ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

Laboratorio Itals (itals@unive.it)
Dipartimento di Studi Linguistici e Culturali Comparati
Università Ca' Foscari, Venezia

Recensione a cura di Paolo Torresan

AUTORE: **N. T. Carr**

TITOLO: **Designing and Analyzing Language Tests**

EDITORE: **Oxford University Press**

LUOGO: **Oxford**

ANNO: **2011**

Il libro scritto da Nathan Carr ha due pregi: (a) illustra i concetti-chiave della valutazione linguistica in maniera chiara (tra i destinatari ci sono, infatti, gli studenti dei suoi corsi di *language testing*); (b) istruisce su come gestire il calcolo dei valori più comuni e diffusi nella valutazione linguistica mediante un foglio di calcolo Excel (più accessibile del costoso SPSS).

Cogliamo l'occasione di questa recensione per passare in rassegna i concetti e gli indici salienti nella valutazione linguistica. Mettiamo in evidenza in giallo nel sommario qui sotto i capitoli che risultano particolarmente interessanti a chi si avvicina al mondo del *testing*.

Cap.	Titoli	Contenuti
I	What are we testing and why?	Coordinate generali sulla valutazione linguistica
II	The building blocks of tests	Tipologie di test
III	Planning and designing the test	Specifiche generali
IV	Writing the specifications for individual test tasks	Specifiche legate alle singole attività
V	Writing the test	Tecniche per la valutazione linguistica; nozioni di *item writing*
VI	Consistency of measurement	Affidabilità
VII	Rating scales for extended production tasks	Griglie analitiche per la valutazione dello scritto/orale
VIII	Validation	Validazione
IX	Test Administration	Gestione del test

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

X	Other important topics in testing	Test informatizzati, *standard setting*, valutare la competenza linguistica di bambini, uso dell'Analisi del Discorso, ecc.
XI	Organization Data in Microsoft Excel	Nozioni generali di Excel
XII	Descriptive Statistics and Standard Scores	Statistica descrittiva e relativi valori
XIII	Creating Graphs and Charts in Excel	Realizzazione di grafici e diagrammi in Excel
XIV	Correlation	Correlazioni
XV	Item Analysis for Norm-Referenced Testing	Analisi degli *item* di test nel contesto di prove riferite a una norma
XVI	Item Analysis for Criterion-Referenced Testing	Analisi degli *item* di test nel contesto di prove riferite a dei criteri
XVII	Distractor Analyses and Scoring Multiple-choice Items in Excel	Analisi dei distrattori tramite Excel
XVIII	Reliability: Scoring Consistency for Norm-referenced Tests	Affidabilità: coerenza dei punteggi nel contesto di prove riferite a una norma
XIX	Reliability: Scoring Consistency for Criterion-referenced Tests	Affidabilità: coerenza dei punteggi nel contesto di prove riferite a dei criteri

Procediamo con ordine nella rassegna dei capitoli evidenziati.

Il capitolo I si intitola "What are we testing and why?". I concetti chiave sono i seguenti.

1. I test sono strumenti pensati in funzione delle decisioni che si devono prendere. Così come un cacciavite è uno strumento che non va bene per tutti i lavori di casa, alla pari i test non sono interscambiabili tra loro. L'esempio dell'autore è chiaro: non possiamo prendere un test pensato per valutare la competenza comunicativa di un controllore di volo e usarlo per valutare la competenza comunicativa di un medico. Tuttavia, ciò non vale solo nel caso delle microlingue; più in generale, anche quando si adottano test generalisti (ad esempio i test di una certificazione) in sede d'aula (al termine di un qualsiasi corso di lingua), occorre chiedersi se il profilo dei candidati a cui si rivolge l'*item writer* della prova di certificazione sia simile a quello dei candidati a cui si intende somministrare la stessa prova.

