ANNA ZANFEI

VALIDAZIONE: PROCESSI E PROCEDURE PER I TEST DI LINGUA

Estratto da
Quaderni di lingue e letterature
33/2008

VALIDAZIONE: PROCESSI E PROCEDURE PER I TEST DI LINGUA

L'attività di validazione può essere definita come l'insieme delle operazioni attraverso le quali si giudica lo scarto esistente fra gli obiettivi di qualità programmati in sede di progettazione e i risultati effettivamente conseguiti. Nel caso dello sviluppo di uno strumento di valutazione della conoscenza o della competenza costruito ex novo è necessario verificare se questo strumento è consono alle finalità per le quali è stato prodotto. Il processo attraverso cui si compie questa verifica è definito validazione e prevede uno studio dell'affidabilità e della coerenza di test o di altri tipi di strumenti di elicitazione atti a raccogliere dati su una particolare abilità o conoscenza. L'analisi della validità dell'uso del test si svolge attraverso un processo che segue metodi di ricerca qualitativi e quantitativi. La validazione permette di costruire un ragionamento interpretativo basato su ipotesi e confutazioni e di raccogliere prove statistiche sulla funzionalità degli strumenti testistici in relazione al *construct*. Perciò l'adeguatezza e l'appropriatezza delle inferenze basate sulle diverse modalità di *assessment* sono avvalorate in termini rigorosi sulla base delle prove statistiche e del fondamento teorico del giudizio valutativo.

Le attività coinvolte nello studio della validazione comprendono due compiti strettamente interrelati tra loro: il primo è l'articolazione di una ipotesi di interpretazione (*construct*) e il secondo è la raccolta di prove statistiche rilevanti in supporto all'accuratezza delle interpretazioni dei punteggi. Gli aspetti di questa concettualizzazione della validazione riguardano la qualità delle interpretazioni dei risultati di un *assessment* come validi misuratori di una particolare abilità, fermo restando che la validità di un test non è mai una misurazione accurata di quella abilità ma piuttosto l'interpretazione più plausibile in base alle prove statistiche ottenute.

Ovviamente per ogni nuovo test o per ogni nuovo uso di test precedentemente utilizzati ad altri fini è necessario ricorrere al processo di validazione in quanto le statistiche raccolte hanno valore interpretativo specifico per l'uso che se ne vuole fare o l'interpretazione che se ne vuole dare. Messick [1] definisce la validità come un concetto unitario e sostanzialmente atta a

1. S.A. Messick, *The once and future issues of validity: assessing the meaning and consequences of*

ponderare la qualità globale del test. Da ciò consegue che non è plausibile frammentare il concetto di validità in validità di contenuto e validità del *construct*: esiste solo un'unica qualità globale del test. Tuttavia Bachman [2] fa presente che esistono vari studi di validazione creati per esempio esclusivamente per analizzare il contenuto del test, mentre altri analizzano le correlazioni tra diverse modalità di misurazione di una stessa abilità linguistica.

Lo studio di validazione qui presentato riporta le analisi statistiche effettuate sui test di lingua per stranieri costruiti in base al *construct* creato in seno al progetto CERCLU (Certificazione dei Centri Linguistici Universitari) [3]. Il caso qui esaminato riguarda un gruppo di quattro test di livello (B1 e B2) costruiti per verificare la competenza (*proficiency*) nelle quattro abilità di base per la lingua italiana e per la lingua inglese: abilità nel parlato, nello scritto, nella lettura e nell'ascolto. Il progetto CERCLU presenta notevoli novità nell'ambito della valutazione. In primo luogo il construct è basato sul *Framework* [4] del Consiglio Europeo per la valutazione delle abilità linguistiche mentre per la costruzione dei compiti si è avvalso dell'impostazione rigorosamente scientifica delle procedure per sviluppare prove testistiche di grande efficacia formulata da Bachman e Palmer [5]. In secondo luogo i test di comprensione dell'ascolto e della lettura sono somministrati al computer quindi fanno parte dei CBT (*Computer Based Test*) e iBT (*internet Based Testing*) che presentano caratteristiche diverse rispetto ai tradizionali test denominati *paper & pencil*. In terzo luogo la grande novità risiede nella valutazione dei compiti di produzione scritta e orale che consistono di una, peraltro riuscitissima, applicazione delle teorie della *discourse analysis*.

