Normalizzazione dei giudizi in sede di valutazione dello scritto e adattamento di una griglia di analisi: il resoconto di un'esperienza

Normalización de los juicios en la evaluación del escrito y adaptación de una tabla de análisis: el informe de una experiencia

Criteria Normalization in Evaluation of Writing and Adaptation of an Analysis Table Based on a Real Evaluation Session: Report of an Experience

Paolo Torresan
piroclastico@gmail.com
Doctorado en Lingüística y Filología Románica. Ca' Forcari, Venecia, Italia
Profesor de Didáctica de Lenguas Modernas en Struttura Speciale di Lingue e Letterature Straniere, Università di Catania, Ragusa, Italia

Abstract

Per valutare un testo prodotto da un allievo, un valutatore può dotarsi di *rating scale* (altrimenti dette *scoring rubric*), vale a dire griglie mediante le quali esaminare l'abilità di riferimento (scritta o orale) nelle sue diverse componenti (*criteri*), graduate a loro volta secondo descrittori, a cui corrispondono diversi punteggi. Al fine di ridurre margini di variabilità legati a interpretazioni soggettive, tornano utili sessioni di confronto tra valutatori. Così facendo si contribuisce all'affidabilità dei giudizi. In questo articolo descriviamo una sessione di confronto tra 9 valutatori, ciascuno dei quali è stato chiamato a esprimere un giudizio su 7 testi composti da studenti di italiano LS, di livello attorno all'A2, secondo le indicazioni del *Quadro di Riferimento Europeo per le Lingue* (Council of Europe, 2001; d'ora in poi, CEFR). La sessione è avvenuta all'interno di un corso di formazione da noi condotto nel mese di novembre 2014 presso L'Istituto Italiano di Cultura di Lima. Durante la sessione ci siamo serviti di una griglia riportata in Spinelli (2014), da noi leggermente modificata. Le differenze di giudizio che sono emerse consentono di aprire un confronto tra i valutatori. In secondo luogo, permettono di avviare una riflessione sull'adeguamento della griglia al contesto in cui operano i valutatori.

Parole-chiave: italiano come lingua straniera, valutazione dello scritto, griglia di valutazione, normalizzazione dei giudizi.

Resumen

Para evaluar un texto producido por un alumno, un evaluador puede adoptar una *rating scales* (también llamados *scoring rubrics*), es decir tablas mediante las cuales examinar la habilidad, escrita u oral, en sus varios componentes (*criterios*), graduados y a su vez según descriptores, a

los que corresponden varios puntajes. Con la finalidad de reducir los márgenes de variabilidad vinculados a interpretaciones subjetivas, resultan útiles sesiones de confrontación entre evaluadores. De este modo se contribuye a generar confianza en los juicios. En este artículo describimos una sesión de confrontación entre 9 evaluadores, a cada uno de los cuales se le pidió que expresara un juicio sobre 7 textos redactados por estudiantes de italiano LS, de nivel aproximado A2, según las indicaciones del Marco Común Europeo de Referencia para las Lenguas (Council of Europe, 2001; en adelante CEFR). La sesión se realizó durante un curso de formación que condujimos, en el mes de noviembre de 2014 en el Instituto Italiano de Cultura de Lima. Durante la sesión usamos una tabla presente en Spinelli (2014), a la que le hicimos ligeras modificaciones. Las diferencias de juicio que surgieron permitieron abrir una confrontación entre los evaluadores. En segundo lugar, permitieron iniciar la reflexión sobre la adecuación de la tabla al contexto en el cual operan los evaluadores.

Palabras clave: italiano como lengua extranjera, evaluación del escrito, tabla de evaluación, normalización de los juicios.

Abstract

In order to evaluate a text produced by a student, an evaluator may adopt rating scales (also called scoring rubrics), which are tables through which writing or speaking skills are assessed in their various components (criteria), being graded at the same time according to their descriptions, to which several scores have been assigned. In order to reduce the limits of variability associated to subjective interpretations, comparing and discussion sessions among evaluators are useful. In this way, in this way, assessment criteria becomes trustworthy. In this article we describe a comparing and discussion session involving 9 evaluators, who were asked to express an individual opinion about seven portions of text written by students of Italian LS, whose average level is A2, in accordance with the Common European Framework of Reference for Languages (Council of Europe, 2001; hereinafter CEFR). This session was held during a training course we conducted in November 2014 at the Italian Institute of Culture [Instituto Italiano de Cultura] in Lima. During the session, we used a table present in Spinelli (2014), on which we made small modifications. The small differences suggested by the evaluators. Secondly, they allowed starting the analysis on adapting the table to the context in which evaluators work.

Key words: Italian as a foreign language, evaluation of written text, evaluation table, standardization of criteria.

Cenni sulla complessità della valutazione della produzione scritta

La valutazione dello scritto di un apprendente di lingua impone la messa a punto di vari strumenti, seguendo diverse fasi:

- Definizione del costrutto
- Scrittura e calibrazione di *prompt*
- Elaborazione/adattamento di una griglia di osservazione
- Normalizzazione dei giudizi (*rater norming* o *rater standardization*)
- Raffinamento della griglia

Nella sua globalità, non si tratta di un processo lineare, bensì circolare, in cui il risultato raggiunto in una fase può indurre a ritarare gli strumenti adottati in precedenza (Weigle, 2002).

Definizione del costrutto

Per *costrutto* è da intendersi l'idea astratta di abilità o di competenza che intendiamo misurare. Scrivono Alderson *et al.* (1995, pp. 16-17):

> Every test is an operationalization of some beliefs about language whether the constructor refers to an explicit model or merely relies upon 'intuition'.
> Every theory contains constructs (or psychological concepts), which are its principal components and the relationship between these components.

Nel nostro caso, il costrutto cui facciamo riferimento ha a che fare con che cosa intendiamo per "produzione scritta": quali sono le componenti che la caratterizzano. I criteri su cui si basa una griglia di valutazione ci restituiscono una 'fotografia' di tali componenti (McNamara, 2000, p. 37).

È bene sottolineare che il *costrutto* è sensibile, come si evidenzia nel passo citato sopra, alla concezione di lingua che si condivide[1]; se si opera nel contesto di un *achievement test*, il *costrutto* è sensibile al contesto istituzionale e al tipo di sillabo. Possiamo dunque riscontrare differenze tra *rating scale* pensate per apprendenti adulti e quelle per adolescenti, o tra quelle ideate per i corsi di lingua in generale e altre confezionate per corsi di microlingua, e così via[2].

Scrittura e calibrazione del *prompt*

Lo stimolo per scrivere (ovvero la consegna: *scrivi su…*) si definisce *prompt*. In merito precisiamo che:

- La formulazione deve risultare adeguata al livello di comprensione dello studente, al fine di non introdurre variabili esterne all'abilità che si intende misurare, le quali possono agire negativamente sulla *performance* del candidato, come può accadere nei seguenti casi:
  - Il *prompt* è lungo e complesso: l'eventuale bassa prestazione del candidato nell'espressione scritta può essere dovuta a scarse capacità di lettura (e in tal caso si incorre in una situazione di invalidità dovuta a una *construct-irrelevant variance*).
  - Si fa riferimento a preconoscenze che non tutti i candidati posseggono (es: descrivere un viaggio in aereo, quando non è detto che tutti gli esaminandi siano saliti su un aereo) o impliciti culturali (si fa riferimento a elementi della cultura straniera di cui non tutti i candidati sono a conoscenza). Anche in questo caso, i dati che si raccolgono possono essere alterati da fattori esterni al costrutto (*construct-irrelevant variance*).
- Il prompt deve risultare calibrato, ovvero in linea con le indicazioni dello strumento regolativo cui ci si attiene. Nel nostro caso, il prompt deve rispettare le indicazioni del CEFR e del Profilo (Spinelli & Parizzi, 2011) o più in generale le indicazioni curricolari (cfr. per esempio, Attestato ADA, 2013) in riferimento al livello per il quale è stato pensato

[1] Possiamo avere, per fare un esempio, *rating scale* relative alla produzione scritta che comprendono, tra i criteri, anche il *layout* o l'appropriatezza sociolinguistica, mentre altre no.

[2] In generale le griglie confezionate per un test di profitto (per esami su larga scala, come sono nell'ambito dell'italiano, PLIDA, CILS, CELI) sono "*theory-based*", mentre quelle pensate per la classe si rapportano maggiormente a un sillabo (ufficiale, o implicito, com'è nei manuali), che a sua volta si può ricollegare a una teoria di riferimento.

- Il *prompt* deve risultare generativo: deve stimolare la produzione, costituendo un'occasione sfidante per l'allievo che si accinge a scrivere (Hedge, 1991); inoltre, deve riflettere scopi pragmatici reali (Weingle, 2002).
- Se articolato in passaggi, il *prompt* deve essere:
  - Organico. A volte le indicazioni di un elenco (*bullet point*) sono dispersive; ci può essere un 'punto' che spinge lo studente 'fuori strada' rispetto al tema indicato dai precedenti; il risultato finale è che al lettore, ignaro della consegna, il testo prodotto può risultare poco coerente e coeso, non già a causa della scarsa competenza testuale dell'autore ma per effetto di com'è impostata la traccia.
  - Equilibrato. Se i punti da trattare sono troppi, i margini di autonomia nell'organizzazione del testo vengono drasticamente ridotti; è facile che lo studente sviluppi una lista di frasi anziché un testo coeso.

Nelle istruzioni che accompagnano il *prompt* dovrebbero essere esplicitati (Tankó 2005):

- L'*audience* (a chi è destinato lo scritto)
- Il ruolo assunto dallo scrivente
- Il tipo di testo
- Lo scopo
- La lunghezza[3]
- Il tempo concesso

Eventualmente si possono esplicitare i criteri di valutazione (Weigle, 2002).

Elaborazione e raffinamento di una griglia di valutazione

Il valutatore può esprimere un giudizio basandosi su una generale impressione circa la *performance* del candidato, o servendosi di alcuni criteri guida, riuniti in una griglia (*rating scale*), che valgono, abbiamo detto, a rappresentare il costrutto dell'abilità che intende misurare.

