Tracciare la competenza linguistica in italiano L2: cosa ci rivela SELF in Comprensione Orale

Triscia Biagiotti, Cristiana Cervini, Monica Masperi

Tracciare la competenza linguistica in italiano L2: cosa ci rivela SELF in Comprensione Orale

Triscia Biagiotti, Université Grenoble Alpes, Cristiana Cervini, Università di Bologna, Monica Masperi, Université Grenoble Alpes

Citation: Biagiotti, Triscia, Cristiana Cervini, Monica Masperi (2021) “Tracciare la competenza linguistica in italiano L2: cosa ci rivela SELF in Comprensione Orale”, in Monica Masperi, Cristiana Cervini, Yves Bardière (eds.) *Évaluation des acquisitions langagières : du formatif au certificatif*, *mediAzioni* 32: A97-A133, http://www.mediazioni.sitlec.unibo.it, ISSN 1974-4382.

1. Introduzione

La progettazione di un test per la valutazione di competenze linguistiche, quale è SELF (Cervini, Jouannaud 2015; Cervini 2016; Cervini, Masperi 2021), comporta un’ampia riflessione sul modello di lingua da proporre e sul concetto di competenza da adottare. Nell’ambito del testing questa riflessione si traduce nella definizione del *costrutto* del test. Con il termine *costrutto* ci si riferisce a tutti quegli aspetti della conoscenza e delle abilità in possesso del candidato che si intendono valutare. In altri termini, identificare un costrutto richiede di esplicitare una “modellizzazione” della competenza (MacNamara 2000; Barni in Vedovelli 2005). Nel caso di SELF, è la *competenza comunicativa* – intesa come competenza ad usare la lingua in un contesto sociale – a costituire il costrutto, oggetto di misura. È ampiamente documentato e anche facilmente intuibile, che un tale costrutto non è *direttamente* misurabile. A questo riguardo, si è ritenuto indispensabile supportare la creazione dei contenuti del test con procedure esplicite e rigorose, non definite a priori, ma elaborate per mezzo di una *ricerca-azione* condotta in concomitanza con lo sviluppo informatico del dispositivo. Il presente contributo è finalizzato alla discussione dei risultati della versione italiana di SELF, utilizzato per il posizionamento degli studenti negli ultimi tre anni

accademici presso le istituzioni francesi. Più precisamente, dirigiamo la nostra attenzione sulla comprensione orale e sul comportamento degli item che testano questa abilità, osservando il rapporto tra costrutto e valori psicometrici.

1.1. L'identità di SELF

SELF è uno strumento didattico concepito come punto d'appoggio alle azioni di formazione con l'obiettivo di tracciare e valorizzare il profilo linguistico dell'apprendente. In quest'ottica, SELF è destinato a fornire in ingresso (a fini di posizionamento) e in itinere (a fine diagnostico) risultati che attestino un'immagine attendibile delle competenze pregresse e dei margini di progressione del candidato. Alla base della progettazione del dispositivo vi è l'esigenza di adottare uno stesso approccio metodologico, qualunque sia la lingua target, in termini di coerenza del costrutto, creazione dei contenuti, processi di validazione e di visualizzazione dei risultati. A questo presupposto si è aggiunta l'altrettanto imprescindibile necessità di approdare alla realizzazione di un dispositivo tecnopedagogico flessibile e adattabile, che tenga conto delle esigenze di tutti gli attori della valutazione formativa a livello istituzionale, sul piano didattico (ricercatori, insegnanti e studenti) e amministrativo. SELF si presenta dunque in sintesi come un sistema operativo completo, dotato di player, tool autore dei task e strumenti di gestione dei risultati e di somministrazione delle sessioni (Cervini, Masperi 2021).

Le abilità linguistiche testate da SELF sono tre: comprensione orale, scritta ed espressione scritta breve[1]. Per progettare il test, ci siamo avvalsi di diversi strumenti di consultazione (referenziali, *learner corpora*, sillabi interni redatti da esperti della disciplina e conoscitori del pubblico di riferimento) che hanno guidato gli autori nell'identificazione dei focus linguistico-comunicativi, in linea con il livello di competenza di un candidato o di un "gruppo prototipico" e conformemente al costrutto stabilito (Cervini 2016). In altri termini, la ricerca iniziale inerente al ciclo

[1] Con la denominazione "espressione scritta breve" ci riferiamo a task che richiedono ai candidati l'inserimento di singole parole o di brevi sintagmi all'interno di un testo scritto autentico.

di progettazione del test (ALTE 2011: 14) e riguardante le fasi di definizione del costrutto, ha consentito di riflettere non solo sul modello di lingua e di cultura che si intende veicolare ma anche su come veicolarlo. Nella fattispecie, tanto i task di SELF considerati nella loro integralità, quanto i singoli item, possono presentare diverse focalizzazioni (morfosintattica, lessicale, comunicativa, pragmatica) e sollecitare diverse operazioni cognitivo-procedurali, tra cui citiamo l'inferenza (di una o più informazioni implicite, di uno stato d'animo, di una intenzione comunicativa), la comprensione estesa o puntuale, l'interazione (simulata), la riformulazione, l'individuazione di un elemento specifico, anche al fine di apportare una correzione o un miglioramento al testo (*ibid: 68*).

1.2. Il costrutto di Comprensione orale in SELF

Il ruolo centrale dell'abilità di ascolto nel processo di acquisizione linguistica ci ha condotti a porre in primissimo piano la progettazione della valutazione di questa abilità in SELF. In particolare, la stretta relazione tra il livello di comprensione orale e il successo in un test di lingua straniera, anche se ancora scarsamente documentata (Hossein 2012[2]; Quintin 2011[3]), ci sembra costituire un indicatore di grande interesse nonché un promettente spunto di indagine in campo diagnostico-formativo.

[2] I risultati a cui approda l'analisi dei dati rilevati dalla somministrazione del test IELTS su un campione di 701 studenti hanno confermato un livello di correlazione altamente significativa tra i risultati della prova di comprensione orale e il risultato complessivo dell'esame IELTS (Hossein 2012: 657-663).

[3] Presentando dei risultati empirici ottenuti tramite un confronto tra il tasso di superamento della certificazione CLES 1 (B1) e CLES 2 (B2) tra il 2010 e il 2011 e i risultati al test diagnostico Dialang. J.J. Quintin rileva una correlazione significativa tra il livello ottenuto in comprensione orale nel test e il livello globale dello studente. In merito al CLES1 sul campione preso in esame, la correlazione è perfetta e biunivoca: se un apprendente ottiene il livello B1 in comprensione orale nel test Dialang, possiamo predire che il livello certificato per mezzo del CLES sarà B1, e viceversa.

Diversamente da quanto avviene generalmente nei test più diffusi, che presentano in fase di valutazione della comprensione orale un'alternanza tra scritto e orale (testo fonte orale – consegne e quesiti in forma scritta), il costrutto di comprensione orale di SELF prevede che tutti gli input (contesto, testo fonte, domande, opzioni di risposta) siano oralizzati. Questa scelta del "tutto all'orale" è motivata dall'esigenza di aderire il più possibile all'autenticità della lingua orale, dove la parola è codificata sotto forma di suono, si svolge in tempo reale con scarse possibilità di ripetizione, ed è sociolinguisticamente marcata da forme che la distinguono dalla lingua scritta. Ritenuta in un primo tempo azzardata, soprattutto riguardo alla progettazione nelle lingue tipologicamente distanti dai sistemi europei, quali il mandarino o il giapponese, questa opzione "tutto all'orale" ha riscosso in fase sperimentale un largo consenso presso il pubblico target, diventando una sorta di "marchio di fabbrica" di SELF.

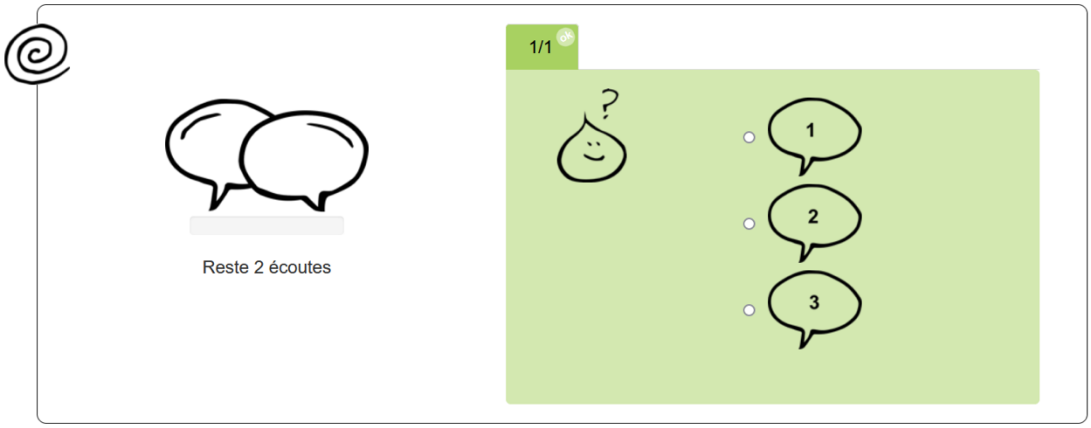

Fig. 1 Esempio di task "tutto all'orale".

Premesso questo, cercheremo ora di capire con quali presupposti e con quali strumenti viene testata la comprensione orale in italiano.

In primo luogo ci sembra importante menzionare la scelta di esporre il candidato a una lingua neo-standard che conceda delle aperture verso una vasta gamma di varietà diatopiche, poco marcate ma riconoscibili. La scelta di supporti audio e video autentici è stata preferita alle registrazioni in studio, qualora le fonti selezionate fossero liberamente fruibili a scopo didattico. All'atto pratico, tuttavia, il materiale autentico si è rivelato scarsamente disponibile per i livelli elementari, per i quali si è reso necessario l'adattamento dei testi e il ricorso a una strumentazione tecnica. La naturalità e il flusso dell'eloquio, come anche le

variazioni prosodiche, sono elementi ai quali si è rivolto un grande interesse: è noto infatti che possono notevolmente incidere sulla complessità del task, e in special modo in riferimento alla comprensione degli elementi pragmatici del discorso veicolati da fatti intonativi.

In situazione reale, l'ascolto è sempre dettato da una ragione o da un particolare interesse. A questo riguardo, e pur concedendo che un'attività di ascolto in ambito formale di valutazione poco ha da spartire con un atto d'ascolto in ambiente "ecologico", ci è sembrato importante esplicitare *il focus* dell'ascolto proposto, valorizzando in primo luogo il *contesto*. A tal fine, si è cercato di privilegiare il ricorso a un'informazione chiara ma breve, fornita in supplemento al testo-fonte, che partecipi a compensare il deficit informativo inerente alla situazione nel quale è posto il candidato al test. In merito al tipo di *ricezione orale*, si è cercato di variare l'esposizione a generi di testo, proponendo supporti monologali e dialogali correlati il più possibile all'autenticità situazionale, e includendo un formato di tipo "dialogo interrotto", che le opzioni di risposta permettono di completare. Tra le attività e i processi che il test intende sollecitare (cfr. infra 1.1.) ci preme sottolineare l'interesse rivolto alla competenza pragmatica, tutt'ora ancora poco indagata in ambito di valutazione (Ferrari *et al*. 2016), e che SELF tenta di apprezzare a partire da un livello elementare di conoscenza della lingua (Higashi, Shirota 2019).

2. La comprensione orale passata al setaccio: osservazioni e analisi dei dati raccolti

2.1. Il processo di validazione di un test di lingua (ciclo del testing) e le analisi successive

Le varie tappe che hanno portato alla creazione della prima versione del test in italiano (e nelle altre lingue del progetto[4]) si basano sulla letteratura di riferimento

[4] L'italiano è, insieme all'inglese, la lingua pilota del progetto IDEFI Innovalangues (ANR-11-IDFI-0024, Masperi, 2011). La versione "posizionamento" di SELF è, allo stato attuale, disponibile nelle seguenti lingue: italiano, inglese, cinese mandarino, giapponese, spagnolo, francese lingua

più nota in ambito di Language Testing (Bachman, Palmer 1996; McNamara 2000; Purpura 2004) e riflettono le raccomandazioni delle principali associazioni di Language Testing. La progettazione ha seguito le tappe illustrate nel ciclo del testing (cfr. fig. 1, Cervini 2016), al fine di rispettare i principi di validità, affidabilità e utilità. Il processo di costruzione di SELF è altresì caratterizzato dalla ciclicità e dall'iteratività, in quanto le decisioni che vengono prese in una determinata fase possono essere rimesse in discussione durante le fasi successive.

