italofono). Pensiamo ancora, in ambito accademico, alle difficoltà di conversione dei voti assegnati a esami sostenuti in altri paesi, all'interno di progetti di mobilità studentesca: qui, sebbene la conversione si basi su criteri di proporzione matematica in apparenza oggettivi, in realtà talora i criteri si rivelano essere fluidi e culturalmente connotati. Un esempio lampante si riscontra nella conversione dei voti tra Italia e Francia, in cui la scala nominale è su base 20, con punto di cesura convenzionalmente fissato a 10. Nella realtà accademica francese il 20 non viene mai utilizzato nella scala di valutazione, al punto che alcune università italiane suggeriscono di convertire il 16 ottenuto nell'esame francese, con la verbalizzazione di un 30 nel piano di studi italiano[6]. Riba [2014] illustra come l'assegnazione di un voto non sia paragonabile alla misurazione di un fenomeno, di un oggetto, di una realtà fisica, per le quali esistono unità di misura chiaramente definite e socialmente accettate[7]. Più di ogni altra cosa, dare un voto significa trasmettere un messaggio, che in quanto tale è complesso, multidimensionale e soggetto a malintesi[8]. Il *giudizio* può essere adeguatamente espresso anche attraverso scale nominali ordinali, come nel caso del ventaglio che va da 'insufficiente' a 'eccellente', oppure da 'principiante' ad 'avanzato', o ancora da decisioni dicotomiche come nel caso di 'ammesso'-'non ammesso'. La variazione della scala di riferimento o l'assenza della scala stessa incide sul tipo di feedback e può veicolare messaggi molto diversi tra loro: spesso gli studenti percepiscono una sanzione, a volte un'accezione punitiva, oppure una ricompensa, un incoraggiamento, un avvertimento [Riba 2015].

Nella prospettiva di Hadji [1993] l'impatto del giudizio varia in base al ruolo di chi lo emette: il *giudizio* dell'*osservatore* è finalizzato a formulare semplici constatazioni di fatto (*Il a mis sa cravate*), il *giudizio* con *funzione prescrittiva* intende orientare il risultato e influire direttamente sull'oggetto valutato (*Tu dois mettre ta cravate*), e infine il *giu-*

[6] Cfr. esempio di tabella di conversione adottata dall'università di Trieste, <https://corsi.units.it/sites/default/files/media/documents/tabella_di_conversione_dei_voti_universitari_esteri-1.pdf> (ultima consultazione, 1 giugno 2020).

[7] Pensiamo al metro o al piede per la lunghezza, al kilo per il peso, al watt per la potenza elettrica, ecc.

[8] « L'apprenant mexicain ne comprend pas la valeur sociale d'un 12/20 attribué par son professeur français, cette note étant considérée comme très mauvaise au Mexique alors qu'elle est tout à fait acceptable en France, ou l'élève déçu n'accepte pas d'être ajourné à un examen qu'il pensait avoir réussi » [Riba 2014: 1].

dizio del valutatore che comporta una presa di posizione, talvolta personale e soggettiva (*Quelle belle cravate !*). Il fragile equilibrio che caratterizza il valutatore e la delicatezza che contraddistingue ogni atto valutativo vengono rappresentati da Hadji con l'immagine di un funambolo che si destreggia tra essere e dover essere, e con quella di un tessitore che stringe nodi tra spirito di osservazione e necessità di prescrizione.

La confusione tra valutazione e misura è diffusa. Doucet [2001], riprendendo Bachman [1990], esplora e chiarisce il rapporto che esiste tra valutazione, test e misura. È infatti possibile valutare senza misurare, come nel caso della valutazione formativa o diagnostica (sfere 1 e 2 in fig. 4), così come è possibile misurare senza valutare, come nel caso delle misure fisiche (sfera 5 in fig. 4), oppure somministrare test senza valutare, come nel caso dei test in ambito di acquisizione linguistica per osservare le evoluzioni dell'interlingua (sfera 4). La prova per la valutazione delle competenze linguistiche si trova all'intersezione delle tre istanze: valutazione, test, misura. Il test è dunque uno strumento che produce una misura con l'obiettivo di esprimere un giudizio o prendere una decisione, cuore dell'attività del valutare (sfera 3 in figura 4).

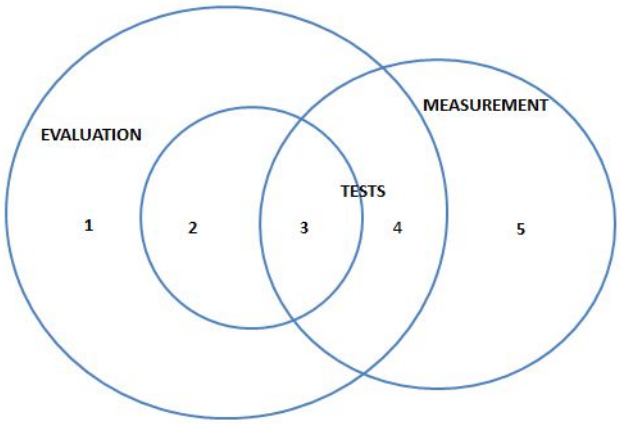

Fig. 4. Rapporto tra valutazione, test e misura (Doucet [2001: 3] da Bachman [1990: 23])