[3] Carson (cit. in Weigle, 2000, p. 103) suggerisce che si stabiliscano i limiti in termini di pagine –mezza pagina, una pagina, ecc. – piuttosto che in unità linguistiche (es. numero di parole), in maniera da rispecchiare l'orientamento generale che ha chi scrive in un contesto extrascolastico. Ad ogni modo, il testo deve essere sufficientemente ampio da consentire un giudizio affidabile (cfr. Hughes, 2003, p. 90).

Questi criteri possono articolarsi in descrittori, accompagnati da un punteggio (per esempio da 1 a 5, o da 0 a 10), in accordo con il sistema di valutazione adottato dall'istituzione di appartenenza.

Nell'*Appendice 1a*, il lettore accede a un esempio di griglia per la valutazione dello scritto di un allievo di livello A2 (menzionata in Spinelli, 2014), secondo i parametri stabiliti dal CEFR. In essa si distinguono le varie tappe per raggiungere il livello. Durante la progettazione di una griglia si deve:

- Evitare, per quanto possibile, di sovrapporre criteri (se due descrittori rimandano a uno stesso comportamento, si penalizzerebbe/premierebbe uno studente due volte)
- Evitare di considerare componenti esterne rispetto a quello che intendiamo valutare (altrimenti il giudizio potrebbe essere invalidato per via di una *construct-irrelevant variance*; fattori esterni possono essere, per esempio, le conoscenze enciclopediche degli studenti)
- Prevedere una proporzione tra i descrittori (per capirci, il 'salto' tra un descrittore 2 e un descrittore 4 dovrebbe essere lo stesso tra un descrittore 4 e un descrittore 6)
- Prevedere un'adeguata discriminazione da parte dei descrittori (altrimenti la prova diventa inaffidabile, dal momento che lascia ampio spazio a interpretazioni soggettive: un valutatore può essere incerto nella scelta tra due o più descrittori facenti capo allo stesso criterio, visto che non presentano differenze significative tra di loro)
- Puntare a un allineamento tra descrittori relativi a criteri diversi (se il descrittore 4, per esempio, relativo al criterio "A" indica una buona competenza, tale aggettivo deve qualificare anche la competenza rappresentata dai descrittori equivalenti dei criteri "B", "C", "D", ecc.)
- Raggiungere un compromesso tra esaustività e praticabilità (i descrittori devono essere sufficientemente dettagliati ma non in maniera tale da complicare l'uso della griglia. Come buona norma, suggeriamo che una griglia possa starci in un foglio A4 o al limite su due).

Nel leggere la griglia, il valutatore può scorrere i descrittori dall'alto in basso, fino a individuare quello che più si addice a caratterizzare la competenza dell'allievo (*"best fit" approach*, Carr, 2011, p. 137).

In sostanza, l'elaborazione di una griglia è un primo passo, seguito da un processo di raffinamento. Il raffinamento può avvenire:

- *ab intra*, ovvero per mezzo di una riflessione sulla stessa, in quanto tale, a partire dal giudizio di esperti, dal confronto con griglie simili, dallo studio della letteratura, ecc.
- *ab extra*, cioè a partire dall'analisi dei risultati che provengono dell'applicazione (I criteri sono rappresentativi? I descrittori sono efficaci? Lo strumento è praticabile?).

Normalizzazione e stabilizzazione dei giudizi

Un'istituzione educativa (dal centro certificatore alla singola scuola) ha interesse che i giudizi forniti dai propri insegnanti/valutatori siano validi e affidabili.
Per quanto concerne la *validità*, occorre *in primis* esplicitare il costrutto e accertarsi che la prova si attenga ad esso.

Per quanto riguarda l'*affidabilità*, essa è messa a rischio dai margini di soggettività che incorrono nella formulazione del giudizio.

L'esplicitazione del costrutto da parte dello *staff* di valutatori, la messa a punto di una griglia efficace (criteri ben definiti; descrittori proporzionati, allineati e calibrati), la scrittura di un *prompt* adeguato (che rimanda a un genere testuale familiare agli esaminandi, in grado di elicitare una quantità sufficiente di *output*, privo di *bias* culturali, di genere, ecc.) si completano, infine, con i processi di *stabilizzazione* e di *normalizzazione* dei giudizi, in modo da ridurre al minimo variazioni intra- e intersoggettive da parte dei valutatori (aumentando, di converso, la *intra-rater reliability* e la *inter-rater reliability*).

Variabili soggettive, infatti, come il grado di stanchezza o l'umore, possono far sì che il giudizio del valutatore non sia stabile: se deve giudicare la stessa *performance* a distanza di tempo, si può trovare nelle condizioni di formulare un giudizio diverso. D'altro canto, lo stesso

valutatore può esprimere, rispetto a una certa *performance*, un giudizio diverso rispetto a quello dei colleghi, dimostrandosi più severo o più tollerante nei confronti della media.

In ambito formativo diventa allora fondamentale prevedere spazi volti alla *stabilizzazione* e alla *normalizzazione*, in vista di una maggiore coerenza (in modo che la mia attribuzione di livello di oggi sia pari a quella di domani, per esempio) e una sintonia tra coloro che valutano (in modo tale che uno studente possa farsi valutare dal valutatore x o dal valutatore y, senza il timore di incappare, in un caso o nell'altro, in quello più intransigente). I processi che, in sede di formazione, si possono avviare a tal proposito sono i seguenti:

- Somministrare campioni di *output* a distanza di tempo a uno stesso valutatore e analizzare eventuali oscillazioni di giudizio: a che cosa sono dovute? Eventuali differenze nell'attribuzione dei punteggi in corrispondenza ai singoli criteri sono eventualmente compensate da un giudizio complessivo stabile?[4]

- Somministrare uno stesso campione di *output* a un gruppo di valutatori e studiare eventuali difformità nei giudizi: emerge un nucleo di valutatori che valuta in maniera omogenea o si nota la presenza di più sottogruppi? I margini di variazione sono trascurabili? Quali sono i criteri in corrispondenza ai quali si danno i *gap* maggiori? Tali problemi sono imputabili a distrazione o a una '*secret agenda'* del valutatore ("*idiosyncratic criteria*", Charney, 1984) o, ancora, rimandano a una griglia poco chiara?

Nei paragrafi che seguono descriviamo i primi passi di un'esperienza di normalizzazione dei giudizi in sede di valutazione dello scritto occorsa in un contesto formativo.

L'esperienza condotta presso l'Istituto di Cultura di Lima

Dal 17 al 19 novembre 2014 abbiamo condotto, in occasione della XXVIII Settimana della lingua italiana nel mondo, un seminario di aggiornamento presso l'Istituto Italiano di Cultura di Lima, dal titolo "*Confezionare, calibrare e validare test di italiano*", della durata complessiva di nove ore.

[4] In tal caso a rimanere costante è l'impressione generale che il valutatore ha del testo.

Gli argomenti che abbiamo trattato sono stati:

- La confezione di test di comprensione validi e affidabili
- La calibrazione degli *item* di test di lettura (*Standard setting*)
- La mappatura del testo (ovvero l'individuazione di nuclei informativi a partire dai quali costruire *item* di comprensione)
- La normalizzazione dei giudizi in sede di valutazione dello scritto e del parlato

Per quanto riguarda l'ultimo punto, con particolare riferimento all'abilità di scrittura, abbiamo agito attraverso la somministrazione previa (due settimane prima del corso) di sette composizioni scritte di studenti di italiano di livello A2 a nove valutatori volontari. Durante il corso, abbiamo analizzato i giudizi espressi. Nei paragrafi che seguono i dati vengono illustrati.

Si premetta, in accordo con quanto già dichiarato, che l'esperienza, oltre a consentire un maggiore equilibrio tra i giudizi dei valutatori, si è rivelata utile per ridefinire la struttura della griglia di analisi, in accordo con le esigenze del contesto peculiare dell'IIC di Lima. Torniamo quindi a sostenere che una sessione di normalizzazione agisce, indirettamente, sul raffinamento/adattamento di una griglia.

La valutazione degli scritti

Grazie alla collaborazione della Scuola Edulingua (San Saverino Marche), abbiamo raccolto sette scritti di studenti di livello A2 (*Appendice 2*). Gli studenti avevano sviluppato un testo breve (era stata data loro l'indicazione: "12-15 righe", ie. tra le 100 e le 150 parole) sulla base delle seguenti istruzioni:

Composizione

Scrivi una lettera ad un/a amico/a sul mese trascorso in Italia, toccando i punti seguenti:

1. Cosa ti è piaciuto di più e cos'altro avresti voluto fare
2. Descrivi le persone che hai incontrato
3. Racconta un episodio curioso e divertente
4. Cosa consigli a una persona che vuole venire in Italia e studiare alla scuola Edulingua?

Il coordinatore didattico dell'Istituto di Cultura, una volta ricevuti gli scritti, li ha inoltrati ai nove insegnanti che si sono offerti di valutare i testi in forma anonima. Ciascun valutatore è stato avvisato del livello degli studenti (A2) ed è stato invitato a servirsi della griglia che elaborammo a partire da quella menzionata in Spinelli, 2014. Il lettore ha accesso a tale griglia nell'*Appendice 1b*.

La formulazione di un giudizio sulla *performance* del candidato da parte del valutatore è passata attraverso i seguenti criteri:

- *contenuto* (realizzazione del compito ed effetto sul lettore)
- *vocabolario* (ampiezza e appropriatezza)
- *accuratezza* (ortografia e morfosintassi)
- *organizzazione* (coerenza e coesione)

Espresso il giudizio, ciascun valutatore ha trasmesso i risultati al coordinatore e questi a noi. I dati raccolti sono presentati nelle tabelle dei paragrafi a seguire. Per ciascuna di esse, la legenda adottata è la seguente:

V: *valutatori*
S: *testi scritti*
m: *valore minimo*
M: *valore massimo*
d: *differenza tra il valore minimo e il valore massimo*
d_1: *differenza tra il secondo voto in ordine di severità e il voto più alto (tale valore è stato da noi considerato alla luce, come si vedrà, del fatto che un valutatore si è generalmente dimostrato di gran lunga il più severo di tutti)*
me: *media aritmetica*
mn: *mediana (valore collocabile a metà tra i giudizi, sotto il quale si hanno i giudizi più severi e sopra il quale i giudizi più tolleranti)*
md: *valore/i più frequente/i*

In ogni tabella abbiamo evidenziato in rosso (vedi Tabella 1) i giudizi che si distaccano dagli altri in un senso positivo (quando cioè il valutatore esprime, da solo, il giudizio più alto); viceversa, i giudizi evidenziati in giallo rappresentano casi isolati di severità (il valutatore esprime, da solo, il giudizio più basso). Mediante un fondino grigio abbiamo evidenziato le differenze apprezzabili in termini di assoluti (*d*; il distacco tra valori minimi e valori massimi) e

relativi (d_1; differenza tra il secondo voto in ordine di severità e il voto più alto). Riteniamo 'apprezzabile', per entrambi questi indici, un margine superiore alle tre unità[5]

Consideriamo, quindi, criterio per criterio, i giudizi espressi, soffermandoci in particolare su quegli scritti che sollevano una particolare divergenza nei giudizi.

