narrativi accessibili come racconti brevi o estratti letterari adattati.

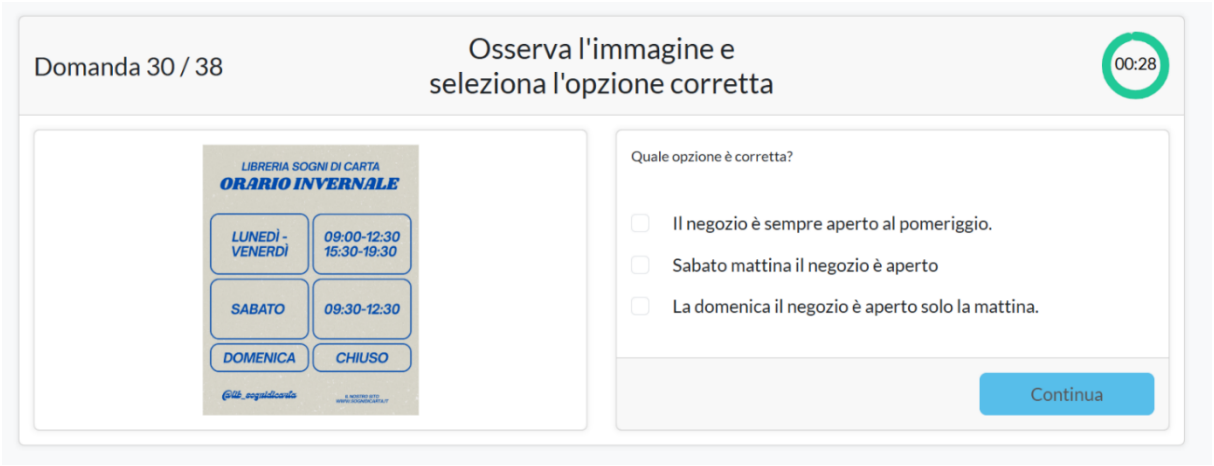


Figura 20:Esempio di task utilizzato per testare la lettura, incentrato sulla comprensione di un volantino riportante informazioni pratiche di prima necessità

Il domino relativo alla comprensione orale include invece l'ascolto di conversazioni quotidiane su argomenti familiari quali esperienze personali, progetti, interessi, opinioni condivise in interazioni sociali informali. Rientra anche la comprensione di informazioni trasmesse oralmente in contesti pubblici – annunci in stazioni, aeroporti, negozi, istruzioni fornite da personale di servizio – essenziali per orientarsi in ambienti sociali (come esemplificato nella Figura 21). Infine, la fruizione di contenuti mediatici accessibili quali brevi video informativi, podcast divulgativi, o programmi radiofonici introduce gli apprendenti a registri e varietà discorsive caratteristiche della comunicazione tramessa contemporanea.

Figura 21:Esempio di task utilizzato per testare l'ascolto, incentrato sulla comprensione di dettagli estrapolati da una conversazione tra un'operatrice museale e un utente

Per la produzione scritta, le situazioni comunicative rilevanti comprendono la redazione di comunicazioni personali e informali, come *e-mail* e messaggi, che costituiscono la forma prevalente di interazione scritta nella vita quotidiana (La Figura 22 mostra la casistica delle *e-mail*). Viene inoltre contemplata l'espressione scritta di opinioni e argomentazioni su temi di interesse generale e la descrizione di esperienze, eventi e situazioni che riguardano in prima persona la vita dell'individuo.

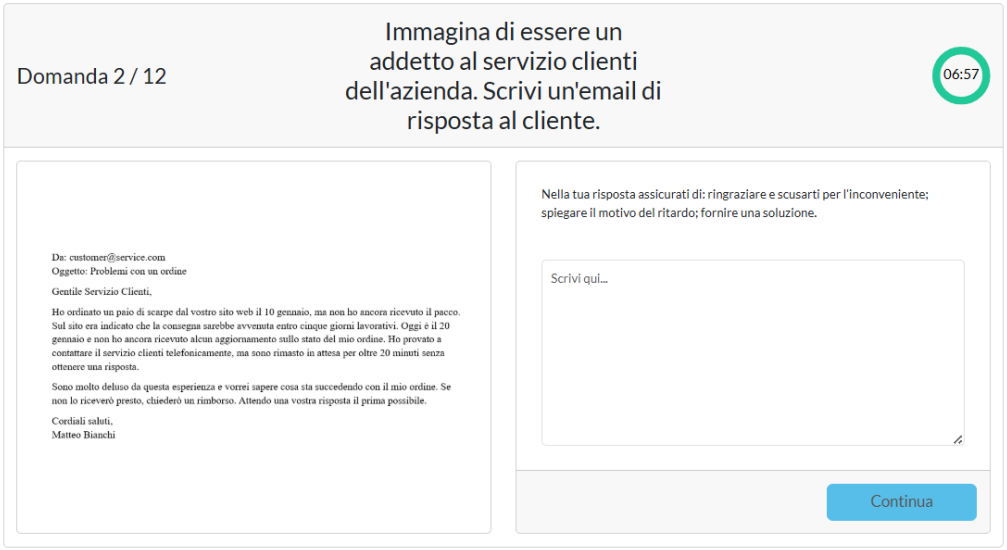


Figura 22:Esempio di task utilizzato per testare la scrittura, con particolare attenzione alla capacità di adeguarsi al registro diamesico e diafasico

Infine, il dominio relativo alla produzione orale comprende la descrizione di persone, luoghi e immagini, competenza essenziale per fornire informazioni referenziali o condividere impressioni visive. Inoltre, la narrazione di esperienze personali e racconti di eventi permette la costruzione di relazioni sociali attraverso condivisione di storie e vissuti. Allo stesso modo, l'espressione e l'argomentazione di opinioni su temi quotidiani consente la partecipazione attiva a discussioni, dibattiti informali e scambi dialogici dove si negoziano posizioni e prospettive (La Figura 23 riporta un esempio di questa tipologia di *task*). Infine, l'interazione dialogica in situazioni comunicative semplici costituisce la forma più basilare e frequente di uso della lingua orale.

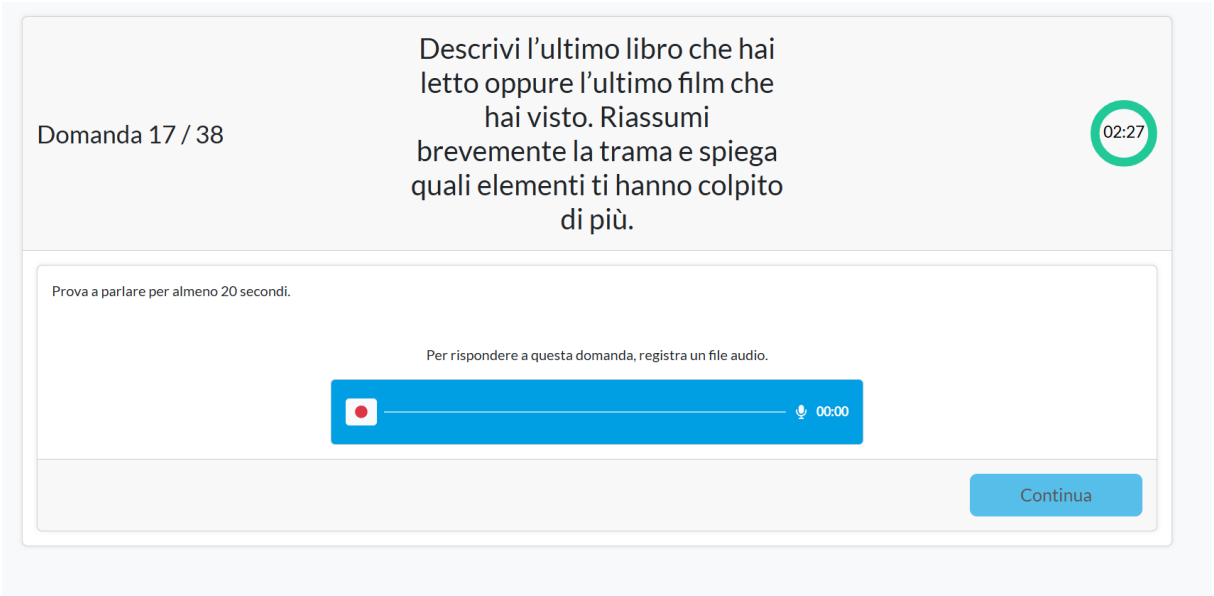


Figura 23:Esempio di task utilizzato per testare il parlato, nello specifico la narrazione di conoscenza

La definizione del costrutto adottata nel test integra il già descritto modello teorico della competenza linguistica comunicativa proposto da Bachman e Palmer (1996) con i descrittori di competenza articolati nel QCER. Questa doppio ancoraggio teorico risponde a due esigenze complementari: da un lato, disporre di un *framework* teorico che specifichi le dimensioni della competenza; dall'altro, allineare lo strumento valutativo a un sistema di livelli ampiamente riconosciuto e utilizzato a livello internazionale. Come discusso nella sezione sulla validità, la definizione del costrutto non costituisce un semplice esercizio teorico astratto ma ha implicazioni concrete per la progettazione dei task. Ciascun compito incluso nel test deve essere

giustificabile in termini di quali dimensioni della competenza intende sollecitare e attraverso quali processi cognitivi.

Il test è stato progettato per una popolazione eterogenea di apprendenti di italiano come lingua seconda, senza restrizioni rispetto a età, lingua materna, competenza o contesto di apprendimento. Questa scelta riflette l'ambizione di costruire uno strumento di valutazione generale e applicabile a contesti diversificati piuttosto che uno strumento specializzato per una nicchia specifica di utenti. I candidati devono tuttavia disporre di un computer o *laptop* connesso a Internet, microfono funzionante per le produzioni orali, e connessione sufficientemente stabile per la somministrazione *online*. Sebbene questa condizione introduca potenzialmente un elemento di iniquità – candidati senza accesso adeguato a tecnologie digitali risultano esclusi – essa costituisce un vincolo accettato consapevolmente in ragione dei benefici in termini di praticabilità e scalabilità offerti dalla piattaforma *web-based*. Si presuppone inoltre che i candidati possiedano un'alfabetizzazione di base e una familiarità sufficiente con interfacce digitali per svolgere il *test*. Per minimizzare l'influenza di queste variabili sulla performance, l'interfaccia è stata progettata con il fine di essere il più intuitiva possibile e le istruzioni sono formulate in modo chiaro e accessibile. La consapevolezza delle caratteristiche della popolazione target ha informato diverse scelte progettuali. Ad esempio, la selezione di argomenti culturalmente neutri o ampiamente accessibili riflette la preoccupazione per l'equità rispetto a *background* diversi.

Come discusso nella sezione teorica, l'utilità complessiva di un *test* è la funzione dell'equilibrio tra sei qualità fondamentali, opportunamente calibrate in relazione al contesto d'uso specifico. Il piano di valutazione dell'utilità definito per ETET ha individuato un bilanciamento ritenuto adeguato agli scopi e ai vincoli operativi del progetto.

In tale quadro, la validità di costrutto è stata considerata la qualità prioritaria, in quanto il *test* deve fornire misure significative delle dimensioni della competenza linguistico-comunicativa definite nel costrutto di riferimento. Tale validità è sostenuta da diverse fonti di evidenza: validità cognitiva (i task sollecitano processi mentali analoghi a quelli attivati in contesti comunicativi reali), validità contestuale (i compiti rappresentano in modo adeguato il *target language use domain*), validità di *scoring* (i punteggi risultano stabili e coerenti) e validità consequenziale (le conseguenze dell'uso del test sono allineate ai valori formativi dichiarati). Nello specifico, la validità cognitiva è garantita dall'ancoraggio del disegno dei task alle evidenze provenienti dagli studi psicolinguistici e socio-cognitivi che verranno discussi nei

prossimi paragrafi; la validità contestuale, oltre a fondarsi su tali evidenze, è ulteriormente assicurata dall'analisi dei domini d'uso linguistico delineati dal QCER.

Ulteriore caratteristica fondamentale è l'affidabilità, denominata nel *framework* di Weir (2005) validità di *scoring*. L'utilizzo di Large Language Models per la valutazione automatizzata delle competenze linguistiche rappresenta un'innovazione metodologica recente per la quale la ricerca non ha ancora raggiunto un consenso sulla validità delle misurazioni né sviluppato dei protocolli di validazione standardizzati e condivisi. Di fronte a questa assenza di *framework* consolidati, si è reso necessario progettare e implementare una strategia di validazione empirica specifica per ETET, garantita attraverso due fasi complementari: una fase di validazione che ha preceduto la somministrazione e una fase post somministrazione.

A luglio 2025, prima della distribuzione del test, è stata condotta una sessione di validazione preliminare per verificare la stabilità e la coerenza interna del sistema di *scoring* automatico in relazione alle domande aperte. Sono state selezionate tre domande di produzione scritta – PS1, PS2, PS3 – e tre di produzione orale – PO1, PO2, PO3 – rappresentative dei formati di task utilizzati nel test. Per ciascuna domanda è stata formulata una risposta di qualità controllata, prodotta da un utente competente a livello target specifico. Ogni risposta è stata sottoposta al sistema per dieci iterazioni consecutive mantenendo invariato l'input, al fine di osservare la stabilità dei punteggi assegnati a parità di contenuto. Per testare l'eventuale presenza di *bias* di genere, le domande di produzione orale sono state valutate sia su produzioni generate da un parlante maschile (PO1m, PO2m, PO3m) sia femminile (PO1f, PO2f, PO3f). Per ciascuna iterazione sono stati registrati i punteggi relativi a tutti i parametri considerati: *structure and grammar* (X_1), *content and argumentation* (X_2), *vocabulary* (X_3), *comprehension and adherence to topic* (X_4), *pragmatics and cohesion* (X_5). Per gli item di produzione orale, in aggiunta ai parametri sopra citati, sono stati analizzati anche i risultati relativi alla *fluency* (X_6), *accuracy* (X_7), *completeness* (X_8) e *pronunciation* (X_9). La coerenza dei punteggi è stata quantificata attraverso il coefficiente di variazione (CV), ottenuto dal rapporto percentuale tra deviazione standard e valore medio per ciascun parametro (vedi formula 1):

$$CV = \frac{\sigma}{\mu} \times 100 \quad (1)$$

Il coefficiente di variazione descrive la dispersione relativa: valori elevati indicano alta variabilità nei risultati e bassa coerenza, valori bassi mostrano invece stabilità; in questo lavoro la soglia di accettabilità è stata fissata al 10%.

I risultati dell'analisi – riportati in Tabella 6 – mostrano come tutte le misurazioni, ad eccezione del parametro X_1 nella domanda PO2 con voce femminile, si collocano al di sotto del limite prestabilito del 10%, evidenziando buona coerenza del modello nell'assegnazione dei punteggi.

Domande	**X_1**	**X_2**	**X_3**	**X_4**	**X_5**	**X_{6-9}**
PS1	4,9%	3,8%	3,9%	6,5%	2,6%	\
PS2	5,1%	3,8%	4,4%	4,1%	2,6%	\
PS3	2,5%	2,8%	3,6%	2,7%	3,9%	\
PO1f	2,5%	3,2%	7,0%	3,9%	4,8%	0,0%
PO2f	10,6%	4,0%	8,1%	4,6%	8,8%	0,0%
PO3f	4,1%	4,1%	5,6%	3,9%	2,8%	0,0%
PO1m	6,7%	4,0%	5,9%	2,3%	6,0%	0,0%
PO2m	7,8%	2,7%	6,5%	3,4%	4,9%	0,0%
PO3m	7,4%	3,9%	5,5%	2,8%	6,0%	0,0%

Tabella 6:Valori CV per ciascun tipo di domanda e parametro

Il confronto tra produzioni maschili e femminili non ha evidenziato differenze sistematiche nei punteggi medi o nella variabilità. Questo suggerisce che il modello sia sostanzialmente indifferente alla variabile di genere, oppure che il genere, combinandosi con altri fattori acustici – come il tono di voce, la distanza dal microfono, la velocità di eloquio, l'eventuale presenza di rumore di sottofondo – non assume rilevanza determinante nel calcolo finale. Tale esito è coerente con l'obiettivo di costruire uno strumento robusto e il più possibile privo di *bias* che penalizzerebbero sistematicamente sottogruppi specifici.

La stabilità interna del modello, indagata nella prima fase di verifica, è condizione necessaria ma non sufficiente per garantire la validità di *scoring*: un sistema può essere perfettamente coerente nell'applicare criteri errati. Per verificare che i punteggi automatici siano allineati con i giudizi umani esperti, è stata condotta una validazione empirica attraverso il confronto sistematico dei risultati prodotti dal sistema con i voti forniti da tre annotatori umani. Questa validazione, documentata dettagliatamente nel paragrafo 4.4 del Capitolo 4, fornisce evidenze sull'accordo *inter-rater* tra sistema automatico e valutatori umani, costituendo la forma più robusta di validazione dell'affidabilità e della validità di *scoring*.

Inerentemente al parametro dell'autenticità, già in parte menzionato in relazione alla validità contestuale, si è cercato di preservarne valori elevati nei limiti dei vincoli tecnologici e operativi. Tuttavia, l'autenticità assoluta è limitata dalla natura inevitabilmente artificiale della situazione di test e dai vincoli imposti dai formati valutativi.

L'interattività è invece variabile a seconda del tipo di task. Le produzioni estese sono progettate per essere altamente interattive, richiedendo mobilitazione integrata di conoscenze linguistiche, competenza enciclopedica e strategie metacognitive. Al contrario, *item* a scelta multipla presentano interattività limitata ma sono giustificati dalla necessità di campionare efficientemente ampie porzioni del dominio linguistico.

Inerentemente al parametro dell'impatto, l'obiettivo principale è quello di massimizzare il *washback* positivo minimizzando conseguenze negative. Il carattere multidimensionale, l'accessibilità tramite piattaforma online e la generazione di *feedback* personalizzato articolato per abilità, dovrebbero orientare gli apprendenti verso pratiche di studio mirate.

Infine, anche la praticabilità è stata una qualità fortemente privilegiata attraverso le scelte tecnologiche attuate. L'automazione completa dello *scoring* – possibile grazie all'integrazione di tecnologie ASR e LLM – elimina la necessità di valutatori umani, riducendo drasticamente tempi e costi. In aggiunta, la somministrazione *web-based* elimina vincoli logistici legati a sedi fisiche e orari prestabiliti.

La progettazione del test è stata condizionata dalle risorse effettivamente disponibili e dai vincoli operativi del contesto di ricerca. Per tali ragioni, le ambizioni progettuali sono state calibrate per non eccedere le capacità effettive di implementazione. Dal punto di vista tecnologico ETET utilizza una piattaforma *web* con capacità di somministrazione *online*, *storage* sicuro dei dati e interfacce utente; un accesso a GPT-4o via API OpenAI per lo *scoring* automatizzato delle produzioni aperte; un sistema ASR basato su Azure per la trascrizione delle produzioni orali; l'attuazione della sintesi vocale tramite ElevenLabs per la generazione degli audio utilizzati negli item di comprensione orale. Per quanto riguarda invece le risorse umane è stato necessario l'intervento del *team* di sviluppo per l'implementazione tecnica, la presenza di un ricercatore principale e due docenti supervisori per la progettazione linguistica, tre annotatori esperti per la validazione empirica dello *scoring* automatico e cinque utenti madrelingua che hanno costituito il gruppo di controllo. In termini di risorse temporali, le tempistiche limitate conseguenti allo sviluppo del lavoro in ambito di un progetto di ricerca magistrale hanno influenzato la scala del *piloting* e il numero di iterazioni possibili nel raffinamento dei *prompt*. Infine, inerentemente alle risorse economiche, è stata prevista

l'erogazione di un incentivo per i partecipanti che hanno scelto di aderire alla ricerca; tuttavia, le limitate risorse, hanno comportato l'esigenza di ricorrere alla soluzione della collaborazione volontaria per quanto riguarda gli annotatori esperti.

3.3.2 Operazionalizzazione

Successivamente alla definizione dello scopo che ha portato alla progettazione del test, all'identificazione del dominio TLU da porre in indagine, alla descrizione dei candidati, all'analisi del costrutto, alla descrizione del piano di valutazione attuato e allo studio delle risorse e dei vincoli si è passati alla progettazione effettiva del test, traducendo questi principi astratti in pratiche e materiali concreti.

La prima fase ha comportato la creazione di un *blueprint* – ovvero, un documento che esplicitasse la struttura organizzativa del test – nel quale sono state operazionalizzate le diverse componenti del costrutto. La Tabella 7 schematizza le scelte adottate nella creazione del test. L'esame risulta così strutturato in cinque sezioni, presentate secondo un ordine progressivo che riflette un criterio di gradualità cognitiva ed emotiva. La prima sezione è dedicata alla competenza grammaticale, seguita dalla comprensione scritta e dalla comprensione orale; il test si conclude con le sezioni di produzione scritta e di produzione orale. L'ordine di somministrazione delle prove è stato definito con l'obiettivo di porre progressivamente l'utente a proprio agio, riducendo l'impatto emotivo associato alle abilità produttive. In particolare, si è scelto di collocare la sezione grammaticale in apertura della prova poiché essa si basa su tipologie di esercizi ampiamente familiari agli utenti, in virtù del loro frequente impiego nei contesti didattici. Tale scelta consente di favorire un primo approccio rassicurante al *test*. Le attività di produzione – che richiedono invece un maggiore sforzo cognitivo, comportano un coinvolgimento emotivo più elevato e sono spesso percepite come più stressanti – sono state collocate nelle fasi finali della prova, in modo tale che l'utente le affrontasse dopo aver avuto modo di familiarizzare con la struttura e le modalità del *test*, riducendo inoltre il rischio che l'ansia scaturita da queste prove compromettesse l'intera performance del *test*.

Ciascuna delle cinque sezioni è stata progettata e bilanciata per coprire l'intero spettro di competenza formalizzato dal QCER presentando un 10% di domande relative ai livelli A1 e C2, mentre tutti gli altri livelli – A2, B1, B2, C1– sono uniformemente rappresentati da un 20% di domande. Tale strutturazione presenta una maggiore densità di *item* nei livelli intermedi, riflettendo la considerazione teorica che la maggior parte degli apprendenti di una L2 si colloca tipicamente in queste fasce. Al contrario, i livelli estremi ricevono allocazione ridotta poiché

rappresentano porzioni più esigue della popolazione. Le sezioni di competenza grammaticale, lettura e ascolto sono composte rispettivamente da dieci *item*; diversamente, la sezione relativa alla scrittura è articolata in cinque item mentre quella di parlato in tre. Il bilanciamento interno del *test* non riguarda solo la difficoltà associata all'*item* ma un insieme di elementi quali la tipologia dei *task*, gli argomenti e le tematiche delle domande e il focus linguistico (Alte 2011).

Il tempo complessivo massimo per completare l'intero *test* è di circa 70 minuti. Diversamente da ciò che avviene nei test tradizionali, caratterizzati da vincoli temporali uniformi per sezione, il sistema adotta un approccio flessibile in cui ciascun *item* ha associato un minutaggio intrinseco. L'allocazione del tempo per ciascuna domanda è stata calcolata tenendo in considerazione il tempo di lettura della consegna in una lingua non materna e la tipologia di *task* proposto. In particolare, ai *task* a risposta chiusa o predefinita è stato attribuito un tempo inferiore rispetto a quelli che richiedono una produzione autonoma da parte dell'utente. Un ulteriore criterio considerato nella distribuzione del tempo è stato la tipologia di sforzo cognitivo richiesto dal compito. Per esempio, nel caso degli esercizi di comprensione scritta, le domande che sollecitavano una lettura di tipo selettivo, finalizzata all'individuazione di informazioni specifiche (quali per esempio, date, luoghi o nomi), sono state associate a un minutaggio più contenuto, al fine di limitare il ricorso a strategie di lettura approfondita. Al contrario, le domande che richiedevano l'attivazione di processi cognitivi più complessi, come l'inferenza o la sintesi delle informazioni, hanno previsto una maggiore disponibilità di tempo.

Sezione	**Numero di item**	**Formati di task principali**	**Tempo (minuti)**	**Modalità di scoring**
Grammatica	10	Scelta multipla, *Gap-filling*	~10	Automatica tramite chiavi
Comprensione scritta	10	Scelta multipla, Scelta multipla visuale, Cloze guidato, Collegamento, Riordinamento	~15	Automatica tramite chiavi
Comprensione orale	10	Scelta multipla, Scelta multipla visuale, *Gap-filling*	~15	Automatica tramite chiavi
Produzione scritta	5	*Gap-filling,* Scrittura guidata, Composizioni	~20	LLM
Produzione orale	3	Conversazioni	~11	LLM + ASR
Totale	38	-	~70	-

Tabella 7:Blueprint complessivo del test

La progettazione dei *task* per ciascuna sezione del *test* è stata orientata da un'analisi approfondita dei principi di validità inizialmente formulati da Weir (2005) e successivamente ampliati e sistematizzati nella serie di volumi pubblicata da Cambridge ESOL. Tale collana si concentra sull'analisi e sulla descrizione dei principi di validità che caratterizzano le quattro abilità linguistiche fondamentali, attraverso una serie di opere di riferimento: *Examining Writing* di Shaw e Weir (2007), *Examining Reading* di Khalifa e Weir (2009), *Examining Speaking*, a cura di Taylor (2011), e infine *Examining Listening* di Geranpayeh e Taylor (2013).

3.3.2.1 Valutare la grammatica

La sezione dedicata alla competenza grammaticale mira a valutare la conoscenza esplicita e la capacità di applicare correttamente regole e strutture proprie dell'italiano. Come discusso in più sezioni di questo lavoro, la competenza grammaticale costituisce una componente fondamentale della competenza organizzativa all'interno del modello di Bachman e Palmer. Sebbene non sia sufficiente per garantire una comunicazione efficace, essa fornisce le risorse strutturali necessarie per codificare e decodificare significati in modo preciso. I fenomeni grammaticali selezionati nel test riflettono le progressioni tipiche dello sviluppo morfosintattico in italiano L2, come teorizzate e descritte negli studi sulle sequenze acquisizionali. A tal proposito, si è fatto affidamento all'opera *Valutare e certificare l'italiano di stranieri* (2003), nella quale Barki – in unione a numerosi altri studiosi – propone una rassegna delle principali sequenze acquisizionali ravvisabili nell'italiano L2: partendo da quella relativa alle distinzioni morfologiche del verbo, per poi passare alla sequenza relativa all'accordo di genere, citando quella inerente allo sviluppo della negazione per poi terminare con la discussione delle sequenze di sviluppo relative alla sintassi. Altra opera fondamentale utilizzata per la classificazione degli elementi grammaticali e lessicali nelle relative fasce di competenza è rappresentata dal *Profilo della lingua italiana* (2010) di Spinelli e Parizzi[39]. La sezione relativa alle liste lessicali proprie di ciascun livello è stata utilizzata tanto per la creazione dei *task* quanto per la selezione del lessico più appropriato nella descrizione delle consegne.

[39] Alcune sezioni dell'opera sono consultabili online al seguente link: https://www.unistrapg.it/profilo_lingua_italiana/site/index.html

I *task* utilizzati nella sezione grammaticale sono principalmente risposte multiple sotto forma di menù a tendina (come mostrato dalla Figura 25) – tipologia utilizzata principalmente per valutare i livelli inferiori – e gli esercizi di completamento nel quale l'utente deve produrre un input linguistico (come indicato in Figura 24). In questo ultimo caso la consegna ha riportato il numero massimo di parole richieste e, qualora non fosse esplicitato nel testo, il lemma che si intendeva testare.

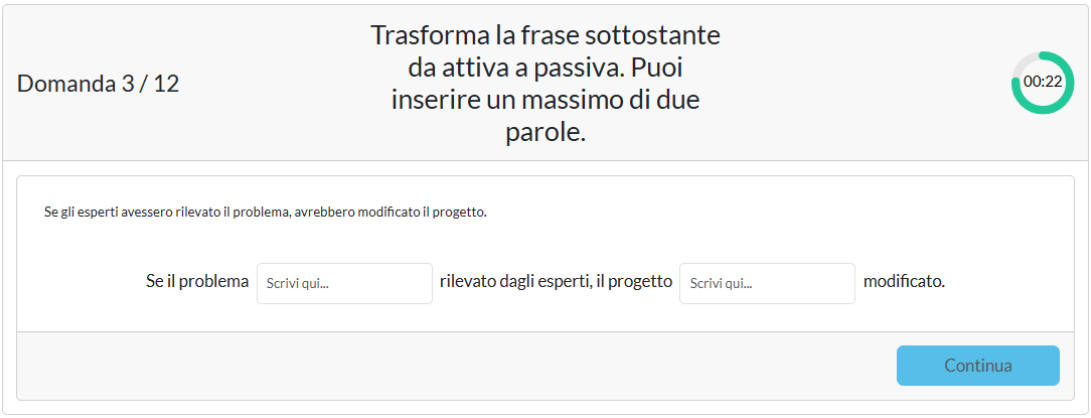


Figura 24: Esempio di task gap-filling utilizzato per testare la grammatica

La ratio progettuale relativa ai task a risposta multipla ha riguardato la scelta di adottare sistematicamente tre opzioni di risposta – di cui una corretta e due distrattori – anziché i formati a quattro o cinque opzioni. Tale decisione riflette evidenze empiriche che dimostrano come tre opzioni rappresentino il punto ottimale di equilibrio tra controllo del *guessing* ed efficienza psicometrica. La probabilità teorica suggerisce che aumentando il numero di opzioni si riduca proporzionalmente la probabilità di risposta casuale corretta: con due opzioni la probabilità teorica di indovinare è del 50%, con tre opzioni scende al 33%, con quattro al 25% e con cinque al 20%. Tuttavia, questa logica presuppone che tutti i distrattori siano ben costruiti ed egualmente plausibili per l'utente. Rodriguez (2005), in una meta-analisi comparativa di studi su test con tre, quattro e cinque opzioni, conclude che la differenza teorica tra tre e quattro opzioni non si traduce in miglioramenti significativi dell'affidabilità del test quando i distrattori sono ben costruiti. Ulteriori analisi empiriche condotte sulle distribuzioni di risposta nei test a risposta multipla rivelano che il numero medio di distrattori effettivamente funzionanti è spesso inferiore a due: candidati anche con competenza limitata scartano facilmente opzioni implausibili. Shizuka et al. (2006), in uno studio sperimentale condotto su test di inglese come L2, hanno evidenziato come le risposte degli utenti si distribuissero in media su circa 2,6

opzioni, indipendentemente dal numero nominale presentato, confermando che la riduzione teorica del *guessing* è largamente illusoria.

