

Sperimentazione e testing

di GIANFRANCO PORCELLI

[pubblicato su *Lingue e Didattica*, a. X, n. 36, aprile 1979, pp. 24-29]

Le prove “oggettive”

Nell'ambito di un'impostazione sperimentale, assume particolare rilievo il problema della raccolta dei dati e della verifica di tutti quei parametri attraverso i quali è possibile giudicare se, e in quale misura, siano stati conseguiti gli obiettivi del progetto. Infatti una sperimentazione, per essere veramente tale, non può prescindere da una verifica oggettiva dei dati di partenza e di quelli terminali, così che i risultati dell'intervento didattico non siano attestati da valutazioni soggettive, basate su impressioni o viziate da un atteggiamento favorevole o sfavorevole verso la sperimentazione, ma emergano invece da rilevazioni affidabili.

In concreto, questa esigenza ci porta a discutere sull'impiego dei test: fino a che punto, e per quali aspetti, si possono ritenere prove “oggettive”? Sul piano tecnico l'oggettività di un test è il prodotto di due fattori primari: la “validità” e la “affidabilità”. Secondo una definizione ormai classica, la validità è l'idoneità di un test a conseguire i suoi fini, ossia l'attitudine a verificare ciò che intende verificare e nient'altro.

È, evidentemente, una definizione troppo generica e in gran parte tautologica; per precisarla prendiamo in esame i vari tipi di validità che sono stati individuati.

La *face validity* (validità “facciale”) è l'appropriatezza del test in rapporto alla situazione, alla popolazione e all'oggetto su cui si opera il controllo. Non è corretto usare un test di profitto finale (“*achievement o proficiency test*) come strumento di verifica iniziale (*assessment*);

un test costruito per studenti di scuola media non potrà, di norma, essere usato nella scuola elementare senza gli opportuni adattamenti; un test che intende controllare la padronanza del lessico non deve coinvolgere altri elementi (per esempio, a livello ortografico o morfosintattico), né può essere impiegato per verificare abilità o contenuti diversi, quali il “saper leggere”.

Un secondo tipo di validità (*content validity o validità “dei contenuti”*) presuppone un'attenta analisi della lingua presa in esame, in rapporto agli obiettivi del corso. Non si tratta quindi soltanto di controllare che non vi siano elementi estranei o spurii, ma piuttosto di garantire che siano presenti tutti gli *items* idonei a verificare il

raggiungimento delle mete didattiche. Molto spesso le scelte si orientano verso quei contenuti che sono più facilmente inseribili in quesiti a scelta multipla o che comunque si prestano meglio a realizzare test di un certo tipo ; al tempo stesso vengono magari tralasciati aspetti e contenuti che sono essenziali, ma che non si sa bene come tradurre in quesiti idonei ad un controllo oggettivo.

Il terzo tipo di validità (*construct validity* o validità “di impostazione”) si riferisce alla capacità di un test di misurare certi tratti caratteristici in coerenza con una data teoria dell'apprendimento linguistico. Se, ad esempio, si parte da un approccio funzionale alla L2, la verifica non deve poi essere condotta attraverso moduli operativi che si collegano invece ad un approccio strutturale o comunque a modelli di apprendimento diversi da quello scelto dall'insegnante.

C'è infine la validità “empirica” o “statistica” (*empirical validity*). Se i risultati di un test vengono confermati da dati affidabili raccolti mediante altri strumenti di verifica, si dice che quel test possiede validità statistica, solitamente espressa per mezzo di indici di correlazione. In questo ambito si usa distinguere tra validità “concomitante” (*concurrent*) e validità “predittiva” (*predictive*). Nel primo caso c'è contemporaneità tra la raccolta dei dati attraverso il test e la rilevazione dei valori che servono come criterio di raffronto; nel caso della validità predittiva si ha invece una conferma dei dati del test attraverso i risultati di prove successive. Questo avviene normalmente quando un test ha funzione prognostica.

L'altro aspetto dell'oggettività di un test, accanto alla validità, è la sua affidabilità, cioè la capacità di condurre a risultati costanti e indistorti. Si ha questo tipo di affidabilità quando un test offre sempre lo stesso punteggio, ovviamente a parità di prestazioni, indipendentemente da chi effettua la correzione. Se due o più correttori giungono a punteggi diversi per un singolo elaborato, o se un test, somministrato due volte allo stesso gruppo di allievi (senza che nel frattempo si siano modificate le variabili verificate) produce una scala di valori significativamente diversa, quel test manca di affidabilità e non è oggettivo.

