PAOLO TORRESAN

# CERTIFICAZIONE PLIDA. AFFIDABILITÀ DEI VALUTATORI DELLA PROVA SCRITTA

Pedro & João editores

# CERTIFICAZIONE PLIDA

## AFFIDABILITÀ DEI VALUTATORI DELLA PROVA SCRITTA

Pedro & João
editores

Un ringraziamento, per il confronto, a Mike Linacre, Everett Smith, Arturo Mendoza Ramos e a Thomas Eckes; ai valutatori e al gruppo PLIDA per i dati trasmessi; a Ana Luísa de Araujo Mhereb e a Elisabetta Santoro per l'introduzione.

Un ringraziamento va anche ai colleghi lo scambio con i quali è prezioso (Clelia, Alberta, Carlo, Franco, Elisabetta, Giovanna, Manuela, Anna e molti altri).

Questa pubblicazione è stata possibile grazie all'appoggio dell'Università Federale Fluminense (Niterói, Rio de Janeiro)

# CERTIFICAZIONE PLIDA

## AFFIDABILITÀ DEI VALUTATORI DELLA PROVA SCRITTA

*Paolo Torresan*

Pedro & João
editores

Paolo Torresan

**Certificazione PLIDA. Affidabilità dei valutatori della prova scritta.** São Carlos: Pedro & João Editores, 2024. 173p. 16 x 23 cm.

**ISBN: 978-65-265-122-8 [Digital]**

1. Avaliação linguística. 2. Ensino de língua estrangeira. 3. Linguística aplicada. I. Título.

CDD – 410

**Capa:** Andersen Bianchi
**Ficha Catalográfica:** Hélio Márcio Pajeú – CRB - 8-8828
**Editores:** Pedro Amaro de Moura Brito & João Rodrigo de Moura Brito



**Pedro & João Editores**
www.pedroejoaoeditores.com.br
13568-878 – São Carlos – SP
2024

A mia madre,
che il prete ha voluto si chiamasse Rita
e che invece si chiama Margherita, che è il nome di un fiore.

A mia madre,
che ha vissuto sotto il fascismo
nella campagna trevigiana;
i partigiani di Mignagola volevano caricarla sul camion,
una suora l'ha salvata.

A mia madre.

## INDICE

## INDICE DELLE TABELLE

INTRODUZIONE



TESTO PRINCIPALE

## INDICE DELLE FIGURE

## PREFAZIONE DELL’AUTORE

In questo libro presentiamo lo studio dei giudizi espressi da sette valutatori in merito agli elaborati redatti da tre gruppi di candidati della *Certificazione PLIDA* di livello B1.

Dopo una introduzione su quanto sia difficile esaminare una *performance*, e più in particolare una *performance* linguistica (**sezione A**), illustriamo le caratteristiche generali di un’analisi statistica a carattere probabilistico, chiamata Rasch (**sezione B**). Applicata a più variabili (*Many-Facet Rash Measurement*), essa consente di verificare la conformità del comportamento del valutatore alle misure attese dal modello. Si tratta di una statistica prescrittiva e che pur prefigura, come ideale, non già un giudizio scevro da oscillazioni (perfettamente coerente e allineato), quanto un giudizio di media conformità (o di inaffidabilità contenuta), come meglio verrà descritto nel corso del libro.

Segue l’analisi vera e propria, nella quale triangoliamo diverse evidenze per stilare il profilo dei valutatori come gruppo e come singoli (**sezione C**).

Chiudono il volume le conclusioni, i riferimenti bibliografici e alcune appendici di approfondimento (**sezione D**).

Il volume può tornare utile agli enti preposti alla certificazione delle competenze linguistiche, al ricercatore che si occupi di valutazione linguistica e più in generale agli insegnanti chiamati a valutare la produzione scritta dei propri allievi.

Treviso,
25 febbraio 2024

Paolo Torresan

# INTRODUZIONE

# VALUTAZIONE E CERTIFICAZIONE DI UNA L2: SFIDE E PROSPETTIVE

di Ana Luísa de Araujo Mhereb & Elisabetta Santoro

È noto che valutare le conoscenze e le competenze in una seconda lingua (L2) non è semplice. Da una parte, la valutazione linguistica è un compito delicato e di grande responsabilità per le conseguenze e l'impatto che ne può derivare sia a livello individuale che sociale. Dall'altra, si tratta di un processo estremamente complesso perché è necessario tener conto di innumerevoli aspetti tra cui i modelli teorici di lingua, di competenza e di misurazione. Risiede qui una delle principali sfide: mentre questi modelli sono caratterizzati da indeterminatezza e variazione, la valutazione della competenza in L2 cerca proprio il contrario, ovvero, esattezza e precisione e ciò crea inevitabilmente una contraddizione intrinseca (Barni, 2005).

Va detto, tuttavia, che la natura indeterminata dei fattori con cui si devono fare i conti non impedisce le operazioni di valutazione (che in molti contesti sono, tra l'altro, indispensabili). Di fatto, sebbene non sia possibile misurare la competenza in L2 in modo assolutamente esatto, si possono sviluppare forme di misurazione che tendono ad essere sempre più precise perché bilanciano l'indeterminatezza con l'esplicitazione. È imprescindibile, per esempio, che valutati e valutatori sappiano esattamente qual è la concezione di lingua che soggiace al tipo di esame o di test che si propone e lo è altrettanto che si chiarisca quali riferimenti teorici relativi alla competenza linguistica e alla misurazione guidano tanto l'elaborazione del test quanto l'attribuzione dei punteggi.

La valutazione di una L2 può avvenire in diverse situazioni e, a seconda dell'obiettivo, si distinguono diverse tipologie. Esistono i test di profitto o *achievement tests* che mirano a misurare le conoscenze acquisite dagli studenti a partire da un determinato insieme di contenuti o abilità specifiche, di solito dopo un periodo di istruzione, come nel caso della valutazione finale di un corso di lingua. Un'altra tipologia è quella dei test di livello o *placement tests* che vengono utilizzati quando un apprendente sta per iniziare un corso di lingua, al fine di determinare la sua competenza linguistica e di collocarlo al livello più appropriato. Esistono poi i test di competenza o *proficiency tests* che valutano il livello generale di competenza di un individuo in una L2, indipendentemente dal suo percorso e dal fatto che l'apprendimento sia avvenuto in un contesto formale o meno.

Dell'ultimo tipo fanno parte le certificazioni di competenza in L2, considerate "esami di alta rilevanza" (*high-stakes exams*), cioè esami il cui risultato ha effetti di grande peso nella vita dei candidati. Tra le sue conseguenze ci sono infatti quella di permettere o impedire l'accesso all'università, l'ottenimento di un lavoro o l'acquisizione della cittadinanza del paese in cui si parla la lingua sottoposta ad esame. Quindi, se per qualcuno, che è di solito chi ha avuto la possibilità di studiare la lingua, di prepararsi per l'esame e di pagare per la prova, può aprire porte; per altri, che per vari motivi non hanno avuto risorse economiche, tempo sufficiente o l'opportunità di imparare la lingua e avere un buon risultato all'esame, può chiuderle (Bachman & Purpura, 2008; McNamara & Shohamy, 2008).

Oltre alle conseguenze sociali, la valutazione certificatoria è caratterizzata dal cosiddetto effetto retroattivo (*washback effect*) che spesso ha sull'insegnamento e sull'apprendimento di una L2, influenzando l'elaborazione di materiali didattici, i contenuti dei corsi di lingua e persino la pratica degli insegnanti in classe (Barni, 2000). Data l'importanza sociale e glottodidattica di questo tipo di valutazione, è cruciale che tutte le decisioni relative al suo processo, dalla costruzione dei test al giudizio da parte dei valutatori, siano prese con molta cautela.

Durante l'elaborazione di un esame di certificazione in L2, sono due i criteri fondamentali che occorre considerare: la validità e l'affidabilità. Anche se complementari, questi aspetti sembrano spesso essere in tensione (Barni, 2005). La validità mira a garantire che il test misuri in modo preciso e corretto ciò che si propone di misurare, come nel caso di abilità linguistiche specifiche. Se, per esempio, si elabora un test che abbia come scopo valutare l'abilità di lettura, non devono essere inclusi item che abbiano finalità diverse. D'altra parte, l'affidabilità ha l'obiettivo di garantire risultati coerenti, anche quando il test viene somministrato in momenti diversi o quando viene corretto da altri valutatori oppure dallo stesso valutatore a distanza di tempo. Nonostante esista un dibattito su quale criterio debba avere la priorità, è consenso che entrambi siano fondamentali e debbano essere considerati in modo complementare. Pertanto, i costruttori di test linguistici devono affrontare un'altra sfida: creare test che siano il più possibile validi e affidabili, ovvero che non solo consentano sia al candidato che al valutatore di identificare con chiarezza che cosa si valuta, ma che, allo stesso modo, diano indicazioni precise su come si attribuiranno i punteggi che porteranno al voto finale. Su questo torneremo tra poco.

Un altro aspetto essenziale è comprendere le principali caratteristiche dei test linguistici che differiscono soprattutto in base all'oggettività e alla soggettività delle risposte e, di conseguenza, della correzione (Bachman, 1990). I test oggettivi, tra cui per esempio i test a scelta multipla o quelli che prevedono il riempimento di una lacuna, hanno risposte fisse, predefinite dai criteri stabiliti al momento dell'elaborazione. L'oggettività permette anche la correzione meccanica, spesso adottata quando il numero di test da correggere è molto elevato. I test soggettivi, che comprendono soprattutto le produzioni orali e scritte più libere, presuppongono, invece, diversi esiti possibili, sebbene alcuni siano spesso più accettabili di altri. Questi test richiedono dunque il giudizio di un valutatore, cosa che rende il processo di valutazione più lungo, oltre che più delicato e complesso, come vedremo meglio di seguito.

I test linguistici vengono a volte caratterizzati anche in base al metodo. McNamara (2000) prende in esame due categorie che chiama *paper-and-pencil tests* e test di performance. Considerati tradizionali, i *paper-and-pencil tests* corrispondono ai test di tipo oggettivo, poiché sono spesso strutturati nel formato di risposta chiusa, in cui il candidato deve scegliere l'opzione corretta tra le possibilità presentate. Al contrario, i test di performance valutano le abilità dell'apprendente attraverso simulazioni di compiti quotidiani in contesti più realistici e sono, dunque, caratterizzati dalla soggettività sia delle risposte che della correzione. Sono test basati su un campione in genere piuttosto esteso di scritto o parlato che viene valutato da uno o più valutatori o *raters*, utilizzando una procedura di valutazione concordata in precedenza.

A seconda degli obiettivi della valutazione, un tipo di test può essere più adatto di un altro. Tuttavia, ai test di performance è stato attribuito, soprattutto in periodi più recenti, un ruolo di particolare rilievo. Si tratta, di fatto, di test particolarmente apprezzati perché, come si è accennato prima, considerano l'uso della lingua in situazioni reali, riflettendo una tendenza crescente anche nell'insegnamento delle L2. La risposta alla sfida di limitare l'inevitabile soggettività legata alla valutazione di questo tipo di test è stata metterlo al centro delle attenzioni e trasformarlo in priorità nella ricerca e nello sviluppo dei test linguistici (si vedano, tra gli altri, Hauptman, LeBlanc & Wesche, 1985; Aschbacher, 1992; Shohamy, 1995; Kane & Mitchell, 1996; McNamara, 1995, 1996 e 1997; Norris *et al*., 1998; Skehan, 2001).

Quando si deve esprimere un giudizio su una performance in L2, è innanzitutto necessario interrogarsi sul tipo di valutazione che verrà utilizzata. Quelli più comunemente utilizzati sono essenzialmente due: la valutazione detta olistica e quella che viene denominata analitica.

Nella valutazione olistica, il *rater* deve formare un'opinione generale e assegnare alla performance un solo punteggio complessivo. Una delle principali critiche a questo tipo di valutazione è che "*raters may be over influenced by a superficial impression formed in the first couple of minutes based upon criteria they are not consciously aware of*" (North, 2003, p. 70). In questo senso, ogni valutatore può dare più importanza a determinati criteri che, non essendo specificati, possono portare a incoerenze tra valutatori diversi.

Nella valutazione analitica, al contrario, si definisce preliminarmente quali aspetti della performance devono essere valutati e a ciascuno di essi viene assegnato un punteggio separato. Uno dei vantaggi di questo tipo di valutazione rispetto a quella olistica è che l'utilizzo di una scala analitica consente a tutti i *raters* di concentrarsi sugli stessi criteri, fornendo anche un metalinguaggio specifico nei casi in cui vi sia disaccordo nel giudizio. Inoltre, l'utilizzo di scale analitiche può contribuire alla formazione dei valutatori, consentendo loro di riflettere in modo più preciso sugli aspetti che compongono la competenza in L2.

La decisione su quali criteri sono rilevanti e comporranno la scala è una questione sempre più centrale per la validità della valutazione poiché "*the heart of the test construct lies here*", come afferma McNamara (2000, pp. 36-37). Infatti, nel caso delle certificazioni, in cui si devono seguire requisiti istituzionali, una delle tappe fondamentali è stabilire criteri chiari per misurare la performance e assegnargli peso e rilevanza definiti. Dopodiché, per ogni criterio si possono includere i descrittori che orienteranno i *raters* nel momento di attribuire il punteggio.

Se da una parte le caratteristiche dei test riflettono la visione che gli elaboratori hanno della lingua e delle necessità comunicative dei candidati, i criteri di valutazione con i loro descrittori esplicitano e rendono operativo il modo in cui comprendono la competenza linguistica e quello che cercano di sottoporre a valutazione (North, 2003). Dunque, anche se non si fa riferimento a una teoria specifica, i parametri stabiliti e il peso attribuito a ciascuno sono indizi delle "basi teoriche" che guidano la valutazione, ovvero, di quali sono le concezioni di lingua e competenza sottostanti al test (McNamara, 1996).

Per fare un esempio, prendiamo i parametri di valutazione adottati dalla Certificazione PLIDA per il livello B1. Ne vengono definiti quattro: *i. contenuto e svolgimento del compito, ii. coerenza e coesione, iii. lessico, iv. grammatica, ortografia, punteggiatura.* Nel caso della Certificazione PLIDA, ciascun criterio ha lo stesso peso e il punteggio massimo da attribuire è sempre di 10 punti. Tuttavia va detto che non sempre è così: capita spesso che i responsabili per l'elaborazione di una griglia decidano che un criterio debba avere un peso maggiore rispetto a un altro. Solo per fare un esempio, si potrebbe decidere di assegnare 10 punti al paramento del contenuto e dello svolgimento del compito e 6 a quello che riguarda la grammatica. Sono appunto questi gli indizi a cui facevamo riferimento poco fa: che cosa si considera più rilevante? Lo svolgimento adeguato del compito o la grammatica? E che cosa rivela questa scelta sull'idea di lingua e comunicazione che guida la valutazione dell'esame?

Indipendentemente dal punteggio massimo stabilito, a ogni criterio possono anche essere associati dei descrittori che hanno la funzione di orientare l'attribuzione adeguata del punteggio secondo la performance del candidato. Si tratta di un altro strumento che ha l'obiettivo di ridurre la soggettività di chi valuta, poiché descrizioni precise sono, in linea di massima, capaci di contribuire a rendere più esplicito e condiviso il modo in cui si interpreta un determinato criterio. Per illustrare questa idea, usiamo ancora una volta un esempio: i descrittori del parametro "grammatica, ortografia, punteggiatura" dell’esame PLIDA B1 (Tab. 1).

Tabella 1. *Descrittori del parametro "grammatica, ortografia, punteggiatura" (PLIDA - B1)*

| Punteggio | Descrittori |
|---|---|
| **10**<br><br>**9** | - Il testo presenta una buona varietà delle strutture previste per il livello, usate in modo corretto e appropriato.<br>- Errori isolati (morfologici, ortografici o di punteggiatura). |
| **8**<br><br>**7** | - Il testo presenta una buona varietà delle strutture previste per il livello.<br>- Gli errori morfologici riguardano singoli elementi della frase e possono essere ripetuti.<br>- Ortografia e punteggiatura sono abbastanza curate; si notano varie incertezze. |
| **6**<br><br>**5** | - Il testo presenta un numero limitato di strutture previste per il livello, non tutte usate con sufficiente padronanza.<br>- Errori (morfologici, ortografici e di punteggiatura) diffusi; in alcuni passaggi la lettura può essere faticosa. |
| **4**<br><br>**3** | - Gli errori (morfologici, ortografici e di punteggiatura) sono numerosi, anche nel caso di strutture elementari; la lettura è molto faticosa. |
| **2**<br><br>**1** | - Gli errori (morfologici e ortografici) impediscono quasi del tutto la comprensione del testo. La punteggiatura è quasi assente. |

Fonte: sito PLIDA (https://plida.dante.global/it/preparati-allesame)

La lettura dei descrittori chiarisce intanto che, per attribuire il punteggio relativo alla grammatica, l'occhio del valutatore non deve rivolgersi a "tutta la grammatica" in generale, ma alle "strutture previste per il livello" e alla loro "varietà". Un'altra questione di fondo è la distribuzione degli errori che possono essere, secondo i descrittori, assenti, isolati, diffusi, numerosi o possono ostacolare la comprensione, la quale entra esplicitamente in gioco solo in negativo ("impediscono quasi del tutto la comprensione del testo") nel caso del punteggio più basso.

Basta questo esempio a confermare il ruolo fondamentale dei descrittori nel processo di attribuzione del punteggio. Se non ci fossero, ogni *rater* potrebbe avere una propria interpretazione dei criteri stabiliti e, di conseguenza, gli esiti dei *test-takers* avrebbero molto probabilmente gravi inconsistenze dovute alle incoerenze nei giudizi.

E arriviamo a un'altra questione essenziale, ovvero, la presenza dei valutatori. Sebbene indispensabile, in particolare quando si tratta di valutare una performance scritta o orale, essa fa sì che la valutazione sia ancora più difficile a causa dell'inevitabile interpretazione personale e della soggettività a cui non si può sfuggire quando il giudizio viene emesso da un essere umano. Come scrive McNamara, "*the rating given to a candidate is a reflection, not only of the quality of the performance, but of the qualities as a rater of the person who has judged it*" (2000, p. 37).

È evidente che la soggettività dei *raters* ha un impatto sull'affidabilità della valutazione (Wigglesworth & Frost, 2017). Non è raro trovare gruppi di valutatori che, pur avendo avuto la stessa formazione e avendo utilizzato la stessa griglia per esprimere il loro giudizio, hanno comunque attribuito punteggi diversi a una stessa performance così da produrre un basso livello di *inter-rater reliability*. Non è, tuttavia, improbabile che la stessa cosa accada persino con un solo valutatore che, a distanza di tempo, a causa delle svariate interferenze che possono entrare in gioco, corre il rischio di attribuire punteggi distinti alla performance degli stessi candidati come dimostra la cosiddetta *intra-rater reliability*. In questo senso, oltre alla continua formazione dei valutatori, uno dei modi più efficaci per limitare la soggettività intrinseca nella valutazione e per controllare che si mantenga l'affidabilità è garantire che i criteri adottati e i rispettivi descrittori siano il più possibile chiari e precisi, visto che servono come punti di riferimento per il giudizio emesso.

***

La questione dell'affidabilità del giudizio dei *raters*, unita all'efficacia della griglia di valutazione, è il tema centrale dello studio di questo libro di Paolo Torresan. Condotta nell'ambito della Certificazione PLIDA, l'investigazione si concentra, in particolare, sulla valutazione della performance dei candidati nei test di produzione scritta dell'esame di livello B1. Un'analisi di questo tipo ha l'obiettivo di individuare eventuali incoerenze presenti nel giudizio dei diversi valutatori, fornendo le basi per cercare in seguito cause e possibili soluzioni.

Partendo dal presupposto che la soggettività ha, soprattutto in questo tipo di test, un notevole peso, è necessario riflettere sulle prospettive che abbiamo a disposizione per limitarla. Torresan ne cita alcune. La prima può essere gestita dal *rater* stesso in maniera autonoma e consiste nella *valutazione ripetuta a distanza di tempo,* quella che si realizza quando un valutatore giudica più volte la stessa performance con un intervallo di tempo tra due o più giudizi. È un procedimento che viene realizzato per identificare incoerenze, oltre che per spingere alla riflessione sui motivi di eventuali discrepanze e su come superarle in futuro.

Si propongono, inoltre, altri due procedimenti per controllare o ridurre la soggettività dei valutatori, in questo caso partendo da un confronto tra più *raters*. Si tratta, da una parte, di *sessioni collettive con un moderatore* e, dall'altra, di *valutazione doppia*. Nonostante richieda un investimento di tempo non sempre possibile, la prima è particolarmente adatta al contesto scolastico, quando due o più docenti valutano insieme la performance di uno stesso apprendente. Nel confrontare il giudizio dato da ciascuno è probabile che emergano criteri impliciti e, di conseguenza, che la discussione tra i docenti consenta di accorgersi di possibili incoerenze o imprecisioni. La valutazione doppia viene, invece, spesso usata in contesto certificatorio: una stessa performance viene valutata da due *raters*, in modo che il voto finale sia calcolato eseguendo la media aritmetica dei due voti assegnati.

L'attribuzione di giudizi uguali o simili depone a favore dell'affidabilità del risultato raggiunto dai candidati; mentre nel caso di una discrepanza significativa, si richiede il giudizio di un terzo *rater*.

Questo tipo di correzione ha ovviamente anche un effetto sulla formazione. Di fatto, mettendola in pratica, si tenderà necessariamente ad affinare sempre di più la propria capacità di identificare gli aspetti linguistici da valutare e di riflettere sui punteggi da attribuire. Inoltre, attività di questo genere servono anche ad acquisire una maggiore sensibilità rispetto alle azioni più o meno consapevoli che compongono l'emissione del giudizio. La grande lezione che di solito si impara è che, quando si è chiamati a valutare, è sempre utile sospettare di sé stessi per accorgersi, in modo via via più consapevole, delle insidie della valutazione.

Esiste, infine, ancora una soluzione per mettere alla prova il livello di (in)affidabilità dei valutatori: si tratta delle analisi statistiche, per le quali occorre il lavoro di un ricercatore, preferibilmente esterno al processo in sé, che abbia, tuttavia, conoscenze approfondite nel campo del *language testing*, oltre che competenze specifiche che gli permettano di analizzare sia il test che i risultati per verificarne soprattutto validità e affidabilità. Il trattamento statistico può essere applicato a diversi aspetti del test: si possono esaminare, tra le altre cose, la prova, la distribuzione del punteggio, le griglie utilizzate per attribuirlo e i punteggi o di uno stesso *rater* a distanza di tempo oppure di un gruppo di *raters*.

Nello studio di Torresan è stata eseguita la *Many-Facet Rasch Measurement,* cioè l'analisi Rasch estesa a più variabili. Tra le variabili esaminate, vi sono le qualità del valutatore (la sua generosità, il suo grado di coerenza, la sua capacità di discriminare i candidati più abili dai meno abili) e del gruppo di valutatori (il loro grado di omogeneità nell'attribuzione dei punteggi), oltre che l'efficacia dello strumento utilizzato per valutare ovvero della griglia di valutazione.

Esaminando diversi aspetti che contribuiscono all'affidabilità dei *raters* della Certificazione PLIDA, Torresan dimostra che si tratta di una fase del processo di valutazione che richiede speciale attenzione. Criteri e descrittori chiari e operativi orientano il giudizio dei valutatori e sono uno strumento di cui tenere conto per garantire risultati sempre più affidabili. Esistono valutatori più o meno generosi, valutatori più o meno coerenti, valutatori più o meno capaci di distinguere i diversi livelli di performance, così come esistono griglie e descrittori più o meno efficaci. Queste caratteristiche vanno riconosciute e discusse per poter essere poi condivise ed esercitate.

Ricerche come quella presentata in questo libro, unite ad attività di formazione, alla condivisione dei criteri e agli strumenti che oggi ci offre la tecnologia, ci permettono di confrontarci e di riflettere sullo spinoso processo della valutazione. Non è sempre fattibile né facile metterle in pratica, ma sono di certo prospettive importanti grazie alle quali potremo avvicinarci a una valutazione che sia, in tutte le sue manifestazioni, sempre più valida e etica. A tutto vantaggio di valutatori e valutati.

## Riferimenti bibliografici

ASCHBACHER, P. A. Performance assessment: State activity, interest, and concerns. *Applied Measurement in Education*, 4(4), 1991, pp. 275-288.

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

BACHMAN, L F.; PURPURA J. E. Language Assessments: Gate-Keepers or Door-Openers? In: SPOLSKY, B.; HULT, F. M. (eds.), *The Handbook of Educational Linguistics*. Blackwell Publishing Ltd., 2008, pp. 456-468.

BARNI, M. La verifica e la valutazione. In: DE MARCO, A. (a cura di), *Manuale di glottodidattica: insegnare una lingua straniera*. Carocci, 2000, pp. 155-174.

BARNI, M. "La valutazione delle competenze linguistico-comunicative in L2". In: VEDOVELLI, M. (a cura di), *Manuale della certificazione dell'italiano L2*. Carocci, 2005, pp. 29-46.

HAUPTMAN, P. C.; LEBLANC, R.; WESCHE, M. B. (eds.), *Second language performance testing*. Ottawa: University of Ottawa, 1985.

KANE, M. B.; MITCHELL, R. (eds.), *Implementing performance assessment: Promises and challenges*. Mahwah, NJ: Lawrence Erlbaum Associates, 1996.

MCNAMARA, T. Modelling performance: Opening Pandora's box. *Applied Linguistics*, 16(2), 1995, pp. 159-179.

MCNAMARA, T. *Measuring Second Language Performance*. Edinburgh Gate: Addison Wesley Longman Limited, 1996.

MCNAMARA, T. Performance testing. In: CLAPHAM, C.; CORSON, D. (eds.), *Encyclopedia of language and education: Volume 7 Language testing and assessment*. Dordrecht, NL: Kluwer Academic, 1997, pp. 131-139.

MCNAMARA, T. *Language Testing*. Oxford University Press, 2000.

MCNAMARA, T.; SHOHAMY, E. Language Tests and Human Rights. *International Journal of Applied Linguistics*, 2008, pp. 89-95.

NORRIS, J.; BROWN, J.; HUDSON, T.; YOSHIOKA, J. *Designing second language performance assessments.* Honolulu, HI: University of Hawai'i, Second Language Teaching & Curriculum Center, 1998.

NORTH, B. Scales for rating language performance: Descriptive models, formulation styles, and presentation formats. *TOEFL Monograph 24*, 2003.

SHOHAMY, E. Performance assessment in language testing. *Annual Review of Applied Linguistics*, 15, 1995, pp. 188-211.

SKEHAN, P. Tasks and language performance assessment. In: BYGATE, M.; SKEHAN, P.; SWAIN, M. (eds.), *Researching pedagogic tasks: Second language learning, teaching and testing*. Harlow, UK: Pearson Education, 2001, pp. 167-185.

WIGGLESWORTH, G.; FROST, K. (2017). Task and performance-based assessment. In: SHOHAMY, E., OR, I. & MAY, S. (eds.), *Language Testing and Assessment.* Encyclopedia of Language and Education, third edition. Springer, 2017, pp. 121-134.

# SEZIONE A

# LA VALUTAZIONE DELLA *PERFORMANCE*. COORDINATE GENERALI

## 1. *Performance* vs prestazione

Il termine *performance* rimanda a un'azione complessa, il cui esito è difficilmente valutabile. Si consideri la *performance* di un/a cuoco/a, di un(')avvocato/a, di un/a ginnasta, di un/a ballerino/a, di un(')insegnante o di un/a cantante. In tutte queste situazioni, il giudizio è formulato da una giuria/commissione; così facendo, le impressioni di un/a giurato/a, giudice, commissario/a vengono mitigate/compensate dal giudizio degli altri.

Il termine *prestazione*, diversamente da *performance* (con cui tuttavia è spesso tradotto), indica invece, nella nostra ottica, un comportamento più semplice. La prestazione di un centometrista o di un ciclista si dice "ottima" in relazione alla velocità, fattore facilmente misurabile (non serve un giudice, in questo caso; è necessario un cronometro).

Rientra nell'ambito della *performance* anche la comunicazione (in qualsiasi forma essa si dia, e qualsiasi sia il tipo di testo/discorso prodotto). Si ha, pure in questo caso, un'azione complessa, difficile da essere valutata. In particolare, se il tipo di testo ha una struttura aperta, ovvero si presta ad ampi margini di interpretazione, il divario tra i giudizi può essere molto ampio. Uno stesso film, per esempio, può essere ritenuto di poco valore da alcuni e osannato da altri; e lo stesso vale per una canzone, per una poesia, per una *pièce* teatrale, ecc.

## 2. È possibile valutare la *performance*?

Stando a quanto detto, sembrerebbe che sia *estremamente* difficile formulare un giudizio obiettivo in generale in merito a una *performance*, posto che ogni attribuzione di valore è soggettiva. In realtà, ciò su cui spesso poggia il valore attribuito alla *performance* è il consenso di più esperti. Anche laddove un'opera d'arte sia rigettata dal pubblico, per esempio, è necessario che vi sia un consenso perlomeno tra i critici e tra gli specialisti affinché possa essere dichiarata tale. Parimenti, in ambito scolastico, un testo ha valore nella misura in cui più insegnanti, ovvero più professionisti dell'educazione, lo riconoscono come tale.

## 3. Insidie nella valutazione della *performance*

Il giudizio di chi è preposto a valutare la *performance* si attiene, in genere, ad una scala, che può essere costituita da etichette (*scarso, sufficiente, discreto, buono, ottimo*) o da numeri. La scala riduce ma non neutralizza il punto di vista del singolo: ciascun valutatore, in effetti, può interpretare a suo modo i punteggi.

Nello schema a seguire, ispirato a McNamara (1996), esemplifichiamo quanto detto. Abbiamo tre valutatori deputati alla valutazione dell'abilità del parlato da parte di apprendenti di una lingua straniera. Ciascuno usa la scala |0÷10| in modo diverso. Mettiamo in risalto le differenze salienti:

- è più facile, per un candidato, raggiungere la sufficienza se a valutarlo è il valutatore [**B**], anziché il valutatore [**A**] o il valutatore [**C**]; il voto 6 attribuito da [**B**] equivale, infatti, al voto 5 attribuito da [**A**] ed approssimativamente al voto 5.7 attribuito da [**C**];
- sia il valutatore [**B**] che il valutatore [**C**] non usano l'intera scala; [**B**] omette il punteggio 1; [**C**] usa un *range* molto ristretto: |4÷9|;

- il valutatore [**A**] assegna il punteggio massimo a compiti che risultano privi di qualsiasi errore/inadeguatezza; il valutatore [**B**], invece, ha un atteggiamento più elastico: ai suoi occhi non è necessario raggiungere la perfezione per raggiungere l'eccellenza;
- nel caso di [**C**] l'intervallo |5÷7| è molto esteso; ciò lascia supporre una propensione all'utilizzo della parte centrale della scala ("tendenza centrale").

**A.** 1 2 3 4 5 **6** 7 8 9 10

**B.** 2 3 4 5 **6** 7 8 9 10

**C.** 4 5 **6** 7 8 9

Per conferire maggiore attendibilità al giudizio, alla scala si possono accompagnare dei descrittori per livello (*scala olistica*) oppure si possono definire più sottoscale, ciascuna delle quali dedicata a una componente specifica e corredata di descrittori, livello per livello (*scala analitica*; cfr. Spinelli 2014). Così, per esempio, se si dovesse valutare la presentazione orale su *Power Point* su un qualsiasi tema da parte di un allievo, si potrebbero assegnare voti distinti a componenti quali

- il contatto visivo con chi ascolta,
- il volume della voce,
- la chiarezza della pronuncia,
- l'elaborazione critica dei contenuti,
- la sequenziazione delle informazioni,
- il rispetto dei tempi,

- l'attinenza al tema, e così via,

per calcolare, infine, una media aritmetica (o ponderata, nel caso in cui si volesse dare maggiore peso a una componente rispetto alle altre).

Il fatto che si scompatti la *performance* in tante componenti, e che a ciascuna di esse corrispondano descrittori distinti per livello, vale a garantire una maggiore precisione al giudizio:

- *si tende ad evitare che il giudizio sottorappresenti il costrutto di riferimento*, ovvero che si fornisca un'immagine parziale (e quindi distorta) dell'abilità a cui rimanda la *performance* (in altre parole, si evita che il giudizio relativo a una componente sia esteso all'intera *performance* – è il caso che si verifica, per esempio, allorquando, nel valutare la qualità di uno scritto, il valutatore si attiene alla sola accuratezza);
- *si tende ad evitare che il giudizio sia influenzato da fattori esterni al costrutto di riferimento* (la simpatia che il valutatore può provare per un candidato, per esempio, è un fattore esterno rispetto all'abilità in questione – a meno che non si stia valutando la *performance* di un comico).

Nonostante una griglia analitica garantisca una maggiore affidabilità del giudizio, margini di soggettività possono pur sempre agire nella formulazione del giudizio. Consideriamo, per esempio, una griglia usata per valutare la *performance* linguistica di un apprendente di una lingua straniera o seconda: la griglia si compone di descrittori che hanno una natura qualitativa, e non quantitativa – i margini tra i descrittori sono quindi elastici, non definiti in maniera univoca. Anche laddove si ricorra a dei quantificatori, essi non sono dei numerali (delle percentuali, delle occorrenze, ecc.) ma degli avverbi che rimandano a nozioni piuttosto vaghe (*molto, abbastanza, poco, di rado, frequentemente, del tutto,* ecc.), soggette a interpretazione.[1]

Più in generale l'affidabilità del giudizio può essere minacciata da fattori esterni al costrutto i quali possono agire al di sotto della soglia della coscienza del valutatore. Si vedano, a tal proposito, nella tabella 1 (pagina che segue), i fattori che possono condizionare il giudizio durante la valutazione di uno scritto in lingua straniera (cfr. Hughes *et al.* 1983; Weigle 2002; Myford, Wolfe 2003, 2004, Eckes 2015). Alcuni di essi tendono ad essere più stabili, mentre altri sono legati alle circostanze; entrambi opacizzano comunque l'oggetto di misurazione, agendo da schermo deformante durante la rilevazione della *performance*.

