Effetti intenzionali di washback nelle prove di valutazione

Odelli, E.; Diadori, P.; Machetti, S.

EFFETTI INTENZIONALI DI *WASHBACK* NELLE PROVE DI VALUTAZIONE

di Enrico Odelli*

1. Introduzione

Gli studi sugli effetti dei test hanno rappresentato negli ultimi trenta anni un terreno particolarmente fertile, attualmente gli esperti sono concordi nel considerare l'insieme delle ricadute delle prove di valutazione come una delle caratteristiche essenziali dei test, in quanto tendono a modificare la natura dell'insegnamento e dell'apprendimento. Nelle ricerche sulla valutazione linguistica Bachman e Palmer (1996) accolgono il concetto di 'impatto' fra le utilità e qualità essenziali di un test. Pur collocandosi nell'ampio settore di questi studi, il presente contributo ha come scopo specifico di indagare gli effetti intenzionali dei test, fenomeno trattato solo più recentemente nella letteratura come *intended washback* (Andrews, 2004). Se l'insegnante è anche costruttore di test e/o valutatore, le prove possono costituire uno strumento particolarmente efficace di condizionamento positivo del comportamento degli apprendenti. In questo studio di caso esponiamo l'esperienza avvenuta attraverso l'uso di test digitali di valutazioni in itinere dell'apprendimento dell'italiano come lingua straniera all'Università di Leida.

1.1 *Origine e definizione del termine*

Nella letteratura sull'argomento ritroviamo svariati termini con cui si designa il medesimo fenomeno, la denominazione più frequente, quella che si è cristallizzata nel tempo e attualmente risulta più comunemente accettata, è senza dubbio *washback* (da qui *wb*). Nella fase iniziale degli studi di questo fenomeno, che pos-

* Università di Leida.

siamo far risalire alla seconda metà degli anni Ottanta, il termine viene utilizzato in concorrenza con *backwash* (Hughes, 1989) e allude alla metafora proveniente dalla nautica del flusso che si crea per l'effetto rotatorio dell'elica del motore di un'imbarcazione. Recentemente l'uso di *backwash* denota invece sempre l'effetto negativo del *wb*. Nella ricerche, dipendentemente dalla loro impostazione, riscontriamo un ventaglio terminologico piuttosto ricco: 'validità consequenziale' (Messick, 1989, 1996), 'validità sistemica' (Frederiksen e Collins, 1989), 'istruzione basata sulla misurazione' (Popham, 1987) o 'allineamento curriculare' (Madaus, 1988; Smith, 1991). Per quanto riguarda il termine 'impatto' (Bachman e Palmer, 1996), la sua accezione è in stretto legame alla rappresentazione degli effetti di livello macro. Secondo questo significato 'impatto' indica pertanto una categoria più generica rispetto al *wb*.

Nelle varie definizioni ravvisiamo un denominatore comune che riguarda gli attori principali su cui si producono effetti di *wb*: gli insegnanti e gli apprendenti. Le definizioni di *wb* oscillano però da formulazioni generiche a formulazioni più mirate. Secondo Bailey (1996) il *wb* rappresenta globalmente l›influenza dei test sull'insegnamento e l'apprendimento. Messick (1996) pone l'accento sulla misura in cui il test influenza gli insegnanti di lingue e gli studenti a "fare cose" che altrimenti non avrebbero necessariamente fatto. Buck (1988), che ha condotto un'indagine sui test di ingresso nelle università giapponesi, propone una spiegazione del fenomeno più complessa, afferma infatti che c'è una tendenza naturale sia per gli insegnanti che per gli studenti ad adattare le loro attività in classe alle esigenze del test, specialmente quando il test è molto importante per il futuro degli studenti e le percentuali di promossi sono utilizzate come misurazione del successo dell'insegnante. Oltre a sottolineare la rilevanza delle ricadute dei test sulla classe, Buck è il primo studioso di *wb* ad individuare la dicotomia fra 'effetto benefico' ed 'effetto dannoso' dei test. Il suo contributo non si limita tuttavia ad evidenziare il concetto di direzionalità. Dalla definizione possiamo infatti estrapolare un secondo principio fondante della teoria del *wb*, il 'grado di intensità' delle ricadute: quelle prove di grande importanza per gli individui e per gli effetti rilevabili anche a livello sociale vengono identificate come *high stakes*, vale a dire come test ad elevata posta in gioco o test particolarmente a rischio per i candidati. A Buck il merito di aver individuato questi due pilastri della teoria del *wb*; gli studi successivi hanno approfondito l'esplorazione delle modalità attraverso le quali si hanno ricadute. Fra le domande ricorrenti ci si chiede se le tracce prodotte nell'ambito di un unico test siano rigidamente separabili in *wb* negativo e *wb* positivo; questa conclusione si riscontra in molti studi, Alderson e Wall (1993) invece ritengono che il medesimo test possa presentare effetti benefici e dannosi contemporaneamente. Una riflessione critica sull'evoluzione degli studi sul *wb* proviene da Cheng e Curtis (2004), che sostengono che la ricerca potrebbe trarre notevole profitto se fosse rivolta allo studio delle cause del fenomeno in relazione all'insegnamento e all'apprendimento, piuttosto che all'analisi della classificazione positiva o negativa del *wb*. Anche