2. Molte nomenclature in uso sono puramente indicative. Si pensi all'opposizione *test oggettivo/test soggettivo*: il primo si riferirebbe a risposte chiuse sulla cui interpretazione non vi è molto da dire, il secondo invece a risposte aperte, in cui invece il giudizio del valutatore esercita un grande peso. Benché gli aggettivi possano ricevere un qualche credito nella misura in cui descrivono l'attribuzione del punteggio, non possiamo dire che un test di lettura o di ascolto o di grammatica che in genere sono definiti "oggettivi" non presentino margini di

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

soggettività: la scelta del testo, del formato, del numero degli *item*, ecc. sono, infatti, frutto di una scelta da parte di chi confeziona la prova. Così, al contrario, la costruzione di un test scritto/orale e la valutazione stessa dell'abilità orale/scritta possono essere molto sorvegliate, e i margini di soggettività possono essere molto contenuti.

Lo stesso discorso può essere ripetuto per l'opposizione *test diretti* (si desume il grado di abilità alla luce dell'uso della lingua) e *test indiretti* (il grado di un'abilità è ricondotto all'esecuzione di compiti molto specifici e guidati (si testa, per esempio, l'abilità di produzione orale ad un livello elementare mediante la sollecitazione di risposte orali a domande scritte). I *test diretti*, legati all'uso, vengono considerati in genere più idonei e autentici; eppure, anche in questo caso, il valutatore non è al riparo da elementi che possono inficiare la prova: fattori come la scarsa familiarità con il tema, l'ansia, il disagio – all'intero di un'interazione - di fronte a un partner eccessivamente prolisso, la timidezza, la mancanza di *rapport* con l'intervistatore, la non comprensione di quanto richiesto nella consegna possono infatti offuscare, anziché rivelare in maniera trasparente, la competenza.

Lo stesso discorso vale, ancora, per l'opposizione *formativo/sommativo*. Uno studente, per dire, può pur trarre vantaggio da un test di fine corso (massima espressione della sommatività), posto che vi sia un adeguato *feedback* (il che garantisce, in ogni caso, una valenza formativa al test).

In definitiva, per Carr è bene interpretare le coppie delle classiche opposizioni come gli estremi di *continuum*; nella realtà delle cose, infatti, le proprietà non sono così esclusive.

3. Nessun test è perfetto. Dobbiamo immaginare l'esito di ogni prova come una diapositiva dai contorni sfumati. Sono inevitabili degli errori di misurazione. Ciò implica due conseguenze: (a) un lavoro tenace per ridurre i margini di inevitabile approssimazione, al fine di generare un giudizio quanto mai rispettoso della reale competenza del candidato (il tema si collega a quello della *fairness*); (b) un *item writer* è portato a maturare un atteggiamento di distacco verso il proprio prodotto, sollecitando lo sguardo critico di altri per poterlo migliorare. Leggiamo nel libro un passaggio significativo a tal proposito (100; il corsivo è nostro):

> "Various members of this test development team will, as a rule, be assigned specific portions of the project, with the rest of the committee reviewing each person's work. Perhaps the most emotionally sensitive portion of this process is the reviewing of test items. *Newcomers to this process must quickly learn to grow thick skin, and not take personally the inevitable criticism of their passage, questions, and prompts*. Indeed, for many people, the first time they are told – by colleagues, even friends – that several of their items were unclear, or not

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

> congruent with the specifications, their reaction is similar to that of a new mother who has been told her baby is ugly. Unfortunately, sometimes the "baby" really is ugly, and the task needs revision or replacement. This should not been taken as a sign of personal criticism, and should be accepted in good grace as port of an important and ongoing learning process. It is worth keeping in mind that the more tasks a particular teacher creates, the more their work will draw criticism.
>
> The process of reviewing items, passages, keys, and prompts needs to be approached collegially. One person should be in charge of recording all changes that are agreed upon, and whenever possible, decision should be made by consensus. If one person is officially in charge of the group, it is important that he or she does not enforce decisions upon the others, unless there are tremendous differences in test-writing expertise. Likewise, being the loudest and more stubborn group member does not make one's ideas more correct than everyone else's. Interestingly enough, there appears to be very little empirical research on the testing committee dynamics (but see Kim *et al.* 2010, for a rare example."

