Franco M. Impellizzeri, *Centro di ricerca interuniversitario in Biomeccanica e scienze motorie, CeBiSM, Rovereto; Dipartimento di ricerca e sviluppo, Schulthess Klinik, Zurigo, Svizzera,* **Pierluigi Fiorella,** *Istituto di Medicina e scienza dello Sport, Coni, Roma; Settore Sanitario F.C. Internazionale,* **Maurizio Fanchini,** *Centro di ricerca interuniversitario in Biomeccanica e scienze motorie, CeBiSM, Rovereto;* **Duccio Ferrari Bravo,** *Facoltà di Scienze motorie, Università degli Studi di Verona,* **Carlo Castagna,** *Corso di Laurea in Scienze motorie, Facoltà di Medicina e chirurgia di Roma Tor Vergata, Roma.*

Foto Calzetti & Mariucci Editori

I TEST DI VALUTAZIONE: QUELLO CHE NON VIENE MAI DETTO

Basi teoriche per la validazione dei test e l'esempio del calcio

Un test per essere considerato valido deve soddisfare una serie di criteri di qualità senza i quali non può essere utilizzato. Riferendole all'esempio del gioco del calcio, si espongono le basi teoriche utili per la validazione di un test mostrando come le tecniche usate per studiare le misure dei costrutti teorici possano, e debbano, essere impiegate per la misura di variabili oggettive (variabili fisiologiche). Il processo di validazione di un test è un percorso scientifico rigoroso che prevede almeno cinque step indispensabili e atti a valutare altrettanti attributi che devono essere presenti nel test. Tali attributi sono rappresentati dal modello teorico, cioè dalla definizione del modello della prestazione di riferimento per l'identificazione delle caratteristiche rilevanti per la performance, allo scopo di verificare se la caratteristica che si vuole misurare è rilevante; dalla validità, cioè la determinazione del grado con cui un test misura ciò che si suppone debba misurare; la ripetibilità, che decreta la stabilità di una misura ripetuta nelle stesse condizioni e nello stesso soggetto (rumore); la responsività interna, cioè il grado con cui il risultato di un test è sensibile ai cambiamenti indotti da un intervento (segnale), ed esterna, rappresentata dal grado con cui i cambiamenti di un test riflettono i cambiamenti della misura di riferimento; la interpretabilità, che è il grado della possibilità di intepretazione di un test (dati di riferimento, cambiamento minimo significativo da un punto di vista pratico e statistico, rapporto rumore: segnale). Prima di considerare valido un test, vanno rigorosamente verificate le sue proprietà e i risultati devono soddisfare in maniera adeguata ognuno di questi cinque attributi. Si forniscono poi indicazioni di come tecniche statistiche aiutino l'interpretazione dei test, sia a livello di gruppo, sia a livello statistico.

Premessa

Quando si parla di test di valutazione tutti più o meno pensano di sapere di cosa si tratta. In realtà l'area dei test (e quindi delle proprietà delle misure) è uno degli ambiti più complessi e spesso sottovalutati delle scienze dello sport e dell'esercizio fisico (Impellizzeri, Marcora 2009, *in press*). A fronte di un aumento dell'interesse verso la valutazione funzionale, non si è registrato lo stesso aumento di interesse – e quindi divulgazione – per i fondamenti teorici che stanno alla base del loro sviluppo, validazione ed interpretazione. Si è, quindi, assistito negli ultimi anni a un proliferare incontrollato di test. In un momento in cui sembra andare di moda il binomio *sport e scienza*, non c'è cosa peggiore e più pericolosa della pseudoscienza. Con questo articolo vogliamo fornire una introduzione *user friendly*, ma scientificamente rigorosa, della teoria che sta alla base dello sviluppo dei test che può essere utile a coloro che vogliono ideare un nuovo test, validare test già esistenti o semplicemente capire se i test che vengono proposti sono validi o no. Faremo riferimento al calcio perchè tra i vari sport è tra quelli che più soffre di questo proliferare di nuovi test, ma i fondamenti teorici presentati sono ovviamente applicabili a tutti gli sport. Non approfondiremo le tecniche statistiche più avanzate per le quali rimandiamo a testi specifici. Daremo invece indicazioni di come le tecniche statistiche possano aiutare a interpretare i test sia a livello di gruppo, ma, soprattutto, come richiesto dagli allenatori e preparatori, a livello individuale.