Tabella 1
Giudizi sugli scritti: il contenuto.

		V															
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**M**	**M**	**d**	**d_1**	**me**	**mn**	**md**
S	**1**	4	3	4	4	5	3	0	3	4	0	5	5	2	3,3	4	4
	2	6	7	7	7	8	8	6	8	8	6	8	2	2	7,2	7	8
	3	4	3	4	2	3	3	0	4	5	0	5	5	3	3,1	3	3/4
	4	5	7	6	4	10	9	4	6	7	4	10	6	4	4,9	5,5	4/6/7
	5	5	5	5	5	4	7	4	4	5	4	7	3	3	4,9	5	5
	6	7	8	9	7	8	9	6	9	7	6	9	3	2	7,8	8	7/9
	7	6	5	6	5	5	7	4	6	6	4	7	3	2	5,6	6	6

S1, S3

Nella tabella riguardante il *contenuto*, emerge un *pattern*: il valutatore V7 si profila come quello più severo, assegnando il valore minimo assoluto (zero) in un paio di occasioni (in corrispondenza agli scritti S1, S3), in disaccordo rispetto ai colleghi. È il suo voto, in entrambe le circostanze, a determinare una differenza (*d*) di ben 5 punti. Se non considerassimo il suo parere,

[5] In una sessione avanzata, con valutatori esperti in termini di normalizzazione, il valore scenderebbe alle due unità.

la differenza interna al gruppo (d_1), in corrispondenza a S3 e S4, si ridurrebbe rispettivamente a 2 e 3 punti: non sarebbe apprezzabile[6].

S4

Di rilievo è lo scarto tra i valutatori in merito a S4: d=6; d_1=4, con un m=4 (condiviso da due *rater*) e un M=10 (assegnato da V5, e in parte sostenuto, con il voto di 9 punti, da V6). Che ci sia una certa dispersione (e un certo disorientamento) è ravvisabile anche dalle mode: 4, 6, 7. In sostanza, abbiamo più gruppi di giudizi in merito alla valutazione del *contenuto*: da chi ritiene che il compito sia realizzato al minimo e che la lettura richieda uno sforzo da parte del lettore (V4, V7), a chi ritiene che la consegna sia stata portata a termine parzialmente (V1, V3, V8), a chi, ancora, ritiene che l'obiettivo sia stato realizzato in maniera adeguata (V2, V9), sino a giungere, infine, a chi è dell'opinione che il testo sia comprensibile e che il tema sia stato affrontato pienamente (V5, V6).

La questione pare rimandare, in questo caso, alle componenti che definiscono il criterio del *contenuto*, e sulla base delle quali sono stati costruiti i descrittori. Analizzando i descrittori, il *contenuto* è declinato in:

- *Chiarezza e comprensibilità del testo* (la comprensibilità è definita, a sua volta, sulla base del tipo di errori)
- *Svolgimento del compito*

Nella tabella sottostante sono messi in evidenza i passaggi dei descrittori che si rifanno alle componenti cui abbiamo appena dato accenno (vedi Tabella 2).

[6] Può essere che la severità di V7 in merito al *contenuto* sia stata determinata, in corrispondenza a S1, S3, per via di altri criteri, quali *lessico* e *accuratezza*.

Tabella 2

Descrittori delle componenti del criterio del contenuto: a) chiarezza e comprensibilità del testo; b) svolgimento del compito.

10 **9**	(**a**) Il contenuto è chiaro e comprensibile. (**b**) L'argomento è stato trattato pienamente.
8 **7**	(**b**) Il compito è stato realizzato adeguatamente; (**a**) il testo si può dire abbastanza chiaro.
6 **5**	(**b**) Il compito è stato realizzato parzialmente. (**a**) Alcuni errori impediscono la comprensione. Il messaggio è solo parzialmente comunicato al lettore.
4 **3**	(**b**) Il compito è stato realizzato al minimo. (**a**) Molti errori impediscono la comprensione e la lettura richiede uno sforzo da parte del lettore
2 **1**	Ai limiti del valutabile. Prestazione minima.
0	Non valutabile

Facendo riferimento alle componenti di cui sopra, S4 risulta un testo "*chiaro e comprensibile*"; inoltre i *bullet point* di cui si compone il *prompt* pare siano stati tutti ripresi (vedi Figura 1).

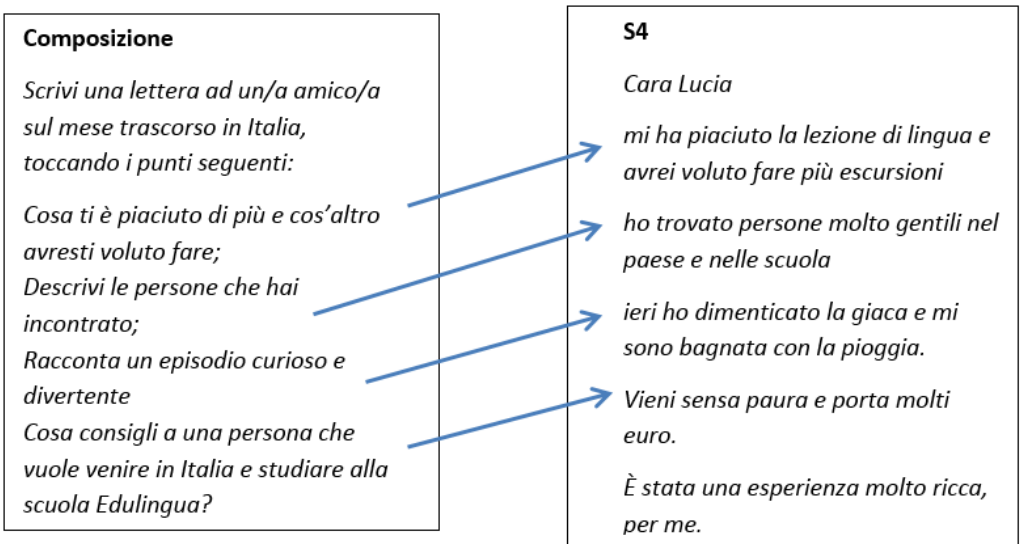


Figura 1. Ripresa dei punti del prompt (bullet point) in S4.

Stando così le cose, la posizione dei valutatori più generosi è giustificata. Tuttavia può essere che i valutatori che si sono espressi con giudizi più bassi, condividano, implicitamente, una nozione di "*task achievement*", vale a dire di "realizzazione del compito", più complessa e articolata, che preveda:

- Uno sviluppo/una spiegazione di ciascun punto (e non una semplice ripresa; si potrebbe parlare tal proposito di "*sviluppo tematico*")
- Un *rispetto dei limiti assegnati*.

Entrambi questi aspetti sono disattesi da S4. Gli argomenti indicati nel *prompt* sono ripresi ma non sviluppati; inoltre il testo è brevissimo: il numero di parole è circa il 40% del limite stabilito nella consegna. La dispersione dei giudizi pare dunque essere dipesa da una '*secret agenda*'. Alcuni valutatori avrebbero avuto in mente altri criteri quando si è parlato di *contenuto*, e ne avrebbero tenuto conto al momento di decidere che voto assegnare[7]. Può tornare utile, in tal senso, una riscrittura dei descrittori alla luce delle componenti di cui sopra (cfr. *Appendice 1c*).

Tabella 3
Giudizi subli scritti: il vocabolario.

		V															
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**m**	**M**	**d**	**d_1**	**me**	**mn**	**md**
S	**1**	4	3	4	4	3	4	0	3	4	0	4	4	1	2,8	4	4
	2	6	6	8	7	8	7	4	9	7	4	9	5	3	6,9	7	7
	3	3	2	4	2	1	3	0	3	5	0	5	5	3	2,3	3	3
	4	5	7	7	3	8	9	4	6	6	3	9	6	6	6,1	6	6/7
	5	4	5	6	5	4	6	4	5	4	4	6	2	2	4,8	5	4
	6	7	8	9	7	8	9	6	9	7	6	9	3	2	7,8	8	7/9
	7	4	5	7	5	6	7	4	7	5	4	7	3	3	5,6	5	5/7

[7] Non è escluso, ad ogni modo, come dimostrato in Lumley, 2002, che in un criterio multidimensionale, com'è appunto il *contenuto*, il valutatore possa affidarsi più all'una o all'altra due dimensioni (es. più alla *chiarezza e alla comprensibilità* che non allo *svolgimento del compito*), nell'atto di formulare il proprio giudizio, specie se nel testo non si dà un'evidenza di quanto espresso nel descrittore.

Anche nell'ambito del *vocabolario*, se non consideriamo il valore isolato del giudice più severo (pure in questo caso V7), le differenze si riducono e non sono in genere apprezzabili (vedi Tabella 3).

S4

È ancora una volta, tuttavia, il testo S4 a generare ampio disaccordo: *d*=6. I giudizi si distribuiscono su quattro livelli: si passa dall' "*ampio e sempre appropriato*" (V6) al "*limitatissimo*" (V4). Per dirimere la questione, occorre far fede all'ampiezza di vocabolario e di strutture attese da uno studente di livello A2. Dal campione esiguo di lingua esibito in S4 (50 parole) è piuttosto difficile ritenere che lo studente abbia dato prova di un lessico "*ampio*": la brevità del testo (che, abbiamo visto, dovrebbe portare a giudizi negativi in merito al *contenuto*) gioca un suo ruolo anche sull'ampiezza lessicale.