L'obiettivo della presente ricerca è incentrato sull'osservazione del comportamento dei task di comprensione orale in SELF. Per le analisi psicometriche sono state adottate la teoria classica del test (*Classical Test Theory*, d'ora in avanti CTT)[5] e la Teoria di Risposta all'Item (*Item Response Theory*, d'ora in avanti IRT) e in particolare il Modello di Rasch a 1 Parametro[6].

La CTT ci ha permesso in particolare di: i) verificare il grado di coerenza interna del test (attraverso il valore espresso dall'alfa di Cronbach, che consente di osservare se tutti gli item stiano misurando la medesima caratteristica, nel nostro caso l'abilità linguistica); ii) analizzare la difficoltà (espressa dal valore P[7]) e il potere discriminante (espresso dall'indice di discriminazione Rit o Rir[8]) di ogni

straniera (FLE). Si auspica inoltre di progettare due ulteriori versioni in tedesco e in FLE internazionale.

[5] Le analisi sono state effettuate tramite il programma Tia Plus (Cito 1998-2013).

[6] Le analisi sono state effettuate tramite il programma Winsteps (Linacre 2012).

[7] Il valore P viene espresso per ogni singolo item ed è dato dalla percentuale di candidati che hanno risposto correttamente all'item. In CTT, tale valore è espressione dell'indice di difficoltà di un item: più il valore è alto, più l'item è considerato facile (dato che una percentuale maggiore di candidati vi risponde correttamente) e viceversa.

[8] I valori Rit e Rir sonon due indici di discriminazione che, sulla base di un calcolo matematico, traducono la misura in cui l'item è in grado di discriminare i candidati più abili (cioè quelli che hanno tendenza a rispondere correttamente) da quelli meno abili (cioè quelli che tendono a rispondere in maniera errata). Il manuale del programma TiaPlus (Cito 2013) indica che il valore Rit (*item test correlation*) calcola la correlazione esistente tra il punteggio del singolo item e il punteggio totale del test. Tale valore può essere compreso tra -1.0 e +1.0. Un Rit > 0 indica che l'item è coerente in quanto i candidati più abili vi hanno risposto correttamente e i meno abili hanno fornito una risposta socrretta; un Rit = 0 indica che non esiste correlazione tra l'item in questione e il resto del test; un Rit < 0 indica che l'item si comporta in maniera incoerente in

singolo item; iii) studiare il comportamento delle risposte corrette (chiavi) e dei distrattori degli item contenuti nei vari task del test.

Come verrà esplicitato più avanti (cfr. infra 3.1), la IRT ci ha consentito di assegnare un valore specifico ad ogni singolo item e ad ogni singolo candidato. Item e candidati sono stati collocati su una scala di misura di tipo lineare[9]. In altre parole, grazie alla IRT è stato possibile ottenere scale di misura che ordinano gli item a seconda della loro difficoltà e i candidati a seconda della loro capacità, ovvero del loro grado di abilità linguistica.

2.2. Analisi del campione

Le analisi psicometriche sono state condotte su un campione di 1499 studenti che hanno effettuato il test di italiano negli anni accademici 2018-19, 2019-20, 2020-21. Procediamo in primo luogo con una descrizione accurata del campione, specificando: l'Università o l'Istituzione di provenienza, l'anno di iscrizione, il livello di competenza linguistica (espresso in SELF tramite un feedback del tipo "*en route vers B2*[10]") che indica anche il gruppo classe di appartenenza dello

quanto candidati abili forniscono una risposta errata e candidati non abili rispondono in maniera corretta. Il volre Rir (*item-rest correlation*) è un indice simile al Rit ma in questo caso il calcolo della correlazione non include il punteggio dell'item in questione. Si tratta del valore di riferimento, per quanto riguarda l'indice di discriminazione in CTT e può essere compreso tra -1 e +1. Un Rir < 0 è sempre considerato non discriminante, per quanto riguarda i valori positivi, le soglie di riferimento variano a seconda del contesto. Per SELF abbiamo considerato "accettabili" gli item con Rir >= 0.15 applicando la seguente distinzione: un item viene considerato molto discriminante se ha un Rir >= 0.30; un valore compreso tra 0.15 e 0.30 significa che il potere discriminante è meno elevato e ciò può essere sintomo di eventuali malfunzionamenti che devono essere confermati o smentiti attraverso l'analisi congiunta di altri valori psicometrici e del contenuto dell'item.

[9] "Una scala di misura è lineare se l'unità di misura è costante lungo tutta la scala" (Penta *et al*. 2005: 56, nostra traduzione).

[10] Il feedback viene elaborato sulla base di un "punteggio aggregato" (*score agrégé* in francese); si tratta di un punteggio complessivo che aggrega, secondo calcoli che variano da lingua a lingua, i risultati ottenuti nelle tre abilità testate da SELF: comprensione orale, comprensione scritta ed espressione scritta breve. L'indicazione che viene fornita non si riferisce al livello di competenza acquisito bensì al livello del corso al quale lo studente è invitato a iscriversi. Per questo motivo è

studente e il livello di competenza linguistica per abilità (comprensione orale, comprensione scritta e espressione scritta breve), il tempo impiegato per lo svolgimento del test (si ricorda infatti che SELF non prevede vincoli di tempo), la lingua materna e le altre lingue di riferimento dei candidati.

2.2.1. Università o Istituzione di provenienza

Il test viene utilizzato in numerose istituzioni universitarie francesi[11]; tuttavia, la maggioranza degli apprendenti (il 63,5%) che hanno adottato SELF italiano sono iscritti a Università e Istituti di istruzione superiore situati nei dipartimenti dell'Isère, della Savoia e Alta Savoia[12]. L'Università di Nizza, l'Università di Corsica Pasquale Paoli e l'Ecole Normale Supérieure di Lione rappresentano le Istituzioni con il maggiore numero di test SELF italiano somministrati, al di là degli Istituti posti sotto la tutela della Académie di Grenoble[13].

2.2.2. Anno d'iscrizione

Rispetto all'anno di iscrizione è interessante notare che il 64% del campione (pari a 960 studenti) è composto da iscritti al primo anno di formazione universitaria. Questo dato indica che la versione italiana del test viene utilizzata principalmente per il posizionamento delle matricole nei corsi di lingua, dunque come test di livello iniziale. In effetti, gli studenti che seguono una formazione linguistica

stato scelto di indicare il gruppo-classe consigliato tramite la dicitura "*en route vers*…", cioè "verso il livello…". Questa scelta riflette una concezione di apprendimento linguistico in quanto processo in evoluzione.

[11] Cfr. il sito istituzionale Innovalangues: https://innovalangues.univ-grenoble-alpes.fr/

[12] Si tratta in particolare di: Université Grenoble Alpes, Ecole Nationale Supérieure d'Architecture de Grenoble, Institut d'Etudes Politiques de Grenoble, Institut Polytechnique de Grenoble, Université Savoie Mont-Blanc.

[13] Si tratta di una circoscrizione che raggruppa gli istituti dei dipartimenti dell'Ardèche, della Drôme, dell'Isère, della Savoia e Alta Savoia.

solitamente ottengono una validazione del loro livello di competenza linguistica alla fine del corso annuale (o comunque dopo aver seguito e superato con successo due semestri di corso del medesimo livello) e l'anno successivo sono ammessi direttamente nei gruppi di livello superiore senza bisogno di effettuare nuovamente il test.

Se ci soffermiamo sul sottocampione di studenti iscritti al primo anno osserviamo che più della metà degli studenti (53%) ottiene un risultato al test pari o superiore a B1 (punteggio aggregato a partire da *en route vers B2.1*, "verso il B2.1").

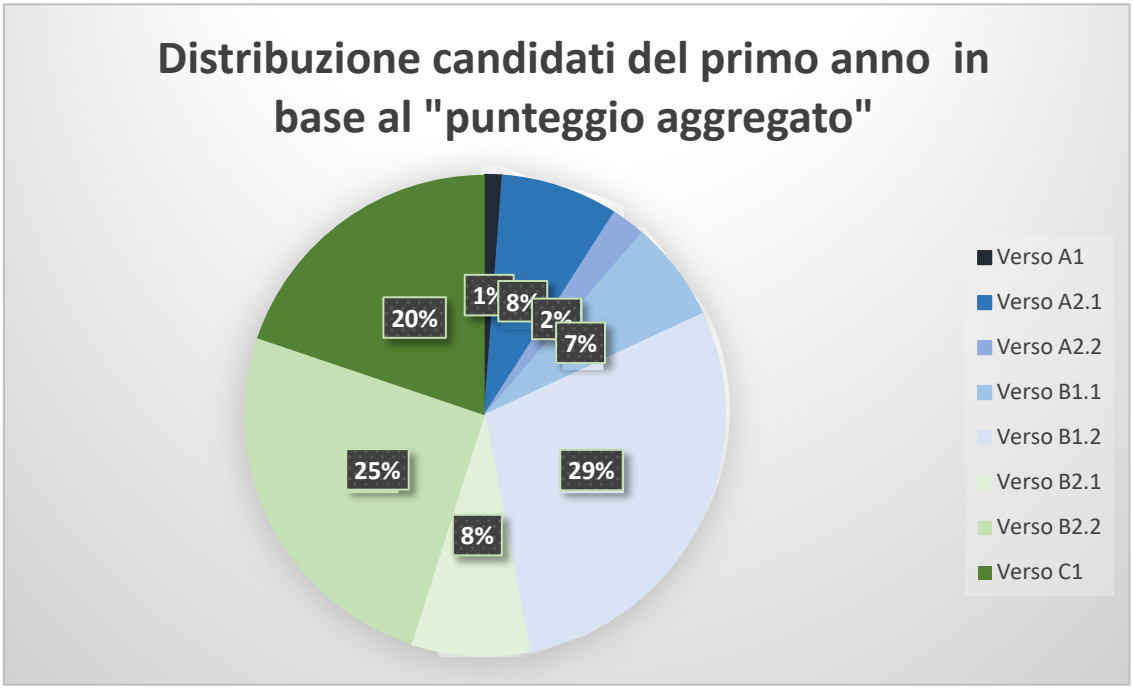

Grafico 1 Distribuzione dei candidati del I anno in base al punteggio aggregato.

Questo dato riflette, seppur in maniera parziale, la diffusione dell'italiano nel sistema scolastico francese: secondo i dati pubblicati annualmente dal rapporto *Repères et références statistiques: enseignement, formation, recherche*[14] l'italiano è scelto come seconda lingua straniera dal 5,3% degli alunni di scuola

[14] Repères et références statistiques : enseignement, formation, recherche 2020. Ministère de l'Education Nationale, de la Jeunesse et des Sports. Disponibile al link: https://www.education.gouv.fr/reperes-et-references-statistiques-2020-1316. Dati riferiti al 2019.

secondaria di primo e di secondo grado[15], mentre rappresenta in assoluto la terza lingua straniera più studiata, scelta dal 36,5% degli alunni di scuola secondaria di secondo grado.

2.2.3. "Punteggio aggregato" e "punteggio per abilità"

Alla fine del test SELF ogni studente riceve i risultati ottenuti e in particolare il punteggio aggregato (cfr. infra 2.2) di competenza linguistica e il livello per ogni abilità testata (nello specifico, comprensione orale, comprensione scritta e espressione scritta breve).

Fig.2 Screenshot dei risultati ottenuti alla fine di SELF.

Il grafico sottostante illustra la ripartizione dell'insieme dei candidati in funzione del gruppo target che si consiglia di frequentare sulla base del risultato ottenuto al test. Come possiamo osservare, nella maggior parte dei casi vengono suggerite formazioni di livello B1.2, B2.2 e C1. Questo significa quindi che il livello che si postula come acquisito è rispettivamente il B1.1, B2.1 e B2.

[15] Nel 96% dei casi, l'inglese è la lingua straniera più studiata come prima lingua (LVA) nella scuola secondaria di primo e secondo grado.