Le prospettive sulla valutazione variano e acquisiscono sfumature diverse in base a chi guida il processo o a chi lo osserva: il docimologo, il sociologo, il didatta. La prospettiva docimologica tende all'oggettività della misura, ma allo stesso tempo è consapevole dei *bias* che influiscono negativamente sulla sua precisione e attendibilità: tra i più noti l'effetto alone, l'effetto pigmalione, l'effetto di contrasto, l'effetto di stereotipia,

l'effetto della distribuzione forzata dei risultati (per approfondimenti De Landsheere [1992]; Demeuse - Strauven [2013: 191-193]). La prospettiva sociologica indaga sui rischi sociali del *teaching to the test* e sugli effetti distorsivi di una eccessiva standardizzazione [Riba 2008: 37]: appiattimento del modello di competenza linguistica sui descrittori, adeguamento cieco alla norma, potere economico e politico degli enti certificatori. Infine, i linguisti e i didatti si sono concentrati sulle nozioni di competenza, performance e conoscenza, concordando sulle tre finalità principali della valutazione: la valutazione predittiva, che consente di fare previsioni, anticipazioni, posizionamento e diagnosi, la valutazione formativa, che dà indizi e informazioni in itinere a studenti e docenti, e infine la sommativa, quella più nota nella rappresentazione tradizionale e che si attua per prendere decisioni alla fine di un percorso di apprendimento o per certificare le competenze di un candidato. Tra detrattori e sostenitori, al centro delle attenzioni passate e recenti dei didatti impegnati nella valutazione, c'è senza dubbio il *Quadro Comune Europeo di Riferimento per le Lingue*, nelle sue due versioni: la prima pubblicata nel 2001 e la seconda a complemento della prima, come si evince dal titolo *Companion Volume* (d'ora in poi CV), pubblicata nel 2016. Se la seconda edizione intendeva colmare le lacune della prima, correggere il tiro sulle critiche ricevute e mettersi al passo con le recenti politiche linguistiche, l'effetto sortito pare essere proprio di altro genere.

3.2. *Spaccare la competenza in quattro? Funzioni e disfunzioni del QCER*

Molto si è detto, nonché accesamente dibattuto in merito al suo ruolo, alle sue funzioni e all'utilità per l'apprendimento, l'insegnamento e la valutazione, come enunciato nel sottotitolo della versione del 2001. Da un lato è innegabile che la prima edizione del QCER si sia configurata come opera di estrema innovazione e importanza nel panorama della didattica delle lingue (*le grand livre rouge de la didactique des langues* per Riba [2008:61]), frutto di un grande sforzo collettivo coordinato dal Consiglio d'Europa. A livello didattico e politico, gli approcci comunicativo-azionali e il plurilinguismo sono i *leitmotiv* al centro di ogni capitolo. L'innovazione non stava tanto nella scientificità dei contenuti, quanto nel loro valore simbolico e politico: armonizzare i livelli e i descrittori di competenza, mettere a disposizione linee guida con un lessico "comune" a tutti i partner europei, a beneficio di una comunità vasta e

plurale di studenti e docenti. Questi ultimi sono rappresentati nel QCER in modo prototipico: lo studente 'cittadino europeo e attore sociale' che in una dimensione complessa di *sapere, saper essere, saper fare e saper apprendere* impara le lingue per interagire con altri cittadini europei, e il docente che, sempre prototipicamente, promuove una dimensione complessa e olistica dell'apprendimento linguistico e della sua valutazione, attraverso la mobilità e il plurilinguismo.

Per molti anni dopo la pubblicazione, l'affermazione 'il test o il manuale è allineato ai descrittori del quadro' ha costituito per la comunità didattica una sorta di garanzia di qualità, di condivisione di un principio comune rassicurante, e anche di modernità. Tuttavia, lo stesso Riba già nel 2008 sagacemente riconosce che

> Le succès du Cadre est tel que ces échelles aux vertus modélisantes revêtent désormais un caractère sacralisant. Plus de méthode, plus de certification, (presque) plus d'article qui n'y fasse directement allusion ou référence, au moins en Europe ; si l'on n'y prend garde, cet ouvrage destiné à la réflexion pourrait bien devenir le grand livre rouge de la didactique des langues, ce qui conduirait inéluctablement à sa proche condamnation [Riba 2008: 61].

Anno dopo anno, il QCER 2001 viene accompagnato dalla pubblicazione di numerosissimi altri progetti (linee guida, manuali, *tool*, ecc.)[9] che hanno l'obiettivo di facilitare l'allineamento dei contenuti didattici e/o valutativi ai descrittori di livello e di accrescere la consapevolezza di insegnanti e redattori di manuali e di test. Tuttavia, la necessità di dare vita a un corollario di progetti parenti presenta numerose ambivalenze. Da un lato, è conferma della sua fama e della sua diffusione, dall'altro ne testimonia la parzialità e la difficoltà di utilizzo nella pratica. Descrittori troppo 'laschi', e quindi inutilizzabili nella valutazione e nella didat-

[9] Solo per citarne alcuni, consideriamo i referenziali per le lingue (*The English Profile, El Plan Curricular*, *Il Profilo della Lingua Italiana, ecc.)* e un numero sempre crescente di progetti per l'allineamento al QCER : CEF-ESTIM GRID (Level estimation Grid for teachers): http://cefestim.ecml.at/, GULT (Guidelines for task-based language Testing): http://gult.ecml.at/GULT/tabid/2311/language/en-GB/Default.aspx, RELEX (Relating Language Examination to the CEF): http://relex.ecml.at/RelEx/tabid/2322/language/en-GB/Default.aspx, WEB-Cef (Collaborative Evaluation of Language Oral Proficiency): http://www.webcef.eu/, ecc. Ultima consultazione, 1 giugno 2020.