Dal poco che possiamo osservare, desumiamo, tuttavia, che l'allievo si sia dimostrato capace di una discreta varietà, grazie alla quale è in grado di trasmettere quel che intende dire. Sa usare il passato prossimo, accompagnato da pronomi (riflessivi e indiretti, pur se in questo caso sbaglia l'uso dell'ausiliare), l'imperativo ("*vieni*", "*porta*"), fino a dimostrare dimestichezza con un condizionale composto (allocabile a un livello B2), che probabilmente ha indotto V5 e V6 a formulare giudizi generosi. A S4 può corrispondere, dunque, un giudizio ai limiti della sufficienza (6) o appena al di sopra (7), in accordo, del resto, con le mediane che emergono dall'indagine.

Tabella 4
Giudizi sugli scritti: l'accuratezza.

		V															
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**M**	**M**	**d**	**d_1**	**m e**	**m n**	**m d**
S	**1**	3	3	4	3	1	3	0	3	4	0	4	4	3	2,7	3	3
	2	6	6	7	7	9	7	4	8	8	4	9	5	3	6,8	7	7
	3	3	2	4	2	2	3	0	4	5	0	5	5	3	2,7	3	2
	4	6	6	7	2	9	8	4	8	7	2	9	7	4	6,3	7	6/ 7/ 8
	5	3	4	6	4	3	7	2	5	4	2	7	5	4	4,2	4	4
	6	7	8	8	6	8	7	4	9	7	4	9	5	2	7,1	7	7/ 8
	7	4	4	6	6	7	7	4	6	5	4	7	3	3	5,4	6	4/ 6

Si notano differenze di giudizio piuttosto accentuate in merito all'*accuratezza* degli scritti S4, S5 (vedi Tabella 4).

S4

Nel primo caso, il valutatore più parsimonioso è V4, il quale peraltro giudica gli scritti S4, S5 in maniera difforme dai colleghi, considerando il primo di qualità peggiore del secondo.

Il comportamento severissimo di V4 in merito a S4 (assegna solo 2 punti: "*ai limiti del valutabile*") ci pare ingiustificato (applicazione poco attenta dei descrittori?): pur nelle dimensioni ridotte del testo, si notano pochi errori in S4 e, oltretutto, non paiono rilevanti.

Ad ogni modo, anche in questo caso, sembra sia proprio l'esiguità del testo ad aver contribuito alle divergenze tra chi giudica.

S5

In merito a S5, si nota un'ampia distribuzione dei giudizi, come si evince dallo schema qui sotto (vedi Figura 2).

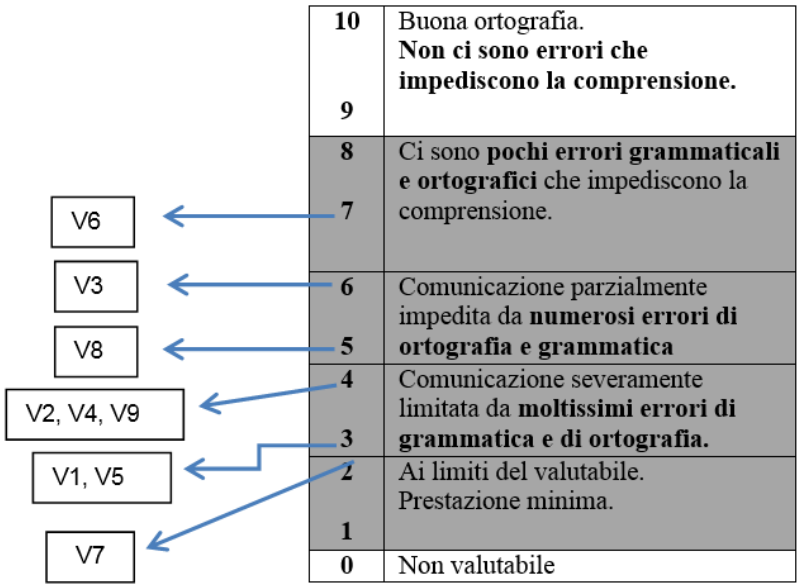


Figura 2. S5, distribuzione dei giudizi.

Ai valutatori non è chiaro se gli errori presenti nel testo impediscano appena la comprensione (V6), o la compromettano in parte (V3, V8) o in maniera considerevole (V2, V4, V9, V1, V5) o addirittura rendano inqualificabile la prova (V7).

Le posizioni estreme (V6, V7) paiono, tuttavia, escludersi a vicenda: il testo si lascia comprendere da un nativo, sebbene con fatica.

La questione che rimane comunque irrisolta è stabilire se la comprensione sia "*parzialmente*" o "*considerevolmente*" compromessa dagli errori.

A ben vedere, S5 contiene errori che non pregiudicano gravemente la comprensione; quindi, stando all'applicazione fedele della griglia, pare abbiano maggior ragione V8 e V3, che hanno assegnato punteggi (rispettivamente 5, 6) corrispondenti al seguente descrittore: "*Comunicazione parzialmente impedita da numerosi errori di ortografia e grammatica*".

Come mai, dunque, la maggioranza dei valutatori (V1, V5, V2, V4, V9) si riconosce in un giudizio più severo, optando per il descrittore di livello inferiore: "*Comunicazione severamente limitata da moltissimi errori di grammatica e di ortografia*"?

Prima di rispondere, c'è da ravvisare la sovrapponibilità, nella griglia di riferimento, tra componenti del criterio *contenuto* e componenti del criterio *accuratezza*.

Nel primo, il *contenuto*, in corrispondenza alla stringa 3-4, leggiamo: "*Molti errori impediscono la comprensione e la lettura richiede uno sforzo da parte del lettore*". Nel secondo, l'*accuratezza*, in corrispondenza alla stringa equivalente (3-4) leggiamo: "*Comunicazione severamente limitata da moltissimi errori di grammatica e di ortografia*". In effetti, il *contenuto* è (in parte) tradotto nella comprensibilità del testo, la quale è una funzione del tipo di errori commessi; d'altro canto, l'*accuratezza* è definita sulla base degli errori che esercitano un impatto sulla comprensibilità del testo (vedi Figura 3).

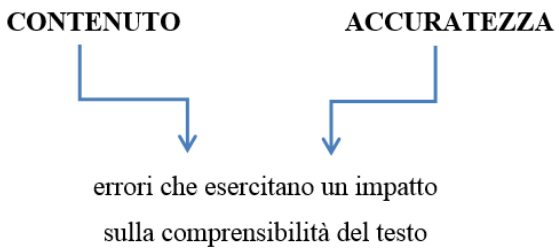


Figura 3. Sovrapposizione tra le componenti dei criteri accuratezza e contenuto.

Così come stanno le cose, pare si rischi di valutare due volte una stessa dimensione. Un adattamento potrebbe essere quello, in merito al criterio dell'*accuratezza*, di considerare in parte gli errori che esercitano un impatto sulla comprensibilità del testo (senza poi considerarli nel *contenuto*), per prestare attenzione anche agli errori regressivi rispetto al livello di competenza morfosintattica e ortografica atteso.

Un allievo può compiere errori pre-sistematici (relativi a forme non ancora acquisite e quindi non valutabili) o sistematici (relativi a forme acquisite; cfr. Cattana & Nesci, 2004). Gli errori sistematici possono riguardare forme recentemente acquisite o *routine* e vocaboli acquisiti al livello più basso.

È probabile che una '*secret agenda*', definita dalla competenza morfosintattica e ortografica attesa da un A2 – o meglio dal sillabo corrispondente a questo livello e costituito dal materiale adottato in Istituto (Balì & Rizzo, 2002) – abbia spinto il nutrito gruppo di giudici ad adottare il descrittore corrispondente al livello 3-4: ("*Comunicazione severamente limitata da moltissimi errori di grammatica e di ortografia*"). In sostanza, costoro avrebbero messo in secondo piano l'aspetto della comprensibilità del messaggio, dimostrandosi sensibili piuttosto alla 'gravità' degli errori, avendo come parametro il sillabo in uscita di un allievo A2, così come esso si presenta nel manuale in uso[8].

In effetti S5 presenta un numero cospicuo di forme elementari errate, che anzi si spera un allievo del livello inferiore (A1) abbia già consolidato; peraltro, nel testo si riscontrano numerosissime inferenze della lingua materna (lo spagnolo):

- "*o prenditu*" (per: "ho imparato")
- "*amice simpatice*" (per: "amici simpatici" o "amiche simpatiche")
- "*di Argentina e Brazil*" (per: "argentine e brasiliane" o "argentini e brasiliani")
- "*ho preguntado*" (per: "ho domandato")

[8] Tra l'altro, si badi, le indicazioni del CEFR relative a *grammatica* e *lessico* sono vaghe. Le indicazioni per l'italiano escono nel 2011 (Spinelli & Parizzi, 2011): nove anni dopo la pubblicazione del manuale tuttora in uso da parte dagli insegnanti che operano presso l'IIC di Lima (Balì & Rizzo, 2002). I valutatori, quindi, si sono presumibilmente attenuti al sillabo morfolessicale di livello A2 desumibile dal manuale, composto però prima dell'uscita dello strumento regolativo voluto dal Consiglio d'Europa, il *Profilo* appunto.

Anche in questo caso, insomma, come nel caso del *contenuto*, pare necessario un adattamento della griglia, mediante riscrittura dei descrittori (cfr. *Appendice 1c*).