Introduzione

I metodi e le tecniche statistiche necessarie per determinare la validità delle prove di valutazione derivano dal dominio della psicologia e della sociologia (Ary et al. 2006). Purtroppo nella loro applicazione nell'area dello sport si sono spesso persi alcuni passaggi fondamentali. Questo ha condotto nel breve volgere di tempo al facile sorgere di una miriade di test da campo che risultano spesso non "validi". Nel presente articolo faremo particolare riferimento alle tecniche clinimetriche (derivate dalla psicometrica), ovvero quell'area scientifica che si occupa della qualità delle misure cliniche e in particolare dei cosiddetti *self-report* (strumenti utilizzati per quantificare la percezione dei pazienti rispetto all'esito di trattamenti o all'impatto di particolari patologie) (de Vet et al. 2003). In pratica un test per essere considerato valido deve soddisfare una serie di criteri di qualità senza i quali il test non può essere utilizzato. In quanto segue verranno riportate le basi teoriche utili per la validazione di un test mostrando come le tecniche usate per studiare le misure dei costrutti teorici possano, e debbano, essere impiegate per la misura delle variabili oggettive (variabili fisiologiche). Il processo di validazione di un test è un percorso scientifico rigoroso che prevede almeno cinque *step* indispensabili e atti a valutare altrettanti attributi che devono essere presenti nel test. Quindi, prima di considerare valido un test, vanno rigorosamente verificate le sue proprietà e i risultati devono soddisfare in maniera adeguata ognuno dei seguenti cinque attributi.

Step	Attributo da verificare	Verifica da operare
1	Modello teorico	Definire il modello della prestazione di riferimento per l'identificazione delle caratteristiche rilevanti per la *performance*. Verificare se la caratteristica che si vuole misurare è rilevante.
2	Validità	Determinare il grado con cui un test misura ciò che si suppone debba misurare.
3	Ripetibilità	Decretare la stabilità di una misura ripetuta nelle stesse condizioni e nello stesso soggetto (rumore).
4	Responsività	*Interna:* Grado con cui il risultato di un test è sensibile ai cambiamenti indotti da un intervento (segnale). *Esterna:* grado con cui i cambiamenti di un test riflettono i cambiamenti della misura di riferimento.
5	Interpretabilità	Grado di interpretabilità di un test (dati di riferimento, cambiamento minimo significativo da un punto di vista pratico e statistico, rapporto rumore: segnale).

Tabella 1 – Attributi che devono essere verificati durante il processo di validazione di un test

Modello teorico

Il modello teorico è il requisito di partenza essenziale per sviluppare o validare un test (Impellizzeri, Marcora 2009, *in press*). È bene non identificare esclusivamente il modello teorico con quello fisiologico dato che non sempre i fattori limitanti la prestazione sono ascrivibili a fattori di natura fisica. Di fatto tale identificazione determinerebbe a priori una esclusione di altre componenti importanti. Quali, ad esempio, gli aspetti psicobiologici, psicologici, biomeccanici, e così via. A dimostrazione di ciò possiamo citare il recente articolo di Marcora e coll. (Marcora et al. 2009) nel quale è stato dimostrato come la fatica mentale costituisca un fattore limitante la prestazione di *endurance*. Sicuramente più adeguato risulta il termine *modello prestativo*, più generico, ma che racchiude in sé la finalità principe degli operatori dello sport, ovvero la prestazione. È di fatto attraverso la definizione di un modello teorico della prestazione che si identificano le componenti misurabili del modello stesso. Per chiarire questa importante procedura faremo riferimento, quale paradigma esemplificativo, al gioco del calcio (Impellizzeri, Marcora 2009, *in press*).