[1] Del resto, se anche si adottasse un criterio numerico il problema non sarebbe risolto, posto che, per esempio, non tutti gli errori e non tutte le inadeguatezze hanno lo stesso peso nella resa del messaggio (cfr. Pallotti 2005) e per di più esistono aspetti della *performance* difficilmente quantificabili.

Tabella 1. *Fattori che incidono sull'affidabilità del giudizio durante la valutazione di uno scritto in lingua straniera*

| **FATTORI PIÙ STABILI** | **FATTORI PIÙ ALEATORI** |
|---|---|
| *il timore di danneggiare lo studente*[2] *e/o di essere contestato*[3]. Esso si traduce nell'uso ristretto della scala (*restrizione di intervallo*[4]; laddove i giudizi gravitino | *il grado di stanchezza* (maggiore è lo sfinimento psico-fisico, meno concentrato è il valutatore; può consultare la griglia in modo approssimativo oppure |

[2] Per Thorndike e Haagen la preoccupazione di arrecare un danno a un candidato spiega il comportamento generoso di molti valutatori. I due scrivono (1977[4]: 448-449; il corsivo è nostro): "We have pointed out that the rater is often as much committed to the people he is rating as he is to the agency for which ratings are being prepared. Over and above this, *there seems to be a widespread unwillingness on the part of raters*, at least in the United States, *to damn a fellow human with a low rating*. The net result is that ratings tend quite generally to pile up at the high end of any scale. The unspoken philosophy of the rater seems to be 'one person is as good as the next, if not a little better,' so that 'average' becomes in practice not the midpoint of a set of ratings but near the lower end of the group. It is a little like the commercial classification of olives, where the tiniest ones are called 'medium,' and they go from there through 'large' and 'extra large' to 'jumbo' and 'colossal'".

[3] Il timore di essere contestato, da parte del valutatore, determina un atteggiamento generoso (Yoder, Staudohar 1982[7]: 219): "a desire to err on the generous side, to avoid controversy by giving each ratee the benefit of the doubt".

[4] La restrizione di intervallo può accompagnarsi ad un eccesso di generosità, qualora i punteggi usati siano solo quelli della parte superiore della scala (*tendenza superiore*). Al contrario, può accompagnarsi ad un eccesso di severità quando i valori usati sono quelli della parte inferiore della scala (*tendenza inferiore*). Laddove i valori si concentrino nella parte intermedia della scala, si parla di *tendenza centrale* (si veda la nota a seguire). In tutti e tre i casi la discriminatività di una prova (ovvero la capacità di distinguere finemente i gradi di competenza) viene meno.

| | |
|---|---|
| attorno alla media, si parla di *tendenza centrale*)[5] | attenersi a parametri personali, anziché ai descrittori della scala) |
| *la scarsa discriminazione tra componenti della scala* (il valutatore tende a uniformare il giudizio tra componenti distinte)[6] | *l'umore* |
| *l'esperienza accumulata nel valutare*[7] | *il momento della giornata* |

[5] Guildford associa la tendenza centrale a una certa insicurezza da parte del valutatore (1954: 278; il corsivo è nostro): "One of the reasons for the error of central tendency is that raters hesitate to give extreme judgments and thus tend to displace individuals in the direction of the mean of the total group. *This is perhaps most common in rating individuals whom the raters do not know very well*". La tendenza centrale può altresì essere dovuta ad un'analisi superficiale della *performance*: si considerano solo gli aspetti macro, più facilmente rilevabili.

[6] Si tratta di un caso specifico di *effetto alone*. Tale effetto si può manifestare quando il giudizio si basa prevalentemente su una componente, trascurando le altre (il concetto, in tal caso, si sovrappone a quello di sottorappresentazione del costrutto), oppure su una impressione generale sul candidato oppure ancora, come dicevamo, quando più tratti della competenza vengono assimilati (cfr. Fisicaro, Vance 1994). A riguardo di quest'ultimo caso, occorre comunque distinguere un *effetto alone vero* da un *effetto alone apparente*: può capitare che, nel definire il costrutto, colui che allestisce la griglia distingua tratti in realtà fortemente correlati: in tale circostanza, la replica dell'attribuzione di giudizio su componenti distinte da parte del valutatore non è un errore logico (qualcosa che pertiene, dunque, al suo modo di intendere), quanto il riflesso dell'interdipendenza tra elementi che fanno capo a un'unica dimensione (Cooper 1981; Murphy 1982; Bartlett 1983; Pulakos *et al.* 1986).

[7] Secondo Weigle (1998) colui che valuta da lunga data tende ad essere più tollerante rispetto al collega che ha meno esperienza; nel caso specifico, però, di un consesso di valutatori che operano insieme, Eckes (2015) è convinto del contrario, posto che il *rater* più esperiente, assumendo il

| | |
|---|---|
| *la propria formazione come apprendente* e/o *modelli culturali di appartenenza*[8] | *il contesto in cui uno lavora* (la temperatura dell'ambiente, che può essere eccessiva o troppo bassa; la presenza di fonti di disturbo; ecc.) |
| *il confronto con sé stesso come modello*[9] | *Il momento in cui un elaborato è valutato rispetto ad una serie*[10] |
| *l'atteggiamento verso la lingua/cultura del candidato* | *il tempo a disposizione* (condizioni di pressione agiscono negativamente sull'attendibilità del giudizio) |
| *il grado di familiarità del valutatore con la lingua di origine del candidato* o più in generale con le lingue note al candidato (un valutatore che conosca a fondo la/le lingua/e nota/e all'apprendente tende a essere tollerante rispetto ad errori di *transfer* linguistico) | *l'atteggiamento verso il candidato* (il grado di simpatia o l'impressione generale che il valutatore si è fatto su di lui, sulla base dell'apparenza o dei risultati ottenuti in *performance* precedenti; ecc.) |
| *l'atteggiamento verso certi tipi di calligrafia*[11] | *il grado di accordo con il contenuto* |
| *la preferenza verso certi generi testuali* | *la qualità dei testi appena valutati* (*effetto contrasto*): ciò significa che un testo mediocre, a fronte di una serie |

ruolo di modello nei confronti di quelli meno esperienti, tende ad essere più severo.

[8] Cfr. Chen *et al.* 1995; Aiken 1996.

[9] Cfr. Murray 1938; Latham *et al.* 1975.

[10] Alcune ricerche avvisano che i testi valutati per primi tendono a ricevere giudizi meno severi rispetto a quelli valutati successivamente (cfr. Godshalk *et al.* 1966; Coffman, Kurfman 1968).

[11] Hughes *et al.* (1983) sostengono che valutatori la cui grafia è chiara tendono ad essere severi con apprendenti la cui grafia è poco chiara.

| | |
|---|---|
| | di testi scarsi, può apparire migliore di quel che effettivamente è[12] |
| Laddove la scala sia composta da fasce di punteggio (es. \|4-5\|; \|5-6\|; \|6-7\|; ecc.), *la tendenza a usare solo uno dei punteggi presenti nella fascia* (denominabile come *effetto localizzazione*) | *il voto espresso su una categoria contigua* (*errore di prossimità*)[13] |

Questi fattori si possono combinare tra loro. Il risultato è un giudizio caratterizzato da

- *disallineamento*, e cioè assegnazione di punteggi maggiori o minori rispetto all'effettiva abilità del candidato;
- *incoerenza* o *erraticità*; il valutatore non dà prova di un comportamento stabile; può rivelarsi ora generoso, ora severo.[14]

È possibile, peraltro, una situazione in cui *disallineamento* ed *erraticità* co-occorrano. Si pensi, per esempio, a un valutatore tendenzialmente severo (*disallineamento*) che però, in alcune circostanze, per una qualche ragione, formuli giudizi eccessivamente generosi (*erraticità*).[15]

---

[12] Cfr. Stalnaker 1936, Guildford 1954.
[13] Immaginiamo una griglia analitica in cui la componente grammaticale preceda la componente lessicale: il voto espresso in merito alla prima può influenzare quello relativo alla seconda (cfr. Stockford, Bissell 1949).
[14] Questo secondo caso è legato, in particolare, all'impatto esercitato dai fattori aleatori.
[15] In un percorso formativo, l'erraticità può essere ritenuta più grave, visto che, mancando una stabilità del giudizio, allo studente viene meno un parametro mediante il quale definire la qualità della propria *performance*.

*Disallineamento* ed *erraticità* inficiano la validità del giudizio, posto che il voto attribuito non rivela la reale competenza del candidato. Se il voto assegnato sottostima la competenza, nei confronti del candidato viene perpetrata un'ingiustizia. Può accadere, in tal frangente, che la percezione che questi ha di sé e del proprio valore venga minacciata e che il suo impegno venga meno. Qualora la prova abbia un carattere certificatorio, un intero progetto di vita può addirittura sfumare.

Una situazione di svantaggio può darsi anche nel caso della sovrastima della competenza. Giudizi eccessivamente generosi possono illudere il candidato di possedere una competenza che non ha: messo alla prova in un contesto di uso spontaneo della competenza, egli può fallire.

## 4. Strategie volte ad accertare e a migliorare la qualità del giudizio

Le serie conseguenze che una valutazione inaffidabile può esercitare a livello psicologico, professionale e sociale dovrebbero indurre gli enti deputati a valutare la *performance* a porsi il problema della verifica dell'affidabilità dei valutatori.

Suggeriamo, a tal proposito, quattro soluzioni. La prima può essere svolta dal valutatore in autonomia, la seconda e la terza prevedono un confronto tra più valutatori; la quarta comporta un'indagine a carico di un ricercatore:

- *la valutazione ripetuta a distanza di tempo* (§ 4.1.);
- *sessioni collettive con un moderatore* (§ 4.2.);
- *la valutazione doppia* (§ 4.3.);
- *analisi statistiche* (§ 4.4.).

---

È come se il valutatore agisse alla pari di una *roulette* russa: non si sa mai cosa ci si può aspettare da lui.

## 4.1. La valutazione ripetuta a distanza di tempo

Per ridurre l'*erraticità*, un valutatore può, autonomamente, valutare più volte una stessa *performance* a distanza di tempo. Per esempio, dopo aver valutato una composizione scritta, può tornare sullo stesso testo a distanza di un paio di settimane, per esprimere un nuovo giudizio. Successivamente appura se si diano delle discrepanze tra il primo e il secondo giudizio, e ragiona su cosa possa averle determinate e su come potervi porre rimedio in futuro.

## 4.2. Sessioni collettive con un moderatore

A scuola, in sede di scrutini, gli insegnanti si confrontano tra loro sui voti da assegnare agli studenti nelle rispettive materie. Si tratta di uno scambio che non è detto incida sull'affidabilità del giudizio, posto che la variazione dei voti riferiti a un allievo può essere esclusivamente attribuita al diverso impegno che questi ha profuso nello studio delle discipline. In altre parole, se lo studente è bravo in matematica, fisica, chimica, biologia e storia, ma non lo è in inglese né in latino, il problema si ritiene riguardi l'apprendente, poco portato per le lingue, e non necessariamente la competenza docimologica dei docenti di inglese e di latino.

Un metodo empirico per escludere che il problema risieda nel docente è quello di indire delle sessioni in cui più docenti *della stessa materia* valutano la *performance* dell'alunno. Nell'esempio del docente di latino o di inglese, si potrebbero organizzare degli incontri nei quali più docenti di latino valutano una certa versione (o più di una) effettuata dall'allievo, e più docenti di inglese valutano una produzione scritta/orale in inglese (o più di una) realizzata dall'allievo. Dal confronto, potrebbe emergere che il giudizio negativo del docente di inglese è influenzato dal comportamento indisciplinato dell'apprendente, o che il giudizio altrettanto poco lusinghiero del docente di latino è dovuto alle versioni troppo difficili somministrate in sede di verifica (il problema, in tal caso, è da attribuire al compito e non al giudizio).

Insomma, durante la discussione tra colleghi possono venire alla luce criteri impliciti che stanno alla base del disallineamento del giudizio.

Nella tab. 2, *infra*, a titolo di esempio, abbiamo tabulato i voti assegnati da insegnanti di italiano (I-IX) che operano presso l'Istituto Italiano di Cultura di Lima (Perù), in riferimento alla componente grammaticale di scritti elaborati da studenti di livello A2 (*a-g*) in risposta ad un medesimo *task*. I valutatori si sono serviti di una griglia analitica. A seguito dell'attribuzione del voto, gli insegnanti sono stati invitati a confrontarsi tra loro; lo scambio ha coinvolto, in particolare, i docenti che hanno espresso i *giudizi estremi* (tali attribuzioni sono evidenziate con un fondo colorato nello schema: in rosso, i punteggi maggiori; in giallo, quelli minori). Per esempio, in riferimento all'elaborato [*d*] sono stati chiamati al confronto i valutatori [**IV**] e [**V**], i quali hanno assegnato rispettivamente il voto più basso (2) e il voto più alto (9).

In generale uno scambio di questo tipo permette a chi valuta di prendere coscienza di dove si colloca il proprio giudizio rispetto alla norma (e cioè al giudizio dei più); nel momento in cui emergono le motivazioni che hanno sostenuto le diverse attribuzioni di giudizio, ciascuno ha modo di confrontare i criteri che hanno guidato la propria decisione con quelli degli altri: può rendersi conto, per esempio, di avere una visione eccessivamente concentrata su una componente della *performance*, o di obbedire a dei parametri eccessivamente rigorosi, e via di seguito.

Tabella 2. *Voti assegnati da diversi valutatori (I-IX) alla componente grammaticale di testi scritti realizzati da diversi candidati (a-g) in riferimento ad uno stesso compito*

| | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** | **IX** |
|---|---|---|---|---|---|---|---|---|---|
| **A** | 3 | 3 | 4 | 3 | 1 | 3 | 0 | 3 | 4 |
| **B** | 6 | 6 | 7 | 7 | 9 | 7 | 4 | 8 | 8 |
| **C** | 3 | 2 | 4 | 2 | 2 | 3 | 0 | 4 | 5 |
| **D** | 6 | 6 | 7 | 2 | 9 | 8 | 4 | 8 | 7 |
| **E** | 3 | 4 | 6 | 4 | 3 | 7 | 2 | 5 | 4 |
| **F** | 7 | 8 | 8 | 6 | 8 | 7 | 4 | 9 | 7 |
| **G** | 4 | 4 | 6 | 6 | 7 | 7 | 4 | 6 | 5 |

### 4.3. La valutazione doppia

In contesto certificatorio, al fine di aumentare il grado di affidabilità del giudizio, una stessa composizione può essere valutata da due valutatori; il voto finale viene poi definito dalla media aritmetica dei voti assegnati dai due. Nel caso in cui vi sia una discrepanza significativa (>20%), è chiamato in causa un terzo valutatore; si calcola, quindi, la media tra il voto assegnato da quest'ultimo e il voto più prossimo (Shaw, Weir 2007).

### 4.4. Indagini statistiche

Si possono condurre studi statistici sui voti assegnati dal medesimo gruppo di valutatori a una serie di *performance*.

L'analisi più semplice è il calcolo dell'*indice di correlazione*. Più complessa, ma al tempo stesso più completa, è l'*analisi Rasch estesa* *a più variabili*, chiamata *Many-Facet Rasch Measurement* (d'ora in poi MFRM). L'analisi si rivolge a più variabili (o "facce", *facets*), alcune delle quali hanno una natura politomica [i dati non sono rappresentati in maniera binaria 1-0 (giusto/sbagliato), come avviene per gli *item* di test chiusi, ma sono riferiti agli intervalli della scala in adozione, es. |0÷10|]. Oltre ad accertare il grado di affidabilità del valutatore, l'analisi MFRM fornisce informazioni sull'efficacia della griglia in adozione.[16] Il *software* che ne permette l'applicazione è Facets®. Quantunque il linguaggio di programmazione sia complesso e la tabulazione dei dati richieda molto tempo, la quantità di indici e di grafici messi a disposizione è notevole. La sezione che segue è dedicata ad esplorare le caratteristiche di tale indagine.

[16] Per approfondimenti, cfr. Prieto, Nieto 2014; Eckes 2015.

# SEZIONE B

# L’ANALISI RASCH ESTESA A PIÙ VARIABILI (MFRM)

## 5. Introduzione

L’analisi Rasch non rientra nell’ambito della *statistica descrittiva* (statistica classica) ma costituisce un esempio di *statistica prescrittiva*: fornisce un modello probabilistico a cui vengono confrontati i dati raccolti.

Nel caso dell’*Item Analysis*, ovvero dell’indagine di *item* di prove chiuse (come per esempio quesiti a scelta multipla volti alla verifica dell’abilità di comprensione), l’Analisi Rasch è applicata a due variabili: il grado di difficoltà degli *item* e l’abilità del candidato; il *software* di riferimento è Winsteps® (cfr. Green 2013; Torresan 2015).

Nel caso della valutazione dei valutatori della *performance* l’Analisi Rasch è applicata a un numero maggiore di variabili, alcune delle quali possono avere un carattere politomico (considerato il *range* relativamente esteso di punteggi presenti in una scala di valutazione, generalmente |0÷10|). In tal caso si parla di *Many-Facet Rasch Measurement* e il *software* di riferimento è Facets®.

Una *Many-Facet Rasch Measurement* stabilisce una scala comune (quindi opera una linearizzazione) alle variabili unidimensionali (ovvero, afferenti a una medesima dimensione, e cioè a uno stesso costrutto) oggetto di indagine (cfr. Bond, Fox 2007).

Dobbiamo concepire, in effetti, il voto espresso dal valutatore della *performance* come un vero e proprio costrutto, ovvero una realtà “costruita” da più elementi o dimensioni (o, nel lessico del software, da “facce”, *facets*). Il voto espresso dal *rater* (voto che rappresentiamo con la lettera **V**) è perlomeno il risultato di tre fattori:

- *la competenza del candidato* (**A**),
- *la generosità/severità del valutatore* (**B**),
- *il grado di complessità/facilità del compito* (**C**).[17]

Noi non vediamo queste "facce": non sappiamo quale sia per esempio l'incidenza del grado di severità del valutatore sul voto assegnato all'allievo.

Per sapere esattamente qual è la competenza del candidato, nel suo valore assoluto ("reale") (**A**), al netto delle altre componenti che concorrono al voto osservato (**V**), dobbiamo operare una sottrazione rispetto al valore osservato:

$$A= V - B - C$$

Un esempio: in una scala |0÷10|, alla *performance* di un candidato è stato assegnato il voto 8.5 da parte di un valutatore che ha un grado di generosità pari a 1.5 (il valutatore tende a regalare 1.5 voti rispetto alla norma), con un compito molto facile, sottocalibrato di due punti rispetto al livello di competenza del candidato. L'abilità vera e propria è calcolata come segue:

$$A= 8.5 - (1.5) - (2.0)$$
$$A= 8.5 - 1.5 - 2$$
$$A=5$$

Benché sia stato giudicato molto positivamente (8.5), il candidato, in questo caso, ha una competenza il cui valore "reale" è al di sotto di 3.5 punti rispetto al valore "osservato" (il valore della competenza è pari a 5).

Facciamo un altro esempio: alla *performance* di un candidato è stato assegnato il voto di 5.7 da parte di un valutatore leggermente

[17] È raro, in effetti, che un valutatore sia perfettamente equo, privo cioè di tratti di generosità/severità, così come è raro il fatto che il compito sia calibrato millimetricamente sul livello di competenza del candidato.

severo, al quale corrisponde un grado di generosità pari a –0.5 (il valutatore tende a togliere 0.5 voti rispetto alla norma), sulla base di un compito leggermente sovracalibrato rispetto al livello di competenza del candidato (–0.2). L'abilità vera e propria è calcolata come segue:

$$A= 5.7 - (-0.5) - (-0.2)$$
$$A= 5.7 + 0.5+0.2$$
$$A=6.4$$

In questa seconda circostanza si dà l'opposto: il candidato è stato penalizzato di 0.7 punti (voto osservato=5.7; voto "reale"=6.4). Sempre in questo caso, qualora il punteggio a cui corrisponde la sufficienza (*cut score*, punto di taglio) nella scala in adozione sia pari a 6 (come avviene in molti contesti scolastici), la *performance* del candidato verrebbe ingiustamente giudicata insoddisfacente.

**6. Il calcolo dei residui**

Come funziona la logica del modello Rasch? Come viene calcolato il valore "reale"?

Nel *software* Facets® vengono inseriti, da parte del ricercatore, i voti assegnati ("observed scores") da diversi valutatori a una serie di *performance* prodotte da più candidati (o più specificamente, nel caso della griglia analitica, a componenti della *performance*). Tali voti, chiamati dal sistema "grezzi" ("raw score"), sono dei punteggi 'agglomerati', frutto cioè, come dicevamo, della somma di più fattori.

Il programma è in grado, attraverso una complessa triangolazione, di stabilire i *valori attesi* ("expected scores") o "reali" ("true scores"), al netto dei contributi degli altri fattori (e cioè del grado di generosità/severità del valutatore e della facilità/difficoltà del compito).

La differenza tra i valori osservati ("observed scores" o "raw scores") e i valori attesi o "reali" ("expected scores" o "true scores") è chiamata *residuo* (“residual”). Tutti gli indici forniti dall’analisi MFRM sono ricavati a partire dalla *somma dei residui*. Maggiore è la somma dei residui, meno affidabile è il giudizio del valutatore.

Spieghiamo meglio attraverso un esempio, ispirato a McNamara 1996. Nella tab. 3, *infra*, nove valutatori (I-IX) formulano un giudizio mediante un voto, su una scala |0÷10|, riguardo alle composizioni di sette allievi (*a*-*g*), riferite al medesimo compito.

Tabella 3. *Voti assegnati da un ipotetico gruppo di valutatori a diversi elaborati*

| | **I** | **II** | **III** | **IV** | **V** | **VI** | **VII** | **VIII** | **IX** |
|---|---|---|---|---|---|---|---|---|---|
| **a** | 5 | 6 | 8 | 9 | ? | 7 | 6 | 8 | 6 |
| **b** | 4 | 3 | 3 | 4 | 5 | 6 | 5 | 6 | 7 |
| **c** | 1 | 3 | 2 | 4 | 4 | 3 | 2 | 5 | 4 |
| **d** | 3 | 4 | 5 | 3 | 2 | 4 | 4 | 3 | 4 |
| **e** | 8 | 10 | 10 | 10 | 8 | 7 | 8 | 9 | 7 |
| **f** | 6 | 7 | 8 | 8 | 6 | 6 | 5 | 5 | 6 |
| **g** | 7 | 6 | 8 | 8 | 6 | 7 | 7 | 6 | 6 |

Immaginiamo che il sistema voglia calcolare il valore “reale” attribuito dal valutatore [**V**] al compito dello studente [a] (casella gialla). Il sistema inferisce questo valore attraverso un calcolo di probabilità che tiene conto dei giudizi espressi dallo stesso valutatore sugli altri compiti e dei voti espressi su quella composizione dagli altri valutatori.

Supponiamo che tale valore “reale”, attribuito dal sistema, sia 7. Può capitare che il valutatore [**V**] abbia assegnato un voto più basso (es. 6). La differenza tra il valore “osservato” e il valore “reale”, il *residuo*, è in questo caso negativa (6-7= –1). Può capitare, al contrario, che il valutatore abbia assegnato un voto più alto, tipo 8; il residuo in questo caso è positivo (8-7= 1). La tendenza ad accumulare *residui positivi* attesta un *disallineamento per eccesso* o *positivo* (il valutatore è generoso); viceversa, la tendenza ad accumulare *residui negativi* attesta un *disallineamento per difetto* o *negativo* (il valutatore è severo). L'*erraticità*, altro caso ancora, è collegata invece a un comportamento instabile: il valutatore può essere talora severo e altre volte generoso.

### 7. La supposizione di un comportamento di media conformità

Va precisato che il modello non prevede un'affidabilità in grado massimo da parte del valutatore. L'analisi prefigura piuttosto un comportamento di media conformità, tanto in termini di allineamento (*intra-rater reliability*), quanto in termini di coerenza (*inter-rater reliability*).

Spieghiamo meglio, servendoci di un esempio (fig. 1, *infra*). Immaginiamo che l'elaborato prodotto da 8 candidati (*a-h*), perfettamente allineati tra loro in termini di competenza, sia valutato da tre valutatori (I-III):

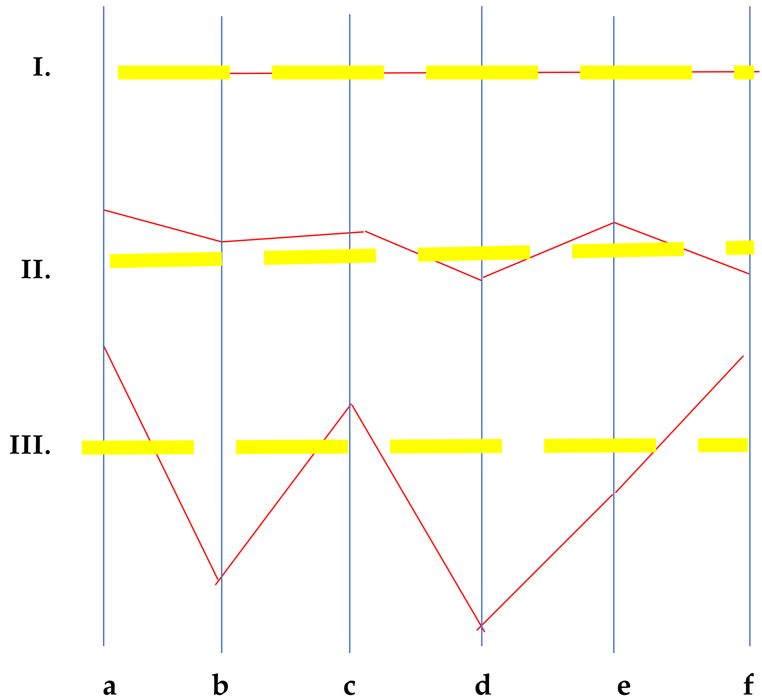
- il valutatore [**I**] è massimamente affidabile (è coerente ed è allineato), il suo giudizio rispecchia l'uniformità della competenza [la linea del giudizio (in rosso), si sovrappone – e nasconde in parte – la linea (tratteggiata) della competenza];
- il valutatore [**II**] presenta un'oscillazione moderata del giudizio;
- al valutatore [**III**] corrisponde, infine, un disallineamento significativo, dai valori opposti (egli è tendenzialmente severo, tuttavia in alcune occasioni non lo è).

Ebbene, il programma si attende che, in termini probabilistici, un valutatore assuma un comportamento del secondo tipo ([**II**]), di media conformità al modello (o di inaffidabilità contenuta). Al contrario, l'eccessiva predittività/uniformità del giudizio, quindi la perfetta coerenza, com'è nel caso [**I**], così come, al contrario, l'impredicibilità del giudizio, come nel caso [**III**], meritano di essere indagate.

In altre parole, benché in linea teorica una somma dei residui pari a zero possa essere ritenuta la situazione migliore in assoluto, di fatto non è lo; anzi, se ciò si verificasse, può destare qualche preoccupazione. È improbabile, in effetti, che un valutatore agisca con una precisione assoluta e costante; il fatto che il suo giudizio oscilli meno del previsto può essere determinato da *bias*, come un uso ristretto della scala.

Il modello probabilistico Rasch, insomma, non prevede un comportamento automatizzato, privo di sfumature da parte di chi giudica la *performance*; al contrario, assume che il giudizio del *rater* abbia dei margini contenuti di variabilità, tanto rispetto alla norma, cioè al giudizio degli altri (*inter-rater reliability*), quanto rispetto a sé stesso (*intra-rater reliability*).

Figura 1. *Modelli di comportamento del valutatore*




## 8. La mappa delle variabili

Abbiamo detto che l'analisi MFRM linearizza le dimensioni che concorrono al giudizio (le varie "facce"), ovvero istituisce una scala comune, in modo che si possano comparare le une alle altre.

Il risultato di questa operazione è una mappa chiamata "mappa delle variabili" ("Variable map", o anche "Vertical ruler" o "Wright map"). Essa è paragonabile a una scala Celsius: è centrata sullo zero e presenta valori, chiamati *logit* (contrazione di *log-odd units*), che possono essere positivi o negativi. Se la misura in *logit* del grado di generosità di un valutatore si colloca in prossimità dello zero, l'apporto del valutatore alla formulazione del voto è pressoché nullo; alla pari, se il valore in *logit* di un compito si colloca in prossimità dello zero, significa che esso è perfettamente calibrato in rapporto alla competenza del candidato; se, infine, la competenza di un candidato si colloca in prossimità dello zero, significa che essa coincide con la competenza media dell'intero campione.

*Il logit* è ottenuto mediante il logaritmo del rapporto tra due probabilità: la probabilità di un candidato di ricevere un certo giudizio da un certo valutatore in riferimento a una determinata componente del costrutto in rapporto alla probabilità di ricevere un voto immediatamente inferiore [$P_{nmijk}/P_{nmij}(k-1)$]. La formula è sviluppata in questo modo:

$$\log(P_{nmijk}/P_{nmij(k-1)})=B_n-A_m-D_i-C_j-F_k$$

dove

$P_{nmijk}$ = la probabilità che un candidato *n*, nel compito *m*, sia valutato nella componente *i* dal valutatore *j* e riceva il voto *k*

$P_{nmij(k-1)}$= la probabilità che lo stesso candidato *n*, nel compito *m*, valutato nella componente *i* dal valutatore *j*, riceva il voto immediatamente inferiore a *k* (*k-1*)

I fattori a destra dell'equazione, $B_n$, $A_m$, $D_i$, $C_j$, $F_k$., rappresentano rispettivamente

$B_n$= l'abilità del candidato *n*

$A_m$= il grado di difficoltà del compito *m*

$D_i$= il grado di difficoltà della componente specifica del costrutto oggetto di analisi (es. grammatica)

$C_j$= il grado di severità del valutatore

$F_k$= la difficoltà di ricevere il voto *k* anziché un voto inferiore (*k-1*)[18]

[18] $F_k$ può anche essere interpretato come la soglia in cui due voti contigui hanno pari probabilità (50%) di essere osservati. Eckes, il quale usa una simbologia diversa dalla nostra ($\tau_k$ anziché $F_k$), scrive (2015: 27): "Specifically, the threshold parameter, or category coefficient, $\tau_k$, is the location on the latent dimension where the adjacent categories, k and k-1, are equally probable to be observed. In other words, $\tau_k$ represents the transition point at which the probability is 50% of an examinee responding in one of two adjacent categories, given that the examinee is in one of those two categories. These transition points are called *Rasch-Andrich thresholds*". Per approfondimenti, cfr. Bond, Fox 2007.

# SEZIONE C

# LA VALUTAZIONE DELL'AFFIDABILITÀ DEI VALUTATORI DELLA PROVA SCRITTA DELLA CERTIFICAZIONE PLIDA

A seguire illustriamo i risultati dell'analisi MFRM applicata ai giudizi dei valutatori dello scritto della *Certificazione PLIDA*. Descriviamo dapprima la *Certificazione PLIDA Nuovo Formato* (§ 9), quindi gli elementi-chiave della ricerca: il gruppo dei valutatori (§ 10.1.1), i compiti (§ 10.1.2), la griglia di valutazione (§ 10.1.3), i risultati dell'indagine (§ 10.2).

## 9. La Certificazione PLIDA Nuovo Formato

Il *Progetto Lingua Italiana PLIDA* rappresenta una tra le azioni mediante le quali la Società Dante Alighieri promuove la lingua e la cultura italiana nel mondo. Sorto agli inizi degli anni 2000, il Progetto dedica una particolare attenzione alla valutazione della competenza comunicativa degli apprendenti di italiano L2/LS, nonché alla formazione degli insegnanti.

Emanazione del Progetto è l'omonima Certificazione, la quale figura tra le quattro certificazioni di lingua italiana riconosciute dal Ministero degli Affari Esteri, riunite nel marchio CLIQ (*Consorzio Lingua Italiana di Qualità*). Alla luce di osservazioni inoltrate dai centri d'esame alla Sede Centrale e in sintonia con la rivisitazione dei *format* di esami di competenza europei, la *Certificazione PLIDA* è stata soggetta a una riforma nel 2014, che ha interessato finora i livelli A2, B1, B2, C1. Le prove sono state riformulate e i criteri di valutazione della *performance* sono stati ricalibrati. L'intera operazione è riassunta in un documento, il *Nuovo Sillabo della Certificazione PLIDA*, a cui si aggiungono i *Quaderni delle Specifiche* (www.plida.it); volumi di simulazione delle prove sono stati pubblicati presso Alma Edizioni.

Una tra le caratteristiche della *Certificazione PLIDA* è il fatto di non contemplare batterie di esercizi per valutare la conoscenza delle componenti linguistiche. Non abbiamo, in altre parole, prove dedicate a lessico e grammatica. La conoscenza linguistica viene desunta, indirettamente, attraverso gli esiti dei candidati nelle prove di produzione/interazione orale e di produzione scritta.

La valutazione della parte orale è svolta da Commissioni in loco; diversamente, la valutazione della parte scritta è realizzata centralmente da un gruppo di 7 valutatori. In entrambi i casi chi valuta la *performance linguistica* è tenuto a seguire un percorso di formazione, grazie al quale apprende sia la logica della costruzione dei *task* sia come usare le griglie analitiche.

## 10. L'indagine

La nostra indagine ha mirato ad accertare l'affidabilità dei sette valutatori deputati alla valutazione della prova scritta. Lo studio ha riguardato, in particolare, la valutazione dei compiti realizzati da candidati di livello B1, mediante la griglia omologa (qui riportata nell'**Appendice 1**).

La ricerca si inserisce nel più ampio progetto di accreditamento della *Certificazione PLIDA* presso l'ALTE (*Association of Language Teachers in Europe*), organizzazione europea la quale, tra le altre funzioni, garantisce standard di qualità. In vista di un *audit* che accerti la qualità della Certificazione da parte di ALTE, il gruppo di ricerca PLIDA ci ha coinvolto nella realizzazione dell'indagine suddetta. Lo studio è stato realizzato nel primo semestre del 2021, all'interno di un progetto di ricerca di post-dottorato.