L'affidabilità viene di solito perseguita mediante quesiti che ammettono una sola risposta esatta, chiaramente determinata secondo criteri di accettabilità precisati esplicitamente, dettagliatamente e preventivamente. È l'aspetto più tipico e più noto dei test, e quello che li fa preferire alle prove non oggettive, nelle quali l'intervento del correttore influisce sui punteggi attribuiti. Il problema dell'affidabilità ha varie implicazioni, a livello sia teorico che pratico, ma in questa sede riteniamo preferibile tornare ad esaminare più dettagliatamente, il versante della validità e affrontare alcuni problemi che si riferiscono alla validazione dei test e al loro impiego in, una sperimentazione.

Problemi di scelta dei contenuti

Per quanto riguarda la validità dei contenuti e dell'impostazione metodologica dei test, vi sono alcuni aspetti specifici del *language testing* su cui va posto l'accento. Ad esempio, siamo soliti distinguere tra lessico attivo e passivo; il concetto in sé è abbastanza semplice e chiaro, ed è certamente utile quando parliamo di metodi, contenuti e obiettivi. Però se cerchiamo di definire esattamente quando un vocabolo debba intendersi acquisito attivamente, ci scontriamo con molteplici difficoltà. Ognuno di noi, se richiesto di definire i criteri coi quali discriminare tra possesso attivo e passivo giungerebbe probabilmente a risposte in certa misura diverse. Ecco quindi che i conti o le liste di vocaboli vanno prese con molta cautela quando su di essi si vogliono poi costruire dei test. Ed è anche per questo che si incontrano difficoltà a preparare dei test su scala nazionale quando l'insegnamento è stato effettuato con materiali prodotti localmente in più centri.

È certamente possibile costruire test di livello “aspecifici”, ossia non legati a particolari materiali didattici (unità didattiche o corsi); questi test di solito si basano sulla preventiva determinazione di un corpus linguistico, in termini di lessico e di strutture morfosintattiche, e di un repertorio di problemi di apprendimento.

Solo recentemente, nel quadro di un approccio funzionale, si è cercato di elaborare dei test muovendo dall'analisi delle abilità linguistiche connesse a determinati impieghi della L2 (ad esempio, per svolgere determinate mansioni in un'azienda). Ma anche in questo caso si tratta di test “d'entrata” e non di test finali. Il loro scopo è di accertare la situazione esistente, su cui intervenire con programmi didattici adeguati. Nell'ambito della scuola elementare si dà invece per acquisito che il punto di partenza, in quanto a profitto nella L2, è uguale per tutti ed è pari a zero. Per ciò che invece si riferisce alle attitudini, alle condizioni psico-fisiche e socio-economiche degli allievi, la situazione di partenza non è affatto uguale, la sua rilevazione richiede l'impiego di test attitudinali, schede tipologiche e altri strumenti, la cui analisi ci porterebbe a varcare i limiti di questo contributo. In fase di controllo del profitto si usano solitamente test “specifici”, ossia basati su particolari materiali didattici, e ciò perché è risultato e-stremamente difficile costruire delle prove tendenzialmente più generiche ma che di fatto non privilegiassero alcuni gruppi di classi o di allievi rispetto ad altri. I problemi risiedono non solo e non tanto nella scelta dei contenuti, quanto piuttosto nella modalità di presentazione e somministrazione. Se, nel quadro di una sperimentazione, si vogliono avere prove comuni, occorre un lavoro preventivo che permetta di concordarle tra le diverse *équipes*, programmato con largo anticipo rispetto al momento in cui i test dovranno essere somministrati.

Interpretazione dei dati

Occorre poi considerare quale valore abbiano i risultati ricavati dalla prima somministrazione dei test. Infatti que-

sti strumenti, che dovrebbero darci la misura del successo (o dell'insuccesso) della sperimentazione, sono essi stessi da verificare, tarare e validare. Le procedure di standardizzazione di un test sono lunghe e complesse, ma qui basterà un breve cenno per avere un'idea dei problemi che vi sono coinvolti.