tica? Oppure descrittori troppo 'vincolanti' e con possibili effetti coercitivi e nefasti sulle politiche linguistiche? L'estrema parcellizzazione nella descrizione minuziosa di competenza, performance e fonti allontana i descrittori dall'autenticità delle situazioni comunicative e dei contesti socioculturali, per loro natura intrinsecamente ibridi e multidimensionali. Non solo la competenza comunicativa viene analizzata e illustrata nelle sue componenti (morfosintassi, lessico, pragmatica, fonetica, ecc.) e per le diverse abilità ricettive e produttive, livello per livello, ma anche in riferimento a tipologie di scambi comunicativi tra loro molto simili (*information exchange, conversation, informal discussion*). Come nella pratica sia davvero possibile operare una distinzione tra 'scambio di informazioni', 'conversazione' e 'discussione informale' è cosa a noi sconosciuta. Al contempo però sempre gli stessi descrittori sono accusati di vaghezza e ambiguità, soprattutto a livello di scelte terminologiche, difetti che ne minano la validità apparente (*face validity*) e li rendono poco usabili. Il dibattito si fa ancora più acceso quando alcuni studiosi constatano che l'edizione del 2016 non solo non tiene in considerazione le critiche avanzate alla prima edizione (Maurer - Puren [2019:1]), ma addirittura ne esacerba alcuni aspetti. In riferimento a vaghezza e ambiguità dei descrittori, Maurer e Puren [2019: 58-60] criticano l'uso di avverbi come 'abbastanza', 'adeguatamente', 'relativamente', che accrescono la genericità delle affermazioni rendendoli di fatto inapplicabili. Per rimediare a questo difetto, le misure intraprese dal gruppo autoriale[10] del QCER consistono nella proposta di descrizioni minuziosissime delle performance attese, come negli esempi, tratti dalla versione francese.

S'adresser à un auditoire, B1 - Peut faire un exposé *non complexe*, préparé, sur un sujet familier *dans son domaine*, qui soit *assez clair* pour être suivi (p. 77).

Compréhension générale de l'écrit, B1 - Peut lire des textes factuels *clairs* sur des sujets relatifs à son domaine et à ses intérêts avec un niveau *satisfaisant* de compréhension (p. 63).

In questo caso l'indeterminatezza – che nel dibattito filosofico sarebbe più precisamente classificata come 'genericità' – risiede nei ter-

[10] Tra le varie critiche avanzate sia al QCER che al *Companion Volume* vi è quella della poca trasparenza sull'autorialità delle opere, con due obiettivi principali: da un lato, mettere in prima posizione l'istituzione (Consiglio d'Europa), dall'altro incrementare la legittimità dei contenuti attraverso l'associazione delle due opere a un gruppo autoriale europeo e multidisciplinare.

mini ‘non complexe’, ‘clair’, ‘satisfaisant’. Che cosa significa non complesso? Qual è l’idea di complessità rappresentata nel descrittore? A che cosa ci si riferisce con ‘soddisfacente’?

Un’altra critica riguarda la disomogeneità nel passaggio tra i diversi descrittori per livello, difficilmente riconducibili alla realtà delle tappe di acquisizione linguistica.

> Une troisième raison de cette difficulté est le fait que les critères de performance ne sont pas systématiquement suivis d’un niveau à l’autre de la même échelle. On se demande pourquoi en B1 et B2 il s’agit de « lire », alors que dans les autres niveaux inférieurs et supérieurs, il s’agit de « comprendre » [Maurer, Puren 2019: 61].

Alla vaghezza che impedisce un buon uso del CV nella didattica, si affiancano critiche di carattere politico-sociale, forse ancora più pungenti, che in questa sede saranno solo accennate. In sintesi, il QCER e il CV sono accusati di essere strumenti di controllo prescrittivo della politica linguistica europea, al servizio degli enti certificatori che erogano test ad alto impatto per la vita dei candidati. La diffusione capillare di questi strumenti avrebbe portato a una concezione standardizzante e dogmatica della didattica delle lingue basata sul modello del locutore nativo[11]. A questo proposito Maurer e Puren [2019: 10] si chiedono: “Cambridge s’aligne-t-il sur le CECR ou le CECR s’aligne-t-il sur Cambridge ? S’il est permis de se poser la question, c’est que ce sont les bases de descripteurs de l’organisme britannique qui ont servi de point de départ au travail du VC”.