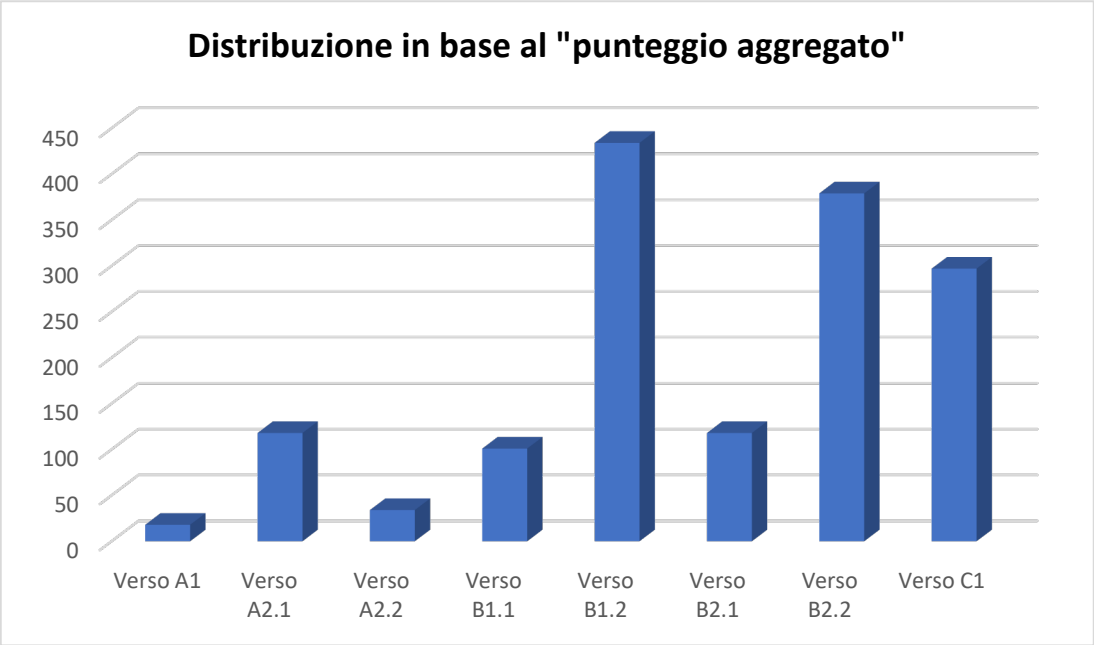

Grafico 2 Distribuzione del campione in base al punteggio aggregato.

Tale dato trova conferma anche per quanto riguarda il livello di competenza linguistica in comprensione orale. La maggior parte del campione ottiene infatti un livello uguale o superiore al B1 (cfr. infra 1.2).

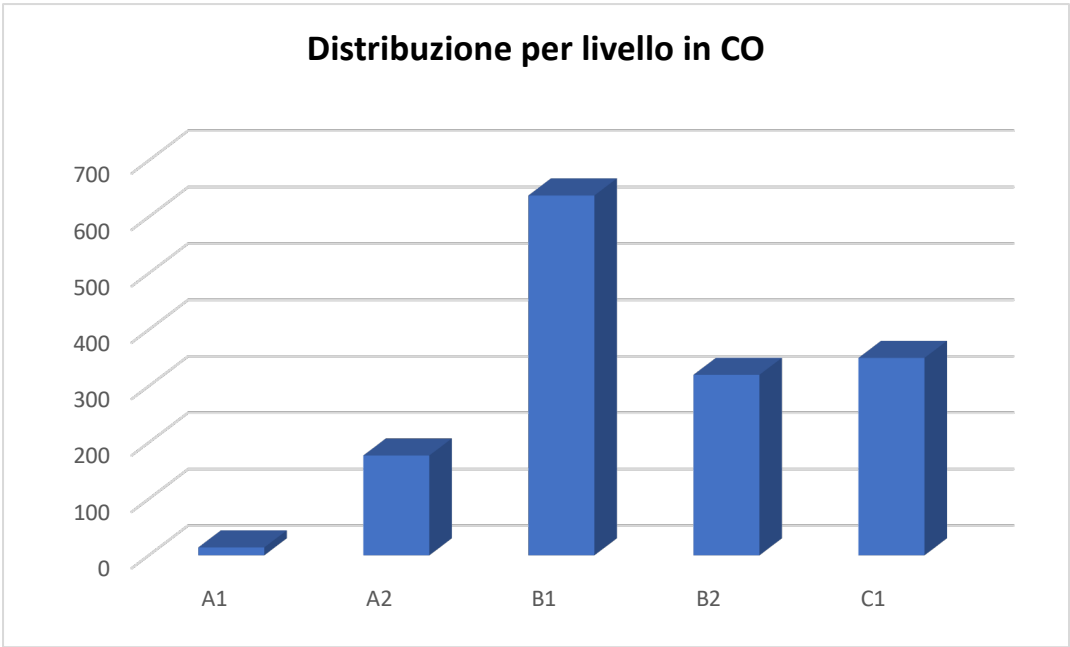

Grafico 3 Distribuzione del campione per livello di comprensione orale.

2.2.4. Durata del test

Osserviamo ora il tempo impiegato dai candidati per lo svolgimento del test. La durata del test costituisce un criterio non trascurabile per garantirne la fattibilità: la durata eccessiva di un test di posizionamento può in effetti incidere negativamente sulla performance degli studenti così come sull'organizzazione logistica del test. Questi due fattori sono stati tenuti in considerazione per fare in modo che il test avesse una durata media di 60 minuti.

Rispetto al campione in esame, i dati raccolti indicano che la maggioranza dei candidati (il 54%) effettua il test in meno di 60 minuti.

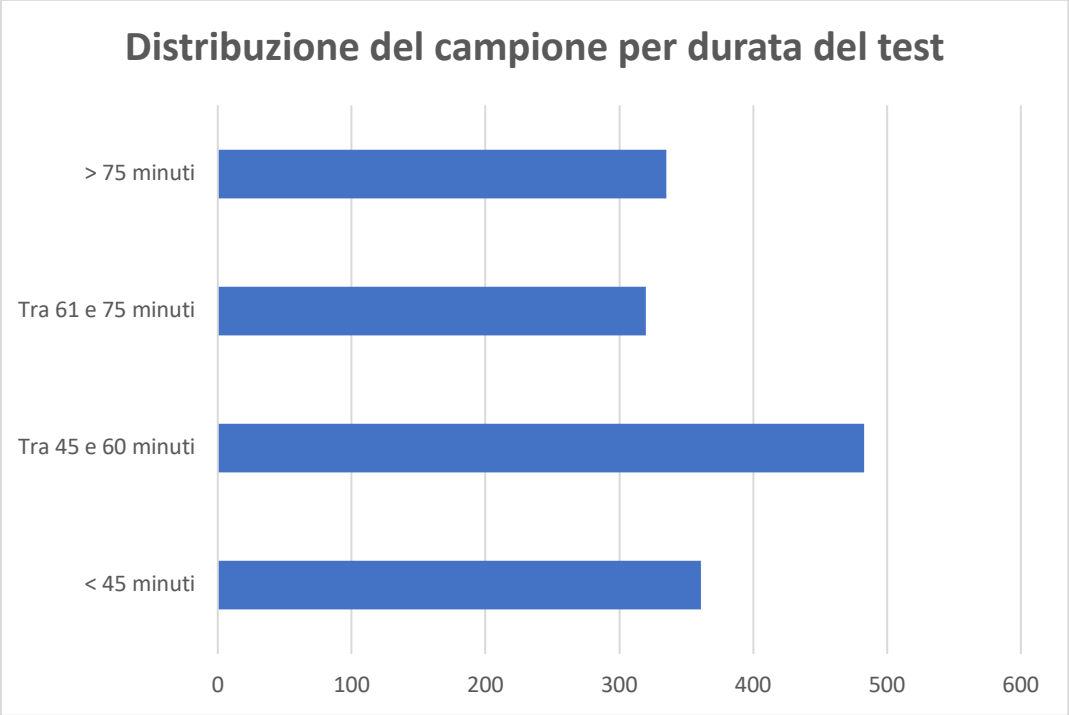

Grafico 4 Distribuzione del campione per durata complessiva del test.

Dal Grafico 4 si osserva che la maggioranza degli studenti effettua il test in un lasso di tempo situato tra i 45 e i 75 minuti e che il 43% dei candidati effettua il test in più di 60 minuti.

Se mettiamo in relazione la durata del test con i risultati ottenuti dai candidati, osserviamo correlazioni interessanti: il 24,6 % di candidati con tempi di svolgimento inferiori ai 45 minuti ha ottenuto "en route vers A1" e "en route vers A2" nel risultato finale. Allo stesso modo il 34,5% di candidati con tempi di

svolgimento inferiori ai 45 minuti ha ottenuto “en route vers B2.2” e “en route vers C1” nel risultato finale. Nel primo caso, si ipotizza che gli studenti di livello inferiore all’A1 o di livello A1, non avendo le capacità necessarie per rispondere a tutti gli item, possono essere portati a rispondere in maniera aleatoria, soprattutto durante la tappa del *minitest*, comune a tutti i candidati, che include item di comprensione orale ed espressione scritta breve calibrati di livello A2, B1 e B2. Nel secondo caso invece i candidati che ottengono dei livelli molto alti svolgono il test in tempi rapidi (meno di 61 minuti).

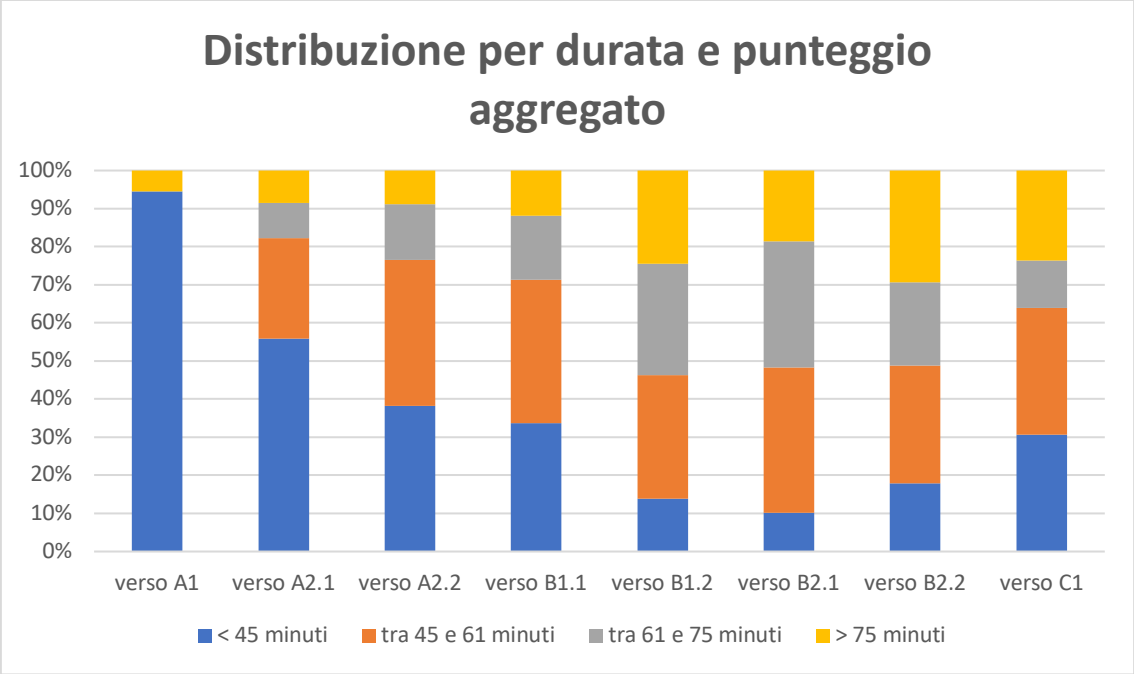

Grafico 5 Distribuzione del campione per durata del test e punteggio aggregato.