Ogni quesito o *item* ha un indice di facilità (IF) e un indice di discriminazione (ID). L'IF è semplicemente il rapporto tra il numero di risposte esatte e il totale delle risposte. La facilità del quesito non dipende solo dal suo contenuto ma anche dalla sua formulazione. A parità di altre condizioni, un quesito a due alternative è più facile di un quesito a scelta tripla; la difficoltà di un test è anche determinata dalla chiarezza delle istruzioni, dalla qualità e comprensibilità di eventuali supporti visivi, dalla velocità di somministrazione, ecc. Un IF troppo alto o troppo basso rivela un item poco significativo, a cui spesso si accompagna un ID insufficiente. L>ID esprime la capacità di un item di operare una differenziazione tra i migliori e i peggiori. Se è troppo basso indica che al quesito rispondono altrettanto bene (o male) sia i migliori che i peggiori, e se è negativo (i peggiori rispondono meglio dei migliori), rivela che il quesito è ambiguo, scorretto o mal formulato.

Ecco un esempio di quel che può evidenziare una *item analysis*. Poniamo il caso di un quesito a scelta quadrupla, a cui 100 esaminati abbiano risposto così:

89 hanno scelto A, la risposta esatta

7 hanno scelto B, uno dei distrattori

4 hanno scelto C, un altro distrattore

0 (nessuno) ha scelto D, l'ultimo distrattore.

A parte l'IF probabilmente troppo alto (89%), l'analisi delle risposte errate indica che a dispetto dell'apparenza superficiale il quesito è in realtà a scelta tripla, in quanto uno dei distrattori è così palesemente errato che nessuno, nemmeno tra i peggiori, l'ha scelto. In altri casi questa indagine ci può rivelare quesiti ambigui, che ammettono due risposte, o comunque anomali e quindi da eliminare o revisionare.

Tecniche e testing

Accanto ai problemi connessi con il ruolo del testing nel quadro di un progetto sperimentale, vi sono quelli legati all'individuazione delle tecniche più valide per il controllo dell'apprendimento. Qui accenneremo soltanto alla verifica del saper comprendere e parlare la lingua straniera come esito di un insegnamento a bambini di scuola

elementare, con un programma basato sull'approccio funzionale¹. Per quest'ultimo aspetto bisogna tuttavia ricordare che siamo ancora in una fase di precisazioni e assestamenti che riguardano i *notional-functional syllabuses* nel loro complesso; sul testing, in particolare, sono in corso studi interessanti e promettenti, ma per il momento non disponiamo di risultati che abbiano sufficiente ampiezza e consistenza. I volumi sul *testing*, compresi quelli di più recente pubblicazione, hanno ancora una impostazione che non è fundamentalmente diversa da quella del Lado (1961), almeno per la parte che riguarda le tecniche di controllo delle abilità linguistiche. Nella ricerca attuale vi sono due filoni principali, che si riferiscono ai test basati sul concetto di ridondanza (come i *cloze test*), e a quelli che impiegano tecniche di tipo performativo, in analogia alle attività di *problem-solving*.

Lo scoglio maggiore è costituito dalla verifica “oggettiva” della produzione, e specialmente della produzione orale. Molti autori hanno preferito aggirare l'ostacolo, affermando che non è indispensabile ricorrere a prove che implicano la produzione, poiché è stato verificato che i risultati di test oggettivi hanno avuto un alto indice di correlazione con giudizi attendibili sulla capacità di produzione. In altre parole, si sostiene che si possono usare dei test che non verificano direttamente la produzione linguistica (orale o scritta) ma che tuttavia ci forniscono dei dati affidabili che possono essere assunti come indice della padronanza del saper parlare o del saper scrivere. Non occorrerebbe quindi far eseguire dei temi o dei riassunti per accertare la capacità di comporre o di riassumere: un test con un congruo numero di quesiti sul lessico e sulla morfosintassi ci può dare, secondo i sostenitori ad oltranza dei test “oggettivi”, un'indicazione precisa sulla misura in cui ciascun esaminato possiede tale capacità. Analogamente, la capacità di dialogare, invece di essere accertata attraverso i dialoghi, potrebbe essere inferita mediante prove oggettive sulla discriminazione di suoni, comprensione di frasi, scelta di locuzioni appropriate, ecc.

Questo tentativo di superare il problema non è convincente per varie ragioni, sia di ordine glottodidattico che statistico-metodologico. Sotto quest'ultimo profilo, occorre ricordare come l'esistenza di una correlazione tra due variabili non indica di per sé un nesso causale, ossia un rapporto di causa ed effetto. Un esempio “classico” è dato dall'andamento, nel corso di un anno, delle vendite di gelato e delle morti per annegamento: posti su un grafico, si evidenzerebbe un andamento pressoché parallelo, e calcolando la correlazione con una delle formule disponibili, si ricaverebbe un indice molto alto. Tuttavia, come è ovvio, ciò non significa che una delle due variabili sia la causa (o l'effetto) dell'altra. Esiste una terza variabile, la temperatura atmosferica, che verosimilmente è la causa di entrambe.