Innanzitutto occorre definire cosa si intenda per *prestazione calcistica*. Questa procedura, necessaria, ma non sufficiente per la validazione di un test, pur apparentemente semplice, presenta le difficoltà tipiche della soluzione sistemica semplificata (modello) a problemi complessi e, di fatto, risulta nella soluzione tutt'altro che immediata. Difatti l'assumere quale paradigma prestativo il costrutto più immediato, ovvero l'*esito di un incontro*" (vittoria-sconfitta-pareggio), renderebbe il risultato troppo influenzabile da quei fattori episodici che condizionano l'esito della partita. Una soluzione "più stabile" potrebbe essere la classifica finale di un Campionato o un torneo che, che per quanto influenzabile da episodi, dovrebbe meglio riflettere il valore di una squadra. In questo contesto un'altro indicatore utile per poter definire la *performance* calcistica potrebbe essere quindi il *ranking* FIFA, che a sua volta è il risultato dei successi o insuccessi delle squadre di ogni Nazione. Qualcuno ovviamente potrà dissentire da ciò, il che già fa comprendere come sia di fatto complicato sviluppare un modello teorico (Taylor et al. 2008). È comunque vero che senza una definizione della prestazione non è possibile definire le componenti che la influenzano o determinano, ed è quindi impossibile sviluppare un test per misurare questi fattori decisivi. Occorre ricordare che un modello è comunque una semplicazione di un fenomeno multifattoriale di vario livello di complessità e che come tale va considerato, senza incorrere nell'errore di banalizzare le risposte da questo afferenti. In ogni caso l'identificazione della *performance* quale

risultato finale in una classifica risulta la base su cui si muove tutta la ricerca applicata al calcio. Partendo da questo assunto il gruppo danese di Bangsbø (Mohr et al. 2003) ha mostrato come i giocatori danesi corressero meno ad alta intensità (>15 km/h) dei giocatori Italiani. Appartenendo i giocatori italiani ad un livello competitivo superiore (più vittoriosi e quindi meglio posizionati nel *ranking* FIFA), si è concluso che la capacità di correre ad alta intensità sia un fattore importante per il calcio. Questa è l'unica evidenza a supporto del modello oggi più usato nel calcio, ma essendo una conclusione abbastanza logica e facilmente condivisibile è stata ed è comunemente accettata. Quindi, si è affermato un modello tri-compartimentale che comprende oltre alla componente fisica quelle relative ai fattori tecnici e tattici. Risulta evidente che qualsiasi tentativo operato nell'intento di fornire un modello integrato e, quindi, "uni-compartimentale" o pseudo-integrato determinerebbe un aumento del livello esplicativo del modello stesso. Il primo di questi tentativi è stato di recente effettuato in uno studio realizzato esaminando il Campionato italiano e nel quale è stato dimostrato come i giocatori delle migliori cinque squadre corressero mediamente meno ad alta intensità dei giocatori che militavano nelle ultime cinque squadre del campionato (Rampinini et al. 2007b). La differenza tra i due gruppi, invece, è stata trovata nell'attività svolta ad alta intensità in possesso di palla, nella quale i giocatori appartenenti alle squadre meglio classificate eccellevano. La stessa cosa è stata confermata in uno studio successivo svolto sulla *Premier League* (Di Salvo et al. 2009). Quanto sopra senza dubbi suggerisce che nella definizione di un modello teorico calcistico bisogna tenere conto delle interazioni tra i tre fattori e non delle singole componenti. Un ulteriore passo nella definizione di un modello integrato della prestazione calcistica è stato compiuto grazie a recenti pubblicazioni, le quali hanno dimostrato come una maggiore abilità di svolgere attività intermittenti ad alta intensità possa avere un effetto indiretto, e non diretto, sulla *performance*. Questo attenuando il decremento della qualità tecnica causato dall'insorgere della fatica sia temporanea che cumulativa (Impellizzeri et al. 2008b; Rampinini et al. 2008). Questa acquisizione scientifica risulta di rilevante importanza dato che sicuramente qualche preparatore fisico si sarà trovato di fronte all'obiezione, esercitata da allenatori che banalizzando la *performance* calcistica e, quindi, il relativo modello, la associano al solo evento tecnico ("giocare al pallone", "mettere la palla in rete"). Tale obiezione pur lecita, risulta indebolita di fatto dalla dimostrazione che i giocatori che hanno risultati inferiori nello *Yo Yo Intermittent Recovery Test* sono anche quelli che sbagliano un maggior numero di passaggi successivamente a uno sforzo intermittente che induce fatica (Impellizzeri et al. 2008b; Rampinini et al. 2008). In pratica questa è la dimostrazione attualmente più forte del potenziale ruolo determinante di una buona preparazione fisica nel calcio. In altre parole, forse il giocatore di successo non corre di più, ma sbaglia meno; e se allenato, tende a sbagliare ancora meno. Ovviamente sull'interazione tra capacità tecniche e fisiche occorrono ancora molti studi, ma l'inizio sembra molto incoraggiante e utile per raffinare il modello teorico di partenza.