L'organizzazione

L*'organizzazione* pertiene alla struttura del testo. Nel descrittore corrispondente al livello 7-8 c'è, a dire il vero, un accenno allo *sviluppo tematico* ("*alcune parti del testo non sono adeguatamente sviluppate*"), il quale però, come abbiamo argomentato in precedenza, dovrebbe ricadere piuttosto sotto il criterio del *contenuto*. Si tratta di un elemento da noi introdotto nella scheda riportata in Spinelli (2014), per una ragione di discriminazione (abbiamo inserito un descrittore in più e abbiamo inteso specificarlo il più possibile), senza accorgerci tuttavia che ciò avrebbe comportato una sovrapposizione con le componenti di altri criteri. Ciò, almeno in linea di principio, ha potuto indurre qualche valutatore a formulare un giudizio 'spurio' (vedi Figura 4).

ottimo	Chiara progressione delle idee che sono ben collegate. Buona introduzione e conclusione.
buono	Basica, con possibili frasi formulaiche e uso di semplici elementi di connessione (ad esempio "e", "ma" e "perché").
adeguato	Basica, con possibili frasi formulaiche. Scarsi tentativi di connessione.
non adeguato	Parole isolate, mancanza di coerenza
debole	Produzione troppo limitata per essere valutata in base all'input richiesto

(ottimo)	**10** **9**	Chiara progressione delle idee: ben collegate. Buona introduzione e conclusione.
(buono)	**8** **7**	Una buona organizzazione, anche se alcune parti del testo non sono adeguatamente sviluppate.
(sufficiente)	**6** **5**	Una organizzazione sufficiente, con frasi formulaiche. Uso di semplici elementi di connessione (ad esempio, "ma", "e", "perché").
(ai limiti della sufficienza)	**4** **3**	Frasi brevi con scarsa connessione; poca fluidità nel testo.
(non adeguato)	**2** **1**	Parole isolate, mancanza di coerenza.
(debole)	**0**	Non valutabile

Figura 4. Il criterio dell' organizzazione nella versione riportata in Spinelli (2014), a siniestra, e nel nostro adattamento, a destra:in evidenza il descrittore introdotto e l'elemento di sovrapposizione.

Ad ogni modo, il criterio dell'*organizzazione* solleva ampie differenze tra i giudici, come si può notare nello schema sottostante (vedi Tabella 5).

Tabella 5
Giudizi sugli scritti: l'organizzaaione.

		V															
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**m**	**M**	**D**	**d_1**	**me**	**mn**	**md**
S	**1**	4	2	4	4	3	3	0	3	4	0	4	4	2	3	3	4
	2	6	4	6	7	9	9	6	9	8	4	9	5	5	7,1	7	6/ 9
	3	4	2	4	2	3	6	0	3	5	0	6	6	4	3,2	3	2/ 3/ 4
	4	4	4	4	3	4	7	4	4	6	3	7	4	4	4,4	4	4
	5	4	3	5	4	3	9	4	6	4	3	9	6	6	4,7	4	4
	6	7	6	10	8	9	9	4	10	7	4	10	6	4	7,8	8	7/ 9/ 10
	7	4	4	4	4	4	9	4	8	5	4	9	5	5	5,1	4	4

Il giudice generalmente più severo, V7, in questo caso si è dimostrato tale solo in un paio di casi, e ancora una volta assegnando un voto negativo assoluto (zero) agli scritti S1 e S3 (com'era avvenuto per il *contenuto*)[9].

Si distacca, invece, per generosità, il giudice V6, che in ben 4 occorrenze formula il giudizio più benevolo.

L'estrema varietà delle scelte dei valutatori ci impone di procedere con una rassegna analitica, scritto per scritto.

S1

L'unico testo che raccoglie un certo consenso è S1. Media e mediana sono pari a 3, la moda è pari a 4: gran parte dei giudizi si concentrano sullo stesso livello: "*Frasi brevi con scarsa connessione; poca fluidità nel testo*".

[9] C'è da chiedersi se, anche in questo caso, la "*non valutabilità della prova*", da lui riconosciuta, sia imputabile all'interferenza di criteri esterni, quali l'*accuratezza* e il *vocabolario*.

A riprova della scarsa coerenza, in S1 si nota che l'ultimo elemento del *bullet point* (il consiglio a un amico di frequentare la scuola Edulingua) si presenta avulso dal resto delle informazioni[10].

S2

A parte il giudizio severo di V4, che non è giustificato, il gruppo si muove in due direzioni, come evidenziano le mode 6 e 9, con uno scarto di tre punti. S2**,** comunque sia, è un testo strutturato; può aver tratto in inganno i giudici più severi la lista di località a metà del testo, per via della quale essi avrebbero potuto sentirsi indotti a ritenere che il testo sia caratterizzato da un'organizzazione povera.

S3

La media aritmetica in merito a S3 è leggermente superiore a quella di S1 (3,2 > 3). I giudizi sono molto variegati, spalmati su 4 livelli (vedi Figura 5).

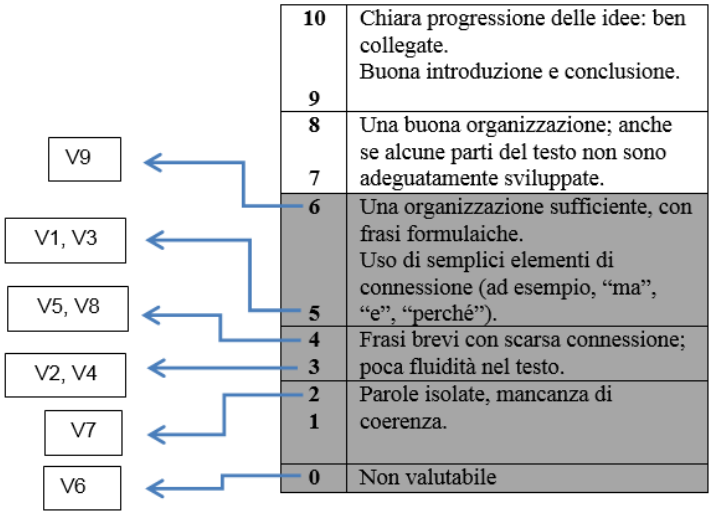


Figura 5. S3, distribuzione dei giudizi.

In questo frangente si rivela necessario procedere a un confronto tra valutatori, nel quale ciascuno esponga e sostenga il proprio punto di vista, al fine di giungere a un giudizio normalizzato, quindi un accordo sul fatto che lo scritto in oggetto sia un abbozzo sconclusionato

[10] Il nostro giudizio, come moderatore, oscilla attorno al 2, quindi ci allineiamo con il parere dei giudici più severi.

(0-2) o una somma di frasi prive di connessione organica (3-4) o presenti, invece, caratteristiche di coerenza e coesione tali da renderlo "*sufficientemente strutturato*" (5-6).

Pare, ad ogni modo, ci si collochi alle soglie della testualità. La struttura è appena abbozzata (3). Chi lo leggesse da esterno può cogliere salti logici: "Io ho incontrate belli personi, et attenti. Io ho dovutto lasciare dil treno perché la maquinetta non fontionaba et il controlore a ditto di lasciare"

S4

In merito allo scritto S4, il valutatore V6 si distacca dal gruppo per il suo giudizio positivo (7). Il resto del gruppo appare compatto, allineandosi attorno al valore 4 (così media, mediana e moda).

S5

Anche per S5, il valutatore V6 rappresenta una voce isolata (9). Come si vede dalla mappa sottostante, il corpo dei giudizi si proietta su due livelli. La moda (è evidente dalla mappa stessa) è pari a 4 (vedi Figura 6).

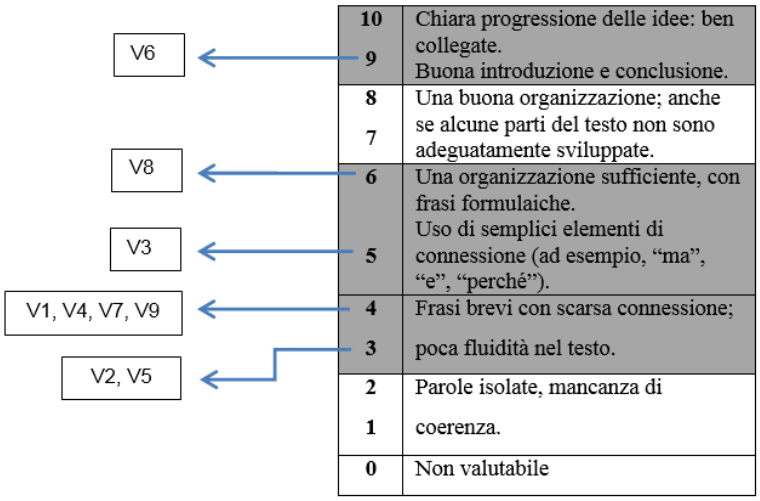


Figura 6. S5, distribuzione dei giudizi.

Anche in questo caso un processo di normalizzazione passa attraverso una discussione sulla qualità di S5. È vero in effetti che gli elementi di connessione sono scarsi ("*e*", "**cuando*"), però è anche vero che sono usati convenientemente. Inoltre, benché le frasi paiano giustapposte,

l'impressione generale che si ha, attraverso la lettura, è che ci si trovi effettivamente di fronte a un testo. In altre parole, benché il brano sia privo di dispositivi che garantiscano fluidità, la costruzione testuale c'è e regge[11], al punto che il lettore può fruire della lettura, senza dover far riferimento alla consegna per potersi orientare (si noti l'accompagnamento al lettore: "*Un episodio divertente e de paura*"). Si potrebbe paragonare il brano a un muro a secco: manca la calce ma le pietre si incastrano tra sé[12].

Un'osservazione che ci riguarda. Nell'adattamento della griglia di cui si ha menzione in Spinelli (2014), omettemmo l'aggettivo "*possibile*", riferito all'uso di "*semplici elementi di connessione",* in merito al descrittore corrispondente al livello 5-6. L'assenza di questa specificazione può generare nel valutatore un atteggiamento condizionato, tale da fargli reputare che, se mancano questi dispositivi di connessione, il testo è irrimediabilmente privo di coesione (così come invece non è).

S6

La prova si distacca rispetto alle altre per qualità. La divisione tra i giudizi è in realtà meno evidente di quanto potrebbe apparire: solo V7 e V8 non risultano allineati; la maggioranza si muove attorno ai due livelli più alti (vedi Figura 7).