relativamente al grado di intensità è opportuna una riflessione critica: benché le ricadute siano più significative e più facilmente percepibili negli insegnanti e negli apprendenti per i test ad alta posta in gioco, sorge la domanda se, al contrario, i test con una posta in gioco limitata non producano per nulla effetti. Intuitivamente presupponiamo che ogni test sia destinato a influenzare il comportamento degli *stakeholder* e che l'osservazione degli effetti riguardi un fenomeno assai complesso, in quanto il grado di intensità è in stretta interazione non solo con le finalità generali del test, ma anche con obiettivi, atteggiamenti e percezioni dipendenti dalle peculiarità dei ruoli degli *stakeholder*.

2. Quadro teorico

Prima di affrontare lo stato dell'arte, si rende necessaria una premessa riguardo a come si manifesta dal punto di vista temporale il *wb*: se si escludono le forme alternative e innovative di valutazione o autovalutazione, ad esempio i test in itinere o il *Portfolio*, la prova di valutazione avviene sempre a conclusione del corso. Dobbiamo però ipotizzare anche altri contesti, considerando che attualmente l'apprendente ha quasi sempre a disposizione un esempio della prova di verifica. Anzi, la tendenza attuale è proprio quella di garantire la massima trasparenza nella valutazione linguistica fornendo esempi, istruzioni sui contenuti delle prove, requisiti minimi per il superamento, criteri di valutazione, ecc. Sarebbero perciò principalmente questi elementi e non tanto la prova di valutazione vera e propria, a produrre effetti positivi e negativi durante il percorso di apprendimento. La dimensione temporale è stata descritta nei dettagli prima nello studio di Hughes (1989) e poi integrata con ulteriori particolari da Pan (2009). Dagli studi si evince che il comportamento dell'apprendente e del docente sono fortemente influenzati dal desiderio di focalizzarsi solo sulla parte del percorso di apprendimento che verrà successivamente somministrata in sede di esame, trascurando di conseguenza le mete didattiche, linguistiche e formative, che non fanno parte direttamente del test (Chan, 2020). Oltre a manifestare effetti anteriormente e posteriormente all'esame il *wb* può avere ricadute al di là della durata del singolo corso: l'insegnante può sottoporre a revisione il curricolo, le attività della classe e la prova di esame, in vista di un adattamento per i corsi futuri. In questo caso gli effetti del *wb* coinvolgono periodi lunghi e più generazioni di studenti. L'intervento dell'insegnante si realizza all'interno del monitoraggio della relazione fra la prova e le attività di apprendimento. A questo proposito Chan (2018) ha proposto un modello ciclico che mostra l'intervento dell'insegnante che adegua la propria didattica in base alle ricadute del test. Agli apprendenti verrà impartito un insegnamento migliore rispetto a quello della generazione precedente di apprendenti e di volta in volta il ciclo si ripeterà.