Un’ulteriore considerazione è quella relativa ai candidati che impiegano più di 75 minuti per effettuare il test. Come si può osservare dal Grafico 5, che mette in evidenza la differenza di durata per ogni livello target, nei due sottogruppi dei livelli più avanzati (“verso B2.2” e “verso C1”) si concentra maggiormente la percentuale, pur minima, di candidati che effettuano il test in più di 75 minuti. Questo dato può trovare un’interpretazione nel fatto che i candidati che vengono posizionati nei livelli più elevati sono sottoposti a una versione del test che include task più complessi e di più ampia durata (soprattutto per quanto riguarda la comprensione orale). Eccezion fatta per il sottogruppo di livello “verso B1.2”, i dati sono complessivamente confortanti quanto alla durata impiegata per

effettuare il test, in quanto in ogni sottogruppo la maggioranza dei candidati termina il test in una durata compresa tra i 45 e i 60 minuti.

2.2.5. Lingua materna

Un ultimo dato che emerge dall'analisi del campione è quello relativo alla biografia linguistica. A tal proposito è doveroso precisare che SELF è stato progettato in un contesto di formazione universitaria francese e dunque per un pubblico target specifico, quello di studenti universitari francofoni. Tale scelta non ne preclude tuttavia l'utilizzo in altri contesti, previe adeguate sperimentazioni sui pubblici target.

Anche se il campo informativo relativo alla lingua materna non è obbligatorio, 1244 candidati (l'82,98% del campione totale) lo hanno compilato: non è sorprendente rilevare che la maggioranza del campione (l'87,62%) sia composta da studenti che indicano come prima lingua di riferimento il francese. Seguono poi l'italiano (5,3%), l'arabo classico o il dialetto marocchino (l'1,68%) e le lingue germanico-slave (1,52%).

Nel grafico seguente viene illustrata la distribuzione del campione per lingua materna. Le lingue sono suddivise in tre categorie: lingue distanti (albanese, arabo classico o dialetto marocchino, cinese/cantonese, greco, ungherese, persiano, peul, turco e wolof), lingue germanico-slave (inglese, lettone, norvegese, polacco, russo, tedesco e ucraino) e lingue romanze (catalano, francese, italiano, spagnolo, malgascio, portoghese e rumeno).

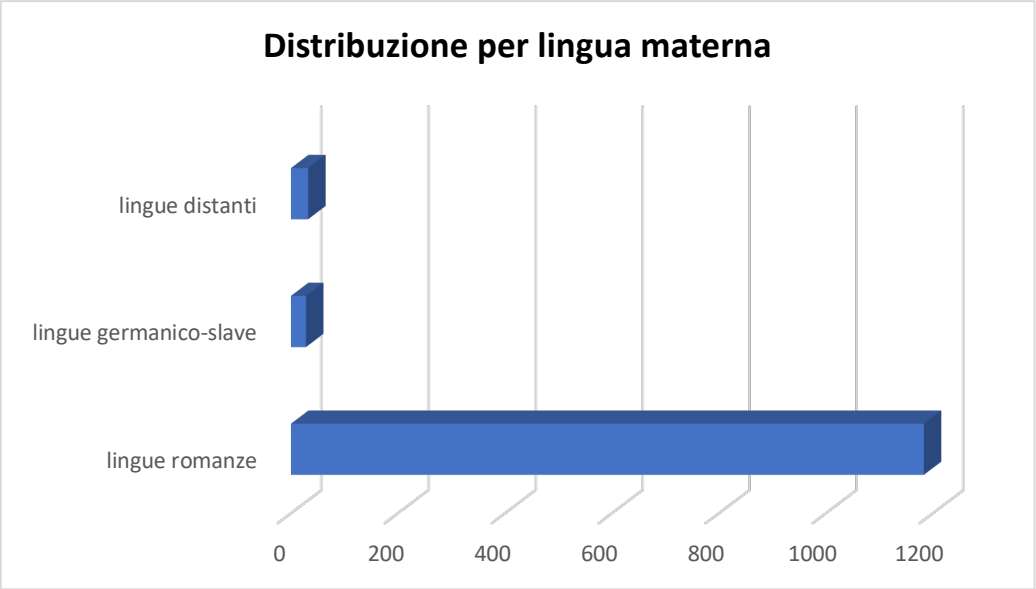

Grafico 6 Distribuzione per lingua materna.

3. SELF e psicometria

Il termine psicometria deriva dall'associazione delle due forme greche *psyché* (*ψυχή*) e *metron* (*μέτρον*), cioè anima, psiche e misura; il significato letterale del termine è dunque "misura dell'anima, della psiche". Inizialmente definita come una branca della psicologia sperimentale che misura fenomeni psichici in relazione ai parametri di intensità, durata e frequenza (Aliotta 1905; Spearman 1907), la psicometria nasce precisamente dall'esigenza di misurare determinate caratteristiche dell'essere umano (Penta *et al*. 2005) e viene applicata in diversi ambiti delle scienze umane (quali ad esempio la psicologia, la sociologia, la medicina o ancora la didattica). Le caratteristiche legate alla *psyché* dell'essere umano non sono direttamente osservabili e la loro "misurazione" deve spesso appoggiarsi su processi inferenziali (Cervini, Martari 2020). Come si possono misurare quindi caratteristiche umane come la sincerità, l'empatia, il raziocinio o, come nel nostro caso, la competenza in L2? In psicometria tali caratteristiche umane vengono chiamate *variabili latenti*, *tratti latenti*, *attributi* o *fattori* (Penta *et al*. 2005) e per poterle quantificare è necessario ricorrere a strumenti di misura che rispondano a tre criteri fondamentali: la validità, l'affidabilità e la fattibilità.

Le tecniche di analisi psicometrica, indispensabili nelle fasi di validazione e di assemblaggio di SELF, possono rivelarsi molto utili anche per il monitoraggio periodico degli item e per approfondire determinati aspetti di ricerca.

Che cosa ci dice dunque la psicometria sugli item di comprensione orale del test SELF in italiano? In che misura consente di aprire nuove prospettive di osservazione? È possibile individuare delle costanti in item che presentano valori psicometrici migliori rispetto ad altri e viceversa? Quali considerazioni possiamo fare a livello didattico a partire dai valori psicometrici rilevati?

3.1. Analisi psicometriche

Per rispondere a queste domande abbiamo applicato la CTT e la IRT avvalendoci dei programmi TiaPlus (Cito 1998-2013) e Winsteps (Linacre 2012). Ai fini della presente ricerca, le analisi si sono concentrate sulle seguenti tappe del test[16]:

- il *minitest*, comune a tutti i candidati e composto in totale da 31 item di livello A2, B1 e B2[17] di comprensione orale e espressione scritta breve;
- la seconda tappa, composta da 40 item in totale di livello A2, B1 e B2 di tutte e tre le abilità;
- la terza tappa, composta da 28 item in totale di livello B2 e C1 di tutte e tre le abilità.

Per le tre tappe abbiamo proceduto, in un primo tempo, all'analisi tramite la CTT e successivamente tramite la IRT. Anche se l'applicazione delle due teorie statistiche ci ha permesso di giungere a valori e conclusioni simili in termini, ad esempio, di difficoltà e discriminazione degli item, la scelta di adottare entrambe

[16] Come esplicitato in precedenza, la maggioranza dei candidati svolge la seconda o la terza tappa; pertanto, il numero dei candidati sulla prima tappa non era sufficiente per ottenere risultati significativi.

[17] L'attribuzione di livello ad ogni singolo item è stata determinata da una procedura di standard setting (bookmark method) sulla base dei valori MEASURE ottenuti tramite l'IRT.

è motivata dalle differenze teoriche alla base di CTT e IRT. La CTT si basa sul postulato secondo cui il risultato ottenuto al test da un candidato (X) risulta dalla somma tra il punteggio reale (T) e l'errore di misura (E) (Leavault, Gregoire 2014; Penta *et al*. 2005). Tuttavia i calcoli matematici che vengono effettuati per generare i valori psicometrici in CTT si basano sulla matrice di punteggi osservati sul campione di candidati in questione (Penta *et al*. 2005). Ciò significa che, secondo i criteri della CTT, tali valori dipendono imprescindibilmente dalle specificità del campione in esame e questo può condurre a trarre conclusioni inesatte. Ad esempio, se il campione è composto da individui con un basso livello di padronanza linguistica, un item avrà un indice di difficoltà (P) basso e sarà perciò considerato difficile ma lo stesso item testato su un campione di soggetti con un alto livello di padronanza linguistica otterrà un indice di difficoltà più alto e risulterà dunque più semplice (Levault, Gregoire 2014). Anche se è possibile ridurre il rischio di incorrere in interpretazioni sbagliate proponendo un campione più ampio e rappresentativo (ALTE 2011: 76), quando si intende effettuare una generalizzazione dei dati raccolti, la IRT si rivela essere uno strumento d'indagine più adeguato. Il programma Winsteps consente di effettuare analisi psicometriche con la IRT e più precisamente con il modello di Rasch a un parametro. Tale modello, di natura probabilistica, si distingue dai modelli di tipo deterministico, come ad esempio il modello di Guttmann (1944, 1950), che prevede che la possibilità di rispondere correttamente a un item dipenda dalla capacità del soggetto[18]. Senza attardarci eccessivamente su questioni legate ai calcoli matematici e statistici, a differenza dei modelli di tipo deterministico, il modello di Rasch consente di generare delle scale di misura lineari in funzione

[18] Confrontando due candidati, uno più abile (candidato a) e uno meno abile (candidato b) che rispondono allo stesso item (item x), il modello di Guttmann prevede che a risponda correttamente e che b fornisca una risposta sbagliata. Il punto critico di tale modello è dovuto al fatto che, in presenza di due candidati di capacità diversa ma comunque superiore alla difficoltà dell'item, risulta impossibile distinguerli tra loro: il modello infatti prevede che entrambi i candidati forniranno la medesima risposta (corretta), pur avendo livelli di capacità differenti (Penta *et al*. 2005: 27). Il modello di Rasch si basa invece sul postulato che "la probabilità di rispondere correttamente a un item aumenta in maniera continua in funzione della capacità della persona. Di conseguenza, se due persone hanno due livelli di capacità differenti ma comunque superiori al livello di difficoltà di un item, è possibile distinguerle grazie a questo modello poiché le loro probabilità di rispondere correttamente all'item sono diverse" (ibidem, nostra traduzione).

della capacità dei candidati al test e del livello di difficoltà degli item. A ciascun candidato e a ciascun item viene infatti assegnato un valore, espresso in logit[19], ed entrambi vengono conseguentemente posizionati su una medesima scala di misura. A differenza della CTT, la IRT e in particolare, il modello di Rasch, consente dunque di ottenere una certa indipendenza dei risultati dal campione di riferimento.

A questo riguardo, dopo aver lanciato le analisi psicometriche con la IRT, abbiamo deciso di comparare i risultati ottenuti al test, espressi tramite il punteggio aggregato con la scala di valori generata dal modello di Rasch in riferimento ai candidati. Come accennato poc'anzi, infatti, il modello assegna ad ogni candidato un valore espresso in *logit* che indica il suo livello di capacità. In Winsteps, tale valore corrisponde all'indice *MEASURE*. Dopo aver sistemato in ordine crescente i valori *MEASURE* dei candidati abbiamo rilevato una corrispondenza tra i logit dei singoli candidati e i risultati ottenuti individualmente al test, in termini di livello target. I candidati con i valori *MEASURE* più bassi vengono in effetti posizionati nei livelli più bassi e all'aumentare del valore *MEASURE* anche il livello verso il quale vengono orientati i candidati aumenta. Questo indica che vi è una corrispondenza tra la scala di valori dei candidati generata dal modello statistico e il punteggio aggregato degli stessi candidati che viene fornito da SELF.

3.1.1. Il *minitest*

L'applicazione della CTT ci ha permesso di verificare innanzitutto che venisse rispettato il criterio dell'unidimensionalità del test. Come menzionato poc'anzi, lo scopo di SELF è misurare una sola dimensione o, per usare il termine psicometrico, un unico *tratto latente*: la competenza linguistica in italiano L2.

[19] "Le logit, acronyme de *log-odds unit*, […] est l'unité probabiliste utilisée pour exprimer la capacité des personnes et la difficulté des items […]. Une unité logit est définie comme la différence entre la capacité d'une personne et la difficulté d'un item pour laquelle la personne a un rapport de vraisemblance de réussite égale à 2.71:1 (soit exp (1):1). L'échelle de mesure exprimée en logits est une échelle d'intervalles" (Penta *et al*. 2005: 37).