### 10.1. Elementi oggetto di indagine

Di seguito illustriamo gli elementi oggetto di analisi:

- *il gruppo dei valutatori* (§ 10.1.1.);
- *i compiti* (§ 10.1.2.);
- *la griglia di valutazione* (§ 10.1.3.).

#### 10.1.1. Il gruppo dei valutatori

Il gruppo di valutatori si compone di sette professionisti: [**H**], [**J**], [**K**], [**W**], [**X**], [**Y**], [**Z**]. Ciascuno di essi vanta un'esperienza pluriennale, sostenuta da aggiornamento da parte del gruppo di ricerca PLIDA.

#### 10.1.2. I compiti

Onde disporre di un campione rappresentativo dell'abilità di scrittura di ciascun candidato (e, quindi, al fine di coprire il costrutto in maniera relativamente estesa), la *Certificazione PLIDA Nuovo Formato* al livello B1 prevede la realizzazione di due compiti, ciascuno dei quali corrisponde ad un *format* ben preciso.

Nella tabella 4, *infra*, illustriamo le caratteristiche dei due *format* (attinte dalle *Specifiche*, accessibili online: Bariviera *et al.* 2021). Il *format* B, come si vede, è più breve ed agevole. Il tempo concesso al candidato per l'esecuzione dei due compiti è di 60 minuti.

Tabella 4. *Format A e Format B nella parte scritta della Certificazione PLIDA Nuovo Formato, livello B1: caratteristiche*

| ***format*** | ***range* di parole** | ***genere*** | **struttura del *prompt* e materiale d'appoggio** | **svolgimento del *task*** | **competenze attese** |
|---|---|---|---|---|---|
| A | 110-150 | Scrittura di un testo di tipo informativo/ narrativo/ descrittivo | Elenco puntato funzionale<br><br>Testi e/o immagini vengono forniti come stimolo | Sulla base delle informazioni contenute in un input scritto (un breve testo informativo o descrittivo su turismo, tempo libero, servizi; e-mail di un amico) o di un input visivo, il candidato scrive un testo sviluppando una serie di funzioni indicate in una scaletta (per es. informare, consigliare, valutare, dare la propria opinione su qualcosa, ecc.) | Trasmettere per iscritto informazioni e idee su argomenti sia astratti sia concreti, verificare le informazioni ricevute, porre domande su un problema o spiegarlo con ragionevole precisione |
| B | 70-100 | Scrittura di un testo di tipo narrativo | Elenco puntato funzionale | Sulla base di un input scritto oppure usando come | Scrivere resoconti di esperienze; descrivere |

| | | | | | |
|---|---|---|---|---|---|
| | | | Testi e/o immagini vengono forniti come stimolo | risorsa le proprie esperienze personali, il candidato scrive un testo sviluppando una serie di funzioni indicate in una scaletta (per es. raccontare, descrivere esperienze, spiegare opinioni, ecc.) | sentimenti e impressioni |

Il Certificatore ha selezionato una rosa di 60 composizioni tra gli elaborati prodotti da tutti i candidati in tre distinte Sessioni PLIDA (giugno 2016, agosto 2016, novembre 2017). Si tratta delle composizioni realizzate da 30 candidati: 30 di esse fanno capo al *format A* e 30 fanno capo al *format B*. Più in particolare, il Certificatore ha raccolto gli elaborati prodotti da 10 candidati per ciascuna Sessione, prestando attenzione affinché ciascun campione fosse rappresentativo di diversi livelli di abilità (cfr. tabella 5, pagina a seguire).

Tabella 5. *Format, compiti, gruppi, sessioni*

| **format** | **compiti** | **candidati** | **gruppi** | **Sessioni PLIDA** |
|---|---|---|---|---|
| A | A | 1-10 | I | giugno 2016 |
| B | B | 1-10 | I | giugno 2016 |
| A | C | 11-20 | II | agosto 2016 |
| B | D | 11-20 | II | agosto 2016 |
| A | E | 21-30 | III | novembre 2017 |
| B | F | 21-30 | III | novembre 2017 |

Successivamente, le 60 composizioni sono state valutate separatamente da ciascun valutatore, al quale è stato chiesto di far fede ai descrittori della griglia analitica prevista per il livello B1 (cfr. **Appendice 1**). I voti ci sono stati infine trasmessi.

In un primo calcolo abbiamo considerato tra loro *omogenei* i compiti facenti parte dello stesso *format*.

Ci siamo resi conto però che la nostra decisione avrebbe potuto sollevare dei dubbi, poiché presuppone l'equivalenza, in termini di difficoltà, dei compiti relativi a ciascun *format*. Lo stesso manuale di Facets® (Linacre 2023: 291) avvisa infatti che

> "When one facet is nested within another facet (without anchoring or group-anchoring), Facets makes an arbitrary allocation of statistical information between the facets, this can lead to unstable estimation"

Nel nostro caso, i *compiti* sono "nidificati" (*nested*) all'interno della "faccia" del *format*.

Tra le soluzioni proposte dal manuale per far fronte a una situazione del genere vi è l'ancoraggio della variabile in questione allo zero *logit* (Linacre 2023: 353-354).

Nell'elaborare le istruzioni da trametttere al *software* una seconda volta, abbiamo dunque fissato allo zero *logit* il grado di difficoltà di ciascun compito. Questo ancoraggio permette di isolare il comportamento dei valutatori al netto dell'influenza della difficoltà del compito (difficoltà non facilmente oggettivabile, ripetiamo, posto che compiti distinti sono stati assegnati a gruppi distinti).[19]

### 10.1.3. La griglia

Per formulare il loro giudizio, i valutatori hanno fatto uso della griglia di valutazione della produzione scritta del livello B1 della *Certificazione PLIDA Nuovo formato*, presentata in Bariviera *et al.* 2021 (cfr. **Appendice 1**). La griglia scompatta il costrutto dell'abilità di scrittura in quattro componenti:

- *contenuto ed efficacia del testo* (d'ora in poi, per ragioni di brevità, *contenuto*),
- *coesione e coerenza*,
- *lessico*,
- *accuratezza morfologica e ortografia* (d'ora in poi, per ragioni di brevità, *grammatica*).

La griglia si compone di una scala distinta in 11 livelli (0÷10). I livelli sono raggruppati in sei fasce, a ciascuna delle quali corrisponde, componente per componente, un descrittore specifico. Possiamo definire le fasce mediante le seguenti etichette:

- fascia 0: non valutabilità/comprensibilità dell'elaborato;
- fascia |1-2|: *performance* scarsa;
- fascia |3-4|: *performance* insufficiente;
- fascia |5-6|: *performance* sufficiente;[20]
- fascia |7-8|: *performance* buona;

[19] A seguito del secondo calcolo, non abbiamo comunque riscontrato significative variazioni dei risultati.

[20] Il punto di taglio (*cut score*), ovvero il punteggio a cui corrisponde la sufficienza, è 5.

- fascia |9-10|: *performance* ottima.

## 10.2. I risultati

L'analisi MFRM ha avuto per obiettivo la rilevazione del grado di affidabilità dei valutatori e l'accertamento della pertinenza della griglia di valutazione della produzione scritta di livello B1.

Più in dettaglio, l'analisi si è focalizzata sui seguenti fattori:

- *l'adeguatezza al modello Rasch dei dati raccolti* (§ 10.2.1);
- *il grado di generosità del valutatore* (§ 10.2.2);
- *il grado di coerenza del valutatore* (cioè la stabilità del *pattern* di valutazione; *intra-rater reliability*) (§ 10.2.3);
- *la capacità del valutatore di discriminare i candidati più abili dai meno abili* (§ 10.2.4);
- *il grado di omogeneità del gruppo di valutatori* (§ 10.2.5);
- *l'efficacia della griglia di valutazione* (§ 10.2.6);
- *bias specifici* (§ 10.2.7).

Nei paragrafi che seguono presentiamo le indagini separatamente, riportando ciascuna in una nuova pagina.

### 10.2.1. L'adeguatezza al modello Rasch dei dati raccolti

La prima indagine ha riguardato la *valutazione globale di adeguatezza dei dati raccolti in riferimento al modello Rasch*. Scrive Eckes (2009: 27):

> "Generally speaking Rash models are idealization of empirical observation. Therefore, empirical data will never fit a given Rasch model perfectly. In other words, with a sufficient large sample of data, any model can be shown to be false (Lord, Novick 1968). The really interesting question concerns the practical utility of a model; that is, we need to know whether the data fit the model usefully, and, when misfit is found, how much misfit there is and where it comes from […]. One way to assess overall data-fit is to examine responses that are unexpected given the assumptions of the model […]. According to Linacre 2008, satisfactory model fit is indicated when about 5% or less of (absolute standard residuals are ≥2, and about 1% or less of (absolute) standardized residuals are ≥3".

Nel nostro caso, il numero totale di risposte valide è pari a 1680; le risposte con residui standardizzati (i residui divisi per la deviazione standard attesa dal modello) ≥2 sono 70 (4.16%); quelle con residui standardizzati ≥3 sono 12 (7 per mille). In termini di generale adeguatezza, il campione si conforma al modello Rasch.

### 10.2.2. Il grado di generosità del valutatore

*La severità/generosità del valutatore* (ovvero il *grado di disallineamento positivo/negativo*) può essere apprezzata ad occhio nudo mediante la *mappa delle variabili*. Grazie ad essa si coglie la distribuzione dei valori degli elementi di ciascuna variabile in riferimento alla scala *logit* (figura 2, *infra*; la scala *logit* è rappresentata nella prima colonna).

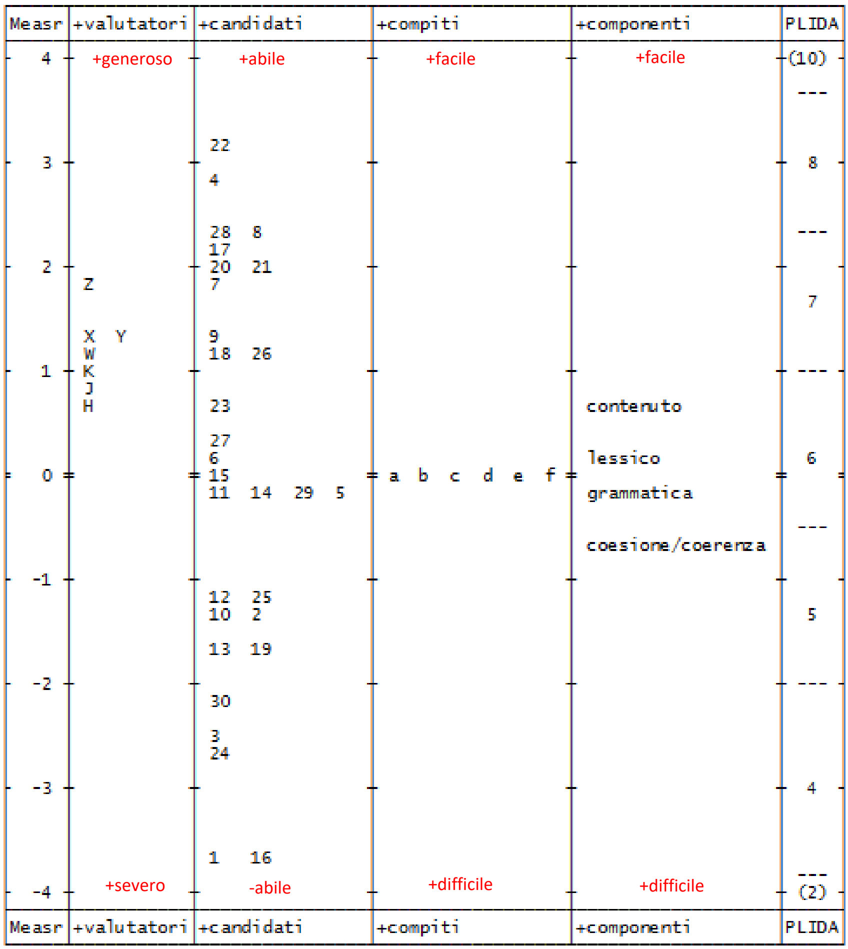
Nell'elaborare la mappa tutte le variabili vengono *centrate* sul valore medio della scala (0 *logit*), eccetto quella oggetto di studio (libera così di "fluire" lungo la scala).[21] Nel nostro caso, poiché l'analisi concerne l'affidabilità dei valutatori, l'unica variabile non centrata è quella del valutatore (seconda colonna). Ricordiamo peraltro che, a causa dello specifico *rating design* (con la distribuzione non uniforme dei compiti ai gruppi), abbiamo ancorato (*anchoring*) allo zero *logit* il valore dei compiti (quarta colonna).

A seconda del posizionamento sulla scala del singolo valutatore (e quindi a seconda del *logit* corrispondente), possiamo evincere il grado di severità: più il valore si avvicina allo zero *logit*, più l'effetto generosità/severità è contenuto; viceversa, più dista dallo zero, in senso positivo o negativo (a dipendere, da com'è orientata la variabile), più abbiamo un'accentuata generosità o un'accentuata severità.

Come si evince dalla mappa (seconda colonna), nel nostro caso tutti i valutatori sono *generosi*: si posizionano ben al di sopra dello zero *logit*.

[21] Vanno redatte mappe diverse, dunque, in base agli oggetti di indagine.

Figura 2. *Mappa delle variabili*




Prima di accedere ai valori *logit* che determinano il posizionamento dei valutatori nella mappa (tab. 6, *infra*), illustriamo altre voci che compaiono in essa.

Come detto, la prima e la seconda colonna rappresentano rispettivamente la scala *logit* e il *grado di generosità/severità dei valutatori*; le successive quattro rappresentano rispettivamente *l'abilità dei candidati; il grado di facilità/difficoltà dei compiti* (nel nostro caso, per via del *rating design* specifico, abbiamo ancorato il valore *logit* dei compiti allo zero), *il grado di facilità/difficoltà delle componenti dell'abilità di scrittura*, riportate nella griglia (*contenuto; lessico, grammatica; coesione e coerenza*) e infine *gli intervalli tra i punteggi della griglia*, così come sono stati usati globalmente dal gruppo dei valutatori.[22]

Al momento di impartire le istruzioni al programma[23] abbiamo deciso di orientare tutte le variabili positivamente: sono collocati, quindi, nella fascia più alta, rispettivamente, i valutatori più generosi, i candidati più abili e la componente più facile (fanno eccezioni i compiti che, torniamo a ripetere, sono stati allineati allo *zero*).

Oltre il *disallineamento positivo* da parte dell'intero gruppo, dalla mappa si evincono un paio di fenomeni:

- un'ampia distribuzione dei candidati (di circa 7 *logit*);
- il *contenuto* è la componente più facile, mentre la *coesione/coerenza* è quella più difficile.

Il primo aspetto va interpretato in un senso positivo: il gruppo PLIDA ha scelto oculatamente i candidati, con *performance* distribuite in termini di qualità.

Il secondo aspetto (il diverso grado di difficoltà delle componenti) merita di essere indagato (lo faremo più avanti, nel § 10.2.7.1.1).

---

[22] I segmenti tratteggiati indicano gli intervalli di mezzo punto della scala decimale (*Rasch half-point thresholds*). Per esempio, il segmento tratteggiato tra il 4 e il 5 corrisponde al punteggio 4.5 (se tracciamo, a partire dallo stesso segmento, una linea orizzontale vedremo come ad essa corrisponda un valore pari a –2 *logit*; ciò significa che il punteggio atteso di un ipotetico candidato la cui abilità sia di –2 *logit* è pari a 4.5/10). Salendo o scendendo rispetto al segmento di mezzo ci si orienta verso i numeri interi (nel nostro caso, rispettivamente, 5 e 4).

[23] Qualora il lettore volesse accedere alle istruzioni trasmesse al programma, lo rimandiamo all'**Appendice 2**.

Riprendiamo la questione del posizionamento dei valutatori nella mappa. Nella tabella 6, a seguire, sono evidenziati in rosso i valori *logit* di ciascuno di essi.

I dati si possono leggere in tre blocchi:

- primo blocco: *le medie*
- secondo blocco: *i valori logit*
- terzo blocco: *l'errore di misurazione*

Tabella 6. *Medie e valori logit relativi al grado di generosità dei valutatori*

Legenda
**MO=** media osservata
**MC=** media corretta secondo il modello Rasch
**SPMC=** scarto progressivo delle medie corrette
**L(v)=** conversione della MC in *logit*
**SPL=** scarto progressivo in *logit*
**ES=** errore standard di misurazione[24]

Primo blocco | Secondo blocco | Terzo

| valutatori | MO | MC | SPMC | L(v) | SPL | ES |
|---|---|---|---|---|---|---|
| H | 6.3 | 6.26 | 0.00 | 0.62 | 0.00 | 0.08 |
| J | 6.5 | 6.44 | 0.18 | 0.87 | 0.25 | 0.08 |
| K | 6.6 | 6.55 | 0.29 | 1.01 | 0.40 | 0.08 |
| W | 6.7 | 6.64 | 0.38 | 1.13 | 0.52 | 0.08 |
| X | 6.8 | 6.75 | 0.49 | 1.28 | 0.66 | 0.08 |
| Y | 6.8 | 6.77 | 0.51 | 1.31 | 0.69 | 0.08 |
| Z | 7.2 | 7.17 | 0.91 | 1.84 | 1.22 | 0.08 |

[24] L'indice è sensibile all'entità dei dati raccolti. Nel nostro caso i valori relativi all'errore standard di misurazione sono relativamente contenuti, considerato il numero esteso di composizioni valutate da ciascun valutatore.

*Primo blocco: le medie*

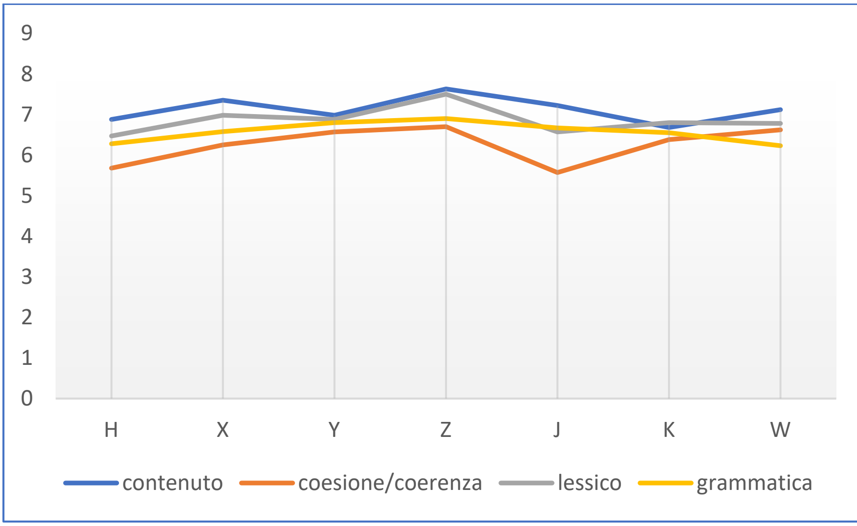
Le medie calcolate dal sistema sono due: quella *osservata* (MO) e quella *corretta secondo il modello Rasch* (MC).

La *media osservata* (MO) è data dalla somma dei voti assegnati divisa per il numero di occorrenze. Qui sotto, alla tabella 7 e quindi alla fig. 3 ad essa relativa, abbiamo rappresentato, più in dettaglio, le *medie osservate* di ciascun valutatore in riferimento alle singole componenti.

Tabella 7. *Medie osservate in rapporto alle componenti*

| **valutatori** | **contenuto** | **coesione/ coerenza** | **lessico** | **grammatica** |
|---|---|---|---|---|
| H | 6.88 | 5.68 | 6.47 | 6.28 |
| X | 7.35 | 6.25 | 6.98 | 6.58 |
| Y | 6.98 | 6.57 | 6.88 | 6.80 |
| Z | 7.63 | 6.70 | 7.50 | 6.90 |
| J | 7.22 | 5.57 | 6.57 | 6.67 |
| K | 6.68 | 6.38 | 6.80 | 6.55 |
| W | 7.12 | 6.62 | 6.78 | 6.23 |

Figura 3. *Medie osservate in rapporto alle componenti*




Ritornando alla tab. 6, *supra*, la seconda voce rappresenta la *media corretta secondo il Modello Rasch* (MC). In merito ad essa, Erguvan e Aksu Dunya scrivono (2020: 10):

> "«fair average» […] is the average rating for each instructor adjusted for the deviation of the instructors in that instructor's sample from the overall student mean"

Detto altrimenti, MC è la media attribuibile al valutatore se, in teoria, scorporassimo dalla *media osservata* l'incidenza dell'abilità dei candidati facenti parte dello studio: insomma, è il dato "puro" che indica effettivamente quale sarebbe il valore medio attribuibile al valutatore, al netto dall'apporto dato dalla competenza di *questo* campione di candidati.

Al fine di comparare i valutatori tra loro è utile ragionare sugli *scarti progressivi tra le medie corrette rispetto al valore minimo* (SPMC). Lo *scarto relativo alla media corretta* tra il valutatore meno generoso del gruppo [**H**] e quello più generoso [**Z**] è pari a 0.91: ciò significa che, in generale, un candidato si troverebbe ad avere quasi un punto di differenza (in riferimento alla scala della griglia di valutazione) nel caso il suo elaborato venga esaminato dall'uno o dall'altro valutatore.

*Secondo blocco: i valori logit*

Tutti i valutatori sono generosi: [**Z**] è il più generoso (+1.84 *logit*); [**H**] è quello meno (+0.62 *logit*). Come interpretare i valori *logit* di ciascuno?

Wilson e Engelhart (2000), ai quali fa eco Mendoza Ramos (2018), suggeriscono che il disallineamento di un valutatore è contenuto allorquando il suo valore *logit* rientra nell'intervallo |±1|. Laddove si riscontrano valori superiori all'unità, il grado di severità o di generosità di un valutatore desta, invece, preoccupazione.

Nel nostro caso, solo due valutatori presentano un valore al di sotto dell'unità: [**H**] e [**J**]; tutti gli altri (evidenziati con fondino giallo nella tab. 6, *supra*), compreso [**K**] il cui valore *logit* è appena al di sopra dell'unità, presentano valori che superano l'intervallo previsto da Wilson e Engelhart (2000).

Nella tabella 8, *infra*, abbiamo ipotizzato, a scopo puramente indicativo, quattro livelli di generosità:

- *entro i limiti,*
- *borderline,*
- *eccessiva,*
- *estrema.*[25]

[25] Il numero di livelli è statisticamente superiore; di fatto, come vedremo nel § 10.2.5, si danno ben sei livelli di severità.

Tabella 8. *Gruppi indicativi di generosità*

| **valut.** | **<1** generosità entro i limiti | **≈1** generosità *borderline* | **1÷1.50** generosità eccessiva | **>1.50** generosità estrema |
|---|---|---|---|---|
| H | 0.62 | | | |
| J | 0.87 | | | |
| K | | 1.01 | | |
| W | | | 1.13 | |
| X | | | 1.28 | |
| Y | | | 1.31 | |
| Z | | | | 1.84 |

A parere di Linacre (2020), qualora tra il grado di generosità/severità di due valutatori vi sia uno scarto di oltre mezzo punto, in termini di *logit*, possiamo parlare di giudizi effettivamente dissimili ("If the rater's differ by more than 0.5 score points, then they are noticeably different").

*Terzo blocco: l'errore standard di misurazione*

ES è *l'errore standard di misurazione*. È una dimensione insita in ogni indagine statistica, ed è tanto maggiore quanto minore è la quantità dei dati raccolti; viceversa, quanto maggiore è il numero dei dati a disposizione, tanto minore è l'errore.[26]

Nel nostro studio gli errori di misurazione sono contenuti (tab. 6, *supra*). Come vanno letti? Consideriamo il valutatore [**H**], con un *logit* pari a 0.62 (tab. 4, *supra*). Ebbene, la *misura effettiva*, al netto dell'errore di misurazione, oscilla tra 0.54 e 0.70 (±0.08; ± una deviazione standard dalla media), con una sicurezza del 68%; tra 0.46 e 0.78 (±0.08x2; ± due deviazioni standard dalla media), con una sicurezza del 95%.

[26] Per questa ragione, sia detto per inciso, al fine di compiere un'indagine sull'affidabilità dei valutatori che sia a sua volta affidabile, si raccomanda di disporre di un campione di almeno trenta candidati (Linacre 1994).

### 10.2.3. Il grado di coerenza del valutatore

L'analisi MFRM restituisce il grado di coerenza del valutatore con sé stesso ("intra-rater reliability") attraverso delle *statistiche di conformità* ("fit statistics"). Sulla base dei *residui standardizzati* (residui divisi per la deviazione standard attesa dal modello) è possibile calcolare due indici:

- l'*indice outfit* ("outlier-sensitive infit statistics"), sensibile ai valori che divergono dalla norma (*outliers*; come può essere, per esempio, un voto estremamente basso assegnato a un elaborato eccellente da parte di un valutatore generoso);[27]
- l'*indice infit* ("information weighted infit statistics"); è un valore ponderato (assegna un peso minore alle valutazioni estreme) e vale a rappresentare problemi sistematici; riflette, dunque, *pattern* di valutazione (Eckes 2005, 2015).[28]

I valori di riferimento degli indici variano in funzione dell'entità della prova (se la prova si rivolge a un bacino d'utenza molto ampio, i margini sono più stringenti). Nell'ambito dello

[27] Scrive Eckes (2015: 76): "Rater outfit is sensitive to "outlying" unexpected ratings ("outfit" is short for "outlier-sensitive fit statistic"). Outlying ratings refer to a situation where the latent variable locations of rater j and the locations of the other elements involved, such as examinees and criteria, are farther apart from one another (e.g., separated by more than 1.0 logits). Thus, when a lenient rater awards harsh ratings to a highly proficient examinee on a criterion of medium difficulty, this rater's outfit will increase".

[28] Scrive Eckes (2015: 76-77): "Rater infit is sensitive to "inlying" unexpected ratings. More specifically, infit is sensitive to unexpected ratings where the locations of rater j and the other elements involved are aligned with each other, that is, where the locations are closer together on the measurement scale (e.g., within a range of about 0.5 logits) [...]. Since such ratings are generally associated with higher estimation precision, infit is commonly considered more important than outfit in judging rater fit".

studio dell'affidabilità dei valutatori, Wright e Linacre (1994) suggeriscono di attenersi all'intervallo |0.60÷1.40|.

Tanto per l'*infit* quanto per l'*outfit*, valori inferiori alla soglia minima (<0.60) rivelano una *predittività* maggiore del previsto ("overfitting", *iperconformità* al modello); valori superiori alla soglia massima (>1.40) indicano, invece, un'*imprevedibilità* superiore al previsto ("outfitting", *iperdifformità* rispetto al modello). Il problema maggiore sta nel caso dell'*iperdifformità*, posto che, in questa circostanza, non abbiamo le basi per stabilire la presenza di un *pattern* vero e proprio [i giudizi del valutatore variano di molto rispetto alle aspettative del modello; ciò significa che nell'ipotesi si ripeta la stessa valutazione con lo stesso campione e con lo stesso strumento (griglia), non è detto che quel valutatore formuli gli stessi giudizi].

Nella tab. 9, *infra*, sono riportate le statistiche di conformità relative al gruppo di studio. Ricordiamo che ciascun valutatore (n=7) ha valutato 60 composizioni (due composizioni per ciascuno dei 30 candidati), su quattro categorie (*contenuto, coesione e coerenza, grammatica, lessico*), per un totale di 240 punteggi assegnati.

Tabella 9. *Statistiche di conformità dei valutatori*

Legenda

**L(v)=** valori *logit* dei valutatori, recuperati dalla tabella 4

**Infit=** valore ponderato (al netto dei valori *outliers*) relativo al *pattern* di giudizio (l'intervallo previsto è |0.6; 1.4|; valori inferiori alla soglia minima indicano un *pattern* rigido, con eventuale presenza di *bias*; valori superiori alla soglia massima indicano un *pattern* erratico)

**i-zstd=** punteggio zeta, indica la precisione statistica delle misurazioni *infit* (l'intervallo previsto è |±2|; valori interni all'intervallo indicano una occasionalità del valore *infit*; valori esterni indicano una significatività statistica)[29]

**Outfit=** valore non ponderato relativo al *pattern* di giudizio, sensibile ai comportamenti *outliers* (l'intervallo di riferimento è|0.6; 1.4|; valori inferiori alla soglia minima indicano una predittività superiore alle attese; valori superiori alla soglia rivelano un'aleatorietà dei giudizi)

**o-zstd=** punteggio zeta che indica la precisione statistica dei misurazioni *outfit* (come sopra, in riferimento a i-zstd)

| **valutatori** | **L(v)** | **Infit** | **i-ZSTD** | **Outfit** | **o-ZSTD** |
|---|---|---|---|---|---|
| H | 0.62 | 0.90 | -.1.00 | 0.93 | -0.70 |
| J | 0.87 | 1.11 | 1.10 | 1.12 | 1.20 |
| K | 1.02 | 0.81 | -2.10 | 0.82 | -2.10 |
| W | 1.13 | 0.80 | -2.40 | 0.80 | -2.30 |
| X | 1.28 | 0.89 | -1.20 | 0.89 | -1.20 |
| Y | 1.31 | 1.03 | 0.30 | 1.03 | 0.30 |
| Z | 1.84 | 1.41 | 4.10 | 1.41 | 4.00 |

[29] Valore ottenuto mediante la divisione tra lo scarto del punteggio del candidato rispetto alla media e la deviazione standard. Può essere positivo o negativo a dipendere dal fatto che il punteggio di partenza sia superiore o inferiore alla media. Va rilevato comunque che, ai fini dell'analisi, il *punteggio zeta* va letto contestualmente agli indici di *infit* o *outfit*: se questi rientrano nella norma, in genere di eventuali valori eccedenti del *punteggio zeta* non si tiene conto.

Meritano una particolare attenzione gli indici *infit* e *outfit* (e relativi punteggi zeta) del valutatore [**Z**]. Il valutatore presenta valori *borderline* (*infit*: 1,41; *outfit*: 1.41), con punteggi zeta oltre l'intervallo (4.10; 4.00). Poiché positivi, i valori *infit* e *outfit* confermano il *trend* di eccesso positivo del giudizio. La prossimità alla soglia superiore di tolleranza (1.4) ci informa di aspetti di *impredicibilità*: ciò vale sia in riferimento ad alcune divergenze estreme rispetto ai valori attesi (*outliers*), come dimostrato dall'indice *outfit*, sia in riferimento ad alcune divergenze ponderate, al netto delle valutazioni estreme, come dimostrato dall'indice *infit* (il quale riflette, abbiamo detto, la sistematicità di un certo comportamento).

In precedenza avevamo affermato che il valutatore [**Z**] è il più generoso. La prossimità degli indici di conformità alla soglia superiore di tolleranza ci impone di aggiungere che, oltre all'eccesso di generosità, si ravvisa una *erraticità*. Ciò significa che, benché vi sia una tendenza a un giudizio eccessivamente positivo, si registrano delle oscillazioni; non è escluso, pertanto, che nell'ipotesi si ripeta la misurazione il valutatore formuli giudizi dissimili.

### 10.2.4. Il grado di discriminazione del valutatore tra i candidati più abili e quelli meno abili

La correlazione tra i giudizi emessi dal singolo valutatore e quelli emessi dagli altri valutatori indica se l'intero gruppo procede nella stessa direzione nell'assegnare i punteggi massimi ai candidati più abili e i punteggi minimi ai candidati meno abili. Come in ogni correlazione, il valore oscilla all'interno dell'intervallo |±1|. Più il valore si avvicina a 1, maggiore è la convergenza nella discriminazione; più si avvicina allo zero, minore è la convergenza; qualora si registri un valore negativo abbiamo, invece, un'interpretazione opposta del costrutto di riferimento da parte del valutatore rispetto a quella del gruppo.

Nel nostro caso, il comportamento dei valutatori, in termini di discriminatività, si presenta abbastanza compatto. Le correlazioni tra il giudizio del singolo valutatore (V) e quello degli altri (VV) oscillano nell'intervallo |0.82÷0.88| (cfr. tab. 10, *infra*). Si tratta di un valore alto: il gruppo lavora all'unisono nel distinguere i candidati più abili da quelli meno abili.[30]

[30] I valori osservati (V/VV-O) sono, peraltro, prossimi ai valori attesi (V/VV-A).

Tabella 10. *Correlazioni tra le valutazioni*

Legenda

**V/VV-O=** correlazione *osservata* tra il giudizio del singolo valutatore e quello degli altri valutatori

**V/VV-A=** correlazione *attesa* tra il giudizio del singolo valutatore e quello degli altri valutatori

| **valutatori** | **V/VV-O** | **V/VV-A** |
|---|---|---|
| H | 0.83 | 0.86 |
| J | 0.86 | 0.85 |
| K | 0.86 | 0.85 |
| W | 0.87 | 0.85 |
| X | 0.86 | 0.85 |
| Y | 0.88 | 0.85 |
| Z | 0.82 | 0.85 |

### 10.2.5. Il grado di omogeneità del gruppo di valutatori

Ai fini della validità di una prova, è importante che al compito prodotto da un candidato sia assegnato un voto pressoché identico da valutatori diversi. Ciò significa che, da parte di un'istituzione che eroga esami di competenza, è importante accertarsi che i valutatori operino con una certa sintonia.

Va in ogni caso precisato che l'omogeneità attesa non è assoluta. Se ciò si verificasse, si potrebbe supporre sia dovuto a una forzatura dei giudizi (i valutatori tendono a conformarsi a una norma comune, senza più apportare un contributo alla misurazione; l'affidabilità estrema, in questo caso, rischia di imporsi sulla validità del giudizio). L'ipotesi del modello Rasch è, infatti, pur sempre quella secondo la quale i valutatori agiscono come degli esperti, con un loro grado di indipendenza, e non in maniera meccanica, come se fossero delle "rating machine", nel qual caso cesserebbero di costituire una variabile (cfr. Linacre 1998). Diremmo, pertanto, come argomentato in precedenza, che il modello si attende un accordo sostenuto, ma non assoluto.