Il modello di Chan (2018) non rappresenta l'unico tentativo di catturare sche-

maticamente la complessità del *wb*, negli studi sull'argomento si ritrovano diversi modelli. Alderson e Wall (1993) propongono, ad esempio, 15 ipotesi di ricadute:

1) un test influenzerà l'insegnamento;
2) un test influenzerà l'apprendimento;
3) un test influenzerà ciò che gli insegnanti insegnano;
4) un test influenzerà il modo in cui gli insegnanti insegnano;
5) un test influenzerà ciò che gli studenti imparano;
6) un test influenzerà il modo in cui gli studenti apprendono;
7) un test influenzerà la velocità e la sequenza dell'insegnamento;
8) un test influenzerà la velocità e la sequenza di apprendimento;
9) un test influenzerà il livello e la profondità dell'insegnamento;
10) un test influenzerà il livello e la profondità dell'apprendimento;
11) un test influenzerà l'atteggiamento nei confronti del contenuto, del metodo, ecc. di insegnamento/apprendimento;
12) i test che hanno conseguenze importanti produrranno *wb*;
13) i test che non hanno conseguenze importanti non produrranno alcun *wb*;
14) i test avranno effetti di *wb* su tutti gli studenti e gli insegnanti;
15) i test avranno effetti di *wb* per alcuni insegnanti e alcuni studenti, ma non per gli altri.

I modelli successivi di Bailey (1996), Pan (2009) e altri studiosi, che per motivi di spazio non possiamo esaminare, mostrano la tendenza a integrare gli elementi coinvolti nel *wb* attraverso una rappresentazione complessa.

Gli studi sul *wb* degli ultimi trent'anni si sono dimostrati di grande utilità, in quanto sono finalizzati a chiarire la sistematicità delle ricadute, a cogliere le modalità con cui gli effetti si manifestano, solo parzialmente ad indagare le cause delle ricadute. Sebbene la difficoltà di esplorare la sua regolarità sia notevole, senza queste ricerche empiriche non saremmo in grado di spiegare poco o nulla di questo fenomeno. Notiamo che gli studi degli ultimi 15 anni sono spesso realizzati in contesti con esami con una posta in gioco alta, molti riguardano la certificazione dell'inglese *TOEFL* (*Test of English as A Foreign Language*) o *IELTS* (*International English Language Testing System*) o certificazioni simili. Più limitati di numero sono gli studi empirici che considerano il *wb* uno strumento a disposizione dei costruttori dei test, degli insegnanti, dei valutatori, della politica per ottenere effetti positivi nell'apprendimento. In questo settore specifico di studi Andrews (2004) distingue due tipi di *wb*: l'*intended* (il *wb* intenzionale) e l'*unintended* (il *wb* non previsto). Sempre improntata su questa diversificazione abbiamo alcuni studi fra cui quello di Qi (2007), che ha analizzato il *wb* intenzionale della produzione scritta del *National Matriculation English Test* (*NMET*) in Cina. Oltre ad aver coinvolto un numero elevato di *stakeholder* (costruttori dei test, insegnanti e studenti), segue un metodo di ricerca basato sulla raccolta di indicazioni provenienti da opinio-

ni, osservazioni in classe e questionari, nonché prestando attenzione tanto al *wb* intenzionale quanto al *wb* non previsto. L'obiettivo del suo studio è indagare se l'innovazione curricolare della produzione scritta possa essere mediata attraverso la modifica dei test. I *task* delle prove scritte erano stati modellati su contesti comunicativi per elicitare una ricaduta sul modo di insegnare e prepararsi alla prova scritta. Qi (2007) arriva alla conclusione che in quel contesto specifico il *wb* intenzionale non ha prodotto gli effetti desiderati: gli insegnanti e gli studenti, nonostante il rinnovamento delle prove di scrittura, per varie ragioni non si sono adattati ai nuovi requisiti della prova.

Il *wb* intenzionale è stato studiato soprattutto per i test ad alta posta in gioco, in contesti in cui si volevano avviare innovazioni curricolari, tanto nell'ambito generale dell'istruzione, tanto in quello specifico dell'educazione linguistica. Potremmo descrivere in modo conciso il *wb* intenzionale come l'intervento mirato nelle prove di valutazione volto ad ottenere ricadute positive di vario tipo. Questa pratica manipolatoria sembra inaugurare un'area di studi in grado di fornire indizi sul funzionamento del *wb* molto più rilevanti rispetto agli studi precedenti, che erano spesso caratterizzati esclusivamente da un'analisi sterile, condotta a posteriori del *wb* non previsto.