Quanto fino ad ora illustrato ha inteso argomentare la natura inferenziale di ogni processo valutativo, in particolare in riferimento alle tappe

[11] Nella lettera aperta “Il progetto di ampliamento del QCER: una pseudo buona iniziativa del Consiglio d’Europa” pubblicata a maggio 2017 a firma di diverse Associazioni tra le critiche leggiamo: “lo sviluppo schiacciante del QCER e la pressione uniformante dello stesso sulla ricerca, l’insegnamento, le politiche pubbliche” [...], “in questo modo si cristallizza e addirittura si rafforza la dominazione di una concezione standardizzante e dogmatica della didattica delle lingue, una concezione cieca che porta alla distruzione di ogni forma di riflessione dinamica e situata [...], “l’apparente “oggettività” e scientificità del QCER, il discorso di tipo conoscitivo che funge da base, gli conferiscono la sagoma di uno strumento inespugnabile, ciò che lo espone particolarmente al rischio di essere strumentalizzato a vari fini politici”.

http://www.transitlingua.org/assets/projet-d-amplification-du-cadre-europ%C3%A9en-de-r%C3%A9f%C3%A9rence-en-langues.pdf (ultima consultazione 1 giugno 2020)

di *attribuzione di valore,* vale a dire di assegnazione di un voto, di un giudizio e delle decisioni che ne conseguono. Tuttavia, è possibile non subire gli effetti controproducenti di soggettività e vaghezza, ma al contrario tenerli sotto controllo, riconoscerli e monitorarli, al fine di prendere decisioni disciplinate e "utili".

4. *Strategie per ridurre vaghezza e ambiguità nel* testing

Le strategie di riduzione di vaghezza e ambiguità nel LA su cui focalizzeremo l'attenzione nei prossimi paragrafi sono centrali nei processi di qualità per la progettazione e l'utilizzo di test che possano ritenersi utili. L'adozione di una *démarche qualité* [Riba, Mègre 2015] mira, tra le altre cose, a disciplinare le decisioni prese nel LA, nonché a trovare risposte adeguate ai quesiti posti nella prima parte dell'articolo, tra i quali: come delimitare *i* (cfr. par. 2)? Come arginare l'area di vaghezza e di ambiguità nella valutazione? Come ridurre l'errore di misura? E soprattutto, come fare in modo che il vago non sia inutilmente arbitrario?

4.1. *Il* testing *come processo circolare*

Un test è utile [Bachman, Palmer 1996] quando tutte le qualità più importanti – validità, affidabilità, autenticità, interattività, fattibilità – sono monitorate e in equilibrio tra loro. Un test utile si basa su un processo circolare e ibrido di validazione, intesa come raccolta progressiva di dati e informazioni che contribuiscono a creare l'argomentazione della sua validità. Il processo di validazione su cui si fonda un test utile è aperto e iterativo: le tappe di ideazione e progettazione dei contenuti precedono e seguono le tappe di analisi quantitativa, di solito di carattere psicometrico, le une beneficiando degli apporti delle altre. In un test utile l'errore di misura è dunque sotto controllo e il vago non è arbitrario.

> Language assessment is a broad term referring to a systematic procedure for eliciting test and nontest data for the purpose of making inferences or claims about certain language-related characteristics of an individual. […]. The goal behind all L2 assessments is to elicit L2 performance from an individual under certain conditions so that performance consistencies can be interpreted and used to produce records such as scores, verbal descriptions, or mental notes. Interpretations from these records are then used as evidence for making decisions [Purpura 2016: 190].

L'argomentazione di validità viene rappresentata in modo esaustivo nello schema in fig. (5) [ALTE 2011:15], basato sui modelli di Kane, Crooks e Cohen [1999]. Le riflessioni, analisi e verifiche, rappresentate in ognuna delle tappe dello schema, sono collegate, e in qualche modo dipendenti, dalle conclusioni raggiunte nelle tappe successive. Allo stesso modo, l'esito dell'ultima tappa (*real life*) potrebbe ricondurre al principio del ragionamento. Vediamo come.

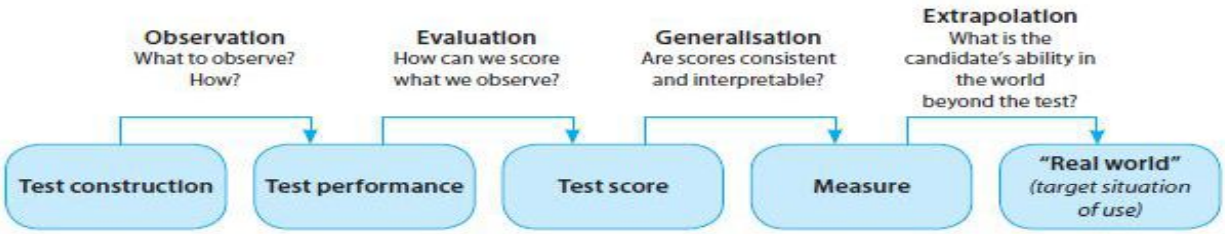

Fig. 5. Argomentazione di validità: rappresentazione grafica delle tappe e dei quesiti associati.

a) *Rapporto tra costrutto e performance* (*test construction* e *test performance*): le produzioni dei candidati devono costituire un campione rappresentativo e interpretabile, coerente con il modello di competenza scelto e veicolato nel test.

> Ogni valutazione linguistica implica una *concezione della lingua*, una concezione pedagogica e, in modo forse più indiretto, una nozione di apprendimento linguistico. Cambiando le teorie linguistiche, quelle pedagogiche e quelle psicolinguistiche (di riferimento) cambiano anche i *tipi di prove*» [Ciliberti 1994: 177].