E la forza? Sicuramente a questo punto qualcuno si sarà chiesto come entra la forza in questo modello, data la controversia esistente tra sostenitori della forza o dell'aerobico. La forza potrebbe rientrare in questo modello nel caso che la si considerasse una determinante della capacità di correre ad alta intensità (che infatti include gli *sprint*). L'allenamento aerobico ad alta intensità è stato mostrato da vari studi influenzare la *performance* fisica e tecnica, e le evidenze di questo sono sostanziali e abbastanza forti (Bravo et al. 2007; Dupont et al. 2004; Helgerud et al. 2001; Impellizzeri et al. 2006; Impellizzeri et al. 2008b; McMillan et al. 2005; Siegler et al. 2003; Stølen et al. 2005). Per la forza non ci sono queste evidenze. Alcuni studi supportano l'utilità della forza per il fatto che un giocatore possa saltare più in alto nei colpi di testa o che possa correre più veloce uno *sprint* arrivando prima sulla palla (Hoff, Helgerud 2004; Wisløff et al. 2004). Uno studio di Wisløff et al. (1998) ha mostrato come, nel Campionato norvegese i giocatori della squadra prima classificata nel Campionato avessero prestazioni nei test di forza massima più elevate dei giocatori dell'ultima squadra del campionato. La stessa cosa non fu invece trovata nei test di salto. Arnason et al. (2004a) hanno mostrato una correlazione tra salto verticale e classifica finale nel Campionato islandese. Al contrario, Cometti et al. (2001) mostrarono che i giocatori amatori francesi saltavano di più dei giocatori di categorie *élite*. Rosh et al. (2000) hanno evidenziato come i giocatori amatori avessero valori di salto verticale più basso in confronto dei *top level* e di terza divisione. Tuttavia, gli stessi autori trovarono test di

salto simili tra *top level*, terza divisione e giocatori di squadre locali (basso livello). Quindi, anche se la forza esplosiva, misurata tramite i salti verticali, appare essere intuitivamente un caratteristica importante per i calciatori non ci sono evidenze scientifiche del ruolo determinante di questa caratteristica sulla *performance* calcistica. Altri ancora suggeriscono che la forza avrebbe un significato in termini di prevenzione degli infortuni. In questo ambito esistono certamente più evidenze (Arnason et al. 2004a; Arnason et al. 2004b; Croisier 2004; Croisier, Crielaard 2001; Croisier et al. 2002; Croisier et al. 2003; Dvorak, Junge 2000; Junge, Dvorak 2004). Di fatto, se si considera che l'infortunio possa influire negativamente sulla prestazione della squadra, la forza andrebbe inserita in un modello che vede l'infortunio tra le determinanti delle prestazione calcistica, così come tutto ciò che ha un ruolo nella prevenzione. In pratica, mentre molti si sono adoperati per sviluppare test per la misura della forza nelle sue varie espressioni e metodi più o meno originali per migliorarla, pochi studi hanno cercato di verificare la relazione della forza con la *performance* (diretta o indiretta), o si sono preoccupati di definire il costrutto che influenzerebbe la *performance* calcistica e al quale la forza sarebbe legata.

Nell'ottimizzazione del modello, l'allenamento della forza, e di tutto ciò che ha, o potrebbe avere un ruolo preventivo, potrebbe influenzare l'incidenza degli infortuni e di conseguenza i tre costrutti di base (*performance* fisica, tecnica e tattica). Infatti, un atleta infortunato non può giocare o se gioca non può esprimere al massimo le tre componenti della *performance*. Queste considerazioni forniscono la base teorica per inserire, ad esempio, i lavori propriocettivi tra i fattori da studiare. Dato che lo scopo nel calcio è vincere, e non avere atleti più forti, fintanto che non verrà chiarito se e come la forza vada inserita nel modello prestativo calcistico, non è possibile spiegare come misurarla ed allenarla in modo appropriato. Questo non vuol dire che la forza, come altre capacità non siano rilevanti in assoluto, significa semplicemente che non si sa quanto e quale espressione di forza sia importante. Questo è più rilevante di quanto si pensi. Se, ad esempio, un domani si scoprisse che è la forza esplosiva ad avere un ruolo determinante nella prestazione, i test di salto avrebbero senso mentre i test di forza massima no, oppure viceversa. Attualmente le evidenze sono discordanti.