3. Contesto dello studio

Fra gli obiettivi dell'apprendimento della lingua italiana nel corso di laurea in Lingua e cultura italiana all'Università di Leida, accanto a quello fondamentale linguistico-comunicativo, è inclusa anche la riflessione metalinguistica. Questa competenza, oltre ad interfacciarsi con la dimensione dell'adeguatezza grammaticale nelle abilità di comunicazione, si rende necessaria come risorsa spendibile nelle altre materie del curriculum, ad esempio nell'analisi dei testi letterari e nell'approfondimento dello studio della linguistica. Sarà utile aggiungere che un numero di laureati del dipartimento si troverà ad insegnare l'italiano come lingua straniera. Per questo gruppo la conoscenza approfondita della grammatica arricchisce il repertorio di strumenti a disposizione nella propria professione. Nel percorso della laurea triennale, relativamente alla suddivisione delle attività di apprendimento in presenza rispetto a quelle in autonomia, siamo dell'opinione che l'apprendimento della grammatica possa avvenire in gran parte come *task* da svolgere a casa attraverso l'utilizzo di piattaforme digitali. Il vantaggio di questa scelta si coglie per quelle attività di acquisizione della lingua per le quali la comunicazione in presenza risulta essere *conditio sine qua non* e per le quali le piattaforme digitali generano forme di *feedback* poco idonee. Per l'apprendimento della grammatica gli studenti del primo anno di *Bachelor* hanno a disposizione *Compito* (acronimo in neerlandese di *Computer Italiaans Onderwijs*, ossia 'Insegnamento dell'italiano al computer') una grammatica interamente online con circa 500 pagine di teoria collegate tra loro secondo una configurazione a grafo

per consentire la consultazione da ogni luogo dell'ipertesto, con 1800 esercizi e 2000 testi di *feedback*. Nel primo periodo in cui *Compito* veniva utilizzato in autonomia, gli studenti tendevano a posticipare lo studio della grammatica a pochi giorni prima dell'esame. In generale possiamo notare come questo tipo di comportamento sia una conseguenza dei sistemi di insegnamento con esami programmati a conclusione del corso. Per l'apprendimento di ogni L2 si ottengono risultati migliori se l'apprendente investe nello studio quotidianamente, in questo modo si crea una sinergia fra l'acquisizione nelle ore in presenza e l'approfondimento della grammatica in autonomia. La realtà mostra che solo un numero limitato di studenti che ha già acquisito le abilità di pianificazione, tende a diluire le ore di studio per tutta la durata del corso. Abbiamo ipotizzato che, sostituendo l'esame finale con test digitali in itinere distribuiti nei singoli moduli, lo studente fosse indotto a iniziare l'apprendimento in autonomia già nelle prime settimane del modulo. Come effetto secondario abbiamo supposto una ricaduta positiva sul rendimento, vale a dire sul rapporto fra i crediti formativi acquisiti e quelli previsti dal programma. Ritardando l'inizio della preparazione di un esame, fra gli effetti negativi più evidenti ritroviamo il mancato superamento delle prove d'esame, l'impossibilità di concludere lo studio entro tre anni, l'eventuale interruzione dello studio e il rendimento negativo del dipartimento. Le difficoltà degli studenti del primo anno di pianificare a breve e a lungo termine il proprio studio sono arcinote e quantunque siano stati adottati diversi espedienti per arginare questa problematica con corsi *ad hoc* o l'uso del *Portfolio*, possiamo affermare che queste iniziative si sono dimostrate assai sterili. Saper pianificare, come abilità specifica rispetto alla più generale capacità di studio, è una tipica *soft skill*. Nella pratica corrisponde all'individuazione dei compiti di preparazione di un esame e alla capacità di rapportarli a una scansione temporale realistica, quindi saper suddividere in sub compiti, saper finalizzare attività per evitare che entrino in conflitto con altre attività, tanto di studio quanto della sfera privata.