Rispetto alla vastità delle pubblicazioni teoriche e applicative sulla validazione nell'ambito generale della psicometria esiste una letteratura molto ridotta sulla validazione dell'*assessment* e dei test di lingua straniera e ancora più ridotta è la letteratura in merito alla validazione di *Computer Based*

---

measurement, in H. Wainer and H.I. Braun (eds.), *Test Validity*, Hillsdale, Lawrence Erlbaum, 1988, pp. 27-46.

2. L. F. Bachman, *Fundamental Considerations in Language Testing*, Oxford, Oxford University Press, 1995, pp. 18-25.

3. Il *Progetto CERCLU* è nato nell'anno 1998 ad opera dei massimi esperti di nove Centri Linguistici di Ateneo italiani (Bergamo, Roma 3, Roma IUSM, Trieste, Padova, Siena, Verona, Cagliari, Bologna) con lo scopo di creare una serie di test per la verifica della competenza linguistica nella lingua italiana e inglese a livello B1 e B2.

4. Association Language Testers in Europe, *Principle of good practice for ALTE examinations*, http://www.altc.org/can_do/index.php, 2007.

5. L. F. Bachman, Palmer A. S., *Language Testing in Practice*, Oxford, Oxford University Press, 1996.

Tests o *internet Based Tests* nonché ai metodi di valutazione delle abilità di produzione rigorosamente legati alla *discourse analysis*. La letteratura più recente sulla ricerca nel *Language Testing* sta concentrando l'attenzione proprio su questi ambiti emergenti come testimoniano le pubblicazioni in seno alle riviste più accreditate in ambito accademico quali *Language Assessment Quarterly*, *Language Testing Journal*, *TESOL Quarterly*. Inoltre alcuni recenti testi si rivelano fondamentali come riferimento ad uno studio di validazione di questa nuova tipologia di test di lingua. In particolare per quanto riguarda l'impostazione data all'interno del CERCLU dei test di produzione Lazaraton [6] fornisce un rigoroso riferimento teorico e applicativo. L'autrice infatti sottolinea l'efficacia dei test di produzione che sono frutto dell'applicazione di metodi di ricerca qualitativa anche se non deve mai mancare lo studio quantitativo per la validazione ed afferma:

> Although language testing is perhaps still the most positivist in orientation of all areas of applied linguistics, [...] There has been a move from positivistic research focusing on properties of tests reflected in score data toward a broader and more critical examination of a wide range of validity issues, [...] and this has required the use of an increasing variety of more qualitative research methods [...]. One approach to test validation that has had a dramatic increase in popularity in recent years is the analysis of test discourse, particularly in relation to tests of second language speaking proficiency [7].

Quanto descritto e auspicato da Lazaraton trova la sua applicazione nei TEST CERCLU oggetto del presente studio di validazione in quanto la creazione della griglia di valutazione rispecchia fedelmente la valutazione di quei fattori di coesione e coerenza del discorso così come sono ampiamente descritti negli studi sulla *discourse analysis*.

Le peculiarità del test somministrato a computer vengono inoltre descritte da Inn-Chull Choi et ali [8] che collocano il CBT di L2 all'interno della ricerca scientifica e applicativa del *language testing* sottolineando la necessità di una rivisitazione in termini di validazione di queste nuove tecniche di elicitazione.

Nel caso dei test CERCLU sono stati costruiti *ad hoc*, da un gruppo di tecnici didattici e tecnici laureati, dei formati di costruzione delle domande, che rispondessero alle esigenze del test di lingua, in quanto *Question Mark Perception*, il programma utilizzato non è nato allo scopo di realizzare test

6. A. Lazaraton, *A Qualitative Approach to the Validation of Oral Language Tests*, Cambridge, Cambridge University Press, 2002.
7. *Ibid.*, p. 42.
8. I.C., Choi, Kim K.S., Boo J., *Compatibility of a paper-based language test and a computer-based language test*, in «Language Testing», 20, 2003, pp. 295-306.

di lingua. Esso tuttavia ha permesso di personalizzare il formato di domande, risposte, presentazione di testi di ascolto e di lettura. Ciò ha avuto come risultato la produzione del test CERCLU come un CBT completamente nuovo rispetto a quelli già in commercio.