Il programma Facets® calcola il grado di omogeneità di un gruppo di valutatori attraverso diversi indici:

a) *la percentuale di convergenza* (osservata e attesa) *del giudizio del valutatore rispetto al giudizio degli altri*;
b) *la percentuale di convergenza* (osservata e attesa) *dei giudizi dei valutatori, presi nella loro totalità*;
c) *l'indice di separazione* ("rater separation ratio"), che informa di quante volte il grado di severità del gruppo di valutatori dista dalla precisione delle loro misurazioni;
d) *l'indice di stratificazione* ("number of strata index"), che informa quanti insiemi omogenei, in termini di severità, si rinvengono nel campione;
e) *la verifica dell'ipotesi nulla* (ovvero dell'ipotesi che vi sia una uniformità assoluta dei giudizi da parte del gruppo di valutatori);

f) *l'indice di affidabilità* ("reliability of rater separation index"), che restituisce il grado di compattezza dei giudizi.

In più, il ricercatore può calcolare per suo conto

g) *l'indice Rasch-Kappa*
h) *l'indice di distribuzione dei giudizi*

A seguire descriviamo ciascuno di essi.[31]

(a-b) *Le percentuali di convergenza*

In termini ideali, *le percentuali di convergenza*, sia del singolo rispetto al gruppo che dei membri del gruppo tra loro, sono pari al 100%. Cioè, in una situazione di perfetta omogeneità, i voti di Tizio, di Caio e di Sempronio coincidono; che il candidato sia valutato dall'uno o dall'altro non genera, dunque, alcuna differenza.

Di fatto, come appena precisato, i margini di soggettività in merito alla valutazione della competenza sono ineliminabili. Il modello Rasch considera, in effetti, come fisiologiche leggere differenze di giudizio.[32] È vero, però, che laddove i giudizi presentino percentuali di convergenza molto basse (nell'ordine del 20%) si ha un *eccesso di eterogeneità*: è assai probabile, in questo caso, che il costrutto relativo alla *performance* non sia stato operativizzato allo stesso modo (la griglia di valutazione viene interpretata in modo dissimile dai *raters*).

---

[31] Tutti questi indici, giacché calcolati a partire dagli stessi dati, ci forniscono una stessa informazione; sono perciò ridondanti. Nell'atto di presentare i risultati della propria rilevazione, non è necessario che il ricercatore compili l'intero elenco. Tuttavia, per motivi di esaustività, in questa sede diamo accenno a ciascuno di essi.

[32] Come avviene con ogni *performance* (artistica, sportiva, ecc.), anche la *performance* linguistica può dar adito ad apprezzamenti distinti, a seconda delle sfumature/degli aspetti della competenza colti ora dall'uno ora dall'altro valutatore.

La tabella 11, *infra*, ci informa delle percentuali di accordo (*osservato* e *atteso*) tra i giudizi di ciascun valutatore e i giudizi del gruppo: si tratta di valori relativamente bassi. Ciò lascia supporre che il gruppo sia caratterizzato da una disomogeneità accentuata.

Tabella 11. *Percentuali di accordo tra i giudizi*

Legenda

**% ACC-O=** percentuale di accordo *osservato* tra il giudizio del singolo valutatore e quello degli altri valutatori

**% ACC-A=** percentuale di accordo *atteso* tra il giudizio del singolo valutatore e quello degli altri valutatori

| **valutatori** | **% ACC-O** | **% ACC-A** |
|---|---|---|
| H | 36.2 | 31.6 |
| J | 38.5 | 32.8 |
| K | 39.5 | 33.2 |
| W | 36.9 | 33.3 |
| X | 37.3 | 33.2 |
| Y | 33.0 | 33.1 |
| Z | 30.1 | 30.2 |

I valori sono riassumibili in una cifra complessiva di accordo osservato pari al 35.9%: si tratta di un valore prossimo al valore atteso (in riferimento a *questo* gruppo di valutatori), che è pari al 32.5%.

(c-d) *L'indice di separazione* e *l'indice di stratificazione*

Il valore ideale dell'*indice di separazione* e dell'*indice di stratificazione* (che dal primo deriva) è pari a 1. Più il valore dista dall'unità, maggiore è l'eterogeneità del gruppo.[33]

[33] Nel caso di estrema eterogeneità, può capitare addirittura che ciascuno dei due indici sia superiore al numero dei valutatori.

Nel nostro caso, *l'indice di separazione* è pari 4.52: l'intervallo di severità è superiore più di 4 volte al grado di precisione da parte del gruppo. Si tratta di una conferma della disomogeneità.

*L'indice di stratificazione* è pari a 6.36: ci informa che si danno almeno 6 livelli di severità nel campione (in sostanza, tutti divergono dagli altri, eccetto [**X**] e [**Y**] che sono allineati tra loro, con una differenza minima di 0.03 *logit*; cfr. la mappa delle variabili, fig. 2, *supra;* tab. 6, *supra*).

(e) *La verifica dell'ipotesi nulla*

L'*ipotesi nulla* viene sconfermata con un indice di probabilità $p<0.05$: in tal caso, si evince la non casualità relativa all'eterogeneità interna al campione di riferimento (in sostanza, si confuta l'idea che la differenza tra i valutatori sia nulla).

Nel nostro caso, *l'ipotesi nulla* è sconfermata ($p<0.05$): si danno differenze di giudizio tra i valutatori.

(f) *L'indice di affidabilità*

L'*indice di affidabilità* è calcolato in maniera simile all'Alfa di Cronbach (applicato nella statistica classica)[34]: valori prossimi allo zero indicano omogeneità; valori prossimi a 1 indicano disomogeneità.

Nel nostro caso, *l'indice di affidabilità* è pari 0.95: si evidenzia, appunto, una significativa divergenza tra i giudizi.

---

[34] L'*Alfa di Cronbach* ci restituisce il grado di omogeneità tra gli *item* ed è calcolata sulla base delle *varianze* (è data dal rapporto tra la varianza data dai singoli *item* e la varianza complessiva dell'esercitazione). La *varianza* è un *indice di distribuzione* ottenuto mediante la divisione della somma del quadrato delle distanze dalla media dei singoli punteggi con il numero delle occorrenze meno un'unità [$\Sigma d^2/(n-1)$].

(g) *L'indice Rasch-Kappa*

L'*indice Rash-Kappa* è simile all'indice Kappa di Cohen, che stima l'affidabilità di una classificazione statistica (cfr. Cohen 1960; Linacre 2014). La formula dell'*indice Rash-Kappa* è la seguente:

Rasch-Kappa= (% di accordo osservato – % di accordo atteso) / (100 – % di accordo atteso).

Laddove il valore si discosti dallo zero, in senso positivo o negativo, si ravvisa una divergenza tra i valutatori (cfr. Eckes 2015: 92).

Nel nostro caso:

Rasch-Kappa= (35.9 – 32.5) / (100 – 32.5).

Rasch-Kappa= 3.4 / 67.5

Rasch-Kappa= 0.050

L'*indice Rasch-Kappa* rivela indipendenza di giudizio.

(h) *L'indice di distribuzione dei giudizi*

In merito alla distribuzione dei giudizi, Eckes (2012: 280) segnala che qualora la somma assoluta del *logit* relativo al valutatore più severo e del *logit* relativo a quello meno severo sia superiore alla quarta parte della somma assoluta dei *logit* relativi ai candidati estremi (più abile e meno abile), abbiamo una discrepanza tra i giudizi dei valutatori.

Nel nostro caso:

- Somma assoluta dei *logit* dei valutatori estremi: 0.62+1.84= 2.46
- Somma assoluta dei *logit* dei candidati estremi: 3.21+3.69= 6.90
- Quarta parte della somma assoluta dei *logit* dei candidati estremi: 6.90/4= 1.725

Poiché 2.46 (somma assoluta dei *logit* dei valutatori estremi) >1.725 (quarta parte della somma assoluta dei *logit* dei candidati estremi), si deduce che vi è una dispersione tra i giudizi.

Ricapitolando, i vari indici confermano la disomogeneità interna al gruppo.

### 10.2.6. La qualità della griglia di valutazione

Il programma Facets® consente di analizzare come i livelli di competenza definiti dagli intervalli della scala di valutazione (cioè i singoli punteggi) sono stati impiegati dal gruppo di valutatori. Le *research questions* sono le seguenti: la scala in adozione (nel nostro caso|0÷10|) è efficace per rilevare la competenza oggetto di studio? I descrittori relativi alle componenti riflettono l'evoluzione del costrutto? Gli intervalli sono pochi? Sono eccessivi?

Il programma assolve alle seguenti funzioni:

- computa il numero di occorrenze registrate per ogni punteggio (quante volte i valutatori hanno impiegato il valore $x$, quante volte il valore $y$, e così via);
- calcola la percentuale di ciascun punteggio rispetto agli altri;
- converte in una scala *logit* le medie relative alle occorrenze *osservate*, congiuntamente alle medie *attese*;
- calcola gli indici di conformità *outfit*;
- calcola i *livelli-soglia Rasch-Andrich* (che rappresentano il discrimine in termini di probabilità tra punteggi contigui).[35]

Il programma prevede un incremento progressivo (monotonico) dei *logit* corrispondenti ai vari livelli della scala, nonché dei *livelli-soglia Rasch-Andrich*. Nel caso ciò sia disatteso, occorre che la griglia di valutazione sia rivista, posto che la differenziazione tra i livelli non riflette adeguatamente lo sviluppo del costrutto della *performance*.

In aggiunta, Linacre (1999, 2004) consiglia che gli intervalli tra i *livelli-soglia Rasch-Andrich* siano contenuti all'interno dell'intervallo |1.4÷5|; nel caso la differenza sia <1.4 siamo avvisati di una eccessiva segmentazione della scala (alcuni livelli, cioè, non

[35] In altre parole, si tratta del valore in *logit* a cui corrisponde la pari probabilità (50%) di ricevere un certo voto (es. 4) o il voto prossimo (es. 5).

sono distintivi; è, quindi, il caso di fonderli); nel caso in cui, invece, la differenza sia >5, vi è un macrolivello che non discrimina convenientemente le fasce di competenza (val la pena, in tale circostanza, operare distinzioni ulteriori in seno alla scala).

Veniamo al nostro studio. Nella tabella 12, *infra*, si illustra la distribuzione dei punteggi assegnati dal gruppo.

Tabella 12. *La scala in adozione. Punteggi decimali, logit, outfit*

Legenda

| | |
|---|---|
| **punt=** | punteggi utilizzati dai valutatori |
| **oc=** | occorrenze |
| **%=** | occorrenze in percentuale |
| **L(o)=** | conversione delle medie *osservate* (corrispondenti ai singoli punteggi) in valori *logit* |
| **L(a)=** | conversione delle medie *attese* (i.e. corrette dal modello Rasch) in valori *logit* |
| **Outfit=** | indice di conformità sensibile ai valori *outliers* |

| **punt** | **Oc** | **%** | **L(o)** | **L(a)** | **Outfit** |
|---|---|---|---|---|---|
| 2 | 2 | ≈0% | -2.72 | -2.99 | 1.2 |
| 3 | 18 | 1% | -2.40 | -2.55 | 1.1 |
| 4 | 137 | 8% | -1.80 | -1.78 | 1.0 |
| 5 | 249 | 15% | -0.75 | -0.73 | 0.90 |
| 6 | 373 | 22% | 0.48 | 0.41 | 1.1 |
| 7 | 333 | 20% | 1.52 | 1.58 | 1.1 |
| 8 | 347 | 21% | 2.62 | 2.67 | 1.0 |
| 9 | 148 | 9% | 3.48 | 3.46 | 0.8 |
| 10 | 73 | 4% | 4.20 | 4.04 | 0.8 |

L'intervallo dei punteggi assegnati consta di 9 punti (|2÷10|), con una concentrazione nella fascia |5÷8| (78% dei voti si collocano su questa fascia). Il punteggio zero e il punteggio 1 non sono mai stati usati. Nella seconda colonna sono elencate le occorrenze dei punteggi, trasformate in percentuali nella terza colonna.

Nella quarta colonna abbiamo la media, convertita in *logit*, di un candidato la cui competenza corrisponde a quella determinata fascia di punteggio; nella quinta abbiamo la media corretta secondo il modello Rasch (al netto della severità del valutatore). È importante che sia l'una che l'altra abbiano un andamento monotonico, e cioè che avanzino con il progredire dei punteggi: significa che le fasce di livello della griglia in adozione riflettono l'evoluzione del costrutto.[36] Se noi scorriamo i valori delle colonne $L_{(o)}$ e $L_{(a)}$ notiamo appunto la crescita costante dei valori: ciò conferma la validità della struttura della griglia in adozione.

Nella sesta colonna abbiamo i valori di *outfit*, relativi a ciascuna fascia di punteggio. Nelle parole di Eckes (2015: 26)

> "This indicator compares the average examinee proficiency measures and the expected examinee proficiency measure, that is, the examinee proficiency measure the model would predict for a given rating category if the data were to fit the model. The greater the difference between the average and the expected measures, the larger the mean-square outfit statistic will be. In general, this statistic should not exceed 2.0"

Nel nostro caso, i valori di *outfit* ruotano attorno all'1, quindi rientrano nella norma (<2).

Nella tabella 13, *infra*, riportiamo i valori riferiti alle *soglie di Rasch-Andrich*. Una *soglia di Rasch-Andrich* esprime il valore *logit* in corrispondenza del quale convergono le curve di probabilità di due punteggi adiacenti (come evidenziato dalle linee tratteggiate nella fig. 4a, riportata ancor più avanti). Detto altrimenti, al valore *logit* di una *soglia di Rasch-Andrich* corrisponde la pari probabilità da

[36] Eckes scrive (2015: 118) "The basic requirement is that average measures advance monotonically with categories; that is, higher average measures produce observations in higher categories, and vice versa. When this requirement is met, it is safe to conclude that higher ratings correspond to more of the variable being measured".

parte di un candidato di ottenere due punteggi contigui. Facciamo un esempio, attingendo alla tabella 13, *infra*: un candidato la cui abilità sia pari a +1.11 *logit* (riga evidenziata con la freccia) ha una pari probabilità di ottenere un punteggio pari a 6 e un punteggio pari a 7.

Pure le *soglie di Rasch-Andrich* (SRA) presentano un incremento monotonico, confermando la qualità della griglia nel discriminare livelli inferiori e superiori di competenza.[37]

Facciamo notare, tuttavia, come alcuni *scarti tra le soglie di Rasch-Andrich* (con asterisco nella tab. 13, *infra*, e nella fig. 4a, *infra*) sono inferiori al valore raccomandato da Linacre di 1.4 (Linacre 1999, 2004). Si deduce che, a tratti, la griglia opera una *discriminazione eccessiva*. Ciò si evince, in particolare, per i valori estremi (in corrispondenza dei quali si contano assegnazioni di punteggio relativamente basse), e cioè tra la soglia 2|3 e la soglia 3|4 da un lato, e tra la soglia 8|9 e la soglia 9|10, dall'altro. Scarti fuori norma, ma di minor rilievo, si registrano anche tra la soglia 4|5 e la soglia 5|6, da un lato, e tra la soglia 6|7 e la soglia 7|8, dall'altro. Il gruppo si è servito relativamente poco, insomma, dei livelli estremi (punteggi 2, 3 e 10) e di alcuni livelli intermedi (5 e 7). Quest'ultimo problema, come vedremo più avanti, dipende dal *sovrauso di alcuni punteggi intrafascia* (6 e 8) da parte di un paio di valutatori (§ 10.2.7.4).

---

[37] Eckes scrive (2015: 119): "When the thresholds do not advance monotonically with categories, that is, when the thresholds are disordered, it can be inferred that the rating scale does not function properly: as one moves up the latent continuum, the categories involved would never be the most likely response to be observed".

Tabella 13. *Soglie di Rasch-Andrich*

Legenda
**D-punt=** discrimine tra i punteggi
**SRA=** soglie di Rasch-Andrich
**ES=** errore standard di misurazione
**S-SRA=** scarti tra le soglie di Rasch-Andrich

| **D-punt** | **SRA** | **ES** | **S-SRA** |
|---|---|---|---|
| 2\|3 | -4.98 | 0.72 | *0.77 |
| 3\|4 | -4.21 | 0.24 | 2.35 |
| 4\|5 | -1.86 | 0.11 | *1.29 |
| 5\|6 | -0.57 | 0.08 | 1.68 |
| 6\|7 | 1.11 | 0.08 | *0.99 |
| 7\|8 | 2.10 | 0.08 | 1.84 |
| 8\|9 | 3.94 | 0.09 | *0.53 |
| 9\|10 | 4.47 | 0.14 | |

Nella mappa delle curve di probabilità (fig. 4a, *infra*) evidenziamo tanto le *soglie* (linee verticali tratteggiate) e relativi valori (in basso), quanto gli *scarti* (in alto, in rosso; quelli fuori norma sono marcati con asterisco).

Come va letto il grafico? Nell'ordinata abbiamo l'indicazione della probabilità (es. 1 significa 100%; 0.9 significa 90%, e così via). L'asse delle ascisse, invece, rappresenta la scala *logit* [l'intervallo complessivo della scala (da -7 circa a +6) è più ampio rispetto al *range* dell'abilità dei candidati (che oscilla tra -3.69 a +3.21 *logit*; cfr. mappa delle variabili, fig. 2, *supra*)].

Abbiamo dieci curve, corrispondenti ai punteggi della scala usati dai valutatori (2÷9). Si veda, a tal proposito, la legenda sottostante (ad es. "category probability 4", in corrispondenza alla curva rosa, significa che ci si riferisce alla curva della probabilità relativa al punteggio 4).

Ciascuna curva assume un picco distintivo all'interno di un certo intervallo. Per esempio, la curva relativa al punteggio 4 assume un picco distintivo nell'intervallo |(-4.21)÷(-1.86)| *logit*. E cioè un candidato la cui abilità è compresa tra (-4.21) e (-1.86) *logit* ha una probabilità maggiore di ricevere un voto 4, rispetto a un voto 3 oppure a un voto 5. Quant'è questa probabilità? Per saperlo occorre tracciare una linea orizzontale e leggere il valore corrispondente sull'asse delle ordinate. Per comodità, andiamo alla figura successiva, 4b, e tracciamo sull'asse delle ordinate (cioè delle probabilità) le linee orizzontali passanti per la cresta della curva 4 e per le *soglie di Rash-Andrich* 3|4 e 4|5.

Ebbene

- in corrispondenza della soglia 3|4 (–4.21 *logit*), il candidato ha una pari probabilità di prendere un voto 3 o un voto 4 (circa 38%), a fronte invece di una percentuale molto bassa di prendere un voto 5, nell'ordine del 4% (linee tratteggiate rosse);
- in corrispondenza della cresta della curva di probabilità 4 (–3 *logit* circa), il candidato ha il 59% di probabilità di prendere un 4, a fronte del 18% di prendere un 5 o un 3 (linee tratteggiate gialle);
- in corrispondenza della soglia 4|5 (–1.86 *logit*), il candidato ha una pari probabilità di prendere un 4 o un 5 (42% circa), mentre la probabilità di prendere un 3 scende a circa il 5% (linee tratteggiate arancioni).

Se ci spostiamo a destra rispetto a questo intervallo e osserviamo la curva della probabilità relativa al punteggio 5 (curva in nero), notiamo che essa si distingue ("cresce") rispetto alle curve adiacenti nell'intervallo |(–1.86)÷(–0.57)| *logit*. A mano a mano che ci muoviamo verso il centro di questo intervallo, aumenta la probabilità di un candidato (dal 42% al 45% circa) di ricevere un voto pari a 5, rispetto a un voto pari a 4 o a un voto pari a 6.

L'andamento monotonico delle medie convertite in *logit* che evidenziavamo nella tabella 12, *supra*, si riflette nelle curve di probabilità (figg. 4a, 4b). Nel complesso abbiamo, infatti, dei profili ordinati: la curva relativa a un punteggio subentra alla curva relativa al punteggio precedente, come la più probabile, a mano a mano che progrediamo nella scala *logit*, da sinistra (valori negativi) verso destra (valori positivi), in maniera, appunto, speculare rispetto all'evolvere dell'abilità dei candidati.

Figura 4a. *Curve di probabilità dei punteggi e soglie di Rasch-Andrich*

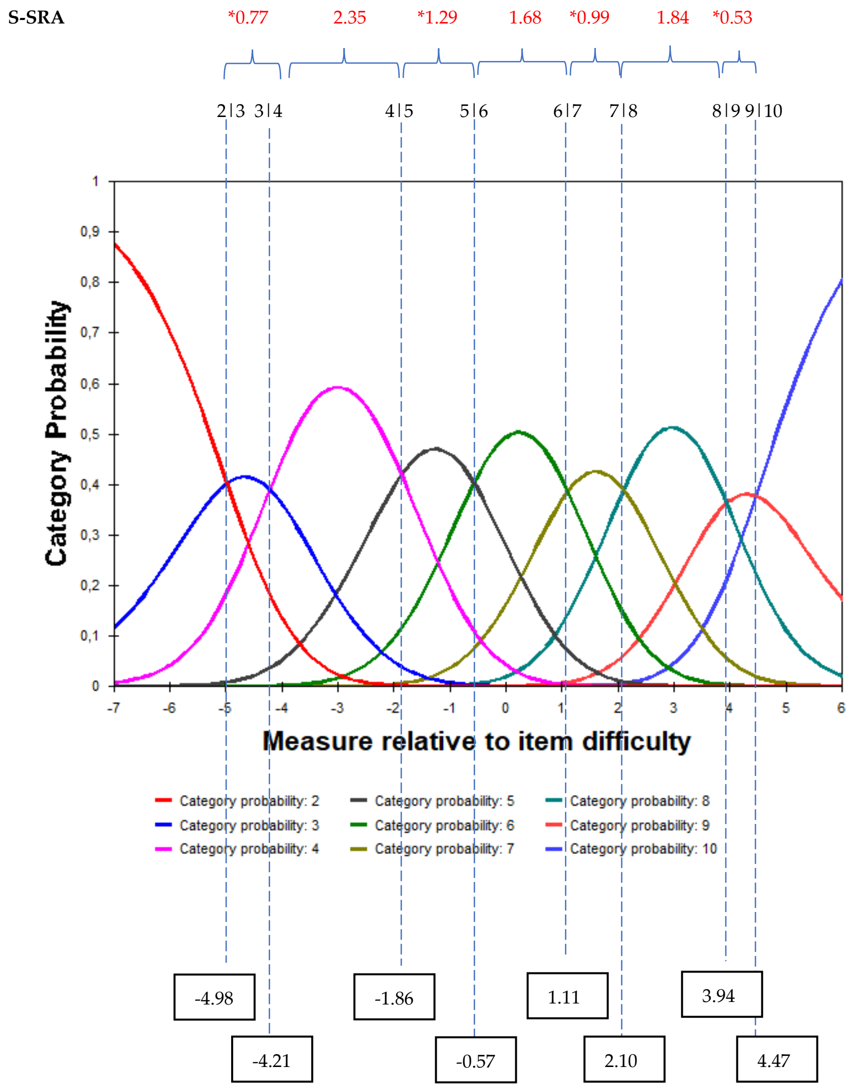
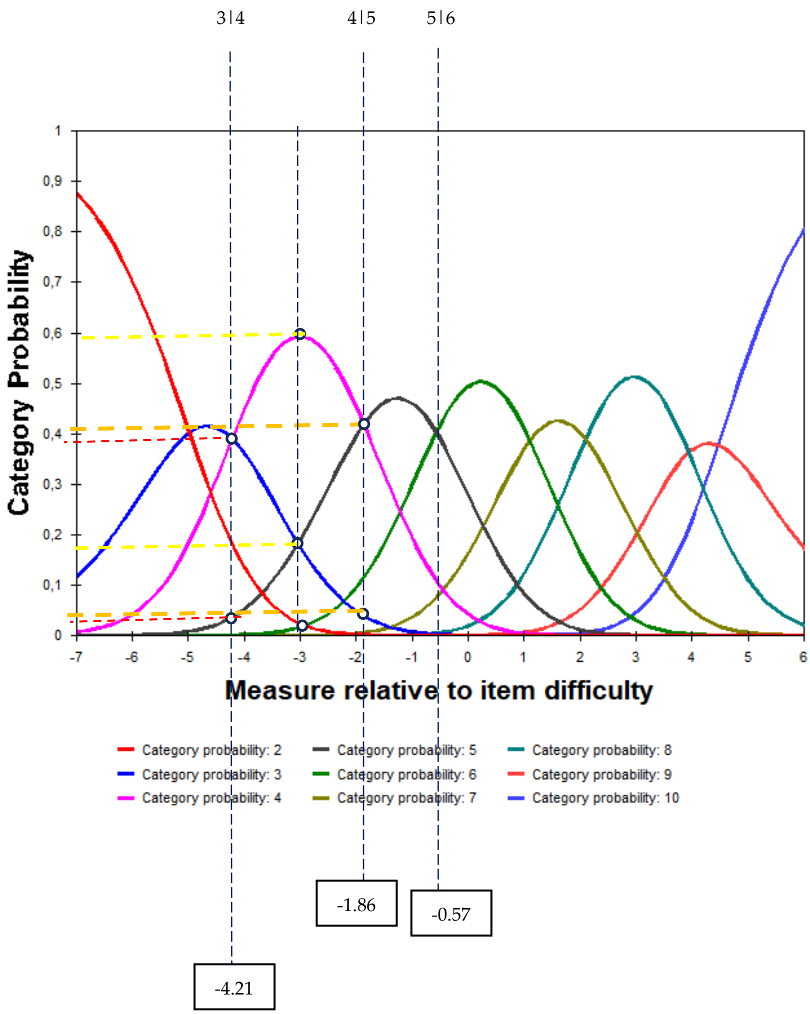


Figura 4b. *Curve di probabilità dei punteggi: esempio di corrispondenza valore logit e probabilità*



Più avanti vedremo, comunque, come in alcuni casi specifici, e cioè nel valutare alcune componenti del costrutto (*contenuto*, in primis), il comportamento di alcuni valutatori dimostri un uso inappropriato della scala (§§ 10.2.7.1.1; 10.2.7.2; 10.2.7.3; 10.2.7.4).

### 10.2.7. *Bias* specifici

Lo studio dei *bias* (letteralmente: distorsioni, alterazioni) ci permette di formulare delle ipotesi circa le cause di giudizi inadeguati (cfr. Kondo-Brown 2002; Myford, Wolfe 2003, 2004). Lo studio prevede la triangolazione di osservazioni di diverso genere; nel nostro caso

- *le statistiche di conformità delle componenti e dei compiti* (§ 10.2.7.1);
- *l'analisi delle interazioni valutatori/componenti* (§ 10.2.7.2);
- *lo studio delle valutazioni inattese* (§ 10.2.7.3);
- *la distribuzione dei punteggi da parte dei valutatori* (§ 10.2.7.4).

Condurremo le indagini separatamente, riportando ognuna a pagina nuova. In un capitolo successivo stileremo il profilo di ciascun valutatore (§ 10.2.8).

#### 10.2.7.1. Statistiche di conformità delle componenti e dei compiti

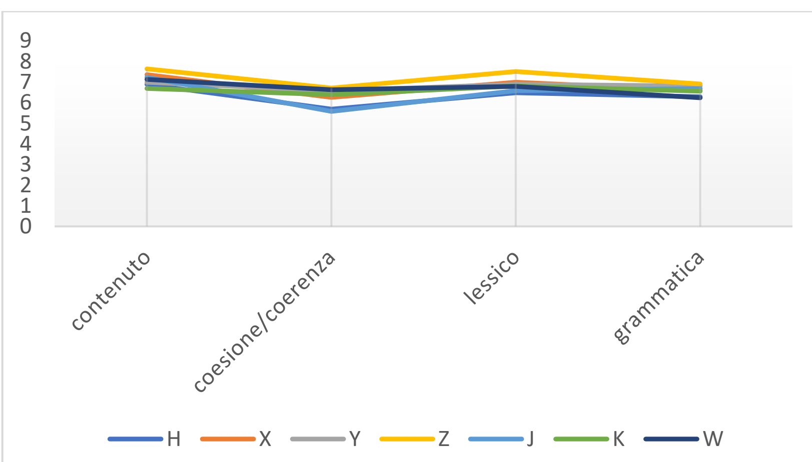
Consideriamo, dapprima, il modo in cui i valutatori – come gruppo – hanno gestito la valutazione delle componenti, riflesse nelle categorie della griglia (*contenuto; lessico; grammatica; coesione/coerenza*). In secondo luogo prendiamo in esame gli indici di conformità dei *compiti*.

##### 10.2.7.1.1. Le componenti

Dalla mappa delle variabili avevamo colto (fig. 2, *supra*) una differenza delle componenti in termini di difficoltà: la *coesione/coerenza* è la più difficile; il *contenuto* è la più facile; *lessico* e *grammatica* sono pressoché uguali in termini di difficoltà.

Ciò lo si evince anche dal profilo delle medie delle valutazioni (*medie* osservate) di ciascuna componente (fig. 5, *infra*). Nel grafico, è facile notare il distacco di [**Z**] (linea gialla), il valutatore più generoso.

Figura 5. *Medie dei punteggi assegnati alle componenti da ciascun valutatore*



A seguire, nella tab. 14, illustriamo i valori *logit* e le *statistiche di conformità* di ciascuna componente.

Tabella 14. *Medie, valori logit e statistiche di conformità relative alle componenti*

Legenda

| | |
|---|---|
| **MO=** | media osservata |
| **MC=** | media corretta nel modello Rasch |
| **L(c)=** | conversione in valori *logit* |
| **ES=** | errore standard di misurazione |
| **Infit=** | indice di conformità ponderato |
| **i-ZSTD=** | punteggio zeta relativo alla precisione statistica dei valori *infit* |
| **Outfit=** | indice di conformità sensibile alle osservazioni *outliers* |
| **o-ZSTD=** | punteggio zeta relativo alla precisione statistica dei valori *outfit* |

| | **MO** | **MC** | **L(c)** | **ES** | **Infit** | **i-ZSTD** | **Outfit** | **o-ZSTD** |
|---|---|---|---|---|---|---|---|---|
| contenuto | 7.1 | 7,10 | 0.60 | 0.06 | 1.34 | 4.6 | 1.36 | 4.8 |
| lessico | 6.9 | 6.82 | 0.22 | 0.06 | 0.90 | -1.5 | 0.92 | -1.1 |
| grammatica | 6.6 | 6.52 | -0.18 | 0.06 | 0.72 | -4.5 | 0.71 | -4.6 |
| coesione e coerenza | 6.3 | 6.19 | -0.64 | 0.06 | 1.01 | 0.2 | 1.00 | 0.0 |

Emergono due osservazioni:

- la componente del *contenuto* presenta *indici di conformità* relativamente alti (*infit*=1.34; *outfit*= 1.36), prossimi alla soglia di tolleranza, la quale – ricordiamo – è pari a 1.4 (considerevoli sono pure i punteggi zeta: 4.6; 4.8). Ciò rivela una *oscillazione dei giudizi* da parte dei singoli valutatori in merito a tale componente;
- la componente della *grammatica* presenta valori i quali, seppur a norma (*infit*= 0.72; *outfit*= 0.71), sono prossimi alla

soglia di *iperconformità* (che è pari a 0.6); alti, anche in questo caso, sono i punteggi zeta (-4.5; -4.6).

Più avanti approfondiremo entrambi gli aspetti (§ 10.2.7.2).

### 10.2.7.1.2. I compiti

In merito ai *compiti*, dicevamo che in una prima indagine (successivamente abbandonata) avevamo compattato i valori di ciascun compito nel *format* ad esso relativo: i compiti *a, c, e* erano stati ricondotti al *format A*; mentre i compiti *b, d, f* erano stati ricondotti al *format B*.

In questa ipotesi di omogeneizzazione risultava che il *format A* e il *format B* erano piuttosto calibrati, con uno scarto di 0.32 *logit*. Il *format* A risultava leggermente più *facile* del *format B* (cfr. tab. 15, infra).

Tabella 15. *Medie e statistiche di conformità relative ai format*

Legenda

| | |
|---|---|
| **format=** | tipologia di compito |
| **MO=** | media *osservata* |
| **MC=** | media *corretta* nel modello Rasch |
| **L(f)=** | conversione in valori *logit* |
| **ES=** | errore standard di misurazione |
| **Infit=** | indice di conformità ponderato |
| **i-ZSTD=** | punteggio zeta relativo alla precisione statistica dei valori *infit* |
| **Outfit=** | indice di conformità sensibile alle osservazioni *outliers* |
| **o-ZSTD=** | punteggio zeta relativo alla precisione statistica dei valori *outfit* |

| **format** | **MO** | **MC** | **L(f)** | **ES** | **Infit** | **i-ZSTD** | **Outfit** | **o-ZSTD** |
|---|---|---|---|---|---|---|---|---|
| A | 6.8 | 6.77 | **0.16** | 0.04 | 0.99 | -0.1 | 0.99 | -0.1 |
| B | 6.6 | 6.54 | **-0.16** | 0.04 | 1.01 | 0.0 | 1.01 | 0.1 |

In un secondo momento, onde evitare distorsioni nei risultati, abbiamo rifiutato l'ipotesi dell'omogeneizzazione, propendendo per l'ancoraggio allo zero *logit* del grado di difficoltà di ciascun compito. A spingerci in questa direzione era anche il fatto che, benché relativamente contenuta, la differenza tra i *format* risultava statisticamente significativa (il *chi-quadrato* era relativamente alto in relazione al numero degli elementi; *chi quadrato*=27.8; *gradi di libertà*=1; $p<0.05$).

L'indagine delle *medie osservate* e dell'*indice di conformità* dei singoli compiti è riportata nella tab. 16, nella pagina che segue (il valore *logit*, come si vede, è ancorato allo 0).

In riferimento alle *medie osservate*, balza all'occhio il valore relativamente alto del compito *e* (evidenziato in giallo). Il leggero disallineamento tra i *format*, che emergeva nei primi calcoli (considerata l'ipotesi dell'omogeneizzazione), si può spiegare probabilmente a partire da questa evidenza.