Se cambiano le teorie linguistiche e pedagogiche di riferimento, cambiano di conseguenza i tipi di item e di task proposti nelle prove. In prospettiva diacronica pensiamo infatti all'impatto che i test fattoriali (*discrete-point test,* [Lado 1961]), rispetto ai test integrati (*integrative test,* [Carroll 1968]), o al *task-based* testing [Norris et al. 1998] degli approcci comunicativo-azionali possono avere sui candidati, rispetto alle operazioni cognitive sollecitate, alle possibili modalità di correzione e feedback. Se cambiano gli obiettivi della valutazione, cambiano (o *dovrebbero cambiare*) i tipi di prove. I test di posizionamento per esempio, noti anche nel panorama italofono come *placement test*, possono produrre risultati validi e affidabili anche solo tramite la somministrazione di un

numero esiguo di item, con indici di discriminazione elevati e con un'ottima correlazione con la competenza globale; molto diverse sono le condizioni per un utilizzo utile ed efficace dei test certificativi, in cui è necessaria una copertura molto ampia dei contenuti e delle abilità per produrre risultati validi e affidabili. Al variare degli obiettivi della valutazione e del tipo di test utilizzato, varieranno (o *dovrebbero variare*) anche trasparenza ed esplicitezza nei criteri, chiarezza ed estensione del feedback, coerenza e rigore nell'utilizzo dei risultati ottenuti (giudizi o punteggi).

b) *Rapporto tra performance e punteggio* (*test performance* e *test score*): In che modo e secondo quali criteri possiamo valutare quanto è stato osservato? la performance orale o scritta dei candidati deve infatti essere associata a un punteggio o a un giudizio. Quali aspetti della performance sono valorizzati o, al contrario, quali sono messi in secondo piano? È questa la tappa in cui si mettono a punto scale e griglie di riferimento, rappresentative del costrutto di competenza del test.

c) *Rapporto tra punteggio e misura* (*test score* e *measure*): questo passaggio riguarda la generalizzazione dei risultati ottenuti nella valutazione di un candidato o di un gruppo di candidati. I risultati prodotti dal test o dalla prova (giudizio, voto, punteggio) sono generalizzabili? Cioè una seconda somministrazione dello stesso test replicherebbe risultati paragonabili alla prima? Oppure un candidato con caratteristiche sociobiografiche e culturali affini otterrebbe un risultato comparabile al primo? Tali quesiti sono indispensabili alla verifica del parametro di affidabilità del test (*reliability*) secondo il quale fattori indipendenti dalla competenza linguistica, come il giorno (più generalmente le coordinate spazio-temporali in cui si somministra la prova), le caratteristiche individuali del candidato (età, biografia linguistica, genere, ecc.), le caratteristiche del valutatore (severità, abitudini, stereotipi, ecc.), la versione del test somministrato, e altri fattori simili, non incidono sul risultato in modo significativo, cioè non determinano errori di valutazione.

d) *Rapporto tra misura e vita reale* (*measure* e *real life*): in questa tappa la finzione che contraddistingue le situazioni di *testing* si confronta con l'autenticità dei comportamenti nella vita reale. Le performance dei candidati nel *testing* si replicano anche nei contesti autentici? Rispecchiano le abilità e i saperi dei candidati nella vita reale? Immaginiamo i

possibili effetti collaterali deleteri nel caso in cui si certifichi una competenza avanzata e specialistica nella lingua target (per esempio linguaggio medico per professioni sanitarie) e questa, al contrario, non corrisponda alle competenze che il candidato riesce ad attivare negli scambi reali lavorativi tra medico e paziente. Un risultato di questo tipo rimetterebbe in seria discussione l'intera argomentazione di validità, riportando il processo all'inizio, cioè alla prima tappa (a).

4.2. *Punti di cesura e* standard setting

La validazione tramite modelli psicometrici applicati a campioni di una certa ampiezza è ormai diffusissima per i test che vogliano uscire da una dimensione locale e artigianale. È altresì obbligatoria per quelli che producono risultati di tipo *high-stake* (cfr. come sancito e dichiarato nei codici etici e nei codici di buone pratiche di tutte le associazioni internazionali per il *testing*). Se la validazione psicometrica è indispensabile alla produzione di misure affidabili, valide ed eque, questa non consente di eliminare integralmente le aree di vaghezza e ambiguità, intrinseche nei processi decisionali del *testing* applicato a fini valutativi. Il modello psicometrico dell'IRT (*Item Response Theory*) permette di ricavare informazioni utili, e a tratti illuminanti, sul comportamento dei candidati e sull'andamento degli item, che sarebbero sfuggite all'osservazione umana di tipo percettivo e impressionistico. Ad esempio, è possibile i) verificare la difficoltà di un item o di una componente del test, in riferimento a una certa popolazione; ii) verificare che un item o un'intera prova possano discriminare correttamente i candidati forti dai candidati deboli; iii) mettere in evidenza il carattere aleatorio e imprevedibile di alcuni distrattori; iv) rilevare errori sistematici[12] dovuti a fattori che nulla hanno a che vedere con il costrutto di competenza linguistica. Per esempio, gli errori sistematici di carattere culturale, indagati attraverso gli indici DIF (*Differential Item Functioning*) dimostrano che alcune tipologie di quesiti sono strettamente influenzate dalle caratteristiche socio-biografiche della popolazione a cui sono proposti. Tuttavia, tutte queste operazioni assolutamente indispensabili e insostituibili, non sembrano dare una risposta univoca ed esaustiva al dilemma trattato nel nostro saggio,

[12] L'errore sistematico si oppone all'errore casuale, non eliminabile nei processi di valutazione e infatti considerato esplicitamente nelle formule psicometriche (cfr. ALTE 1998: 298).