La decisione di utilizzare tre opzioni allinea lo strumento con le raccomandazioni *evidence-based* riportate dalla letteratura contemporanea sul *test design*. Tale scelta ha permesso di risparmiare tempo ed energie che sono state invece impiegate per assicurare che i due distrattori presenti costituissero effettivamente errori plausibili caratteristici di candidati al livello target, piuttosto che riempitivi evidentemente implausibili inseriti per raggiungere un numero arbitrario di opzioni.

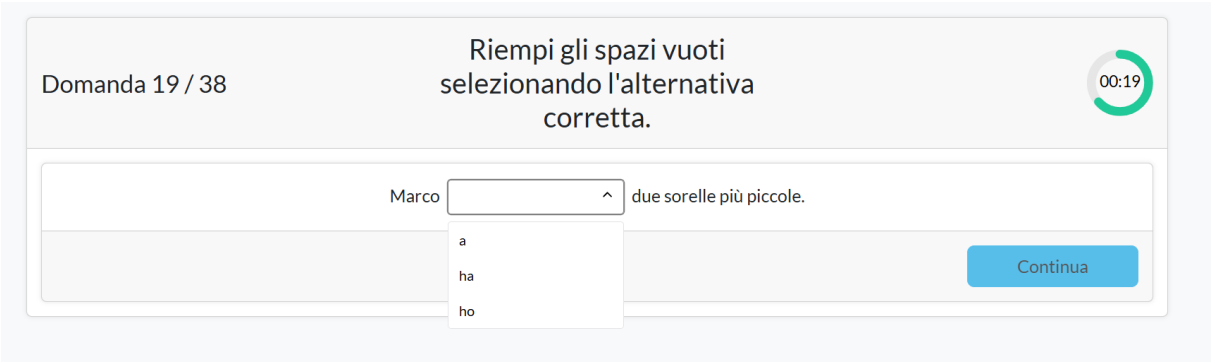


Figura 25:Esempio di task multiple choice utilizzato per testare la grammatica

3.3.2.2 Valutare la lettura

La sezione dedicata alla valutazione della comprensione scritta mira a sollecitare processi mentali che lavorano a diversi livelli di elaborazione e che concernono sia operazioni locali su singole frasi sia la costruzione di modelli mentali coerenti dell'intero testo. È prassi nel campo del LT scomporre l'abilità di lettura in lettura globale e lettura locale. A tal proposito Weir spiega:

> «La comprensione globale si riferisce alla comprensione di proposizioni oltre il livello di microstruttura, ovvero qualsiasi macro-proposizione nella macrostruttura, comprese le idee principali e i dettagli importanti. La comprensione locale si riferisce alla comprensione di proposizioni a livello di microstruttura, ovvero il significato di elementi lessicali, riferimenti pronominali, ecc.» (2005: 89; traduzione mia)[40].

[40] «Global comprehension refers to the understanding of propositions beyond the level of microstructure, that is, any macro-propositions in the macrostructure, including main ideas and important details. Local comprehension refers to the understanding of propositions at the level of microstructure, i.e., the meaning of lexical items, pronominal reference, etc.».

Urquhart e Weir (1998) associano questa bipartizione ai concetti di lettura attenta e lettura rapida, teorizzando una matrice (rappresentata in Tabella 8) che combina i diversi processi che si possono realizzare.

#	Global Level	Local Level
Careful Reading	Establishing accurate comprehension of explicitly stated main ideas and supporting details Making propositional inferences[41].	Identifying lexis Understanding syntax
Expeditious Reading	Skimming quickly to establish discourse topic and main ideas, or structure of text, or relevance to needs. Search reading to locate quickly and understand information relevant to predetermined needs[42].	Scanning to locate specific points of information.

Tabella 8:Tipi di lettura (Weir 2005: 90)

Nel caso della lettura attenta locale il lettore elabora sequenzialmente il testo costruendo il significato proposizionale a livello di frase. Questo processo richiede decodifica ortografica, accesso lessicale, *parsing* sintattico e integrazione semantica. È il tipo di lettura caratteristico dell'apprendimento dei testi, in cui il lettore cerca di processare la maggior parte delle informazioni presenti. La tipologia della lettura attenta globale spinge invece l'utente a costruire una rappresentazione coerente dell'intero testo, integrando informazioni tra frasi e paragrafi e identificando la macrostruttura: il tema principale, lo sviluppo argomentativo e l'eventuale posizione dell'autore. Passando ai casi di lettura rapida si può citare la lettura di ricerca, in cui il lettore individua informazioni specifiche su argomenti predeterminati (nomi, date, cifre, dati particolari) senza necessariamente seguire la linearità del testo. Il processo è selettivo: porzioni del testo vengono ignorate se non contengono le informazioni richieste. Infine, nel caso dello *skimming*, il lettore costruisce rapidamente una comprensione del tema generale e della struttura

[41] Stabilire una comprensione accurata delle idee principali esplicitamente espresse e dei dettagli di supporto. Fare inferenze proposizionali.

[42] Scorrere rapidamente per stabilire l'argomento del discorso e le idee principali, la struttura del testo o la pertinenza alle esigenze. Cercare nel testo al fine di individuare rapidamente e comprendere le informazioni rilevanti per esigenze predeterminate.

del testo campionando selettivamente sezioni ritenute rilevanti ed evitando dettagli. L'obiettivo è cogliere "il succo del discorso" investendo il minimo sforzo cognitivo.

Le diverse modalità di lettura teorizzate negli studi relativi alla validità cognitiva hanno guidato la creazione e selezione degli item, cercando di aderire quanto più ai principi di validità contestuale. La tipologia di *task* a scelta multipla è stata utilizzata per valutare sia la comprensione locale sia quella globale. Per i *task* di lettura attenta, gli *item* sono stati ordinati secondo la sequenza lineare delle informazioni nel testo. Come argomenta Hughes (2003), tale allineamento è coerente con il processo cumulativo di costruzione della rappresentazione mentale e riduce varianza casuale che comprometterebbe l'affidabilità. Per *task* di lettura rapida l'ordine sequenziale risulta meno vincolante poiché il processo stesso non è lineare. La Figura 26 mostra un esempio di lettura attenta testata tramite *task multiple choice*.

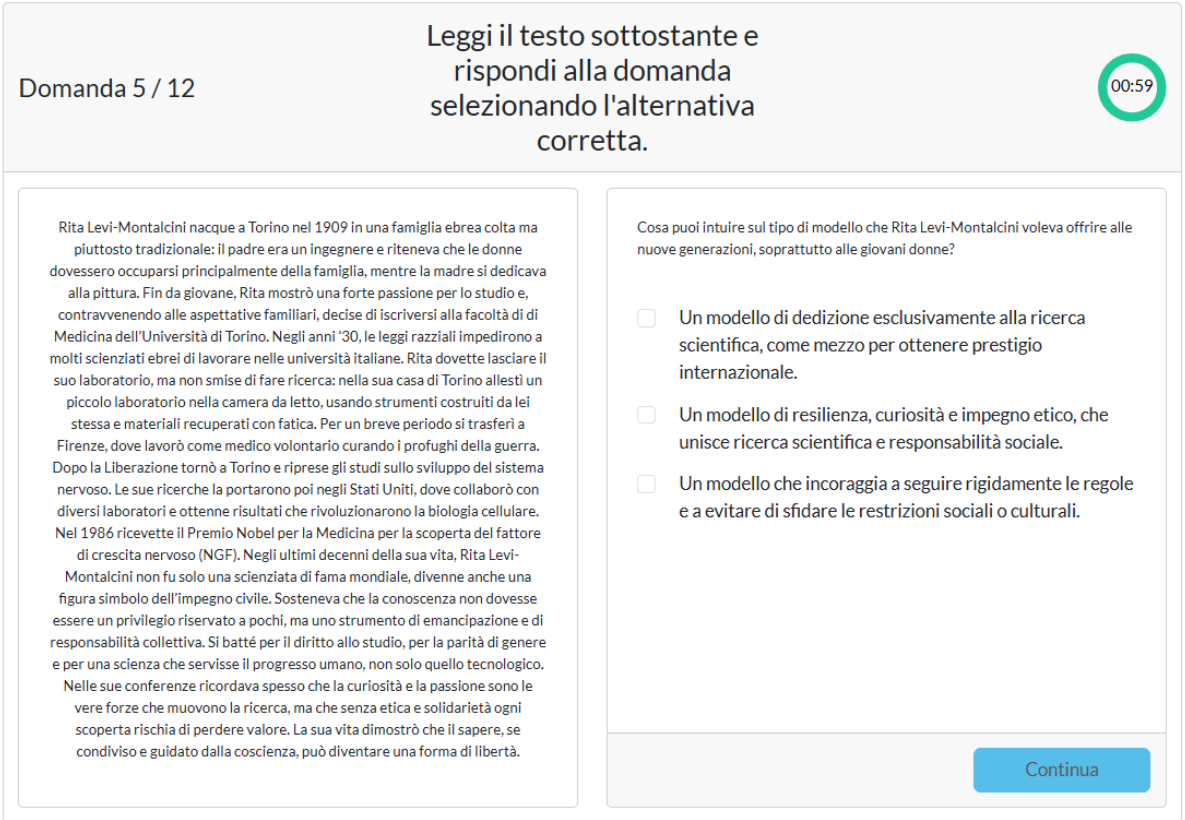


Figura 26:sempio di item che mira a sollecitare la lettura globale attenta ed in particolare la lettura atta a cogliere il succo del discorso

Per la valutazione dei livelli più bassi si è invece scelto di fare affidamento alla tipologia di scelta multipla visuale (*visual multiple choice*): tale decisione è dovuta al fatto che la riduzione del carico linguistico presente nell'*item* permette agli apprendenti di focalizzarsi principalmente

sul testo in input e sollecitare principalmente elaborazione referenziale locale. La Figura 27 mostra un esempio di domanda a scelta multipla visuale.

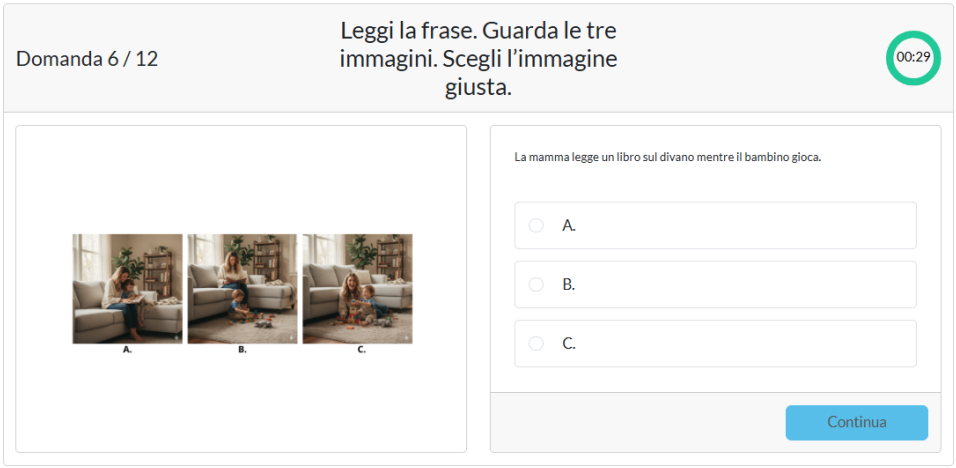


Figura 27:Esempio di task visual multiple choice

Un'altra tipologia di domanda è rappresentata dal *cloze* guidato, nel quale viene fornito un testo di riferimento a cui sono state eliminate alcune parole. L'utente – come si può vedere in Figura 28 – ha a disposizione un menù a tendina che presenta tre possibilità di risposta, di cui una sola risulta corretta.

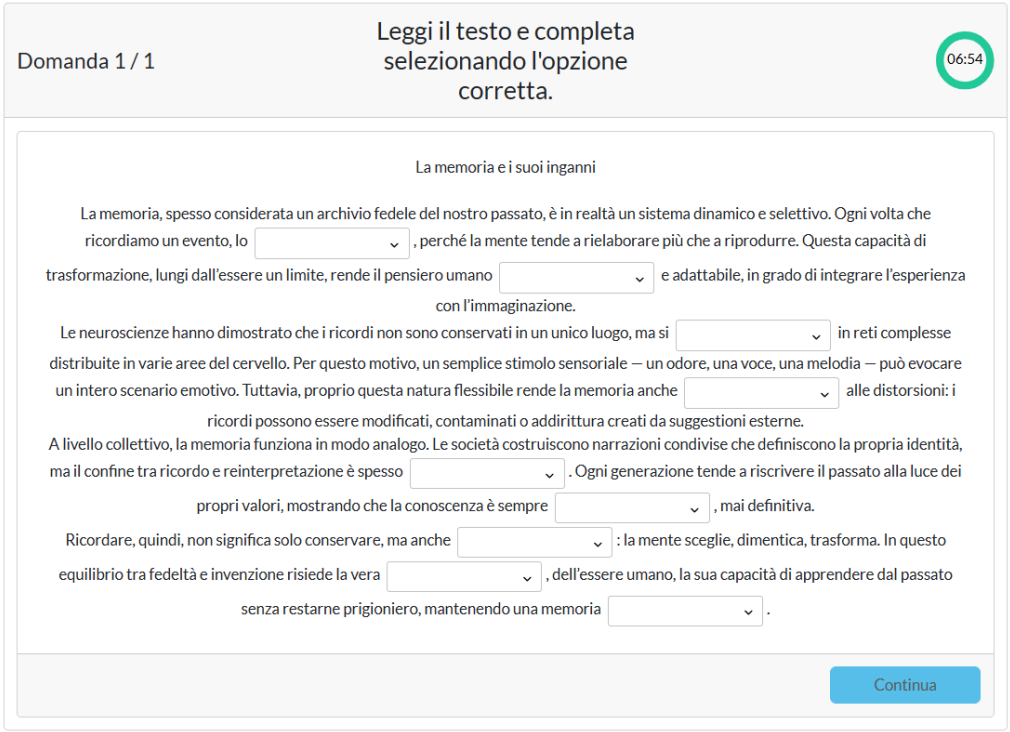


Figura 28:Esempio di domanda atta a verificare le capacità di lettura attraverso la tipologia del cloze

Ulteriore task utilizzato per verificare la capacità di lettura è la tipologia di domanda a collegamento (rappresentata dalla Figura 29), nella quale l'utente deve unire una parola o un concetto con la relativa descrizione o definizione.

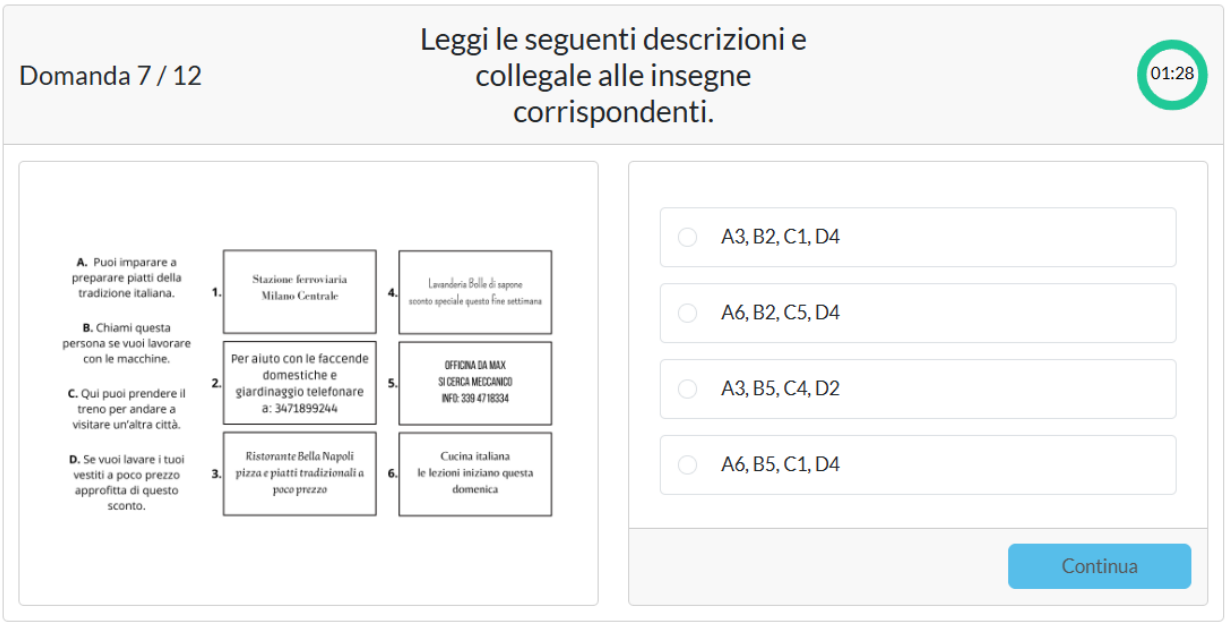


Figura 29:Esempio di tipologia di domanda a collegamento

Infine, l'ultima tipologia di *task* utilizzata nella sezione di *reading* è quella del riordinamento, esemplificata dalla Figura 30. All'utente vengono forniti diversi paragrafi che compongono un testo ma sono presentati in ordine sparso; obiettivo dell'utente sarà quello di organizzare il testo in modo tale da garantirne logicità, coesione e coerenza interna. Questa tipologia, data la difficoltà e la quantità di input linguistico da elaborare, è stata riservata per il livello C2.

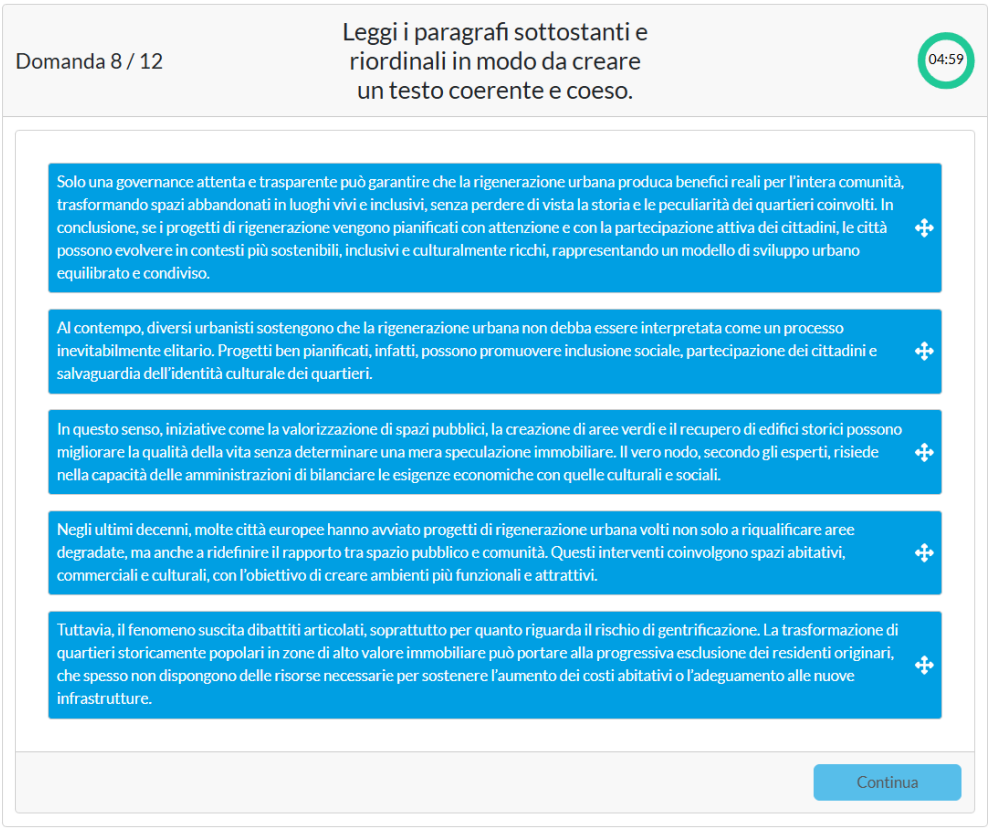


Figura 30:Esempio di domanda atta a verificare le capacità di lettura globale tramite il task di riordinamento

3.3.2.3 Valutare l'ascolto

La sezione relativa alla comprensione orale mira a valutare la capacità dell'utente di elaborare il linguaggio parlato in tempo reale. Tale processo risulta cognitivamente più complesso rispetto alla lettura a causa dell'evanescenza dell'input, del fatto che la velocità di elocuzione viene imposta dall'esterno, che le pause e la segmentazione prosodica possono non coincidere con i confini sintattici e che numerosi fenomeni fonetici possono interferire aumentando la complessità del riconoscimento lessicale. Inoltre, concorrono ad aumentare le richieste cognitive imposte al candidato le caratteristiche proprie della registrazione, le tipologie di *task* utilizzate e la forma degli *item*, che si presentano in una modalità diversa dal costrutto testato. Nel loro insieme, tali elementi rappresentano una sfida significativa anche per gli sviluppatori di *test*, andando a delinearsi «probabilmente [come] la più complessa delle quattro abilità linguistiche da testare» (Field, 2013: 84).

Field scorpora l'abilità dell'ascolto in sei livelli di elaborazione. Nella fase di decodifica dell'input l'ascoltatore converte i segnali acustici in unità sillabiche; segue la fase di ricerca lessicale in cui le forme fonologiche vengono associate alle rappresentazioni lessicali già

presenti nella mente dell'apprendente. Nella successiva fase di analisi sintattica il materiale lessicale viene messo in relazione con il co-testo in cui si inserisce. La fase di costruzione del significato comporta l'elaborazione del contenuto proposizionale attraverso l'attivazione di conoscenze enciclopediche e di meccanismi inferenziali. Infine, nella fase di costruzione del discorso, l'utente valuta la pertinenza delle informazioni appena elaborate rispetto al contesto comunicativo più ampio e ne integra il contenuto in una rappresentazione mentale coerente dell'intero evento di ascolto.

Analogamente alla comprensione scritta anche nella comprensione orale si distinguono modalità di ascolto locale, focalizzato sull'individuazione di informazioni specifiche, e globale, incentrato sulla costruzione di una comprensione complessiva. A questi si aggiungono i processi di decodifica dell'informazione attraverso il riconoscimento di elementi espliciti e di costruzione dell'informazione tramite l'inferenza. La Tabella 9 riporta le principali pratiche di ascolto teorizzate da Field (2013).

General Focus	**Definition**
Gist (G)	Listening selectively to identify the overall idea or the macro-proposition[43].
Listening for specific information (SI)	Listening selectively to identify names, dates, places, numbers, acronyms and so on[44].
Listening for important details (ID)	Listening selectively to identify words / phrases which are important in the sound file[45].
Search listening (SL)	Listening for words that are in the same semantic field. For example, the word "doctor" might bring to mind such words as "hospital", "clinic", "accident", "university", "health", "medicine" and so on[46].
Listening for main ideas and supporting details (MISD)	Listening carefully in order to understand explicitly stated main ideas and supporting details[47].

[43] Ascoltare in modo selettivo per identificare l'idea generale o la macro-proposizione.
[44] Ascoltare in modo selettivo per identificare nomi, date, luoghi, numeri, acronimi e così via.
[45] Ascoltare in modo selettivo per identificare parole/frasi importanti nel file audio.
[46] Ascoltare parole che appartengono allo stesso campo semantico. Ad esempio, la parola "dottore" potrebbe far venire in mente parole come "ospedale", "clinica", "incidente", "università", "salute", "medicina" ecc.
[47] Ascoltare attentamente per comprendere le idee principali espresse in modo esplicito e i dettagli di supporto.

Listening to infer (propositional) meaning (IPM)	Listening carefully to understand implicit meaning. For example, listening to infer the speaker's attitude towards a particular line of argument[48].

Tabella 9:Focus generale dell'ascolto

I formati di task utilizzati nella comprensione orale sono analoghi a quelli della comprensione scritta: *visual multiple choice* per testare l'ascolto degli utenti con livelli più bassi; domande a risposta multipla per valutare sia la decodifica dell'informazione sia la costruzione dell'informazione e il riempimento autonomo (cfr. Figura 21) per analizzare la capacità di cogliere dettagli importanti. Al fine di valutare la comprensione dei candidati con una competenza elevata è stata utilizzato un task di *gap filling* nel quale il file audio e l'input visivo riportavano le medesime informazioni generali, sebbene queste ultime fossero espresse, nel testo scritto, in forma parafrasata rispetto a quanto contenuto nel messaggio orale. Al candidato veniva richiesto di comprendere il significato globale del testo e di completare le frasi utilizzando le parole fornite come indizi nella consegna (cfr. Figura 31).

[48] Ascoltare attentamente per comprendere il significato implicito. Ad esempio, ascoltare per dedurre l'atteggiamento dell'oratore verso una particolare linea argomentativa.

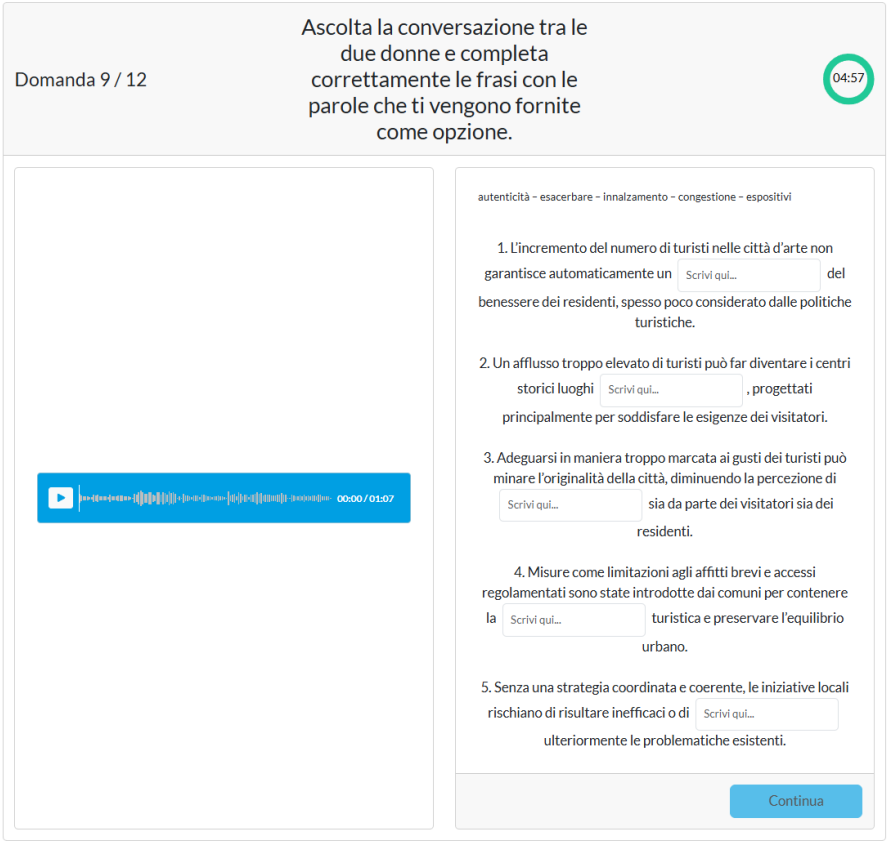


Figura 31:Esempio di task utilizzato per valutare l'ascolto globale pensato per testare i livelli di competenza più alti

Gli input forniti nei *task* di comprensione orale sono stati progettati bilanciando numerosi parametri – quali la tipologia di scambio tra i partecipanti, il *framework* di ascolto preso in indagine, il numero e il genere dei partecipanti, la durata della registrazione, la velocità di eloquio, la tipologia di compito e la complessità linguistica e tematica – al fine di garantire la più elevata presenza di validità contestuale e varietà di situazioni di ascolto. La Tabella 10 mostra la *ratio* dietro la costruzione della sezione di ascolto.