TiaPlus permette di verificare l'unidimensionalità attraverso l'analisi fattoriale delle singole tappe che compongono il test e ne fornisce una rappresentazione sia numerica che grafica.

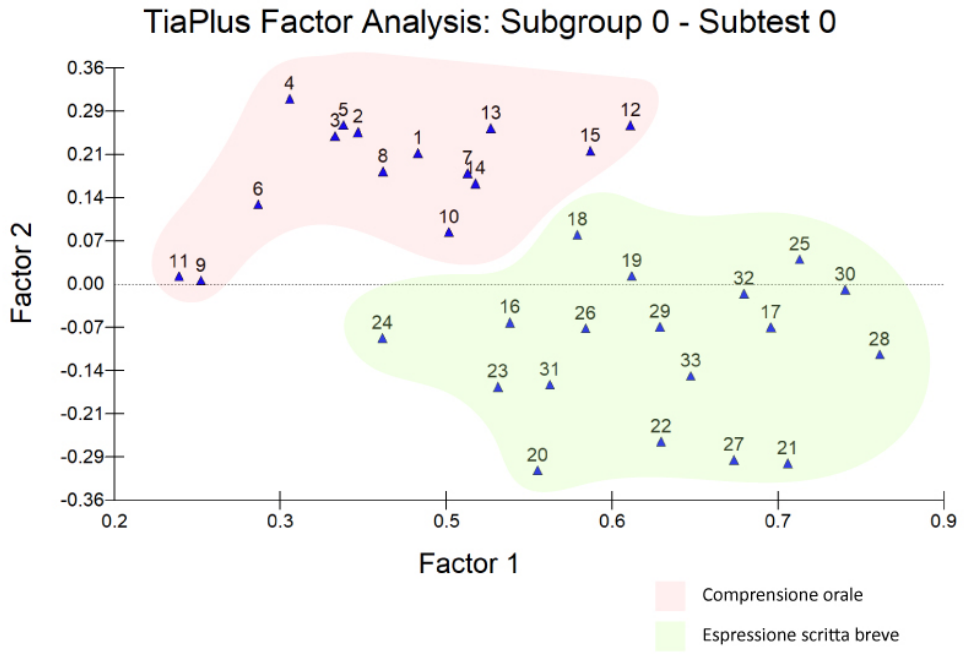

Fig.3 Analisi fattoriale degli item del minitest di SELF italiano

Come si può osservare dalla Figura 3, tutti gli item del *minitest* sono situati nella parte a destra degli assi cartesiani. Questo tende a confermare che il criterio dell'unidimensionalità viene rispettato: il test misura effettivamente un unico *tratto latente* che corrisponde al costrutto del test (cfr. infra 1).

È inoltre interessante notare la disposizione degli item all'interno del grafico: dopo aver proceduto all'identificazione degli item in funzione dell'abilità testata, è emerso, come già osservato in casi simili (van der Silk, Weideman 2005; Le, du Plessis, Weideman 2011; Weideman 2019), che pur rispettando il criterio dell'unidimensionalità vi può essere una scissione e un raggruppamento visibile tra le abilità testate, che non si sovrappongono e creano degli insiemi distinti. Nel nostro caso sono emersi due insiemi che abbiamo messo in evidenza con colori distinti (rosa *vs*. verde).

La configurazione grafica ottenuta tende ad indicare che la risposta dei candidati è determinata non solo dal grado di competenza linguistica ma anche dal grado di competenza nelle singole abilità, suggerendo l'esistenza appunto di due dimensioni sottostanti corrispondenti ai due tipi di abilità testate nel *minitest*. Osserviamo inoltre che il *Factor 1* rappresenta 1/3 della varianza. Procedendo ad un confronto con i valori relativi agli indici di difficoltà e di discriminazione di ciascun item, è emerso che il *Factor 1* potrebbe rappresentare una sintesi di questi due valori. L'analisi fattoriale in formato grafico posiziona gli item del test rispetto alla loro correlazione con il *Factor 1* e quindi ci permette di osservare quali item svolgono al meglio il loro lavoro, cioè confortano maggiormente il criterio di validità del test. È interessante notare che si tratta soprattutto di item di espressione scritta breve, per i quali la correlazione con il *Factor 1* risulta essere la più elevata. Per quanto riguarda la comprensione orale invece gli item che accreditano maggiormente la validità del test sono gli item 12, 13, 14 e 15, che passeremo in rassegna nelle pagine che seguono. Al contrario, gli item 11 e 9, molto vicini all'asse delle ordinate e situati quasi allo stesso livello, sono gli item che hanno ottenuto gli indici di discriminazione più bassi (Rir di 0,18 e 0,21).

Un'altra nozione importante in CTT è quella di coerenza interna, legata al concetto di affidabilità (Frisbie 1988) e che "valuta l'omogeneità dello strumento di misura attraverso la correlazione di ogni item con gli altri item e con il punteggio totale" (Penta *et al*. 2005: 122). Di solito, per misurare la coerenza interna di un test in statistica viene utilizzato il coefficiente alpha di Cronbach. Questo coefficiente può variare da 0 a 1 (un'alpha di Cronbach uguale a 1 indica una coerenza interna ideale) ed è empiricamente accettabile a partire da 0,70. Nel nostro caso, il *minitest* ha ottenuto un'alpha di Cronbach di 0,87 che corrisponde dunque ad una coerenza interna più che accettabile.

Passando all'*analisi degli item*, i valori rilevati sia tramite la CTT che tramite la IRT indicano che l'insieme degli item del *minitest* di comprensione orale hanno ottenuto buoni indici di discriminazione (Rir e Rit). Per quanto riguarda la CTT, il Rir varia infatti da 0,18 a 0,46 e il coefficiente di discriminazione in IRT (il PTMA) da 0,22 a 0,51. Questo significa che gli item di comprensione orale proposti nel

minitest sono complessivamente in grado di discriminare i candidati più abili da quelli meno abili.

Gli item che ottengono i valori più discriminanti sono in particolare quattro (12, 13, 14 e 15), due dei quali appartenenti allo stesso task. La prima constatazione da fare rispetto a questi item è che si tratta degli ultimi quattro della serie di comprensione orale del *minitest*. Il *minitest* è in effetti composto da 33 item (corrispondenti a 15 task), 15 di comprensione orale (9 task) seguiti da 18 item di espressione scritta breve (6 task). Gli studenti non sono a conoscenza della composizione del test e non sono dunque in grado di sapere se stanno svolgendo gli ultimi task di una determinata abilità. Che cosa hanno in comune questi item? E perché sono quelli che riescono a discriminare meglio degli altri item i candidati al test?

Nel primo caso si tratta di un'attività di risposta a scelta multipla di tipo *discourse completion task*, che prevede una chiave e due distrattori. Ai candidati viene infatti richiesto di selezionare, tra le opzioni proposte, quella più adatta per completare il dialogo del testo fonte.

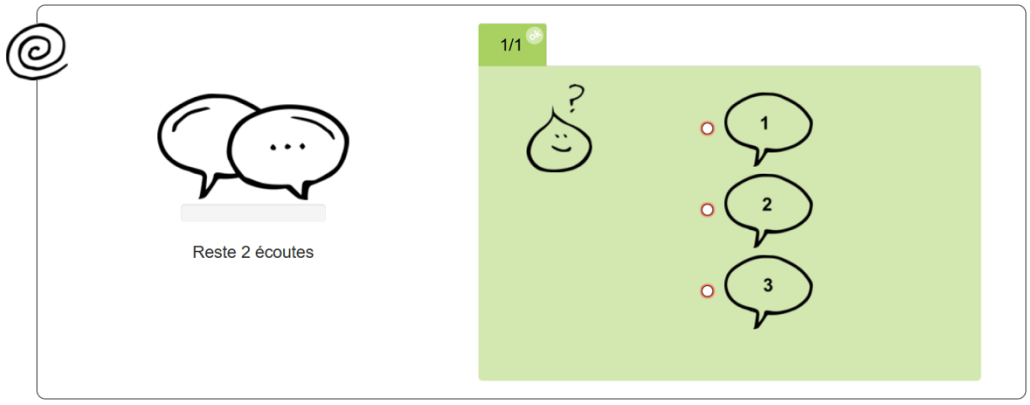

Fig. 4 Screenshot del task in questione (item 12).

Il task si compone, come tutti i task di SELF, di un contesto, di un testo fonte, di uno *stem* e delle opzioni di risposta. Eccezion fatta per il testo fonte, che può essere ascoltato massimo due volte, tutti gli elementi che compongono il task prevedono un numero di ascolti illimitato. In questo caso specifico, il testo fonte è un audio di 18 secondi e si tratta di una conversazione informale (contesto: *"conversazione tra amici"*) a due voci, una maschile e una femminile. L'item ha

un focus morfosintattico, in particolare riguarda la concordanza dei tempi verbali al passato; l'informazione necessaria per rispondere correttamente è esplicitamente menzionata nel testo fonte (informazione di tipo endoforico) ma non presenta ridondanza.

L'immagine sottostante illustra il comportamento dei distrattori e della risposta corretta dell'item in questione.

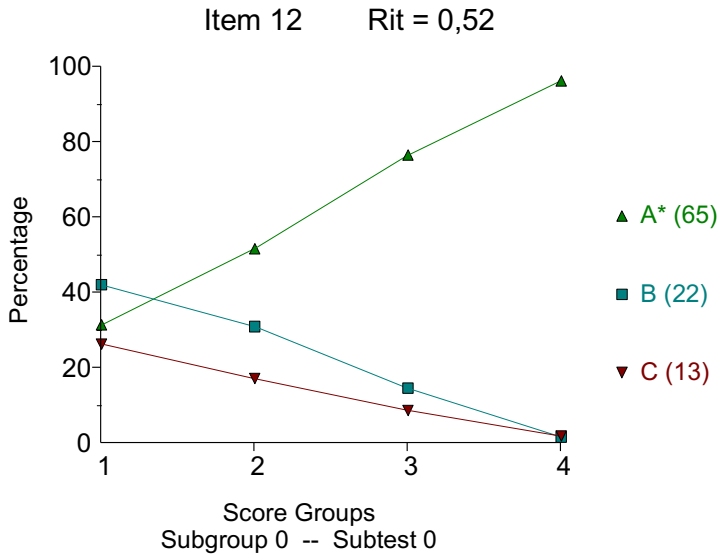

Fig. 5 Comportamento delle opzioni di risposta dell'item 12.

Sull'asse delle ascisse vengono riportati i quattro gruppi di candidati identificati da Tiaplus a partire dai risultati ottenuti al test sull'insieme del campione testato (N=1499) secondo la CTT. Il valore 1 rappresenta il gruppo dei candidati più "deboli" e il valore 4 il gruppo dei più "forti". Sull'asse delle ordinate viene invece indicata la percentuale di candidati che hanno selezionato una determinata opzione di risposta. Infine a destra del grafico, vengono esplicitate la chiave (A, contrassegnata da un asterisco) e i distrattori (B e C) e per ogni opzione viene specificata tra parentesi la percentuale di risposte ottenute. Come si può osservare dal grafico, la percentuale più alta di candidati ad aver selezionato la risposta corretta appartiene al gruppo 4, mentre per quanto riguarda i due distrattori le risposte arrivano maggiormente da candidati del gruppo 1.

Rispetto ai valori della IRT, Winsteps fornisce i valori di Infit e Outfit utili a "quantificare la conformità degli item e delle persone ai requisiti del modello"

(Penta *et al*. 2005: 94, nostra traduzione). Per entrambi i valori sono inoltre disponibili due diverse forme: il quadrato medio (mean square, MNSQ) e i quadrati medi convertiti in una statistica t (ZSTD) (Penta *et al*. 2005: 95)[20]. L'item in questione presenta dei buoni valori MNSQ vicini a 1 (Infit MNSQ = 0,93 e Outfit MNSQ = 0,83) ma dei valori ZSTD inferiori a -2, il che significa che i dati sono troppo prevedibili.

Nel secondo caso, i due item (13 e 14) appartengono allo stesso task; il testo fonte è un dialogo audio informale (contesto: "*dialogo tra amici"*) a due voci, una maschile e una femminile. Si tratta di una domanda con risposta a scelta multipla con una chiave e due distrattori. La durata è superiore rispetto all'altro item analizzato (1.01 minuti) e gli ascolti disponibili per il testo fonte sono due.