cioè la determinazione del confine tra *i* e non-*i*. Tradotto in termini pratici, i risultati psicometrici non consentono di determinare il punto di cesura (*cut score*) tra un valore e un altro valore. La scelta del termine *valore*, vago e inclusivo, non è casuale: le decisioni da prendere si esprimono infatti con linguaggi diversi: passaggi da un livello a un altro (da B1 a B2, da principiante a falso principiante), cesura tra un 'ammesso' e un 'non-ammesso', o ancora attribuzione di un voto su una scala ordinale. Come noto, non sono le prove o le performance evidentemente eccellenti o evidentemente scarse a comportare difficoltà decisionali; sono piuttosto le gradazioni intermedie. Il punto di cesura, "le plus petit score qu'un candidat doit avoir pour qu'on puisse lui attribuer un niveau ou un classement dans un test ou un examen" (RelEx 2011: 99), è arbitrario in quanto assente in natura e frutto del consenso tra esperti. Più ampiamente, tale consenso deve essere socialmente accettabile e accettato dagli utilizzatori diretti (per esempio studenti e docenti) e indiretti (enti educativi o la società più in generale). Il punto di cesura non è mai un valore assoluto né oggettivamente esatto, come potrebbe essere per una misura fisica, ma sempre associato a una zona di incertezza che, nelle formule, viene identificata come errore di misura. Più il consenso è ampio tra gli esperti e maggiore è il numero di esperti coinvolti, meno rilevante sarà l'errore di misura. Inoltre i punti di cesura non sono universalmente validi, per diverse ragioni: da un lato perché si basano su criteri di riferimento condivisi in uso in una specifica comunità (pensiamo ancora al QCER), dall'altro perché la definizione dei livelli e dei modelli di competenza evolve e si trasforma senza sosta, così come avviene comunemente nella ricerca scientifica e nella definizione dei campi del sapere[13]. L'individuazione dei punti di cesura avviene attraverso un'importantissima tappa di carattere qualitativo, nota come *standard setting*[14], che segue di norma le tappe di validazione psicometrica, tipicamente di carattere quantitativo. I numeri e le statistiche acquisiscono così nello *standard setting* la loro massima utilità ed espressività grazie alla lettura da parte di occhi esperti, disposti a confrontarsi e a discutere delle proprie rappresentazioni e credenze sulla competenza linguistica.

[13] Le considerazioni qui riportate in merito ai punti di cesura e allo standard setting sono il frutto di riflessioni scaturite da formazioni tenute dal dott. Vincent Folny, Responsable expertise pédagogique, Département évaluation et certifications, CIEP.

[14] Per approfondimenti sulle diverse pratiche possibili per lo standard setting e per le procedure di formazioni e documentazione, cfr. Relex 2011, Blais 2008, Cizek 2001 e 2007.

Etablir des standards en éducation est une opération qui vise essentiellement à étalonner des valeurs subjectives (…). A la différence que pour la mise en place de standards de performance en éducation il n'y pas de référent externe objectif et stable, comme un mètre, auquel pourraient se référer ceux et celles qui sont appelés comme juges ou experts pour échelonner les apprentissages et les compétences selon des standards de performance [Blais 2008: 103].

Per Cizek [2001: 5] la definizione dei punti di cesura costituisce il momento cruciale dell'intera validazione di un test: frutto della discussione ragionata da parte di esperti panelisti, nelle sessioni di *standard setting* questi dovranno confrontarsi e mettere a nudo i valori culturali, politici e artistici che guidano e determinano le loro decisioni in ambito valutativo.

La costruzione di un argomento di validità è un processo iterativo e sempre aperto, da cui dipende l'accettazione sociale dei test, dei risultati ottenuti e degli usi che si faranno di tali risultati nella società.

5. *Riflessioni conclusive*

Siamo partiti dalla formulazione di due domande, alle quali possiamo ora rispondere sinteticamente: a che condizioni le componenti di soggettività e di vaghezza non compromettono l'attendibilità dell'attribuzione di un valore? In che modo la soggettività viene considerata in alcuni ambiti significativi del *language testing*?

5.1. *Soggettività e vaghezza nell'attribuzione di valore*

Il quadro teorico di riferimento della *fuzzy logic* ci ha permesso di considerare la nozione di soggettività e quella di vaghezza come profondamente connaturate al LA, ponendo la distinzione tra due casi che si situano, nella realtà, come estremi di un *continuum*: un livello di competenza linguistica può essere infatti definito come *i* (un intervallo) o come *x* (un valore preciso) e più facilmente come un giudizio che si situa su un punto mediano tra *i* e *x*. Più la condizione dell'attribuzione di valore si trova collocata verso il polo *i* più l'attribuzione del giudizio comprenderà di fatto l'insieme di tutti i valori che "non possono facilmente essere falsi", sebbene questa situazione comporti una certa dose di ambiguità. Del resto, un'attribuzione di giudizio efficace, è una valutazione in cui i valori ambigui sono il più possibile ridotti. Abbiamo inoltre sostenuto

che l'arbitrarietà è un'azione talora necessaria nella disambiguazione degli inevitabili casi *borderline*, purché tale atteggiamento corrisponda a una forma di *convenzionalità* e cioè sia esplicitamente accettato da parte di tutti gli attori del processo di valutazione. Infatti, certi valori *x* sono attribuibili o no a *i* solo in relazione alle differenti possibili situazioni nelle quali la valutazione viene prodotta. A queste condizioni l'attendibilità dell'attribuzione di valore non è compromessa dalla soggettività e della vaghezza dei casi *borderline*.