Ritornando al tema dell'apprendimento linguistico, riscontriamo qualcosa di analogo. Se ci richiamiamo agli studi

¹ Per una descrizione del Progetto ILSSE (Insegnamento delle Lingue Straniere nelle Scuole Elementari) si vedano i contributi di M. L. SALA, G. RUGGIERO BOELLA, e C. UMANA CEVO in “Lingue e Didattica”, a. IX n. 33, Ottobre 1978.

di Carroll² e al test di attitudine alle lingue moderne da lui elaborato, vediamo che esso abbraccia fattori specificamente linguistici, quali la capacità di discriminare suoni e di associare suoni e simboli, o la capacità di apprendere forme legate tra loro da particolari regole morfo-fonematiche, ma coinvolge anche altri fattori assai meno specifici, quali la memoria e l'aspetto verbale dell'intelligenza. Per la presenza di questi fattori, dobbiamo attenderci un certo grado di correlazione positiva non solo tra le varie abilità nella L2 osservabili in un allievo, ma anche fra il grado di apprendimento di una lingua straniera e le capacità di espressione linguistica in genere. Per questo non è sufficiente avere dei dati-indice correlabili, ma è importante verificare direttamente le abilità linguistiche che costituiscono gli obiettivi di un dato programma.

Stimoli per la produzione orale

Se oggi non accettiamo più acriticamente certe posizioni dei fautori del *testing*, dobbiamo d'altra parte renderci conto che non possiamo accertare la produzione orale usando tecniche rigorosamente oggettive. Fortunatamente l'alternativa non è tra i test, da una parte, e i dialoghi “a ruota libera” dall'altra. Per una verifica sufficientemente strutturata del saper parlare, capace di condurre a dati abbastanza attendibili e quindi comparabili, esiste una strategia che, sul versante dell'allievo, tende a fargli produrre quelle frasi che a noi interessano, e sul versante dell'esaminatore gli fornisce parametri di giudizio che possano limitarne la soggettività.

Per far parlare l'esaminato esistono tre tipi principali di stimoli: quelli verbali nella L2, quelli visivi, e le tecniche di conversazione guidata. Nel primo gruppo rientrano anzitutto le domande nella lingua straniera; è la tecnica più “naturale” ma è assai limitata, in quanto le risposte sono spesso ridotte a brevissimi enunciati (“sì”, “no” o sintagmi elementari) e, nel caso di enunciati completi, si tratta normalmente di frasi dichiarative, prevalentemente affermative. Esistono varie tecniche che consentono di far produrre sia frasi negative che interrogative e, in pratica qualsiasi tipo di enunciato. Alcune di queste tecniche sono del tutto analoghe a quelle degli esercizi strutturali i quali, come è noto, consentono di spaziare in tutta la gamma delle strutture linguistiche. Si tratta però di un'attività logico-formale che va al di là delle normali capacità dei bambini nella scuola elementare, e che inoltre rimane nell'ambito di un'impostazione strutturalista e non funzionale.

La seconda via passa attraverso l'impiego di immagini come stimolo per la produzione orale. Nemmeno questo ap-

² J. B. CARROLL e S. M. SAPON, *Modern Language Aptitude Test*, New York, The Psychological Corporation, 1959 (versione italiana a cura di M. FERENCICH, Firenze, Organizzazioni Speciali, 1964).

proccio è esente da inconvenienti, che sono connessi soprattutto con la polisemia dell'immagine; infatti è estremamente difficile che un'immagine da sola possa condurre inequivocabilmente alla produzione di una frase prestabilita. Un cenno di saluto può essere interpretato come “Ciao!” in italiano, ma come “*Hello!*” oppure “*Bye-bye!*” in inglese (a seconda che il saluto avvenga nel momento in cui ci si incontra o in cui ci si lascia). Per superare queste difficoltà si possono proporre due alternative principali:

a) Sostituire l'immagine singola con una sequenza di immagini; così facendo è possibile eliminare certe ambiguità, in quanto l'immagine successiva seleziona e precisa certi elementi dell'immagine precedente, ma si corre il rischio di introdurre altre ambiguità se la sequenza non è studiata con la massima cura.

b) Inserire simboli vari, già noti agli esaminati, che precisino l'immagine. Da anni in Francia si è percorsa questa via, raggiungendo un notevole livello di sofisticazione che però in alcuni casi giunge all'ermetismo. Certi simboli sono di facile e immediata interpretazione (come, ad esempio, il punto interrogativo per indicare che si deve produrre una domanda), altri invece implicano processi mentali molto complessi, che talora presuppongono la sicura padronanza di alcuni concetti di analisi grammaticale e logica.