Il progetto di validazione

Il test di lingua straniera è uno strumento che permette di raccogliere punteggi numerici interpretabili come indice di ciò che il soggetto a cui è stato somministrato il test è in grado di fare con le sue competenze linguistiche. Interpretare i punteggi di una prova di lingua straniera significa pertanto creare un legame tra i risultati ottenuti da un candidato in un compito e la sua competenza o abilità in alcuni ambiti d'uso della lingua. Da ciò deriva che la valutazione di un candidato si basa su prove concrete quali i punteggi e sulla loro corretta interpretazione ed è proprio con la validazione che si attua un processo che permette di fornire prove a sostegno dell'interpretazione corretta dei punteggi. Le procedure di validazione si basano in primo luogo sul *construct* e quindi sugli obiettivi e sui criteri di verifica nonché sulle *specifications* del test, cioè sulla descrizione delle modalità e dei contenuti della verifica stessa. Il *construct* viene creato sulla base di teorie linguistiche sulle abilità di comunicazione considerate e sulla base di un *framework* di riferimento. In secondo luogo si ricorre alle metodologie statistiche più adatte al caso per ottenere un'obiettiva ponderazione dello strumento creato dopo aver affinato e ottimizzato lo strumento stesso attraverso un processo che pesa l'affidabilità del test attraverso il rapporto di co-variazione con i singoli compiti che compongono lo strumento di verifica. Nel caso qui analizzato questo processo ha portato all'eliminazione o alla rielaborazione di alcuni compiti che effettivamente dall'analisi statistica risultavano non essere conformi agli obiettivi del test stesso o semplicemente risultavano troppo difficili per il livello di competenza oppure mostravano ambiguità di interpretazione nella composizione del compito.

Bachman [9] fa presente che il processo di validazione include l'interpretazione del punteggio mediante inferenze sulle abilità che il test deve misurare. Per poter procedere all'interpretazione per inferenze è necessario che esse siano sostenute da prove empiriche e da argomentazioni interpretative. La

9. L.F. Bachman, *Statistical Analysis for Language Assessment*, Cambridge, Cambridge University Press, 2005, pp. 262-4.

prima fase del processo di validazione è atta a determinare la correttezza dell'interpretazione dei punteggi come indice dell'abilità o delle abilità che si vogliono misurare. La seconda fase consiste nel determinare l'uso dei punteggi nella valutazione dell'abilità e la rilevanza dell'abilità che testiamo nell'ambito formativo o professionale. Un esempio tipico di questa fase, descritto da Bachman, è l'uso del punteggio di un test sulla competenza lessicale come test di piazzamento per i corsi di produzione scritta. È evidente che per poter utilizzare questa procedura è necessario provare che la conoscenza del lessico sia effettivamente rilevante per la produzione scritta. Inoltre è necessario provare che il punteggio ottenuto dalla somministrazione di questo particolare test sia rilevante ai fini del piazzamento. Un altro esempio è dato dallo studio di validazione per l'uso di un test di produzione del parlato che era stato progettato per testare il livello di *proficiency* di ESL (*English as a Second Language*) all'interno del TOEFL (*Test of English as a Foreign Language*) nel momento in cui questo stesso test è stato utilizzato per fare uno *screening* di *proficiency* degli insegnanti di ESL [10].

La terza fase prende in considerazione l'utilità del test e quindi l'effetto *washback* nell'iter formativo e la qualità del suo impatto nei confronti della popolazione a cui viene somministrato il test. Queste considerazioni sono rilevanti in quanto un test che non è ben compreso nella sua utilità formativa e che si presenta come strumento valutativo poco coerente avrà un impatto negativo sugli utenti del test stesso. A questo proposito è comprensibile che vi siano corsi studiati appositamente per apprendere le tecniche per affrontare test come il TOEFL e i test del Cambridge Syndicate PET (*Preliminary English Test*) e FCE (*First Certificate English*) che da una parte utilizzano un *test facet* particolare rispetto ad altre prove più tradizionali e dall'altra vengono somministrati in molti paesi diversi per cui hanno bisogno di ridurre l'impatto negativo dovuto all'alta eterogeneità dei candidati.

L'approccio quantitativo

L'approccio quantitativo permette di fornire prove empiriche per avvalorare l'interpretazione logico-razionale dei dati elicitati con un test specifico. La validazione è quindi, secondo questa impostazione, definibile come la co-

10. X. Xi, *Validating TOEFL*® iBT Speaking and Setting Score Requirements for ITA Screening, in «Language Assessment Quarterly», 4, 4, 2007, pp. 18-31.

struzione di un *case study* basata su argomentazioni e prove che avvalorano una corretta interpretazione dei punteggi di un test.