5.2. *La soggettività nel* language testing

La valutazione è dunque un processo di tipo inferenziale finalizzato a produrre decisioni autentiche e utili, fondate su principi di esattezza. Come in ogni processo decisionale, arbitrarietà ed errore di misura possono essere ridotti, limitati ed attenuati ma non eliminati completamente. Al fine di controllare al meglio eventuali errori casuali, e non produrre errori sistematici, è necessario che gli attori della valutazione abbiano piena consapevolezza delle condizioni che stanno alla base di un test utile, valido e affidabile, e pertanto possano attuare le giuste strategie di verifica e controllo della qualità, lungo tutta la filiera del *testing*, dalla progettazione dei contenuti al risultato finale. La valutazione e gli effetti che questa produce sulle istituzioni, sui candidati, e sulla società in generale, sono fortemente ancorati a uno specifico contesto sociale, culturale e comunicativo, così come lo sono i descrittori di riferimento adottati per progettare i contenuti e per valutare le *performance* degli studenti. Se soggettività e arbitrarietà non sono quindi eliminabili, è possibile però ridurne gli effetti collaterali che incidono sull'attendibilità dei risultati, attraverso l'applicazione delle buone pratiche di LA, in cui la dimensione psicometrica viene ridiscussa e interpretata da valutatori esperti nell'ambito delle sessioni di *standard setting*.

RIFERIMENTI BIBLIOGRAFICI

ALTE [2011], *Manual for Language Test Development and Examining*, Council of Europe, https://www.alte.org/resources/Documents/ManualLanguageTest-Alte2011_EN.pdf (ultima consultazione 3 giugno 2020).

ALTE [1998], *Multilingual Glossary of Language Testing Terms*, Cambridge, University Press.

Austin, J.L. [1962], *How to do things with words*, Oxford, University Press.

Bachman, L.F. - Palmer, A.S. [2010], *Language assessment in practice*, Oxford, University Press.

— [2002], *Some reflections on task-based language performance assessment*, «Language Testing» 19(4), pp. 453-476.

— [1990], *Fundamental considerations in language testing,* Oxford, University Press.

Barni, M. [2000], La verifica e la valutazione, in A. De Marco (a cura di), *Manuale di glottodidattica*, Roma, Carocci.

Barkaoui, K. [2010], *Variability in ESL essay rating processes: the role of the rating scale and rater experience*, «Language Assessment Quarterly» 7, no. 1, pp. 54–74.

Bezzi, C. [2014], *Il nuovo disegno della ricerca valutativa*, Milano, Franco Angeli.

Blais, J-G. [2008], Les standards de performance en éducation, *Mesure et évaluation en éducation* 31, 2, pp. 93-105.

Burns, L. [1991], *Vagueness: An Investigation into Natural Languages and the Sorites Paradox*, Dordrecht, Kluwer Academic Publishers.

Carroll, J.B. [1968], *The psychology of language testing*, in A. Davies (ed.), *Language Testing symposium: A psycholinguistic approach*, Oxford, University Press, pp. 46-69.

Centre Européen pour les Langues Vivantes [2011], *Relier les examens de langues au Cadre européen commun de référence pour les langues: apprendre, enseigner, évaluer (CECR)*, Editions du Conseil de l'Europe, Strasbourg.

Cervini, C. - Martari, Y. [in preparazione], *Chi ha paura della soggettività? Studio di caso su complessità e sfocatezza nella valutazione della performance orale in italiano L2.*

Chapelle, C.A. [1998], *Construct definition and validity inquiry in SLA research*, in Bachman, L.F. - Cohen, A.D. (Eds.), *Interfaces between second language acquisition and language testing research* (pp. 32–70), Cambridge, University Press.

Ciliberti, A. [1994], *Manuale di glottodidattica: per una cultura dell'insegnamento linguistico*, Firenze, La Nuova Italia.

Cizek, G.J. (dir.) [2001], *Setting performance standards: Concepts, methods and perspectives*, Mahwah, NJ, Lawrence Erlbaum.

Cizek, G.J. - Bunch, M.B., [2007], *Standard setting: a guide to establishing and evaluating performance standard on tests*, Thousand Oaks, CA, Sage.

Consiglio d'Europa [2001], *Il Quadro Comune europeo di Riferimento per le Lingue: insegnamento, apprendimento, valutazione*, Firenze-Oxford, La Nuova Italia.

Consiglio d'Europa [2018], CEFR – *Companion Volume with new descriptors*, <http://rm.coe.int/cecr-volume-complementaire-avec-de-nouveaux-descripteurs/16807875d5 (versione francese)

De Landsheere, G. [1992], *Dictionnaire de l'évaluation et de la recherche en éducation*, Paris, PUF.

Demeuse, M. - Strauven, C. [2013], *Développer un curriculum d'enseignement ou de formation: Des options politiques au pilotage*, De Boeck Supérieur Louvain-la-Neuve, Belgique.

Deygers, B. - Van Gorp, K. - Demeester, T. [2018], The B2 Level and the Dream of a Common Standard, «Language Assessment Quarterly» 15, 1, pp. 44-58.