In aggiunta al carattere della perfettibilità, ogni test si regge su un sottile compromesso tra le seguenti esigenze descritte da Bachman e Palmer (1996):

- *Validità* (il rispetto del costrutto di riferimento);
- *Affidabilità* (coerenza della prova in sé stessa, riduzione dei margini di errore di misurazione);
- *Praticabilità* (attenzione alle esigenze logistiche);
- *Impatto* (conseguenze sociali del test; in particolare, se riferite all'apprendimento/insegnamento, si parla di *washback*);
- *Autenticità* (il fatto che il test non misuri tanto una competenza esercitata tra le pareti scolastiche ma sia specchio di operazioni che lo studente può condurre in ambito extrascolastico; ciò contribuisce alla cosiddetta 'ecologia' del test).

Potremmo aggiungere, peraltro, una sesta caratteristica: *la trasparenza*. Il candidato deve sapere preventivamente come sarà valutato; inoltre ha il diritto di ottenere le spiegazioni che giustifichino il perché di un certo punteggio nel quale egli non si riconosce.

Il capitolo II si intitola "Task – the Building Block of Tests".

Carr ricorre al concetto di *task* come iperonimo che abbraccia tutti i tipi di tecniche/attività che un item *writer* può elaborare. In realtà, sarebbe più

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

appropriato parlare di "tecniche" o di "attività", come avviene in Alderson (2000) (Bachman [1990], a sua volta, parla di "test method", termine infelice posto che rimanda al concetto di metodo, che ha una portata molto ampia in didattica; Famularo [2008] parla, a sua volta, di "response format").

Facendo fede alle classiche distinzioni tra

- attività chiuse o strutturate ("selected response task");
- attività semistrutturate o semiaperte o a risposta limitata ("limited production task");
- attività aperte ("extended production task");

Carr evidenzia quanto segue

1. Le attività semiaperte sono più difficili rispetto a quelle chiuse (è più difficile formulare la risposta a una domanda aperta rispetto al fatto di scegliere un'opzione di una scelta multipla; su questo punto si vedano anche Brown 1980; Shohamy 1984; Chapelle, Abraham 1990; Wolf 1993). Ciò è vero, considerato anche – aggiungiamo noi - il fatto che gli esercizi di discriminazione prevedono pur sempre un margine di "guessing".

2. Carr osserva inoltre come (31) "selected response tasks […] tend to be less authentic than their limited production equivalents". **Vi è quindi un deficit di autenticità nelle attività strutturate**.

3. In riferimento ad attività semiaperte vale la pena prevedere un *corpus* di chiavi accettabili (eventualmente tale *corpus* può essere compilato a partire da risposte fornite dai candidati cui la prova è stata somministrata in fase di *pre-testing* o di pilotaggio).

4. **Il C-test favorisce un certo stile di apprendimento** (quello di chi è indipendente da campo, quindi di chi è più analitico, rispetto a quello che è più dipendente, e quindi più olistico). Se così è, ciò lascia supporre che fattori esterni al costrutto in riferimento al quale questa tecnica è ideata (il costrutto è la comprensione scritta) possano esercitare una insidia alla validità di questa tecnica. Tali considerazioni possono del resto essere estese a tutte le forme di riempimento lessicale, inclusi i *cloze lessicali facilitati*.

5. La cura da prestare nella formazione delle coppie in sede di valutazione dell'interazione orale (come ben illustrato, peraltro, nella ricerca condotta in Ockey 2009).

6. L'inappropriatezza, al fine di valutare la competenza comunicativa, di test basati su **interventi orali preconfezionati**. I candidati possono, in tal caso, memorizzare per filo e per segno un discorso, salvo poi essere completamente

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

spiazzati di fronte a una domanda improvvisata posta dall'interlocutore/insegnante.

7. Riprendendo un'osservazione di Xi (2005), Carr sottolinea che **eventuali grafici che accompagnano i *prompt* devono essere semplici**, onde evitare di confondere i candidati che hanno meno dimestichezza con questo genere di testi.