Il progetto di validazione è quindi un processo attraverso cui si scopre e si determina se la prestazione dei candidati, nei test che sono oggetto dell'indagine, è attribuibile alle abilità che questi test vogliono misurare o se vi siano altri fattori, fuorvianti e non appartenenti alle abilità, che influiscono sui punteggi alterando così l'efficacia dello strumento testistico.

Alcuni di questi fattori possono essere fonte di un'interpretazione non corretta dei punteggi, mentre altri fattori, pur non appartenendo all'abilità, non sono rilevanti per l'interpretazione corretta dei punteggi. Nel caso qui esaminato ad esempio sono stati somministrati i test a gruppi di studenti che condividevano tutti una stessa madrelingua e una stessa formazione culturale. Il fatto cioè che gli studenti a cui sono state somministrate le prove dei test CERCLU B1 e B2 avessero tutti un iter scolastico equivalente, oltre ad appartenere alla stessa fascia d'età e a condividere il fatto che fossero studenti universitari, permette di affermare che questi fattori non possono essere fonte di diversificazione nel punteggio ottenuto nel campione esaminato.

L'altro tipico fattore preso in considerazione come possibile elemento di incidenza sul punteggio di un test di lingua straniera è lo stile cognitivo del candidato. Nel caso qui esaminato il campione a cui è stato somministrato il test era caratterizzato da studenti universitari della stessa età e questa caratteristica avvicina molto gli stili cognitivi dei soggetti: nel caso del B1 e B2 di lingua inglese i soggetti erano tutti italofoni mentre nel caso di B1 e B2 di Lingua italiana le provenienze geolinguistiche erano eterogenee.

Un altro fattore è dato dalla familiarità del soggetto con la tipologia dei compiti che gli sono richiesti. Questo fattore è stato ovviato tramite la presentazione al soggetto di diverse tipologie di compiti relativa ad una stessa abilità. Inoltre anche il fattore della varietà lessicale è stato preso in considerazione attraverso l'inserimento di testi che trattano argomenti diversi proponendo schemi concettuali e domini lessicali non omogenei. Infine per garantire che il test così formulato (contenente cioè compiti diversi costruiti con tecniche di elicitazione diverse e basati su tematiche diverse, evitando così di avere una misura della difficoltà e discriminazione data dal metodo del test più che della abilità che si vuole misurare) non fosse affetto da altri fattori fuorvianti, un gruppo di esperti ha costruito e selezionato i compiti più adatti per le singole abilità. A questo punto l'operazione necessaria era quella di sottoporre il test ad una serie di misure statistiche che permettessero di operare un processo di ottimizzazione sull'affidabilità interna ed esterna del test.

La raccolta di prove quantitative

Il progetto di studio per la validazione quantitativa del costrutto teorico impone comunemente un approccio statistico correlazionale che prevede tre possibilità:

1. la somministrazione di diversi test ad un unico campione costituito da un numero elevato di soggetti
2. la progettazione di uno studio sperimentale con gruppi equivalenti che abbiano conseguito un trattamento o un percorso di studio di tipo diverso
3. la somministrazione di diversi test a campioni costituiti da gruppi di livello di abilità differente che dovrebbero produrre prestazioni differenti.

Nel nostro studio si è scelta la procedura di cui al punto 2. Le valutazioni statistiche si sono basate sulle variabili che i *raters*, cioè i valutatori esperti linguistici, dovevano esaminare e alle quali dovevano assegnare un punteggio in relazione alla qualità delle variabili linguistiche prese in esame. Le qualità di ogni variabile veniva descritta in maniera dettagliata e corrispondeva quindi ad uno specifico punteggio. In questo modo le procedure di analisi dei contenuti si concentravano su un numero di qualità linguistiche prestabilite e sulla fascia di livello di ognuna permetteva di raggiungere una valutazione delle prove sistematica e obiettiva. In particolare si è proceduto all'analisi del registro, della punteggiatura e dell'ortografia, del lessico, della grammatica, della coesione, del contenuto e della sua organizzazione. Questi elementi sono aspetti dell'abilità e della competenza tipicamente presenti nella valutazione della produzione in lingua da parte di apprendenti la cui madrelingua è diversa dalla lingua *target*.