Livello	**Framework**	**Tipologia**	**Partecipanti**	**Genere**	**Durata**	**Velocità di eloquio**	**Tipologia di task**	**Domanda**	**Durata**
A1	ricerca lessicale	monologo	1	M	breve	lenta	visual multiple choice	IT-L-A1-1: comprensione lessicale [lavanderia a bolle di sapone]	25 sec

A2	comprensione globale	monologo	1	F	breve	lenta	riempimento guidato	IT-L-A2-9: comprensione globale [giornata al mare]	45 sec
A2	Analisi sintattica	dialogo	2	M e F	breve	lenta	multiple choice	IT-L-A2-1: comprensione [cosa per cena]	45 sec
…	…	…	…	…	…	…	…	…	…

Tabella 10:Esempio di blueprint utilizzato per la costruzione della sezione di ascolto

Gli audio alternano tra situazione di ascolto presentate sotto forma di monologo e dialogo poiché le due tipologie impongono richieste cognitive parzialmente differenti. In particolare, i dialoghi introducono un livello di complessità aggiuntivo rispetto ai monologhi: i candidati devono gestire sovrapposizioni occasionali tra parlanti, seguire cambi di turno e discriminare tra voci diverse. Ai livelli bassi, nei quali le esigenze cognitive legate alla normalizzazione percettiva di voci multiple rappresentano un carico significativo, è stato garantito che un solo parlante trasmettesse le informazioni essenziali per rispondere agli *item*, permettendo ai candidati di focalizzarsi su una voce principale pur essendo esposti alla struttura dialogica. Al contrario, nei livelli di competenza più elevati, le informazioni rilevanti sono state distribuite tra entrambi gli interlocutori, richiedendo monitoraggio simultaneo di voci multiple. Infine, nei casi in cui fossero presenti due interlocutori, ai livelli più bassi è stato garantito che le voci risultassero chiaramente distinguibili attraverso la combinazione di parlanti di genere maschile e femminile. Nei livelli più elevati, invece, sono state introdotte anche interazioni tra parlanti dello stesso genere; in questi casi, sono stati selezionati parlanti caratterizzati da una differenziazione evidente nella frequenza fondamentale, al fine di preservare la discriminabilità vocale.

Il ritmo del parlato costituisce un parametro cruciale nella determinazione del livello di difficoltà degli item di ascolto. Come documentato da Field (2013), la velocità di eloquio influenza direttamente i processi di segmentazione lessicale e accesso al lessico mentale: un parlato troppo rapido può impedire ai candidati con una competenza bassa di identificare confini tra parole o attivare rappresentazioni lessicali prima che un nuovo input arrivi, causando sovraccarico cognitivo. Nel *test*, la velocità adottata nei materiali audio è stata calibrata in modo progressivo in relazione all'avanzare dei livelli di competenza. Coerentemente con tale

impostazione, anche la durata delle registrazioni associate agli *item* è aumentata in maniera sistematica: nei livelli più bassi gli audio possono avere una durata di circa dieci, venti o trenta secondi, laddove nei livelli più elevati questi possono raggiungere anche i due minuti e mezzo. Questa progressione riflette vincoli cognitivi ampiamente documentati in letteratura: ascoltare audio in L2 impone considerevole carico alla memoria di lavoro, specialmente per ascoltatori con esposizione limitata alla lingua *target*.

Una caratteristica inevitabile dei *test* di ascolto somministrati tramite piattaforma digitale è l'assenza del supporto visivo che sarebbe invece disponibile in molti contesti comunicativi reali. Per far fronte a questa limitazione è stato deciso di permettere agli utenti di riascoltare gli audio presenti nelle domande. Questa scelta costituisce un compromesso consapevole tra autenticità assoluta ed equità.

3.3.2.3.1 Sintesi vocale

In linea con l'obiettivo di automatizzazione completa del test, anche gli input audio presenti nell'esame sono stati realizzati interamente attraverso tecnologie di sintesi vocale *text-to-speech* (TTS). Questa scelta ha permesso di eliminare la necessità di organizzare sessioni di registrazioni con parlanti nativi, metodologia estremamente dispendiosa in termine di tempo e denaro. Allo stesso modo, è stato possibile attuare un controllo totale sulle caratteristiche acustiche degli audio, progettando e calibrando ogni item acustico per ciascun livello QCER; così facendo è stata garantita la riproducibilità e la scalabilità nella produzione dei materiali.

Le tracce sono state prodotte utilizzando ElevenLabs, una piattaforma che utilizza sistemi di sintesi vocale basati su architetture di *deep learning*. In particolare, ElevenLabs impiega modelli di sintesi *end-to-end* che integrano componenti di rappresentazione testuale, modellazione prosodica e generazione del segnale acustico; questi consentono la produzione di *output* vocali caratterizzati da un'elevata naturalezza timbrica. Nella piattaforma sono disponibili diversi modelli TTS, differenziati per la copertura linguistica, la qualità espressiva e la robustezza *cross-lingua*; questi includono sia modelli progettati per gestire input monolingue in inglese sia modelli multilingue, che mantengono coerenza fonetica e stabilità prosodica. Per il presente lavoro è stato selezionato il modello *Eleven Multilingual v2*, scelto per la maggiore stabilità nella resa fonetica dell'italiano. La Figura 33 rappresenta la schermata di lavoro della piattaforma ElevenLabs.

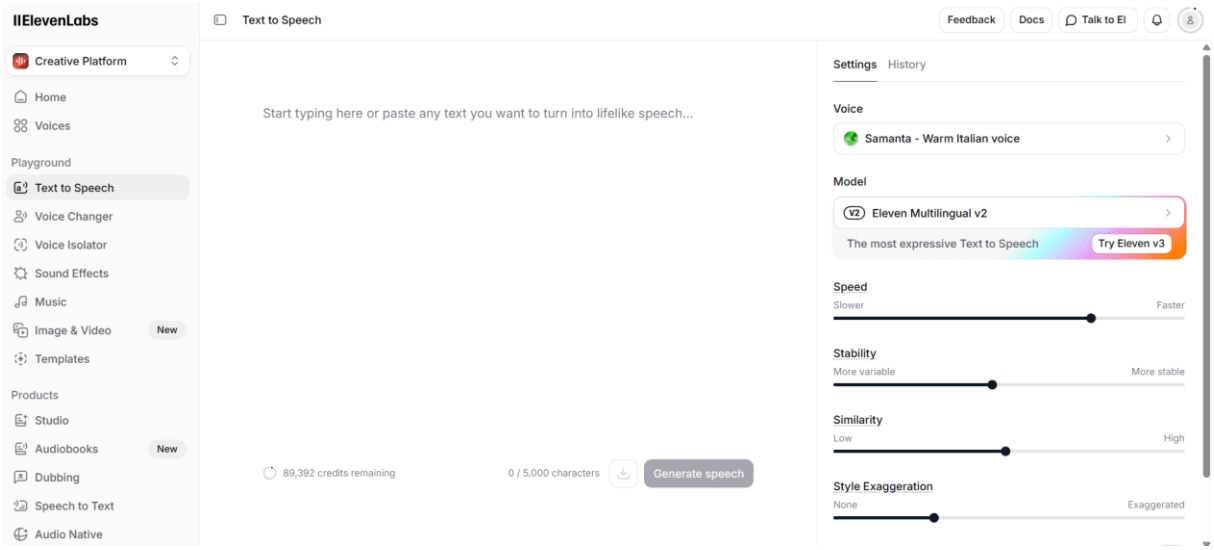


Figura 32:Schermata operativa di ElevenLabs. A sinistra si trovano i tool offerti dalla piattaforma mentre a destra è presente la libreria contenente i modelli disponibili, le voci utilizzabili e i parametri da settare.

Il processo di generazione degli audio ha seguito una procedura standardizzata articolata in quattro fasi. In un primo momento sono stati realizzati i contenuti testuali degli audio attraverso le tecnologie generative. Agli LLM sono stati forniti prompt strutturati che specificavano il livello QCER target, l'argomento ed eventuali informazioni a riguardo, il genere discorsivo atteso, la lunghezza approssimativa richiesta e specifici vincoli linguistici. Una volta ottenuta la trascrizione dell'audio, questa è stata inserita nell'interfaccia di ElevenLabs e dalla libreria sono state scelte le voci pre-addestrate che si intendevano utilizzare, selezionando quelle che garantivano maggiore naturalezza prosodica. Successivamente, sono stati settati i parametri della voce calibrandone la velocità (*speed*), la stabilità (*stability*), la somiglianza (*similarity*) e l'esagerazione stilistica (*style exaggeration*). Nello specifico, la velocità controlla il ritmo dell'elocuzione; come analizzato in precedenza il parametro è stato calibrato in relazione al livello: valori inferiori a 1.0 (considerati come rallentamenti) sono stati utilizzati per i livelli A1 e A2, laddove valori prossimi e superiori al 1.10 sono stati attuati per i livelli C1 e C2. La stabilità garantisce invece la consistenza prosodica e tonale della voce generata. Valori elevati producono parlato più uniforme e prevedibile; valori bassi introducono maggiore variabilità prosodica che può risultare più naturale ma allo stesso tempo meno controllata. Per l'operazionalizzazione relativa ad ETET sono stati utilizzati valori intermedi al fine di bilanciare naturalezza e chiarezza. Infine, la similarità valuta quanto fedelmente il modello riproduce le caratteristiche della voce target mentre l'esagerazione stilistica controlla l'intensità

delle peculiarità espressive e prosodiche. Questi due parametri sono stati calibrati a seconda delle caratteristiche intrinseche presentate dalle diverse voci.

Per quanto riguarda la realizzazione degli audio dialogici, ciascuna battuta è stata generata separatamente utilizzando la voce selezionata, producendo file audio distinti per ciascun turno conversazionale. I file individuali sono stati successivamente assemblati in una traccia unica utilizzando *Audacity*, un software di elaborazione audio. Questo ha permesso di controllare precisamente la temporizzazione tra turni, introducendo – per esempio per gli audio generati per i livelli inferiori – la possibilità di distanziare maggiormente le battute.

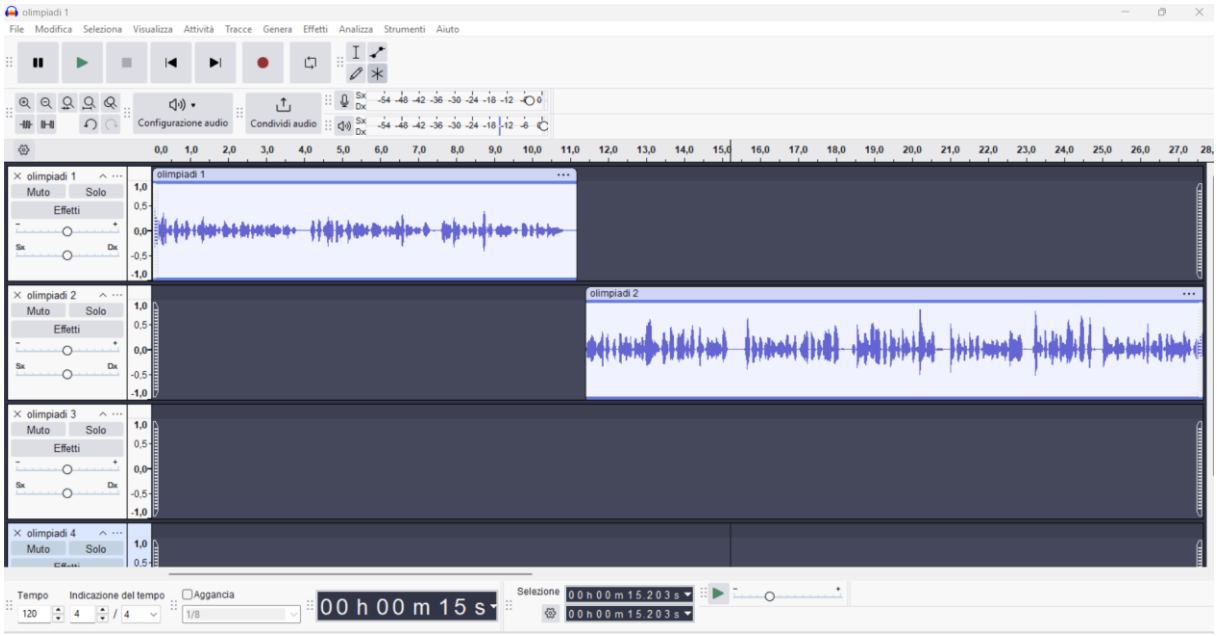


Figura 33:Interfaccia del software Audacity utilizzato per l'assemblaggio degli audio

L'automatizzazione completa del processo ha offerto molteplici benefici: in primis ha permesso di ridurre le tempistiche e il dispendio economico che sarebbero stati insiti nelle sessioni di registrazione con parlanti nativi; secondariamente, la possibilità di settare i parametri ha permesso una massima precisione nella calibrazione acustica per ciascun livello; tutti i parametri utilizzati sono stati documentati sistematicamente, permettendo di replicare esattamente le condizioni di generazione per produrre materiali addizionali o forme parallele con caratteristiche acustiche identiche. Nonostante i numerosi vantaggi non si possono tacere le limitazioni presenti in questo approccio. La produzione automatica in lingua italiana non tiene conto delle numerose variazioni diatopiche caratteristiche della situazione linguistica della penisola. Questa uniformità elimina una dimensione della competenza di ascolto

potenzialmente rilevante, specialmente nei livelli C1 e C2, dove i descrittori QCER indicano che i candidati dovrebbero essere capaci di comprendere «un'ampia gamma di materiale registrato o trasmesso via radio, anche se qualche parte è in lingua non standard» (Concilio d'Europa 2020:55). Inoltre, nonostante le tecnologie TTS siano in costante aggiornamento, le sintesi neurali attuali non sono ancora in grado di replicare fedelmente tutte le sfumature prosodiche del parlato umano, o comunque, risulta particolarmente difficile comunicare al modello l'intonazione attesa.

In conclusione, nel bilanciamento tra autenticità e controllo sperimentale, si è optato per l'utilizzo di voci sintetiche accettando il compromesso sulla naturalezza assoluta in favore della standardizzazione e della riproducibilità.

3.3.2.4 Valutare la scrittura

La sezione relativa alla valutazione della produzione scritta mira a valutare la capacità degli utenti di generale testo nella lingua target. La scrittura in una L2 è un'abilità che richiede un'orchestrazione complessa di numerosi processi cognitivi: generazione di idee, traduzione mentale nella lingua target, selezione lessicale, codifica morfosintattica, pianificazione retorica, monitoraggio e revisione (Hayes, Flower 1980). Weigle (2002) distingue tre tipologie di elaborazione cognitiva nella scrittura: la riproduzione, l'organizzazione di informazioni note e la generazione di nuove idee.

Fanno riferimento alla prima tipologia la scrittura di informazioni già codificate linguisticamente – come il dettato – che richiedono principalmente una trascrizione ortografica accurata senza elaborazione contenutistica significativa. L'organizzazione di informazioni note riguarda invece la sistemazione e presentazione di informazioni già possedute o che vengono esplicitate nell'input fornito, richiedendo l'utilizzo di strategie di selezione e sequenziamento ma non di creazione di contenuti originali. In ultimo, la strategia di generazione di idee nuove richiede all'utente un'elaborazione profonda.

Parallelamente alla proposta di Weigle, Scardamalia e Bereiter (1987) distinguono due approcci strategici alla scrittura: uno inerente alla narrazione della conoscenza, caratteristico degli scrittori principianti in cui la pianificazione risulta limitata e la focalizzazione riguarda la generazione lineare di contenuto già presente in memoria o presentato sotto forma di input nella domanda; l'altro relativo alla trasformazione della conoscenza, tipologia adatta per gli scrittori esperti. Tale metodologia richiede consapevolezza dei problemi retorici e contenutistici.

La distinzione cognitiva proposta sopra ha guidato la creazione e calibrazione dei *task* di produzione scritta. Seguendo le indicazioni contenute in letteratura, per i livelli inferiori sono stati selezionati *task* con focus specifico che richiedono produzioni brevi focalizzate su dimensioni discrete come il lessico (come riportato in Figura 34) o la morfosintassi relativa a frasi semplici (cfr. Figura 35).

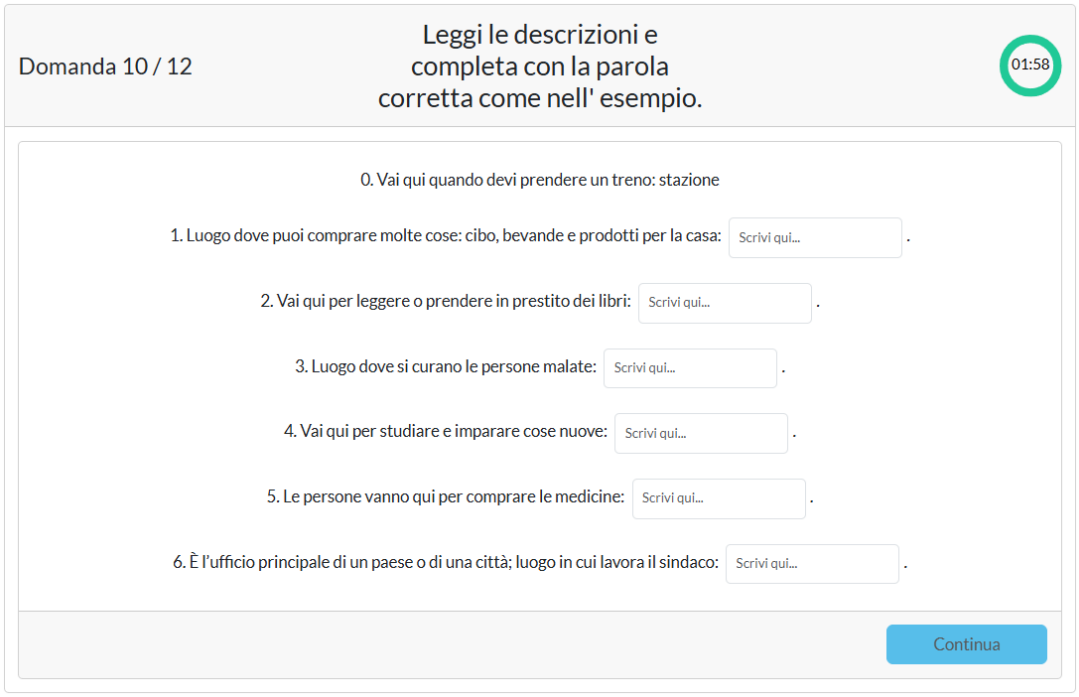


Figura 34:Esempio di task utilizzato per valutare la scrittura mirata, atto a elicitare la conoscenza lessicale

Questi task sollecitano principalmente produzione locale senza richieste retoriche complesse, appropriati per candidati che stanno ancora consolidando risorse linguistiche di base.

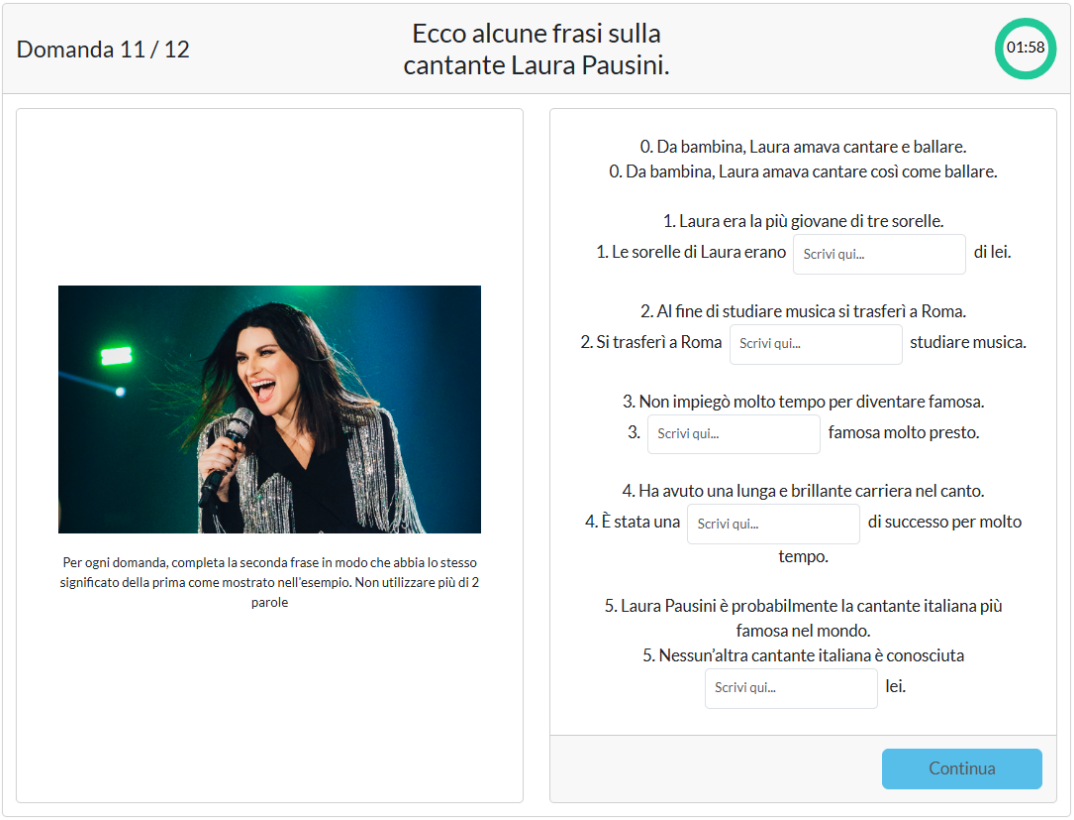


Figura 35:Esempio di task di lettura che mira ad indagare la conoscenza morfosintattica elementare

Per i candidati appartenenti ai livelli superiori sono stati progettati task che richiedevano invece una produzione linguistica estesa. Nel primo caso, tramite la risposta a una *e-mail* (cfr. Figura 22), si è cercato di testare la tipologia di narrazione della conoscenza, attraverso la creazione di un contesto chiaro grazie ai numerosi input forniti. Questa tipologia ha permesso inoltre di testare l'appropriatezza sociolinguistica, valutando la capacità dell'utente di adeguarsi alle richieste retoriche imposte dal genere testuale specifico e dalle esigenze intrinseche al registro formale, pur mantenendo contenuti relativamente concreti e personali. Per i livelli più elevati è stato invece prevista la stesura di un *opinion essay*, ovvero un testo che contemplasse la generazione di conoscenza attraverso la discussione di una tematica generale. Questa tipologia sollecita il pensiero e la valutazione critica di posizioni diverse e l'attuazione di un'organizzazione retorica persuasiva; il candidato deve operare simultaneamente nello spazio retorico e contenutistico.

3.3.2.5 Valutare il parlato

La sezione relativa alla produzione orale valuta la capacità dell'utente di generare linguaggio parlato fluente, fonologicamente comprensibile, grammaticalmente appropriato e pragmaticamente efficace in situazioni comunicative diverse. Il modello classico per descrivere il *framework* del linguaggio è quello proposto da Levelt (1989), che descrive la produzione linguistica come composta da sei fasi sequenziali (rappresentati nella Figura 36): una prima fase di concettualizzazione nella quale l'utente genera l'idea che vuole esprimere; una fase di codifica grammaticale in cui avviene la costruzione della struttura sintattica e la selezione degli elementi lessicali; le fasi di codifica fonologica e fonetica nelle quali si attua la conversione della rappresentazione astratta in sequenze di foni e l'adattamento della sequenza fonologica in istruzioni per gli organi articolatori; la fase di articolazione vera e propria che prevede la produzione dell'enunciato e, in ultima, la fase di automonitoraggio che prevede la verifica dell'accuratezza e appropriatezza di quanto enunciato.

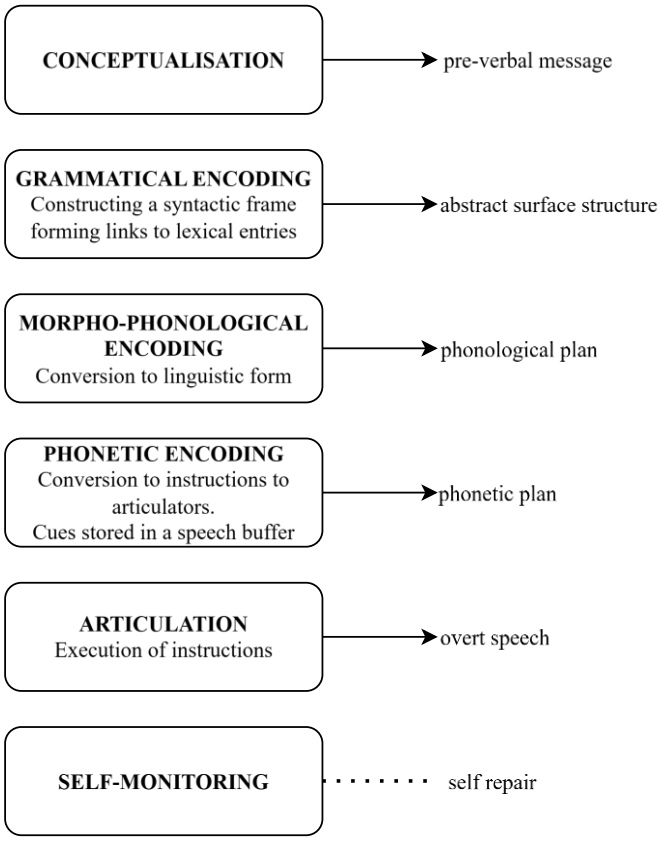


Figura 36:Adattamento del modello di Levelt relativo alla produzione orale

L'elaborazione del parlato presenta numerose sfide specifiche rispetto alle altre abilità: prima fra tutte la produzione orale necessita di un'elaborazione in tempo reale senza possibilità di revisione estensiva. Nella costruzione delle prove è necessario valutare la misura in cui può essere consentita la pianificazione; questa permette di organizzare in maniera più sofisticata il discorso ma ne riduce di conseguenza l'autenticità, poiché, nella comunicazione orale spontanea, spesso non viene concesso tempo di preparazione. Inoltre, un altro aspetto da tenere in considerazione è la natura dell'interazione che sussiste tra parlante e interlocutore: nel caso del monologo viene richiesto all'utente un alto grado di pianificazione autonoma e mantenimento prolungato del discorso; diversamente, il dialogo richiede consapevolezza e capacità di gestione dei turni e di negoziazione del significato. In ultimo, come già descritto per il *framework* relativo alla scrittura, anche il parlato può essere suddiviso in strategie di riproduzione, narrazione della conoscenza e trasformazione della conoscenza.

La sezione di parlato include tre *task* progettati per sollecitare tipologie diverse di discorso orale. Il primo esercizio riguarda la descrizione di una foto; essendo progettato per testare le competenze di utenti appartenenti ai livelli inferiori è stata selezionata un'immagine che contemplasse la presenza di una scena familiare nella quale potesse essere utilizzato un lessico domestico ma che, al tempo stesso, potesse suscitare nei livelli più elevati riflessioni più approfondite. Il *task*, essendo pensato per sollecitare una descrizione guidata, è stato corredato da input e suggerimenti di analisi. La Figura 37 rappresenta un esempio di questa tipologia di *task*.

Figura 37: Esempio di task di parlato che sollecita la produzione attraverso la descrizione guidata

Per quanto riguarda il *task* relativo alla narrazione della conoscenza è stato chiesto ai candidati di descrivere l'ultimo libro letto o film visto. Questa tipologia di domanda sollecita il linguaggio narrativo richiedendo l'uso di tempi verbali al passato, un ampio utilizzo di connettivi temporali e causali e l'impiego di un lessico astratto per esprimere emozioni e valutazioni personali.

Se queste prime tipologie di domande sollecitano produzioni monologiche il terzo *task* è stato progettato per approssimare, nei limiti tecnologici della piattaforma, la produzione dialogica interattiva. A causa delle limitazioni intrinseche di un sistema automatizzato senza esperti presenti in tempo reale, non è possibile valutare autenticamente l'interazione vera e propria – che richiederebbe adattamento reciproco dinamico, negoziazione del significato, gestione collaborativa dei turni, e *backchanneling*. Tuttavia, si è cercato di preservare parzialmente la validità cognitiva e l'autenticità contestuale della comunicazione dialogica attraverso una struttura che simula la sequenza turno-domanda-risposta caratteristica delle conversazioni reali. Concretamente, il task mostrato nella Figura 38, presenta un input audio contenente un dialogo tra due interlocutori; questo termina con una domanda esplicitamente rivolta al candidato su una tematica di interesse generale. L'utente deve dapprima comprendere la domanda posta oralmente attivando processi di comprensione orale, successivamente prendere posizione sul tema in indagine ed infine argomentare la propria visione fornendo esempi o elaborazioni a supporto. Sebbene manchi la vera reciprocità – poiché l'utente non può richiedere chiarimenti o sviluppare ulteriormente lo scambio – tale scelta progettuale costituisce un compromesso consapevole tra autenticità ideale e vincoli operativi dell'automazione.

Figura 38:Esempio di task di produzione orale atto a simulare la situazione dialogica

3.3.2.6 Considerazioni contestuali generali

La costruzione dei task per ciascuna sezione del test ha richiesto una calibrazione sistematica di molteplici parametri contestuali che influenzano direttamente la difficoltà interna degli item. Nello specifico, tale modellizzazione ha operato simultaneamente su cinque dimensioni principali: la lunghezza degli input, la complessità sintattica, le richieste lessicali, gli argomenti trattati e lo scopo comunicativo.