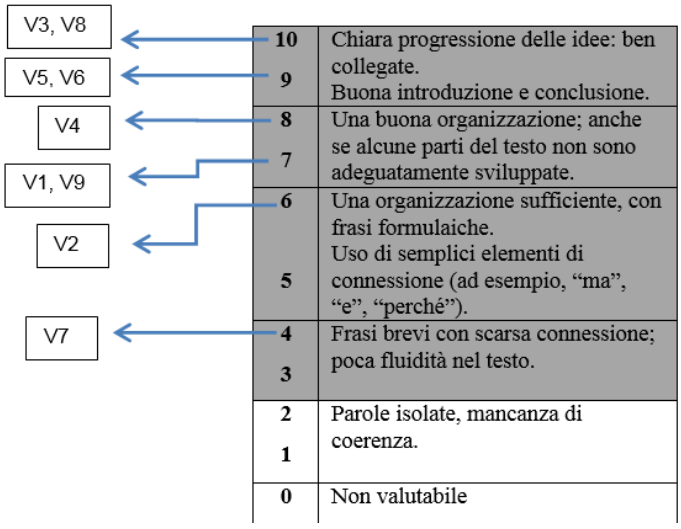


Figura 7. S6, distribuzione dei giudizi.

[11] Cfr. a tal proposito Widdowson (1983).
[12] Come moderatori, siamo quindi in sintonia con il giudizio singolo di V3 (5), che del resto riflette la media aritmetica.

Anche in questo caso una discussione interna al gruppo dei valutatori può sancire se si tratta di una prestazione "*buona*" o "*ottima*" in termini di *organizzazione*[13].

S7

I giudizi si distribuiscono su quattro livelli, con un'evidente concentrazione attorno al descrittore che rappresenta un livello "*ai limiti della sufficienza*" (4). La moda è pari a 4, così è la mediana; la media aritmetica (5,1) è sensibile ai giudizi più generosi (V6, V8), (vedi Figura 8).

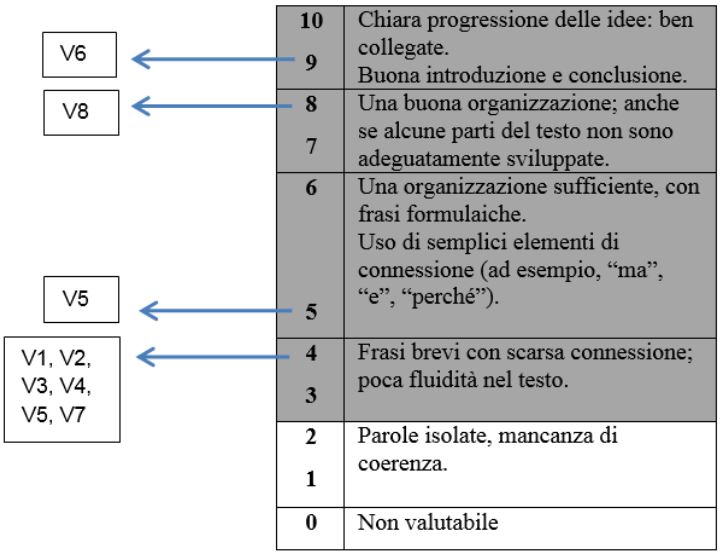


Figura 8. Distribuzione dei giudizi.

Lo scritto, in effetti, fa fatica a decollare (le prime tre proposizioni appaiono piuttosto slegate), e anche se nella seconda metà si riscontra un leggero recupero, rimane in dubbio la coesione dell'elemento finale ("*un episodio più divertente*") rispetto a quanto precede. Anche in questo caso la discussione tra i *rater* può risultare risolutiva[14].

[13] L'impressione dello scrivente è che si tratti di un testo elegante; il nostro giudizio si muove attorno all'8 e al 9. Il solo passaggio che potrebbe parere poco coeso è quello che collega la scuola (ubicazione, persone incontrate, ecc.) alla descrizione di Firenze; tuttavia, a ben vedere, nell'introduzione vi è un'anticipazione del tema del viaggio: "*Ho conosciuto tanti luoghi e persone che seranno sempre nella mia testa e cuore*".

[14] Come moderatori, ci pare che l'intervallo sancito dagli indicatori statistici (4-5) costituisca il margine all'interno del quale la discussione debba essere impostata.

Riflessioni conclusive

La sessione di normalizzazione realizzata presso l'IIC di Lima ha costituito una duplice occasione.

Da un lato essa ha consentito un confronto tra i valutatori, in modo tale che ciascuno di essi ha potuto:

- Raggiungere una consapevolezza circa il proprio posizionamento in riferimento alla linea seguita dal gruppo (eventualmente mettere alla prova le convinzioni che si nutrono circa la propria 'generosità' o 'severità' in veste di giudice)
- Ricalibrare il proprio giudizio alla luce di aspetti notati da altri o a cui altri hanno dato maggiore importanza o, alternativamente, sostenere le proprie ragioni nonostante la divergenza rispetto al resto del gruppo.

Dall'altro lato essa ha portato a una ristrutturazione della griglia di analisi (non si è trattato tanto di una validazione, quanto di una riarticolazione del costrutto, considerati i *belief* dei *rater* e il contesto di insegnamento). Più in dettaglio, argomentiamo quanto segue.

Il confronto tra i valutatori

Il confronto tra pari costituisce un passaggio di rilievo per controllare la qualità dei giudizi in sede di *performance assessment* (Shohamy *et al.*, 1992; Lim, 2011). Ciascun esaminatore ha modo di avere uno specchio relativo alla propria capacità di gestione della complessità e di acquisire uno spirito critico maggiore (cfr. Weigle, 1994).

Dal comportamento dei valutatori si è evinta una generale coerenza interna nella formulazione dei giudizi (*intra-rater reliability*), con una costante severità di V7 (che sovrastima l'abilità di uno studente di livello A2), e una generosità del valutatore V6 in merito al criterio dell'*organizzazione* in particolare. Ciò lo si evince facilmente attraverso la Figura 9, nella quale abbiamo riportato i giudizi massimi e minimi isolati, ovvero le attribuzioni di punteggio per le quali i valutatori si sono distaccati dal resto del gruppo in un senso positivo o negativo.

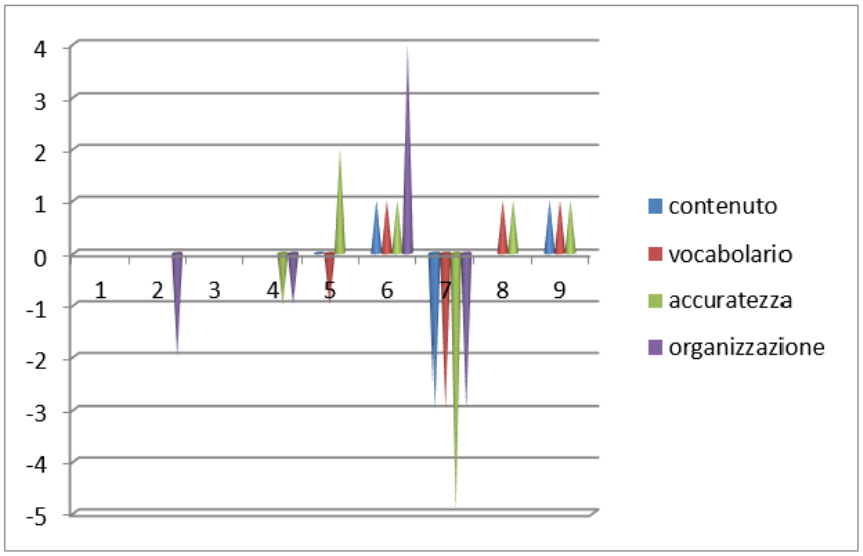


Figura 9. Giudizi massimi e minimi isolati.

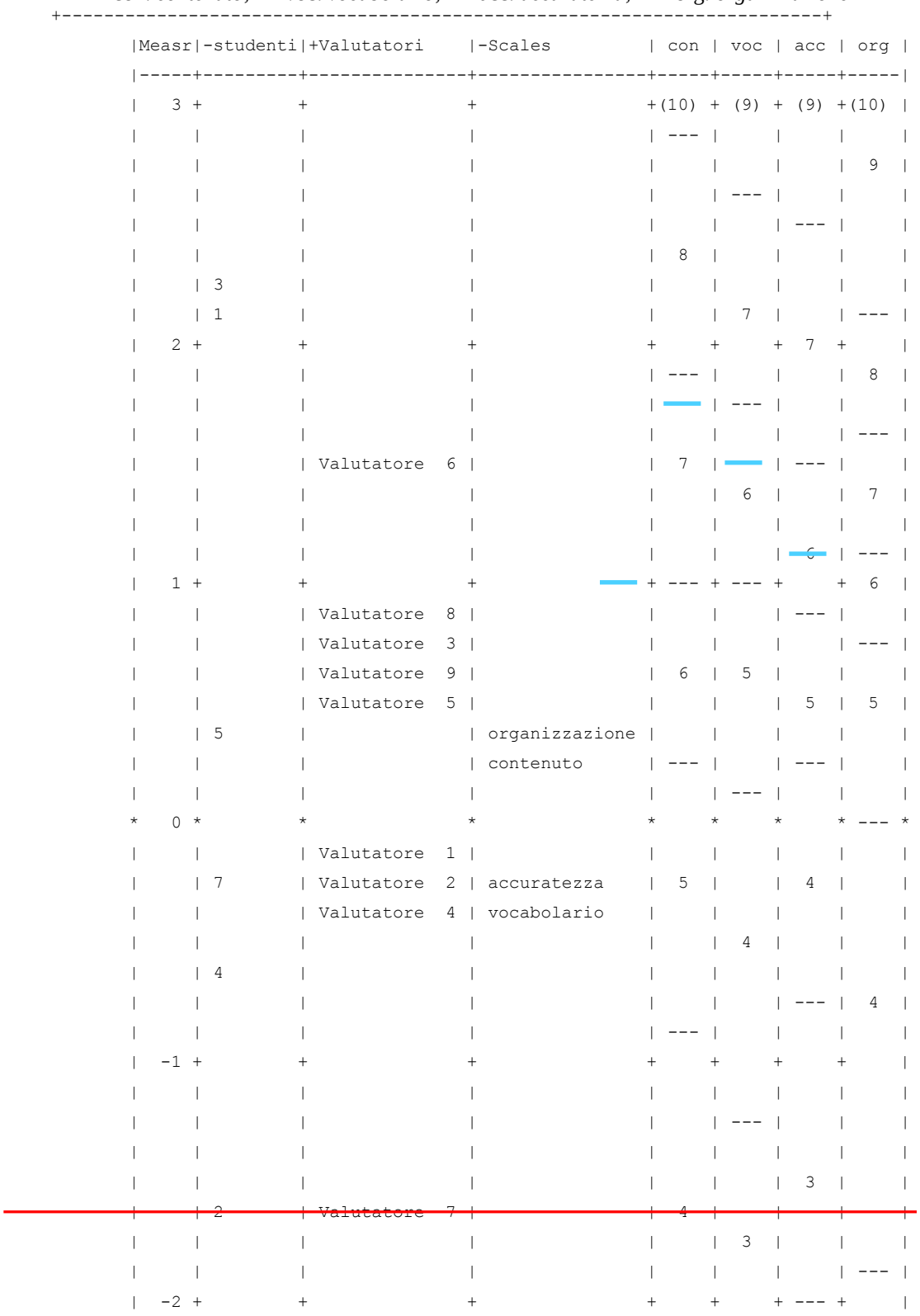
A un´ analisi Rasch risulta, ad ogni modo, che i valutatori, in generale, hanno interpretato i descrittori in maniera disallineata, dimostrando maggiore severità per *accuratezza* e *vocabolario* rispetto a *contenuto* e *organizzazione*.