Con i bambini della scuola elementare non è pensabile che si possa percorrere questo itinerario che, oltre certi limiti molto ristretti, richiede tutta una serie di condizionamenti di tipo pavloviano ai vari simboli impiegati. Bisogna d'altronde riconoscere che è difficile, in molti casi, individuare una terza alternativa tra l'immagine ambigua e quella condizionante.

Il terzo approccio alla produzione orale si ricollega alle tecniche della conversazione guidata (drammatizzazione, “dialogo aperto” e *role-playing*). All'esaminato viene descritta una situazione e gli viene assegnato un ruolo ; ad esempio, gli stimoli possono essere del tipo : “Incontri un amico per la strada: salutalo”. Questo permette il controllo non solo della correttezza formale, ma anche della scelta delle espressioni appropriate rispetto alla situazione. Sono evidenti sia i collegamenti con l'approccio funzionale, sia le possibilità di realizzazione pratica e di presentazione del test sotto forma di gioco. Tuttavia nemmeno questo approccio è privo di prevedibili inconvenienti. Anzitutto richiede tecniche di somministrazione tra le più “aperte”, il che significa che se non sono ben condotte possono facilmente degenerare nella conversazione “a ruota libera”, non strutturata, da cui è molto difficile ricavare valutazioni uniformi e comparabili. In secondo luogo comporta l'uso della L1 per la presentazione delle istruzioni (comprendendo in questo termine la descrizione della situazione, l'assegnazione del ruolo, ecc.) ; occorre prestare attenzione a formulare gli stimoli in modo tale che le frasi poi prodotte nella L2 non siano la traduzione diretta. Solo a livelli più avanzati sarà possibile fornire le istruzioni nella L2, facendo attenzione a che non vi siano problemi di comprensione che interferiscano con la valutazione della produzione.

I tre approcci non si escludono necessariamente a vicenda, ma al contrario possono integrarsi opportunamente. Vi sono molteplici modi di associare immagini e stimoli nella L2, oppure descrizioni di situazioni e stimoli visivi, ecc. Le scelte operative concrete varieranno caso per caso, a seconda degli obiettivi, del livello di preparazione, dei materiali disponibili e anche, in certa misura, tra una lingua e un'altra.

Il giudizio sulla produzione orale

Come accennato all'inizio della sezione precedente, una riduzione della soggettività del giudizio passa anche attraverso la definizione di parametri e di criteri di giudizio. I parametri a cui si fa più spesso riferimento sono:

a) la “appropriatezza” della risposta, ossia la pertinenza della reazione dell'esaminato rispetto allo stimolo che gli viene somministrato;

b) la correttezza dell'enunciato a livello fonologico (per quanto si riferisce sia ai fonemi segmentali che ai valori prosodici, intonazione, accento, ritmo, ecc.);

c) la correttezza dell'enunciato a livello morfosintattico e lessicale;

d) la prontezza della reazione agli stimoli: un'eccessiva lentezza nella produzione può indicare che l'esaminato sta traducendo mentalmente o che comunque non ha sufficientemente interiorizzato certi meccanismi.

In quanto ai criteri di giudizio e alle tecniche di assegnazione del punteggio, la procedura più semplice, ma anche più rozza, è quella del “tutto o niente”: si stabilisce un livello di accettabilità, al disopra del quale si dà un giudizio positivo e al di sotto del quale il giudizio è negativo, senza gradi o sfumature. Fino a qualche tempo fa l'alternativa era, di solito, tra un enunciato perfettamente costruito e appropriato e un enunciato imperfetto e quindi non accettato come valido, anche se solo lievemente scorretto. Oggi il criterio discriminante che si preferisce assumere è l'idoneità dell'enunciato a realizzare la comunicazione, eventualmente anche in presenza di errori non gravi.