La comparazione

Un esempio che riguarda la comparazione di due test tra i più usati al mondo per testare la competenza dell'*English as a Foreign Language* è lo studio effettuato da Bachman [11] sulla comparazione tra TOEFL e FCE, basato sul calcolo della media del punteggio dato da due esperti per ogni *Test Method Facet* (TMF). Sulla falsariga del metodo comparativo di Bachman, per la validazione del CERCLU ogni esperto impiegato nel nostro processo di attuazione ha dovuto valutare le variabili caratterizzanti il metodo usato per elicitare i dati sulle abilità linguistiche quali la lunghezza del testo, il les-

11. L. F. Bachman, *Statistical Analysis for Language Assessment*, cit., pp. 79-85.

sico impiegato, il tipo di informazione presentato nel testo (astratta o concreta, positiva o negativa, reale o irreale), l'argomento, il genere, la grammatica, la coesione, l'organizzazione retorica ed il registro linguistico. Inoltre gli esperti dovevano valutare le abilità di comunicazione linguistica necessarie per ogni compito. Per trovare similarità e differenze sono state calcolate le medie e le deviazioni standard di ogni media dei punteggi per ogni TMC e CLA. Se la differenza tra le medie si mostrava maggiore della deviazione standard, allora la differenza veniva valutata come significativa.

Il problema che il giudizio degli esperti non sia sempre coerente è stato ampiamente dimostrato nelle ricerche sul *Language Testing*. Tuttavia nel caso qui esaminato questa problematica è stata risolta utilizzando un *framework* di riferimento delle categorie di contenuto e nei criteri di *rating* per quanto riguarda la produzione scritta e parlata. Tuttavia le ragioni di una possibile incoerenza possono insorgere anche a causa di altri fattori che non permettono una chiara valutazione della pertinenza del contenuto. Gli elementi che possono incidere sulla debolezza della coerenza sono essenzialmente tre:

1) i campi dell'uso della lingua non sono ben definiti, per cui è impossibile identificare tutti i compiti che rappresentano un particolare campo d'uso

2) le inferenze basate solo sul contenuto di un certo campo d'uso sono limitate ad esso, quindi si possono fare inferenze solo su ciò che il candidato conosce e sa fare in rapporto alle aree coinvolte nel test che è stato somministrato

3) dato che ogni candidato è diverso, diverse sono le prestazioni di fronte a uno stesso compito e a compiti che presentano lo stesso contenuto [12].

Le differenze nelle strategie e nei processi usati dai candidati sono state oggetto di ricerche qualitative e quantitative nell'ambito della linguistica applicata e hanno dimostrato come la tecnica di analisi delle descrizioni verbali dei processi usati dai candidati sia assolutamente utile nel determinare quali strategie vengono usate quando i candidati affrontano il test. In particolare lo studio di validazione qualitativa e quantitativa effettuato da Alderson [13] ha dimostrato che nei compiti più difficili i candidati ricorrono in misura maggiore alle strategie relative al tipo di test a cui sono sottoposti, mentre nei compiti facili prevale l'uso di strategie legate all'abilità che il test vuole misurare.

12. *Ibid.*, pp. 88-97.

13. J.C. Alderson, Clapham C., Wall D., *Language Test Construction and Evaluation*, Cambridge, Cambridge University Press, 1995.

Lo studio correlazionale

Il concetto di correlazione è essenziale per il *testing* ed è alla base anche degli studi di validazione e di affidabilità di misurazione. Bachman [14] sostiene inoltre che le procedure statistiche che sono state sviluppate per studiare questo concetto hanno fornito le basi per nuove strade di ricerca. Essenzialmente si parla di correlazione qualora due misure tendono a muoversi insieme. Essa rappresenta quindi un indice di co-variazione di due variabili. Quando due distribuzioni differenti variano assieme significa che co-variano; maggiore è la covarianza e maggiore è la relazione tra due gruppi di punteggi. Il concetto di relazione prevede che le variazioni in una entità corrispondano alle variazioni nell'altra entità esaminata; esse possono essere due variabili come l'abilità di parlare e di ascoltare ma anche due strutture o due concetti come la motivazione e l'apprendimento delle lingue. In questo caso si tratta di due strutture concettuali diverse, se invece c'è una correlazione tra un test di grammatica e uno di lessico si parla di una relazione tra due variabili.

La correlazione è quindi una relazione tra due entità che, nel caso qui studiato, sono variabili di uno stesso *construct* ma anche *construct* diversi, che variano in termini di forza e direzione. Ciò significa che abbiamo preso in considerazione il grado di correlazione tra le varie entità esaminate e la positività o negatività della covarianza.