Così, per esempio, le linee di demarcazione evidenziate in azzurro, nella Figura 10, tra il punteggio 6 e il punteggio 7 ci dimostrano che un candidato può raggiungere più agevolmente un voto 7 in termini di *organizzazione* e *contenuto*, rispetto all'impegno che invece gli è richiesto in termini di *accuratezza* e *vocabolario*.

Emergono inoltre *cut score* differenziati: il punto di discrimine tra il livello A1 e il livello A2, per esempio, appare diverso tra i valutatori (si veda la linea rossa, nella Figura 10, in corrispondenza della quale e sotto la quale si collocano le composizioni che, a detta di V7, esprimono un'abilità di scrittura di livello A2).

Legenda

con: *contenuto*; voc: *vocabolario*; acc: *accuratezza*; org: *organizzazione*



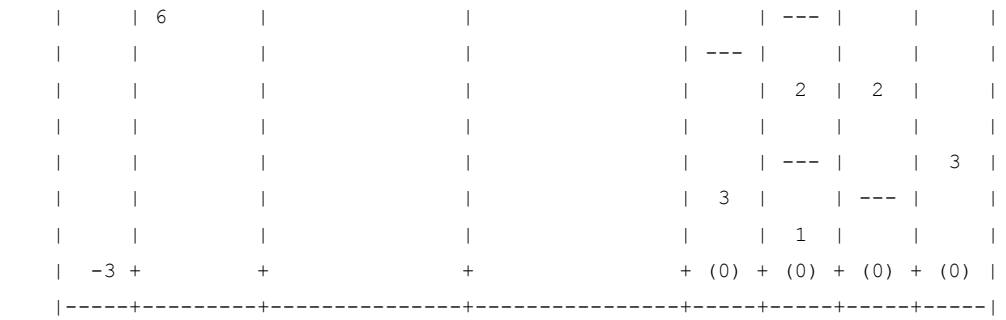


*Figura 10***.** Analisi Rasch: *vertical ruler.*

La riscrittura della griglia di valutazione

Il valutatore può affinare lo strumento di osservazione attraverso un confronto con colleghi esperti, sia prima della somministrazione di una prova (cfr. *Appendice 1b*) che in un momento successivo, ovvero alla luce del *feedback* che costoro gli trasmettono una volta applicata la griglia (cfr. *Appendice 1c*).

Nel nostro caso il confronto tra valutatori è avvenuto contestualmente a una sessione di normalizzazione dei giudizi. La conclusione cui siamo giunti è che siano necessari:

- Un'integrazione, per quanto concerne il criterio del *contenuto*, di componenti quali:
  - Il *rispetto dei limiti previsti nella consegna*
  - Lo *sviluppo tematico*
- Un arricchimento dei descrittori del *vocabolario*
- Una considerazione, per quanto riguarda il criterio dell'*accuratezza*, di errori al livello basico (A1)
- Una discriminazione maggiore tra alcuni descrittori inerenti l'*organizzazione*

La conclusione metodologica cui giungiamo è la seguente: l'esperienza ci ha indicato come una riflessione sulla struttura di una griglia di analisi può non bastare per accertare l'adeguatezza della stessa. I risultati che provengono dal confronto tra coloro che la usano (nel nostro caso, un confronto orientato ad accertare l'affidabilità dei valutatori) può rivelarsi, altresì, di estrema utilità per accertare l'idoneità dello stesso strumento di rilevazione. In sostanza rivendichiamo la possibilità di un processo circolare e di ricalibrazione reciproca: una griglia consente di accertare l'adeguatezza dei giudizi dei valutatori, tanto quanto i giudizi degli stessi apportano informazioni

utili per confermare l'adeguatezza dello strumento di rilevazione, in riferimento al contesto per il quale lo stesso strumento è stato pensato/adattato.

Limiti dell'esperienza e prospettive future

L'esperienza descritta presenta tuttavia alcuni limiti:

- Il *prompt*
  - È molto denso, considerati i margini ai quali lo studente si deve attenere
  - ha una scansione poco usuale, anticipando la valutazione dell'esperienza ("*cosa ti è piaciuto di più e cos'altro avresti voluto fare"*) alla descrizione e alla narrazione
  - Presenta parti che implicano funzioni e strutture superiori all'A2 ("*cosa ti è piaciuto di più e cos'altro avresti voluto fare*"; cfr. Spinelli & Parizzi, 2011)
- La griglia iniziale (menzionata in Spinelli, 2014) è forse pensata per altri contesti: può non riferirsi all'ambito della classe (quindi non valere come *achievement test*); è probabile si riferisca a un *proficiency test* (non abbiamo però accesso alle *test specification*). In questo senso l'operazione condotta può presentare alcune forzature
- Il numero degli scritti è esiguo

Gli sviluppi per eventuali analisi future possono essere i seguenti:

- Ripetere l'esperienza su un *corpus* di scritti più ampio e considerare se le divergenze si riducono
- Far analizzare la griglia ai valutatori *prima* della valutazione degli scritti, in modo che eventuali perplessità possano essere discusse a monte
- Concedere sufficiente tempo ai valutatori di riflessione personale in merito a proposte di modifica della griglia, per passare poi a un confronto con gli altri, spostando alla fine l'intervento del formatore. Di contro, affidarsi alla sola condivisione *in plenum* può favorire il valutatore più persuasivo e inibire quello più accomodante, così come la sola analisi *a tergo* da parte del formatore può condurre a interpretazioni nelle quali non tutti i partecipanti si riconoscono.

Riferimenti bibliografici

Alderson, C., Clapham, C. & Wall, D. (1995). *Language Test Construction and Evaluation*. CUP: Cambridge.

Attestato, ADA (2013). Alma: Firenze.

Balì, M. & Rizzo, G. (2002). *Espresso. Corso di italiano 2*. Alma: Firenze.

Carr, N.T. (2011). *Designing and Analyzing Language Tests*. OUP, Oxford.

Cattana A. & Nesci, M.T. (2004). *Analizzare e correggere gli errori*. Guerra: Perugia.

Charney, D. (1984). The Validity of Using Holistic Scoring to Evaluate Writing: A Critical Overview. *Research in the Teaching of English*, 18, 65-81.

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. CUP: Cambridge.

Hedge, T. (1991). *Writing.* OUP: Oxford.

Hughes, A. (2003). *Testing for Language Teachers*. CUP: Cambridge.

Lim, G. S. (2011). The Development and Maintenance of Rating Quality in Performance Writing Assessment: A Longitudinal Study of New and Experienced Raters. *Language Testing*, 28, 543-560.

Lumley, T. (2002). Assessment Criteria in a Large-scale Writing Test: What do They Really Mean to Raters? *Language Testing*, 19, 246-276.

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

Shohamy, E., Gordon, C. & Kraemer, R. (1992). The Effect of Rater's Background and Training on the Reliability of Direct Writing Task. *Modern Language Journal*, 76, 27-33.

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

Spinelli, B. & Parizzi, F., (2011). *Profilo della lingua italiana*. La Nuova Italia: Firenze.

Tankó, G. (2005). *Into Europe: The Handbook of Writing*. The British Coucil: Budapest.

Widdowson, H. G. (1983). *Learning Purpose and Language Use*. OUP: Oxford.

Weingle, S. C. (1994). Effects of Training on Raters of ESL Composition. *Language Testing*, 11, 197-223.

Weingle, S. C. (2002). *Assessing Writing*. OUP: Oxford.

Appendice 1

Scheda di valutazione per la produzione scritta, livello A2 (menzionata in Spinelli, 2014).

a) Versione originale

Giudizio	CONTENUTO (realizzazione del compito ed effetto sul lettore)	VOCABOLARIO (ampiezza e appropriatezza)	ORGANIZZAZIONE (coerenza e coesione)	ACCURATEZZA (ortografia e errori di grammatica)
Ottimo	Il contenuto è chiaro e comprensibile. Gli obiettivi del compito sono stati pienamente raggiunti. L'argomento è stato trattato pienamente. Il messaggio è stato chiaramente trasmesso al lettore.	Vocabolario ampio e sempre appropriato.	Chiara progressione delle idee che sono ben collegate. Buona introduzione e conclusione.	Buona ortografia. Non ci sono errori che impediscono la comprensione.
Buono	Il compito è stato realizzato con errori minimi che non compromettono la chiarezza del testo. In generale, il messaggio è stato comunicato chiaramente al lettore.	Vocabolario in generale appropriato con alcune carenze lessicali.	Basica, con possibili frasi formulaiche e uso di semplici elementi di connessione (ad esempio, "ma", "e", "perché").	Ci sono pochi errori grammaticali e ortografici che impediscono la comprensione.
Adeguato	Il compito è stato realizzato solo parzialmente. Alcuni errori impediscono la comprensione. Il messaggio è solo parzialmente comunicato al lettore.	Vocabolario limitato che impedisce parzilamente la comunicazione.	Basica, con possibili frasi formulaiche. Scarsi tentativi di connessione.	Comunicazione parzialmente impedita da errori di ortografia e di grammatica.
Non adeguato	Il compito non è stato realizzato. Molti errori impediscono la comprensione e la lettura richiede uno sforzo da parte del lettore.	Troppo limitato per la realizzazione del compito.	Parole e isolate, mancanza di coerenza.	Comunicazione severamente limitata da errori di ortografia e di grammatica.
Debole	Il compito non è stato realizzato. Non ci sono abbastanza informazioni per essere comprensibile.	Troppo limitato per essere valutato.	Produzione troppo limitata per essere valutata in base all'input richiesto.	Non sufficiente per essere valutata.