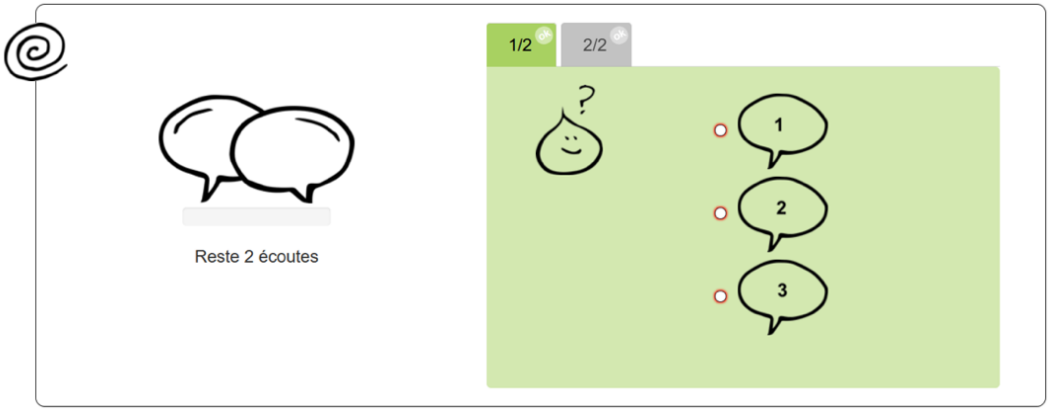

Fig. 6 screenshot del task in questione (item 13 e 14).

Il focus del primo item è lessicale e si tratta in particolare di una comprensione dettagliata, l'informazione è ridondante con riferimento endoforico. Il secondo item ha invece un focus morfosintattico e anche in questo caso lo sforzo di

[20] Quando gli item sono perfettamente conformi ai requisiti del modello i valori di Infit e Outfit sono, per quanto riguarda l'indice MNSQ uguali a 1 e vicini allo 0 per l'indice ZSTD (Penta *et al.,* ibidem). I valori di Infit e Outfit MNSQ sono sempre positivi; tuttavia, valori inferiori a 1 indicano che "la risposta della persona (item) è più deterministica di quanto preveda il modello" (Penta *et al*., 2005: 98, nostra traduzione) e valori superiori a 1 indicano che "la risposta della persona (item) è più aleatoria di quanto preveda il modello" (Penta *et al*., 2005: 99, nostra traduzione). Per quanto riguarda invece i valori di Infit e Outfit ZSTD, i dati si aggiustano perfettamente alle proprietà del modello quando sono uguali a 0, un valore positivo indica un underfit (la risposta all'item è meno prevedibile) e uno negativo un overfit (la risposta all'item è più prevedibile di quanto possa prevedere il modello) (Penta *et al*., ibidem).

comprensione si concentra su un elemento particolare contenuto nel testo fonte. Gli indici di discriminazione sono molto buoni in entrambi i casi:

	Rir (CTT)	PTMA (IRT)
Item 13		0,42
Item 14		0,41

Anche per quanto riguarda il comportamento di chiavi e distrattori, i due item presentano caratteristiche simili e coerenti.

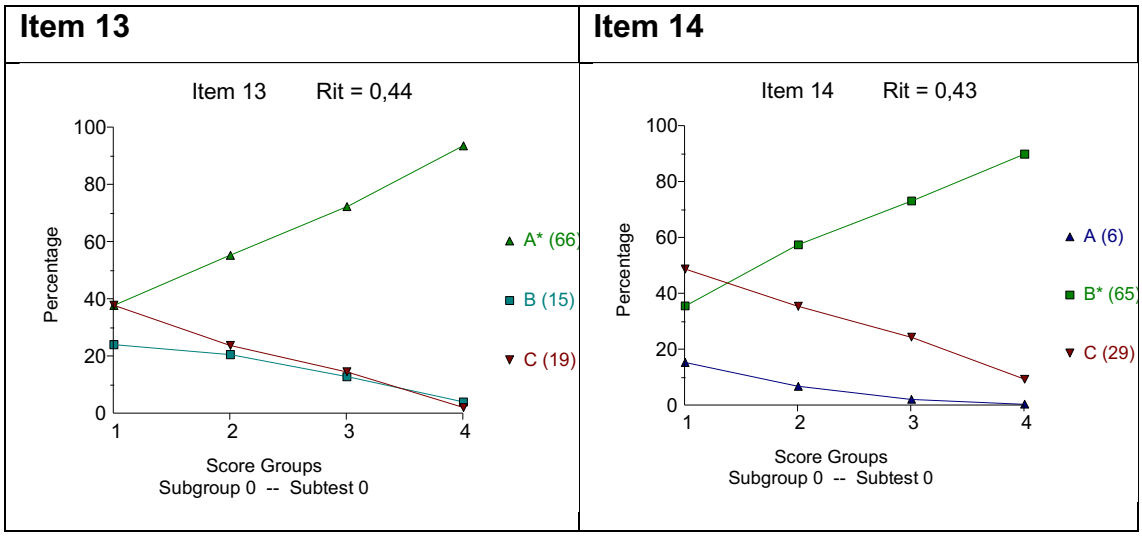

Fig. 7 Comportamento delle opzioni di risposta degli item in questione.

In entrambi i casi infatti la chiave (contrassegnata da un asterisco) viene selezionata prevalentemente dai candidati dei gruppi più "forti" mentre i candidati dei gruppi più "deboli" hanno tendenza a scegliere uno dei due distrattori.

Il quarto item più discriminante (item 15) è ancora una volta una domanda a risposta multipla con una chiave e due distrattori. Il testo fonte è però, in questo caso, un video di una pubblicità di Poste Italiane a proposito dei servizi postali online (contesto: *pubblicità creativa*). Il testo fonte ha una durata di 0.46 secondi ed è possibile visionare il video una sola volta. Questo item è, insieme al primo esaminato, il più discriminante del *minitest* e l'ultimo della serie di comprensione orale. Se ci si basa sull'indice di discriminazione della IRT, risulta essere addirittura il più discriminante in assoluto (PTMA = 0,51), dato molto interessante poiché l'item si differenzia dagli altri per il formato audiovisivo.

Inoltre, nella scala lineare generata dal modello di Rasch, questo item ha il valore più elevato (MEASURE = 0,79) rispetto agli altri tre. Ciò significa che, secondo i parametri della IRT, è l'item con l'indice di difficoltà più alto e dunque è, tra i quattro item di comprensione orale del *minitest* che ottengono i migliori valori psicometrici, quello più difficile.

Item	MEASURE
13	-0,36
12	-0,32
14	-0,31
15	0,79

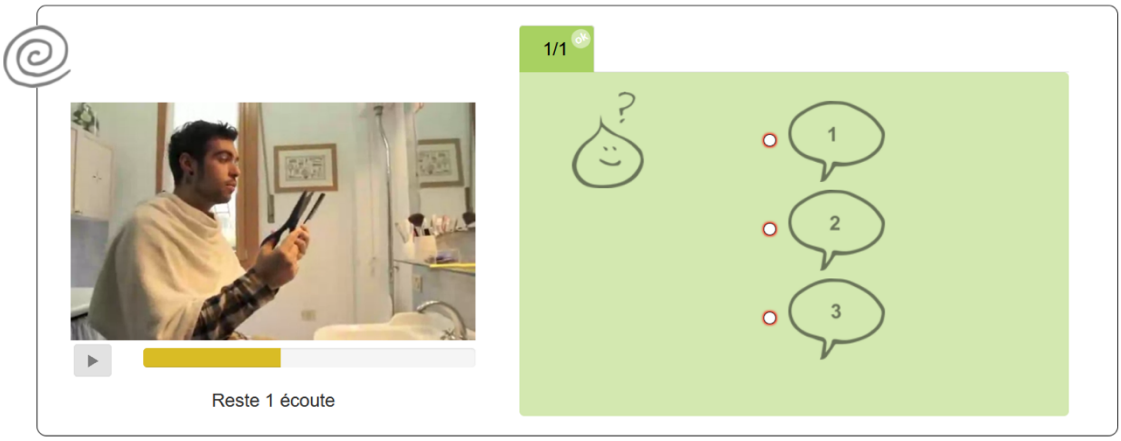

Fig. 8 screenshot del task in questione (item 15).

Il focus è lessicale e la comprensione è ancora una volta dettagliata, l'informazione non è ridondante e il riferimento è anche in questo caso endoforico. Anche i valori di Infit e Outfit sono molto buoni ma è interessante osservare il comportamento di chiave e distrattori:

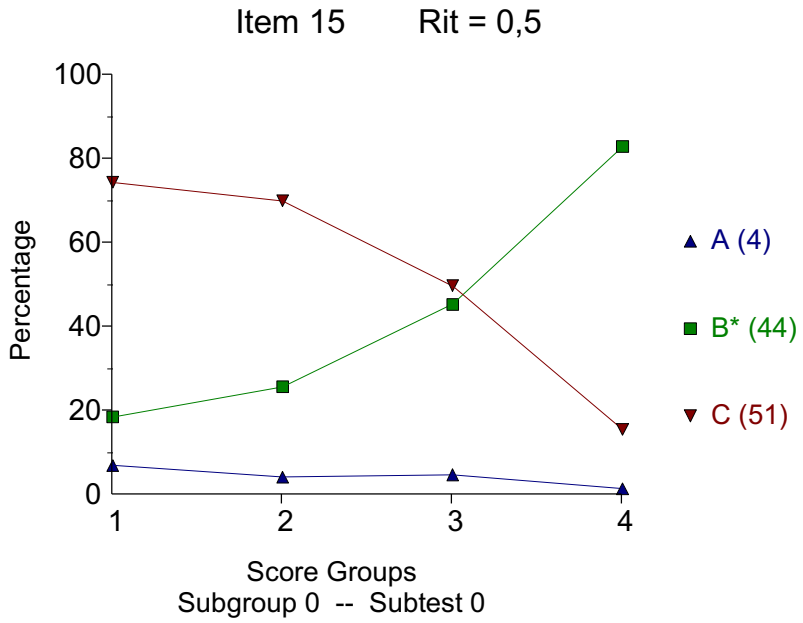

Fig.9 Comportamento delle opzioni di risposta dell'item 15.

Come si può vedere dal grafico, la chiave (B, contrassegnata da un asterisco) ha ottenuto il 44% di risposte, contro il 51% del distrattore C. In presenza di altri valori non accettabili questo comportamento anomalo potrebbe rappresentare un campanello d'allarme. Tuttavia in questo caso, il fatto che l'item presenti ottimi valori in termini di indici di discriminazione come di Infit e Outfit è rassicurante. Per di più, come illustrato nella Figura 8, anche se il distrattore C è stato selezionato anche da candidati del gruppo 4, la maggioranza di questi ha individuato la risposta corretta e la ragione dell'elevato tasso di selezione del distrattore C trova una spiegazione nelle alte percentuali di selezione dello stesso da parte dei candidati dei gruppi 1 e 2.

Per riassumere quindi, gli item più discriminanti di comprensione orale contenuti nel *minitest* hanno sì caratteristiche differenti, ma anche diversi punti in comune, come illustrato nella tabella sottostante.

	Item 12	Item 13	Item 14	Item 15
Tipologia di attività	Risposta a scelta multipla (una chiave e due distrattori)	Risposta a scelta multipla (una chiave e due distrattori)	Risposta a scelta multipla (una chiave e due distrattori)	Risposta a scelta multipla (una chiave e due distrattori)
Focus	Morfosintattico	Lessicale	Morfosintattico	Lessicale
Comprensione	Dettagliata	Dettagliata	Dettagliata	Dettagliata
Riferimento	Endoforico	Endoforico	Endoforico	Endoforico
Formato	Audio Registrato in studio	Audio Registrato in studio	Audio Registrato in studio	Audiovisivo
Dialogo/Dialogo interrotto/Monologo	Dialogo interrotto	Dialogo	Dialogo	Monologo
Registro	Informale	Informale	Informale	Informale
Voci maschili/Voci femminili	1 voce maschile e 1 voce femminile	1 voce maschile e 1 voce femminile	1 voce maschile e 1 voce femminile	2 voci maschili
Durata	18''	1.01''	1.01''	46''
Numero di ascolti del testo fonte	2	2	2	1

Tab.1 Riepilogo delle caratteristiche dei quattro item di comprensione orale del minitest con i valori psicometrici migliori.