Per la comparazione col test CERCLU di livello B2 per la lingua inglese sono stati scelti due altri test di grande diffusione presso i centri linguistici italiani e cioè il *Quick Placement Test* e il *First Certificate of English*. Come Bachman [15] ha dimostrato le differenze tra il FCE ed il TOEFL sono minime e comunque non riguardano l'essenza dei due test come strumento di valutazione di *English language proficiency*. Si può quindi affermare che una buona correlazione con FCE implica anche una affinità con il TOEFL.

Il TOEFL iBT (*computer-based format* usato per l'*internet based testing*) ed il FCE sono entrambi strumenti che verificano, come il CERCLU, le quattro abilità di base anche se alcuni compiti richiedono l'uso di più di una abilità per cui il candidato, nel parlato, dovrà anche ascoltare e nella prova di ascolto dovrà leggere le risposte prima di selezionare quella più adeguata. La versione iBT del TOEFL ha dei descrittori di competenza per ogni *skill* che interpretano e il punteggio comprende le penalizzazioni per le ri-

14. L. F. Bachman, *Statistical Analysis for Language Assessment*, cit., pp. 79-85.
15. *Ibid.*, pp. 95-100.

sposte errate a differenza di ciò che avviene nel CERCLU e nel FCE. Inoltre è necessario precisare che sia TOEFL che FCE sono test che necessitano di un corso sulle tecniche del test e su ciò che esattamente viene richiesto ai candidati soprattutto nella produzione scritta e orale. Il CERCLU si presenta come un test non propriamente legato ad una preparazione su un lessico specifico o su uno stile particolare di produzione, tuttavia i criteri di attribuzione del punteggio nelle prove di produzione sono certamente fissati per valutare il raggiungimento di un livello di *proficiency* adeguato alle esigenze accademiche. Gli altri test sottoposti all'analisi statistica sono stati comparati a test dello stesso livello di comprovata efficacia. Su questo punto è necessario precisare che sia il FCE che il PET, prodotti dalla Cambridge, sono strumenti testistici non computerizzati e che inoltre sono stati ideati prima dell'uscita del *Framework* Europeo per le lingue. Quindi è ovvio che si sarebbero trovate comunque alcune differenze con il CERCLU il quale invece segue in maniera estremamente accurata *the European Framework*. Le differenze reali sono soprattutto date dalle statistiche di tendenza centrale che se da un lato non riportano grandi differenze nel risultato globale del test, sottolineano però la difficoltà maggiore della prova dedicata alla lettura nel CERCLU rispetto alla stessa prova ideata nel sistema Cambridge e viceversa per quanto riguarda la prova di ascolto. Per la validazione del CERCLU di italiano per stranieri si è scelto di utilizzare il CELI (Certificato della conoscenza della Lingua Italiana) che ha un obiettivo un po' diverso dal CERCLU, cioè quello di testare dei professionisti. Le differenze che si notano sono attribuibili alla conformità del CERCLU al *Framework* Europeo.

Lo studio sulle abilità

Le prove di ascolto e di lettura queste sono state somministrate al computer attraverso la costruzione degli *item*, cioè dei singoli compiti in formato elettronico attraverso l'ausilio del sistema autore *Perception* non propriamente congegnato per le esigenze del test di lingua, ma comunque portatore di alcuni vantaggi fondamentali, tra i quali:

1) la sicurezza che il sistema di *open scripting architecture* permette di integrare media differenti come appunto l'audio e lo scritto,

2) la possibilità di creare dei formati di compiti differenti da quelli presenti nel *template* offerto dal pacchetto,

3) la generazione automatica di statistiche utili alla fase di *pretesting*,

4) la possibilità di creare una randomizzazione delle domande e delle risposte in fase di somministrazione,

5) la totale sicurezza del sistema sul server quando viene somministrato attraverso la rete che risulta inviolabile,

6) la rigorosità nell'accesso da parte del candidato di cui il sistema tiene una storia completa.

La fase di *pretesting* ha permesso di ovviare alle ambiguità di taluni *item* a scelta multipla le cui risposte davano adito a interpretazioni diverse. Questo è un tipico effetto collaterale delle domande a scelta multipla che non è assolutamente prevedibile aprioristicamente e quindi era necessario rivelare le ambiguità attraverso una prima somministrazione utile solo allo scopo di studiare statisticamente il funzionamento delle risposte.