Tabella 16. *Medie osservate e statistiche di conformità relative ai compiti*

Legenda

| | |
|---|---|
| **format=** | tipologia di compito |
| **Task=** | compito specifico assegnato nelle diverse sessioni |
| **MO=** | media osservata |
| **MA=** | media ancorata al valore centrale |
| **LA=** | *logit* ancorato al valore centrale |
| **ES=** | errore standard di misurazione |
| **Infit=** | indice di conformità ponderato |
| **i-ZSTD=** | punteggio zeta relativo alla precisione statistica dei valori *infit* |
| **Outfit=** | indice di conformità sensibile alle osservazioni *outliers* |
| **o-ZSTD=** | punteggio zeta relativo alla precisione statistica dei valori *outfit* |

| format | task | MO | MA | LA | ES | Infit | i-ZSTD | Outfit | o-ZSTD |
|---|---|---|---|---|---|---|---|---|---|
| A | A | 6.80 | 6.65ª | 0.00 | 0.07 | 0.96 | -0.4 | 0.95 | -0.50 |
| A | C | 6.50 | 6.65ª | 0.00 | 0.07 | 1.11 | 1.3 | 1.13 | 1.50 |
| A | E | 7.10 | 6.65ª | 0.00 | 0.07 | 0.90 | -1.1 | 0.90 | -1.20 |
| B | B | 6.50 | 6.65ª | 0.00 | 0.07 | 1.10 | 1.1 | 1.09 | 1.00 |
| B | D | 6.40 | 6.65ª | 0.00 | 0.07 | 1.11 | 1.3. | 1.15 | 1.70 |
| B | F | 6.80 | 6.65ª | 0.00 | 0.07 | 0.77 | -2.9 | 0.77 | -2.90 |

Nella tabella 17, *infra*, evidenziamo la *media aritmetica* tra le *medie osservate* relative *a compiti che pertengono alla stessa sessione.* Presi complessivamente, i *format* della sessione di novembre 2017 hanno riscosso punteggi relativamente alti (si tratta, ad ogni modo, di un dato poco significativo, posto che i calcoli sono operati a partire da gruppi differenti di candidati).

Tabella 17. *Medie aritmetiche relative ai format somministrati nella medesima sessione*

<table>
<tr><th>format</th><th>task</th><th>media osservata</th><th>media tra i compiti di una stessa sessione</th><th>candidati</th><th>gruppi</th><th>Sessioni PLIDA</th></tr>
<tr><td>A</td><td>A</td><td>6.80</td><td rowspan="2">6.65</td><td rowspan="2">1-10</td><td rowspan="2">I</td><td rowspan="2">giugno 2016</td></tr>
<tr><td>B</td><td>B</td><td>6.50</td></tr>
<tr><td>A</td><td>C</td><td>6.50</td><td rowspan="2">6.45</td><td rowspan="2">11-20</td><td rowspan="2">II</td><td rowspan="2">agosto 2016</td></tr>
<tr><td>B</td><td>D</td><td>6.40</td></tr>
<tr><td>A</td><td>E</td><td>7.10</td><td rowspan="2">6.95</td><td rowspan="2">21-30</td><td rowspan="2">III</td><td rowspan="2">novembre 2017</td></tr>
<tr><td>B</td><td>F</td><td>6.80</td></tr>
</table>

### 10.2.7.2. Analisi delle interazioni valutatori-componenti

L'indagine che abbiamo condotto fino ad ora ritrae il comportamento *generale* dei valutatori.

Possiamo, tuttavia, operare uno studio più approfondito considerando l'accumulo di residui occorsi durante la valutazione dei singoli candidati, delle singole componenti o del singolo compito.

Questo genere di analisi, che riguarda le interazioni tra un paio di variabili (valutatore-candidato; valutatore-compito; valutatore-componente), si chiama *analisi delle interazioni*, o più tecnicamente "differential facet functioning", DFF (per approfondimenti, cfr. **Appendice 3**). In sostanza, possiamo pur sapere che in genere un dato valutatore si presenta come generoso, ma se vogliamo scoprire come la sua generosità vari al variare dei candidati, oppure delle componenti o dei compiti, dobbiamo munirci di una lente di ingrandimento. Ecco dunque a cosa serve un'analisi delle interazioni: essa ci restituisce i *bias locali.*

In questa sede riportiamo i dati relativi all'*interazione tra valutatori e componenti.* Tale indagine ci permette di cogliere i *bias* attivi durante la valutazione di una *componente*. La tabella di riferimento è la n. 18, *infra*; essa presenta, tra gli altri, i seguenti dati:

- i valori *logit* relativi a ciascuna componente [**L(c)**] (ripresi dalla tab. 14)
- i valori *logit* dei *residui* relativi a ciascuna *componente* [(**B**) *bias size*)[38]
- i valori *logit* della *generosità generale* di un valutatore, ripresi dalla tab. 6, *supra* [**L(v)**];
- i valori *logit* della *generosità riferita alla singola componente* [(**L(vc)**] (frutto della somma algebrica tra i valori *logit* dei

[38] Ricordiamo che il residuo è positivo quando il valutatore è più generoso del previsto, mentre è negativo nel caso contrario.

residui relativi alla componente e i valori *logit* corrispondenti al grado di *generosità generale* del valutatore);

- le *statistiche di conformità* (i valori ≥1.40, che palesano una *iperdifformità*, sono in giallo; i valori ≤0.60, che attestano, invece, una *iperconformità*, sono in arancione).

Occorre precisare che i valori con un *t statistico* ≥±2 (le caselle sono evidenziate con un fondo azzurro chiaro) sono statisticamente significativi; a *t statistico* ≥±2.6 corrisponde un'elevata significatività (le caselle sono evidenziate con un fondo azzurro scuro).[39]

Tabella 18. *Interazioni valutatori-componenti*

| Legenda | |
|---|---|
| **Comp=** | componente |
| **L(c)=** | valore in *logit* della componente (come da tab. 14, supra) |
| **Val=** | valutatore |
| **L(v)=** | generosità media (*logit* del valutatore, come da tab. 6, *supra*) |
| **∑=** | valore totale *osservato* (punteggio totale accumulato nella valutazione della componente) |
| **∑C=** | valore totale *corretto* (punteggio totale accumulato nella valutazione della componente corretto secondo il modello Rasch) |
| **oc=** | numero di occorrenze |
| **B=** | *bias* relativo alla singola componente |
| **L(r)=** | conversione in *logit* del residuo |
| **ES=** | errore standard di misurazione |
| **t=** | valore di verifica dell'ipotesi |
| **p=** | indice di probabilità |
| **L(vc)=** | generosità in riferimento alla singola componente |
| **Infit=** | indice di conformità ponderato |
| **Outfit=** | indice di conformità sensibile alle osservazioni *outliers* |
| CONT= | contenuto |
| CC= | coesione e coerenza |
| GRAM= | grammatica |
| LEX= | lessico |

[39] Cfr. Linacre 2012: 13.

| Comp | L(c) | Val | L(v) | ∑ | ∑C | Oc | B | ES | T | p | L(vc) | Infit | Outfit |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CC | -0.64 | **W** | **1.13** | 397 | 374.3 | 60 | **0.55** | 0.15 | 3.55 | 0.0008 | **1.68** | 0.80 | 0.80 |
| GRAM | -0.18 | **J** | **0.87** | 400 | 382.60 | 60 | **0.42** | 0.15 | 2.70 | 0.0089 | **1.29** | 0.60 | 0.60 |
| CONT | 0.60 | **J** | **0.87** | 433 | 415.5 | 60 | **0.41** | 0.15 | 2.69 | 0.0094 | **1.30** | 1.60 | 1.60 |
| CC | -0.64 | **K** | **1.01** | 383 | 369.4 | 60 | **0.33** | 0.16 | 2.13 | 0.0375 | **1.34** | 0.90 | 0.90 |
| CC | -0.64 | **Y** | **1.31** | 394 | 381.4 | 60 | **0.30** | 0.15 | 1.96 | 0.0543 | **1.61** | 1.00 | 1.00 |
| LEX | 0.22 | **Z** | **1.84** | 450 | 440.4 | 60 | **0.23** | 0.16 | 1.48 | 0.1443 | **2.07** | 1.70 | 1.60 |
| CONT | 0.60 | **H** | **0.62** | 413 | 404.08 | 60 | **0.19** | 0.15 | 1.27 | 0.2090 | **0.81** | 1.50 | 1.60 |
| CONT | 0.60 | **X** | **1.28** | 441 | 433.0 | 60 | **0.19** | 0.15 | 1.24 | 0.2214 | **1.47** | 1.40 | 1.50 |
| GRAM | -0.18 | **Y** | **1.31** | 408 | 400.8 | 60 | **0.17** | 0.15 | 1.12 | 0.2694 | **1.48** | 0.90 | 1.00 |
| GRAM | -0.18 | **H** | **0.62** | 377 | 372.2 | 60 | **0.12** | 0.16 | 0.75 | 0.4543 | **0.74** | 0.40 | 0.40 |
| GRAM | -0.18 | **K** | **1.01** | 393 | 388.60 | 60 | **0.11** | 0.15 | 0.69 | 0.4931 | **1.12** | 0.40 | 0.40 |
| LEX | 0.22 | **K** | **1.01** | 408 | 405.4 | 60 | **0.06** | 0.15 | 0.41 | 0.6865 | **1.07** | 0.50 | 0.60 |
| LEX | 0.22 | **X** | **1.28** | 419 | 416.7 | 60 | **0.05** | 0.15 | 0.35 | 0.7280 | **1.33** | 0.60 | 0.70 |
| CONT | 0.60 | **Z** | **1.84** | 458 | 456.3 | 60 | **0.04** | 0.16 | 0.27 | 0.785 | **1.92** | 1.80 | 1.80 |
| CONT | 0.60 | **W** | **1.15** | 427 | 426.7 | 60 | **0.01** | 0.15 | 0.06 | 0.9552 | **1.16** | 0.70 | 0.70 |
| LEX | 0.22 | **H** | **0.63** | 388 | 388.7 | 60 | **-0.02** | 0.15 | -0.11 | 0.9111 | **0.61** | 0.87 | 0.80 |
| CC | -0.64 | **Z** | **1.87** | 402 | 403.7 | 60 | **-0.04** | 0.15 | -0.26 | 0.7973 | **1.83** | 1.30 | 1.30 |
| LEX | 0.22 | **W** | **1.13** | 407 | 410.4 | 60 | **-0.08** | 0.15 | -0.52 | 0.6028 | **1.07** | 0.60 | 0.60 |
| LEX | 0.22 | **Y** | **1.31** | 413 | 417.7 | 60 | **-0.11** | 0.15 | -0.72 | .04715 | **1.20** | 1.1 | 1.1 |
| GRAM | -0.18 | **X** | **1.28** | 395 | 399.8 | 60 | **-0.11** | 0.15 | -0.73 | 0.4664 | **1.17** | 0.70 | 0.70 |
| LEX | 0.22 | **J** | **0.87** | 394 | 399.3 | 60 | **-0.13** | 0.15 | -0.82 | 0.4169 | **0.74** | 1.00 | 1.00 |
| CC | -0.64 | **X** | **1.28** | 375 | 380.4 | 60 | **-0.13** | 0.16 | -0.84 | 0.4057 | **1.15** | 0.70 | 0.70 |
| GRAM | -0.18 | **Z** | **1.84** | 414 | 423,4 | 60 | **-0.22** | 0.15 | -1.44 | 0.1554 | **1.62** | 0.80 | 0.80 |
| CC | -0.64 | **H** | **0.62** | 341 | 353.30 | 60 | **-0.31** | 0.16 | -1.93 | 0.0582 | **0.31** | 0.80 | 0.80 |
| CONT | 0.60 | **Y** | **1.31** | 419 | 434 | 60 | **-0.35** | 0.15 | -2.29 | 0.0255 | **0.96** | 0.90 | 0.80 |
| GRAM | -0.18 | **W** | **1.13** | 374 | 393.5 | 60 | **-0.47** | 0.16 | -3.02 | 0.0038 | **0.66** | 0.70 | 0.70 |
| CONT | 0.60 | **K** | **1.01** | 401 | 412.6 | 60 | **-0.48** | 0.15 | -3.15 | 0.0026 | **0.53** | 1.1 | 1.1 |
| CC | -0.66 | **J** | **0.87** | 334 | 363.50 | 60 | **-0.73** | 0.16 | -4.63 | 0.0000 | **0.14** | 0.60 | 0.60 |

Il valore-chiave nella tabella è quello riferito al *bias* (B); esso traduce i residui accumulati a causa di una gestione delle componenti non in linea con le aspettative del modello Rasch.

Facciamo un esempio con l'analisi dell'interazione tra il *valutatore* [**W**] e *la componente della coesione/coerenza* (prima riga della tabella). Il punteggio riscosso durante la valutazione della componente nei 60 *task* valutati da [**W**] è pari a 397 (∑); il punteggio totale atteso dal modello è pari a 374.3 (∑C). La differenza ammonta a 22.7: ciò rappresenta l'insieme dei residui prodotti durante le valutazioni di [**W**] in riferimento alla componente della *coesione e coerenza*. Convertito in *logit*, questo valore si traduce nella misura di 0.55 (B).

A cosa va attribuita questa cifra? Alla componente o al valutatore? Riteniamo sia di pertinenza del valutatore; il problema cioè non riguarda tanto la calibrazione della componente rispetto alle altre ma il valutatore.

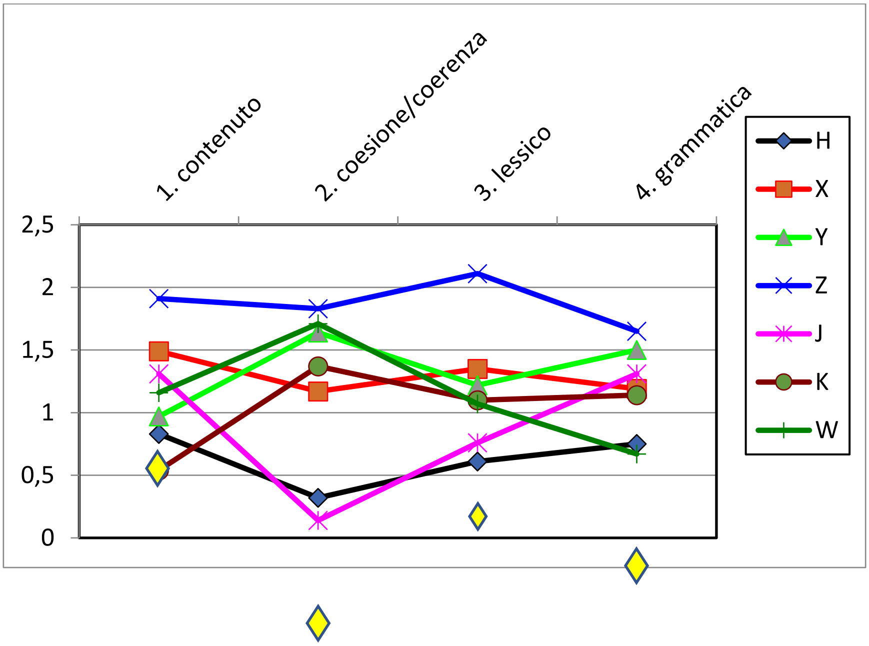
Per cogliere la generosità manifestata dal valutatore [**W**] durante la valutazione della *coesione e coerenza* dobbiamo sommare la cifra relativa al *bias* (0.55) a quella della generosità di [**W**] (1.13). Il risultato (1.68) ci fornisce un valore [L(vc)], grazie al quale possiamo apprezzare il grado di generosità di [**W**] in riferimento alla componente specifica.

Alla figura 6, *infra*, i valori [L(vc)] riportati nella tabella 18, *supra*, sono riproposti in forma di grafico.

Il rombo giallo ◊ evidenzia il posizionamento di ciascuna componente sulla scala *logit*, al netto dell'influenza della generosità di chi valuta (i valori sono ripresi dalla tabella 14, *supra*, e sono i seguenti: *contenuto*= 0.60; *coesione e coerenza*= –0.64; *lessico*= 0.22; *grammatica*= –0.18).

Si noti come solo il valutatore [**K**], nell'atto di giudicare il *contenuto*, si avvicini al valore *logit* della componente.

Figura 6. *Valori della generosità dei valutatori in relazione alle componenti*




Quali altre osservazioni si possono fare a partire dalla figura?

In primo luogo, come ampiamente notato, emerge la generosità diffusa di [**Z**], la cui spezzata è ben al di sopra di tutte le altre. In seconda istanza si evince una distinzione interna al gruppo in merito alla valutazione della *coesione e coerenza*: i valutatori [**H**] e [**J**] hanno uno standard più severo rispetto al gruppo, e più prossimo al valore *logit* della componente. In sostanza, questi due valutatori vedono dei difetti nella gestione della *coesione e coerenza* da parte dei candidati che i colleghi non percepiscono.

Nella tabella 19, *infra*, abbiamo riassunto le violazioni delle statistiche di conformità durante la valutazione delle componenti rilevate nella tabella 18, *supra*.

Come si vede

- quattro valutatori su sette presentano una *impredicibilità* nella valutazione del *contenuto* (la quale fa supporre una generale difficoltà a interpretare i descrittori della griglia relativi a tale componente);
- cinque valutatori su sette presentano una *iperconformità* distribuita su varie componenti.

Vale la pena precisare, ad ogni modo, che solo per il valutatore [J] si danno valori statisticamente significativi ($t>2$; $p<0.05$), cioè non occasionali: essi concernono l'*iperdifformità* in relazione alla valutazione della componente del *contenuto* e l'*iperconformità* in relazione sia alla componente della *coesione e coerenza* che alla componente della *grammatica*.

Tabella 19. *Violazioni delle statistiche di conformità in relazione alle componenti*

Legenda

▲= occorrenza statisticamente significativa ($t \geq 2$)

iperdifformità dal modello ("misfitting")

iperconformità ("overfitting)

| | [H] | [X] | [Y] | [J] | [Z] | [K] | [W] |
|---|---|---|---|---|---|---|---|
| contenuto | | | | ▲ | | | |
| coesione/ coerenza | | | | ▲ | | | |
| lessico | | | | | | | |
| grammatica | | | | ▲ | | | |

#### 10.2.7.3. Analisi delle valutazioni inattese

La nostra indagine procede in modo sempre più dettagliato: dalla valutazione delle *componenti* e dei *compiti* da parte del gruppo (§ 10.2.7.1) siamo passati all'analisi delle *interazioni valutatore-componente* (§ 10.2.7.2), e giungiamo, infine, alla considerazione delle *valutazioni inattese.* Analizziamo cioè le occasioni puntuali (in merito all'attribuzione di un punteggio a una certa componente in riferimento all'elaborato prodotto da un certo candidato) in cui si sono registrate le *maggiori* deviazioni dal modello (e quindi si sono verificati i maggiori *residui*).

L'esito di questa ricognizione è rappresentato nella tab. 20, *infra*. I *residui* ivi riportati corrispondono a valori *standardizzati* (ovvero divisi per la deviazione standard stimata) ≥±2, una soglia oltre la quale la divergenza dal modello (in senso positivo o negativo) è *significativa* (se il residuo standardizzato è ≥±3 il dato si presenta come *altamente significativo*, cioè il modello Rasch qualifica quel particolare giudizio come *molto inaspettato*).[40]

Per comodità del lettore i residui sono divisi in due blocchi:

- nel primo abbiamo i *residui negativi,* che indicano cioè valutazioni nelle quali il valutatore è stato più severo del previsto (le voci sono in ordine decrescente a dipendere dal *residuo standardizzato*);
- nel secondo abbiamo i *residui positivi*, in corrispondenza dei quali si sono registrati i giudizi più generosi del previsto (le voci sono in ordine crescente a dipendere dal *residuo standardizzato*).

---

[40] Scrive Eckes 2015: 75: "Standardized residuals with absolute values greater than 2 have $p < .05$ under Rasch-model conditions, and so indicate significant departure in the data from the Rasch model. Those observations are commonly considered significantly unexpected and may be subjected to closer inspection". Per approfondimenti, cfr. Engelhard, 2002; Myford & Wolfe, 2003.

Si noterà come molti residui riguardino il *contenuto* (le occorrenze sono marcate in rosso).[41] Colpisce poi il fatto che i residui maggiori rimandino ad elaborati specifici:

- l'*elaborato b* del *candidato 8* ha riscosso i *maggiori residui negativi* (marcati in giallo), a carico di tutti i valutatori, ad eccezione di [**Y**];
- gli *elaborati c* e *d* del *candidato 16* hanno riscosso i *maggiori residui positivi* (marcati in verde) durante la valutazione operata da [**J**], [**H**] e [**Z**].

Tabella 20. *Valutazioni inattese*

| Legenda | |
|---|---|
| **Comp=** | componente |
| **Task=** | tipo di compito |
| **Cand=** | candidato |
| **Valut=** | valutatore |
| **VO=** | valore *osservato* |
| **VA=** | valore *atteso* |
| **Re=** | residuo |
| **ReSt=** | residuo standardizzato |
| CONT= | contenuto |
| CC= | coesione e coerenza |
| GRAM= | grammatica |
| LEX= | lessico |

[41] Quasi la metà dei residui standardizzati con un valore compreso tra|2÷3| concerne questa componente; oltre la soglia del 3, la quasi totalità dei residui rimanda al *contenuto*.

| Comp | Task | Cand | Valut | VO | VA | Re | ReSt |
|---|---|---|---|---|---|---|---|
| CONT | B | 8 | **H** | 5 | 8.4 | **-3.4** | -3.9 |
| CONT | B | 8 | **X** | 6 | 8.9 | **-2.9** | -3.4 |
| CONT | B | 8 | **W** | 6 | 8.8 | **-2.8** | -3.2 |
| CONT | C | 13 | **J** | 3 | 5.7 | **-2.7** | -3.2 |
| CONT | B | 8 | **K** | 6 | 8.7 | **-2.7** | -3.1 |
| CONT | B | 8 | **J** | 6 | 8.6 | **-2.6** | -3.0 |
| CONT | B | 8 | **Z** | 7 | 9.3 | **-2.3** | -3.0 |
| LEX | D | 16 | **J** | 2 | 4.2 | **-2.2** | -3.0 |
| CONT | B | 9 | **Z** | 6 | 8.6 | **-2.6** | -2.9 |
| LEX | D | 16 | **H** | 2 | 4.1 | **-2.1** | -2.9 |
| CC | E | 22 | **K** | 6 | 8.4 | **-2.4** | -2.8 |
| CONT | F | 21 | **K** | 6 | 8.5 | **-2.5** | -2.8 |
| CONT | C | 13 | **Z** | 4 | 6.3 | **-2.3** | -2.7 |
| CC | F | 22 | **J** | 6 | 8.3 | **-2.3** | -2.7 |
| LEX | A | 5 | **Z** | 5 | 7.2 | **-2.2** | -2.6 |
| CONT | D | 20 | **Z** | 7 | 9.1 | **-2.6** | -2.4 |
| LEX | D | 12 | **Y** | 4 | 6.1 | **-2.1** | -2.5 |
| LEX | A | 2 | **Y** | 4 | 6.0 | **-2.0** | -2.4 |
| CC | B | 7 | **Z** | 6 | 8.1 | **-2.1** | -2.4 |
| CC | B | 4 | **J** | 6 | 8.0 | **-2.0** | -2.4 |
| CC | C | 11 | **K** | 4 | 6.0 | **-2.0** | -2.4 |
| CONT | C | 13 | **X** | 4 | 5.9 | **-1.9** | -2.3 |
| CONT | C | 11 | **K** | 5 | 6.9 | **-1.9** | -2.2 |
| LEX | D | 16 | **Z** | 3 | 4.8 | **-1.8** | -2.2 |
| GRAM | B | 30 | **Z** | 3 | 4.6 | **-1.6** | -2.1 |
| LEX | B | 22 | **J** | 4 | 5.7 | **-1.7** | -2.1 |
| CONT | C | 12 | **Z** | 5 | 6.8 | **-1.8** | -2.1 |
| CC | D | 14 | **H** | 4 | 5.8 | **-1.8** | -2.1 |
| LEX | D | 11 | **Y** | 4 | 5.8 | **-1.8** | -2.1 |
| CONT | F | 22 | **X** | 5 | 6.8 | -1.8 | -2.1 |
| CONT | A | 24 | **Z** | 8 | 9.5 | -1.5 | -2.1 |
| LEX | E | 25 | **H** | 4 | 5.7 | **-1.7** | -2.0 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CONT | E | 24 | **Z** | 4 | 5.7 | **-1.7** | -2.0 |
| GRAM | F | 27 | **X** | 5 | 6.8 | **-1.8** | -2.0 |

Seconda parte

| **Comp** | **Task** | **Cand** | **Valut** | **VO** | **VA** | **Re** | **ReSt** |
|---|---|---|---|---|---|---|---|
| GRAM | B | 9 | **J** | 9 | 7.3 | **1.7** | 2.0 |
| LEX | C | 14 | **Z** | 9 | 7.3 | **1.7** | 2.0 |
| LEX | D | 14 | **Z** | 9 | 7.3 | **1.7** | 2.0 |
| CC | A | 5 | **Y** | 8 | 6.2 | **1.8** | 2.1 |
| CC | A | 2 | **W** | 7 | 5.3 | **1.7** | 2.1 |
| CC | B | 2 | **W** | 7 | 5.3 | **1.7** | 2.1 |
| CONT | C | 15 | **X** | 9 | 7.2 | **1.8** | 2.1 |
| GRAM | C | 14 | **J** | 8 | 6.2 | **1.8** | 2.1 |
| CONT | c | 16 | **J** | 6 | 4.4 | **1.6** | 2.1 |
| LEX | d | 18 | **Z** | 10 | 8.2 | **1.8** | 2.1 |
| GRAM | d | 14 | **J** | 8 | 6.2 | **1.8** | 2.1 |
| CC | c | 11 | **X** | 8 | 6.2 | **1.8** | 2.2 |
| LEX | d | 13 | **H** | 7 | 5.2 | **1.8** | 2.2 |
| CONT | e | 30 | **H** | 7 | 5.2 | **1.8** | 2.2 |
| CC | e | 21 | **Z** | 10 | 8.1 | **1.9** | 2.2 |
| CC | f | 24 | **K** | 6 | 4.4 | **1.6** | 2.2 |
| LEX | b | 11 | **K** | 6 | 4.3 | **1.7** | 2.3 |
| LEX | c | 19 | **Z** | 8 | 6.1 | **1.9** | 2.3 |
| CC | c | 15 | **K** | 8 | 6.1 | **1.9** | 2.3 |
| CC | c | 11 | **W** | 8 | 6.1 | **1.9** | 2.3 |
| LEX | d | 12 | **X** | 8 | 6.1 | **1.9** | 2.3 |
| CC | d | 13 | **Y** | 7 | 5.1 | **1.9** | 2.3 |
| LEX | d | 13 | **Z** | 8 | 6.1 | **1.9** | 2.3 |
| CONT | f | 24 | **K** | 7 | 5.1 | **1.9** | 2.3 |
| CC | a | 8 | **Y** | 10 | 8.0 | **2.0** | 2.4 |
| CONT | a | 5 | **J** | 9 | 6.8 | **2.2** | 2.5 |
| CONT | b | 9 | **J** | 10 | 7.8 | **2.2** | 2.5 |
| CONT | e | 27 | **Z** | 10 | 7.9 | **2.1** | 2.5 |
| GRAM | a | 3 | **Y** | 7 | 5.0 | **2.0** | 2.6 |
| CC | e | 28 | **W** | 10 | 7.8 | **2.2** | 2.6 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| CC | a | 7 | **Y** | 10 | 7.7 | **2.3** | 2.7 |
| CONT | e | 30 | **X** | 8 | 5.7 | **2.3** | 2.8 |
| CONT | d | 16 | **J** | 7 | 4.4 | **2.6** | 3.5 |
| CONT | c | 16 | **H** | 7 | 4.3 | **2.7** | 3.7 |
| CONT | c | 16 | **Z** | 8 | 5.0 | **3.0** | 3.7 |
| CONT | d | 16 | **H** | 7 | 4.3 | **2.7** | 3.7 |

Poiché le maggiori deviazioni da parte del gruppo (ad eccezione di [**Y**] e, in misura minore, di [**W**]) riguardano il *contenuto* e considerato che esse sono sia per *difetto* (punteggi attesi nelle fasce più elevate sono stati riportati a valori intermedi) che per *eccesso* (punteggi attesi nelle fasce insufficienti sono stati riportati su fasce sufficienti e, al tempo stesso, punteggi attesi su fasce sufficienti sono stati riportati sulle fasce più alte), si conferma una *erraticità* generale nell'attribuzione dei valori in rapporto a questa componente. Ciò fa supporre una difficoltà, da parte dei *raters*, ad interpretare i descrittori relativi. È probabile, cioè, che la griglia non riesca a catturare tutte le caratteristiche testuali relative alla categoria "efficacia e contenuto", o perlomeno che certi tipi di testi non siano facilmente valutabili con i descrittori a disposizione. Ulteriori indagini dovrebbero essere realizzate.

### 10.2.7.4. La distribuzione dei punteggi da parte dei valutatori

Esuliamo per un momento dall'analisi MFRM e procediamo con uno scrutinio delle percentuali dei voti assegnati dal gruppo, nell'ottica di scorgere *pattern* di giudizio.

Nella tabella 21a, *infra*, abbiamo evidenziato un paio di fenomeni:

- alcuni punteggi non sono stati usati (0, 1); altri sono stati utilizzati molto poco (2, 10;[42] si vedano a tal proposito le occorrenze nulle contrassegnate da fondo grigio);[43]
- vi è una *concentrazione dei giudizi* (≥67%, cioè ≥2/3) *in terne salienti di punteggi* (a volte contigui, a volte no) o in *coppie salienti di punteggi non contigui*. Tali "poli di attrazione" sono evidenziati con fondo giallo. La scala, evidentemente, è stata usata in modo *selettivo* – un fatto facile da cogliersi soprattutto in riferimento alla valutazione del *contenuto*.

[42] È una considerazione, del resto, che avevamo già espresso nel § 10.2.6.

[43] Vale la pena rilevare, in ogni caso, l'uso più esteso della scala in riferimento al *lessico* rispetto alle altre componenti.

Tabella 21a. *Percentuali dei punteggi assegnati. I nuclei*

CONTENUTO

| **Fasce di punteggio** | **Punteggi** | [H] | [X] | [Y] | [Z] | [J] | [K] | [W] |
|---|---|---|---|---|---|---|---|---|
| Prima | 0 | | | | | | | |
| Seconda | 1 | | | | | | | |
| | 2 | | | | | | | |
| Terza | 3 | | | | | 2 | | |
| | 4 | 5 | 3 | 3 | 5 | 7 | 8 | 5 |
| Quarta | 5 | 18 | 8 | 20 | 5 | 8 | 7 | 8 |
| | 6 | 10 | 22 | 18 | 20 | 18 | 42 | 22 |
| Quinta | 7 | 33 | 12 | 18 | 12 | 17 | 5 | 27 |
| | 8 | 20 | 28 | 23 | 30 | 23 | 28 | 18 |
| Sesta | 9 | 10 | 25 | 7 | 7 | 15 | 10 | 13 |
| | 10 | 3 | 2 | 10 | 22 | 10 | | 7 |

COESIONE/COERENZA

| **Fasce di punteggio** | **Punteggi** | [H] | [X] | [Y] | [Z] | [J] | [K] | [W] |
|---|---|---|---|---|---|---|---|---|
| Prima | 0 | | | | | | | |
| Seconda | 1 | | | | | | | |
| | 2 | | | | | | | |
| Terza | 3 | | | | 2 | 2 | | |
| | 4 | 13 | 13 | 18 | 12 | 12 | 15 | 7 |
| Quarta | 5 | 33 | 13 | 12 | 12 | 33 | 13 | 18 |
| | 6 | 30 | 35 | 20 | 22 | 30 | 30 | 25 |
| Quinta | 7 | 18 | 13 | 18 | 20 | 13 | 3 | 23 |
| | 8 | 5 | 23 | 15 | 17 | 7 | 37 | 12 |
| Sesta | 9 | | 2 | 10 | 10 | | 2 | 13 |
| | 10 | | | 7 | 7 | | | 2 |

(segue)

LESSICO

| Fasce di punteggio | Punteggi | [H] | [X] | [Y] | [Z] | [J] | [K] | [W] |
|---|---|---|---|---|---|---|---|---|
| Prima | 0 | | | | | | | |
| Seconda | 1 | | | | | | | |
| | 2 | 2 | | | | 2 | | |
| Terza | 3 | 3 | 2 | 2 | 2 | 5 | 3 | 3 |
| | 4 | 2 | 5 | 8 | 5 | 7 | 3 | 3 |
| Quarta | 5 | 13 | 10 | 18 | 8 | 10 | 10 | 17 |
| | 6 | 27 | 13 | 8 | 22 | 17 | 32 | 20 |
| Quinta | 7 | 32 | 32 | 23 | 7 | 35 | 12 | 22 |
| | 8 | 18 | 22 | 23 | 25 | 12 | 27 | 18 |
| Sesta | 9 | 2 | 12 | 8 | 10 | 10 | 12 | 3 |
| | 10 | 2 | 3 | 8 | 22 | 3 | 2 | 3 |

GRAMMATICA

| Fasce di punteggio | Punteggi | [H] | [X] | [Y] | [Z] | [J] | [K] | [W] |
|---|---|---|---|---|---|---|---|---|
| Prima | 0 | | | | | | | |
| Seconda | 1 | | | | | | | |
| | 2 | | | | | | | |
| Terza | 3 | | | | 2 | 2 | | |
| | 4 | 8 | 13 | 15 | 3 | 7 | 8 | 12 |
| Quarta | 5 | 20 | 13 | 10 | 5 | 23 | 22 | 25 |
| | 6 | 23 | 13 | 12 | 38 | 13 | 22 | 17 |
| Quinta | 7 | 33 | 30 | 32 | 8 | 18 | 12 | 27 |
| | 8 | 13 | 22 | 13 | 33 | 22 | 28 | 15 |
| Sesta | 9 | 2 | 8 | 12 | 8 | 13 | 8 | 5 |
| | 10 | | | 7 | 2 | 2 | | |

Nella tabella 21b, *infra*, abbiamo raggruppato i dati in un unico prospetto, disposto in orizzontale.[44]

Nella maggior parte dei casi, coppie e terne si concentrano su due fasce di punteggio: la quarta |5÷6| e la quinta |7÷8|, riquadrettate in rosso nella tabella. Si manifesta cioè una *restrizione di intervallo*, evidente in particolare nelle valutazioni di [**H**], [**X**] e [**K**] (riquadri in azzurro). Considerando un *range* leggermente più

[44] Le abbreviazioni delle componenti sono le seguenti: ct= *contenuto*; cc=*coerenza e coesione*; lx=*lessico*; gr=*grammatica.*

ampio, e cioè $|4 \div 9|$, il 99.9% delle osservazioni di [**H**], [**X**] e [**K**] ricade all'interno di esso.