6. **L'inadeguatezza della traduzione per valutare singole abilità** (la lettura o la scrittura). Benché possa correlare con esse, la traduzione riflette un costrutto a sé: un'abilità di mediazione intertestuale. Pertanto non è opportuno ricorrervi per valutare una generale competenza comunicativa.

I capitoli III-IV trattano uno stesso argomento, con un grado di approfondimento via via maggiore. Il III si intitola "Planning and Designing the test"; il IV, invece "Writing the Specifications for Individual Test Tasks". I punti di maggiore rilievo dei due capitoli sono i seguenti.

1. La definizione delle funzioni delle Specifiche. Le Specifiche sono un testo che accompagna una prova e che ne spiega ogni elemento e giustifica ogni scelta del Certificatore. Esplicitare le caratteristiche di un test, in termini di

- (a) *scopo* (a cosa serve il test);
- (b) *costrutto* (definendolo e chiarendo come la prova lo operativizzi);
- (c) *natura* (la prova fa riferimento a una norma, ovvero il giudizio dell'abilità del singolo viene espresso facendo riferimento all'abilità degli altri, o a dei criteri, vale a dire a degli standard?);
- (d) *domini* (o ambiti di esperienza in cui viene praticata la lingua; es. ambito familiare, accademico, ecc.);
- (e) *profilo del campione* (lingue conosciute, età, bisogni, ecc.);
- (f) *definizione del punteggio a cui è attribuita la sufficienza*;
- (g) *risorse disponibili* (tempi, spazi, materiali, persone)

torna utile per motivi di

- (I) trasparenza verso gli *stakeholders* (tutte le persone interessate);
- (II) controllo interno (nell'atto stesso di esplicitare i criteri, ci si può rendere conto di eventuali incongruenze);
- (III) replicabilità (il *framework* vale come riferimento anche versioni future del test, evitando così che si siano discrepanze tra le sessioni).

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

2. Carr precisa inoltre **la natura aperta delle Specifiche**. Si tratta di un documento destinato a evolvere sulla base delle retroalimentazioni che provengono da diverse fonti (amministratori, partecipanti a pilotaggi, revisori, ecc.).

3. È opinione dell'autore che sia necessario **redigere delle Specifiche anche nel caso si adottino/adattino prove pre-esistenti**, al fine di rendere chiaro perché e come tali prove si conformino al contesto in cui si opera (tale operazione di teorizzazione a posteriori viene chiamata "reverse engineering").

4. Oltre a dimensioni macro, le Specifiche devono anche avere un livello di dettaglio. È necessario cioè **siano esplicitate**

- **le linee-guida in riferimento ai testi oggetto di comprensione** (lunghezza, tipologie, indicazioni inerenti la redazione/recitazione su traccia), alla luce delle abitudini di lettura/ascolto dei candidati;
- **le linee-guida in riferimento alle tecniche;**
- **le linee-guida relative all'assegnazione del punteggio,**
- **le linee-guida relative alla redazione di *prompt* per l'elicitazione di produzioni individuali**, di coppia o di gruppo (con una massima contestualizzazione di ruoli, scopi, funzion), nonché all'**uso di *rating* scale**.

Il capitolo V ha per titolo "Writing the Test". Riguarda la parte più artigianale della valutazione linguistica: il processo di *item writing*. Confezionare prove ben calibrate, prive di incongruenze è un'abilità che si costruisce con il tempo. Un test è, infatti, un genere di testo, e così come costa fatica diventare abili nel redigere un testo giuridico o un testo amministrativo o un testo scientifico, alla pari la redazione di una prova passa per un lento apprendistato. Nel bollettino di qualche anno fa esponemmo una serie di raccomandazioni di cui è bene un *item* writer tenga conto. Qui sintetizziamo alcuni suggerimenti di massima espressi da Carr.