4. Metodologia

Come primo passo è stato realizzato un *database* con prove di valutazione sommativa suddivise per grado di difficoltà. Le prove comprendono le abilità ricettive (comprensione orale e scritta) e due aree relative alla conoscenza (lessico e grammatica). Sono stati creati test digitali per tre dei quattro moduli di lingua del primo anno di *Bachelor*, per ogni prova sono disponibili tre versioni diverse, ma calibrate per grado di difficoltà allo stesso modo. L'assetto di questo intervento presenta caratteristiche distinte rispetto alla maggioranza degli studi precedenti sul *wb* intenzionale, dove le modifiche ai test sono quasi sempre di tipo contenutistico. Per questo studio di caso l'intervento è costituito da una scomposizione dell'esame finale in due o tre test di un'ora e mezza ciascuno e dalla loro ridistribuzione temporale. Ad ogni prova è assegnato un valore percentuale e il voto corrisponde alla

media ponderata, considerando il peso di ciascuna prova. Riportiamo per comodità lo schema completo dei test dei primi tre moduli del primo anno di *Bachelor* del dipartimento di Lingua e cultura italiana all'Università di Leida:

Competenza linguistico-comunicativa Ia (semestre 1, durata 6 settimane): 3 test digitali
Test 1 alla terza settimana di lezione: peso 20%
Test 2 alla quinta settimana di lezione: peso 30%
Test 3 alla sesta settimana di lezione: peso 50%
(Esame digitale sessione di recupero alla settima settimana: peso 100%)

Competenza linguistico-comunicativa Ib (semestre 1, durata 7 settimane): 2 test digitali
Test 1 alla quarta settimana di lezione: peso 20%
Test 2 alla settima settimana di lezione: peso 40%
Progetto 1 (*Take Home* di produzione scritta): peso 10%
Produzione e interazione orale: peso 30%
(Esame digitale sessione di recupero all'ottava settimana: peso 60%)

Competenza linguistico-comunicativa Ic (semestre 2, durata 6 settimane): 3 test digitali
Test 1 alla terza settimana di lezione: peso 20%
Test 2 alla quinta settimana di lezione: peso 30%
Test 3 alla sesta settimana di lezione: peso 50%
(Esame digitale sessione di recupero alla settima settimana: peso 100%)

Rispetto agli studi nell'ambito della certificazione con un numero elevato di *stakeholder* e ad elevata posta in gioco, questo studio di caso coinvolge un numero di partecipanti inferiore, inoltre la posta in gioco è notevolmente più ridotta. Un altro elemento distintivo riguarda i ruoli di costruttore delle prove, di insegnante e di valutatore: per questo studio vengono tutti ricoperti da un'unica persona, nelle certificazioni queste funzioni sono sempre distinte.

Lo studio di caso considera un arco di tempo della durata di 16 anni, i test sono stati sottoposti ad aggiustamenti ciclici secondo la linea proposta da Chan (2018). Poiché nel corso di questi 16 anni sono intervenuti numerosi cambiamenti nei test, nella tecnologia di supporto e nella dimensione sociale, non opteremo per una ricerca con un'elaborazione statistica dei dati. Lo studio quantitativo richiederebbe un esame dettagliato dell'esito di tutti i test in relazione a ogni modifica, non aggiungerebbe tuttavia elementi decisivi per l'interpretazione delle ricadute. Per l'analisi dei dati e la loro spiegazione ci affideremo invece a una presentazione qualitativa, organizzata in base alle ricadute più evidenti di tipo positivo e negativo, ne discuteremo inoltre le cause. In coerenza con la tradizione degli studi sul *wb* abbiamo raccolto dati in forma mista, provenienti da interviste individuali a

studenti, osservazioni del docente e questionari. Le domande dei questionari sono state predisposte da *Ecole*, la sezione che si occupa del sostegno informatico e dalla commissione paritetica docenti-studenti del dipartimento di Lingua e cultura italiana all'Università di Leida, che è incaricata del controllo della qualità della didattica del dipartimento.