Tabella 21b. *Percentuali dei punteggi assegnati. La restrizione di intervallo*

| | [H] | | | | [X] | | | | [Y] | | | | [Z] | | | | [J] | | | | [K] | | | | [W] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR |
| 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | | 2 | | | | | | | | | | | | | | | | 2 | | | | | | | | | |
| 3 | | | 3 | | | | 2 | | | | 2 | | | 2 | 2 | 2 | 2 | 2 | 5 | 2 | | | 3 | | | | 3 | |
| 4 | 5 | 13 | 2 | 8 | 3 | 13 | 15 | 13 | 3 | 18 | 8 | 15 | 5 | 12 | 5 | 3 | 7 | 12 | 7 | 7 | 8 | 15 | 3 | 8 | 5 | 7 | 3 | 12 |
| 5 | 18 | 33 | 13 | 20 | 8 | 13 | 10 | 13 | 20 | 12 | 18 | 10 | 5 | 12 | 8 | 5 | 8 | 33 | 10 | 23 | 7 | 13 | 10 | 22 | 8 | 18 | 17 | 25 |
| 6 | 10 | 30 | 27 | 23 | 22 | 35 | 13 | 13 | 18 | 20 | 8 | 12 | 20 | 22 | 22 | 38 | 18 | 30 | 17 | 13 | 42 | 30 | 32 | 22 | 22 | 25 | 20 | 17 |
| 7 | 33 | 18 | 32 | 33 | 12 | 13 | 32 | 30 | 18 | 18 | 23 | 32 | 12 | 20 | 7 | 8 | 17 | 13 | 35 | 18 | 5 | 3 | 12 | 12 | 27 | 23 | 22 | 27 |
| 8 | 20 | 5 | 18 | 13 | 28 | 23 | 22 | 22 | 23 | 15 | 23 | 13 | 30 | 17 | 25 | 33 | 23 | 7 | 12 | 22 | 28 | 37 | 27 | 28 | 18 | 12 | 18 | 15 |
| 9 | 10 | | 2 | 2 | 25 | 2 | 12 | 8 | 7 | 10 | 8 | 12 | 7 | 10 | 10 | 8 | 15 | | 10 | 13 | 10 | 2 | 12 | 8 | 13 | 13 | 3 | 5 |
| 10 | 3 | | 2 | | | 2 | 3 | | 10 | 7 | 8 | 7 | 22 | 7 | 22 | 2 | 10 | | 3 | 2 | | | 2 | | 7 | 2 | 3 | |

Nella tabella 21c, *infra*, abbiamo evidenziato i livelli maggiormente usati all'interno di ogni fascia della scala. Emerge un fenomeno di *localizzazione*, ovvero di *selezione prevalente di alcuni livelli delle fasce*. Tale concentrazione è contrassegnata con un fondo arancio (il caso in cui si ravvisa una percentuale di punteggi identica a quella del livello contiguo è marcato in giallo).

In sostanza, alcuni valutatori tendono a privilegiare un certo livello all'interno delle fasce (e ciò avviene a scapito di una fine discriminazione):

- [**Z**] manifesta la preferenza per i livelli superiori intrafascia;
- [**K**] manifesta la preferenza per i livelli superiori nelle fasce |3÷4|, |5÷6| e |7÷8| e per il livello inferiore nella fascia |9÷10|;
- [**W**] manifesta la preferenza per i livelli superiori nelle fasce |3÷4| e |5÷6| e per i livelli inferiori nelle fasce |7÷8| e |9÷10|.

Tabella 21c. *Percentuali dei punteggi assegnati. La localizzazione*

| | [H] | | | | [X] | | | | [Y] | | | | [Z] | | | | [J] | | | | [K] | | | | [W] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GM | CT | CC | LX | GR | CT | CC | LX | GR | CNT | CC | LX | GR |
| 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | | 2 | | | | | | | | | | | | | | | | 2 | | | | | | | | | |
| 3 | | | 3 | | | | 2 | | | | 2 | | | 2 | 2 | 2 | 2 | 2 | 5 | 2 | | | 3 | | | | 3 | |
| 4 | 5 | 13 | 2 | 8 | 3 | 13 | 15 | 13 | 3 | 18 | 8 | 15 | 5 | 12 | 5 | 3 | 7 | 12 | 7 | 7 | 8 | 15 | 3 | 8 | 5 | 7 | 3 | 12 |
| 5 | 18 | 33 | 13 | 20 | 8 | 13 | 10 | 13 | 20 | 12 | 18 | 10 | 5 | 12 | 8 | 5 | 8 | 33 | 10 | 23 | 7 | 13 | 10 | 22 | 8 | 18 | 17 | 25 |
| 6 | 10 | 30 | 27 | 23 | 22 | 35 | 13 | 13 | 18 | 20 | 8 | 12 | 20 | 22 | 22 | 38 | 18 | 30 | 17 | 13 | 42 | 30 | 32 | 22 | 22 | 25 | 20 | 17 |
| 7 | 33 | 18 | 32 | 33 | 12 | 13 | 32 | 30 | 18 | 18 | 23 | 32 | 12 | 20 | 7 | 8 | 17 | 13 | 35 | 18 | 5 | 3 | 12 | 12 | 27 | 23 | 22 | 27 |
| 8 | 20 | 5 | 18 | 13 | 28 | 23 | 22 | 22 | 23 | 15 | 23 | 13 | 30 | 17 | 25 | 33 | 23 | 7 | 12 | 22 | 28 | 37 | 27 | 28 | 18 | 12 | 18 | 15 |
| 9 | 10 | | 2 | 2 | 25 | 2 | 12 | 8 | 7 | 10 | 8 | 12 | 7 | 10 | 10 | 8 | 15 | | 10 | 13 | 10 | 2 | 12 | 8 | 13 | 13 | 3 | 5 |
| 10 | 3 | | 2 | | | 2 | 3 | | 10 | 7 | 8 | 7 | 22 | 7 | 22 | 2 | 10 | | 3 | 2 | | | 2 | | 7 | 2 | 3 | |

In aggiunta, nella tab. 21d, *infra*, abbiamo messo in risalto due ulteriori aspetti:

- il valutatore [**J**] presenta una *tendenza centrale* in corrispondenza della valutazione della *coerenza e coesione* (riquadro celeste);
- in [**X**] e in [**Z**] si nota una duplicazione dei punteggi in riferimento a componenti distinte (per [**X**]: *lessico* e *grammatica* per [**Z**]: *lessico* e *contenuto*) (riquadri gialli). Si tratta di un *effetto alone*, in virtù del quale il voto dato a una componente viene esteso a una seconda componente.[45]

[45] L'attribuzione, da parte di un valutatore, dello stesso voto a componenti diverse costituisce un caso di effetto alone (cfr. § 1). Può essere che il candidato abbia difficoltà a distinguere le componenti oppure che giudichi una saliente e l'altra dipendente o ancora che il giudizio sia sottoposto ad una impressione generale (Fisicaro, Lance 1990). Qualsiasi sia la causa, "examinees have less independent opportunities to demonstrate their proficiency" (Eckes 2015: 87).

Tabella 21d. *Percentuali dei punteggi assegnati. Effetto alone e tendenza centrale*

| | [H] | | | | | [X] | | | | [Y] | | | | [Z] | | | | [J] | | | | [K] | | | | [W] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | CT | CC | LX | GR | | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GR | CT | CC | LX | GM | CT | CC | LX | GR | CT | CC | LX | GR |
| 0 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 1 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
| 2 | | | 2 | | | | | | | | | | | | | | | | | 2 | | | | | | | | | |
| 3 | | | 3 | | | | | 2 | | | | 2 | | | 2 | | 2 | 2 | 2 | 5 | 2 | | | 3 | | | | 3 | |
| 4 | 5 | 13 | 2 | 8 | | 3 | 13 | 15 | 13 | 3 | 18 | 8 | 15 | 5 | 12 | 5 | 3 | 7 | 12 | 7 | 7 | 8 | 15 | 3 | 8 | 5 | 7 | 3 | 12 |
| 5 | 18 | 33 | 13 | 20 | | 8 | 13 | 10 | 13 | 20 | 12 | 18 | 10 | 5 | 12 | 8 | 5 | 8 | 33 | 10 | 23 | 7 | 13 | 10 | 22 | 8 | 18 | 17 | 25 |
| 6 | 10 | 30 | 27 | 23 | | 22 | 35 | 13 | 13 | 18 | 20 | 8 | 12 | 20 | 22 | 22 | 38 | 18 | 30 | 17 | 13 | 42 | 30 | 32 | 22 | 22 | 25 | 20 | 17 |
| 7 | 33 | 18 | 32 | 33 | | 12 | 13 | 32 | 30 | 18 | 18 | 23 | 32 | 12 | 20 | 7 | 8 | 17 | 13 | 35 | 18 | 5 | 3 | 12 | 12 | 27 | 23 | 22 | 27 |
| 8 | 20 | 5 | 18 | 13 | | 28 | 23 | 22 | 22 | 23 | 15 | 23 | 13 | 30 | 17 | 25 | 33 | 23 | 7 | 12 | 22 | 28 | 37 | 27 | 28 | 18 | 12 | 18 | 15 |
| 9 | 10 | | 2 | 2 | | 25 | 2 | 12 | 8 | 7 | 10 | 8 | 12 | 7 | 10 | 10 | 8 | 15 | | 10 | 13 | 10 | 2 | 12 | 8 | 13 | 13 | 3 | 5 |
| 10 | 3 | | 2 | | | | 2 | 3 | | 10 | 7 | 8 | 7 | 22 | 7 | 22 | 2 | 10 | | 3 | 2 | | | 2 | | 7 | 2 | 3 | |

### 10.2.8. Il profilo dei valutatori

Arrivati a questo punto, incrociando i dati raccolti e le osservazioni effettuate fino ad ora, siamo in grado di stilare il profilo dei valutatori.

Nelle pagine che seguono ciascun profilo è riportato ad una nuova pagina; al testo si accompagnano delle raccomandazioni rivolte al Certificatore (riquadri con fondo giallo).

In un paragrafo conclusivo presentiamo in un unico prospetto le caratteristiche del gruppo (tab. 32, *infra*), in modo che il lettore possa rendersi conto, con un colpo d'occhio, di quanto siano diffusi i *bias*.

#### 10.2.8.1. Il valutatore [H]

Il valutatore [**H**] presenta il valore più equilibrato in termini di *generosità* rispetto all'intero gruppo (+0.62 *logit*, tab. 6, *supra*).

Dall'analisi dell'*interazione valutatore-componenti* (tabb. 18, 19 *supra*) sono emerse violazioni delle statistiche di conformità in riferimento al *contenuto*, legate a diverse risposte inattese (tab. 20, *supra*).

Vi è una tendenza a un *uso ristretto della scala*; essa spiega l'*iperconformità* nella valutazione della componente della *grammatica* (*infit* e *outfit*=0.40), riportata nelle tabb. 18, 19, *supra*.

Si notino, nella tab. 22, *infra*, le percentuali relativamente alte (>30%) assegnate ad alcuni punteggi della scala (in particolar modo al punteggio 7): abbiamo dei veri e propri "poli di attrazione" in fase di attribuzione del giudizio.

Tabella 22. *Percentuali dei punteggi assegnati da* [**H**]

| | **cont** | **cc** | **lex** | **gram** |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | 2 | |
| 3 | | | 3 | |
| 4 | 5 | 13 | 2 | 8 |
| 5 | 18 | 33 | 13 | 20 |
| 6 | 10 | 30 | 27 | 23 |
| 7 | 33 | 18 | 32 | 33 |
| 8 | 20 | 5 | 18 | 13 |
| 9 | 10 | | 2 | 2 |
| 10 | 3 | | 2 | |

Si notino, ancora, nella tab. 23, *infra*, le percentuali altissime relative alle attribuzioni di giudizio nelle fasce in cui si concentrano i *poli di attrazione*: |5÷6| e |7÷8|.

Tabella 23. *Percentuali dei punteggi assegnati da* [**H**] *nelle fasce* |5÷6| e |7÷8|

| **fascia** | **punteggio** | **cont** | **cc** | **lex** | **gram** |
|---|---|---|---|---|---|
| quarta | 5 | 18 | 33 | 13 | 20 |
| | 6 | 10 | 30 | 27 | 23 |
| quinta | 7 | 33 | 18 | 32 | 33 |
| | 8 | 20 | 5 | 18 | 13 |
| | | **81** | **86** | **90** | **89** |

*Raccomandazioni*: in sede di aggiornamento, è bene che il Certificatore riveda, assieme al valutatore, l'applicazione dei descrittori della categoria del *contenuto*.
Inoltre, è necessario che al valutatore sia resa nota la sua *tendenza all'uso ristretto della scala*. Al fine di sollecitare il *rater* all'uso dell'intera scala, si consiglia di fargli prendere visione di elaborati la valutazione delle cui componenti si colloca ai valori estremi.

### 10.2.8.2. Il valutatore [X]

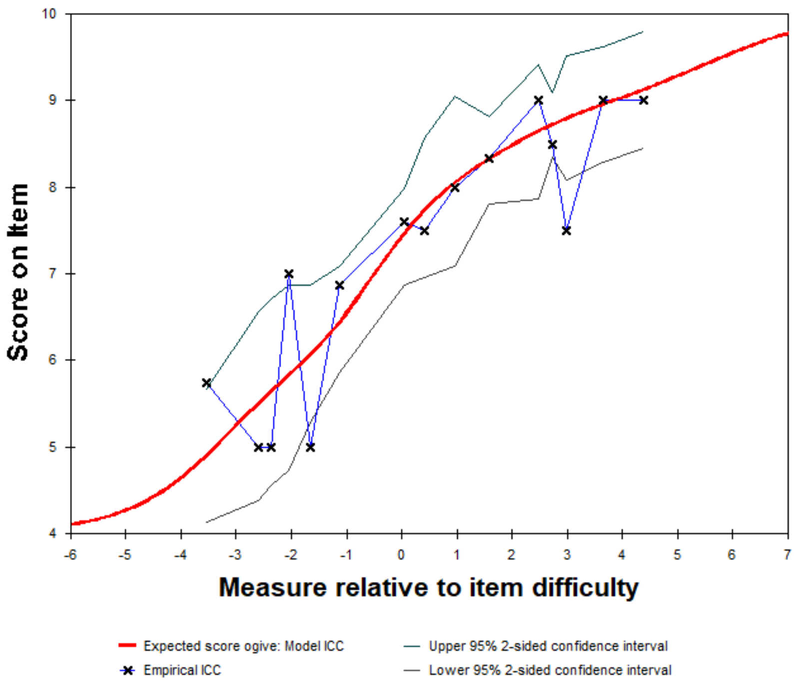
Il valutatore [**X**] è generoso (+1.28 *logit*).

Dall'analisi dell'*interazione valutatore-componenti* (tab. 18, *supra*) emergono valori di conformità statistica in riferimento al *contenuto* (*infit*: 1.40; *outfit* 1.50) che indicano una certa *erraticità* (cfr. anche tab. 20, *supra*).

Al fine di illustrare tale fenomeno, ci serviamo della figura 7, *infra*. L'asse delle ascisse corrisponde alla scala *logit*; l'asse delle ordinate, invece, corrisponde alla porzione di scala decimale usata dal *rater*. La linea azzurra congiunge i punteggi assegnati. La linea rossa indica, invece, lo sviluppo delle misure attese; essa è compresa tra due linee spezzate all'interno delle quali si raccoglie il 95% delle osservazioni nell'ipotesi di una distribuzione normale. Oltre questa fascia (definita *di controllo*, "confidence interval") si danno osservazioni *outliers*, inaspettate.[46] Si notino le diverse oscillazioni, alcune delle quali oltrepassano la fascia di controllo: tale è appunto l'evidenza dell'aleatorietà del giudizio.

[46] I valori sono leggermente diversi da quelli presenti nella tabella delle valutazioni inattese (tab. 20, *supra*), dato che il grafico rappresentato nella fig. 7 è stato ricavato mediante un modello ibrido (cfr. **Appendice 3**); ciononondimeno, la mappa ben rappresenta la distribuzione dei residui.

Figura 7. *Punteggi attribuiti dal valutatore* [**X**], *curva delle valutazioni attese e fascia di controllo in riferimento alla componente del* contenuto




In [**X**], come ravvisato in [**H**] e come si vedrà meglio in [**K**], si riscontra la tendenza ad un *uso ristretto della scala* (fig. 21b, *supra*).

In aggiunta, abbiamo rilevato un *effetto alone*, per via dell'assimilazione dei punteggi assegnati alle componenti di *lessico* e *grammatica* (cfr. tab. 21d, *supra*).

*Raccomandazioni*: il valutatore va avvisato del fatto che il suo giudizio tende ad essere *molto generoso* e che, nelle sue attribuzioni di valore, si evince la tendenza a un *uso ristretto della scala*.
La duplicazione dei voti tra le categorie di *lessico* e *grammatica* (*effetto alone*) potrebbe essere evitata, assicurandosi che la

differenza tra le componenti sia chiara agli occhi del valutatore. Gli si può chiedere di soffermarsi sui descrittori di ciascuna categoria, prima di passare a quelli dell'altra. Eventualmente, onde scongiurare che l'*effetto alone* sia legato ad un *errore di prossimità* (la vicinanza delle categorie di *lessico* e *grammatica* nella griglia), al valutatore può essere suggerito di non seguire l'ordine delle categorie durante l'attribuzione del giudizio: pur cominciando coll'esaminare il *contenuto*, può passare al *lessico*, per poi analizzare la *coesione e coerenza* e, da ultimo, giungere alla *grammatica*. Questo espediente potrebbe scongiurare una valutazione in automatico della *grammatica* a partire dal *lessico*, o viceversa.

#### 10.2.8.3. Il valutatore [Y]

Il valutatore [**Y**] è molto generoso (+1.31 *logit*). Non si ravvisano altre criticità.

*Raccomandazione*: il valutatore va avvisato del *disallineamento per eccesso*.

### 10.2.8.4. Il valutatore [Z]

Il valutatore [**Z**] si distacca per la *generosità estrema* (+1.84); opera, rispetto ai colleghi, una sovrastima della competenza su tutte e quattro le componenti. In particolare, il *disallineamento per eccesso* riguarda le componenti di *contenuto* e *lessico*: a quasi 1/4 del campione viene assegnato il punteggio massimo (10) (si vedano i cerchi verdi nella tab. 24, *infra*).

Tabella 24. *Percentuali dei punteggi assegnati da* [**Z**]. *Aspetti del disallineamento*

| | cont | cc | lex | gram |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | | |
| 3 | | 2 | 2 | 2 |
| 4 | 5 | 12 | 5 | 3 |
| 5 | 5 | 12 | 8 | 5 |
| 6 | 20 | 22 | 22 | 38 |
| 7 | 12 | 20 | 7 | 8 |
| 8 | 30 | 17 | 25 | 33 |
| 9 | 7 | 10 | 10 | 8 |
| 10 | 22 | 7 | 22 | 2 |

Il valutatore dimostra, inoltre, un *comportamento incoerente* (tab. 9, *supra*; valori *infit* e *outfit borderline* ed errore di misurazione notevole), con particolare riferimento alla componente del *contenuto* e a quella del *lessico* (cfr. tabb. 18, 20, *supra*).

A riprova del profilo erratico del *rater*, nelle tabelle a seguire (25, 26), si può notare come, nella valutazione delle componenti di *contenuto* e *lessico*, non abbiamo un incremento monotonico dei valori *logit* corrispondenti ai punteggi assegnati: in un paio di occasioni l'attribuzione di punteggio da parte del valutatore non riflette l'incremento della competenza (valori asteriscati). Questo avanzamento atipico manifesta un uso inappropriato della griglia (si notino, peraltro, i valori *oufit* fuori norma; i.e. >2, evidenziati con fondo giallo).

Tabella 25. *Uso della scala da parte del valutatore* [**Z**] *nella valutazione della componente del* contenuto

Legenda

| | |
|---|---|
| **punt=** | punteggi della griglia di valutazione |
| **oc=** | occorrenze dell'assegnazione di punteggio |
| **%=** | conversione in percentuale delle occorrenze |
| **L(o)=** | valore in *logit* del punteggio assegnato (*osservato*) |
| **L(a)=** | valore in *logit* del punteggio *atteso* dal modello |
| **Scarto=** | differenza tra il valore in *logit* del punteggio assegnato e il valore in *logit* del punteggio atteso |
| **Outfit=** | valore dell'indice di conformità |

| **punt** | **oc** | **%** | **L(o)** | **L(a)** | **Scarto (Lo-La)** | **Outfit** |
|---|---|---|---|---|---|---|
| 4 | 3 | 5 | **-1.88** | -2.21 | +0.33 | 2.0 |
| 5 | 3 | 5 | **-0.88** | -1.70 | +0.82 | 2.4 |
| 6 | 12 | 20 | **-0.96*** | -0.91 | -0.05 | 1.4 |
| 7 | 7 | 12 | **1.18** | 0.04 | +1.14 | 5.1 |
| 8 | 18 | 30 | **1.05*** | 1.24 | -0.19 | 1.6 |
| 9 | 4 | 7 | **1.90** | 2.71 | -0.81 | 2.0 |
| 10 | 13 | 22 | **3.52** | 3.84 | -0.32 | 1.3 |

Tabella 26. *Uso della scala da parte del valutatore* [**Z**] *nella valutazione della componente del* lessico

| **Punt** | **oc** | **%** | **L(o)** | **L(a)** | **Scarto (Lo-La)** | **Outfit** |
|---|---|---|---|---|---|---|
| 3 | 1 | 2 | **-2.65** | -2.59 | -0.07 | 0.6 |
| 4 | 3 | 5 | **-2.13** | -2.09 | -0.04 | 0.7 |
| 15 | 5 | 8 | **-0.03** | -1.25 | +1.22 | 2.4 |
| 6 | 13 | 22 | **-0.78*** | -0.31 | -0.47 | 1.0 |
| 7 | 4 | 7 | **1.96** | 0.64 | +1.32 | 2.1 |
| 8 | 15 | 25 | **1.43*** | 1.64 | -0.21 | 1.9 |
| 9 | 6 | 10 | **2.02** | 2.84 | -0.82 | 1.4 |
| 10 | 13 | 22 | **3.91** | 3.69 | +.022 | 0.8 |

La questione viene riflessa dall'andamento delle curve di probabilità. Nelle figg. 8 e 9, *infra*, l'abilità di un candidato, in termini rispettivamente di *contenuto* e di *lessico*, è messa in relazione con la probabilità di ricevere un determinato punteggio. Sull'asse delle ascisse sono riportati i valori della scala *logit*, mentre sull'asse delle ordinate è riportato il valore di probabilità che oscilla tra 0 e 1 (in termini percentuali, tra 0% e 100%). Abbiamo evidenziato nelle figure, mediante delle linee rosse, l'intervallo di competenza dei candidati: da -3.69 a +3.21 *logit*.

Nella fig. 8 si noti la definizione chiara e netta delle curve corrispondenti ai punteggi 6 e 8, con ogive ampie e distese (a conferma, peraltro, della *localizzazione*, illustrata nella tab. 21c, *supra*).[47] Le curve corrispondenti ai valori 5, 7, e 9, per contro, non emergono: non hanno un profilo distinto. Si presti attenzione pure all'anomala progressione in termini di probabilità. Nella figura abbiamo evidenziato con una linea tratteggiata verticale (in

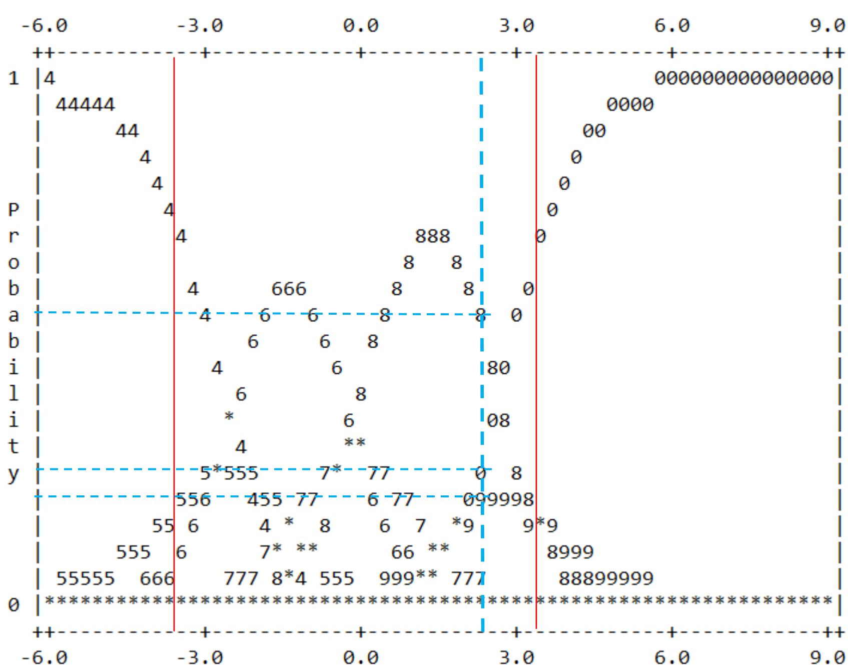
[47] Tali addensamenti coinvolgono in particolare la valutazione della *componente grammaticale*, con il 68% dei giudizi distribuiti sui livelli 6 e 8 (cfr. tab. 21c, *supra*).

azzurro) l'abilità pari a +2.50 *logit* di un ipotetico candidato. Ebbene, egli ha all'incirca

- il 25 % di probabilità di ricevere un 9;
- il 30 % di probabilità di ricevere un 10;
- il 55 % di probabilità di ricevere un 8.

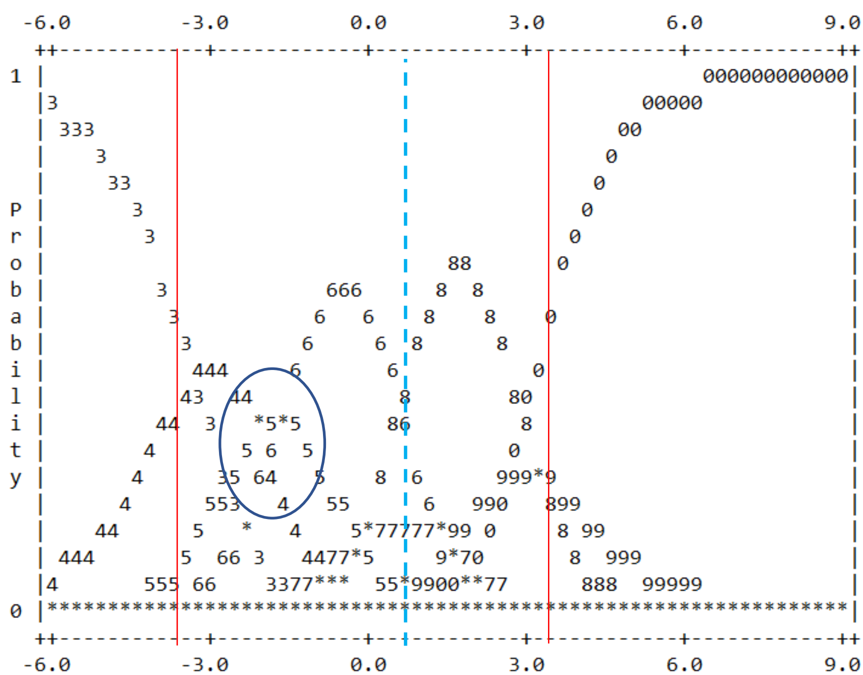
Questo "profilo a U" delle probabilità (9-10-8) riflette un'erraticità nell'attribuzione dei punteggi.

Figura 8. *Curve di probabilità in riferimento alla valutazione della componente del* contenuto *da parte del valutatore* [**Z**]




Passiamo alla figura 9, *infra*, relativa alla componente del *lessico*. Benché la *polarizzazione dei giudizi* sia meno marcata (l'ogiva relativa alla curva di probabilità del punteggio 8 è meno distesa), si registra, alla pari, un incremento anomalo in merito alla probabilità di attribuzione di punteggio [per esempio, un candidato con un'abilità approssimativa di 0.60 *logit* (linea verticale tratteggiata in azzurro) ha una probabilità bassa di ricevere un voto 7, una probabilità maggiore di ricevere un voto 8 oppure, alla pari, un voto 6]. In aggiunta, vi è una crasi delle curve di probabilità in riferimento ai punteggi 4, 5, 6 nell'intervallo |(-2), (-3)| *logit* (evidenziata con il cerchio), a indicare la difficoltà del valutatore ad operare una fine discriminazione in riferimento a questo livello di competenza (è un livello assai critico, peraltro, poiché in prossimità del punto di taglio, che è 5).

Figura 9. *Curve di probabilità in riferimento alla valutazione della componente del* lessico *da parte del valutatore* [**Z**]



A carico di [**Z**] va imputato, infine, un *effetto alone*, per via dell'assimilazione dei giudizi riferiti alla componente del *contenuto* a quelli riferiti alla componente del *lessico* (cfr. tab. 21c, *supra*, ripresa nella tab. 27, *infra*).

Tabella 27. *Percentuali dei punteggi assegnati da* [**Z**]. *Effetto alone*

| | **cont** | **cc** | **lex** | **gram** |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | | |
| 3 | | 2 | 2 | 2 |
| 4 | 5 | 12 | 5 | 3 |
| 5 | 5 | 12 | 8 | 5 |
| 6 | 20 | 22 | 22 | 38 |
| | | | | |
| 8 | 30 | 17 | 25 | 33 |
| 9 | 7 | 10 | 10 | 8 |
| 10 | 22 | 7 | 22 | 2 |

*Raccomandazioni*: è bene che il Certificatore interpelli il valutatore per accertare le cause delle *valutazioni erratiche* (*il valutatore ha valutato i testi in una condizione di pressione o di stanchezza? I descrittori della griglia gli erano chiari? Ha avuto difficoltà a valutare alcuni testi rispetto ad altri? Prima della griglia in adozione, ne usava un'altra? Se sì, può essere stato influenzato da quella?*) e della *generosità abnorme*.
Inoltre, al valutatore va resa nota la tendenza alla *localizzazione* (preferenza per alcuni punteggi intrafascia su tutte le componenti) e, infine, il fatto che egli *tenda ad assegnare lo stesso voto alla componente del contenuto e a quella del lessico* (*effetto alone*).

### 10.2.8.5. Il valutatore [J]

Al valutatore [J] corrisponde un valore accettabile in termini di generosità (+0.87 *logit*).

Egli presenta tuttavia *un'incoerenza*, statisticamente significativa, in riferimento alla valutazione del *contenuto*; inoltre, ravvisiamo un'*iperconformità*, parimenti statisticamente significativa, in riferimento alla valutazione tanto della *coerenza/coesione* quanto della *grammatica* (tabb. 18, 19, *supra*)

L'*iperdifformità* in riferimento alla valutazione del *contenuto* può essere spiegata alla luce delle numerose valutazioni inattese, di segno opposto (positive e negative), con un *range* considerevole di valori [residui standardizzati | (+3.5)÷(–3.2)|] (cfr. tab. 20, *supra*).

La tabella 28, *infra*, illustra l'uso della scala da parte del valutatore [J] in riferimento alla componente del *contenuto*.

Tabella 28. *Uso della scala da parte del valutatore* [J] *in riferimento alla componente del* contenuto

Legenda

| | |
|---|---|
| **punt=** | punteggi della griglia di valutazione |
| **oc=** | occorrenze dell'assegnazione di punteggio |
| **%=** | conversione in percentuale delle occorrenze |
| **L(o)=** | valore in *logit* del punteggio assegnato (*osservato*) |
| **L(a)=** | valore in *logit* del punteggio *atteso* dal modello |
| **scarto=** | differenza tra il valore in *logit* del punteggio assegnato e il valore in *logit* del punteggio atteso |
| **Outfit=** | valore dell'indice di conformità |

| **punt** | **oc** | **%** | **L(o)** | **L(a)** | **scarto (Lo-La)** | **Outfit** |
|---|---|---|---|---|---|---|
| 3 | 1 | 2 | **-1.27** | -3.61 | +2.34 | 2.4 |
| 4 | 4 | 7 | **-2.59*** | -2.77 | +0.18 | 0.3 |
| 5 | 5 | 8 | **-1.23** | -1.68 | +0.45 | 0.5 |
| 6 | 11 | 18 | **-0.33** | -0.60 | +0.27 | 2.8 |
| 7 | 19 | 17 | **-0.36** | 0.50 | -0.86 | 3.1 |
| 8 | 14 | 23 | **1.44** | 1.73 | -0.29 | 0.7 |
| 9 | 9 | 15 | **2.97** | 2.67 | +0.30 | 1.6 |
| 10 | 6 | 10 | **3.42** | 3.14 | +0.28 | 1.6 |

Si notino

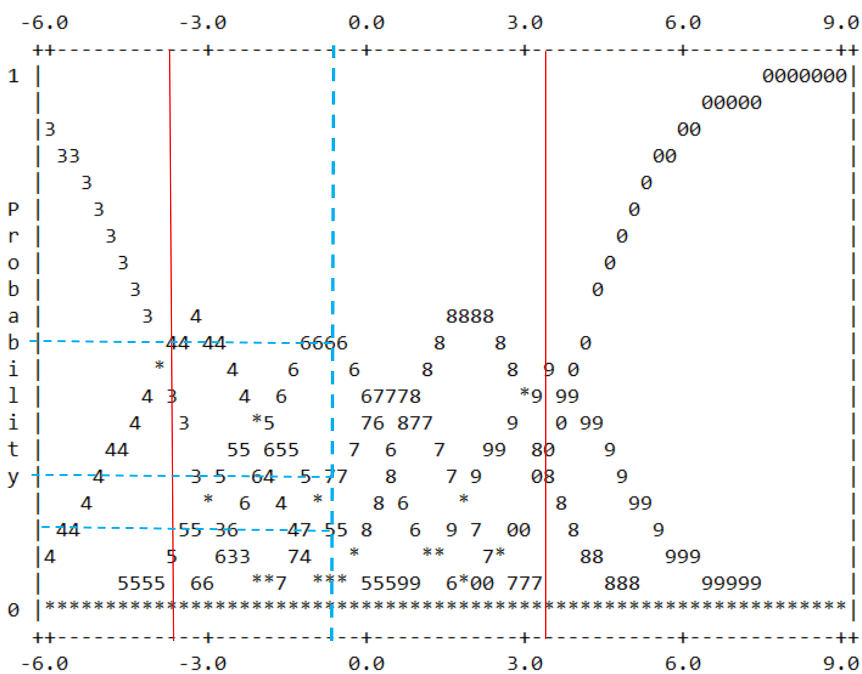
- la violazione dell'incremento monotonico dei valori *logit* (contrassegnata da asterisco);
- i valori *logit* molto prossimi tra loro dei punteggi 6 e 7 (riquadrettati in rosso), a evidenziare il fatto che i due livelli tendono ad essere assimilati;
- i valori *outfit* oltre la soglia (evidenziati in giallo).