Un fattore significativo che contribuisce alla previsione del grado di difficoltà della domanda è la quantità di informazioni che devono essere processate. In relazione ai descrittori del QCER, si osserva infatti una progressione nella lunghezza dei testi dai livelli di competenza più bassi a quelli più elevati. In ETET tale principio è stato tradotto in un aumento sistematico dell'estensione dei testi scritti. Parallelamente, anche i materiali audio crescono in durata, passando da circa 10 secondi nei livelli più bassi fino a circa 2 minuti e mezzo nei livelli più elevati. Questa progressione riflette la capacità crescente di gestire quantità maggiori di informazioni simultaneamente, mantenere rappresentazioni mentali più complesse, e sostenere un'attenzione prolungata su input estesi.

Anche la struttura grammaticale e morfosintattica presente nei testi costituisce una dimensione predittiva della difficoltà. Se nei livelli inferiori dominano frasi brevi caratterizzate da coordinazione paratattica e subordinate esplicite elementari, nei livelli intermedi si

introducono progressivamente subordinate più complesse fino a giungere alla presenza di periodi articolati con subordinate implicite, costruzioni passive, inversioni sintattiche ed ellissi, caratteristiche dei livelli superiori. Tuttavia, come osserva Weir (2005), la relazione tra lunghezza sintattica e difficoltà non è sempre lineare: frasi brevi caratterizzate da ellissi e uso idiomatico possono risultare più difficili da elaborare rispetto a periodi più lunghi ma grammaticalmente espliciti. Per questa ragione, nei livelli avanzati sono stati inclusi anche testi con sintassi apparentemente semplice ma semanticamente dense o pragmaticamente complesse.

La densità e la tipologia di vocabolario costituiscono forse la caratteristica principale per indagare la difficoltà testuale (Nassaji 2006). In ETET, la progressione lessicale opera su tre assi complementari: frequenza, concretezza o astrattezza e polisemia e linguaggio figurato. Le istruzioni e i testi presenti nell'esame sono stati costruiti attraverso un confronto costante con le liste lessicali proposte da Spinelli e Parizzi (2010)[49]. Nello specifico, nei livelli A1 e A2 il lessico è ancorato alle parole più frequenti e familiari, nei livelli intermedi B1 e B2 vengono introdotti progressivamente termini meno frequenti, ed infine nei livelli C1 e C2 aumenta la presenza di vocabolario a bassa frequenza, includendo terminologia semi-specialistica o settoriale. Inoltre, se ai livelli inferiori dominano termini con referenti concreti, ai livelli intermedi si introducono concetti più astratti ma ancora legati ad esperienze comuni, è nei livelli superiori che inizia a comparire nei testi lessico altamente astratto con significati meno ancorati all'esperienza diretta. Infine, è caratteristico della valutazione delle fasce più basse l'utilizzo delle parole nel loro significato principale, mentre con il progredire dei livelli aumenta sistematicamente la presenza di usi ambigui, metaforici, di espressioni idiomatiche e di collocazioni.

Ulteriore parametro indicativo del livello di difficoltà della domanda è la scelta dell'argomento e del tema trattato nell'item. La selezione tematica riflette infatti la progressione dei domini presente nei descrittori QCER. Nei livelli A1 e A2 gli argomenti sono circoscritti a situazioni personali immediate e quotidiane come le presentazioni personali, la famiglia, la casa e la descrizione di routine giornaliere. Nel livello B1 la gamma di situazioni si amplia includendo argomenti generali di interesse comune ma ancora legati all'esperienza personale: viaggi, tempo libero, lavoro, eventi passati. Nel livello B2 i candidati affrontano un'ampia varietà di aree tematiche con rilevanza globale come ambiente, tecnologia, tendenze sociali, salute, educazione, ma la trattazione non è ancora specialistica. Infine, nei livelli C1 e C2

[49] Consultabili anche in formato digitale al seguente link: https://www.unistrapg.it/profilo_lingua_italiana/site/liste_lessicali_a1.html

emergono argomenti progressivamente più tecnici e settoriali, astratti o accademici: dibattiti etici, analisi culturali, questioni scientifiche divulgative e argomentazioni complesse. Allo stesso tempo però, anche ai livelli alti si mantiene il principio di indipendenza da conoscenze specifiche: gli argomenti trattati non presuppongono conoscenze specialistiche pregresse o familiarità con riferimenti culturalmente specifici che penalizzerebbero candidati con *background* diversi.

Seguendo la tassonomia delle funzioni del linguaggio proposta originariamente da Jakobson (1960), i testi sono stati selezionati per rappresentare scopi comunicativi progressivamente diversificati. Ai livelli A1, A2 e B1 predominano testi con una funzione referenziale, con l'obiettivo di informare e descrivere fatti concreti; dal livello B1 si introducono progressivamente testi con funzione emotiva per veicolare la finalità di trasmettere sentimenti o stati d'animo; nei livelli B2 e C1 emergono testi con funzione poetica per esprimere la finalità di intrattenere e veicolare un uso estetico del linguaggio; infine, nei livelli C1 e C2 acquistano rilevanza testi con funzione conativa che svolgono la funzione di persuadere o trasmettere ironia. Questa progressione riflette l'ipotesi secondo la quale la comprensione di testi con scopi comunicativi complessi richieda competenze pragmatiche più avanzate rispetto alla comprensione di testi puramente informativi.

3.3.2.7 Sviluppo dei sistemi di scoring

Lo *scoring* costituisce il processo attraverso cui le *performance* osservabili nei task vengono tradotte in punteggi quantitativi interpretabili. Come discusso nella sezione sulla validità di scoring, la qualità di questo processo influenza direttamente l'affidabilità dei risultati del test. Il sistema di scoring di ETET utilizza approcci differenti a seconda delle tipologie di *item* in indagine: un confronto con chiavi di correzione per le risposte chiuse e una valutazione generata tramite LLM per le risposte aperte.

Nelle sezioni di grammatica, lettura e ascolto sono stati impiegati principalmente *item* a risposta selezionata – come il vero e falso, il *cloze*, le *multiple choice* e le *visual multiple choice* – o item a risposta vincolata – come il *gap-filling* – nei quali il processo di valutazione opera attraverso *matching* algoritmico con chiavi di correzione predefinite. Nel caso degli *item* a scelta binaria o multipla viene attuato uno *scoring* dicotomico: la corrispondenza tra l'opzione selezionata dal candidato e quella prevista dalla chiave comporta l'assegnazione di 1 punto;

diversamente il punteggio assegnato risulta pari a 0. Nel caso degli item di *gap-filling*, la valutazione avviene attraverso il confronto tra la stringa inserita dal candidato e la stringa presente nella chiave di correzione. In questo contesto, è possibile prevedere sia una singola risposta corretta sia un insieme di risposte accettabili. Tale scelta è motivata dalla possibilità che esistano più soluzioni linguisticamente corrette, ad esempio in presenza di sinonimi o di forme flesse equivalenti, nonché dalla volontà di non penalizzare eccessivamente i candidati di livello più basso per errori ortografici quando il focus dell'esercizio è la verifica della conoscenza linguistica generale.

Questo approccio garantisce oggettività ed efficienza computazionale, pur presentando come principale limite una certa rigidità del sistema: risposte non previste dallo sviluppatore ma potenzialmente accettabili vengono valutate dal sistema come errate. Al fine di mitigare tale criticità, è stata condotta una fase di *piloting* preliminare, che ha portato ad un ampliamento e a una revisione delle chiavi di correzione.

Le domande aperte relative alle abilità di produzione scritta e orale richiedono una valutazione multidimensionale che consideri simultaneamente la correttezza morfosintattica, la ricchezza lessicale, l'organizzazione testuale, la coerenza contenutistica e l'appropriatezza pragmatica; nel caso della valutazione del parlato, a questi parametri si aggiungono anche il giudizio sull'accuratezza e sulla fluenza. Questo tipo di valutazione è stata tradizionalmente affidata al giudizio di valutatori umani esperti, introducendo costi e tempistiche di correzione elevati e potenziali problemi legati alla soggettività di tale scoring. La web-app ETET ovvia a questi problemi attraverso l'utilizzo innovativo degli LLM come valutatori.

La valutazione delle risposte tramite LLM avviene attraverso la definizione di un valutatore parametrizzato, costituito da un *prompt* strutturato e da un insieme di criteri di valutazione associati a pesi espliciti. Il *prompt* è stato sviluppato e progressivamente ottimizzato mediante una serie di analisi comparative, durante le quali sono state sperimentate diverse strategie di *prompt engineering*. In particolare, sono state testate tre principali tipologie di *prompt*: una formulazione che esplicitava le fasce del QCER, fornendo in input una rielaborazione dei descrittori contenuti nelle relative tabelle; un prompt di tipo descrittivo e *language-specific*, focalizzato sulle caratteristiche strutturali e funzionali dell'italiano; e un prompt a elevata libertà interpretativa, in cui al modello veniva lasciato maggiore margine decisionale nella valutazione delle risposte. L'analisi comparativa di tali approcci è stata condotta mediante la somministrazione al modello di input testuali il cui livello atteso era stato precedentemente determinato da un annotatore umano. Sono state effettuate diverse iterazioni

per ciascuna tipologia di prompt e i risultati prodotti dal modello sono stati sistematicamente registrati e successivamente confrontati con le valutazioni umane di riferimento. Le iterazioni hanno evidenziato differenze rilevanti tra i diversi approcci, in particolare in termini di stabilità dell'output, granularità della valutazione e aderenza ai costrutti linguistici oggetto di misurazione.

Il *prompt* basato sulle fasce QCER si fonda su una mappatura esplicita dei descrittori di competenza proposti dal QCER su intervalli numerici in scala 0 - 100. I principali vantaggi di tale approccio sono la trasparenza metodologica e il chiaro allineamento ai livelli certificativi ufficiali. Tuttavia, l'uso di scaglioni rigidi e di descrittori predefiniti tende a indurre il modello ad un comportamento di *pattern matching*, comportando una riduzione della sensibilità alle sfumature linguistiche reali del testo. In particolare, è stata osservata una tendenza alla sovra-standardizzazione delle valutazioni, con output uniformi e poco attenti a variazioni sottili della qualità grammaticale, lessicale o pragmatica. I descrittori generali del QCER sono infatti formulati in termini ampi e relativamente astratti, non sempre traducibili in criteri valutativi precisi. Inoltre, essendo *language-neutral*, non specificano quali fenomeni morfosintattici, lessicali o pragmatici siano caratteristici e specifici di ciascun livello inerentemente alla lingua italiana. Infine, il modello ha mostrato la tendenza a produrre valutazioni con limitata differenziazione all'interno di ciascuna fascia e, in generale, caratterizzate da *anchoring bias*, ovvero le risposte venivano collocate in una fascia intermedia anche in presenza di evidenze contrastanti.

Il *prompt* libero è caratterizzato da un numero molto limitato di istruzioni, al fine di massimizzare l'autonomia decisionale del LLM. L'approccio consente al modello di sfruttare appieno le proprie capacità inferenziali, evidenziando tuttavia le criticità ancora significative nel contesto valutativo. In assenza di vincoli espliciti sui parametri in analisi e sui criteri di assegnazione dei punteggi, le valutazioni appaiono frequentemente instabili e difficilmente replicabili. Questo comportamento riduce inevitabilmente l'affidabilità e l'utilizzabilità del sistema in un contesto di testing standardizzato, in cui la coerenza valutativa risulta un requisito fondamentale.

Infine, il *prompt* descrittivo incentrato sulle caratteristiche dell'italiano (*language-specific*), rappresenta una soluzione intermedia tra le opzioni precedentemente analizzate. Grazie al bilanciamento tra controllo e flessibilità da esso consentite, tale strategia è stata selezionata come motore valutativo adottato nel presente lavoro. Il seguente approccio non vincola il modello a soglie numeriche o descrittori di livello espliciti, ma guida l'analisi

attraverso una definizione dettagliata dei parametri linguistici rilevanti, con un *focus* attento alle specificità dell'italiano come lingua seconda. La descrizione puntuale di fenomeni grammaticali, lessicali e pragmatici consente al modello di effettuare una valutazione più fine e linguisticamente informata. Rispetto al *prompt* QCER, questo riduce la classificazione meccanica, incasellando meglio la valutazione in una scala numerica 0-100. Allo stesso tempo, rispetto al *prompt* libero, restituisce una minor variabilità dell'output e una giustificazione più solida rispetto ai diversi criteri valutativi.

In conclusione, la scelta di costruire i *prompt* seguendo un approccio descrittivo e *language-specific* risponde opportunamente all'esigenza di bilanciare rigore metodologico e capacità inferenziali del modello. Resta tuttavia aperta la questione relativa alla dipendenza dal modello specifico utilizzato e la necessità di una conoscenza profonda e specialistica di ogni lingua in indagine per produrre un *prompt* linguisticamente valido.

Il *prompt* utilizzato richiede di valutare le produzioni scritte secondo cinque parametri analitici, ciascuno analizzato separatamente su una scala 0 - 100. Figura 39 mostra la realizzazione del valutatore.

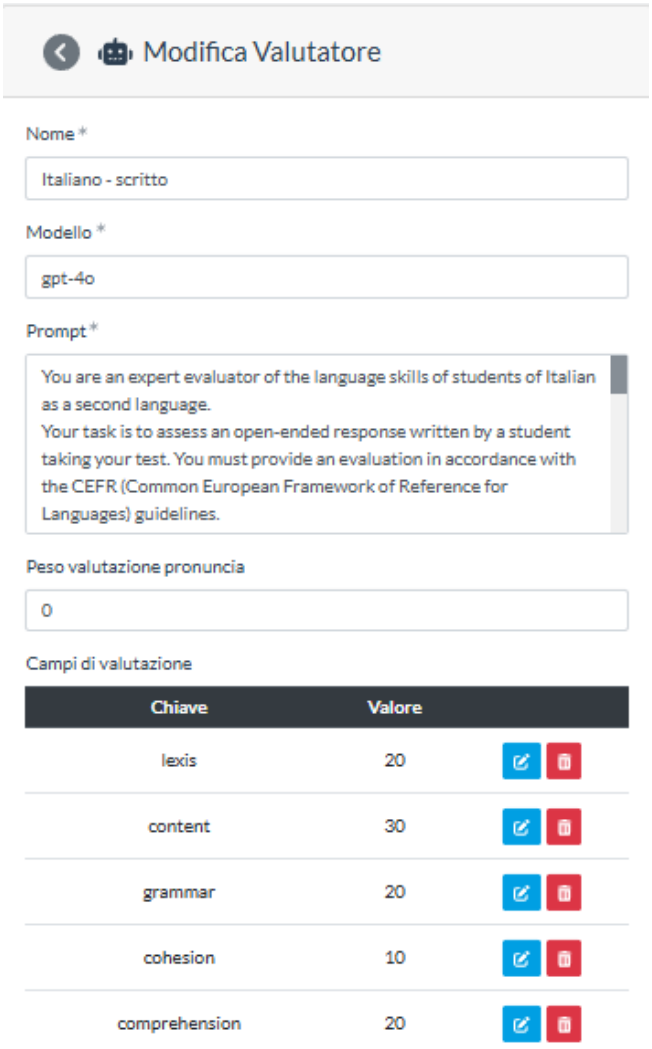


Figura 39: Interfaccia del valutatore contenente nome, modello utilizzato, prompt, parametri in indagine e relativi pesi

Il punteggio complessivo, riportato in centesimi dal modello, viene calcolato sulla base di cinque parametri di valutazione, ciascuno dei quali contribuisce in misura diversa al risultato finale: *structure and grammar* (20%), *content and argumentation* (30%), *vocabulary* (20%), *comprehension and adherence to the topic* (20%) e *pragmatics and cohesion* (10%). La distribuzione dei pesi riflette la concezione comunicativa e orientata all'azione della competenza linguistica, come teorizzata da Bachman e Palman e successivamente dal QCER. Il peso maggiore viene pertanto assegnato alla voce "*Content and Argumentation*", riconoscendo che l'obiettivo primario della produzione linguistica è comunicare significati in modo efficace. Per ciascun parametro il modello fornisce un punteggio numerico e una relativa giustificazione testuale che spiega il punteggio citando evidenze dalla produzione del candidato, come mostrato in Figura 40.

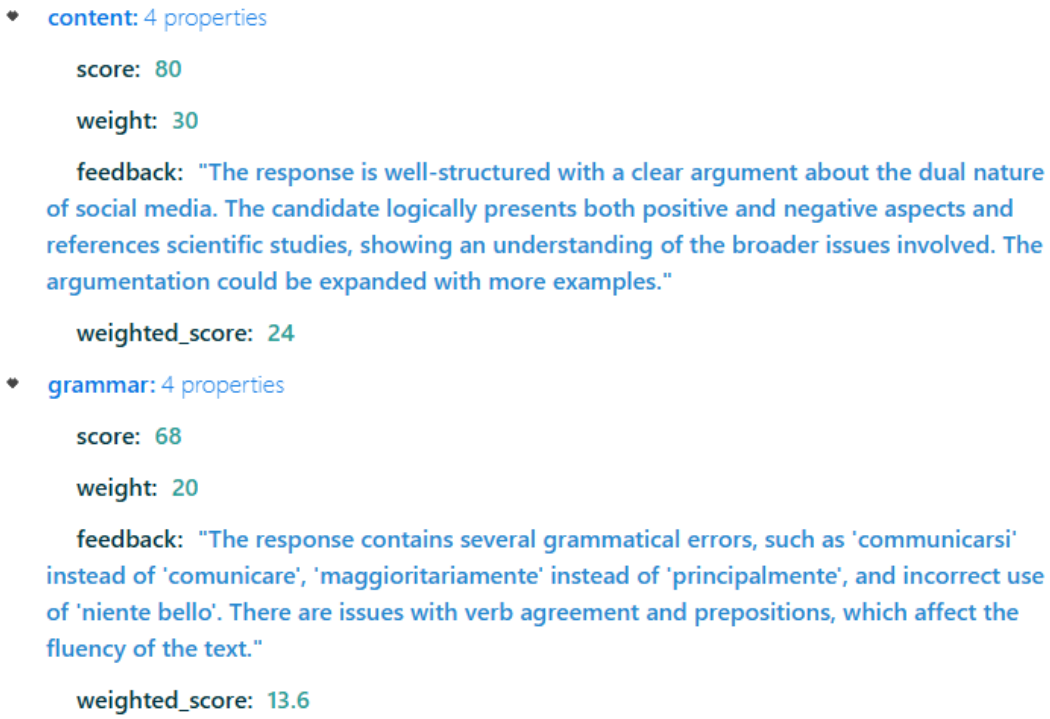


Figura 40:Esempio della valutazione di due parametri. A schermo vengono riportati il punteggio ottenuto, il peso del relativo parametro, il feedback testuale con relative evidenze e il punteggio pesato del singolo parametro

Il punteggio, riportato tramite *range* 0-100, viene successivamente mappato sui livelli QCER attraverso una scala di conversione proprietaria calibrata empiricamente (non divulgabile per ragioni commerciali).

La valutazione delle produzioni orali richiede una pipeline più complessa che integra molteplici componenti tecnologiche. Il *file* audio catturato dal sistema viene inviato all'API di Azure che effettua un'analisi acustica attraverso le dimensioni di *fluency* e *accuracy*. Parallelamente, il file audio viene trascritto attraverso il sistema di *Automatic Speech Recognition* basato su Whisper integrato in Azure. La trascrizione testuale, accompagnata dai punteggi di *fluency* e *accuracy* ottenuti da Azure, viene fornita come input a GPT-4o, incaricato della valutazione linguistica complessiva della produzione orale. I parametri di valutazione sono gli stessi utilizzati per le produzioni scritte ma con pesi ricalibrati e l'aggiunta dell'indice della valutazione della pronuncia per riflettere le specificità dell'oralità. Tabella 11 mostra la distribuzione dei pesi dei parametri nelle domande di parlato.

Chiave	Valore espresso in percentuale
Lexis	20
Content	25
Grammar	20
Cohesion	10
Comprehension	20
Pronunciation Assesment (Fluency + Accuracy)	5

Tabella 11: Parametri e pesi utilizzati nella valutazione delle domande di produzione orale

Dato che il modello è chiamato a produrre un giudizio sul parlato a partire da un riferimento scritto, è stato necessario arricchire il *prompt* esplicitando la necessità di interpretare il testo non come una produzione scritta, bensì come una trascrizione di un discorso orale, in cui appaiono inevitabilmente fenomeni tipici dell'oralità quali esitazioni, ripetizioni, autocorrezioni, riempitivi, segnali discorsivi e frasi incomplete. Questa strategia di *prompting* mira ad evitare che vengano penalizzati tratti che, nel parlato spontaneo, rappresentano caratteristiche fisiologiche e non necessariamente indici di bassa competenza.

Tale procedura non è tuttavia esente da criticità strutturali. La trascrizione testuale costituisce infatti una rappresentazione parziale dell'input orale, non preservando informazioni prosodiche fondamentali come l'intonazione, il ritmo, l'accento, la durata vocalica o eventuali variazioni di intensità. Sebbene alcune di queste informazioni vengano considerate attraverso i punteggi di *fluency* e *accuracy* forniti da Azure, la valutazione operata dell'LLM rimane prevalentemente ancorata alla modalità testuale: aspetti come la gestione dell'intonazione, la naturalezza del ritmo o la segmentazione del discorso non possono essere pienamente integrati nel processo valutativo. Secondariamente, l'intermediazione del sistema ASR introduce un ulteriore grado di complessità e distorsione: errori di trascrizione, omissioni o normalizzazioni operate dal sistema di riconoscimento vocale possono alterare la produzione linguistica del testo, influenzando la valutazione grammaticale e lessicale effettuata dal modello.

Dato il contesto, un approccio basato su un modello multimodale rappresenterebbe una soluzione più adeguata alla valutazione delle produzioni orali, almeno sul piano teorico. Un modello in grado di elaborare simultaneamente input audio e testuale potrebbe infatti integrare direttamente le informazioni acustico-prosodiche con quelle linguistiche, riducendo la perdita informativa dovuta alla trascrizione e consentendo una valutazione più fedele della competenza orale complessiva. In particolare, la disponibilità del segnale audio permetterebbe di valutare in modo più accurato aspetti quali la fluidità percepita, la gestione delle pause, l'intonazione

pragmatica e la naturalezza dell'eloquio, elementi che nel presente sistema vengono solo parzialmente catturati tramite metriche esterne.

In conclusione, la soluzione scelta rappresenta un compromesso tra fattibilità tecnica e accuratezza valutativa. Pur consentendo una valutazione automatica strutturata delle produzioni orali, si evidenziano limiti intrinseci legati alla riduzione a forma testuale del parlato. Tali criticità mostrano una chiara direzione per sviluppi futuri del sistema, orientata verso modelli pienamente multimodali per la valutazione integrata delle competenze orali.

Nel test non tutti gli item contribuiscono in egual misura alla definizione del punteggio finale. Nello specifico «la ponderazione si verifica quando un numero diverso di punti massimi viene assegnato a un elemento, compito o componente del test al fine di modificarne il valore in relazione ad altre parti del test» (Weir 2005: 63; traduzione mia)[50]. Si è scelto di implementare un sistema di pesatura differenziale – schematizzato nella Tabella 12 – che assegna un valore diverso a ciascun item in funzione di tre caratteristiche: il livello QCER associato all'item, la tempistica assegnata alla domanda e il carico cognitivo imposto dal formato di *task*. Il livello QCER conferisce un peso crescente da 1 a 3 in base alla fascia di riferimento; la durata assegna un peso progressivo in funzione del tempo a disposizione (a durata minore di 60 secondi viene associato 1 punto; a durata minore di 120 secondi corrispondono 2 punti; a durata maggiore di 120 secondi sono previsti 3 punti); la tipologia di *task* è classificata in base al carico cognitivo richiesto e può conferire da 1 a 4 punti (qualsiasi tipologia di domanda di tipo *multiple choice* riceve 1 punto; nelle domande di *gap-filling* in cui l'utente deve produrre un input linguistico minimo vengono associati 2 punti; le domande di produzione scritta ricevono 3 punti e quelle di produzione orale 4 a causa dell'elevato sforzo richiesto). A titolo esemplificativo, una domanda di livello A1 della durata di 30 secondi basata su una domanda a scelta multipla avrà un peso complessivo pari a 3 punti (1+1+1), mentre una produzione orale di livello C2 con diversi minuti a disposizione può raggiungere un peso massimo pari a 10 punti (4+3+3).

[50] «Weighting is concerned with the assignment of a different number of maximum points to a test item, task or component in order to change its relative contribution in relation to other parts of the same test».

Livello	Punti assegnati
A2	1
A2	1
B1	2
B2	2
C1	3
C2	3

Durata	Punti assegnati
< 60	1
< 120	2
> 120	3

Tipologia di task	Punti assegnati
MCQ	1
Gap-filling	2
Produzione scritta	3
Produzione orale	4

Tabella 12:Parametri che influenzano l'assegnazione del punteggio

Una volta che l'utente ha completato il test, il sistema calcola i punteggi ottenuti per ciascuna domanda sulla base delle procedure sopra descritte e restituisce un voto finale per ogni abilità linguistica. Il punteggio di ciascuna abilità è calcolato come rapporto tra i punti ottenuti e il totale dei punti realizzabili per quella specifica competenza (cfr. Formula 2).

$$score = \frac{Pt_{fatti}}{pt_{tot}}$$

(2)

Il punteggio finale del test è invece ottenuto dalla media aritmetica dei punteggi relativi alle singole abilità, in modo da uniformarne l'importanza complessiva (come riportato in Formula 3, dove *n* rappresenta il numero di abilità testate).

$$score = \frac{\sum_{i=1}^{n} 5score}{n} \%$$

(3)

Per garantire l'allineamento con i sistemi di certificazione ufficiali, il punteggio percentuale risultante viene infine mappato sulle fasce di competenza stabilite dal QCER. L'output generato dal sistema fornisce all'utente sia un punteggio numerico in percentuale sia una valutazione espressa in termini di livello QCER (da A1 a C2) per ciascuna abilità linguistica, oltre a un livello complessivo relativo all'intero test come mostrato in Figura 41.

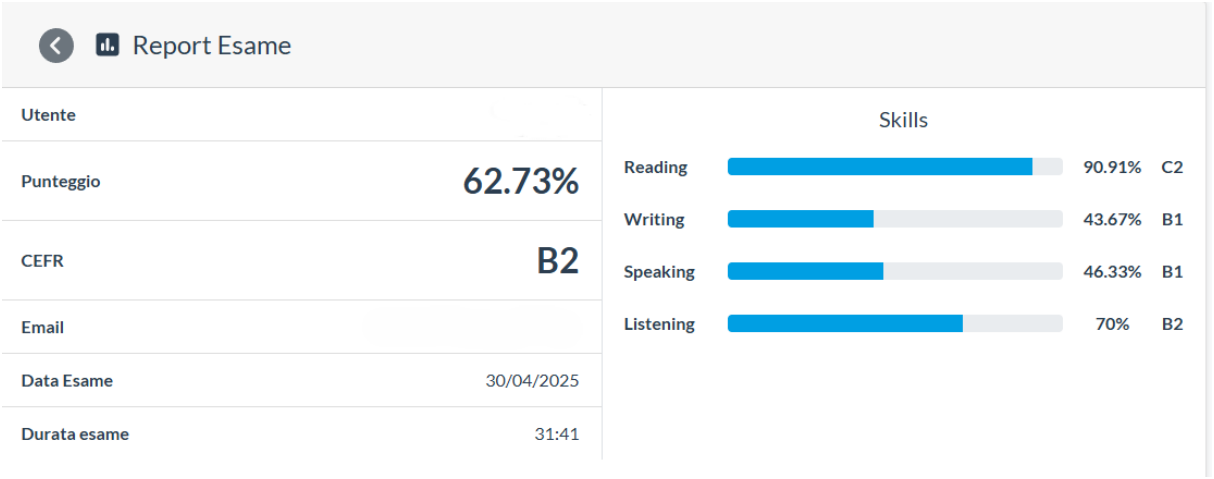


Figura 41:Report finale dell'esame, espresso sia in termini di competenza generale, sia per singola abilità

3.3.3 Somministrazione

Preliminare alla somministrazione pilota su larga scala, è stata condotta una fase di *pre-testing* su un campione ristretto di utenti al fine di raccogliere informazioni utili al perfezionamento dello strumento di valutazione. Tale processo ha permesso di individuare eventuali criticità, quali istruzioni poco chiare, compiti non adeguati o tempistiche di esecuzione non congrue.