1. Fare in modo che **la risposta a prove di comprensione sia ancorata alla comprensione del testo**, transiti cioè per un processo di decodificazione/interpretazione ("passage dependence") e non sia formulata su altre basi (es. conoscenze di cui il lettore/l'ascoltatore è in possesso);

2. Evitare la sovrapposizione tra gli *item*, e cioè che la risposta ad un *item* possa essere legata alla risposta all'*item* precedente ("maintaining item independence");

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

3. Evitare *item* criptici (i quesiti-tranello), cioè formulati in maniera capziosa, ambigua, concettosa; evitare peraltro, per quanto possibile, formulazioni al negativo: confondono;

4. Evitare opzioni (in una scelta multipla, per esempio) **simili** (es. iperonimo/iponimo; sinonimi): anch'esse possono indurre a confusione (aumentano in ogni caso la difficoltà dell'item);

5. Evitare cloze a intervalli regolari (il *cloze* in senso stretto, in cui viene espunta una parola ogni x parole), poiché inaffidabili.

Il capitolo VI ha per titolo "Consistency of Measurement". L'oggetto del capitolo è una qualità astratta: l'affidabilità ("realiability"). I contenuti sono ripresi nel capitolo XVIII, dove vengono rapportati al procedimento di calcolo in Excel. Vediamo come l'autore articola il suo pensiero in merito all'affidabilità.

1. Spesso il concetto di affidabilità è tradotto in termini di stabilità dei risultati, nell'ipotesi (poco plausibile, peraltro) che un test ripetuto in occasioni diverse (*test-retest hypothesis*) produca risultati identici (o meglio, non troppo divergenti). Al contrario, un test è poco affidabile quando, al variare delle condizioni di somministrazione, la prestazione dell'allievo ne risente (si immagini un candidato che abbia la disponibilità di leggere un testo su carta e, in un'altra occasione, sia costretto a leggere lo stesso test proiettato sul muro: è una situazione realmente avvenuta, a distanza di un anno, in alcune scuole medie brasiliane ed è lecito aspettarsi che la modifica abbia un impatto negativo sull'abilità di lettura dei ragazzini). Parliamo in tal caso di **affidabilità esterna di un test** (nell'esempio riportato, sarebbe più opportuno parlare di *in*affidabilità esterna del test).

Esiste una **dimensione interna dell'affidabilità**, che si riferisce alla "coerenza" (*consistency*) tra le parti che compongono un test: in breve, gli *item.* E cioè si ragiona sul fatto se gli *item* concorrano o meno a una stessa dimensione (il costrutto di riferimento).

Tanto più un test è affidabile (internamente ed esternamente) tanto più è preciso, cioè scevro da fattori che possono deformare il giudizio. Ora, per far cogliere il legame affidabilità=precisione, Carr invita a pensare a una dimensione più tangibile rispetto alla competenza linguistica. Immaginiamo – egli suggerisce– di misurare l'altezza di 100 persone. È inevitabile che ci siano dei margini di errore: una persona per esempio, nel momento in cui le viene misurata l'altezza, può star in punta di piedi, un altro può essere leggermente ingobbito, e così via. Se però noi misurassimo molte volte l'altezza di una stessa persona e facessimo una media, ridurremmo di molto i margini di errore. Ci approssimeremmo cioè al valore "vero"

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

(un concetto astratto ma che ci aiuta a comprendere il fatto che ogni misurazione è un'approssimazione). Maggiore è lo scarto tra il valore osservato e il valore "vero" (che di fatto è, ripetiamo, un valore ideale, astratto), minore è l'affidabilità del giudizio in riferimento alla competenza di un candidato. Perciò è importante che, allo scopo di calcolare l'affidabilità interna di una prova nel restituire la competenza di un candidato, ci si basi su un campione relativamente ampio di candidati e ci si serva di un campione relativamente ampio di *item*. A un maggior numero di rilevazioni corrisponde una maggiore approssimazione verso il valore "vero".