La rappresentazione delle curve delle probabilità (fig. 10, *infra*), inoltre, ci avvisa di progressioni anomale. Un ipotetico candidato con un'abilità pari a –0.70 *logit* (linea azzurra tratteggiata) ha all'incirca

- il 15% di probabilità di ricevere un voto pari a 5;
- il 25% di probabilità di ricevere un voto pari a 7;
- 50% di probabilità di ricevere un voto pari a 6.

Come affermato in precedenza, l'andamento irregolare (5-7-6) riflette l'*erraticità del giudizio*.

Figura 10. *Curve di probabilità in riferimento alla valutazione della componente del* contenuto *da parte del valutatore* [**J**]




Da ultimo, come ulteriore *bias*, ricordiamo l'*iperconformità* in riferimento sia alla valutazione della componente della *coesione* e *coerenza* che alla valutazione della componente della *grammatica*. Nel primo caso, l'*iperconformità* può essere spiegata alla luce della *tendenza centrale*, ovvero dell'*uso sovraesteso della fascia della sufficienza* (in tale fascia, si concentra il 63% delle osservazioni; la restrizione del campo è evidente anche se si considera l'intervallo più ampio |4÷7|, a cui corrisponde l'88% delle osservazioni). Nel secondo caso, l'*iperconformità* è collegata da una concentrazione di giudizi (76%) nell'intervallo |5÷8|, ovvero nella *fascia medio-superiore* della scala (cfr. tab. 29, *infra*).

Tabella 29. *Percentuali dei punteggi assegnati da* [J]

| | **cont** | **cc** | **lex** | **gram** |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | 2 | |
| 3 | 2 | 2 | 5 | 2 |
| 4 | 7 | 12 | 7 | 7 |
| 5 | 8 | 33 | 10 | 23 |
| 6 | 18 | 30 | 17 | 13 |
| 7 | 17 | 13 | 35 | 18 |
| 8 | 23 | 7 | 12 | 22 |
| 9 | 15 | | 10 | 13 |
| 10 | 10 | | 3 | 2 |

*Raccomandazioni*: il *disallineamento per eccesso* è contenuto. Destano invece preoccupazione l'*erraticità* in riferimento alla componente del *contenuto* e l'*iperconformità* dei giudizi relativi alla componente della *coesione e coerenza* e della componente della *grammatica*. In tutte e tre i casi abbiamo una significatività statistica dei dati. È cioè assai probabile che, alle prese con la valutazione di nuovi elaborati, [J] continuerà a dimostrare un comportamento poco coerente nella valutazione del *contenuto* ed estremamente controllato nella valutazione delle componenti della *coesione e coerenza* e della *grammatica*. Le cause della limitazione del *range* dei giudizi (che gravitano attorno alla *sufficienza*, nel caso della componente della *coesione e coerenza* [tendenza centrale], e a un livello leggermente più alto, invece, nel caso della componente della *grammatica* [tendenza medio-superiore]) dovrebbero essere indagate: *il valutatore ha sperimentato una certa difficoltà ad interpretare i descrittori della scala corrispondenti ai livelli estremi? Ha bisogno di prendere visione di esempi di testi le cui caratteristiche corrispondono ai livelli estremi?*

### 10.2.8.6. Il valutatore [K]

Il valutatore [**K**] ha un grado di generosità attorno all'1 *logit* (+1.01): una situazione, dicevamo, *borderline* (cfr. tab. 7, *supra*), e quindi di scarso rilievo.

Benché dall'analisi dell'*interazione valutatore-componente* non emergano violazioni alle statistiche di conformità in relazione al *contenuto* (tabb. 18, 19, *supra*), il valutatore ha registrato diverse risposte inattese, con residui di segno opposto (tab. 20, *supra*).

*L'analisi dell'interazione valutatore-componente* ha inoltre rilevato un doppio problema di *iperconformità*, in relazione sia alla valutazione del *lessico* che a quella della *grammatica* (tabb. 18, 19, *supra*). L'*iperconformità* può essere letta alla luce di una generale tendenza a *un uso ristretto della scala* (alla pari di [**H**] e di [**X**]; cfr. tab. 21b, *supra*), con una polarità altissima dei giudizi nei livelli 6 e 8 (come in [**Z**]).

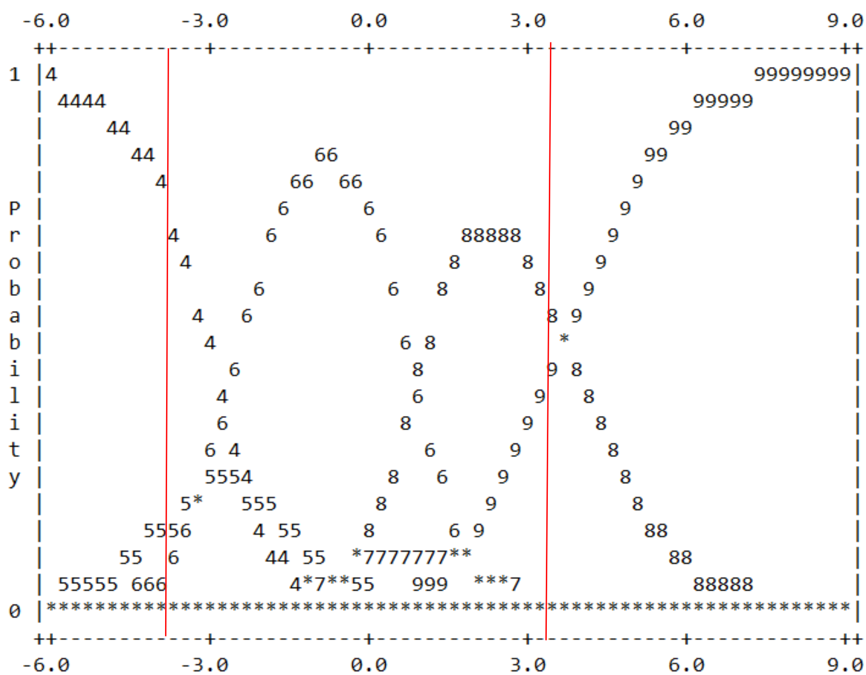
La *localizzazione dei giudizi*, ovvero la loro concentrazione in punti specifici della scala, presentata nella tabella 21c (*supra*), viene ripresa nella tabella 30, *infra*. Le assegnazioni di punteggio superiori al 30% sono cerchiate in rosso; i valori appena al di sotto della soglia del 30% sono cerchiati in azzurro.

Tabella 30. *Percentuali dei punteggi assegnati da* [**K**]

| | **cont** | **cc** | **lex** | **gram** |
|---|---|---|---|---|
| 0 | | | | |
| 1 | | | | |
| 2 | | | | |
| 3 | | | 3 | |
| 4 | 8 | 15 | 3 | 8 |
| 5 | 7 | 13 | 10 | 22 |
| 6 | 42 | 30 | 32 | 22 |
| 7 | 5 | 3 | 12 | 12 |
| 8 | 28 | 37 | 27 | 28 |
| 9 | 10 | 2 | 12 | 8 |
| 10 | | | 2 | |

La *localizzazione delle attribuzioni* è apprezzabile anche considerando il profilo delle curve di probabilità. Si osservino, per esempio, le curve di probabilità relative al *contenuto* (fig. 11, *infra*). Le creste delle curve corrispondenti ai punteggi 6 e 8 sono molto distanziate, i picchi sono alti e le ogive prodotte dall'intersezione sono molto ampie: oltre 2/3 dei candidati, in effetti, sono raggruppati in questi due livelli.

Figura 11. *Curve di probabilità in riferimento alla valutazione del* contenuto *da parte del valutatore* [**K**]




L'*uso ristretto* potrebbe essere spiegato per via della volontà del valutatore di muoversi in una zona di relativa sicurezza, onde evitare di distaccarsi dai giudizi formulati dai colleghi. In effetti, in generale, la percentuale di accordo tra i giudizi del valutatore con quella degli altri valutatori (cfr. tab. 11, *supra*) è la più alta del gruppo (39.5). Non solo: tale valore è superiore a quello che il sistema si prefigura sulla base delle caratteristiche di [**K**], che è pari a 33.2. Abbiamo cioè uno scarto positivo molto alto tra l'accordo osservato e l'accordo atteso (+6.3), che corrobora l'ipotesi di una tendenza alla conformità.[48]

48 Scrive R. Green (2013: 224): "When the observed figures [relative all'accordo effettivo, ndt] are higher than the expected ones, the raters

*Raccomandazioni*: la *generosità* del valutatore è appena al di sopra della soglia di tolleranza; non desta, perciò, una preoccupazione eccessiva.
Vero e proprio *bias* è invece l'*iperconformità* relativa alla valutazione delle componenti di *lessico* e di *grammatica*, collegata ad un *uso ristretto della scala*. A ciò si accompagna una spiccata *localizzazione dei giudizi*, con una concentrazione nei livelli 6 e 8.
Anche in questo caso, è bene che il Certificatore condivida con il *rater* elaborati le valutazioni delle cui componenti attingano ai valori estremi.

may be considered as being too predictable (they are rating in a clone-like fashion rather than as independent experts)".

### 10.2.8.7. Il valutatore [W]

Il grado di generosità del valutatore [**W**] è di poco al di sopra del valore-soglia (+1.13).

Dall'analisi dell'*interazione valutatore-componenti* (tab. 18, *supra*) emerge un valore *borderline* in termini di *iperconformità* nella valutazione del *lessico* (0.60). Non si tratta, tuttavia, di un dato statisticamente significativo.

Abbiamo ravvisato, infine, una propensione alla *localizzazione* (cfr. tab. 21c, *supra*).

Raccomandazioni: monitorare il comportamento del valutatore per vedere se il *disallineamento positivo* si riduce. Far prestare attenzione alla tendenza alla *localizzazione*.

### 10.2.8.8. Sintesi

Ricomponiamo, infine, in un unico quadro le analisi effettuate (tab. 31, *infra*). In esso il lettore trova rappresentate le criticità che caratterizzano ciascun valutatore secondo un diverso grado: quelle in rosso appaiono più evidenti e/o statisticamente significative, mentre quelle in giallo lo sono di meno. Laddove le deviazioni dal modello sono poco significative o appaiono meno evidenti abbiamo tralasciato di evidenziarle.

Abbiamo specificato, infine, i casi in cui la criticità sia vincolata a una componente in particolare; negli altri il *bias* assume invece un carattere trasversale.

Tabella 31. *Criticità nel comportamento dei valutatori*

| | [H] | [X] | [Y] | [Z] | [J] | [K] | [W] |
|---|---|---|---|---|---|---|---|
| *disallineamento marcato* (per eccesso) | | | | | | | |
| *erraticità* | *contenuto* | *contenuto* | | | *contenuto* | *contenuto* | |
| *restrizione di intervallo* I (*tendenza medio-superiore*) | | | | | *grammatica* | | |
| *restrizione di intervallo* II (*tendenza centrale*) | | | | | *coesione e coerenza* | | |
| *localizzazione* (concentrazione dei giudizi in livelli intrafascia) | | | | | | | |
| *effetto alone* (replica dei voti su componenti distinte) | | | | | | | |
| **N. tot. *bias*** | **2** | **3** | **1** | **4** | **3** | **3** | **2** |

Emerge, in primo luogo, il *disallineamento per eccesso*: la maggior parte dei valutatori presenta valori fuori norma (4/7).

Va rilevata, poi, un'*erraticità diffusa* (5/7), in gran parte legata alla valutazione del *contenuto* (4/7).

Si ravvisa, ancora, l'*uso ristretto della scala*: in alcuni casi nella fascia medio-alta (3/7), in un caso attorno ai *valori intermedi* (*tendenza centrale*; 1/7).

La *localizzazione dei giudizi*, vale a dire la tendenza a concentrare le valutazioni in alcuni livelli intrafascia, indipendentemente della componente, è un fenomeno ricorrente (3/7).

Si riscontra, infine, un *effetto alone*: un paio di valutatori formula giudizi molto simili su componenti distinte (2/7). Ciò lascia supporre un'inadeguata operativizzazione del costrutto: è probabile che la distinzione tra le due componenti non risulti chiara ai loro occhi.

Con una certa sommarietà, possiamo suddividere il gruppo dei valutatori in quattro sottogruppi, qui riportati in ordine decrescente di affidabilità:

- il primo include il valutatore [**Y**], il cui giudizio è il più attendibile di tutti (è in difetto per la sola *sovrastima della competenza*);
- il secondo concerne il valutatore [**W**], il cui profilo è soddisfacente (è necessario che egli risolva la *sovrastima della competenza* e la *localizzazione*);
- il terzo comprende i valutatori [**H**], [**X**], [**J**] e [**K**], al cui carico sono imputabili diversi *bias*, alcuni dei quali di forte impatto: tutti e quattro presentano una *erraticità* in riferimento alla componente del *contenuto*; il primo, in più, manifesta una fortissima *restrizione di intervallo*; i giudizi del secondo e del terzo, oltre ad essere estremamente *generosi*, sono contraddistinti rispettivamente da un *effetto alone* e da una certa *localizzazione*; il quarto, infine, usa *la scala in modo ristretto* nella valutazione di due componenti (*tendenza centrale*, in un caso, e *tendenza medio-superiore*, nell'altro);

- infine, [**Z**] è il *rater* meno affidabile; le sue attribuzioni di valore sono caratterizzate da *generosità abnorme, erraticità diffusa*, *localizzazione* ed *effetto alone*.

# SEZIONE D

# CONCLUSIONI
# RIFERIMENTI BIBLIOGRAFICI
# APPENDICI

## 11. Conclusioni

Abbiamo definito la *performance* come un'azione complessa. Pressoché infinita è la serie di *performance* a cui è possibile pensare: dalle strategie per accattivare il consenso degli elettori da parte di un politico alla tattica di un giocatore di calcio, dal rapporto di fiducia che un terapeuta è in grado di instaurare con il paziente alla recitazione di un attore, e così via.

Ogni *performance* è valutabile e il giudizio può assumere una doppia funzione/un doppio valore, a seconda che lo stesso giudizio sia riferito

- alla *performance* in sé e per sé,
- alla competenza.

Nel primo caso il giudizio si limita a definire la qualità della *performance*. Si pensi, per esempio, a un brano cantato al Festival di Sanremo o a una composizione in lingua straniera: il posto assegnato in classifica al primo e così il voto assegnato alla seconda non costituiscono un parametro mediante il quale sono definite in assoluto le qualità canore del cantante, in un caso, e l'abilità di scrittura in una lingua altra, nell'altro.

Altre volte, però, a partire dalla *performance*, chi giudica estrapola il grado di competenza del candidato. La commissione di un concorso, per esempio, a partire dalla *performance* del candidato, unitamente alla disamina di alcuni documenti (il cv, il portfolio, ecc.), sancisce se questi è idoneo a svolgere un certo ruolo; in caso positivo definisce il suo posto all'interno di una graduatoria.

In termini di impatto, il caso dell'estrapolazione è più delicato rispetto al giudizio circoscritto alla *performance*: mentre, infatti, un giudizio negativo sulla *performance* in sé e per sé può, al massimo, intaccare l'autostima dell'esaminato, l'estrapolazione di un giudizio negativo ostacola, talora in modo irrimediabile, l'accesso a un percorso di studio o impedisce l'esercizio di una professione. Da qui, la vigilanza estrema a cui sono tenute le istituzioni preposte alla valutazione della competenza. Ogni istituzione che sia incaricata di estrapolare un giudizio sulla competenza a partire dall'analisi della *performance* deve garantire il rispetto in grado massimo de

- la validità/affidabilità della prova (ovvero dei *task* che la compongono);
- l'adeguatezza dello strumento di misurazione, laddove esso sia usato (la *griglia*);
- l'affidabilità di colui che formula un giudizio (il *valutatore*).

Per ovviare, in particolare, al problema dell'inaffidabilità del *rater*, si può ricorrere, abbiamo visto, a diverse azioni. Tra queste, la più complessa è l'indagine *Many-Facet Analysis Measurement* (MFMR), ovvero un'analisi Rasch relativa a una serie estesa di variabili, alcune delle quali politomiche.

In questa sede abbiamo applicato quest'indagine alle valutazioni effettuate da un gruppo di valutatori che operano in seno alla Certificazione di italiano come lingua straniera PLIDA. Le valutazioni hanno avuto per oggetto composizioni di livello B1 redatte da gruppi di candidati in tre sessioni distinte (giugno 2016; agosto 2016; novembre 2017). I risultati ci informano di diversi problemi a carico dei valutatori, variamente distribuiti: *disallineamento positivo, erraticità, uso ristretto della scala* (tendenza centrale e tendenza medio-superiore), *localizzazione*, *effetto alone*. È emersa, peraltro, la necessità, da parte del Certificatore, di verificare la chiarezza, la pertinenza e l'informatività dei descrittori della griglia della produzione scritta di livello B1 relativi alla componente del *contenuto*.

Il nostro studio può costituire la base per una ricalibrazione dei parametri di giudizio da parte dei valutatori. Resi consapevoli dei rispettivi *bias*, ed eventualmente coinvolti in un successivo *re-training*, essi sono messi nelle condizioni di affinare il loro giudizio (cfr. Wigglesworth 1993; Elder *et al.* 2005; Carlsen 2009).

## Riferimenti bibliografici

AIKEN, L. R., *Rating scales and checklists: Evaluating behavior, personality, and attitudes,* John Wiley and Sons, New York, 1996.

BARIVIERA, F.; CARDILLO, G.; DI TOMASSI, A; MENZINGER, C.; VECCHIO, P. (eds.), *Livello B1 PLIDA. Quaderno delle Specifiche*. *II edizione.* Progetto PLIDA, Società Dante Alighieri, Roma, 2021 <https://plida.dante.global/it/preparati-allesame#B1>

BARTLETT, C. J., "What's the difference between valid and invalid halo? Forced-choice measurement without forcing a choice", *Journal of Applied Psychology*, 68, 1983, pp. 218-226.

BOND, T. G.; FOX, C. M., *Applying the Rasch Model: Fundamental measurement in the human science*, Lawrence Erlbaum Associates, Mahwah, NJ, 2007.

CARLSEN, C., *Guarding the guardians: Rating scale and rater training effects on test scores*, VDM, Saarbrucken, 2009.

CHEN, C.; LEE, S.; STEVENSON, H. W., "Response style and cross-cultural comparisons of rating scales among East Asian and North American students", *Psychological Science*, 6, 1995, pp. 170-175.

COFFMAN, W. E.; KURFMAN, D. "A comparison of two methods of reading essay examinations", *American Educational Research Journal*, 5, 1968, pp. 99-107.

COHEN, J., "A coefficient of agreement for nominal scales", *Educational and Psychological Measurement*, 20, 1960, pp. 37-46.

COOPER, W. H., "Ubiquitous halo", *Psychological Bulletin*, 90, 1981, pp. 218-244.

ECKES, T., "Examining rater effects in TestDaF writing and speaking performance assessments: A many-FACETS Rasch analysis", *Language Assessment Quarterly*, 2, 2005, pp. 197-221.

ECKES, T., "Many-FACETS Rasch Measurement", in S. TAKALA (ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment* (*Section* H), Council of Europe/Language Policy Division, Strasbourg, 2009. https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=0900001680667a23.

ECKES, T., "Operational rater types in writing assessment: Linking rater cognition to rater behavior", *Language Assessment Quarterly*, 9, 3, 2012, pp. 270-292.

ECKES, T., *Introduction to Many-FACETS Rasch measurement. Analyzing and evaluating Many-FACETS Rasch measurement*, Peter Lang, Frankfurt am Main, 2015.

ELDER, C.; KNOCH, U.; BARKHUIZEN, G.; von RANDOW, J., "Individual feedback to enhance rater training: Does it work?", *Language Assessment Quarterly*, 2, 2005, pp. 175–196.

ENGELHARD, G., "Monitoring raters in performance assessments", in G. TINDAL; T. M. HALADYNA (eds.), *Large-scale assessment programs for all students: Validity, technical adequacy, and implementation*, Erlbaum, Mahwah, NJ, 2002, pp. 261-287.

ERGUVAN, I. D.; AKSU DUNYA, "Analyzing rater severity in a freshman composition course using Many Facet Rasch measurement", *Language Testing in Asia*, 1, 2020, pp. 1-20.

FISICARO, S. A.; LANCE, C. E., "Implications of three causal models for the measurement of halo error", *Applied Psychological Measurement*, 14, 1990, pp. 419-429.

FISICARO, S. A.; VANCE, R. J., "Comments on the measurement of halo", *Educational and Psychological Measurement*, 54, 1994, pp. 366-371.

GODSHALK, F. I.; SWINEFORD, F.; COFFMAN, W. E., *The measurement of writing ability,* College Entrance Examination Board, New York, 1966.

GREEN, R., *Statistical analysis for language teachers*, Palgrave, Basingtoke, 2013.

GUILFORD, J. P., *Psychometric methods*, McGraw Hill, New York, 1954[2].

HUGHES, D. C.; KEELING, B.; TUCK, B. F., "Effects of achievement expectations and handwriting quality on test scores", *Journal of Educational Measurement*, 20, 1983, 1, pp. 65-70.

KONDO-BROWN, K. "A FACETS analysis of rater bias in measuring Japanese second language writing performance", *Language Testing,* 19, 2002, 1, pp. 3–31. <https://doi.org/10.1191/0265532202lt218oa>

LATHAM, G. P., WEXLEY, K. N.; PURSELL, E. D., "Training managers to minimize rating errors in the observation of behavior", *Journal of Applied Psychology*, 60, 1975, pp. 550-555.

LINACRE, M., "Sample size and item calibration [or person measure] stability", *Rasch Measurement Transactions*, 7, 1994, p. 328.

LINACRE, M., "Rating, judges and fairness", *Rasch Measurement Transactions*, 12, 1998, pp. 630-631.

LINACRE, M., "Investigating rating scale category utility", *Journal of Outcome Measurement*, 3, 1999, pp. 103-122.

LINACRE, M., "Optimizing rating scale category effectiveness", in E. V. SMITH; R. M. SMITH (Eds.), *Introduction to Rasch measurement*, JAM Press, Maple Grove, MN, 2004, pp. 258-278.

LINACRE, M., *FACETSs Rasch model computer program* [Software manual], 2008, <Winsteps.com>.

LINACRE, M., *FACETSs Tutorial 2. 1-40*. (2012), <http://www.winsteps.com/a/ftutorial2.pdf>

LINACRE, M. *A user's guide to FACETSs: Rasch-model computer programs*, 2014, <http://www.winsteps.com/FACETSs.htm>.

LINACRE, M., "FACETSs - raters information", *Rasch Forum Thread*, 2020, <https://raschforum.boards.net/thread/3321/FACETSs-raters-information>

LINACRE, M., *A user's guide to FACETS Rasch-Model computer programs*, 2023, < https://www.winsteps.com/a/Facets-Manual.pdf>.

LORD, F. M.; NOVICK, M. R., *Statistical theories of mental test scores*, Addison-Wesley, Reading, MA, 1968.

MURPHY, K. R., "Difficulties in the statistical control of halo", *Journal of Applied Psychology*, 67, 1982, pp. 161-164.

McNAMARA, T., *Measuring second language performance*, Longman, London, 1996.

MENDOZA RAMOS, A., "El uso de Many-FACETS Rasch Measurement para examinar la calidad del proceso de corrección de pruebas de desempeño", *Revista mexicana de investigación educativa,* 23, 2018, 77, pp. 597-625.

MURRAY, H. A., *Explorations in personality*, OUP, New York, 1938.

MYFORD, C. M.; WOLFE, E. W., "Detecting and measuring rater effects using many-FACETS Rasch measurement: Part I", *Journal of Applied Measurement*, 4, 2003, pp. 386-422.

MYFORD, C. M.; WOLFE, E. W., "Detecting and measuring rater effects using many-FACETS Rasch measurement: Part II", *Journal of Applied Measurement*, 5, 2004, pp. 189-227.

PALLOTTI, G., "Le ricadute didattiche delle ricerche sull'interlingua", in E. JAFRANCESCO (ed.), *L'acquisizione dell'italiano L2 da parte di immigrati adulti*, Edilingua, Atene, 2004, pp. 43-59.

PRIETO, G.; NIETO, E., "Analysis of rater severity on written expression exam using Many Faceted Rasch Measurement", *Psicológica*, 35, 2014, 2, pp. 385-397.

PULAKOS, E. D.; SCHMITT, N.; OSTROFF, C., "A warning about the use of a standard deviation across dimensions within ratees to measure halo", *Journal of Applied Psychology*, 71, 1986, pp. 29-32.

SHAW, S. D.; WEIR, C. J., *Examining writing. Research and practice in assessing second language writing*, CUP, Cambridge, 2007.

SPINELLI, B., Il "dialogo" tra insegnante e studente nella produzione, revisione e valutazione della scrittura in italiano L2", *Officina.it*, 22, 2014 <www.almaedizioni.it>.

STALNAKER, J. M., "The problem of the English examination", *Educational Record*, 17, 41, 1936.

STOCKFORD, L.; BISSELL, H. W, "Factors involved in establishing a merit-rating scale", *Personnel*, 256, 1949, pp. 94-118.

THORNDIKE, R. L.; HAGEN, E. P., *Measurement and evaluation in psychology and education*, John Wiley and Sons, New York, 1977[4].

TORRESAN, P., "Analisi (classica, Rasch, dei distrattori) di una prova di lettura a scelta multipla della certificazione di italiano per stranieri CILS", *Euro-American Journal of Applied Linguistics and Languages*, 2, 1, 2015, pp. 20-55 <http://www.ejournals.eu/pliki/art/1627/> doi: 10.21283/2376905X.2.27

WEIGLE, S. C., "Using FACETSs to model rater training effects", *Language Testing*, 15, 2, 1998, pp. 263-287.

WEIGLE, S. C., *Assessing writing*, CUP, Cambridge, 2002.

WIGGLESWORTH, G., "Exploring bias analysis as a tool for improving rater consistency in assessing oral interaction", *Language Testing*, 10, 1993, pp. 305-335.

WILSON, M.; ENGELHARD, G., *Objective measurement: Theory into practice*, Vol. 5, Stanford, Ablex, CT, 2000.

WRIGHT, B. D.; LINACRE, J. M., "Reasonable mean-square fit values", *Rasch Measurement Transactions,* 8, 1994, 3, pp. 369-386. <https://www.rasch.org/rmt/rmt83b.htm>

YODER, D.; STAUDOHAR, P. D., *Personnel management and industrial relations*, Prentice-Hall, Englewood Cliffs, NJ, 1982[7].

## Appendice 1. La griglia B1 dello scrivere della Certificazione PLIDA Nuovo Formato

Nelle pagine che seguono il lettore accede alla griglia della produzione scritta di livello B1 in uso all'interno della Certificazione PLIDA (tab. 32, *infra*; ricordiamo che il punto di taglio è pari a 5).

Corredano la lista le *condizioni di non valutabilità di un testo* e la lista delle *strutture ricorrenti al livello B1.* Quest'ultima, utile alla valutazione della componente grammaticale e di quella lessicale, è frutto di uno spoglio di testi prodotti in seno alla stessa Certificazione PLIDA e alla Certificazione PLIDA Juniores.

*Condizioni di non valutabilità di un testo*

Un testo *non è valutabile* se si verifica almeno una di queste condizioni:

- il numero di parole di una o di entrambe le parti è inferiore al limite minimo indicato;
- una o entrambe le prove non sono state svolte;
- la prova contiene brani copiati o trascritti a memoria di cui è stato possibile rintracciare la fonte;
- nella prova sono evidenti gli interventi di più persone: sono riconoscibili diverse grafie e/o diversi livelli di competenza;
- il candidato si limita a riprodurre i testi che ha ricevuto con il fascicolo d'esame (istruzioni, traccia, testi di appoggio, ecc.);
- la prova è stata svolta a matita o corretta con il bianchetto.

*Lista delle strutture ricorrenti nelle prove di produzione scritta del livello B1 del PLIDA*

## Morfologia

### Aggettivi

- Indefiniti *nessuno, ogni*
- Gradi dell'aggettivo: comparativi regolari e irregolari
- Superlativi relativi

### Pronomi

- Pronomi personali atoni complemento diretto e indiretto
- Uso dei pronomi atoni nei tempi composti
- Pronomi atoni combinati
- *Ci* locativo
- *Ne* nelle formule *che ne dici/pensi?*
- Enclisi dei pronomi atoni con l'infinito e con l'imperativo

### Preposizioni

- La preposizione *di* con funzione comparativa

### Verbi

- Scelta dell'ausiliare *avere* o *essere* nelle costruzioni transitive e intransitive di *iniziare/cominciare* e *finire*
- Trapassato prossimo
- Futuro (valore temporale e modale)
- Indicativo imperfetto (forme e uso in contrapposizione al passato prossimo)
- Condizionale presente
- Imperativo formale e informale
- Costruzione impersonale del verbo con *si*
- Congiuntivo presente del verbo essere in costruzioni di alta frequenza come *penso/spero/(mi) sembra che*

**Sintassi**

- Coordinate introdotte da *però*, *invece*, *oppure*, *dunque*, *quindi*, *perciò*, *infatti*, *cioè*
- Completive introdotte da *di*
- Temporali introdotte da *mentre*
- *Senza* + infinito
- Interrogative indirette introdotte da *se* e *come*
- Relative introdotte da *che* e *dove*
- Oggettive esplicite (con il congiuntivo di verbi di alta frequenza) e implicite con *di* + infinito rette da verbi che esprimono opinioni, speranze, sentimenti
- Periodo ipotetico di primo tipo (della realtà)

Tabella 32. *Griglia relativa alla produzione scritta. Certificazione PLIDA Nuovo Formato. Livello B1*

| | CONTENUTO E SVOLGIMENTO DEL COMPITO |
|---|---|
| 10<br>9 | ▪ Affronta tutti i punti della scaletta in modo adeguato e sufficientemente dettagliato.<br>▪ Le caratteristiche del testo (tipologia, registro, formule, ecc.) rispondono pienamente alla richiesta.<br>▪ Il testo può presentare esempi pertinenti, precisazioni, spiegazioni, opinioni o narrazioni secondarie. |
| 8<br>7 | ▪ Affronta tutti i punti in modo generalmente adeguato, ma alcuni possono essere meno sviluppati di altri.<br>▪ Le caratteristiche del testo (tipologia, registro, formule, ecc.) sono adatte alla richiesta. |
| 6<br>5 | ▪ Affronta a grandi linee tutti i punti oppure ne sviluppa solo alcuni in maniera adeguata.<br>▪ Le caratteristiche del testo (tipologia, registro, formule, ecc.) rispondono abbastanza a quanto richiesto; possono comparire piccole incongruenze. |
| 4<br>3 | ▪ Tenta di rispondere alla consegna, ma il testo dà a chi legge l'impressione di un abbozzo.<br>▪ Le caratteristiche del testo (tipologia, registro, formule, ecc.) non sono adatte alla richiesta. |
| 2<br>1 | ▪ Il testo non risponde alla consegna.<br>▪ Il testo è costituito quasi per intero da ripetizioni, elenchi o informazioni irrilevanti. |
| 0 | ▪ Il testo è incomprensibile o non valutabile. |

| COERENZA E COESIONE | |
|---|---|
| 10<br><br>9 | ▪ Le informazioni sono organizzate secondo una progressione coerente, precisa e abbastanza articolata.<br>▪ I coesivi e i connettivi previsti per il livello* vengono usati in modo corretto, esteso e appropriato. |
| 8<br><br>7 | ▪ Le informazioni sono organizzate secondo una progressione generalmente coerente.<br>▪ I coesivi e i connettivi previsti per il livello* vengono usati in modo abbastanza esteso e quasi sempre appropriato.<br>▪ Talvolta le relazioni logiche possono non essere del tutto chiare. |
| 6<br><br>5 | ▪ Le informazioni sono organizzate in modo elementare; alcuni punti del testo possono risultare incoerenti.<br>▪ Usa alcuni coesivi e connettivi previsti per il livello*, anche se non sempre in modo corretto. |
| 4<br><br>3 | ▪ L'organizzazione del testo non è ben definita (digressioni, salti logici, dispersioni, contraddizioni, uso poco ragionato di liste).<br>▪ Usa solo connettivi semplici per collegare le frasi.<br>▪ La scarsa conoscenza dei meccanismi coesivi costringe il candidato a ripetersi. |
| 2<br><br>1 | ▪ Il testo presenta uno schema organizzativo difficile da interpretare.<br>▪ I meccanismi di coesione sono quasi assenti; si limitano per lo più a unire parole o gruppi di parole, non sempre con successo. |
| 0 | ▪ Il testo è incomprensibile o non valutabile. |

*Si veda la lista delle strutture ricorrenti nelle prove di produzione scritta del livello B1 del PLIDA.

| Punteggio | GRAMMATICA, ORTOGRAFIA, PUNTEGGIATURA |
|---|---|
| 10<br>9 | ▪ Il testo presenta una buona varietà di strutture*, usate in modo corretto e appropriato.<br>▪ Errori isolati (morfologici, ortografici o di punteggiatura). |
| 8<br>7 | ▪ Il testo presenta una buona varietà di strutture*.<br>▪ Gli errori (grammaticali, ortografici e di punteggiatura) riguardano singoli elementi della frase e possono essere ripetuti. |
| 6<br>5 | ▪ Il testo presenta un numero limitato di strutture*, non tutte usate con sufficiente padronanza.<br>▪ Errori (morfologici, ortografici e di punteggiatura) diffusi; in alcuni passaggi la lettura può essere faticosa. |
| 4<br>3 | ▪ Il testo presenta solo strutture del livello precedente, non tutte usate con sufficiente padronanza.<br>▪ Gli errori (morfologici, ortografici e di punteggiatura) sono numerosi, anche nel caso di strutture elementari; la lettura è molto faticosa. |
| 2<br>1 | ▪ Il testo presenta solo strutture di base, non tutte usate con sufficiente padronanza.<br>▪ Gli errori (morfologici e ortografici e di punteggiatura) impediscono quasi del tutto la comprensione del testo. |
| 0 | ▪ Il testo è incomprensibile o non valutabile. |

*Si veda la lista delle strutture ricorrenti nelle prove di produzione scritta del livello B1 del PLIDA.