3.1.2. Un confronto tra gli item della seconda e della terza tappa

La CTT e la IRT sono state adottate anche per l'analisi degli item contenuti nella seconda e nella terza tappa del test (cfr. infra 3.1). A differenza del *minitest* queste due tappe contengono item che testano la competenza linguistica in italiano su tre abilità: comprensione orale, comprensione scritta ed espressione scritta breve. Inoltre, le due tappe hanno in comune sei item di comprensione orale.

L'analisi fattoriale delle due tappe ci consente di effettuare immediatamente alcune considerazioni importanti.

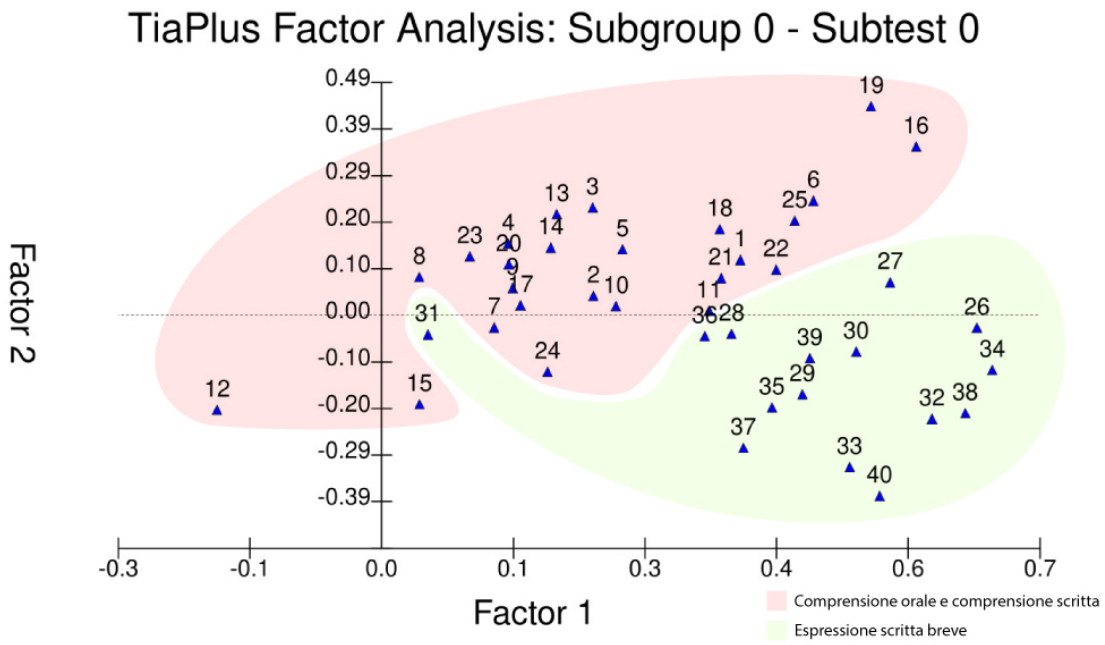

Grafico 7 Analisi fattoriale degli item della seconda tappa del test.

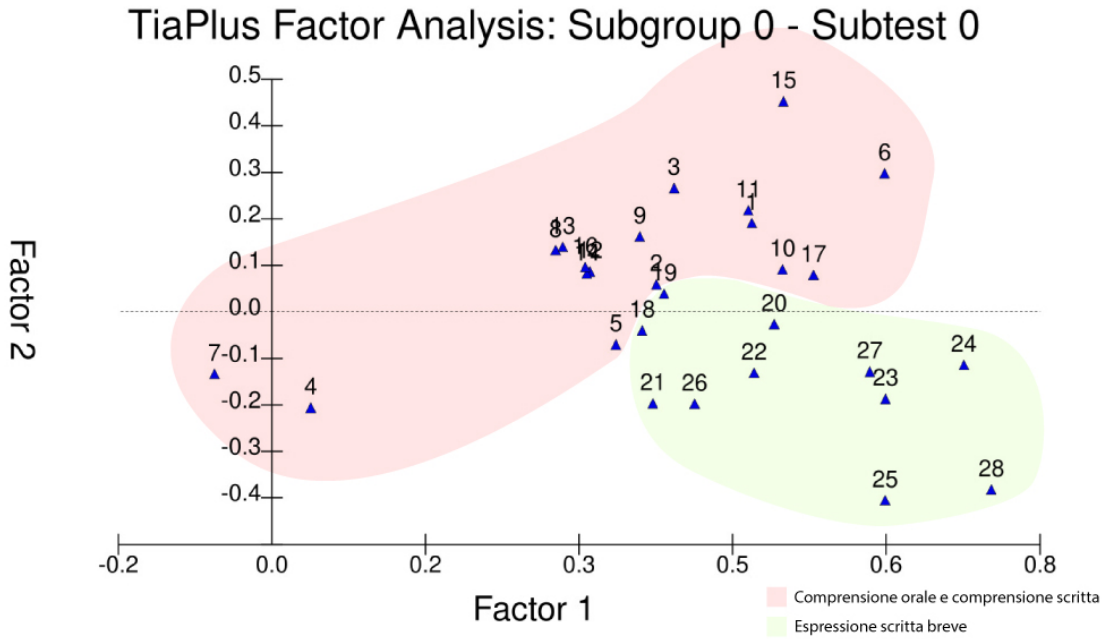

Grafico 8 Analisi fattoriale degli item della terza tappa del test.

Innanzitutto notiamo che, a differenza del *minitest*, in entrambe queste tappe sono presenti item che sono situati nella parte sinistra del grafico (Factor 2). Questo significa che gli item in questione stanno testando qualcosa di diverso rispetto a tutti gli altri item (Factor 1); potrebbe trattarsi di un fattore che non rientra nel costrutto del test. Ci riferiamo all'item 12 della seconda tappa e all'item 7 della terza tappa. Inoltre, nel grafico della seconda tappa (Grafico 7) notiamo

che ci sono diversi item situati molto vicino all'asse delle ordinate e quindi al secondo fattore (Factor 2), si tratta in particolare degli item 8, 15 e 31. Quanto alla terza tappa, l'unico item che si avvicina molto al secondo fattore è il 4 (cfr. Grafico 8). L'insieme di questi item richiede un'analisi più approfondita da svolgersi non solo a partire dai valori psicometrici ma anche in riferimento ai contenuti dei task in questione.

I grafici delle analisi fattoriali di queste due tappe del test ci permettono di effettuare un'ulteriore considerazione: anche in questi due casi la disposizione degli item all'interno del grafico si differenzia a seconda delle abilità. Gli item che testano la comprensione (orale e scritta) sono situati, nella maggior parte dei casi nella parte in alto a sinistra del grafico (Factor 1) e presentano valori positivi, mentre quelli che testano la produzione (espressione scritta breve) si trovano piuttosto in basso a destra e presentano valori negativi. Questo rispecchia quanto già affermato in precedenza per quanto riguarda gli item del *minitest* (cfr. infra 3.1.1).

Gli item che sono situati a sinistra dell'asse delle ordinate (item 12 della seconda tappa e item 7 della terza tappa) fanno parte di uno stesso task, comune alle due tappe.

Ordine degli item nel task	**Ordine degli item nella tappa 2**	**Ordine degli item nella tappa 3**
1	12	6
2	13	7
3	14	8

Tab. 2 Numerazione degli item 12 e 7 nelle tappe 2 e 3.

Si tratta in particolare di item di tipo 'Vero/Falso/Non è possibile dirlo', una tipologia di esercizio che richiede uno sforzo differente rispetto alle altre tipologie di task e nella quale l'abilità inferenziale viene esplicitamente sollecitata. Il testo fonte è un audio di 1.15 minuti e si tratta di un'intervista radiofonica registrata in studio a due voci, una femminile e una maschile. Il task si compone di tre item di comprensione dettagliata con focus lessicale e riferimento endoforico.

È interessante notare innanzitutto che gli item ottengono risultati differenti a seconda della tappa in cui si trovano. Come si osserva dalla Tabella 2 il primo item del task in questione non ha ottenuto risultati soddisfacenti nella seconda tappa (item 12) ma non ha riscontrato invece particolari problemi nella terza tappa (item 6); viceversa, il secondo item del medesimo task ottiene buoni risultati nella tappa 2 (item 13) ma non nella tappa 3 (item 7). Queste due tappe del test non vengono effettuate dagli stessi candidati e chi effettua gli item della seconda tappa ha un livello globalmente inferiore rispetto ai candidati che vengono sottoposti agli item contenuti nella terza tappa. Potremmo dunque ipotizzare che la differenza dei valori ottenuti dal task in questione potrebbe essere dovuta in parte alla differenza del campione, tuttavia è necessario procedere ad un'analisi più approfondita di valori psicometrici e dei contenuti. L'item 12 ottiene, sia in CTT che in IRT, valori psicometrici insoddisfacenti: l'indice di difficoltà è molto basso (P = 12), questo indica che soltanto una minima percentuale di candidati ha selezionato la risposta corretta; l'indice di discriminazione è negativo e dunque l'item non è in grado di discriminare candidati più "forti" da candidati più "deboli". Quali potrebbero essere le ragioni alla base di un tale malfunzionamento? Una prima ipotesi riguardo la tipologia di esercizio non pare fondata, in quanto è una tipologia piuttosto comune in altri test fondati su misure psicometriche.

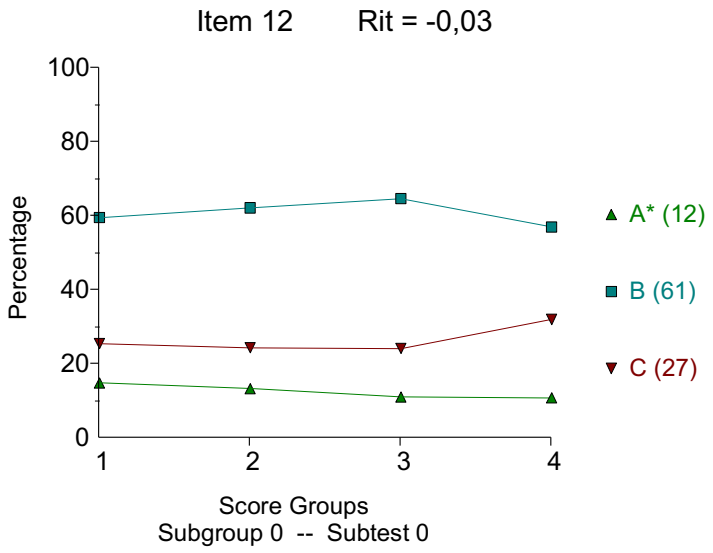

Fig. 10 Comportamento delle opzioni di risposta dell'item 12.

L’item illustrato nella Figura 11 presenta un funzionamento incoerente e come possiamo osservare dal grafico qui sopra la chiave (A, che corrisponde all’opzione “Vero”) è stata selezionata in maggior misura da candidati dei gruppi 1 e 2 (più “deboli”) rispetto a quelli dei gruppi 3 e 4. Per di più la maggioranza dei candidati (il 61%), compresi quelli dei gruppi più “forti”, ha erroneamente identificato la risposta corretta nel distrattore B (cioè l’opzione “Falso”) e non nella chiave A. È altresì interessante notare che il distrattore C (opzione “Non è possibile dirlo”) viene selezionata nella maggioranza dei casi da candidati del gruppo 4, vale a dire i più “forti” secondo il calcolo elaborato dal modello statistico. L’ipotesi più plausibile che spiegherebbe il funzionamento anomalo dell’item e il suo posizionamento nella parte a sinistra degli assi cartesiani è quella relativa ad un difetto di ordine redazionale e quindi contenutistico. Da un’analisi sul contenuto si evince che l’item sta testando effettivamente qualcosa di diverso rispetto agli altri item del test, in quanto la selezione della risposta corretta è determinata dalla conoscenza pregressa di una locuzione di bassa frequenza che viene espressa con un tono aulico e al contempo scherzoso (“chiedere venia”). Un candidato X, pur avendo un livello di competenza linguistica elevato, potrebbe non conoscere tale espressione. In altre parole la conoscenza di una determinata locuzione di bassa frequenza non implica un livello di competenza linguistica elevato dei candidati. È tuttavia doveroso rimarcare che lo stesso item nella tappa 3, nella quale troviamo candidati di livello superiore rispetto alla tappa 2, ottiene buoni risultati. L’item ha un indice di difficoltà accettabile (P = 29) e un indice di discriminazione molto elevato (RIR = 41): si tratta cioè di un item difficile ma in grado di discriminare candidati “forti” e “deboli”.