Nel nostro caso la valutazione dell'affidabilità è stata attuata secondo la procedura denominata *test-retest reliability*. In primo luogo è stato misurato l'indice di *inter-item consistency*, che viene calcolato automaticamente dal sistema autore *Perception*, per passare poi all'analisi dell'indice alfa, misurato con l'ausilio del programma SPSS e in seguito alla *parallel-form reliability* nella fase di analisi delle correlazioni.

La coerenza tra test e *item* singoli è una misura che è utile soprattutto in fase di pretesting che permette di affinare lo strumento testistico che è esattamente ciò che si ottiene con la misurazione del coefficiente *Alfa*. Quindi queste due misurazioni, *Alfa* e *inter-item consistency*, sostanzialmente hanno la stessa funzione, tuttavia sono state usate entrambe proprio per ottenere una verifica dell'affidabilità al di fuori del programma *Perception*.

L'ANALISI AUTOMATICA DEGLI *ITEM* E L'ANALISI QUALITATIVA DEL DISCORSO SCRITTO E ORALE

Perception fornisce due indici denominati *item-test correlation* e *facility value* e sostituisce il *discrimination index* con la deviazione standard. Inoltre nell'analisi statistica dei singoli *item* tiene sotto controllo la capacità di ciascuno di essi di contribuire in maniera efficace al test nella sua completezza.

La tabella denominata "*Perception Statistics: Inglese B2 Listening*" mostra un esempio delle statistiche relative all'analisi degli *item* per quattro prove di ascolto facenti parte di uno dei test di *proficiency* di livello B2, per l'appunto quello relativo all'abilità di ascolto-comprensione:e.

Perception Statistics: Inglese B2 Listening comprehension			
Question description	Standard deviation	Facility	Correlation
b woman	0,310	0,504	0,112
c drugs	0,026	0,581	0,260
e teacher	0,307	0,511	0,023
g programme	0,348	0,352	0,169
h meeting	0,241	0,598	0,017
b end lesson	0,106	0,678	0.034
c explaining	0,266	0,571	0,023
d dep	0,022	0,615	0,101
e economics	0,003	0,369	0,230
f lectures	0,095	0,681	0,126
g buildings	0,033	0,469	0,199
b miss	0,214	0,622	0,019
c copy lesson	0,018	0,644	0,209
d advise	0,261	0,577	0,029
e friend	0,206	0,627	0,218
f girls	0,344	0,296	0,242
g film	0,169	0,651	0,143
h girls2	0,022	0,620	0,265
b globe theatre	0,343	0,407	0,398
c Henry VIII	0,307	0,510	0,238
d copy	0,140	0,665	0,016
e building	0,156	0,658	0,124
f roof	0,169	0,651	0,178
g rains	0,024	0,596	0,188
h groundings	0,226	0,061	0,288
i authorities	0,224	0,613	0,308
l Globe	0,149	0,661	0,159
m riots	0,112	0,676	0,117

Questa tabella contiene un estratto delle statistiche, rese automaticamente da *Perception*, dei dati ottenuti dalla somministrazione del test in una delle prime fasi di *pretesting*, per 28 *item* facenti parte di quattro *listening-comprehension* (per l'inglese L2) del CERCLU B2 computerizzato: il punteggio minimo raggiunto dal campione a cui è stato somministrato il test, la deviazione standard, il grado di facilità dell'item e la correlazione con punteggio globale del test di cui i compiti fanno parte. Si osserva che

8 su 30 degli item qui riportati hanno un indice di correlazione tra lo 0,2 e 0,3 e quindi sono prossimi alla soglia di accettabilità e 2 item hanno un indice ottimale, cioè superiore a 0,3. In base a questi risultati si sono modificati gli item che mostravano una correlazione inferiore allo 0,2. Tuttavia nessuno dei 30 *item* qui considerati ha una correlazione di valore negativo perciò nessuno di loro doveva essere sostituito o scartato. Infatti è noto un indice di correlazione negativa rende il compito completamente fuorviante rispetto alla valutazione dei candidati e solo in quel caso gli *item* vanno scartati.

L'indice di facilità indica la proporzione dei candidati che rispondono correttamente all'*item* rispetto al totale dei soggetti facenti parte del campione a cui è stato somministrato il test. Di norma l'indice di facilità ottimale è tra 0,3 e 0,7 (mentre un indice tra 0,2 e 0,8 è considerato buono) e nella tabella relativa alla *listening comprehension* CERCLU B2 si nota che 28 *item* dei 30 considerati ricadono in questo intervallo. Ciò significa che la prova non è facile ma nemmeno troppo difficile. La misura di dispersione è data dagli indici della deviazione standard che nei 30 item qui presentati non è alta rimanendo all'interno di un intervallo minimo (da 0,02 a 0,34).