## Appendice 2. Istruzioni per il calcolo

Di seguito riportiamo le istruzioni impartite al programma Facets® nel calcolo degli indici. Nel modello viene applicato l'ancoraggio della terza variabile: le *componenti dell'abilità di scrittura* ovvero le *categorie della griglia* (Vertical= 1,2, 3A,4).

```
Title= B1_scritto
Output= B1_scritto.out.6
Facets= 4; 1: valutatori; 2: candidati; 3: compiti; 4: componenti
Positive= 1,2,3,4
Non-centered= 1
Unexpected = 2                  ; report ratings if standardized residual
                                >=|2|
Usort = (1,2,3),(3,2,1),(Z,3)   ; sort and report unexpected ratings
                                several ways
Arrange = m,0f                  ; arrange tables by measure-descending
                                for all facets,
                                ; and 0f = Z-score-descending for facet 0
                                (bias interactions)
Inter-rater = 1                 ; facet 1 is the rater facet
pt-biserial=measure             ; point-measure correlation
Vertical= 1,2, 3A,4,
Model= ?,?,?,?, PlidaB1         ; valutatore, candidato, compito,
                                componenti
Rating scale= PlidaB1,R10
1 = lowest                      ; name of lowest observed category
5 = middle                      ; no need to list unnamed categories
10 = highest                    ; name of highest observed category
*
Labels=
1, valutatori                   ; (elementi: 7)
1= H
2= X
3= Y
4= Z
5= J
```

6= K
7= W
*
2, candidati ; (elementi: 30)
1-30
*
3, compiti, A; (elementi: 6) ; Prompts all anchored to be the same difficulty
1=a, 0
2=b, 0
3=c, 0
4=d, 0
5=e, 0
6=f, 0
*
4, componenti ; (4 categorie)
1=contenuto
2=coesione/coerenza
3=lessico
4=grammatica
*
Data=
1, 1, 1, 1-4, 4, 4, 3, 4
1, 1, 2, 1-4, 5, 4, 5, 4
1, 2, 1, 1-4, 5, 4, 5, 6
1, 2, 2, 1-4, 6, 5, 5, 5
1, 3, 1, 1-4, 5, 4, 5, 5
1, 3, 2, 1-4, 5, 4, 5, 5
1, 4, 1, 1-4, 8, 7, 8, 7
1, 4, 2, 1-4, 8, 7, 9, 8
1, 5, 1, 1-4, 8, 5, 6, 6
1, 5, 2, 1-4, 6, 5, 6, 6
1, 6, 1, 1-4, 8, 7, 6, 7
1, 6, 2, 1-4, 7, 5, 6, 6
1, 7, 1, 1-4, 9, 6, 7, 7
1, 7, 2, 1-4, 8, 7, 7, 7
1, 8, 1, 1-4, 10, 8, 8, 8
1, 8, 2, 1-4, 5, 8, 7, 7

1, 9, 1, 1-4, 9, 6, 7, 7
1, 9, 2, 1-4, 7, 6, 7, 7
1, 10, 3, 1-4, 7, 6, 6, 6
1, 10, 4, 1-4, 5, 5, 6, 6
1, 11, 3, 1-4, 7, 6, 7, 7
1, 11, 4, 1-4, 7, 5, 7, 6
1, 12, 3, 1-4, 7, 5, 6, 5
1, 12, 4, 1-4, 7, 6, 5, 5
1, 13, 3, 1-4, 4, 5, 6, 5
1, 13, 4, 1-4, 5, 4, 7, 6
1, 14, 3, 1-4, 7, 5, 7, 7
1, 14, 4, 1-4, 5, 4, 7, 6
1, 15, 3, 1-4, 7, 6, 7, 7
1, 15, 4, 1-4, 7, 5, 7, 6
1, 16, 3, 1-4, 7, 4, 3, 4
1, 16, 4, 1-4, 7, 5, 2, 4
1, 17, 3, 1-4, 8, 7, 8, 7
1, 17, 4, 1-4, 9, 6, 8, 7
1, 18, 3, 1-4, 8, 6, 8, 8
1, 18, 4, 1-4, 7, 6, 7, 8
1, 19, 3, 1-4, 6, 5, 6, 5
1, 19, 4, 1-4, 6, 5, 6, 5
1, 20, 5, 1-4, 8, 6, 8, 7
1, 20, 6, 1-4, 8, 6, 8, 8
1, 21, 5, 1-4, 8, 6, 8, 8
1, 21, 6, 1-4, 7, 7, 7, 7
1, 22, 5, 1-4, 10, 7, 10, 9
1, 22, 6, 1-4, 9, 7, 8, 8
1, 23, 5, 1-4, 8, 7, 7, 7
1, 23, 6, 1-4, 6, 7, 7, 7
1, 24, 5, 1-4, 4, 5, 6, 5
1, 24, 6, 1-4, 5, 5, 5, 5
1, 25, 5, 1-4, 7, 5, 4, 5
1, 25, 6, 1-4, 7, 6, 6, 6
1, 26, 5, 1-4, 8, 6, 7, 7
1, 26, 6, 1-4, 7, 6, 7, 7
1, 27, 5, 1-4, 7, 6, 7, 7
1, 27, 6, 1-4, 6, 6, 6, 6

1, 28, 5, 1-4, 9, 7, 8, 8
1, 28, 6, 1-4, 9, 8, 8, 7
1, 29, 5, 1-4, 5, 5, 6, 6
1, 29, 6, 1-4, 7, 5, 6, 6
1, 30, 5, 1-4, 7, 5, 5, 4
1, 30, 6, 1-4, 5, 5, 6, 5
2, 1, 1, 1-4, 5, 4, 4, 5
2, 1, 2, 1-4, 6, 4, 4, 4
2, 2, 1, 1-4, 6, 6, 7, 7
2, 2, 2, 1-4, 6, 6, 6, 6
2, 3, 1, 1-4, 5, 5, 5, 4
2, 3, 2, 1-4, 5, 4, 5, 4
2, 4, 1, 1-4, 9, 8, 9, 7
2, 4, 2, 1-4, 9, 7, 10, 7
2, 5, 1, 1-4, 8, 6, 6, 6
2, 5, 2, 1-4, 7, 5, 7, 7
2, 6, 1, 1-4, 8, 8, 8, 7
2, 6, 2, 1-4, 7, 6, 7, 8
2, 7, 1, 1-4, 9, 8, 8, 8
2, 7, 2, 1-4, 9, 8, 8, 8
2, 8, 1, 1-4, 9, 9, 9, 9
2, 8, 2, 1-4, 6, 8, 9, 9
2, 9, 1, 1-4, 9, 6, 8, 8
2, 9, 2, 1-4, 9, 6, 7, 8
2, 10, 1, 1-4, 6, 5, 5, 5
2, 10, 2, 1-4, 6, 5, 6, 5
2, 11, 3, 1-4, 8, 8, 7, 7
2, 11, 4, 1-4, 7, 6, 7, 6
2, 12, 3, 1-4, 7, 6, 7, 6
2, 12, 4, 1-4, 8, 7, 8, 7
2, 13, 3, 1-4, 4, 5, 5, 5
2, 13, 4, 1-4, 5, 5, 6, 5
2, 14, 3, 1-4, 8, 6, 8, 7
2, 14, 4, 1-4, 6, 6, 8, 7
2, 15, 3, 1-4, 9, 6, 7, 7
2, 15, 4, 1-4, 7, 6, 7, 7
2, 16, 3, 1-4, 6, 4, 4, 4
2, 16, 4, 1-4, 6, 5, 3, 4

2, 17, 3, 1-4, 9, 8, 9, 9
2, 17, 4, 1-4, 9, 7, 9, 9
2, 18, 3, 1-4, 7, 7, 7, 8
2, 18, 4, 1-4, 8, 6, 7, 7
2, 19, 3, 1-4, 6, 4, 6, 4
2, 19, 4, 1-4, 5, 4, 7, 5
2, 20, 3, 1-4, 8, 7, 9, 8
2, 20, 4, 1-4, 8, 7, 8, 8
2, 21, 5, 1-4, 9, 8, 8, 8
2, 21, 6, 1-4, 8, 8, 8, 8
2, 22, 5, 1-4, 10, 8, 10, 9
2, 22, 6, 1-4, 8, 8, 9, 8
2, 23, 5, 1-4, 9, 6, 7, 6
2, 23, 6, 1-4, 7, 6, 8, 7
2, 24, 5, 1-4, 4, 4, 5, 4
2, 24, 6, 1-4, 6, 4, 5, 4
2, 25, 5, 1-4, 8, 6, 6, 6
2, 25, 6, 1-4, 8, 6, 6, 6
2, 26, 5, 1-4, 9, 8, 7, 7
2, 26, 6, 1-4, 8, 7, 7, 7
2, 27, 5, 1-4, 9, 7, 8, 7
2, 27, 6, 1-4, 6, 6, 7, 6
2, 28, 5, 1-4, 8, 8, 8, 8
2, 28, 6, 1-4, 9, 8, 7, 7
2, 29, 5, 1-4, 8, 6, 7, 8
2, 29, 6, 1-4, 8, 6, 7, 7
2, 30, 5, 1-4, 8, 6, 6, 5
2, 30, 6, 1-4, 6, 5, 6, 5
3, 1, 1, 1-4, 4, 4, 4, 4
3, 1, 2, 1-4, 5, 4, 5, 4
3, 2, 1, 1-4, 6, 5, 4, 5
3, 2, 2, 1-4, 5, 4, 5, 5
3, 3, 1, 1-4, 5, 6, 5, 7
3, 3, 2, 1-4, 5, 5, 4, 5
3, 4, 1, 1-4, 10, 9, 10, 9
3, 4, 2, 1-4, 10, 10, 10, 9
3, 5, 1, 1-4, 8, 8, 8, 7
3, 5, 2, 1-4, 8, 7, 8, 7

3, 6, 1, 1-4, 7, 7, 7, 7
3, 6, 2, 1-4, 8, 7, 7, 7
3, 7, 1, 1-4, 9, 10, 10, 9
3, 7, 2, 1-4, 9, 8, 8, 9
3, 8, 1, 1-4, 10, 10, 9, 10
3, 8, 2, 1-4, 8, 9, 8, 9
3, 9, 1, 1-4, 8, 8, 8, 8
3, 9, 2, 1-4, 9, 8, 8, 8
3, 10, 1, 1-4, 6, 6, 7, 7
3, 10, 2, 1-4, 6, 6, 7, 7
3, 11, 3, 1-4, 7, 6, 6, 7
3, 11, 4, 1-4, 6, 5, 5, 6
3, 12, 3, 1-4, 5, 5, 5, 5
3, 12, 4, 1-4, 5, 4, 4, 5
3, 13, 3, 1-4, 7, 6, 7, 7
3, 13, 4, 1-4, 7, 7, 7, 7
3, 14, 3, 1-4, 6, 6, 6, 6
3, 14, 4, 1-4, 6, 6, 7, 6
3, 15, 3, 1-4, 6, 6, 6, 6
3, 15, 4, 1-4, 6, 6, 7, 6
3, 16, 3, 1-4, 5, 4, 4, 4
3, 16, 4, 1-4, 5, 4, 3, 4
3, 17, 3, 1-4, 8, 7, 8, 7
3, 17, 4, 1-4, 8, 8, 9, 8
3, 18, 3, 1-4, 7, 7, 7, 7
3, 18, 4, 1-4, 8, 7, 8, 7
3, 19, 3, 1-4, 6, 4, 5, 4
3, 19, 4, 1-4, 5, 4, 5, 4
3, 20, 3, 1-4, 7, 8, 8, 8
3, 20, 4, 1-4, 8, 9, 9, 9
3, 21, 5, 1-4, 8, 9, 10, 9
3, 21, 6, 1-4, 9, 8, 8, 8
3, 22, 5, 1-4, 10, 9, 10, 10
3, 22, 6, 1-4, 10, 10, 9, 10
3, 23, 5, 1-4, 8, 7, 8, 7
3, 23, 6, 1-4, 7, 7, 8, 7
3, 24, 5, 1-4, 4, 4, 5, 4
3, 24, 6, 1-4, 5, 4, 5, 4

3, 25, 5, 1-4, 7, 5, 6, 6
3, 25, 6, 1-4, 5, 5, 5, 6
3, 26, 5, 1-4, 8, 8, 8, 8
3, 26, 6, 1-4, 7, 7, 7, 7
3, 27, 5, 1-4, 8, 7, 8, 8
3, 27, 6, 1-4, 7, 6, 7, 7
3, 28, 5, 1-4, 10, 9, 9, 10
3, 28, 6, 1-4, 8, 8, 7, 8
3, 29, 5, 1-4, 7, 6, 7, 7
3, 29, 6, 1-4, 6, 6, 7, 7
3, 30, 5, 1-4, 6, 5, 6, 5
3, 30, 6, 1-4, 5, 4, 5, 4
4, 1, 1, 1-4, 4, 4, 4, 5
4, 1, 2, 1-4, 6, 3, 6, 3
4, 2, 1, 1-4, 6, 6, 5, 6
4, 2, 2, 1-4, 6, 6, 5, 6
4, 3, 1, 1-4, 6, 6, 6, 6
4, 3, 2, 1-4, 7, 6, 6, 6
4, 4, 1, 1-4, 10, 9, 10, 9
4, 4, 2, 1-4, 9, 8, 10, 8
4, 5, 1, 1-4, 8, 7, 5, 6
4, 5, 2, 1-4, 7, 6, 8, 6
4, 6, 1, 1-4, 8, 7, 7, 6
4, 6, 2, 1-4, 8, 7, 8, 8
4, 7, 1, 1-4, 10, 7, 10, 9
4, 7, 2, 1-4, 8, 6, 8, 8
4, 8, 1, 1-4, 10, 10, 9, 9
4, 8, 2, 1-4, 7, 8, 8, 9
4, 9, 1, 1-4, 8, 7, 7, 8
4, 9, 2, 1-4, 6, 6, 7, 7
4, 10, 1, 1-4, 6, 6, 5, 6
4, 10, 2, 1-4, 6, 5, 6, 6
4, 11, 3, 1-4, 7, 6, 8, 8
4, 11, 4, 1-4, 8, 7, 8, 7
4, 12, 3, 1-4, 5, 7, 8, 6
4, 12, 4, 1-4, 8, 5, 6, 7
4, 13, 3, 1-4, 4, 4, 6, 6
4, 13, 4, 1-4, 5, 4, 8, 6

4, 14, 3, 1-4, 7, 5, 9, 8
4, 14, 4, 1-4, 7, 5, 9, 8
4, 15, 3, 1-4, 9, 8, 8, 8
4, 15, 4, 1-4, 8, 8, 6, 6
4, 16, 3, 1-4, 8, 5, 4, 5
4, 16, 4, 1-4, 6, 4, 3, 4
4, 17, 3, 1-4, 10, 8, 10, 8
4, 17, 4, 1-4, 10, 8, 10, 8
4, 18, 3, 1-4, 10, 7, 9, 8
4, 18, 4, 1-4, 8, 7, 10, 8
4, 19, 3, 1-4, 8, 6, 8, 6
4, 19, 4, 1-4, 6, 4, 6, 6
4, 20, 3, 1-4, 8, 9, 10, 8
4, 20, 4, 1-4, 7, 9, 10, 8
4, 21, 5, 1-4, 10, 10, 10, 9
4, 21, 6, 1-4, 10, 9, 10, 8
4, 22, 5, 1-4, 10, 9, 10, 10
4, 22, 6, 1-4, 10, 10, 10, 8
4, 23, 5, 1-4, 9, 8, 8, 6
4, 23, 6, 1-4, 8, 7, 8, 7
4, 24, 5, 1-4, 4, 5, 6, 6
4, 24, 6, 1-4, 6, 5, 6, 6
4, 25, 5, 1-4, 8, 6, 5, 5
4, 25, 6, 1-4, 6, 6, 6, 6
4, 26, 5, 1-4, 8, 9, 9, 8
4, 26, 6, 1-4, 8, 7, 8, 8
4, 27, 5, 1-4, 10, 8, 9, 8
4, 27, 6, 1-4, 8, 8, 7, 6
4, 28, 5, 1-4, 10, 10, 10, 8
4, 28, 6, 1-4, 10, 8, 8, 7
4, 29, 5, 1-4, 8, 6, 6, 6
4, 29, 6, 1-4, 9, 7, 8, 6
4, 30, 5, 1-4, 5, 4, 4, 6
4, 30, 6, 1-4, 6, 4, 6, 4
5, 1, 1, 1-4, 4, 3, 3, 4
5, 1, 2, 1-4, 4, 3, 3, 3
5, 2, 1, 1-4, 6, 5, 6, 7
5, 2, 2, 1-4, 6, 6, 4, 5

5, 3, 1, 1-4, 5, 4, 4, 5
5, 3, 2, 1-4, 4, 5, 4, 5
5, 4, 1, 1-4, 10, 8, 10, 9
5, 4, 2, 1-4, 9, 6, 9, 9
5, 5, 1, 1-4, 9, 5, 5, 6
5, 5, 2, 1-4, 8, 5, 5, 6
5, 6, 1, 1-4, 8, 6, 7, 6
5, 6, 2, 1-4, 8, 6, 7, 7
5, 7, 1, 1-4, 10, 7, 9, 9
5, 7, 2, 1-4, 9, 6, 7, 7,
5, 8, 1, 1-4, 10, 8, 9, 9
5, 8, 2, 1-4, 6, 7, 9, 9
5, 9, 1, 1-4, 9, 6, 8, 8
5, 9, 2, 1-4, 10, 6, 7, 9
5, 10, 1, 1-4, 5, 5, 6, 6
5, 10, 2, 1-4, 5, 5, 6, 5
5, 11, 3, 1-4, 7, 5, 7, 6
5, 11, 4, 1-4, 7, 5, 7, 6
5, 12, 3, 1-4, 6, 5, 5, 5
5, 12, 4, 1-4, 7, 5, 5, 5
5, 13, 3, 1-4, 3, 5, 6, 5
5, 13, 4, 1-4, 5, 5, 7, 6
5, 14, 3, 1-4, 8, 5, 7, 8
5, 14, 4, 1-4, 6, 4, 7, 8
5, 15, 3, 1-4, 7, 5, 7, 7
5, 15, 4, 1-4, 6, 5, 7, 7
5, 16, 3, 1-4, 6, 4, 3, 4
5, 16, 4, 1-4, 7, 4, 2, 4
5, 17, 3, 1-4, 8, 7, 8, 8
5, 17, 4, 1-4, 9, 7, 8, 8
5, 18, 3, 1-4, 8, 6, 8, 8
5, 18, 4, 1-4, 8, 6, 7, 8
5, 19, 3, 1-4, 7, 4, 5, 5
5, 19, 4, 1-4, 5, 4, 6, 5
5, 20, 3, 1-4, 9, 6, 9, 8
5, 20, 4, 1-4, 8, 6, 9, 9
5, 21, 5, 1-4, 9, 7, 8, 8
5, 21, 6, 1-4, 7, 7, 7, 8

5, 22, 5, 1-4, 9, 7, 10, 10
5, 22, 6, 1-4, 10,6, 8, 9
5, 23, 5, 1-4, 8, 7, 7, 7
5, 23, 6, 1-4, 8, 6, 7, 7
5, 24, 5, 1-4, 4, 4, 4, 5
5, 24, 6, 1-4, 6, 5, 5, 5
5, 25, 5, 1-4, 7, 5, 6, 5
5, 25, 6, 1-4, 7, 5, 6, 5
5, 26, 5, 1-4, 8, 6, 7, 7
5, 26, 6, 1-4, 8, 6, 7, 8
5, 27, 5, 1-4, 8, 6, 7, 7
5, 27, 6, 1-4, 6, 6, 6, 7
5, 28, 5, 1-4, 9, 8, 8, 8
5, 28, 6, 1-4, 10, 8, 7, 8
5, 29, 5, 1-4, 6, 6, 7, 7
5, 29, 6, 1-4, 8, 5, 7, 6
5, 30, 5, 1-4, 7, 4, 6, 4
5, 30, 6, 1-4, 6, 5, 6, 5
6, 1, 1, 1-4, 4, 4, 4, 4,
6, 1, 2, 1-4, 6, 5, 6, 5
6, 2, 1, 1-4, 6, 4, 6, 5
6, 2, 2, 1-4, 6, 6, 6, 6
6, 3, 1, 1-4, 6, 4, 5, 5
6, 3, 2, 1-4, 5, 4, 5, 5
6, 4, 1, 1-4, 8, 8, 9, 9
6, 4, 2, 1-4, 8, 8, 9, 9
6, 5, 1, 1-4, 6, 6, 7, 6
6, 5, 2, 1-4, 6, 6, 6, 6
6, 6, 1, 1-4, 6, 5, 6, 6
6, 6, 2, 1-4, 6, 6, 6, 7
6, 7, 1, 1-4, 8, 8, 8, 8
6, 7, 2, 1-4, 8, 8, 8, 7
6, 8, 1, 1-4, 9, 9, 9, 9
6, 8, 2, 1-4, 6, 8, 8, 8
6, 9, 1, 1-4, 8, 8, 8, 8
6, 9, 2, 1-4, 7, 8, 8, 8
6, 10, 1, 1-4, 6, 5, 6, 6
6, 10, 2, 1-4, 6, 5, 6, 5

6, 11, 3, 1-4, 5, 4, 6, 5
6, 11, 4, 1-4, 6, 6, 6, 7
6, 12, 3, 1-4, 6, 6, 6, 5
6, 12, 4, 1-4, 6, 6, 5, 5
6, 13, 3, 1-4, 5, 6, 6, 5
6, 13, 4, 1-4, 6, 5, 6, 6
6, 14, 3, 1-4, 6, 5, 7, 6
6, 14, 4, 1-4, 6, 6, 7, 6
6, 15, 3, 1-4, 8, 8, 8, 7
6, 15, 4, 1-4, 6, 6, 6, 6
6, 16, 3, 1-4, 4, 4, 3, 4
6, 16, 4, 1-4, 4, 4, 3, 4
6, 17, 3, 1-4, 9, 8, 8, 8
6, 17, 4, 1-4, 9, 8, 9, 8
6, 18, 3, 1-4, 8, 8, 8, 8
6, 18, 4, 1-4, 8, 7, 8, 8
6, 19, 3, 1-4, 8, 6, 7, 5
6, 19, 4, 1-4, 6, 5, 6, 5
6, 20, 3, 1-4, 9, 8, 9, 8
6, 20, 4, 1-4, 9, 8, 9, 8
6, 21, 5, 1-4, 8, 8, 8, 8
6, 21, 6, 1-4, 6, 8, 8, 8
6, 22, 5, 1-4, 9, 6, 10, 9
6, 22, 6, 1-4, 8, 8, 9, 9
6, 23, 5, 1-4, 8, 7, 7, 7
6, 23, 6, 1-4, 6, 8, 7, 7
6, 24, 5, 1-4, 4, 4, 4, 4
6, 24, 6, 1-4, 7, 6, 5, 5
6, 25, 5, 1-4, 6, 6, 6, 6
6, 25, 6, 1-4, 6, 6, 6, 6
6, 26, 5, 1-4, 8, 8, 8, 8
6, 26, 6, 1-4, 8, 8, 8, 8
6, 27, 5, 1-4, 8, 8, 8, 8
6, 27, 6, 1-4, 6, 6, 6, 6
6, 28, 5, 1-4, 8, 8, 8, 8
6, 28, 6, 1-4, 8, 8, 8, 8
6, 29, 5, 1-4, 6, 6, 6, 6
6, 29, 6, 1-4, 7, 6, 7, 7

6, 30, 5, 1-4, 5, 5, 5, 5
6, 30, 6, 1-4, 4, 4, 5, 4
7, 1, 1, 1-4, 5, 5, 4, 5
7, 1, 2, 1-4, 5, 4, 5, 4
7, 2, 1, 1-4, 7, 7, 7, 7
7, 2, 2, 1-4, 7, 7, 6, 7
7, 3, 1, 1-4, 4, 5, 5, 4
7, 3, 2, 1-4, 5, 5, 5, 5
7, 4, 1, 1-4, 10, 9, 10, 8
7, 4, 2, 1-4, 9, 9, 9, 7
7, 5, 1, 1-4, 8, 8, 7, 7
7, 5, 2, 1-4, 7, 6, 7, 6
7, 6, 1, 1-4, 7, 6, 6, 6
7, 6, 2, 1-4, 6, 5, 6, 6
7, 7, 1, 1-4, 9, 7, 9, 8
7, 7, 2, 1-4, 8, 7, 8, 7
7, 8, 1, 1-4, 9, 9, 9, 7
7, 8, 2, 1-4, 6, 7, 8, 8
7, 9, 1, 1-4, 9, 7, 9, 8
7, 9, 2, 1-4, 8, 6, 8, 7
7, 10, 1, 1-4, 6, 6, 6, 5
7, 10, 2, 1-4, 6, 5, 6, 5
7, 11, 3, 1-4, 7, 8, 7, 6
7, 11, 4, 1-4, 7, 7, 7, 7
7, 12, 3, 1-4, 7, 6, 6, 5
7, 12, 4, 1-4, 7, 6, 6, 6
7, 13, 3, 1-4, 5, 4, 5, 4
7, 13, 4, 1-4, 6, 5, 5, 4
7, 14, 3, 1-4, 7, 6, 7, 7
7, 14, 4, 1-4, 7, 7, 8, 7
7, 15, 3, 1-4, 7, 7, 6, 5
7, 15, 4, 1-4, 7, 7, 6, 5
7, 16, 3, 1-4, 4, 4, 3, 4
7, 16, 4, 1-4, 4, 4, 3, 4
7, 17, 3, 1-4, 9, 8, 8, 8
7, 17, 4, 1-4, 9, 9, 8, 9
7, 18, 3, 1-4, 8, 7, 8, 7
7, 18, 4, 1-4, 8, 7, 8, 7

7, 19, 3, 1-4, 6, 6, 5, 5
7, 19, 4, 1-4, 7, 5, 6, 5
7, 20, 3, 1-4, 8, 8, 9, 8
7, 20, 4, 1-4, 9, 9, 9, 8
7, 21, 5, 1-4, 8, 8, 7, 9
7, 21, 6, 1-4, 9, 8, 8, 7
7, 22, 5, 1-4, 10, 9, 10,
7, 22, 6, 1-4, 10, 9, 9, 8
7, 23, 5, 1-4, 8, 8, 7, 7
7, 23, 6, 1-4, 7, 6, 7, 6
7, 24, 5, 1-4, 5, 5, 4, 4
7, 24, 6, 1-4, 6, 5, 5, 5
7, 25, 5, 1-4, 6, 6, 5, 5
7, 25, 6, 1-4, 6, 5, 6, 5
7, 26, 5, 1-4, 8, 7, 7, 6
7, 26, 6, 1-4, 8, 7, 8, 7
7, 27, 5, 1-4, 7, 6, 7, 6
7, 27, 6, 1-4, 6, 6, 6, 5
7, 28, 5, 1-4, 10, 10, 9, 8
7, 28, 6, 1-4, 8, 9, 8, 7
7, 29, 5, 1-4, 6, 6, 7, 6
7, 29, 6, 1-4, 7, 6, 7, 6
7, 30, 5, 1-4, 6, 6, 5, 5
7, 30, 6, 1-4, 6, 5, 5, 5
*

## Appendice 3. Modelli di analisi

La linearizzazione messa in atto dal programma Facets® viene eseguita mediante una serie di istruzioni immesse dal ricercatore (cfr. **Appendice 2**). Il passaggio più importante riguarda la definizione del modello di analisi.

In una ricerca contraddistinta da quattro variabili:

- Variabile 1= *generosità del valutatore*
- Variabile 2= *abilità dei candidati*
- Variabile 3= *difficoltà del compito*
- Variabile 4= *difficoltà delle componenti*

e nell'ipotesi di una scala comune (R), i modelli possibili sono riportati nella tabella 33, *infra* (cfr. Myford, Wolfe 2004: 412).

Il primo viene assunto nell'ipotesi che i valutatori usino la scala allo stesso modo (*modello di default* o *rating scale model*): ci dà uno sguardo generale sul comportamento dei valutatori.

Il secondo modello, che è una variabile del primo, prevede un ancoraggio di una componente ad un certo valore. Nel nostro studio, considerata la disomogeneità dei compiti assegnati ai candidati, abbiamo fissato al valore di 0 *logit* il grado di difficoltà dei compiti.

Il modello *analisi delle interazioni* è applicato nello studio dei bias *locali*, riferiti al rapporto tra due variabili (DFF, *differential facet functioning*). Nel nostro caso siamo ricorsi a questo modello dello studio dell'interazione *valutatore-componente*.

I *modelli ibridi* (sarebbe più intuitivo dire "dettagliati") sono applicati nell'ipotesi che i valutatori usino la scala in maniera difforme gli uni dagli altri. In genere, a meno che non si dispongano di dati molto estesi, tali da registrare almeno 10 occorrenze per la valutazione di ogni categoria, si preferisce usare il *modello di default*, poiché più stabile (Eckes 2015). Ad ogni modo i *modelli ibridi* possono essere integrati al *modello di default* per ragioni diagnostiche, ovvero ai fini di un'indagine su specifiche osservazioni abnormi (Myford, Wolfe 2003, 2004).

Per comodità del lettore abbiamo contraddistinto le variabili con colori diversi: la prima è in rosso (la generosità del valutatore), la seconda è in verde (l'abilità dei candidati), la terza è in celeste (la difficoltà del compito), la quarta è in marroncino (la difficoltà delle componenti); la R finale rimanda alla griglia in adozione (la scala comune).

Tabella 34. *Modelli di analisi*

| **TIPO DI MODELLO** | **STRINGA DI ISTRUZIONI** | **RESEARCH QUESTION** |
|---|---|---|
| **Modello di default** (*rating scale model*) | Model= ?, ?, ?, ?, R | Come il gruppo di valutatori valuta le componenti di una serie di elaborati, attinenti a più compiti e prodotti da diversi candidati, facendo riferimento alla scala R? |
| **Modello con ancoraggio di una componente** | Model= ?, ?, ?A, ?, R | Come sopra, con ancoraggio di una componente a un valore dato (nel nostro caso, la componente del compito è stata ancorata allo zero *logit*). |
| **Analisi delle interazioni** (variazione del modello *rating scale model* per lo studio di interazioni specifiche) | Model= B?, B?, ?, ?, R<br><br>Model= ?, B?, ?, B?, R<br><br>Model= ?, B?, B?, ?, R<br><br>Model= B?, ?, B?, ?, R | Quali sono i dati inattesi (*bias*) nell'interazione tra la variabile 1 (i valutatori) e la variabile 2 (i candidati)?<br><br>Quali sono i dati inattesi (*bias*) nell'interazione tra la variabile 2 (i candidati) e la variabile 4 (le componenti)?<br><br>Quali sono i dati inattesi (*bias*) nell'interazione tra la variabile 2 (i candidati) e la variabile 3 (i compiti)?<br><br>Quali sono i dati inattesi (*bias*) nell'interazione tra |

| | | |
|---|---|---|
| | Model= B?, ?, ?, B?, R | la variabile 1 (i valutatori) e la variabile 3 (i compiti)?<br><br>Quali sono i dati inattesi (*bias*) nell'interazione tra la variabile 1 (i valutatori) e la variabile 4 (le componenti)? |
| **Modello ibrido # 1** | Model= ?, ?, ?, #, R | Cosa emerge nella valutazione delle *singole* componenti da parte del gruppo dei valutatori? |
| **Modello ibrido # 2** | Model= #, ?, ?, ?, R | Come *ciascun* valutatore usa la scala? |
| **Modello ibrido # 3** | Model= #, ?, ?, #, R | Come *ciascun* valutatore valuta le *singole* componenti? |

Valutare la competenza comunicativa di un apprendente è un'operazione complessa. L'indagine presentata in questo volume ci informa di alcuni problemi emersi durante la valutazione di elaborati in seno alla Certificazione di italiano come lingua straniera PLIDA. Alla fine del volume sono suggerite alcune azioni che potrebbero ridurre tali difficoltà.

Pedro & João Editores
pedroejoaoeditores.com.br
ISBN 978-65-265-1212-8