Combinando i due criteri si ottiene una scala di tre punti (enunciato corretto/ enunciato scorretto ma comprensibile/enunciato incomprensibile) che può essere di pratica utilità in molte situazioni di *testing*. La costruzione di scale più sofisticate richiede un'attenta precisazione e puntualizzazione dei parametri; ad esempio, i quattro parametri sopra indicati, se esplicitati meglio e collocati in ordine di importanza, possono dare luogo ad una scala di cinque o sei punti, che va dal livello minimo (frase non pertinente o incomprensibile) al livello massimo (frase corretta anche nell'intonazione e nel ritmo, eseguita con immediatezza), passando per i vari gradi. Queste scale sono difficili da usare “sul campo”; il loro impiego richiede di solito che gli enunciati siano registrati e che l'esaminatore possa lavorare in condizioni ottimali, riascoltando l'esaminato tante volte quante è necessario per

attribuire il punteggio correttamente.

Conclusioni

In queste brevi note non si è inteso delineare una strategia di *testing*, che richiederebbe ben altro spazio e approfondimento, ma soltanto accennare ad alcuni temi che si sono rivelati di preminente importanza ed interesse nell'ambito di un progetto sperimentale di insegnamento delle lingue a livello di scuola elementare. Ci ripromettiamo di ritornare sull'argomento in maniera più dettagliata in una prossima occasione, riferendo su specifiche esperienze di *testing*, già in corso o che stanno per essere avviate. In tale sede sarà possibile una valutazione più accurata del ruolo e dell'utilità di un programma di *testing* nella scuola elementare.

Il mito del “test oggettivo” come panacea che risolve tutti i problemi di valutazione è oggi al tramonto. Nel campo delle lingue straniere sono ora ben chiari i limiti e le precondizioni affinché le prove non tradizionali siano valide ed efficaci, ed è questa consapevolezza, che ha preso il posto di un'accettazione acritica, che lascia intravedere nuove prospettive di sviluppo per i test.

Bibliografia

A) Verifica dell'apprendimento linguistico.

A. AMATO, *11 Testing nella didattica linguistica*, in “Rassegna italiana di linguistica applicata”, VI, 1-2, Gennaio-Agosto 1974, Bulzoni, Roma (numero monografico a cura di A. Amato).

J. P. B. ALLEN e ALAN DAVIES (Eds.), *Testing and Experimental Methods*, ECAL vol. 4, OUP, London 1977.

J. C. CATFORD, *Some Aspects of Linguistics in Language Testing*, in “English Teaching Forum”, XIII, 3-4, 1975, pp. 316-319.

A. DAVIES (ed.), *Language Testing Symposium*, Oxford University Press, Londra 1968.

D. B. HARRIS, *Testing English as a Second Language*, McGraw-Hill, New York 1969.

J. B. HEATON, *Writing English Language Tests*, Longman, Londra 1975.

R. LADO, *Language Testing*, Longman, Londra 1961.

J. E. LOWE, *Are you Fair in Your Grading ?* in “English Teaching Forum”, XIII, 3-4, 1975, pp. 322-327.

G. D. PICKETT, *A Comparison of Translation and Blank-filling as Testing Techniques*, in “English Language Teaching”, XXIII, 1, Ottobre 1968.

G. PORCELLI, *II Language Testing: problemi e tecniche*. Minerva Italica, Bergamo 1975.

J. A. UPSHUR, *Objective Evaluation of Oral Proficiency*, in "English Teaching Forum", XIII, 3-4, 1975, pp. 328-333.

B) *Metodologia della ricerca. Costruzione e validazione dei test. Elaborazione dei test, ecc.*

D. ADKINS WOOD, *Test Construction*, Mer-rill Books, Columbus, Ohio 1961.

J. R. AMOS, F. LLOYD BROWN, O. G. MINK, *Statistical Concepts*, Harper& Row, New York 1965.

L. CALONGHI, *Tests e esperimenti*, P.A.S., Torino 1956.

—————, *La redazione dei quesiti nelle prove*, in "Orientamenti Pedagogici", XII, 3, 1965.

M. GATTULLO, *Didattica e docimologia*. Armando, Roma 1971.

—————, *Il controllo delle risposte a caso nei test di profitto*, in "Scuola e Città", XXIII, 6, Giugno 1972.

F. F. GOROW, *Misure Statistiche*, O.S., Firenze 1962.

E. F. LINDQUIST (ed.), *Educational Measurement*, American Council on Education, Washington D.C. 1951 (v. in particolare il contributo di J. C. FLANAGAN, *Units, Scores and Norms*),

G. PORCELLI, *Obiettivi comportamentistici nello studio delle lingue*, in "Lingua e Civiltà", V, 2, Maggio 1972, pp. 15-19.

R. TITONE, *Introduzione alla metodologia della ricerca nell'insegnamento linguistico / Methodology of Research in Language Teaching*, Minerva Italica, Bergamo 1974.