La fase effettiva di *piloting* è stata progettata per raccogliere dati su una popolazione eterogenea, riflettendo la natura del test come strumento per valutare la *proficiency* generale. Il campionamento è avvenuto tramite la procedura definita di *snowball sampling* o a grappolo (Berruto, Cerruti 2019) nella quale, a partire dai soggetti immediatamente contattabili, ciascun informatore ha potuto arruolare i membri della propria rete sociale che rispettassero i criteri di inclusione prestabiliti. Parallelamente, sono state contattate formalmente molteplici istituzioni che operano con apprendenti di italiano L2, sia tramite *gatekeeper* (Patton 2002) – ovvero un

esperto del settore facente parte delle organizzazioni ha facilitato l'integrazione delle figure idonee – sia tramite contatto istituzionale diretto. Sono stati raggiunte scuole e istituti privati di lingua italiana per stranieri, centri linguistici di ateneo (CLA) di diverse università della penisola ed organizzazioni del terzo settore operanti in contesti multiculturali. Questa diversificazione istituzionale ha permesso di raggiungere candidati con *background* di apprendimento variabili (per esempio differenziando contesti di acquisizione e di apprendimento, e l'esposizione intensiva da quella estensiva) e motivazioni diverse (accademiche, professionali, di integrazione sociale).

Il reclutamento è avvenuto sia attraverso canali digitali – tramite la condivisione del link che riportava ad un modulo *Google Forms* contenente la spiegazione del progetto e la possibilità di adesione – sia tramite volantini cartacei con *Qrcode* distribuiti fisicamente che conducevano al medesimo modulo. Una volta espressa la volontà di partecipazione al progetto, gli utenti sono stati contattati via *e-mail* e sono stati informati dell'uso che sarebbe stato fatto dei dati raccolti, delle istruzioni tecniche necessarie per registrarsi alla piattaforma ETET e per completare il test, della necessità di compilare un questionario sociolinguistico e della possibilità di lasciare un commento finale tramite l'ulteriore questionario di *feedback*. I dati raccolti nel questionario sociolinguistico – che verranno opportunamente analizzati nel Capitolo 4 – comprendevano: una sezione relativa al profilo personale del candidato, che includeva informazioni come età, genere, livello di istruzione, professione, paese di residenza attuale, eventuale permanenza in un territorio italofono per almeno tre mesi; una sezione relativa al *background* linguistico in cui è stata indagata la lingua materna, le ulteriori lingue conosciute, l'età del primo contatto con l'italiano, gli anni totali di studio o esposizione, il tipo di acquisizione e un'autovalutazione relativa alle competenze scritte e orali; una sezione relativa ai contesti d'uso delle lingue, atta ad indagare quali lingue venivano utilizzate nella quotidianità e come si distribuissero nei vari domini sociali e la frequenza di utilizzo quotidiano della lingua italiana; in ultimo, sono state indagate le motivazioni principali che hanno portato all'apprendimento dell'italiano.

Dopo un iniziale interessamento espresso da 164 persone, solo 40 hanno effettivamente completato il test integralmente (3 utenti si sono ritirati durante il test). Dal *pool* di 40 esami completati, è stato necessario effettuare una selezione per costituire il campione oggetto di questa ricerca (selezione opportunamente descritta nel paragrafo 4.2). Tale selezione è stata motivata dalla necessità di eliminare test che riportavano difetti tecnici nella registrazione delle produzioni orali come audio non riproducibili, trascrizioni troncate, o brusii di fondo e

condizioni ambientali pessime che compromettevano l'intelligibilità e generavano trascrizioni ASR altamente inaccurate.

Parallelamente alla somministrazione del test ad apprendenti di italiano L2, sono stati coinvolti anche cinque parlanti nativi di italiano, che hanno svolto la funzione di gruppo di controllo. Ciò ha consentito di esaminare le prestazioni del modello su un campione di riferimento composto da soggetti con competenza effettivamente nativa nella lingua target.

Per condurre la valutazione manuale sono stati reclutati tre annotatori esperti[51]. Il numero selezionato risponde a standard metodologici su studi di affidabilità *inter-rater*: un minimo di due è necessario per calcolare l'accordo, ma la triplice presenza permette di identificare l'eventuale esistenza di *outlier* (se due annotatori concordano sistematicamente mentre il terzo diverge, questo segnala un potenziale problema con quell'annotatore) e calcolo di metriche più robuste. Gli annotatori selezionati possedevano le medesime caratteristiche di essere laureandi o laureati magistrali in linguistica, avere familiarità con la struttura del QCER e i sei livelli di riferimento, possedere un *background* teorico sulla *Second Language Acquisition* e di essere paralanti madrelingua di italiano. Sebbene gli annotatori non fossero valutatori professionali certificati da enti ufficiali, la loro solida formazione linguistica teorica li qualificava per condurre valutazioni informate e metodologicamente consapevoli. Come discusso nella sezione sulle risorse, questa scelta rappresenta un compromesso necessario tra rigore ideale e fattibilità effettiva in contesto di ricerca accademica con budget limitato.

Gli annotatori hanno ricevuto una formazione specifica attraverso la spiegazione degli obiettivi generali della ricerca e lo scopo specifico della validazione umana. Successivamente è stato fornito loro il prompt utilizzato da GPT-4o per la valutazione, la relativa spiegazione dei cinque parametri di valutazione e i relativi pesi. Il modello e gli annotatori umani hanno pertanto ricevuto in input le medesime istruzioni. Durante l'intero processo di valutazione è stato imposto ai valutatori di non comunicare e confrontarsi sui punteggi o sulle valutazioni assegnate. Questa procedura garantisce che l'accordo misurato rifletta una convergenza spontanea dei giudizi piuttosto che coordinazione o influenza reciproca. Nello specifico, gli annotatori non avevano accesso ai punteggi assegnati automaticamente da GPT-4o, non conoscevano le valutazioni assegnate dagli altri due annotatori umani e ricevevano le produzioni dei candidati in formato anonimizzato, al fine di prevenire *bias* inconsci (come, per

[51] Un sentito ringraziamento a tutti gli annotatori che hanno permesso, con le loro annotazioni, la creazione del gold standard, il cui contributo è stato fondamentale per la validazione dello strumento oggetto di indagine in questo lavoro.

esempio, pregiudizi inconsci verso candidati di determinate nazionalità). Per concludere, ogni valutatore ha ricevuto il prompt del task – ovvero le istruzioni originali fornite al candidato, specificando cosa dovesse produrre, con quale scopo, in quanto tempo e con quali vincoli – la produzione effettiva del candidato, ovvero il testo nel caso della produzione scritta e l'audio originale nel caso della produzione orale; in ultimo, è stata fornita la griglia di valutazione strutturata per l'analisi dei cinque parametri oggetto di valutazione. La Figura 42 rappresenta la griglia di annotazione con le informazioni fornite agli annotatori.

ID DOMANDA	CODICE DOM	DOMANDA	ALLEGATO
W1	IT-W-I-9: composizione libera guidata [servizio clienti]	Immagina di essere un addetto al servizio clienti dell'azienda. Scrivi un'email di risposta al cliente.	Da: customer@service.com Oggetto: Problemi con un ordine Gentile Servizio Clienti, Ho ordinato un paio di scarpe dal vostro sito web il 10 gennaio, ma non ho ancora ricevuto il pacco. Sul sito era indicato che la consegna sarebbe avvenuta entro cinque giorni lavorativi. Oggi è il 20 gennaio e non ho ancora ricevuto alcun aggiornamento sullo stato del mio ordine. Ho provato a contattare il servizio clienti telefonicamente, ma sono rimasto in attesa per oltre 20 minuti senza ottenere una risposta. Sono molto deluso da questa esperienza e vorrei sapere cosa sta succedendo con il mio ordine. Se non lo riceverò presto, chiederò un rimborso. Attendo una vostra risposta il prima possibile. Cordiali saluti, Matteo Bianchi
W2	IT-W-A-7: composizione libera [social media]	I social media hanno rivoluzionato il nostro modo di comunicare, di lavorare, di informarci e più in generale di vivere. Fornisci una breve opinione personale sull'argomento, evidenziandone gli aspetti positivi e negativi.	

Utente	Domanda	Risposta	Lexis	Content	Grammar	Coherence	Comprehensio
U1	W1	Gentile cliente, La ringrazio per La sua e-mail. Ci dispiace molto per l'inconveniente. Il ritardo è stato causato da un problema tecnico che abbiamo avuto in questi giorni. In effetti, per questo motivo, non siamo riusciti a contattare nessuno dei nostri clienti per avvisarli dell'imprevisto e porre le nostre scuse. Nei prossimi giorni, il Suo ordine sarà inviato all'indirizio indicato da Lei. Le offriamo un buono con lo sconto che può applicare alle prossime acquiste dei nostri prodotti. Qualsiasi dubbio abbia, ci può contattare con la chat virtuale disponibile al il nostro sito www.divanimond.it. Cordiali saluti, Marco Belloni					
U1	W2	Negli ultimi anni, la società sta vivendo i processi di digitalizzazione. In effetti, i social media rappresentano un parte importante di questo nuovo universo virtuale che si è creato. L'uso di social media ha saputo rivoluzionare il nostro modo di informarci. Grazie agli account informativi riusciamo a sapere la situazione politica e economica anche dall'altra parte del monde con un'immediatezza impresssionante. Sebbene sia uno strumento di comunicazione molto efficiente, ha diversi svantaggi. Diverse ricerche scientifiche nell'ambito della psciologia hanno rivelato che sono principalmente i social a provocare le ansie. Non è un caso che tanti adolscenti adesso si sentono sempre preoccpuati, non sono sicuri di loro e hanno tantissime paure. Inoltre, questo porta al continuo paragone che essi fanno con i coetanei					
U2	W1	Gentile Matteo Inanzitutto vorrei chiedere scuse da parte di tutta la nostra azienda, ultiammente stiamo avendo dei problemi con il sevizio di consegna gia che fa trippo freddo, e i raider ci chiedono di fornirli con dei capellini, mutande e calzini più adatti per queto momento dell'anno. Faremmo il possibile per fare si ch ele tue scarpe arrivino il prima possibile anceh se devo andare io personalmente. Scusa ancorca per il disaggio e ti ringrazio per credere in noi. Bacci e ti auguro una buona giornatta <3 Cordiali saluti.					
U2	W2	Certamente, a giorno d'oggi, big 2025, quasi 2026. Nessuno h apiù voglia di spostarsi di casa per fare niente per cui tutti utiliziamo i socia per parlare con gli amici, la famiglia, laborare o semplicemente ordinare da mangiare. Questo magari un giorno ci portera alla nostra propria auto destruzione. Sicuramente. Ma si sta molto comodi adesso per pensarcia la futuro					
		Gentile Signor Bianchi, Mi dispiace l'inconveniente accaduto. La web è entrata in riparazione la settimana scorsa, ragione per					

Figura 42: Esempio di informazioni fornite agli annotatori umani

CAPITOLO 4. ANALISI E DISCUSSIONE DEI RISULTATI

Il presente capitolo sottopone a verifica empirica i risultati ottenuti dalla sperimentazione pilota condotta sul modello ETET, con l'obiettivo di valutare la solidità delle scelte teoriche, metodologiche e tecniche illustrate nei capitoli precedenti. Dopo aver delineato il profilo della popolazione, l'attenzione si concentra sulla qualità e sull'integrità dei dati raccolti, con particolare riferimento alla sezione di produzione orale. L'elevata incidenza di malfunzionamenti tecnici ha infatti imposto una riflessione metodologica preliminare in merito alla definizione dell'unità di analisi e ai criteri di esclusione adottati per la costruzione del campione. A seguito di tale procedura, il campione effettivamente utilizzato nelle analisi si compone di 95 risposte orali valide e 90 risposte scritte. Su questo insieme è stata costruita la base empirica per la verifica della validità di scoring: il gold standard, definito attraverso l'accordo tra annotatori umani misurato mediante Kappa pesata, costituisce il parametro di riferimento per la valutazione delle prestazioni del sistema automatico. Il confronto tra modello e benchmark esperto evidenzia una differenza strutturale tra le due macro-abilità. Nella sezione relativa alla produzione scritta l'accordo risulta moderato e relativamente stabile; in quella relativa alla produzione orale emergono invece scarti più ampi e una tendenza alla sottostima sistematica. Tale divergenza appare riconducibile, almeno in parte, alla mediazione della trascrizione ASR e alla conseguente perdita di informazioni prosodiche, che incide in modo particolare su alcune dimensioni valutative. L'analisi dell'accordo tra modello e autovalutazioni dei candidati introduce inoltre una prospettiva complementare sulla validità relativa ai criteri. Il sistema si colloca sistematicamente in una posizione intermedia tra il benchmark esperto – più elevato e calibrato secondo standard professionali – e l'autopercezione dei candidati, generalmente più conservativa. Questa configurazione suggerisce che la criticità principale non risieda nella capacità discriminativa del modello, quanto piuttosto nella sua calibrazione assoluta rispetto a scale interpretative differenti. In ultimo, verrà preso in esame il *feedback* degli utenti, che fornisce indicazioni rilevanti circa l'esperienza d'uso della piattaforma.

4.1 Analisi qualitativa della popolazione

La sperimentazione pilota ha coinvolto complessivamente 40 partecipanti che hanno costituito la popolazione che ha svolto il test. Prima della somministrazione, ciascun partecipante ha accettato un modulo di consenso informato nel quale venivano descritte le finalità della ricerca, le modalità di trattamento dei dati personali nel rispetto del Regolamento Generale sulla Protezione dei Dati (GDPR) e la natura volontaria e anonima della partecipazione. Contestualmente alla compilazione del consenso, i partecipanti sono stati invitati a compilare un questionario sociolinguistico finalizzato alla raccolta di informazioni relative al profilo demografico, al repertorio linguistico, ai contesti di utilizzo delle lingue conosciute, alle motivazioni che hanno spinto l'acquisizione della lingua italiana e all'autovalutazione della propria competenza secondo i livelli del QCER. Tutti e 40 i partecipanti hanno rilasciato il consenso informato e compilato il questionario; le analisi descrittive che seguono si riferiscono pertanto all'intera popolazione di 40 soggetti.

Inerentemente alla composizione demografica, la popolazione si articola in 23 partecipanti di genere femminile e 17 di genere maschile. La distribuzione per fasce d'età mostra una netta prevalenza di giovani adulti: la maggior parte dei partecipanti ha un'età compresa tra i 18 e i 34 anni, con 15 soggetti nella fascia 18-24 e 16 nella fascia 25-34. Le fasce di età successive risultano progressivamente meno rappresentate: 4 partecipanti rientrano nella fascia 35-44, 2 nella fascia 45-54 e 1 partecipante rispettivamente in ciascuna delle tre fasce rimanenti. La Tabella 13 riporta la distribuzione incrociata per genere e fascia anagrafica.

Genere	18-24	25-34	35-44	45-54	55-64	65-74	75-99	Totale
Femminile	9	6	4	2	1	0	1	23
Maschile	6	10	0	0	0	1	0	17
Totale	15	16	4	2	1	1	1	40

Tabella 13: Distribuzione della popolazione per genere e fasce d'età

La prevalenza di giovani adulti è coerente con i canali di reclutamento informale utilizzati, basati principalmente sui social media e sulla distribuzione di volantini in luoghi di

studio e di aggregazione inerenti alla vita universitaria. Analogamente, anche i partecipanti raggiunti tramite contatti formali con scuole e istituti privati di italiano L2, centri linguistici di ateneo e organizzazioni multiculturali risultano in larga parte under 35, in linea con il profilo tipico di apprendenti che si apprestano a sostenere certificazioni linguistiche o a intraprendere percorsi di studio e di mobilità internazionale.

Per quanto riguarda il titolo di studio, la popolazione presenta una distribuzione relativamente equilibrata tra i livelli di istruzione secondaria e terziaria. Come illustrato nella Tabella 14, dieci partecipanti hanno completato la scuola secondaria di secondo grado, mentre i rimanenti 30 possiedono almeno un titolo universitario. Tra questi, 12 hanno una laurea triennale, otto una laurea magistrale, sette un master universitario e tre un dottorato di ricerca. Il livello di istruzione relativamente elevato della popolazione in esame costituisce un fattore rilevante nell'interpretazione dei risultati: l'esperienza accademica viene generalmente associata ad un maggiore sviluppo di competenze metacognitive complesse e a una più ampia familiarità con pratiche valutative formali. Va inoltre segnalato che cinque partecipanti possiedono una laurea in lingue, lettere o linguistica e otto dichiarano di aver insegnato o svolto attività di ricerca in ambito linguistico, non necessariamente riferita all'italiano.

Titolo di studio	Numero
Scuola secondaria di secondo grado	10
Laurea triennale	12
Laurea magistrale	8
Master universitario	7
Dottorato di ricerca	3

Tabella 14: 14 Distribuzione della popolazione per titolo di studio

In relazione alla provenienza geografica dei partecipanti, la popolazione si rivela marcatamente eterogenea, con una provenienza da 26 paesi differenti, appartenenti a quattro continenti. La Tabella 15 riporta la distribuzione per area geografica. L'Europa costituisce il gruppo più numeroso (17 partecipanti) seguita dell'Asia e dal Sud America (rispettivamente con nove partecipanti) e infine dall'Africa con (cinque partecipanti).

Continente	Paese	Numero partecipanti
Europa	Francia, Germania, Grecia, Paesi Bassi, Polonia, Regno Unito, Romania, Spagna, Svizzera, Ucraina, Ungheria	17

Asia	Azerbaijan, Cina, Indonesia, Iran, Libano, Pakistan, Russia	9
Sud America	Argentina, Colombia, Brasile	9
Africa	Algeria, Egitto, Ghana, Nigeria	5

Tabella 15: Distribuzione geografica dei partecipanti

Sul piano linguistico, la popolazione presenta un'elevata eterogeneità: i partecipanti dichiarano complessivamente 23 lingue o varietà di L1, riconducibili a 11 famiglie linguistiche. Il gruppo più rappresentato è quello delle lingue romanze – castigliano, catalano, francese, portoghese, spagnolo, romeno – un dato atteso in virtù della prossimità tipologica con l'italiano, che facilita il *transfer* interlinguistico e può incidere positivamente sia sui tempi di acquisizione sia sulle performance orali. A queste lingue della cosiddetta Romània storica si affiancano varietà della Romània nuova, come lo spagnolo dell'Argentina e della Colombia e il portoghese del Brasile.

Sempre nell'ambito indoeuropeo compaiono le lingue slave (con il polacco, il russo e l'ucraino), le germaniche (con il danese, l'inglese e il tedesco), il neogreco come lingua isolata e le lingue indo-iraniche (con l'urdu e il persiano). Sono inoltre rappresentate famiglie non indoeuropee, tipologicamente più distanti dall'italiano: l'uralica con l'ungherese, l'altaica con l'azero, l'afroasiatica con l'arabo, la Niger-Congo con il bini e lo yoruba, la sinotibetana con il mandarino e l'austronesiana con l'indonesiano. Tali lingue condividono con l'italiano un grado limitato di affinità strutturale e, nella maggior parte dei casi, sistemi di scrittura differenti, andando a configurare condizioni di maggiore distanza tipologica.

L'eterogeneità delle L1 rappresenta al contempo un punto di forza del campione, in termini di ampiezza della rappresentatività, e una variabile complessa nell'interpretazione dei punteggi. La distanza tipologica dalla lingua *target* influenza i profili di errore, i tempi di acquisizione e la distribuzione delle difficoltà tra i diversi parametri valutativi. Sebbene un'analisi sistematica delle correlazioni tra famiglia linguistica di appartenenza e performance al test esuli dagli obiettivi del presente studio pilota, essa costituisce una promettente direzione di ricerca per sviluppi futuri del sistema. In particolare, implementazioni multimodali orientate alla valutazione globale del parlato potrebbero richiedere l'indagine di eventuali *bias* legati a specifiche L1.

Dal punto di vista del repertorio linguistico, la popolazione presenta un profilo marcatamente plurilingue: ogni partecipante dichiara di parlare in media due lingue oltre all'italiano e alla propria lingua madre, con un massimo di quattro lingue aggiuntive; nessun partecipante dichiara di essere monolingue. Questo profilo plurilingue implica che per molti partecipanti l'italiano non è la seconda lingua in senso assoluto, bensì una delle lingue del repertorio con funzioni e domini d'uso specifici. L'analisi di questi domini evidenzia un plurilinguismo diffuso e funzionalmente articolato, nel quale la scelta della lingua varia in base all'interlocutore e al contesto comunicativo. La lingua di origine è prevalentemente impiegata nelle interazioni con familiari e amici del paese di provenienza, mentre l'italiano assume il ruolo di lingua principale nei contesti sociali del paese di residenza, in particolare sul luogo di lavoro e nelle relazioni con amici e coinquilini. L'inglese compare con frequenza come lingua franca, e non manca inoltre l'uso di varietà dialettali. Le risposte fornite dai partecipanti illustrano chiaramente tale distribuzione funzionale. U10, ad esempio, dichiara di utilizzare «olandese con i figli, inglese con il partner, italiano con gli amici», evidenziando una netta compartimentazione dei domini linguistici. Analogamente, U16 riferisce di parlare «inglese con i miei coinquilini ed amici, italiano/inglese con il mio partner e arabo con i miei genitori», configurando un repertorio dinamico e adattivo. In altri casi emerge una coesistenza di lingua nazionale e varietà dialettali: U12 riporta l'uso del «portoghese» con i familiari, affiancato dal «dialetto talian (con meno frequenza)», mentre U24 distingue tra «dialetto di Shanghai» e «cinese» con i genitori, alternando «italiano» e «cinese» con la partner e «cinese» e «inglese» con i coinquilini. Nel complesso, il quadro che emerge è quello di repertori plurilingui stratificati, nei quali l'italiano svolge per la maggior parte del campione una funzione comunicativa autentica e quotidiana.

Le modalità di acquisizione dell'italiano e i contesti effettivi di utilizzo della lingua costituiscono variabili centrali nella configurazione della competenza linguistica degli apprendenti. Per quanto riguarda il percorso di apprendimento formale, 35 partecipanti su 40 dichiarano di aver studiato l'italiano in contesti didattici strutturati (a scuola, in corsi privati o in università), mentre cinque riferiscono un'acquisizione esclusivamente informale, basata sull'esposizione spontanea. L'età media di inizio dello studio è pari a circa 23 anni, con un intervallo compreso tra i 12 e i 50 anni; quasi il 70% dei rispondenti che hanno fornito questa informazione ha iniziato in età adulta (≥18 anni), configurandosi quindi come apprendenti tardivi, categoria per la quale la letteratura SLA documenta traiettorie di acquisizione parzialmente differenti rispetto agli apprendenti precoci, in particolare per quanto concerne il

raggiungimento di livelli fonologici prossimi a quelli nativi. Il 60% dei partecipanti indica inoltre che la pronuncia è stata oggetto di insegnamento esplicito, mentre i restanti dichiarano di aver sviluppato le abitudini fonologiche unicamente attraverso l'uso.

Il dato più rilevante concerne tuttavia il contesto di residenza: 35 partecipanti su 40 vivono attualmente in Italia e l'ottanta percento ha soggiornato nel paese per un periodo continuativo di almeno tre mesi, segnalando un'elevata esposizione alla lingua nel suo ambiente d'uso naturale. Ne consegue che, pur essendo l'italiano stato appreso prevalentemente tramite istruzione formale, la maggior parte degli apprendenti ha beneficiato anche di condizioni di immersione linguistica, fattore noto per esercitare un beneficio sui processi acquisizionali. Questa predominanza di apprendenti in contesti di immersione è confermata dalla frequenza d'uso orale dell'italiano: il 65% dichiara di utilizzare la lingua italiana quotidianamente o spesso nella conversazione, e una quota analoga lo impiega regolarmente in ambito lavorativo o accademico. La Tabella 16 sintetizza i tipi di esposizione, le modalità di apprendimento e i contesti d'uso dell'italiano riportati dai partecipanti.

Indicatore	Utilizzo orale quotidiano	Contesto lavorativo o di studio	Soggiorno ≥3 mesi	Residenza attuale in Italia
Quotidianamente	18	17	32	35
Spesso	8	8	-	-
A volte	9	4	-	-
Una volta ogni tanto	5	4	-	-
Mai	-	7	8	5

Tabella 16: Indicatori di esposizione e utilizzo dell'italiano

Come illustrato nel paragrafo 1.6, la letteratura sull'attrattività e sui pubblici dell'italiano L2 delinea un panorama di apprendenti eterogeneo, caratterizzato da motivazioni all'apprendimento tra loro differenziate. La popolazione coinvolta nella presente sperimentazione riflette in larga misura tale varietà, offrendo una rappresentazione trasversale dei principali profili di utenti individuati negli studi di settore (De Mauro 2002; Vedovelli 2021). Le motivazioni dichiarate dai partecipanti sono state ricondotte, mediante un processo di codifica tematica, a tre macrocategorie: integrativa, strumentale e affettivo-identitaria. La Tabella 17 ne sintetizza la distribuzione.

Categoria	Descrizione	Numero di partecipanti
Integrativa	Trasferimento in Italia, comunicazione quotidiana, inserimento nella comunità, partner italiano/a	12
Strumentale	Lavoro, studio universitario, dottorato, Erasmus, tirocinio	16
Affettivo-identitaria	Interesse culturale e linguistico, eredità familiare, motivazioni politico-culturali	13

Tabella 17:Motivazioni all'apprendimento della lingua italiana come L2

La motivazione integrativa, costituita da 12 partecipanti, si articola principalmente attorno alla residenza o al trasferimento in Italia e alla necessità di comunicare efficacemente nei contesti della vita quotidiana. Rientrano in questa categoria apprendenti che acquisiscono l'italiano come strumento per inserirsi nel tessuto sociale e lavorativo del Paese, nonché soggetti coinvolti in dinamiche di plurilinguismo familiare, ad esempio con partner italofoni.

La motivazione strumentale, composta da 16 utenti, comprende apprendenti per i quali la padronanza dell'italiano è funzionale al raggiungimento di obiettivi accademici o professionali specifici: studenti in mobilità Erasmus, allievi che includono l'italiano nel proprio curriculum universitario all'estero, dottorandi trasferitisi in Italia e lavoratori che necessitano di competenze terminologiche specifiche nel proprio ambito specialistico.

Infine, la motivazione affettivo-identitaria presenta profili di particolare interesse teorico. Vi rientrano tre casi riconducibili alla categoria degli *heritage speakers*: un partecipante svizzero del cantone germanofono con nonni veneti che studia l'italiano per riconnettersi con le proprie radici familiari; un partecipante brasiliano discendente di emigrati italiani che conserva il dialetto «talian» nelle interazioni domestiche; e un terzo soggetto che richiama esplicitamente l'«ereditarietà di famiglia» come leva motivazionale. In questa stessa area si collocano motivazioni di natura simbolico-culturale, come nel caso di U24 che associa l'apprendimento dell'italiano a un'identificazione politico-valoriale: «O bella ciao-antifascismo», evidenziando come la lingua possa fungere da veicolo di appartenenza ideale oltre che da strumento comunicativo. Analogamente, alcune risposte richiamano elementi estetici, quali il "suono" o la "melodia" dell'italiano, nonché la percezione di una relativa facilità nelle fasi iniziali dell'apprendimento.

Prima della somministrazione del test, ai partecipanti è stato richiesto di autovalutare la propria competenza linguistica in riferimento ai livelli del QCER, sia in termini globali sia con specifico riferimento alla produzione orale. La Tabella 18 riporta la distribuzione delle risposte per entrambe le dimensioni.

Livello	Competenza globale	Percentuale	Competenza relativa all'oralità	Percentuale
A1	3	7,5%	4	10%
A2	3	7,5%	4	10%
B1	11	27,5%	11	27,5%
B2	14	35%	11	27,5%
C1	6	15%	7	17,5%
C2	3	7,5%	3	7,5%

Tabella 18: 18 Autovalutazione della competenza linguistica secondo i livelli QCER

L'autovalutazione globale evidenzia una concentrazione nei livelli intermedi, con il B2 come livello *target* (35%), seguito da B1 (27,5%) e C1 (15%). I livelli estremi risultano meno rappresentati: A1 e A2 coprono complessivamente il 15% del campione, mentre il C2 si attesta al 7,5%.

Tale profilo appare coerente con la composizione del campione descritta nelle sezioni precedenti: la maggioranza dei partecipanti risiede in Italia e utilizza l'italiano in contesti quotidiani, condizioni che tendono a collocare la competenza e la relativa percezione nei livelli medio-alti della scala. Lo sbilanciamento verso i livelli intermedi risulta inoltre prevedibile alla luce di due fattori convergenti. Da un lato, la popolazione degli apprendenti di L2 si distribuisce generalmente con maggiore densità proprio nelle fasce centrali di competenza. Dall'altro, la partecipazione volontaria a procedure valutative è soggetta a meccanismi di autoselezione: numerosi studi mostrano che apprendenti con competenza percepita più bassa tendono a evitare contesti di testing per timore di esiti negativi e ansia linguistica, mentre soggetti più competenti manifestano maggiore disponibilità a esporsi alla valutazione (MacIntyre et. al. 1997). Questo fenomeno, noto come *self-selection bias* nella partecipazione ai test, implica che differenze sistematiche tra chi accetta di essere valutato e chi no possano incidere sul bilanciamento del campione, determinando una sovrarappresentazione di apprendenti con livelli di competenza medio-alti e una sottorappresentazione dei livelli più bassi.