In pedagogia, così come in psicologia, l'indice a cui ci si appella è l'Alfa di Cronbach. Esso mette in relazione la somma delle singole varianze riferite a un *item*[1], con la varianza dell'intero test. Per item dicotomici, il procedimento di calcolo usato è il cosiddetto Kr-20 (al posto delle varianze fa riferimento all'indice di facilità, che corrisponde alla percentuale di risposta corrette; il valore cui si giunge è in ogni caso lo stesso).

Il valore dell'Alfa di Cronbach può oscillare tra 0 e 1: valori prossimi allo zero indicano che l'errore di misurazione ha un effetto altamente distorcente (l'immagine che noi abbiamo della competenza è massimamente distorta; come potrebbe essere qualora un candidato rispondesse a caso, senza comprendere i quesiti); con valori prossimi all'unità siamo indotti a ritenere che i dati, all'opposto, siano 'limpidi', non inquinati da distorsioni[2]. Il livello soglia accettabile nei testi di competenza è ≥0.8[3] (≥0.7 nei test in classe).

A partire dal valore di Alfa, si può calcolare il margine di errore di ciascun *item* (errore standard di misurazione), attraverso la formula SD $\sqrt{1-\alpha}$ (dove SD sta per deviazione standard e α sta per Alfa di Cronbach). Nell'ipotesi di una distribuzione normale il valore "vero" (ideale!) ha il 68% di probabilità (che corrisponde a una deviazione standard) di trovarsi a ±1 errore standard dal valore osservato; e il 96% di probabilità (che corrisponde a due deviazioni standard) di trovarsi a ±2 errori standard dal valore osservato.

[1] La *varianza* è data dal quadrato della *deviazione standard*. La *deviazione standard* (o scarto quadratico medio) è data dalla radice delle somme degli scarti tra la media e i valori osservati.
[2] Occorre in ogni caso fare attenzione. Valori prossimi a 1 possono anche dipendere da *item* equivalenti (riferiti alla stessa informazione, nel caso di attività di comprensione).
[3] Altri autori suggeriscono comunque come soglia il valore di 0.70 (cfr. Green 2013).

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

Il capitolo VII ha per titolo "Rating Scales for Extended Production Tasks". L'oggetto è la costruzione e il controllo di qualità delle griglie di valutazione. I concetti chiave sono i seguenti.

1. L'autore sottolinea **il carattere aperto, ciclico, iterativo, del processo di costruzione di una griglia di valutazione della *performance* scritta/orale**. Una griglia ben fatta innanzitutto si modella sulla base delle componenti del costrutto di riferimento (il quale, tra le altre cose, è sensibile al profilo dei candidati); inoltre prevede una serie ragionata di livelli della *performance* (non troppi né troppo pochi), in riferimento ai quali sono previsti dei descrittori. Tali descrittori devono

- *essere chiari*;
- *rispecchiare i contenuti del curricolo* (se il test è di profitto) o di un *framework* (se il test è di competenza);
- *essere indipendenti* (non far riferimento ai descrittori di livello superiore o inferiore, né far riferimento a altre categorie);
- *essere allineati per livello*;
- *essere soggetti a revisione sulla base del feedback di coloro che usano la griglia.*

2. **È bene che i valutatori siano formati all'uso della griglia** (gli devono essere spiegate le categorie che riflettono le componenti, i descrittori; gli devono essere mostrati esempi di *performance* che si inquadrano a un certo livello).

In caso di situazioni *borderline*, anziché voti intermedi (es. 6,5), che sono da evitare, poiché privi di descrittore, è bene scegliere quale criterio adottare:

- "the best fit"; si sceglie il voto più alto (7, nell'esempio)
- "the lowest sustained level"; si sceglie il voto più basso (5, nell'esempio)

La decisione dipende anche dal tipo di test, e quindi dal tipo di rigore che il valutatore si impone (nel caso per esempio di una professione in cui l'uso della lingua assolve una funzione molto importante, come può essere quello di un controllore di volo, è meglio attenersi al secondo criterio). In ogni caso la scelta da adottare dev'essere assolutamente chiara e condivisa tra i valutatori.