(Scheda di valutazione per la produzione scritta di livello A2)

b) Versione adattata e sottoposta ai valutatori che operano presso l'IIC di Lima

	CONTENUTO (realizzazione del compito ed effetto sul lettore)	VOCABOLARIO (ampiezza e appropriatezza)	ACCURATEZZA (ortografia ed errori di grammatica)	ORGANIZZAZIONE (coerenza e coesione)
10 **9**	**Il contenuto è chiaro e comprensibile**. **L'argomento è stato trattato pienamente.**	**Vocabolario ampio e sempre appropriato**.	Buona ortografia. **Non ci sono errori che impediscono la comprensione.**	**Chiara progressione delle idee**: ben collegate. Buona introduzione e conclusione.
8 **7**	**Il compito è stato realizzato adeguatamente**; il testo si può dire abbastanza chiaro.	**Vocabolario in generale appropriato con alcune carenze lessicali**.	Ci sono **pochi errori grammaticali e ortografici** che impediscono la comprensione.	Una buona organizzazione; anche se **alcune parti del testo non sono adeguatamente sviluppate.**
6 **5**	**Il compito è stato realizzato parzialmente**. Alcuni errori impediscono la comprensione. **Il messaggio è solo parzialmente comunicato al lettore.**	**Vocabolario limitato**.	Comunicazione parzialmente impedita da **numerosi errori di ortografia e grammatica**	Una **organizzazione sufficiente**, con frasi formulaiche. **Uso di semplici elementi di connessione** (ad esempio, "ma", "e", "perché".
4 **3**	**Il compito è stato realizzato al minimo**. Molti errori impediscono la comprensione e **la lettura richiede uno sforzo da parte del lettore**	Vocabolario limitatissimo	Comunicazione severamente limitata da **moltissimi errori di grammatica e di ortografia.**	Frasi brevi con scarsa connessione; poca fluidità nel testo.
2 **1**	Ai limiti del valutabile. Prestazione minima.	Ai limiti del valutabile. Prestazione minima.	AI limiti del valutabile. Prestazione minima.	Parole isolate, mancanza di coerenza.
0	Non valutabile	Non valutabile	Non valutabile	Non valutabile

c) Versione soggetta a riscrittura (le modifiche sono rappresentate in rosso)

	CONTENUTO (realizzazione del compito ed effetto sul lettore)	VOCABOLARIO (ampiezza e appropriatezza)	ACCURATEZZA (ortografia ed errori di grammatica)	ORGANIZZAZIONE (coerenza e coesione)
10 **9**	Il contenuto è chiaro e comprensibile. I punti sono stati ripresi e sviluppati. Limiti rispettati (+/-10%).	Vocabolario ampio e appropriato. Ci possono essere difficoltà legate a limitazioni lessicali (soprattutto quando si trattano argomenti non familiari).	Buona ortografia. Non ci sono errori che impediscono la comprensione: comunica con ragionevole correttezza. Gli errori sono soprattutto legati alla volontà di trasmettere pensieri complessi. Ci possono essere interferenze da altre lingue.	Elegante e snella progressione delle idee: ben collegate. Buona introduzione e conclusione.
8 **7**	Il compito è stato realizzato adeguatamente. La maggioranza dei punti è stata sviluppata. Il testo si può dire abbastanza chiaro. Limiti rispettati (+/-10%).	Vocabolario appropriato con alcune carenze, che possono essere sopperite da ripetizioni, semplificazioni, calchi.	Ci sono pochi errori grammaticali e ortografici che impediscono la comprensione. Sporadici errori di base (A1). Ci possono essere interferenze da altre lingue.	Una organizzazione adeguata; il testo presenta caratteristiche di sufficiente compattezza.
6 **5**	Il compito è stato realizzato parzialmente. Solo qualche punto della consegna è stato sviluppato. Il messaggio è solo parzialmente comunicato al lettore. Limiti rispettati (+/-10%).	Vocabolario limitato. Utilizza prevalentemente espressioni memorizzate che gli/le consentono di comunicare in risposta a bisogni di tipo concreto (parlare di sé, dare e chiedere informazioni, esprimere *routine*).	Comunicazione parzialmente impedita da numerosi errori di ortografia e grammatica. Controllo non sufficiente delle strutture di base (A1), come per esempio l'uso dei tempi e gli accordi. Ci possono essere interferenze da altre lingue.	Una organizzazione elementare, con frasi formulaiche. Possibile uso di semplici elementi di connessione (ad esempio, "ma", "e", "perché"). In alcune parti, può darsi una scarsa tenuta (l'autore può parere non riesca a gestire il filo del discorso).
4 **3**	Il compito è stato realizzato al minimo. Nessun punto è stato sviluppato adeguatamente. La lettura richiede uno sforzo da parte del lettore: è difficile farsi un'idea di cosa voglia dire il testo, se non si legge la consegna. Limiti non rispettati (-30%).	Vocabolario molto limitato. Repertorio formato da espressioni semplici.	Comunicazione severamente limitata da moltissimi errori di grammatica e di ortografia. Scarso controllo delle strutture di base (A1). Numerosissime interferenze da altre lingue.	Frasi brevi con scarsa connessione; poca fluidità nel testo.
2 **1**	Ai limiti del valutabile.	Ai limiti del valutabile.	AI limiti del valutabile.	Parole isolate, mancanza di coerenza.
0	Non valutabile	Non valutabile	Non valutabile	Non valutabile

Appendice 2

Le prove scritte di studenti di livello A2, forniteci dalla Scuola Edulingua.

S1

Carino amicci, Roma è la citta più grossa di italia, avere molta storia di arte, molto culturale, avere molta escultura, molti monumenti ma è il coliseo simbolo de cristo. Citta e un percorso lungo con molto salite e scale un po stanco per cuesto

Lo megliore venire en otogno e primavera ma in cuesta citta sono encontrato un persono ha reconuciuto ero argentino me ha ditto maradona buona cocaica per cuesto conocere il paige no è buono.

Ci benite a italia a estudiare la sua lingua dovete estudiare in scola Edulingua è lo megliore que potere fare, estudio, viaggio, il trato de tutti, il personal è bellissimo.

S2

Cara Daniela,

Fra tanto tempo che non ci vediamo. Io vorrei raccontare la mia esperienza in Italia.

Io sono andata in Italia per studiare italiano nella scuola di lingue cultura Edulingua a San Severino Marche.

Ho conosciuto a persone di tante paese, adesso ho amici per tutti il mondo. Anche gli insegnanti sono molto gentile.

Abbiamo visitato con la scuola a Cingoli, a Roma, a Venezia, a Firenze, a Siena, a Pisa ed a Perugia.

Mi è piacuto di più la città di Firenze perché è una città d'arte, cultura e cibo italiano buonissimo.

Mi sono perduta in Firenze per 5 ore, ma al fine io ho contrato a mia amica e tutto bene. Al posto tuo Daniela, io andrei a Firenze e anche a Venezia per prendere una gondola. Un bacio

S3

Cara Isa,

io sonno revenuta della Italia. Sonno andatta con i mie Maritto, esere in la scuola. Ma piaciutto molto, mi sono divertita.

Ma io a la mattina mi sonno alzatti a le 7 tutti i giorni, per potere studiare, questo è molto bene per me.

Io ho incontrate belli personi, et attenti. Io ho dovutto lasciare dil treno perché la maquinetta non fontionaba et il controlore a ditto di lasciare.

Quando noi ci vediamo io e te racontare cosi di interesante per tu viaje future.

S4

Cara Lucia

mi ha piaciuto la lezione di lingua e avrei voluto fare più escursioni

ho trovato persone molto gentili nel paese e nelle scuola

ieri ho dimenticato la giaca e mi sono bagnata con la pioggia.

Vieni sensa paura e porta molti euro.

È stata una esperienza molto ricca, per me.

S5

Ciao Emilia, sono molto felice di stare in Edulingua e o prenditu molto. La classe è dinamica e divertente. Ho encontrato amice di Argentina e Brazil sono simpatice.

Un episodio divertente e de paura e che io me ho dormito di Castelraimundo a San Severino, cuando ho preguntado al autista me ha ditto che San Severino è 15 km prima, che paura lui a chiamato a Rosella lei ha mandato un autista per me.

Aspetto che tu retornerebbe a Edulingua in vacanza. Ciao amico.

S6

Cara Sofia,

ti voglio racontare il mio mese a San Severino Marche. Tutto è stato incredibile! Ho conosciuto tanti luoghi e persone che seranno sempre nella mia testa e cuore.

Prima di tutto, debi sapere che queste paessino è bellisimo e tranquilisimo, la scuola è dentro un Palazzio antico ma molto carino. Tutte le persone che ho conosciuto nella scuola sono buone, lor hanno fatto il mese più allegro.

La città che mi ha piaciuto di più è Firenze, è belisima e piena di vita. Ho mangiato tutti il tempo. Principalmente gelati pizze e pannini! Quasi dimentico la NUTELLA!!! I civi è stato meraviglioso in tutti i luoghi. Tu devi venire a questa scuola per imparare la lingua e tante cose sul la vitta! Il miglio viaggio di tutti! Bacci.

S7

Cara Cristina, il mio viaggio in Italia è stato molto allegro.
Nell'aeroporto ricevimento di Ana e Alessandra più buono.

Elisa mia insegnante è sempre felice e pronti a rispondere dubbio, sue lezioni sono divertenti e con partecipazioni de tutti gli studenti.

Gli escursioni organizzata sono interessanti per conoscere la regione e la cultura.

Io ritornerò al Brasile con molto amore per questo Paese.

Un episodio più divertente "il bagno di Elza in Venezia" en una nave di ritorno, tipo, ti spiego meglio. Baci e abbracci