L'autovalutazione della sola produzione orale mostra invece uno spostamento verso i livelli inferiori: il B1 diviene il livello di riferimento condiviso con il B2 (entrambi al 27,5%),

la percentuale di A1 aumenta al 10% e quella del C1 sale al 17,5%. Nel complesso, 28 partecipanti su 40 attribuiscono lo stesso livello QCER alla competenza globale e quella orale; 8 partecipanti valutano invece la propria produzione orale un livello inferiore rispetto alla competenza complessiva, mentre 4 la collocano a un livello superiore.

Lo scostamento osservato suggerisce che una parte dei partecipanti percepisce l'oralità come una dimensione relativamente più debole della propria competenza linguistica complessiva, un fenomeno ampiamente documentato in letteratura. Tale tendenza è generalmente ricondotta, in primo luogo, al filone di studi sull'ansia linguistica, che evidenzia come le attività di speaking attivino livelli più elevati di tensione e vulnerabilità rispetto alle abilità ricettive (Horwitz et. al. 1986); in secondo luogo, alle ricerche sulla percezione di competenza comunicativa, secondo cui gli apprendenti tendono ad autovalutarsi meno competenti nelle situazioni di interazione orale, maggiormente esposte al giudizio sociale (Young 1991); infine, ai contributi che sottolineano il maggiore carico cognitivo della produzione orale, caratterizzata da elaborazione in tempo reale e dalla necessità di coordinare simultaneamente risorse lessicali, grammaticali e pragmatiche (Li 2023). In controtendenza, i pochi casi in cui la produzione orale è autovalutata a un livello superiore rispetto alla competenza complessiva sembrano riconducibili a profili di apprendimento caratterizzati da un'esposizione prevalentemente immersiva e da una minore formalizzazione della competenza, nei quali è tipico osservare uno sviluppo relativamente più avanzato dell'oralità rispetto alle altre abilità.

I dati di autovalutazione verranno ripresi nel corso del capitolo come punto di riferimento per la valutazione della validità relativa ai criteri. Il confronto tra i livelli auto percepiti e quelli assegnati dal sistema ETET costituisce un indicatore della plausibilità e della calibrazione del modello rispetto alle aspettative dei candidati.

4.2 Criteri di esclusione e creazione del campione

Il test utilizzato nella fase pilota si articolava in due sezioni produttive, una dedicata alla composizione scritta e composta da due domande e una parte inerente alla produzione orale articolata in tre quesiti. A differenza degli altri task presenti nel test e come illustrato nel Capitolo 3, tali sezioni sono state valutate mediante LLM o tramite una *pipeline* integrata basata su Azure per la trascrizione automatica e la valutazione del *pronunciation assessment* e la

successiva valutazione tramite LLM. La somministrazione dell'esame ha coinvolto 40 partecipanti; il numero di risposte attese ammontava pertanto a 80 per il writing e a 120 per lo speaking. A queste si aggiungono le produzioni del gruppo di controllo, costituito da 5 parlanti nativi italiani sottoposti al medesimo protocollo: ulteriori 10 risposte di scrittura e 15 di parlato. Complessivamente, il corpus teorico atteso sarebbe dovuto essere composto da 90 risposte scritte e 135 risposte orali.

Tuttavia, a causa di criticità emerse durante la somministrazione, il campione effettivamente utilizzabile si è ridotto in modo significativo, con una concentrazione esclusiva dei problemi nella sezione di speaking. Su 135 risposte orali attese, 39 hanno presentato malfunzionamenti di diversa natura, riconducibili sia a limitazioni tecniche della piattaforma sia a comportamenti errati degli utenti. La presenza di tali anomalie ha reso necessaria un'analisi sistematica preliminare delle risposte orali al fine di distinguere le valutazioni tecnicamente attendibili da quelle generate a partire da input parziali o mancanti. Le sezioni seguenti illustrano pertanto la classificazione e l'analisi di tali problematiche, che costituiscono la premessa metodologica per la definizione del campione di risposte valide impiegato nelle analisi successive.

L'analisi di dettaglio delle anomalie riscontrate nella sezione di speaking ha permesso di precisarne la distribuzione interna. I malfunzionamenti, infatti, non risultano omogeneamente distribuiti tra i partecipanti ma tendono a concentrarsi in un sottogruppo specifico. In particolare, 26 test su 45 non hanno presentato alcuna criticità, mentre i restanti 19 hanno registrato almeno una risposta problematica. Questo dato indica che le anomalie non si configurano come episodi isolati bensì come eventi che colpiscono selettivamente determinate sessioni di somministrazione. La Tabella 19 riepiloga la distribuzione a livello di test e di singola domanda.

Categoria	Numero test	Numero Domande
Test senza problemi	26	78
Test con almeno un problema	19	57
Domande senza problemi	-	17
Domande con problemi	-	40

Tabella 19: 19 Panoramica dei problemi nella sezione di speaking

Considerando i soli 19 test con malfunzionamenti, le 57 corrispondenti domande di speaking mostrano una proporzione di criticità particolarmente elevata: solo 17 su 57 (il 29,8%) risultano prive di problemi, mentre le restanti 40 (il 70,2%) presentano almeno un'anomalia. Questo dato indica che all'interno del sottogruppo di test non andati a buon fine, i malfunzionamenti non sono sporadici ma tendono a colpire la maggior parte delle domande dello stesso test. Ne emerge un quadro compatibile con cause sistemiche legate al singolo partecipante o alla specifica sessione di registrazione, piuttosto che errori isolati del sistema. I 40 casi problematici sono stati classificati in cinque tipologie distinte sulla base dell'analisi dei file audio, delle trascrizioni disponibili e dei metadati della registrazione. La Tabella 20 riporta la distribuzione per tipologia degli errori.

Tipologia	Descrizione	Numero di domande
Trascrizione troncata	Il file audio è presente e di durata normale, ma il sistema ASR interrompe la trascrizione prima della fine della registrazione	19
Risposta non presente	Il file audio non è mai pervenuto al sistema; la domanda risulta priva di qualsiasi dato	11
Risposta saltata intenzionalmente	L'utente ha registrato un file audio di pochi secondi per superare la domanda senza effettivamente produrre una risposta	6
Audio vuoto	L'utente non ha configurato correttamente il microfono come richiesto a inizio test; il sistema riceve un file privo di segnale	3
Allontanamento dal microfono	L'utente ha iniziato a rispondere regolarmente, ma si è allontanato dal microfono durante la registrazione causandone il degrado	1

Tabella 20: Categorie di errori nelle domande con problemi

La tipologia più frequente è quella della trascrizione troncata con 19 casi, in cui il file audio risulta presente e di durata regolare ma il sistema di *speech-to-text* interrompe la trascrizione prima della fine della registrazione. Questa categoria è anche quella con le implicazioni più rilevanti per la validità dello *scoring*, poiché produce trascrizioni parziali su cui GPT-4o opera come se fossero complete, generando valutazioni linguistiche che non rappresentano la produzione effettiva del partecipante.

La seconda categoria per frequenza è costituita dalle risposte non pervenute 11 casi, in cui il file audio non è mai pervenuto al sistema, lasciando la domanda priva di qualsiasi dato sia per la componente Azure sia per la valutazione GPT-4o. Le cause risultano eterogenee e comprendono interruzioni di connessione durante l'*upload*, errori di gestione della sessione lato browser e, in alcuni casi segnalati nel questionario di *feedback*, il mancato avvio della registrazione.

La terza tipologia è quella delle risposte saltate intenzionalmente con 6 casi: l'utente ha registrato un file audio di pochi secondi al solo scopo di sbloccare il passaggio alla domanda successiva. Questo comportamento configura una forma di aggiramento del sistema che produce punteggi di fatto nulli. Pur non costituendo un problema tecnico del sistema in senso stretto, questi casi rappresentano una criticità metodologica rilevante in quanto introducono nel dataset risposte non genuine che abbassano artificialmente la media dei punteggi speaking.

Le ultime due tipologie risultano meno frequenti ma ugualmente informative. I tre casi di audio vuoto riguardano lo stesso partecipante (U37), che non ha configurato correttamente il microfono nonostante le istruzioni fornite ad inizio test: il sistema ha ricevuto file audio privi di segnale per tutte e tre le domande. Il caso singolo di allontanamento dal microfono riguarda invece il partecipante U30 nella domanda S1: la trascrizione si è avviata regolarmente ma si è interrotta quando la qualità del segnale è degradata a causa del movimento dell'utente. Questo caso è stato classificato separatamente dalla troncatura sistematica in quanto la causa è attribuibile al comportamento dell'utente e non a un limite del sistema.

I 19 casi di trascrizione troncata sono stati analizzati nel dettaglio al fine di verificare l'eventuale presenza di pattern sistematici nel comportamento del sistema di ASR. Per ciascun caso sono stati rilevati il momento di interruzione della trascrizione espresso in secondi e la durata totale del file audio, consentendo il calcolo della percentuale di audio effettivamente trascritta. La Tabella 21 riporta i dati per ciascun caso, mentre la Tabella 22 ne sintetizza le statistiche descrittive.

Utente	Domanda	Momento di interruzione (s)	Durata totale della registrazione (s)	Percentuale trascritta	Note
U1	S2	30	106	28,3%	-
U4	S1	22	102	21,6%	-
U14	S1	30	45	67%	Due domande troncate
U14	S3	32	40	80%	-
U19	S1	30	80	37,5%	Tre domande troncate
U19	S2	22	98	22,4%	-
U19	S3	32	93	34,4%	-
U20	S1	6	67	9%	Due domande troncate
U20	S3	2	97	2,1%	-
U24	S2	15	85	17,6%	-
U26	S3	10	34	29,4%	-
U30	S1	10	65	15,4%	Tre domande troncate *
U30	S2	27	55	49,1%	-
U30	S3	23	35	65,7%	-
U31	S1	5	73	6,8%	Due domande troncate
U31	S2	30	117	25,6%	-
U34	S1	7	86	8,1%	Due domande troncate
U34	S3	10	93	10,8%	-
U39	S2	30	67	44,8%	Due domande troncate
U39	S3	8	78	10,3%	-

Tabella 21: Analisi dei casi in cui è intercorsa un'interruzione nella trascrizione

Variabile	Media	Mediana	Deviazione standard
Momento di interruzione (secondi)	18,5	22,0	10,8
Percentuale di audio trascritta (%)	27,3	22,4	20,8

Tabella 22: Statistiche descrittive relative ai casi di interruzione della trascrizione

Le statistiche mostrano una distribuzione del momento di interruzione con media di 18,5 secondi, mediana di 22 secondi e deviazione standard di 10,8 secondi, con valori che spaziano da un minimo di 2 secondi a un massimo di 32 secondi. La percentuale di audio trascritta ha invece media del 27,3% e mediana del 22,4%, con deviazione standard del 20,8%. Nella maggior parte dei casi il sistema ha quindi trascritto meno di un quarto della registrazione, producendo testi del tutto insufficienti per una valutazione linguistica attendibile.

L'analisi della distribuzione del momento di interruzione rivela la coesistenza di due pattern distinti che la statistica aggregata tende a mascherare. Il primo pattern riguarda 6 casi su 19 in cui l'interruzione avviene tra i 28 e i 32 secondi, con il valore massimo di 32 secondi che funge da soglia apparente. La concentrazione di casi attorno a questo intervallo suggerisce l'esistenza di un limite tecnico del servizio di ASR attorno ai 30 secondi di elaborazione continuativa. In questi casi la causa della troncatura è quindi attribuibile a una limitazione del sistema indipendente dalla qualità dell'audio o dal comportamento dell'utente. Il secondo pattern riguarda gli 8 casi con interruzione al di sotto dei 15 secondi, in cui la troncatura molto precoce suggerisce per esempio problemi di connessione durante la trasmissione del flusso audio. I restanti 5 casi si collocano in una fascia intermedia (15-27 secondi) che non consente una classificazione univoca.

Quanto alla percentuale di audio trascritta, i dati confermano l'assenza di una correlazione lineare tra la durata totale del file e il momento di interruzione: file di durata simile mostrano percentuali di trascrizione molto diverse (si confrontino ad esempio l'utente U19 nella domanda S2 con il 22,4% di testo trascritto e U30 nella domanda S3 con il 65,7%, entrambi su file di circa 93-98 secondi), mentre file di durata diversa possono mostrare percentuali analoghe.

Un ulteriore elemento di analisi riguarda la distribuzione dei casi di troncatura tra i partecipanti. Sei utenti su nove presentano troncature in più di una domanda: U19 e U30 mostrano troncature in tutte e tre le domande, mentre U14, U20, U31, U34 e U39 ne presentano

due. Questo pattern suggerisce che la causa della troncatura non risiede esclusivamente in fattori casuali legati alla singola domanda, ma è in parte determinata da caratteristiche della sessione del partecipante quali la configurazione del dispositivo, la velocità della connessione o le impostazioni del *browser*.

Alla luce dei problemi emersi, le 135 risposte di produzione orale sono state classificate in 95 valide e 40 non valide sulla base dei criteri illustrati nella Tabella 19. A differenza di quanto avviene tipicamente nei sistemi di valutazione con campionamento per soggetto, in cui l'unità di esclusione è il partecipante nella sua interezza, in questa sede si è scelto di adottare la singola domanda come unità di analisi. Questa scelta consente di preservare tutte le risposte tecnicamente attendibili indipendentemente dal numero di problemi riscontrati nelle altre domande dello stesso test, massimizzando la quantità di dati validi disponibili per l'analisi.

In termini pratici, la distribuzione delle risposte valide per utente mostra che 26 partecipanti dispongono di tutte e tre le risposte di speaking valide, 7 ne presentano due, 4 ne posseggono una sola e 8 non hanno alcuna risposta valida. Questi ultimi contribuiscono all'analisi esclusivamente attraverso i dati di produzione scritta. Le analisi sui punteggi di produzione orale presentate nelle sezioni successive si baseranno pertanto sul campion di 95 risposte valide.

4.3 Costruzione del gold standard

Al fine di verificare empiricamente l'effettiva validità di *scoring* (cfr. paragrafo 2.2.1) e l'affidabilità (cfr. paragrafo 2.2.2) del modello creato, sono state condotte analisi per indagare la stabilità, la coerenza e l'assenza di distorsioni sistematiche nel processo di assegnazioni dei punteggi. Nel caso di sistemi di valutazione automatica la questione della validità di punteggio assume una rilevanza centrale: l'automazione introduce infatti ulteriori livelli di mediazione tra performance e punteggio, rendendo necessario verificare che tali mediazioni non generino distorsioni sistematiche nel processo valutativo. In questa prospettiva, la validazione non può limitarsi alla sola verifica della stabilità interna del sistema (comunque analizzata e descritta nel paragrafo 3.3.2), ma deve includere un confronto esterno con un riferimento indipendente. Seguendo un orientamento consolidato nella letteratura sulla validazione dei sistemi di *scoring* automatizzato, la presente analisi adotta come criterio di riferimento un *gold standard* costruito a partire dal calcolo dell'accordo ottenuto dai risultati delle annotazioni realizzate da tre valutatori umani esperti, denominati V1, V2 e V3.

Il campione di domande utilizzato per quest'analisi è costituito – come motivato nel paragrafo 4.2 – da 95 domande di produzione orale e 90 di produzione scritta. Per ciascuna domanda ogni annotatore ha assegnato un punteggio in scala 0-100 per ognuno dei parametri oggetto d'indagine, *ovvero: Lexis, Content, Grammar, Coherence e Comprehension* (per entrambe le sezioni), con l'aggiunta di *Fluency* e *Accuracy* per il solo *speaking*. In funzione dei valori assegnati in ciascuna categoria, è stato successivamente calcolato il punteggio Totale della domanda, in accordo con i pesi descritti nel paragrafo 3.3.2.7.

Per poter realizzare il *gold standard* è stato necessario verificare il grado di accordo tra i tre annotatori al fine di ottenere la migliore approssimazione del costrutto misurato. Il *gold standard* si configura come un ulteriore parametro di valutazione relativa ai criteri, discussa nel paragrafo 2.2.1; in questa prospettiva, la validità non è intesa come proprietà intrinseca dello strumento, ma come grado di coerenza delle inferenze che da esso derivano rispetto a un quadro di riferimento condiviso.

Ad una prima analisi è stato possibile osservare una marcata divergenza tra i risultati prodotti dai vari annotatori umani. Si è pertanto deciso di procedere con una *pairwaise comparison* tra gli annotatori al fine di identificare la coppia caratterizzata dal maggiore grado di accordo, assunta successivamente come riferimento per la costruzione del *gold standard*.

Per valutare tale accordo si scelto di far riferimento alla *Kappa di Choen*, in particolare la *Kappa di Cohen pesata con pesi quadratici*. La scelta di questa specifica metrica è direttamente motivata dalla natura della variabile oggetto di analisi: i livelli di competenza linguistica rappresentano categorie ordinali gerarchicamente strutturate (A1-C2), nelle quali la distanza tra i livelli ha un significato interpretativo sostanziale. L'impiego della Kappa semplice sarebbe stato metodologicamente inadeguato, poiché tale indice è concepito per variabili nominali e tratta ogni disaccordo come equivalente, indipendentemente dall'entità dello scarto tra categorie. Nel nostro caso, invece, un disaccordo tra livelli adiacenti non è comparabile a uno scostamento di due o più livelli; era quindi necessario un indicatore capace di incorporare la struttura ordinale della scala. Poiché il *gold standard* è stato costruito sulla base dei giudizi dei due annotatori con il più alto livello di convergenza, l'analisi dell'accordo è stata condotta su coppie di valutatori. Tale impostazione metodologica giustifica l'utilizzo della *Kappa di Cohen*, invece di indici quali la *Kappa di Fleiss* o *l'Alfa di Krippendorff*, generalmente impiegati in presenza di più di due annotatori (Jezek, Sprugnoli 2023).

La Kappa pesata (weighted kappa; k_w) misura la riduzione proporzionale del disaccordo rispetto a quello atteso in condizioni di indipendenza tra annotatori, attribuendo pesi

differenziati ai diversi gradi di disaccordo; in questo modo, l'indice non si limita a distinguere tra accordo e disaccordo, ma tiene conto dell'entità dello scarto tra le categorie assegnate. Formalmente, l'indice è definito come:

$$k_w = 1 - \frac{D_O}{D_e}$$

(4)

dove D_o rappresenta il disaccordo osservato pesato e D_e il disaccordo atteso pesato.

Viene identificato con n_{ij} il numero di casi classificati nel livello i dal primo annotatore e nel livello j dal secondo, mentre con N il numero totale di osservazioni. Si possono dunque definire le proporzioni osservate come $O_{ij} = \frac{n_{ij}}{N}$.

Le distribuzioni marginali dei due annotatori sono date da $p_i = \sum_j O_{ij}$ e da $p_j = \sum_i O_{ij}$; queste consentono di costruire la matrice delle proporzioni attese sotto indipendenza $E_{ij} = p_i \cdot p_j$.

Introducendo una matrice di pesi w_{ij}, il disaccordo osservato e quello atteso sono calcolati rispettivamente come:

$$D_o = \sum_{i,j} w_{ij} O_{ij} \qquad D_e = \sum_{i,j} w_{ij} E_{ij}$$

(5)

Nel caso specifico sono stati adottati pesi quadratici,

$$w_{ij} = \left(\frac{i-j}{k-1}\right)^2$$

(6)

dove k è il numero totale di livelli della scala. Tale formulazione assegna peso nullo ai casi di accordo perfetto $i=j$ e penalizza in modo crescente e non lineare gli scostamenti più ampi.

La scelta dei pesi quadratici risponde all'esigenza di riflettere in modo più adeguato la gravità degli errori: uno scarto di due o tre livelli rappresenta una divergenza sostanziale nella valutazione della competenza e deve incidere maggiormente sull'indice rispetto a un disaccordo di un solo livello. Data l'eccessiva granularità di una scala 0-100 e al fine di ridurre la dimensione del problema è stato scelto di riportare i punteggi assegnati in una scala 1-6, coerente con la scala A1-C2. Tutti i punteggi assegnati sono dunque stati tradotti in questo nuovo sistema di misura secondo la scala convenzionalmente utilizzata da ETET per l'assegnazione dei livelli.

La *Kappa* è stata calcolata per ogni parametro (*Lexis, Content, Grammar, Coherence, Comprehension*, *Fluency* e *Accuracy*) e sul punteggio totale di ciascuna domanda. I risultati ottenuti sono stati interpretati adottando le soglie teorizzate da Landis e Koch (1977): $k < 0{,}00$ = scarsa (*poor*); 0,00-0,20= lieve (*slight*); 0,21-0,40 = debole (*fair*); 0,41-0,60 = moderata (*moderate*); 0,61-0,80 = sostanziale (*substantial*); $k > 0{,}80$ = quasi perfetta (*almost perfect*). Tabella 23 riporta i risultati del confronto tra le coppie nella sezione di produzione orale.

Categoria	**A1-A2**	**Classificazione dell'accordo**	**A2-A3**	**Classificazione dell'accordo**	**A1-A3**	**Classificazione dell'accordo**
Lexis	0,311	Debole	0,344	Debole	0,040	Scarsa
Content	0,521	Moderata	0,207	Debole	0,038	Scarsa
Grammar	0,292	Debole	0,432	Moderata	0,083	Scarsa
Coherence	0,344	Debole	0,289	Debole	0,033	Scarsa
Comprehension	0,658	Sostanziale	0,170	Lieve	0,071	Scarsa
Fluency	0,336	Debole	0,343	Debole	0,073	Scarsa
Accuracy	0,204	Debole	0,345	Debole	0,029	Scarsa
Totale	0,437	Moderata	0,273	Debole	0,040	Scarsa

Tabella 23: Confronto dell'accordo tra le varie coppie di annotatori nella sezione di produzione orale

Le valutazioni di A3 si distinguono sistematicamente per la tendenza ad assegnare punteggi più bassi rispetto agli altri due annotatori: la media del punteggio Totale assegnato da A3 è di 3,48 (su scala 1-6), contro 5,92 per A1 e 5,35 per A3. Tale divergenza va a costituirsi come un errore sistematico che potrebbe minare la validità di *scoring*, riflettendo divergenze

nelle prospettive di assegnazione dei punteggi dei valutatori piuttosto che differenze nella competenza dei candidati. Al fine di scongiurare questo scenario, le valutazioni di A3 non sono state utilizzate per la costruzione del *gold standard*. L'accordo pressoché nullo tra A3 e A1 – con $\kappa = 0{,}040$ sul Totale, classificabile come accordo lieve – indica che i due annotatori producono distribuzioni di punteggi tra loro inconciliabili. Anche l'accordo tra A3 e A2 risulta debole ($\kappa = 0{,}273$), mentre la coppia A1-A2 raggiunge un accordo moderato ($\kappa = 0{,}437$), con valori più elevati in particolare nei parametri di *Comprehension* e *Content*.

Nella sezione relativa alla produzione scritta i risultati mostrano un *pattern* analogo ma con livelli di accordo complessivamente superiori. Anche in questo caso la coppia A1-A3 presenta l'accordo più basso – con $\kappa = 0{,}293$ sul Totale, rappresentando un accordo debole – mentre la coppia A1-A2 raggiunge valori notevolmente più alti: $\kappa = 0{,}828$ sul Totale, classificabile come accordo quasi perfetto e in cui tutti i parametri si collocano in una fascia elevata (0,698–0,787). Tabella 24 riporta i risultati dell'accordo tra gli annotatori relativi alla sezione di produzione scritta.

Categoria	**A1-A2**	**Classificazione dell'accordo**	**A2-A3**	**Classificazione dell'accordo**	**A1-A3**	**Classificazione dell'accordo**
Lexis	0,772	Sostanziale	0,454	Moderata	0,387	Lieve
Content	0,787	Sostanziale	0,369	Debole	0,383	Lieve
Grammar	0,757	Sostanziale	0,505	Moderata	0,369	Lieve
Coherence	0,732	Sostanziale	0,355	Debole	0,253	Lieve
Comprehension	0,698	Sostanziale	0,330	Debole	0,133	Scarsa
Totale	0,828	Quasi perfetta	0,407	Moderata	0,293	Lieve

Tabella 24: Confronto dell'accordo tra le varie coppie di annotatori nella sezione di produzione scritta

Alla luce dei seguenti risultati la coppia A1-A2 è stata identificata come quella con il maggiore accordo reciproco in entrambe le sezioni ed è stata pertanto selezionata per la costruzione del *gold standard*; l'annotatore A3 è stato invece escluso dall'analisi principale. Il *gold standard* per ciascuna risposta e ciascuna categoria è stato calcolato come la media aritmetica arrotondata dei punteggi assegnati dai due annotatori selezionati, ottenendo un valore intero su scala 1-6.

4.4 Analisi dell'accordo tra gold standard e modello

Una volta costruito il *gold standard*, è stato possibile valutare le performance del modello ETET mettendo a confronto i suoi punteggi con quelli prodotti dagli annotatori umani. Tale comparazione è stata effettuata in maniera separata su entrambe le sezioni produttive attraverso il calcolo della Kappa di Cohen pesata con pesi quadratici, del *Mean Absolute Error* (MAE), del *bias* medio e della distribuzione delle distanze tra il punteggio del modello e il *gold standard*.

Il MAE misura la distanza media assoluta tra due valutazioni; in questo caso, codificando i livelli su scala numerica 1-6, si configura come:

$$MAE = \frac{1}{N} \cdot \sum_{i=1}^{N} |x_i - y_i|$$

(7)

dove x_i è il punteggio assegnato dal modello, y_i è il *gold standard* e N è il numero totale di osservazioni. L'indice restituisce quindi lo scarto medio in livelli, senza distinguere la direzione dell'errore.

Oltre al calcolo del MAE, che misura semplicemente quanto, in media, i punteggi differiscono tra loro senza considerare la direzione dello scarto, è stato calcolato anche il *bias* medio definito come:

$$bias = \frac{1}{N} \cdot \sum_{i=1}^{N} (x_i - y_i)$$

(8)

dove x_i rappresenta il punteggio del modello e y_i il punteggio del *gold standard*. A differenza del MAE, il *bias* conserva la direzione dello scarto e consente quindi di identificare eventuali tendenze sistematiche del modello a sovrastimare (*bias* positivo) o sottostimare (*bias* negativo) il *benchmark* umano.

La distinzione tra MAE e *bias* riveste un'importanza metodologica significativa, in quanto i due indicatori offrono informazioni di natura diversa circa il tipo di errore commesso.

Il MAE misura di quanto i punteggi assegnati dal sistema automatico si discostino in media da quelli attribuiti dai valutatori esperti, fornendo una stima globale della distanza tra le due valutazioni. Il *bias* consente invece di stabilire se tale distanza sia distribuita casualmente oppure se il sistema tenda in modo sistematico a posizionare i candidati su un livello superiore o inferiore rispetto al riferimento umano. Quando il *bias* si avvicina al valore del MAE, significa che la maggior parte dello scarto si orienta nella medesima direzione: il sistema risulta, ad esempio, costantemente più severo o più generoso del valutatore umano. Se invece il *bias* è prossimo allo zero ma il MAE rimane elevato, ciò indica che gli errori si distribuiscono in entrambe le direzioni – con il sistema che alterna sovrastime e sottostime – senza che emerga una tendenza stabile. In quest'ultimo caso, la problematica riguarda essenzialmente la precisione dello strumento, e non la presenza di una distorsione sistematica nella valutazione.

4.4.1 Disamina dei risultati relativi alla sezione di speaking

I risultati del confronto tra il modello e il *gold standard* nella sezione di produzione orale sono riportati nella Tabella 25. I valori di kappa pesata si collocano complessivamente nella fascia lieve-debole, con un valore nella categoria Totale pari a 0,23. Le categorie con accordo relativamente più elevato sono *Comprehension* (k = 0,255) e *Grammar* (k = 0,225), mentre *Fluency* (k = 0,060) e *Accuracy* (k = 0,034) registrano i valori più bassi, classificabili come accordo lieve.

Categoria	Kappa pesata	MAE	Bias	Accordo esatto	+1	≥+2	-1	≤-2
Lexis	0.22	1.44	-1.40	5%	0%	1%	47%	46%
Content	0.21	1.55	-1.55	8%	0%	0%	40%	52%
Grammar	0.22	1.73	-1.64	1%	1%	1%	36%	61%
Coherence	0.15	2.03	-2.01	1%	1%	0%	19%	79%
Comprehension	0.26	1.21	-1.19	13%	1%	0%	59%	27%
Fluency	0.06	0.71	-0.05	48%	9%	8%	29%	4%
Accuracy	0.03	0.61	-0.04	49%	14%	5%	31%	1%
Totale	0.23	1.48	-1.46	4%	1%	0%	46%	48%

Tabella 25:Confronto dei risultati ottenuti dal modello e dal gold standard nella sezione di produzione orale

I valori del MAE ci permettono di confermare e articolare questo quadro. Nel parametro *Totale*, il modello si discosta in media di 1,48 livelli QCER dal *gold standard*, con una distribuzione delle distanze che vede meno del 5% di accordo esatto, circa il 47% di scostamenti di un livello e il 48% di scostamenti di due o più livelli. Il valore del *bias* medio per la categoria

Totale conferma che tale scarto non è distribuito simmetricamente attorno al *benchmark* umano: il *bias* negativo si avvicina in valore assoluto al MAE, indicando che l'errore osservato è prevalentemente sistematico e non il risultato di fluttuazioni casuali. In particolare, si vede come il modello tenda a sottostimare la performance del candidato rispetto al *gold standard*, con particolare enfasi su *Grammar* e *Coherence*. Figura 43 mostra la distribuzione della distanza dall'accordo esatto per la categoria *Totale*.