3. **Onde garantire una maggiore affidabilità del giudizio, è bene che ogni candidato venga valutato da almeno due valutatori.** Nel caso di discrepanza, anziché ricorrere a un compromesso (che potrebbe risolversi a favore del valutatore più assertivo, ma non necessariamente più affidabile) è bene coinvolgere un terzo valutatore.

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

Il capitolo VIII ha per titolo "Validation". Il tema è il processo di validazione. I concetti chiave sono i seguenti.

1. Il concetto di validità pertiene al rispetto del costrutto di riferimento (ogni ulteriore specificazione della validità [es. validità di contenuto; validità di ricezione o *face validity*] rimette sostanzialmente alla validità di costrutto).

2. La validità non è tanto un attributo quanto un processo (la "validazione" appunto) **di raccolta di dati** (argomenti" in Kane 1992, 2001) **che valgono a giustificare/sostenere la bontà della prova**.

Il capitolo XII ha per titolo "Descriptive Statistics and Standard Scores". Il tema è la statistica descrittiva. I concetti chiave sono i seguenti

1. Dopo aver precisato come si calcolino i valori di moda, mediana e media, l'autore precisa come il **riferimento alla mediana è importante laddove si abbiano campioni ridotti** (sotto i 30-35 casi) e/o con **distribuzioni asimmetriche** (disallineate a sinistra ovvero con asimmetria positiva, nel caso di test molto difficili; disallineate a destra, ovvero con asimmetria negativa, nel caso di test estremamente facili).

2. Con un campione di oltre 35 individui è possibile che si definisca una distribuzione normale. In realtà molto spesso una situazione di perfetta normalità nel *language testing* è rara (quindi la normalità perfetta è… eccezionale). Spesso cioè abbiamo un'asimmetria dei dati (a sinistra o a destra a dipendere se il test è difficile o facile per la maggioranza). Inoltre il picco può essere molto acuto (curva leptocurtica) o di scarso rilievo (curva platicurtica), cioè possiamo avere una curtosi estrema.[4]

3. Si definiscono i **valori legati alla distribuzione dei dati**:

- **la deviazione standard** o scarto quadrato medio (la formula varia a dipendere se il campione di riferimento è parte di un'intera popolazione o costituisce già da sé una popolazione intera). Essa rappresenta uno scarto attorno alla media. Nell'intervallo ±1 deviazione standard rispetto alla media si concentra il 68,26% dei dati; a un intervallo di ±2 deviazione standard rispetto alla media corrisponde il 95,44% dei dati; a una tripla deviazione il 99,74% dei dati.

[4] Ad ogni modo la distribuzione può essere assimilata, in termini di calcolo, alla condizione di normalità laddove gli indici di asimmetria e di curtosi son contenuti dentro l'intervallo ± 2.

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

- **il punteggio z**. Mentre la deviazione standard è un valore ancorato a un campione di riferimento, il punteggio z è un valore mediante il quale si comparano dati che afferiscono a campioni differenti (e quindi con distribuzioni differenti). Il punteggio zeta è ottenuto dividendo lo scarto tra il valore e la media con la deviazione standard. Il valore che si ottiene ci informa della distanza del punteggio osservato dalla media in termini di deviazioni standard;
- il **punteggio t**, ottenuto riparametrando il punteggio z su una scala 0-100 (si tratta, quindi, sempre di un valore positivo).

Il capitolo XV ha per titolo "Item Analysis for Norm-referenced Testing". Vengono illustrati dei valori riferiti alle proprietà degli item in termini di facilità e di discriminazione; parliamo rispettivamente di

1. Indice di facilità, ovvero della **percentuale dei candidati che hanno risolto l'item**. Per Carr gli item di test afferenti a una prova di competenza devono essere compresi nell'intervallo 0.30-0.70.