La tabella denominata "*Perception Statistics: Reading Inglese B2*" mostra un esempio delle statistiche relative all'analisi degli *item* per 3 prove di lettura facenti parte di uno dei test di *proficiency* di livello B2, per l'appunto quello relativo all'abilità di lettura-comprensione.

Perception Statistics: Reading Comprehension: Inglese B2										
Question description	Question type	Maximum score	Mean score	St. dev. of score	facility	Corr.	Significance of correlation	Outcome analysis		
cigarette	Pull Down List	5	4.324	0,598	0,600	0,470	1%	Outcome name	Percent. answer	Mean for outcome
								1	99.58%	87.78%
								3	77.73%	90.84%
								2	71.01%	91.01%
								4	87.82%	88.9%
								5	96.22%	88.36%
generation_ europe	Pull Down List	10	8.458	1.48	0,587	0,605	1%	Outcome name	Percent. answer	Mean for outcome
								1	91.6%	88.51%
								3	57.14%	91.1%

								2	90.76%	88.4%
								4	92.02%	88.84%
								5	89.92%	89.44%
								6	91.18%	88.96%
								7	74.37%	91.07%
								8	81.51%	90.39%
								9	97.06%	88.33%
								10	80.25%	90.24%
hotspot	Drag and Drop	5	4.769	0,511	0,662	0,287	1%	Outcome name	Percent. answer	Mean for outcome
								right1	93.7%	88.74%
								right3	96.64%	88.39%
								right4	94.96%	88.38%
								right2	95.8%	88.31%
								right	95.8%	88.4%
								wrong	0.42%	60%

Nella tabella relativa alle *reading comprehension* CERCLU B2 (per inglese L2), che riporta i dati forniti dal programma *Perception* si nota che l'indice di correlazione delle 3 *reading* è superiore allo 0,3 e quasi uguale a 0,3 nella terza *reading comprehension*, risultando così ottimale. Anche il livello di difficoltà è assestato attorno allo 0,6 per le tre *reading comprehension* e infine la deviazione standard è minima, presentando così una alta stabilità nel punteggio di tutto il gruppo. Ciononostante l'analisi delle singole risposte e dei distrattori non risulta chiara poiché dalle tabelle fornite dal programma non risulta quale distrattore sia risultato ambiguo e quale invece mostri un comportamento più coerente. Dal tabulato sembra che la media sia altissima ma ciò non è coerente con il grado di facilità che invece risulta, almeno in queste tre prove di abilità di lettura, perfettamente bilanciato.

Perception non fornisce automaticamente l'indice di discriminazione e quindi i dati ottenuti dei punteggi ottenuti con i test informatizzati sono stati trasferiti sul programma SPSS e si sono così utilizzati i valori statistici delle misure di tendenza centrale da cui dipende il valore discriminativo dell'intera prova e le misure della dispersione.

L'analisi qualitativa è stata attuata per la produzione orale e scritta per la lingua italiana e inglese a livello B2.

Il test di produzione orale si compone di tre parti denominate monologica, monologica macroarea e interattiva ognuna dei quali è stata valutata da un osservatore seguendo la stessa griglia di valutazione analitica. In ag-

giunta a questi dati compare una valutazione sulla componente denominata scioltezza o fluency data dall'esaminatore.

I valori medi delle componenti linguistiche sono pesate secondo lo schema seguente:

pronuncia, accento, intonazione 10%,
accuratezza morfosintattica 20%,
lessico 20%,
contenuti e organizzazione del discorso 20%,
appropriatezza espressiva nella produzione monologica o interattiva 20%
scioltezza 10%

Per lo scritto invece i fattori analizzati comprendono: l'ampiezza dell'uso del lessico, accuratezza grammaticale e appropriatezza lessicale, adeguatezza della punteggiatura, la coesione e coerenza del testo, l'organizzazione dei contenuti, l'appropriatezza del registro linguistico. Le percentuali seguono lo schema dell'orale.

Conclusione

Questo articolo pur riportando solo un esempio delle procedure statistiche effettuate dimostra che il processo di validazione basato su rigorose tecniche qualitative e quantitative è garante della corretta interpretazione dei punteggi ed è necessario nella costruzione di uno strumento equo di valutazione anche nell'ambito delle lingue straniere.

Anna Zanfei