I risultati ci permettono di affermare che il problema principale del sistema non risiede tanto nella capacità di distinguere livelli adiacenti, quanto nella calibrazione assoluta rispetto al *benchmark* esperto. Il modello sembra infatti in grado di preservare l'ordine relativo dei candidati, collocando le performance più solide al di sopra di quelle più deboli, ma tende a posizionarle su una scala sistematicamente disallineata rispetto a quella adottata dai valutatori umani.

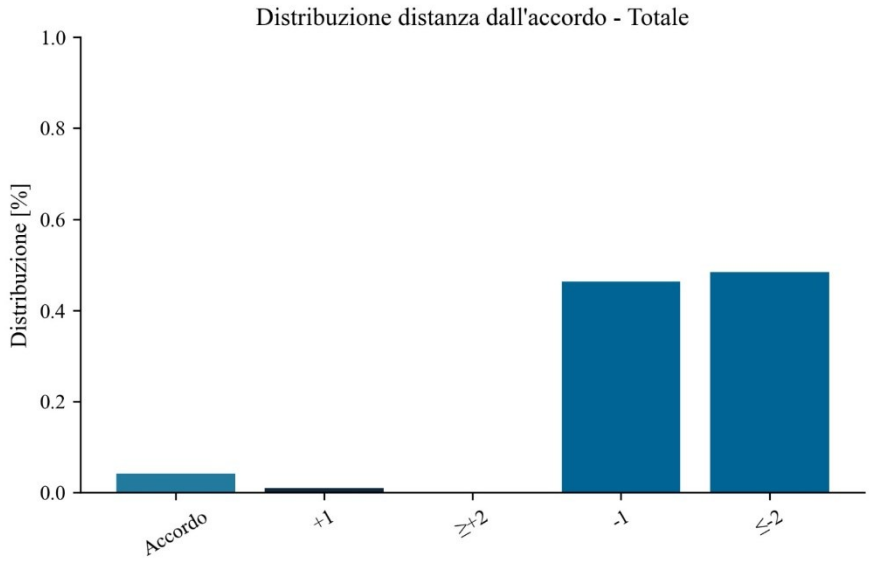


Figura 43: Distribuzione della distanza dall'accordo esatto per la categoria Totale.

Le differenze tra le categorie risultano molto marcate. *Coherence* è la dimensione più problematica registrando il MAE più elevato (2,03 livelli) e la distribuzione degli errori più sfavorevole, con solo l'1,1% di accordo esatto e il 78,9% di scostamenti pari ad almeno due livelli.

Questo risultato è coerente con quanto discusso nel paragrafo 3.3.2.7 sulle limitazioni strutturali della pipeline di valutazione orale. La coerenza discorsiva del parlato spontaneo richiede una comprensione dell'organizzazione del testo che va ben oltre la decodifica lessicale e

morfosintattica, ed è precisamente quella dimensione che la trascrizione ASR non è in grado di restituire fedelmente, privando il modello delle informazioni prosodiche e paralinguistiche necessarie.

A ciò si aggiunge un ulteriore fattore cruciale: le tecnologie STT rendono visibili nel testo scritto fenomeni tipici del parlato spontaneo come false partenze, riformulazioni, segnali discorsivi, ripetizioni e autocorrezioni che nell'oralità non compromettono necessariamente la percezione di coerenza. Nell'ascolto, tali elementi sono integrati nel flusso comunicativo e interpretati come strategie di pianificazione in tempo reale. Trasposti in forma scritta, invece, perdono il supporto prosodico che ne attenua l'impatto e finiscono per frammentare la progressione tematica, producendo un effetto di disorganizzazione più marcato. Inoltre, attuando una trascrizione su un parlato L2 il sistema di riconoscimento vocale può incontrare difficoltà nel riconoscimento di realizzazioni fonetiche non standard. Questo può generare errori di trascrizioni e l'introduzione nel testo di elementi grammaticalmente scorretti o semanticamente incoerenti. In questo modo, la trascrizione non solo perde informazioni rilevanti, ma può anche aggiungerne elementi fuorvianti, alterando artificialmente la coerenza interna del discorso e minando la valutazione dell'effettiva competenza del parlante.

All'estremo opposto si collocano *Fluency* e *Accuracy*, che presentano i MAE più contenuti (0,71 e 0,61 livelli). In questi casi la *pipeline* integra direttamente i punteggi forniti da Microsoft Azure, che analizza il segnale audio e non esclusivamente la trascrizione testuale. L'accesso diretto all'informazione acustica riduce la perdita informativa e migliora l'allineamento con il *gold standard* umano. Inoltre, in queste due categorie il *bias* medio è significativamente più ridotto. Questo indica che la minore distanza dal *benchmark* non è soltanto quantitativa – ovvero l'errore medio è più basso – ma anche qualitativa poiché l'errore non mostra una direzione sistematica stabile. Mentre in altre dimensioni emerge una tendenza marcata alla sottostima, in *Fluency* e *Accuracy* lo scarto è più bilanciato. L'accesso al segnale audio sembra quindi contribuire non solo a ridurre l'ampiezza dell'errore, ma anche la sua componente sistematica, favorendo una calibrazione più stabile rispetto alla valutazione esperta.

Un ulteriore elemento metodologico rafforza questa interpretazione: annotatori umani e modello non hanno lavorato sugli stessi input nel caso delle domande di *speaking*. Gli annotatori hanno valutato direttamente gli audio originali l'laddove il modello ha ricevuto una trascrizione e dei punteggi esterni per le categorie di *Fluency* e *Accuracy*. Ne deriva un'asimmetria informativa strutturale: nelle dimensioni che dipendono maggiormente

dall'organizzazione discorsiva globale (come *Coherence*), il modello parte da una rappresentazione meno completa e potenzialmente distorta rispetto a quella disponibile agli annotatori. Al contrario, nelle categorie in cui può contare su un accesso diretto (anche se mediato) al segnale audio, il divario si riduce sensibilmente. Questa differenza di input costituisce quindi una variabile decisiva per interpretare le divergenze osservate tra valutazione automatica e valutazione umana.

4.4.2 Disamina dei risultati relativi alla sezione di writing

Il confronto tra i risultati generati dal modello e quelli del *gold standard* umano fornisce, nella sezione di produzione scritta, dei risultati sostanzialmente migliori rispetto a quelli della produzione orale. I valori di kappa pesata si collocano uniformemente nella fascia moderata per tutte le categorie analitiche, con valori compresi tra 0,499 (riconducibile alla sezione di *Grammar*) e 0,559 (pertinente invece al parametro di *Content*) e un *Totale* di 0,524. L'omogeneità dei risultati tra le categorie indica che il modello mostra un comportamento consistente nelle diverse dimensioni della valutazione scritta, senza le marcate asimmetrie osservate per lo *speaking*. Tabella 26 riporta i risultati complessivi relativi alla sezione di produzione scritta.

Categoria	Kappa pesata	MAE	Bias	Accordo esatto	+1	≥+2	-1	≤-2
Lexis	0.51	0.98	-0.89	32%	0%	2%	38%	28%
Content	0.56	1.04	-0.96	24%	4%	0%	44%	27%
Grammar	0.50	1.12	-1.01	27%	3%	1%	37%	32%
Coherence	0.54	1.08	-0.97	26%	6%	0%	38%	31%
Comprehension	0.55	0.83	-0.83	42%	0%	0%	38%	20%
Totale	0.52	1.03	-0.97	26%	3%	0%	43%	28%

Tabella 26: Confronto dei risultati ottenuti dal modello e dal gold standard nella sezione di produzione scritta

Nonostante le migliori performance rispetto alla sezione di produzione orale, il MAE evidenzia ugualmente uno scostamento medio di 1 livello in quasi tutte le categorie osservate. L'analisi del *bias* medio consente di qualificare ulteriormente tale scarto: per tutti i parametri il *bias* risulta negativo e prossimo al livello -1, indicando una tendenza sistematica del modello a collocare i candidati in una fascia leggermente inferiore rispetto a quella indicata dal *gold standard*. La distribuzione degli scarti conferma tale direzionalità, mostrando una prevalenza di differenze negative, coerenti con una sottostima sistematica rispetto ai livelli attribuiti dal

gold standard. Figura 44 riporta la distribuzione di variazione nel livello misurato per il punteggio totale.

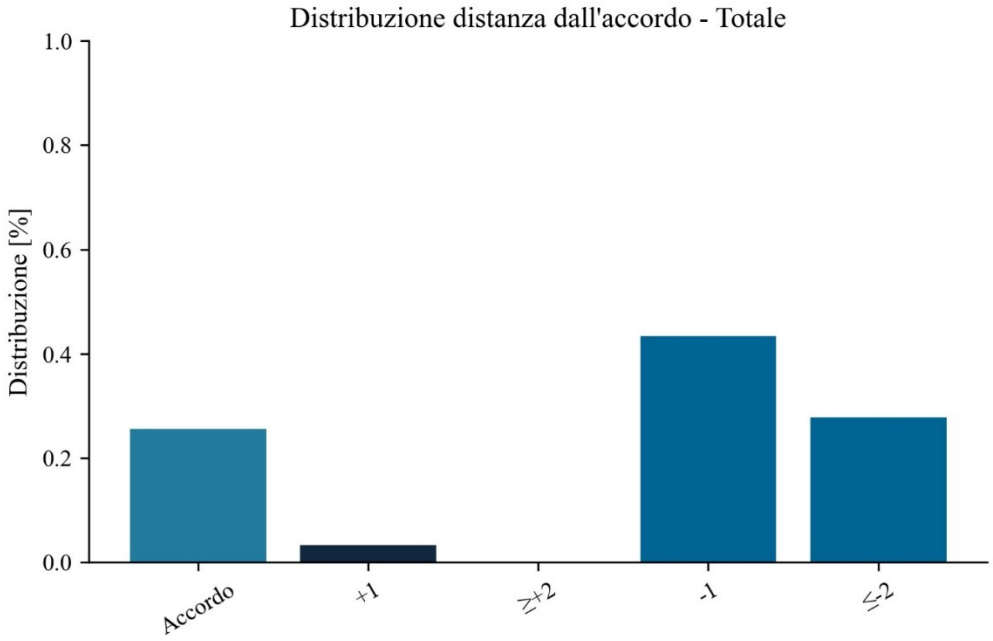


Figura 44: Distribuzione della variazione nel livello misurato per il punteggio totale

Il confronto tra *speaking* e *writing* evidenzia una differenza non solo quantitativa, ma anche qualitativa nella natura dell'errore. Nella sezione di produzione orale, lo scarto rispetto al *gold standard* appare quasi interamente sistematico: il *bias* medio coincide sostanzialmente con il MAE, indicando una tendenza stabile e strutturale alla sottostima. Nella produzione scritta, invece, la distanza media è più contenuta e meno polarizzata: pur essendo presente un *bias* negativo, esso non spiega interamente l'errore osservato, che risulta quindi meno unidirezionale e più distribuito.

Il divario tra le *performance* del modello nelle due sezioni può essere ricondotto alle differenti scelte metodologiche relative alla natura dell'input nei diversi task, espresse nel capitolo 3. Nel writing, il modello valuta direttamente il testo prodotto dal candidato, senza mediazioni ulteriori. Nello speaking, al contrario, il modello non opera sull'audio originale, ma su una trascrizione generata da sistemi ASR di Microsoft Azure. Questo passaggio intermedio introduce inevitabilmente errori e distorsioni che possono propagarsi nella fase valutativa. Si tratta di una conseguenza diretta del compromesso tra fattibilità tecnica e accuratezza della misurazione, esplicitamente discusso nel paragrafo 3.3.2.7.

4.5 Analisi dell'accordo tra autovalutazioni e modello

Un ulteriore elemento preso in esame per valutare le performance del modello è stato lo studio della validità relativa ai criteri, ovvero la correlazione dei risultati con una fonte esterna che possiede proprietà intrinseche riconosciute (cfr. paragrafo 2.2.1). Questo confronto è stato possibile grazie alle informazioni fornite dagli utenti nel questionario sociolinguistico; come discusso nelle sezioni iniziali di questo capitolo, gli utenti hanno espresso un giudizio sul proprio livello QCER relativo alla competenza linguistica generale (denominata Overall) e a quella specifica relativa alla produzione orale (chiamata Speaking). Il confronto con i risultati delle autovalutazioni permette di valutare la coerenza percepita dal candidato tra il proprio livello soggettivo e la stima prodotta dal sistema. Tali risultati sono riportati nella Tabella 27.

Categoria	**Kappa pesata**	**MAE**	**Bias**	**Accordo atteso**	**+1**	**≥+2**	**-1**	**≥-2**
Speaking	0.68	0.69	0.55	51%	24%	18%	7%	0%
Overall	0.52	1.04	-0.11	31%	24%	11%	18%	16%

Tabella 27: Confronto tra le autovalutazioni degli utenti e i punteggi assegnati dal modelo

Il confronto tra autovalutazione orale specifica e modello mostra un accordo sostanziale (k = 0,673) con un MAE di 0,70 e il 50% di corrispondenza esatta, risultato superiore a quello ottenuto nel confronto tra modello e *gold standard* umano. L'analisi del *bias* medio indica che, nel confronto con *Speaking*, il *bias* positivo (+0,55) segnala una lieve tendenza del modello a collocare i candidati su un livello leggermente superiore rispetto alla loro autovalutazione. A differenza del confronto con il *gold standard*, in questo caso la direzionalità dell'errore si inverte, evidenziando una lieve sovrastima rispetto alla percezione soggettiva dei partecipanti.

Questo fenomeno riflette la diversa calibrazione dei punteggi tra le varie fonti di valutazione, ovvero, il differente modo in cui i vari valutatori interpretano e applicano la stessa scala. Se il *gold standard* umano tende a posizionare i candidati più in alto, l'autovalutazione dei candidati è più conservativa – perché le persone tendono sistematicamente a sottovalutare la propria competenza – mentre il modello occupa infine una fascia intermedia. La maggiore convergenza tra modello e autovalutazione non implica che entrambi i sistemi sovrastimino rispetto al *benchmark* umano, ma riflette un diverso punto di riferimento. La distanza dal *gold standard* appare quindi riconducibile più a differenze di calibrazione tra criteri valutativi che a una sovrastima sistematica dei candidati.

Il confronto tra autovalutazione globale e modello evidenzia invece un accordo moderato ($\kappa = 0{,}513$) con un MAE di 1,07 e solo il 29,5% di corrispondenza esatta. Per Overall, il *bias* medio è vicino a zero (−0,11), indicando l'assenza di una distorsione direzionale marcata. Lo scarto medio osservato deriva prevalentemente dalla dispersione attorno all'autovalutazione, piuttosto che da uno spostamento sistematico verso l'alto o verso il basso.

Nel complesso, i dati suggeriscono che il modello tende a posizionarsi sistematicamente al di sotto del *gold standard* umano ma al di sopra dell'autovalutazione dei candidati, configurando una collocazione intermedia nella gerarchia delle stime. Questa dinamica permette di mettere in luce anche i limiti legati all'autovalutazione: sebbene utile per comprendere come i partecipanti percepiscono la propria competenza, non sempre riflette accuratamente le loro abilità reali. Come osservato da James C. McCroskey e da Linda L. McCroskey:

> «Nel caso della competenza comunicativa, le scale di autovalutazione possono essere molto utili se vogliamo sapere quanto una persona pensa di essere competente a livello comunicativo. Se vogliamo sapere quanto una persona sia effettivamente competente, tali scale potrebbero essere totalmente inutili, perché molto probabilmente la persona non lo sa. Molte persone pensano di essere comunicatori molto competenti, quando in realtà non lo sono. Altri credono di essere carenti di competenza, quando in realtà sono comunicatori molto adeguati» (McCroskey 1988:110; traduzione mia)[52].

Per concludere, la Figura 45 sintetizza la direzionalità dello scarto medio del modello rispetto ai tre *benchmark* considerati: *gold standard* nello *speaking*, *gold standard* nel *writing* e autovalutazione dei candidati. La rappresentazione consente di visualizzare in modo immediato non soltanto l'ampiezza dello scarto, ma soprattutto la sua direzione rispetto allo zero, distinguendo tra sottostima e sovrastima sistematica. Emerge chiaramente come il modello tenda a collocarsi al di sotto del *benchmark* esperto, in misura marcata nello speaking e più contenuta nel writing, mentre nel confronto con l'autovalutazione si osserva una lieve sovrastima. Il grafico conferma quindi la natura prevalentemente calibrativa del problema, più che una difficoltà nella discriminazione relativa tra candidati.

---

[52] « In the case of communication competence, self-report scales may be very useful if we want to know how communicatively competent a person thinks he/she is. If we want to know how competent the person actually is, such scales may be totally useless, because the person very likely does not know. Many people think they are very competent communicators, when in fact they are not. Others believe they are lacking in competence, when in fact they are very adequate communicators».

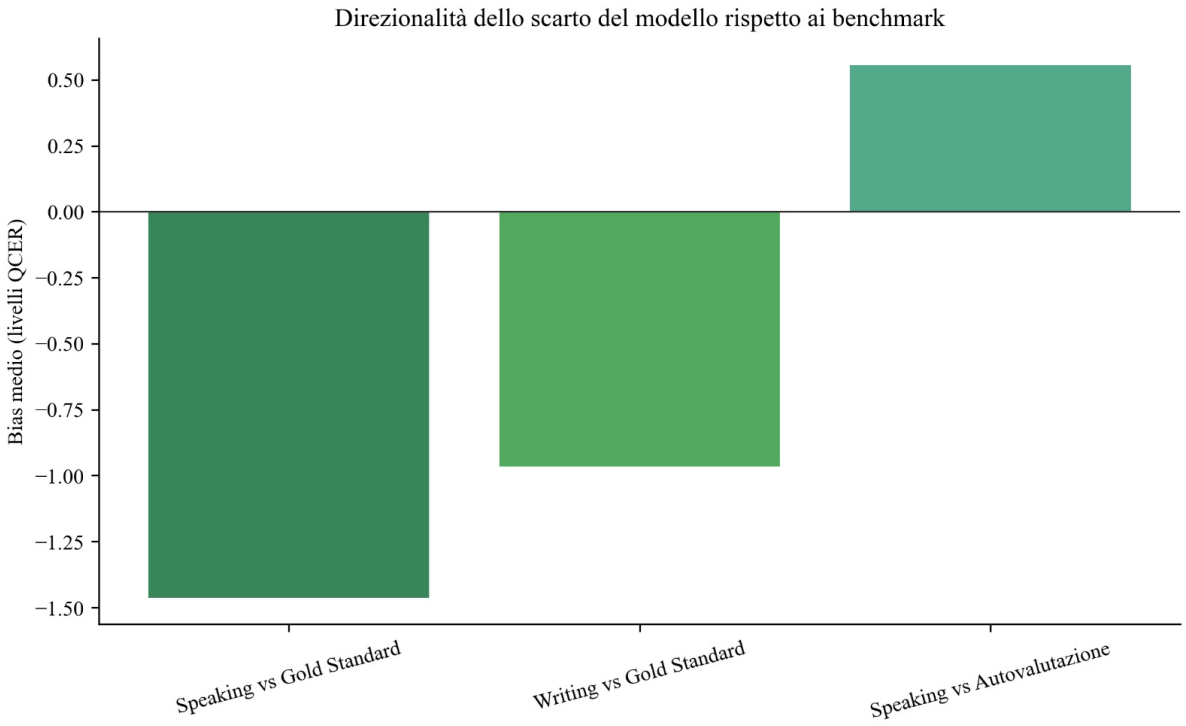


Figura 45:D direzionalità dello scarto medio del modello rispetto ai tre benchmark considerati

4.6 Analisi dell'esperienza utente e dell'usabilità della piattaforma

Al termine della somministrazione del test, i partecipanti sono stati invitati a compilare un questionario di feedback relativo all'esperienza di test e alla piattaforma ETET. Il questionario comprendeva domande a risposta chiusa – dicotomiche o basate su scala Likert – e domande aperte per la raccolta di osservazioni qualitative[53]. Su 40 partecipanti che hanno completato il test, 28 hanno risposto al questionario di feedback; le analisi che seguono si basano pertanto su questo sottogruppo di 28 rispondenti. Tabella 19 riporta la distribuzione delle risposte alle domande chiuse dicotomiche, mentre la Tabella 20 sintetizza i punteggi medi relativi alle scale Likert.

[53] La scala Likert è una modalità di misurazione che associa a ciascun *item* un punteggio numerico collocato lungo un *continuum* di accordo, generalmente compreso tra 0, corrispondente alla posizione di completo disaccordo, e 5, indicante il massimo livello di accordo

Indicatore	Sì	No	Percentuale di risposte affermative
Istruzioni chiare e intuitive	28	0	100,0%
Leggibilità dei caratteri adeguata	24	4	85,7%
Domande formulate in modo chiaro	26	2	92,9%
Domande orali ritenute utili	27	1	96,4%
Difficoltà tecniche riscontrate	4	24	14,3%
Tempo adeguato	10	18	35,7%
Numero di domande adeguato	26	2	92,9%
Quiz adatto al livello	25	3	89,3%
Copertura della comunicazione orale	26	2	92,9%
Test utile per valutare competenze L2	24	4	85,7%

Tabella 28: Distribuzione delle risposte alle domande chiuse del questionario di feedback

Indicatore	Media	Mediana	Range
Usabilità e grafica della piattaforma	4,18	4,0	2–5
Naturalezza delle voci TTS (listening)	3,89	4,0	1–5
Utilità complessiva del test (valutazione)	3,86	4,0	1–5

Tabella 29:Distribuzione delle risposte alle domande aperte del questionario di feedback

L'esame dei *feedback* degli utenti permette di analizzare come lo strumento venga percepito sotto tre aspetti fondamentali: l'usabilità e la chiarezza della piattaforma, la credibilità dello strumento come misura delle competenze comunicative e le difficoltà pratiche legate alla modalità di somministrazione digitale. Combinando i dati quantitativi e qualitativi, emerge un bilancio globalmente favorevole, pur in presenza di alcune problematiche che comportano implicazioni in termini di validità e di equità dello strumento.

Sul piano dell'usabilità, il giudizio complessivo è positivo: tutti i partecipanti hanno trovato le istruzioni chiare e accessibili e quasi la totalità ha ritenuto le domande formulate in modo comprensibile; il punteggio medio di usabilità (4,18 su 5) conferma che l'interfaccia non costituisce un impedimento alla corretta esecuzione del test. In termini di validità, questo dato è particolarmente significativo poiché indica che il livello di familiarità digitale richiesto

dall'uso dello strumento non funge da fattore discriminante tra gli utenti. La quasi totalità dei partecipanti, infatti, non ha segnalato difficoltà tecniche o di navigazione, suggerendo che l'uso dello strumento non introduce un carico cognitivo aggiuntivo legato alle abilità digitali.

Accanto a questo quadro positivo emergono tuttavia alcune criticità di accessibilità. Quattro utenti hanno segnalato problemi di leggibilità, dovuti principalmente alle dimensioni ridotte del testo in alcune sezioni dell'interfaccia e in specifici esercizi. Un'ulteriore criticità, distinta dalla precedente ma non meno rilevante per la praticabilità del test, riguarda la gestione della tastiera in modalità a schermo intero: un partecipante dotato di una tastiera non italiana ha riferito di non essere riuscito a digitare caratteri speciali della lingua italiana poiché la visualizzazione a schermo intero impediva il passaggio a una tastiera virtuale alternativa. Sebbene ci si aspetti che un utente che deve sostenere un esame di lingua si premuri di possedere gli strumenti necessari per scrivere nella lingua in esame, questo limite tecnico può essere facilmente superabile attraverso l'inserzione di un avviso preventivo prima dell'inizio del test o tramite l'integrare una tastiera virtuale nell'interfaccia. In un sistema destinato a un'utenza plurilingue e geograficamente distribuita, l'attenzione a questi aspetti di accessibilità operativa assume particolare rilievo.

Un risultato degno di nota concerne la ricezione della componente orale. La quasi totalità degli utenti riconosce il valore dell'inclusione nell'esame della lingua parlata e ritiene che il test rappresenti in modo adeguato la comunicazione orale, a conferma che l'impostazione multimodale dello strumento è coerente con una visione comunicativa della competenza linguistica in cui la dimensione orale occupa una posizione rilevante. Le voci sintetiche utilizzate negli esercizi di comprensione orale ottengono un punteggio medio di 3,89 su 5. Il posizionamento su una fascia medio-alta della scala suggerisce una percezione generalmente soddisfacente della qualità del TTS; tuttavia, la presenza di valutazioni basse indica che la naturalezza delle voci rappresenta un'area di potenziale miglioramento.

La maggioranza dei partecipanti considera inoltre il test adeguato al proprio livello e giudica appropriato il numero di domande. Questo indica una buona calibrazione generale della difficoltà e della lunghezza dello strumento Alcuni commenti qualitativi segnalano tuttavia la presenza di lessico percepito come tecnico o di attività considerate molto complesse. Tali osservazioni risultano coerenti con il fatto che una parte dei partecipanti ha dichiarato dei livelli di competenza relativamente bassi; questo dato suggerisce che la difficoltà percepita possa riflettere l'eterogeneità dei profili linguistici piuttosto che uno squilibrio strutturale del test.

Il dato più rilevante dell'intero questionario di feedback riguarda la gestione del tempo: quasi due terzi dei partecipanti lo giudicano insufficiente e numerosi commenti descrivono una pressione temporale percepita, soprattutto nelle attività di produzione scritta e nelle domande grammaticali. Si tratta della criticità più segnalata, con un'incidenza nettamente superiore rispetto a qualsiasi altro problema. Due utenti propongono esplicitamente il passaggio da un minutaggio per ciascuna domanda a un *timer* globale per l'intero test, che consentirebbe una gestione autonoma del tempo da parte del candidato in funzione della difficoltà percepita dei singoli item. Sebbene si tratti di una soluzione metodologicamente interessante, più vicina alle modalità valutative tradizionali, tale opzione non è attualmente supportata dalla piattaforma, nella quale ogni domanda viene costruita assegnando una tempistica relativa (cfr. 3.2). Alcuni partecipanti hanno inoltre posto l'attenzione sul fatto che i tempi stringenti avessero scaturito situazioni di ansia: «Per alcuni esercizi il tempo mi sembrava veramente poco, ti mette moltissima ansia. Andrebbe meglio forse avere il tempo generale e poter decidere come sfruttarlo». Questo aspetto è estremamente rilevante poiché la pressione temporale può incidere sulla prestazione indipendentemente dalla competenza linguistica del candidato, introducendo una potenziale fonte di distorsione valutativa.

Alcuni utenti riportano inoltre problemi nella registrazione della produzione orale. Pur trattandosi di un numero limitato di casi, la questione è da attenzionare poiché la componente orale dipende dalla disponibilità e dal funzionamento dei dispositivi. Questo aspetto introduce una potenziale forma di disuguaglianza, rappresentata dal differente livello di alfabetizzazione digitale e dagli strumenti a disposizione dei partecipanti. Ciò è determinante in situazioni di valutazione automatizzata poiché le pratiche di equità impongono che le condizioni di partenza dei partecipanti siano il più possibile uniformi. Per strumenti di valutazione automatizzata, questo aspetto è cruciale: l'equità del test richiede che le condizioni tecniche siano il più possibile uniformi.

La maggioranza dei partecipanti considera il test utile per valutare la capacità comunicativa, sebbene la valutazione media ricevuta da questa domanda sia leggermente inferiore rispetto ad altre dimensioni. Questo dato suggerisce che, pur essendo percepito come valido, lo strumento potrebbe beneficiare di una maggiore trasparenza dei criteri valutativi o di un feedback più esplicito sulle competenze misurate.

Nel complesso, l'analisi dei dati indica che lo strumento venga percepito come coerente con una valutazione comunicativa dell'italiano L2, grazie alla multimodalità e alla presenza della componente orale. Le criticità emerse – in particolare inerenti alla gestione del tempo e,

in misura minore, ad aspetti di leggibilità e tecnici – riguardano prevalentemente le condizioni di somministrazione piuttosto che il costrutto valutato. Tali criticità non mettono in discussione l'impostazione generale del test, ma individuano aree di ottimizzazione necessarie per garantire una maggiore equità e accuratezza valutativa. Nel complesso, i risultati supportano la validità d'uso dello strumento e suggeriscono che, con adeguate revisioni operative, esso possa costituire un mezzo efficace per la valutazione automatizzata della competenza in italiano L2.

Conclusioni

Il presente lavoro prende avvio dall'osservazione di una crescente domanda di strumenti per la valutazione linguistica dell'italiano L2, in particolare nei contesti caratterizzati da un'elevata richiesta di certificazione. A questa esigenza si contrappone tuttavia la sostanziale assenza di sistemi automatizzati integrati in grado di valutare in modo unitario l'insieme delle abilità linguistico-comunicative. La web-app ETET è stata progettata come risposta a questa lacuna, con l'obiettivo di dimostrare come l'automazione della valutazione linguistica non comporti necessariamente un indebolimento del quadro teorico di riferimento, ma possa essere realizzata mantenendo un solido ancoraggio alle teorie acquisizionali della *Second Language Acquisition*, ai principi metodologici del *Language Testing* e agli strumenti più recenti della Linguistica Computazionale.