Il modello teorico, quindi, necessiterebbe anch'esso di validazione e possibilmente questa validazione dovrebbe essere sperimentale per non renderci conto tra un decennio di aver perso tempo a misurare e allenare variabili che hanno un minimo impatto sulla *performance*. Mentre in altre discipline si dedica molto tempo alla costruzione del modello teorico, nel calcio, e nello sport in genere, questo aspetto è sottovalutato, rendendo il tipo di test e i metodi di allenamento dettati più da mode momentanee (e quindi transitorie e in genere cicliche) che da evidenze scientifiche (solide nel tempo).

Sempre partendo dal modello teorico, lo sviluppo di un nuovo test deve essere motivato. Prima di ideare un nuovo test occorre spiegare perché i test esistenti non sono adeguati e possibilmente dimostrarlo. Una motivazione potrebbe essere, ad esempio, che i test a disposizione non sono validati. Anche in questo caso occorrerebbe spiegare perché sia necessario crearne uno nuovo invece di validare, e capire come usare, un test già esistente. Purtroppo, spesso l'introduzione di un nuovo test costituisce la manifestazione della ricerca di popolarità e di notorietà del suo propositore, più che di una reale necessità pratica. Un esempio positivo in questo senso può essere quello offerto da una recente ricerca che si è interessata della *repeated sprint ability* (*RSA*) nel calcio, ovvero della abilità del giocatore di reiterare sprint con brevi pause di recupero e con il minimo deterioramento della prestazione. In questo studio infatti, invece di "inventare" un nuovo test per la determinazione della *RSA*, gli Autori hanno ritenuto di importante rilevanza pratica verificare la validità e ripetibilità del popolare test a navetta di Capanna, già diffuso nel calcio italiano (Impellizzeri et al. 2008a; Rampinini et al. 2007a).

Validità del test

La validità è l'abilità di un test nel misurare ciò che si suppone debba misurare. Ci sono vari modi per verificare la validità di un test e quindi diversi tipi di questa. La validità non è un concetto assoluto. Un test può essere valido per uno scopo e non per un altro. I più frequenti tipi di validità utilizzati nelle scienze dello sport sono quella di facciata (*face validity*), logica o di contenuto, di costrutto e di criterio.

Validità di facciata

È la forma più debole di validità perché non dimostrabile in modo oggettivo e, quindi, troppo soggetta ad interpretazioni personali. Purtroppo è anche il tipo di validità che più viene utilizzato per "dare" validità ai test e viene spesso identificata con la specificità del test. Si dice che un test ha validità

di facciata quando *appare* misurare quello che l'ideatore vorrebbe misurare. Ad esempio, se voglio misurare nel giocatore la capacità di effettuare sprint ripetuti e utilizzo un test che consiste nel fare 6 sprint da 40 m con 20 secondi di recupero (il test di Capanna ad esempio), questo test ha validità di facciata perché appare misurare la capacità di ripetere gli sprint (specificità apparente). In genere si assume, senza dimostrarlo, che un test avente "specificità apparente" sia sicuramente migliore e di conseguenza valido. In pratica, utilizzando un test che riproduce movimenti e azioni dello sport in questione, assumo che la sua specificità gli conferisca una validità di facciata, e che questa validità sia sufficiente per validare il test in relazione alla *performance*. Purtroppo questo non è il caso. Recentemente è stato pubblicato uno studio che ha verificato la ripetibilità di una serie di valutazioni funzionali per il calcio, incluso un test per valutare quanto lontano un giocatore lanci la palla durante una rimessa laterale (Mirkov et al. 2008). Sebbene di fatto la rimessa laterale costituisca un elemento tecnico del gioco del calcio (specificità apparente), l'introduzione di un test in grado di valutare l'abilità di un giocatore in questo gesto risulta giustificato solo nel caso che si verifichi la sua rilevanza nel determinismo prestativo del gioco. Di fatto la validità di facciata non apporta per definizione nulla rispetto alla rilevanza del test stesso per la *performance*. Ancora una volta il modello teorico alla base dello sviluppo del test risulta essenziale. Alla luce di quanto appena esemplificato una erronea interpretazione della vali-