Dietz, R. - Moruzzi, S. (ed.), [2009], *Cuts and Clouds: Essays on the Nature and Logic of Vagueness*, Oxford, University Press.

Domenici, G. [2002], *Manuale della valutazione scolastica*, Roma-Bari, Laterza.

Doucet, P. [2001], Pour un test utile, «ASp» 34, pp. 13-33.

Gan, Z. - Leung, C. - He, J. - Nang, H. [2018], Classroom Assessment Practices and Learning Motivation: A Case Study of Chinese EFL Students, «TESOL Quarterly» 53(2), pp. 514-529.

Hadji, C. [1993], *L'évaluation, règles du jeu. Des intentions aux outils*, Parigi, Edition ESF.

ILTA [2018], *ILTA Bibliography of Language Testing*. <https://www.iltaonline.com/> <https://cdn.ymaws.com/www.iltaonline.com/resource/resmgr/docs/bibliographies/lang_testing_bilbio_2018cat.pdf>

Kane, M. - Crooks, T. - Cohen, A. [1999], *Validating measures of performance*, «Educational Measurement: Issues and Practice» 18 (2), pp. 5-17.

Kolen, M.J. - Brennan, R.L. [2014], *Test Equating, Scaling, and Linking: Methods and Practices*, New York, Springer.

Kosko, B. [1993], *Fuzzy Thinking: The New Science of Fuzzy Logic*, New York, Hyperion.

Lado, R. [1961], *Language testing: The construction and use of foreign language tests*, New York, McGraw-Hill Book Company.

Levi, T. - Inbar-Lourie, O. [2020], Assessment Literacy or Language Assessment Literacy: Learning from the Teachers, «Language Assessment Quarterly» 17,2, pp. 168-182.

Li, S. - Taguchi, N. - Xiao, F. [2019], Variations in Rating Scale Functioning in Assessing Speech Act Production in L2 Chinese, «Language Assessment Quarterly» 16,3, pp. 271-293.

Machetti, S. [2006], *Uscire dal vago. Analisi linguistica della vaghezza nel linguaggio*, Roma-Bari, Laterza.

Magni, E. [2020], *L'ambiguità delle lingue*, Roma, Carocci.

Maurer, B. - Puren, C. [2019], *CECR: par ici la sortie!*, France, Editions des archives contemporaines.

McNamara, T. [1998], *Policy and Social Considerations in Language Assessment*, «Annual Review of Applied Linguistics» 18, pp. 304-319.

— [2011], *Managing learning: and language assessment*, «Language Teaching» 44(4), pp. 500-515.

— [2019], *Language and Subjectivity*, Cambridge, University Press.

Menken, K. [2017], High-Stakes Tests as De Facto Language Education Policies. In Shohamy et al. [2017], pp. 385-396.

Moruzzi, S. [2012], *Vaghezza. Confini, cumuli e paradossi*, Roma-Bari, Laterza.
Norris, J. - Brown, J. D. - Hudson, T. -Yoshioka, J. [1998], *Designing second language performance assessments*, in *SLTCC Technical Report 18. Honolulu: Second Language Teaching & Curriculum Center*, University of Hawai'i at Manoa.
Nouwen, R. - van Rooij, R. - Sauerland, U. - Schmitz, H-C., [2009], *International Workshop on Vagueness in Communication (ViC; held as part of ESSLLI)*, LNAI, vol. 6517, New York, Springer.
Purpura, J.E. [2016], *Second and Foreign Language Assessment*, «The modern language Journal. Supplement», pp. 190-208.
Qarayeva, G. Q. [2018], *On The Investıgatıon Of Speech Acts. Advances*, «Social Sciences Research Journal» 5(2). <https://doi.org/10.14738/assrj.52.4088>.
Riba, P. [2008], Quelques réflexions sur l'évaluation en langues : comment faire mieux ?, in *Chemins actuels, revue de l'association des professeurs de français du Mexique*, AMIFRAM.
— [2014], *De la modélisation didactique à l'évaluation en contexte : un processus dynamique*, «Synergie Mexique» 5, Universidad nacional autonoma de Mexico, Mexico DF.
— [2015], *Convertir l'évaluation en un processus de dialogue*, «Français dans le monde» 400, juillet-août 2015.
Riba, P. - Mègre, B. [2015], *Démarche qualité et évaluation en langues*, Paris, Hachette.
Sbisà, M. (ed,) [1995], *Gli atti linguistici. Aspetti e problemi di filosofia del linguaggio*, Feltrinelli, Milano.
Seden, K. - Svaricek, R. [2018], *Teacher subjectivity regarding assessment: exploring English as a foreign language teachers'conceptions of assessment theories that influence student learning* «CEPS Journal» 8, 3, pp. 119-139.
Shohamy, E. - Or I.G. - May, S. (ed) [2017], *Language Testing and Assessment*, New York, Springer.
Shohamy, E. [2017], *Critical language testing*, in Shohamy et al. [2017], pp. 441-453.
Williamson, T. [1994], *Vagueness*, London - NewYork, Routledge.
— [2007], *Knowledge within the Margin for Error* «Mind» 116, pp. 723-26.
Zadeh, L.A. [1965], *Fuzzy Sets*, «Information and Control» 8, pp. 338-353.

Vorremmo ringraziare Sebastiano Moruzzi, che ha letto una prima versione dell'articolo aiutandoci con alcune preziose osservazioni. Ogni imprecisione o lacuna resta naturalmente a carico nostro.