5. Analisi e interpretazione dei dati

Nel periodo 2004-2020 sono stati somministrati in totale 2703 test, in questo conteggio fanno parte i test in itinere e il primo tentativo dell'esame nella sessione di recupero. Ad un confronto con il periodo 1991-2002 in cui veniva somministrato un unico esame finale, il rendimento è salito mediamente dell'11%: attualmente è su una media del 69%, nel periodo precedente preso a riferimento rimaneva in media al 58%. Si tratta in questo caso di un paragone ibrido, se pensiamo alla grande quantità di fattori di influenza in continuo sviluppo. Ciò nonostante, poiché i lassi di tempo a confronto sono piuttosto lunghi e filtrano tali fattori di influenza, possiamo concludere che questo effetto positivo sia stato realmente indotto da una diversa distribuzione nel tempo dei test e dalla peculiarità del sistema di prove in itinere. Oltre a questo *wb* intenzionale tuttavia, si osservano effetti non previsti sia positivi sia negativi, che sono stati rilevati durante le revisioni cicliche dei test nel corso degli anni e che caso per caso hanno comportato riparazioni.

5.1. *Effetti positivi non previsti*

Una ricaduta positiva, difficilmente calcolabile a priori, riguarda il risparmio di tempo realizzato grazie alla correzione automatica della piattaforma; nel corso di 16 anni il risparmio totale è di circa 900 ore, paragonato alle condizioni di una correzione del medesimo test in versione cartacea, che era pari a 20 minuti per ogni singolo test. Per contro l'investimento nella creazione del *database* con circa 4500 domande ha impegnato il docente per 120 ore e uno studente nel ruolo di assistente per 500 ore. Le piattaforme utilizzate, per effetto degli sviluppi tecnologici risultano essere molto rapidamente obsolete: nei 16 anni sono state impiegate tre piattaforme, considerando che l'ultima (*Remindo*) è stata introdotta nel 2018, possiamo dedurre che la vita media di una piattaforma è dagli otto ai dieci anni. La conversione degli *item* dei test da una piattaforma all'altra avviene manualmente e richiede in media 100 ore di lavoro. Il *database* si è dimostrato oltre a tutto efficace nel periodo dell'emergenza sanitaria Covid-19, i test si sono potuti tenere regolarmente senza rinvii, con l'aiuto dei sistemi di sorveglianza digitali *ProctorExam* e *Proctorio*, che venivano attivati contemporaneamente durante la somministrazione dei test.

Un altro effetto positivo ripetutamente menzionato dagli studenti nelle interviste individuali e nelle valutazioni della qualità della didattica condotte dalla

commissione paritetica è rappresentato dalla disponibilità immediata dei risultati del test, nonché dal *feedback* pressoché costante che accompagna lo studente per tutto lo sviluppo della lingua durante il primo anno di studio. L'insegnante può individuare rapidamente le aree con cui gli studenti incontrano maggiori difficoltà e può intervenire con attività specifiche di sostegno. Osservando il comportamento dello studente in classe, si nota una maggiore attenzione nel monitorare l'adeguatezza grammaticale nei momenti di comunicazione in lingua straniera. Altri effetti benefici notevoli si possono percepire ugualmente in classe: le lezioni ruotano primariamente intorno a quelle attività per le quali il computer si dimostra un mezzo sterile o artificioso nel generare *feedback* per l'apprendimento in autonomia. Sono stati necessari più interventi ciclici di riparazione ai test per contenere effetti negativi di selettività che rischiano di provocare il ritiro dello studente già dopo poche settimane di studio. Attualmente, dopo diverse modifiche nel corso degli anni, per la prima prova è consentito il voto insufficiente, per la seconda il voto insufficiente minimo è 5 (su una scala da 1 a 10), per la terza il voto deve essere sempre sufficiente. Il voto finale è dato dalla media ponderata che deve corrispondere a una sufficienza (voto da 6 a 10). Nella terza prova vengono valutati argomenti della grammatica già presenti nella prima e nella seconda prova. L'esame della sessione di recupero è basato sui contenuti di tutte e tre le prove.