dità di facciata, con la conseguente assunzione di comprovata pertinenza del test (rimessa laterale), potrebbe indurre alcuni preparatori fisici ad allenare la forza degli arti superiori, per esempio, due giorni alla settimana per aumentare di 30 cm la lunghezza della rimessa laterale. Questo senza sapere se questi 30 cm in più influenzeranno veramente la *performance*. Una cosa risulterebbe certa da questa strategia, ovvero la sottrazione di tempo all'allenamento di caratteristiche fisiche magari più rilevanti per la prestazione in oggetto. Un altro esempio di segno contrario è quello offertoci dall'applicabilità al calcio dello *Yo-Yo Intermittent Recovery Test,* il quale consiste nel completare a velocità crescente navette da 20 + 20 m con cambi di direzione di 180° intercalati da 10 secondi di recupero. Qualcuno ne ha contestato lecitamente la specificità dato che, pur contemplando come nel calcio sforzi intermittenti (e per questo più specifico di test continui a navetta come il *Legér*), il protocollo previsto dallo *Yo-Yo Intermittent Recovery Test* risulta differente dalle attività fatte in partita (intermittenza casuale). Tuttavia numerosi studi ne hanno sancito la validità in modo forte e consistente (Bangsbø et al. 2008). In questo caso, come si può notare, si sarebbe addotta una non validità di facciata (protocollo intermittente ad esaurimento contro intermittenza casuale) per dimostrare l'incongruenza dello *Yo-Yo Intermittent Recovery Test* nel calcio. Le verifiche, tuttavia, hanno mostrato, come questa apparente scarsa validità di facciata non ne abbia inficiato la validità.

Uno studio ancora in corso (dati non pubblicati) ha tentato di modificare il test rendendolo più specifico, aggiungendo cambi di direzione di diverse angolazioni, movimenti tipici del calcio come slalom etc. Il risultato ad oggi è che la correlazione tra *Yo Yo* tradizionale e la versione modificata è superiore a 0,90. Questo indicherebbe, senza ombre di dubbio, che i due test sono perfettamente equivalenti e che, di fatto, misurano le stesse variabili. In pratica *la maggiore specificità introdotta complica di fatto solo la fattibilità del test.* Fare riferimento solo alla validità di facciata è un processo soggettivo, e pertanto rischia di condurre lo sviluppatore del test in direzioni fuorvianti nella spasmodica ricerca della specificità che, per quanto desiderabile, non sempre è sinonimo di maggior validità o, in ogni caso, questa presunta maggior validità va dimostrata.

Validità di contenuto

Nel caso dell'esempio riferito al test a navetta (6x40 m) dicevamo che esso ha validità di facciata. Abbiamo anche ricordato che la validità di facciata è un procedimento soggettivo e quindi non oggettivabile. Si potrebbe infatti obiettare sulla sua validità di facciata con l'osservazione che in partita i giocatori raramente fanno *sprint* di 40 m, assai di rado effettuano cambi di senso di 180° e che i recuperi tra uno *sprint* e l'altro sono diversi dai 20 secondi utilizzati nel test. In pratica stiamo mettendo in dubbio la sua *validità di contenuto*. Questo tipo di validità è più oggettivabile. Ad esempio, nel caso del test a navetta posso utilizzare i dati misurati in partita (attraverso analisi video: *match analysis*) per verificare la lunghezza, frequenza e durata dei recuperi durante gli *sprint* e in particolare delle fasi più intense di gioco dato che in genere è bene indagare non un andamento medio, ma le situazioni più critiche di questo. In realtà pochi sono i dati reperibili in questo contesto. In uno studio datato (Withers et al. 1982) è stato mostrato come nelle fasi più intense i giocatori arrivino a compiere sforzi intensi come gli sprint con un rapporto tra sforzo e recupero di 1 a 3, rapporto che si avvicina al test a navetta. Il test a navetta prevede *sprint* da 40 m che non sono frequenti nel calcio. Tuttavia il test prevede un'andata e un ritorno, e 40 m è la distanza totale. Quindi si sta parlando in realtà di *sprint* da 20 metri, i quali sono più frequenti durante la partita. Il punto più debole è il cambio di senso di 180° che avviene raramente in partita. Tuttavia per capire se è accettabile questo tipo di cambio di direzione occorrerebbe vedere se un test a navetta con cambi di direzione di 180° sia correlato con il risultato di un test a navetta con cambi più specifici rispetto a quelli che avvengono in partita. Studi su questo sono in corso sia per il test a navetta che per altri test come lo *Yo-Yo*. Perché, quindi, non utilizzare direttamente cambi di direzione più specifici? La risposta potrebbe essere perché un test con cambi di senso di 180° risulta più facile da eseguire, e nello sviluppo di un test la sua fattibilità è uno degli elementi importanti di cui tenere conto. In ogni caso, come per la validità di facciata, diversi studi hanno mostrato la validità (vedi sezione successiva) nonostante gli apparenti problemi di validità di contenuto. Altro esempio di validità di contenuto è quello fornito dalle simulazioni che vengono in genere sviluppate per studiare le risposte fisiologiche alla partita. Una buona simulazione della *performance* fisica della partita ovviamente deve prendere in considerazione le attività realmente effettuate in partita.
Confrontando le proporzioni di ciascuna attività fisica contenuta nel test con i dati disponibili di *match-analysis* è possibile verificare in modo quantitativo la validità di contenuto.