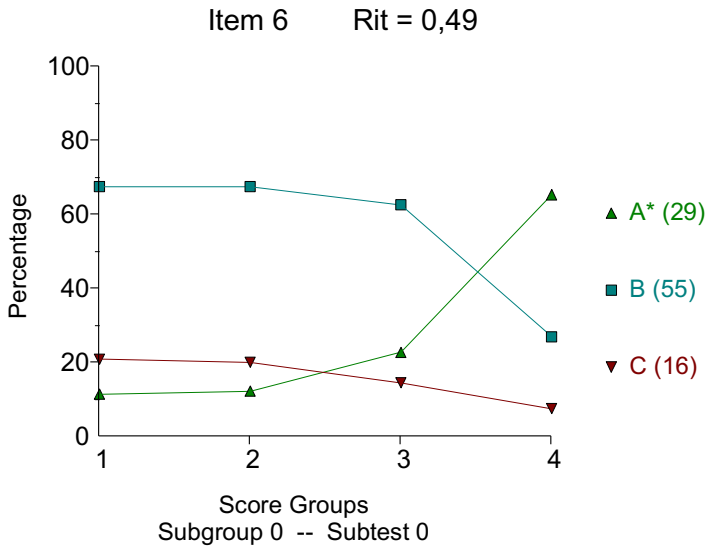

Fig. 11 Comportamento delle opzioni di risposta dell'item 6.

Anche per quanto riguarda il secondo item del task, item 7 che dal grafico dell'analisi fattoriale della tappa 3 risulta essere situato nella parte più vicina all'asse delle ordinate, il focus è lessicale ma in questo caso l'item non testa una conoscenza pregressa. I valori psicometrici indicano infatti che il maggiore problema di questo item è che non riesce a discriminare i candidati (indice di discriminazione negativo, RIR = -3) e risulta addirittura essere troppo semplice (indice di difficoltà P = 80).

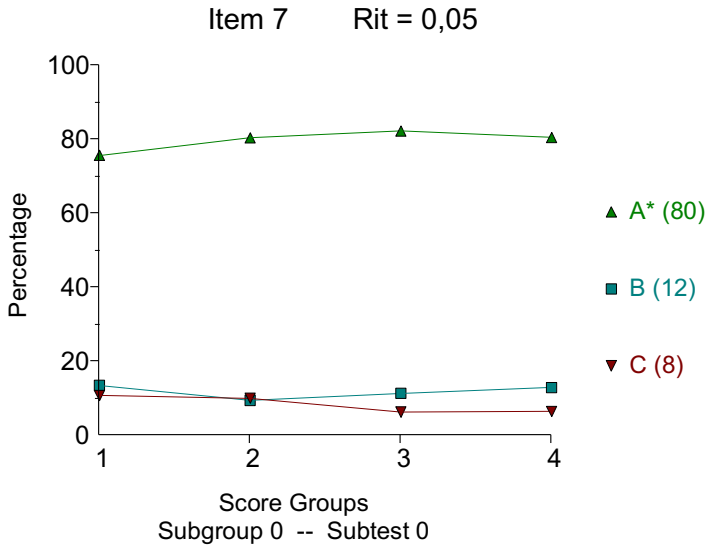

Fig.12 Comportamento delle opzioni di risposta dell'item 7.

L’analisi fattoriale ci mostra che l’item sta testando qualcosa di diverso dal costrutto di SELF, i valori psicometrici ci confermano una situazione critica. È alla luce di tali premesse che l’analisi del contenuto ci permette di avanzare un’ipotesi più concreta. L’item non è risultato affatto discriminante perché il compito che viene richiesto al candidato non è tanto quello di mobilitare risorse linguistico-cognitive per comprendere un dato contenuto espresso oralmente, ma piuttosto quello di associare un termine presente nel testo fonte ad un suo sinonimo esposto nello stem dell’item.

4. Per concludere

Il presente contributo intende illustrare i risultati di uno studio di post-validazione del test SELF italiano, condotto per la prima volta su un ampio campione di candidati (N=1499).

Come abbiamo potuto rilevare, in particolar modo grazie al contributo della psicometria, il test rispetta effettivamente il criterio di validità e la quasi totalità degli item contribuisce in modo efficace a misurare la competenza linguistica in italiano L2 così come è stata definita dal costrutto. L’analisi fattoriale ci ha permesso inoltre di fare alcune considerazioni interessanti circa il contributo che ciascun item apporta al corretto funzionamento del test. È stato quindi possibile osservare che gli item che contribuiscono maggiormente a discriminare i candidati al test sono gli item di espressione scritta breve.

Per quanto riguarda gli item di comprensione orale, le analisi confermano globalmente il loro corretto funzionamento. Si è osservato che gli item con i migliori valori psicometrici condividono tratti comuni che possono essere replicati nel processo di creazione di nuovi contenuti.

Procedendo peraltro a un’analisi più dettagliata degli item di comprensione orale ci siamo posti le seguenti domande: gli item con valori psicometrici ottimali rispecchiano il costrutto definito a priori e dunque testano effettivamente ciò che ci eravamo preposti di testare? Perché alcuni item non ottengono buoni valori

psicometrici? Che cosa stanno testando se non la competenza linguistica in italiano L2? Queste interrogazioni ci permettono di formulare alcune considerazioni in merito alla qualità redazionale, al formato e al focus degli item. Ci offrono inoltre alcuni spunti di riflessione riguardo agli orientamenti legati alla dimensione formativa e diagnostica di SELF dei task di comprensione orale.

Infine ci proponiamo di mettere in relazione i risultati delle analisi psicometriche con alcune caratteristiche individuali dei candidati al test. In particolare, sarebbe opportuno tentare di mettere in relazione il risultato ottenuto dai candidati con le caratteristiche dei singoli item per avviarci verso la progettazione di un SELF a orientamento diagnostico. Allo stesso modo, occorrerebbe procedere ad un'indagine più accurata delle ragioni per le quali alcuni item non ottengono risultati soddisfacenti mentre altri, in proporzione esigua e malgrado il superamento della validazione psicometrica preliminare in fase di pilotaggio e pretest, sembrano valutare i candidati su capacità che non rientrano nel costrutto del test.

Bibliografia

Aliotta, A. (1905) *La misura in psicologia sperimentale,* Firenze: Tipografia Galletti e Cocci.

ALTE (2011) *Manuel pour l'élaboration et la passation de tests et examens de langue –A utiliser en liaison avec le CECR*, Strasbourg: Division de Politiques linguistiques DG II – Service de l'éducation Conseil de l'Europe.

Bachman, L.F. e A.S Palmer (1996) *Language Testing in Practice,* Oxford: Oxford University Press.

Barni, M. (2005) "Etica e valutazione delle competenze il L2", in M. Vedovelli (ed.) *Manuale della certificazione dell'italiano L2*, Roma: Carocci Editore.

Cervini, C. e M-P. Jouannaud (2015) "Ouvertures et tensions liées à la conception d'un système d'évaluation en langues, numérique, multilingue et en

ligne, dans une perspective communicative et actionnelle", *AIsic* 18(2), doi: https://doi.org/10.4000/alsic.2821

------ (2016) "Approcci integrati nel testing linguistico: esperienze di progettazione e validazione in prospettiva interlinguistica", Bologna: *Quaderni del CeSLiC, Atti di Convegni CeSLiC – 4*.

------ e Y. Martari (2020) "Soggettività, vaghezza e ambiguità nel language assessment. Note sui confini dell'attribuzione", *Quaderni di semantica*, 2020, special issue: 101–130.

------ e M. Masperi (2021) "Designing a Multilingual Large-Scale Placement Test with a Formative Perspective: A Case Study at the University of Grenoble Alpes", in B. Lanteigne, C. Coombe, and J. D. Brown (eds.), *Challenges in Language Testing Around the world. Insights for language test users*, Singapore: Springer, 243–253.

Cito (1998-2013©) TiaPlus® Users Manual, M. & R. Department, Cito, Arnhem, NL.

Direction de l'évaluation, de la prospective et de la performance (2020) *Repères et références statistiques. Enseignement, Formation, Recherche*, https://www.education.gouv.fr/reperes-et-references-statistiques-2020-1316

Ferrari S., E. Nuzzo e G. Zanoni (2016) "Sviluppare le competenze pragmatiche in L2 in rete: problemi teorici e soluzioni pratiche nella progettazione dell'ambiente multimediale LIRA" in C. Cervini (ed.), *Interdisciplinarità e apprendimento linguistico nei nuovi contesti formativi. L'apprendente di lingue tra tradizione e innovazione*, Bologna: Quaderni del CESLiC, http://amsacta.unibo.it/5069/

Frisbie, D.A. (1988) "Reliability of scores from teacher-made tests", *Educational Measurement: Issues and Practical, National Council on Measurement in Education* 7(1): 25-35.

Guttmann, L. (1944) “A basis for scaling qualitative data”, *American Sociological Review* 9: 139-150.

Guttmann, L. (1950) “The basis for scalogram analysis” in S.A. Stouffer, L. Guttmann, E.A., Suchman, P.F., Lazarsfield, S.A. Star and J.A. Clausen (eds.), *Measurement and prediction,* Princeton: Princeton University Press.

Higashi, T. e C. Shirota (2019) “Kôdô chyûshin apurôchi ni motozuita Yôroppa ni okeru nihongo onrain tesuto no kaihatsu (Développement du test de japonais en ligne en Europe basé sur l'approche actionnelle)” in Tosaku and Lee (eds.), *ICT x Nihongo kyôiku* (*ICT x Japanese Language Education: Theory and Practice*), Tokyo: Hitsuji Shobô, 150-165.

Hossein, B. (2012) “The Relationship between Listening and Other Language Skills in International English Language Testing System”, *Theory and Practice in Language Studies* 2 (4): 657-663.

Le, P.L., C. du Plessis e A. Weideman (2011) “Test and context: The Use of the Test of Academic Literacy Levels (TALL) at a tertiary institution in VIETNAM”, *Journal for Language Teaching* 42(2), doi: http://dx.doi.org/10.4314/jlt.v45i2.7

Laveault, D. e J. Grégoire (2014) *Introduction aux théories des tests en psychologie et en sciences de l’éducation* (3ème éd.), Louvain-la-Neuve: De Boeck Supérieur.

Linacre, J. M. (2012) *Many-Facet Rasch Measurement: Facets Tutorial*, http://winsteps.com/tutorials.htm

Masperi, M. (2011) *Innovalangues: Innovation et transformation des pratiques de l’enseignement-apprentissage des langues dans l’enseignement supérieur*. Programme IDEFI (ANR-11-IDFI-0024), https://hal.archives-ouvertes.fr/hal-02004250/document

McNamara, T. (2000) *Language Testing*, Oxford: Oxford University Press.

Penta, M., C. Arnoul e C. Decruynaere (2005) *Développer et interpréter une échelle de mesure,* Bruxelles: Mardaga.

Purpura, J. (2004) *Assessing Grammar*, Cambridge Language Assessment Series, Cambridge: CUP.

Quintin, J-J. (2011) “Comparatif des résultats CLES 1 et 2 2010-2011. Journée Pédagogique LANSAD, Université Stendhal Grenoble 3”, 1 juillet 2011, Communication, https://videos.univ-grenoble-alpes.fr/video/3269-comparatif-des-resultats-cles-1-2-20102011/

Rasch, G (1960) *Probabilistic models for some intelligence and attainment tests*, Chicago: MESA Press.

Spearman, C. (1907) “Demonstration of formulae for true measurement of correlation”, *The American Journal of Psychology* 18(2): 161-169, doi: https://doi.org/10.2307/1412408.

Van der Silk, F. e A. Weideman (2005) “The refinement of a test of Academic Literacy”, *Per Linguam* 21(1): 23-35.

Weideman, A. (2019) *Validation and the further disclosures of language test design*, doi: https://doi.org/10.19108/KOERS.84.1.2452