5.2. *Effetti negativi non previsti*

Dai questionari sul controllo della qualità didattica della commissione paritetica è emerso che alcuni studenti si chiedono se sia veramente necessario dover sostenere un numero così elevato di test. Non è escluso che questo commento possa provenire proprio dal gruppo di studenti già capace di pianificare il proprio studio. Qualche studente invece manifesta un certo disagio derivato dal carattere coercitivo del sistema. Inizialmente avevamo supposto che lo studente fosse in grado di trasferire l'esperienza degli effetti positivi dei test in itinere per pianificare meglio le ore di studio. Al contrario abbiamo notato una scarsa propensione alla pianificazione anche negli studenti del secondo anno di *Bachelor*. Il sistema di test in itinere sembra preservare lo studente dall'esposizione dei benefici della pianificazione. Per far fronte a questa ricaduta negativa, si è reso necessario estendere il sistema di test di lingua in itinere anche al secondo anno di *Bachelor*. Affrontando sempre ciclicamente la riparazione, abbiamo deciso successivamente di introdurre un numero inferiore di test intermedi al secondo anno. Con questa progressione discendente del numero di test rispetto al primo anno, lo studente si abitua a un'autonomia maggiore nella pianificazione. Al terzo anno di *Bachelor* gli studenti al contrario sembrano avere più affinità con la capacità di pianificare: osserviamo infatti che tendono a rispettare in misura maggiore le scadenze delle consegne dei propri lavori scritti. Non è escluso che la capacità di pianificare proceda di pari passo con il processo di maturità (gli studenti olandesi arrivano all'università un

anno prima rispetto a quelli italiani) e con un aumento della consapevolezza che questa capacità possa portare notevoli benefici nella gestione del tempo relativamente al rapporto fra obblighi di studio e la sfera privata. Dal punto di vista del dipartimento gli effetti positivi si notano nel numero di studenti che riesce a concludere lo studio in un tempo relativamente breve.

Benché il sistema preveda la possibilità di recuperare dopo un'eventuale falsa partenza, molti studenti si sentono demoralizzati e abbandono lo studio già dopo poche settimane dall'inizio. Proprio per questo motivo siamo intervenuti più volte nella riparazione ciclica della media ponderata del primo e del secondo test. Oltre al pericolo di abbandono dello studio abbiamo constatato che nel comportamento dello studente può subentrare un atteggiamento calcolatore: sapendo di poter recuperare le insufficienze dei primi due test, si impegna solo per l'ultima prova. L'ultimo effetto palesemente negativo, notato anche negli studi precedenti di *wb* (Buck, 1988; Chan, 2020), è che gli apprendenti si concentrano solo sui test e mettono sotto pressione l'insegnante con la richiesta di esercizi supplementari per preparare le prove durante le ore in presenza.

6. Conclusioni

In questa ricerca sugli effetti intenzionali di *wb* abbiamo supposto che sostituendo un unico esame finale con più test in itinere, lo studente fosse indotto ad affrontare la parte dello studio della lingua in autonomia già a partire dalle prime settimane del modulo. Abbiamo inoltre ipotizzato che, grazie ad una migliore sinergia fra acquisizione in classe e apprendimento in autonomia, le attività di classe potessero trarre notevoli benefici. L'intervento ha prodotto una ricaduta positiva per entrambe le ipotesi, a questo si aggiunge un aumento medio dell'11% del rendimento, dato quantitativo che conferma la ricaduta positiva. Oltre ad avere un *wb* intenzionale positivo, le prove di valutazione hanno prodotto effetti non prevedibili tanto di tipo positivo quanto negativo (Andrews, 2004). Attraverso l'intervento ciclico di riparazione dei test, come è stato suggerito da Chan (2018), possiamo ridurre o controllare gli effetti negativi. In questa ottica il *wb* non viene esaminato in modo statico, ma come un processo in corso.

Questo studio mostra inoltre, come era stato proposto anche da Alderson e Wall (1993), che i test possono generare contemporaneamente *wb* benefico e dannoso. La complessità del fenomeno richiede di dover disegnare una mappatura degli effetti, poiché spesso tendono a presentarsi in un legame di interazione dinamica. Nella valutazione del *wb* intenzionale in relazione a quello non previsto, sarà inevitabile sviluppare una scala graduata delle priorità e soppesare le singole ricadute in un rapporto di interconnessione e come interagenti fra di loro. Difficilmente sarà possibile eliminare integralmente tutte le ricadute negative. Al fine di ottimizzare l'individuazione del *wb* è consigliabile osservare gli effetti dei test per

un periodo prolungato (Chan, 2018). Solo in questo modo si potranno individuare le ricadute ricorrenti e i fattori di influenza verranno filtrati.