Nel primo capitolo di questo lavoro è stata delineata la cornice teorica di riferimento, analizzando i principali modelli di competenza linguistico-comunicativa e il ruolo svolto dalle politiche linguistiche europee nella standardizzazione della valutazione. Allo stesso tempo, è stata ricostruita l'evoluzione tecnologica che ha interessato il *Language Testing*, mostrando come il passaggio dai test cartacei ai sistemi digitali e, più recentemente, alle tecnologie basate su intelligenza artificiale abbia progressivamente ampliato le possibilità di progettazione e somministrazione dei test linguistici.

Su queste basi teoriche, il secondo capitolo si è concentrato sulla definizione degli elementi da tenere in considerazione nella costruzione di un test linguistico. L'attenzione è stata rivolta in particolare alle caratteristiche che ne determinano l'utilità complessiva – quali la validità, l'affidabilità, l'autenticità, l'interattività, l'impatto e la praticabilità – e alle diverse fasi che guidano il processo di sviluppo di uno strumento valutativo. Questa analisi ha fornito il quadro di riferimento per la progettazione del test implementato nella piattaforma ETET, orientando sia la definizione del costrutto sia le scelte operative relative ai formati dei task e alle procedure di *scoring*.

Il terzo capitolo ha presentato il caso studio ETET, una piattaforma web-based progettata per la valutazione automatizzata dell'italiano L2. Il sistema integra tecnologie di trattamento automatico del linguaggio, modelli di machine learning e strumenti di riconoscimento vocale per consentire la costruzione e somministrazione scalabile di test linguistici e la valutazione automatizzata delle produzioni scritte e orali.

Il quarto capitolo ha infine sottoposto a verifica empirica il sistema attraverso una sperimentazione pilota, volta a indagare l'affidabilità e la validità di *scoring* del modello mediante il confronto con un gold standard costruito a partire dalle valutazioni fornite da tre annotatori umani esperti. Attraverso il calcolo della kappa di Cohen pesata con pesi quadratici è stato misurato il grado di accordo tra le coppie di annotatori, selezionando come riferimento la coppia con l'accordo più elevato. Il *gold standard* così ottenuto è stato quindi utilizzato come parametro di confronto per valutare le prestazioni del sistema automatico nell'assegnazione dei punteggi nelle domande di produzione scritta e orale.

La valutazione delle prestazioni è stata condotta attraverso un insieme di metriche complementari che permettono di osservare il comportamento del modello da prospettive differenti. In particolare, l'analisi si è basata sul calcolo della Kappa di Cohen pesata con pesi quadratici (K_w), del Mean Absolute Error (MAE) e del *bias* medio, indicatori che consentono di valutare rispettivamente il grado di accordo tra valutatori differenti, l'entità media assoluta dello scarto tra i punteggi assegnati dal modello e quelli del *gold standard* e la direzione dell'errore mantenendo il segno dello scarto tra le valutazioni.

Per la produzione scritta, il confronto ha evidenziato un accordo moderato e uniforme con il *benchmark* esperto ($\kappa = 0{,}524$ sul Totale), con un bias sistematico negativo che indica una tendenza alla sottostima contenuta da parte del modello. Per la produzione orale invece, i risultati sono sensibilmente più deboli ($\kappa = 0{,}231$ sul Totale), con un MAE medio di 1,48 livelli QCER e un *bias* che coincide quasi interamente con il valore dell'errore; questo dato segnala una sottostima sistematica nella valutazione piuttosto che fluttuazioni casuali nell'assegnazione del punteggio. Un'eccezione significativa è rappresentata dalle categorie di *Fluency* e *Accuracy*, per le quali la pipeline integra direttamente i punteggi acustici di Azure ottenendo MAE contenuti e bias prossimi allo zero.

La divergenza tra i risultati delle due abilità produttive non è riconducibile a una debolezza del modello di scoring in sé, ma a una criticità strutturale della *pipeline*: il sistema relativo alla sezione di speaking non opera sull'audio originale, bensì su trascrizioni generate dal modulo ASR di Microsoft Azure, che provoca una progressiva perdita di informazioni prosodiche indispensabili per valutare dimensioni come la coerenza discorsiva – che registra il MAE più elevato (con 2,03 livelli) – e la naturalezza dell'eloquio.

Accanto a questa problematiche tecniche, lo studio presenta ulteriori limiti metodologici. La dimensione del campione – 45 partecipanti, con 95 risposte orali valide dopo i criteri di esclusione – è sufficiente per una validazione pilota ma non consente generalizzazioni robuste,

in particolare per i livelli QCER meno rappresentati. La strategia di campionamento mista, che ha combinato contatti istituzionali e *snowball sampling*, introduce infatti un probabile *self-selection bias* verso apprendenti motivati, con maggiore competenza linguistica e dimestichezza tecnologica.

Nel complesso, i risultati della sperimentazione forniscono evidenze preliminari a sostegno della fattibilità di sistemi di valutazione automatizzata per l'italiano L2. Pur con alcune limitazioni, il modello mostra una capacità significativa di replicare *pattern* valutativi umani, confermando il potenziale delle tecnologie di intelligenza artificiale come supporto ai processi di *assessment* linguistico. Allo stesso tempo, i livelli di accordo osservati tra gli annotatori umani nel processo di costruzione del gold standard mettono in luce come la valutazione linguistica, anche quando condotta da esperti, presenti inevitabilmente margini di variabilità interpretativa. In questa prospettiva, l'impiego di sistemi automatizzati può contribuire a ridurre la componente di soggettività nelle procedure di *scoring*, favorendo una maggiore coerenza e uniformità nell'assegnazione dei punteggi.

I limiti finora evidenziati definiscono con precisione le direzioni di sviluppo da intraprendere per la realizzazione di versioni future del sistema. La priorità riguarda la sezione relativa alla valutazione della produzione orale: la sostituzione della *pipeline* basata su trascrizione ASR con un'architettura multimodale, capace di operare direttamente sul segnale audio, consentirebbe di recuperare le informazioni prosodiche oggi perdute – aumentando la coerenza interna e l'affidabilità delle valutazioni in relazione ai parametri di *coherence* e *grammar* – e di ridurre il *bias* sistematico documentato. In parallelo, un *fine-tuning* di GPT-4o su produzioni di apprendenti di italiano L2 annotate da esperti potrebbe migliorare la calibrazione del sistema sia per la scrittura sia per il parlato, riducendo la tendenza alla sottostima rispetto al *benchmark* umano. Infine, un'ulteriore linea di sviluppo riguarda la necessità di un monitoraggio sistematico delle prestazioni dei modelli linguistici impiegati nella pipeline di valutazione. Considerata la rapida evoluzione delle tecnologie e la continua introduzione di nuovi modelli sul mercato, risulta opportuno prevedere procedure periodiche di validazione, al fine di aggiornare progressivamente l'architettura del sistema e garantire nel tempo livelli adeguati di affidabilità e validità dello *scoring* automatico.

In conclusione, il presente lavoro evidenzia il potenziale innovativo dell'integrazione tra i quadri teorici della SLA, del LT e delle tecnologie di intelligenza artificiale, proponendo un contributo concreto allo sviluppo di sistemi per la valutazione automatizzata delle abilità produttive, ambito ancora poco esplorato nella valutazione dell'italiano L2. In questa

prospettiva, il caso studio ETET rappresenta un tentativo di tradurre tale integrazione teorica in uno strumento concreto di valutazione linguistica automatizzata. Strumenti di questo tipo non sono pensati per sostituire i sistemi certificativi tradizionali, ma piuttosto per affiancarli, ampliandone le possibilità di utilizzo. L'obiettivo è offrire soluzioni scalabili e accessibili, soprattutto nei contesti in cui le modalità di valutazione tradizionali risultano difficili da implementare. Nel complesso, la ricerca mostra come rigore metodologico e innovazione tecnologica non costituiscano obiettivi in contraddizione, ma possano essere perseguiti congiuntamente nello sviluppo di strumenti per la valutazione dell'italiano L2.

Bibliografia

Alderson J. C., 1990. *Learner-centered testing through computers: Institutional issues in individual assessment*, In J. H. A. L. de Jong & D. K. Stevenson (eds.) Individualizing the assessment of language abilities (pp. 20–27). Clevedon, UK: Multilingual Matters.

Alderson J. C., Clapham C., Wall D., 1995, *Language Test Construction and Evaluetion*, Cambridge: Cambridge University Press.

Al-Obaydi, Liqaa Habeb, Pikhart, Marcel e Klimova, Blanka. 2023. *ChatGPT and the General Concepts of Education: Can Artificial Intelligence-Driven Chatbots Support the Process of Language Learning?* International Journal of Emerging Technologies in Learning (iJET) 18, n. 21: 39-50.

Alte, 2011, *Manual for Language Test Development and Examining*, Council of Europe.

Attali Y., Runge A., LaFlair G. T., Yancey K., Goodwin S., Park Y., von Davier A. A., 2022. *The interactive reading task: Transformer-based automatic item generation*. Frontiers in Artificial Intelligence, 5.

Bachman L.F., Cohen A.D., 1998, *Interfaces between Second Language Acquisition and Language Testing Research*, Cambridge: Cambridge University Press.

Bachman L.F., 1990, *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press.

Bachman L.F., Palmer A., 1996, *Language Testing in Practice*, Oxford, Oxford University Press.

Bachman L. F., Palmer A., 2010, *Language Assessment in practice*, Oxford: Oxford University Press.

Baker F. B., 1985, *The basics of item response theory*, Portsmouth, NH: Heinemann.

Baker R., Siemens G., 2014, *Educational Data Mining and Learning Analytics*, R. Keith Sawyer (ed), Cambridge Handbook of the Learning Sciences. Cambridge University Press, pp. 253-274.

Balboni P.E., 2002, *Le sfide di Babele. Insegnare le lingue nelle società complesse*, Torino: UTET.

Barki P., Gorelli S., Marchetti S., Sergiacomo M. P., Strambi B., 2003, *Valutare e Certificare l'italiano di Stranieri. I livelli iniziali*, Perugia: Guerra Edizioni.

Barni M., 2005, *La valutazione delle competenze linguistico-comunicative in L2*, in Vedovelli M., Manuale della certificazione dell'italiano L2, Roma: Carocci Editore.

Bellugi U, 1967, *The acquisition of negation. Unpublished doctoral dissertation*, Cambridge, MA: Harvard University Press.

Benmamoun E., 2021, *Heritage Language Research and Theoretical Linguistics*, in Heritage languages and Linguistics, Cambridge University Press.

Berruto G., Cerruti M., 2019, *Manuale di Sociolinguistica*, Torino: UTET.

Birnbaum A., 1968, *Estimation of an ability*. In F. M Lord & M. R. Novick (Eds.), Statistical theories of mental test scores (pp. 423–479). Reading, MA: Addison-Wesley.

Bourdin D., Sichtmann C., Davvetas V., 2023, *The Influence of Employee Accent on Customer Participation in Services*.

Bridgeman B., Trapani C., Attali Y., 2012*, Comparison of human and machine scoring of essays: Differences by gender, ethnicity, and country*, Applied Measurement in Education, 25(1), 27–40.

Brown R., 1973, *A first language: The early stages*, Cambridge, MA: Harvard University Press.

Brown J. D., Hudson T., 2002, *Criterion-Referenced Language Testing*, Cambridge: Cambridge University Press.

Brown T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Amodei D., 2020, *Language Models are Few-Shot Learners*. arXiv.

Calvet L.J., 2002, *Le marché aux langues. Les effets linguistiques de la mondialisation*, Paris: Plon.

Canale M. e Swain M., 1980, *Theoretical bases of comunicative approaches to second language teaching and testing*, in Applied Linguistics 1 (1), pp. 1-47.

Carroll J. B., 1961, *Fundamental considerations in testing for English language proficiency of foreign students*, in Testing the English Proficiency of Foreign Students, 31–40. Washington, DC: Center for Applied Linguistics.

Chapelle C. A., 2001*, Computer applications in Second Language Acquisition*. Foundations for teaching, testing and research.

Chapelle C. A., Douglas D., 2006, *Assessing language through computer technology*. Cambridge, UK: Cambridge University Press.

Chen X., Zou D., Xie H. R., Su F., 2021, *Twenty-five years of computer-assisted language learning: A Topic modeling analysis*, Language Learning & Technology, 25(3), 151–185

Chomsky N., 1965, *Aspects of the theory of syntax*, Cambridge, Mass.: MIT Press.

Combei C.R., 2023, *Speaking Italian with a Twist. A Corpus Study of Perceived Foreign Accent*, Milano: FrancoAngeli s.r.l.

Consiglio d'Europa, 2001, *Quadro comune europeo di riferimento: apprendimento, insegnamento, valutazione*, La Nuova Italia.

Corder S. P., 1967, *The Significance of Learners' Errors*, International Review of Applied Linguistics in Language Teaching, 5, 161-170.

Davies A., 1990, *Principles of language testing*, Basil Blackwell, Oxford-Cambridge

Devlin J., Chang M. W., Lee K., Toutanova K., 2019, *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), 4171–4186.

De Mauro T., 1991, *Storia linguistica dell'Italia unita*, Laterza.

De Mauro T, Vedovelli, M. Barni, M., Miraglia, L., 2002, *Italiano 2000. Indagine sulle motivazioni e sui pubblici dell'italiano diffuso tra stranieri,* Roma: Bulzoni.

Dokukina I., Gumanova Y., 2020, *The role of chatbots in language learning*. Education and Information Technologies, 25, 1–17.

Donati N., Periani M., Di Natale P., Savino G., Torroni P., *Generation and evaluation of English grammar multiple-choice cloze exercises*, in F. Dell'Orletta, A. Lenci, S. Montemagni, R. Sprugnoli (Eds.), Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024), Pisa, Italy, 2024, pp. 325–334.

Elman J.L., 1990, *Finding structure in time*, in «Cognitive Science», 14, 2, pp.179-211.

Ferrari S., 2019, *Valutare le competenze orali in italiano L2. Variazione longitudinale e situazionale in apprendenti a livello avanzato*, Canterano: Gioacchino Onorati editore.

Field J., 2013, *Cognitive validity*, In Geranpayeh A., Taylor L. Examining Listening Research and practice in assessing second language listening, Cambridge: Cambridge University Press.

Firth J.R., 1957, *Model of Meaning*, in Papers in Linguistics, 1934-1951, London, Oxford University Press, pp. 190-215.

Fulcher G., 2000, *Computers in Language Testin*, In Brett, P. and G. Motteram (Eds) A Special Interest in Computers. Manchester: IATEFL Publications, 93–107.

Gardner R. C., Lambert W. E., 1972, *Attitudes and Motivation in Second Language Learning*, Rowley: Newbury House.

Geranpayeh A., Taylor L., 2013, *Examining Listening: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 35, Cambridge: UCLES/Cambridge University Press.

Green A., 2020, *Exploring Language Assessment and Testing. Language in Action*, x: Routledge.

Guerini F., Dal Negro S., 2007, *Contatto. Dinamiche ed esiti del plurilinguismo*, Roma: Aracne.

Gulliksen H., 1950, C*Theory of mental tests.* John Wiley & Sons, Inc.

Hadi M. U., Tashi Q. A., Qureshi R. et al., 2025, *A Survey on Large Language Models: Applications, Challenges, Limitations, and Practical Usage.*

Halliday M., 1976, *The form of a functional grammar*, in Kress G., System and Function in Language, Oxford University Press.

Hamp-Lyons L., 1997, *Washback, impact and validity: ethical concerns*. Language Testing, 14: 295-303.

Harris Z.S., 1954, *Distributional structure*, in «Word», 10, 23, pp.146-162.

Hayes J. R., Flower L. S., 1980, *Identifying the organisation of writing processes*, in Gregg, L W and Steinberg, E R (Eds) Cognitive processes in writing, Hillsdale, NJ: Lawrence Erlbaum Associates, 3-30.

Heift T., 2021, *Intelligent Computer Assisted Language Learning*, H. Mohebbi and C. Coombe (eds.), Research Questions in Language Education and Applied Linguistics, Springer Texts in Education.

Hochreiter S., Schmidhuber J, 1997, *Long short-term memory*, in «Neural Computation», 9, pp. 1735-1780.

Horwitz E. K., Horwitz M. B., Cope J., 1986, *Foreign Language Classroom Anxiety*, The Modern Language Journal, Summer, 1986, Vol. 70, No. 2 (Summer, 1986), pp. 125-13.

Huang W., Zou D., Cheng G., Chen X., Xie H., 2023, *Chatbots for language learning: A meta-analysis*. Computer Assisted Language Learning, 36(1–2), 1–30.

Hubbard P., 2009, *Computer Assisted Language Learning*, Volume 4: Present Trends and Future Directions in CALL. Critical Concepts in Linguistics Series. New York: Routledge.

Hughes A., 2003, *Testing for Language Teachers*. Cambridge: Cambridge University Press.

Hylton K., Levy Y., Dringus L. P., 2016, *Utilizing webcam-based proctoring to deter misconduct in online exams*, Computers & Education, Volumes 92-93, January-February 2016, Pages 53-63.

Hymes D. H., 1972, *On communicative competence*. In Pride, J.B. & Holmes, J. (eds),

Sociolinguistics: Selected Readings, 269–93. Harmondsworth, UK: Penguin.

Jakobson R., 1960, *Closing statement: Linguistics and poetics*, in Sebeok, T A (Ed.) Style in Language, Cambridge, MA: MIT Press, 350–377.

Ježek E., Sprugnoli R., 2023, *Linguistica computazionale. Introduzione all'analisi automatica dei testi*, Bologna: Il mulino.

Jurafsky D., Martin J. H., 2026, *Speech and Language Processing.*

Katinskaia A., 2025, *An overview of artificial intelligence in computer-assisted language learning.*

Khalifa H., Weir C. J., 2009, *Examining Reading: Research and Practice in Assessing Second Language Reading*, Studies in Language Testing volume 29, Cambridge: UCLES/Cambridge University Press.

Krashen S. D., 1981, *Second language acquisition and second language learning*. Oxford:
Pergamon Press.

Krashen S. D., 1985, *The input hypothesis: Issues and implications*, London: Longman.

Kurdi G., Leo J., Parsia B., Sattler U., Al-Emari S., 2020, *A systematic review of automatic question generation for educational* purposes. Int. J. Artificial Intell. Educ. 30, 121–204.

Labov W., 1972, *Some principles of linguistic methodology. Language in Society*, 1(1): 97-120.

Lado R., 1961, *Language Testing*. London: Longman Group Limited.

Landis J. R., Koch G. G., 1977, The measurement of observer agreement for categorical data, in «Biometrics», 33, 1, pp. 159-174.

Lenci A., Montemagni S., Pirelli V., 2005, *Testo e computer. Elementi di linguistica computazionale*, Roma: Carocci editore S.p.A.

Lenci A, 2013, *Linguistica computazionale*, In La linguistica italiana all'alba del terzo millennio (1997-2010).

Levelt W. J. M., 1989, *Speaking*, Cambridge, MA: MIT Press. Lynch S., 2022, Adapting Paper-Based Tests for Computer Administration: Lessons Learned from 30 Years of Mode Effects Studies in Education, Practical Assessment, Research & Evaluation, Volume 27 Number 22, August 2022.

Li S., 2023, *Working Memory and Second Language Oral Production: A Five-Year Systematic Review*, Proceedings of the International Conference on Global Politics and Socio-Humanitie.

Lord F. M., 1952, *A theory of test scores,* Psychomet Monogr 7.

MacIntyre P. D., Noels K. A., Clément R., 1997, *Biases in Self-Ratings of Second Language Proficiency: The Role of Language Anxiety*, Language Learning 47:2, June 1997, pp. 265-287.

Manning, C. D., & Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Mantovani S., 1998, *La ricerca sul campo in educazione. I metodi qualitativi*, Milano: Bruno Mondadori.

McCroskey J. C., McCroskey L. L., 1988*, Self-report as an approach to measuring communication competence,* Communication Research Reports, 5(2), 108-113.

McNamara T., 2000, Language Testing, Oxford: Oxford University Press

McNamara T., 2004, *Language Testing*, In Davies A., Elder C., The Handbook of Applied Linguistics.

McNamara T., Roever C., 2006, *Language Testing*: *The Social Dimension*, Hoboken: Wiley-Blackwell.

Messick S., 1989, *Validity*. In R. Linn (ed.), Educational Measurement. New York:

Macmillan, pp. 13–103.

Minaee S., Mikolov T. , Nikzad N., Chenaghlu M., Socher R., Amatriain X., Gao J., 2025, *Large Language Models*: A Survey, *ArXiv, abs/2402.06196.*

Mosqueira-Rey E., Hernández-Pereira E., Alonso-Ríos D., Bobes-Bascarán J., Fernández-Leal Á., 2022, *Artificial Intelligence Review*, 56:3005–3054.

Nassaji H., 2006, *The relationship between depth of vocabulary knowledge and L2 learners' lexical inferencing strategy use and success*, Modern Language Journal 90 (3), 387–401.

Noijons, J., 1994, *Testing computer assisted language tests: Towards a checklist for CALT,* CALICO Journal, 12(1), 37-58.

Novello A., 2009, *Valutare una lingua straniera: le certificazioni europee*, Libreria Editrice Cafoscarina srl.

Oller J.W., 1979, *Language tests at school*, London: Longman UK Group Limited.

Patton M. Q., 2002, *Qualitative research & Evaluation methods*, Thousand Oaks, Sage Publications.

Popham W. J., 1981, *Modern educational measurement*, Englewood Cliffs, NJ: Prentice-Hall.

Porcelli G., 1992, *Educazione linguistica e valutazione*, Torino: Petrini.

Radford A., Narasimhan K., Salimans T., Sutskever I., 2018, *Improving Language Understanding by Generative Pre-Training,* OpenAI technical report.

Radford A., Wu J., Child R., Luan D., Amodei D., Sutskever I., 2019, *Language Models are Unsupervised Multitask Learners (GPT-2)*, OpenAI.

Raffel C., Shazeer N., Roberts A., Lee K., Narang S., Matena M., Zhou Y., Li W., Liu P. J., 2020, *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*, Journal of Machine Learning Research 21, pp. 1-67.

Rasch G., 1960, *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Nielsen & Lydiche.

Rod E., 1997, *Second Language Acquisition,* Oxford: Oxford University Press.

Rodriguez M. C., 2005, *Three Options Are Optimal for Multiple-Choice Items: A Meta-Analysis of 80 Years*, Educational Measurement: Issues and PracticeVolume 24, Issue 2 pp. 3-13 of Research.

Roever C., 2001, *Web-Based Language Testing, Language Learning & Technology*, May 2001, Vol. 5, Num. 2 pp. 84-94, Carsten Roever University of Hawai'i at Manoa.

Rumelhart D., Hinton G., Williams R., 1986, *Learning representations by back-propagating errors*, in «Nature», 323, pp. 533-536.

Sahlgren M., 2008, *The distributional hypothesis*, Italian Journal of Lingusitics.

Sari H. İ., Manley H. A. C., 2016, *Computer Adaptive Multistage Testing: Practical Issues, Challenges and Principles*, Journal of Measurement and Evaluation in Education and Psychology, Vol. 7, Issue 2, Winter, 388-406.

Scardamalia M., Bereiter C., 1987, *Knowledge telling and knowledge transforming in written composition*, in Rosenberg, S (Ed.) Advances in Applied Psycholinguistics, Volume 2: Reading, writing and language learning, Cambridge: Cambridge University Press, 142-75.

Shaw S. D., Weir, C.J, 2007, *Examining Writing: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 26, Cambridge: UCLES/Cambridge University Press.

Shermis M. D., Burstein J., 2013, *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY: Routledge.

Shizuka T., Takeuchi O., Yashima T., Yoshizawa Y., 2006, *A comparison of 3 and 4 option English tests for university entrance selection purposes in Japan*, Lang Test, 23(1), 35–57.

Shohamy E., 1998, *How can language testing and SLA benefit from each other?The case of discurse*. In Bachman L.F., Cohen A.D, Interfaces between Second language Acquisition and Language Testing Research, Cambridge: Cambridge University Press.

Selinker R., 1969, *Language transfer.* General Linguistics 9 (2), 67-92.

Spinelli B., Parizzi F., 2010, *Profilo della lingua italiana. Livelli del QCER A1, A2, B1 e B2*. Firenze: La Nuova Italia.

Spolsky B., 1995, *Measured Words, The Development of Objective Language Testing*, Longman.

Sprugnoli L., 2005, *La costruzione delle prove*, In Vedovelli M., Manuale della Certificazione dell'Italiano L2, Roma: Carrocci.

Suvorov R., Hegelheimer V., 2014, *The Companion to Language Assessment*, John Wiley & Sons.

Suvorov R., 2024, *The use of eye tracking in validating L2 listening assessments*. in Yu G. & Xu J., Language Test Validation in a Digital Age.

Taylor R. P., 1980, Introduction. In R. P. Taylor (Ed.), *The computer in school: Tutor, tool, tutee* (pp. 1-10). New York: Teachers College Press.

Taylor L., 2011, (Ed.) *Examining Speaking: Research and Practice in Assessing Second Language Speaking*, Studies in Language Testing volume 30, Cambridge: UCLES/Cambridge University Press.

Ullman M. T., 2001, *The Declarative/Procedural Model of Lexicon and Grammar*, in Journal of Psycholinguistic Research, 30, 1, pp. 37-69.

Urquhart A.H., Weir C. J., 1998, *Reading in a Second Language: Process, Product and Practice*, Harlow: Longman.

van Ek J. A., 1973, *Analysis of the Problems Involved in Defining, in Operational Terms, a Basic Competence Level in Foreign language Learning by Adults*, in J.L.M. Trim, R. Richterich, J. van Ek e D.A. Wilkins, Systems Development in Adult Language Learning, Strasburgo, Consiglio d'Europa.

van Ek J.A., 1975, *The Threshold Level*, Strasburgo, Consiglio d'Europa.

Van Dijk T., 1977, *Text and Context: Explorations in the Semantic and Pragmatics of Discourse*, London: Longman.

VanPatten B., Smith M., Benati A., 2020, *Key Questions in Second Language Acquisition. An Introduction*, Cambridge: Cambridge University Press.

Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A. N., Kaiser Ł., Polosukhin I., 2017, *Attention is All You Need*, in U. von Luxburg et al (a cura di), Proceedings of the 31st International Conference on Neural Information Processing Systems, Red Hook, NY, Curran Associates, pp. 6000-6010.

Vedovelli M., 2021, *Storia linguistica dell'emigrazione italiana nel mondo*, Carrocci.

Vignoli A., Combei C.R., Zappulla F., (2025), *Verso la valutazione automatizzata dell'italiano L2: ETET tra LLM e tecnologie vocali*, Proceedings fo the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025), Cagliari, Italy, 282-291.

Wang Z., Arndt A. D., Singh S. N., Biernat M., Liu F., 2013, "*You Lost Me at Hello": How and when accent-based biases are expressed and suppressed*, International Journal of Research in Marketing, Volume 30, Issue 2, Pages 185-196.

Weigle S. C., 2002, *Assessing Writing,* Cambridge University Press: Cambridge.

Weir C. J., 1993, *Understanding and Developing Language Tests*, New York: Prentice Hall.

Weir C. J., 2005, *Language Testing and Validation: An Evidence-based Approach*, Basingstoke: Palgrave Macmillan.

Widdowson H.G., 1972, *The Teaching of English as Communication*, «English Language Teaching», XXVII (1), pp. 15-19.

Wu X., Xiao L., Sun Y., Zhang J., Ma T., He L., 2022, *A Survey Of Human-In-The-Loop For Machine Learning*, Future Generation Computer Systems.

Yavuz F., Çelik Ö., Yavaş Çelik G., 2024, *Utilizing large language models for EFL essay grading: An examination of reliability and validity in rubric-based assessments*. British Journal of Educational Technology, 56(1), 150–166.

Young D. J., 1991, *Creating a Low-Anxiety Classroom Environment: What Does Language Anxiety Research Suggest?*, The Modern Language Journal, Winter, 1991, Vol. 75, No. 4 (Winter, 1991), pp. 426-439.

Zechner K., Evanin K., 2020, *Automated Speaking Assessment. Using Language Technologies to Score Spontaneous Speech*, Routledge, Abingdon: UK.

Sitografia

https://www.alte.org/ (Ultimo accesso 10/03/2026)

https://www.coe.int/en/web/common-european-framework-reference-languages/reference-level-descriptions (Ultimo accesso 10/03/2026)

http://ethnologue.com/language/ita/ (Ultimo accesso 10/03/2026)

https://cvcl.unistrapg.it/ (Ultimo accesso 10/03/2026)

https://www.unistrasi.it/ (Ultimo accesso 10/03/2026)

https://certificazioneitaliano.uniroma3.it/ (Ultimo accesso 10/03/2026)

https://plida.dante.global/it (Ultimo accesso 10/03/2026)

https://www.ibm.com/history/805-scoring-test (Ultimo accesso 10/03/2026)

https://calico.org/ (Ultimo accesso 10/03/2026)

https://www.unistrapg.it/profilo_lingua_italiana/site/index.html (Ultimo accesso 10/03/2026)