2. La correlazione punto biseriale ($r_{p\text{-}bis}$; altrimenti definita anche CICT, *corrected item-total correlation*) mediante la quale emerge se e quanto il singolo *item* correla con l'intera prova. Il valore è contenuto nell'intervallo ±1; valori al di sopra di 0.30 indicano che gli *item* discriminano opportunamente gli studenti più abili dai meno abili. Valori negativi significano, invece, che l'*item* agisce all'opposto di quanto sperato (i più abili sbagliano, i meno abili indovinano).

Carr ci tiene a precisare che, **purché l'*item* discrimini convenientemente, si può talora decidere di andare in deroga ai parametri suggeriti in merito all'indice di facilità** (103):

> Generally speaking, in NRT analyses, discrimination should probably be prioritized somewhat over difficulty. This does not mean that difficulty should be ignored, but if an item has very high discrimination, and the difficulty is a little outside the target range, the item should probably be a lower priority for revision Any changes, in fact, run the risk of reducing the high discrimination value.

Il capitolo XVII ha per titolo "Distractor Analysis and Scoring Multiple-choice Items in Excel". L'oggetto riguarda lo studio dei distrattori. I punti che hanno richiamato la nostra attenzione sono due.

1. La definizione di una **soglia <u>minima</u>** (che potremmo definire "indice di attrattività") **stabilita attorno al 10%** (riprendendo una proposta presente in

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

Bachman 2004). In altre parole, un distrattore che attiri meno del 10% dei candidati non "distrae" propriamente;

2. L'osservazione che **ogni distrattore dovrebbe avere un indice di discriminazione negativo** (i meno abili lo scelgono).

Riferimenti bibliografici

ALDERSON, J. C., *Assessing Reading*, CUP, Cambridge.

BACHMAN, L., 1990, *Fundamental Considerations in Language Testing*, OUP, Oxford.

BACHMAN, L., 2004, *Statistical Analyses for Language Assessment*, CUP, Cambridge.

BACHMAN, L.; PALMER, A., *Language Assessment in Practice*, OUP, Oxford.

BROWN, J. D., 1980, "Relative Merits of Four Methods for Scoring Cloze Tests", *Modern Language Journal*, 64, 311-317.

CHAPELLE, C. A.; ABRAHAM, R. G., 1990, "Cloze Method: What a Difference Does it Make?", *Language Testing*, 7, 121-146.

FAMULARO, L., 2008, *The Effects of Response Format and Test Taking Strategies on Item Difficulty: A Comparison of Stem-equivalent Multiple-choice and Constructed-response Test Items*, Doctoral Dissertation, Boston College, Boston, MA (*Dissertation Abstracts International*, 68, 10, 4268A; UMI N. AAI3283877).

GREEN, R., 2013, Statistical Analyses for Language Teachers, Palgrave, Basingtoke.

KANE, M. T., 1992, "An Argument-based Approach to Validity", *Psychological Bulletin*, 112, 3, 527-535.

KANE, M. T., 2001, "Current Concerns in Validity Theory", *Journal of Educational Measurement*, 38, 4, 319-342.

ISSN 2724-5888
Bollettino Itals
Anno 21, numero 101
Novembre 2023

KIM, J.; CHI, Y.; HENSCH, A.; JUN, H., LI, H.; ROULLION, V., 2010, "A Case Study on an Item Writing Process: Use of Test Specifications, Nature of Group Dynamics, and Individual Item Writers' Characteristics", *Language Assessment Quarterly*, 7, 2, 160-174.

OCKEY, G. J., 2009, "The Effects of a Test Taker's Group Members' Personalities on the Test Taker's Second Language Group Oral Discussion Test Scores", *Language Testing*, 26, 2, 161-186.

SHOHAMY, E., 1984, "Does the Testing Method Make a Difference? The Case of Reading Comprehension", *Language Testing*, 1, 147-170.

WOLF, D. F., 1993, "A Comparison of Assessment Tasks Used to Measure FL Reading Comprehension", *The Modern Language Journal*, 77, 473-489.

XI, X., 2005, "Do Visual Chunks and Planning Impact Performance on the Graph Description Task in the SPEAK Exam?", *Language Testing*, 22, 463-508.