Possiamo essere d'accordo sul fatto che i test ad elevata posta in gioco tendono ad avere ricadute più significative per gli individui e per la dimensione sociale (Buck, 1988; Alderson e Wall, 1993). Dobbiamo però considerare che tutti i test sono destinati a produrre effetti, la mancata osservazione delle ricadute può semplicemente dipendere dalla volontà di non volerle rilevare.

In questa ricerca abbiamo appurato di nuovo un fenomeno frequentemente segnalato negli studi precedenti: l'attenzione per la preparazione dei test da parte di insegnati e apprendenti rischia di entrare palesemente in contrasto con le mete didattiche, linguistiche e formative (Buck, 1988; Chan, 2020); un numero elevato di test può costituire una seria minaccia per quegli aspetti dell'apprendimento che non rientrano direttamente nei contenuti delle prove somministrate, ma che sono tuttavia importanti.

Nel caso del docente che è anche costruttore di test e/o valutatore le prove di verifica possono realmente rappresentare uno strumento per influenzare in modo positivo il comportamento degli apprendenti.

Riferimenti bibliografici

Alderson J.C., Wall D. (1993), Does washback exist?, *Applied Linguistics*, 14 (2), pp. 115-129.

Andrews S. (2004), "Washback and curriculum innovation", in L. Cheng, Y. Watanabe (edited by), *Washback in language testing: Research, context and method*, London, Erlbaum, pp. 37-52.

Bachman L.F., Palmer A.S. (1996), *Language testing in practice: designing and developing useful language tests*, Oxford, Oxford University Press.

Bailey K.M. (1996), Working for washback: a review of the washback concept in language testing, *Language Testing*, 13 (3), pp. 257-279.

Buck G. (1988), Testing listening comprehension in Japanese university entrance examinations, *JALT Journal*, 10, pp. 15-42.

Chan K.L.R. (2018), "Washback in English pronunciation in Hong Kong: Hong Kong English or British English?", in K.G.D. Chan *et al.* (edited by), *Motivation, identity and autonomy in foreign language education*, Singapore, NUS Centre for Language Studies, pp. 27-40.

Chan K.L.R. (2020) Washback in education: a critical review and its implications for language teachers, *Journal of Foreign Language Education and Technology*, 5 (1), pp. 108-124.

Cheng L., Curtis A. (2004), "Washback or backwash: a review of the impact of testing on teaching and learning", in L. Cheng, Y. Watanabe (edited by), *Wa-*

shback in language testing: Research, context and method, London, Erlbaum, pp. 3-18.
Frederiksen J.R., Collins A. (1989), A systems approach to educational testing, *Educational Researcher*, 18 (9), pp. 27-32.
Hughes A. (1989), *Testing for language teachers*, Cambridge, Cambridge University Press.
Madaus G.F. (1988), "The influence of testing on the curriculum", in L.N. Tanner (edited by), *Critical issues in curriculum: eighty-seventh yearbook of the National Society for the Study of Education*, Chicago, University of Chicago Press, pp. 83-121.
Messick S. (1989), "Validity", in R.L. Linn (edited by), *Educational Measurement*, New York, ACE-Macmillan, pp. 13-103.
Messick S. (1996), Validity and washback in language testing, *Language Testing*, 13 (3), pp. 241-256.
Pan Y.C. (2009), Review of washback and its pedagogical implications, *VNU Journal of Science*, *Foreign Languages*, 25, pp. 257-263.
Popham W.J. (1987), The merits of measurement driven instruction, *Phi Delt Kappan*, 68 (9), pp. 679-682.
Qi L. (2007), Is testing an efficient agent for pedagogical change? Examining the intended washback of the writing task in a high-stakes English test in China, *Assessment in Education*, 14 (1), pp. 51-74.
Smith M.L. (1991), Put to the test: the effects of external testing on teachers, *Educational Researcher*, 20 (5), pp. 8-11.