Validità di costrutto

La validità di costrutto si riferisce all'idoneità di un test nel misurare un concetto teorico che si suppone debba misurare, ovvero il *costrutto*. Nelle scienze dello sport risulta difficile per motivi culturali pensare alle caratteristiche fisiche e alla prestazione come a costrutti teorici. Il fatto che nello sport vi siano molte più variabili e parametri misurabili in modo oggettivo ha fatto spesso dimenticare che molte *performance* sportive rappresentano di fatto costrutti teorici (Atkinson 2002). La performance del calcio è un costrutto: la classifica finale in un Campionato è un surrogato del costrutto, e non il costrutto stesso. Lo stesso ragionamento vale anche per la *performance* fisica durante la partita: la distanza corsa ad alta intensità in partita è un indicatore del costrutto *performance* fisica, e non il costrutto stesso. Altri esempi di costrutti sono la forza e le capacità aerobiche. Da questi esempi risulta edivente, quindi, quanto la validità di costrutto sia importante per verificare se il test che abbiamo sviluppato o stiamo validando misuri effettivamente il costrutto di interesse. In questo contesto risulta di estrema importanza la oggettiva definizione del modello teorico, la quale ci indica se il costrutto di interesse è rilevante per la *performance*. Esistono vari metodi per verificare la validità di costrutto e qui di seguito verranno considerati i più utilizzati.

DIFFERENZA NOTA TRA GRUPPI (KNOWN-GROUP DIFFERENCE)

Uno dei metodi più diffusi per esaminare la validità di costrutto è quello di confrontare il risultato di un test tra gruppi che si suppone differire nel costrutto in questione. Un esempio tipico, riportato anche nei libri di testo (Thomas, Nelson 2001), è quello di un ipotetico nuovo test per misurare la capacità anaerobica. Partendo dall'assunto che i velocisti siano più anaerobici dei mezzofondisti, posso confrontare i risultati nel test dei velocisti con i mezzofondisti. Se il risultato nel test è più alto negli *sprinter* posso dedurre che effettivamente il test misuri le capacità anaerobiche. Tornando al calcio, è tipico confrontare gruppi di livelli competitivi diversi per verificare se la capacità fisica valutata sia un prerequisito per poter competere a più alto livello. Ad esempio, posso effettuare un test su calciatori di Serie A e confrontare il risultato con quello di test effettuati su giocatori di categorie inferiori come ad esempio la Serie B o la Serie C. Se trovo differenze nel test posso ipotizzare che la caratteristica valutata è importante per poter giocare ad alto livello. Questo metodo di validazione può però risultare influenzabile da fattori non controllabili. Ricerche condotte con questo metodo (confronto tra gruppi) vengono chiamati *ex post facto*, termine che indica la condizione in cui si confrontano gruppi per i quali gli avvenimenti che hanno determinato l'appartenenza sono già accaduti, e non sono controllabili (si dice che c'è poco controllo sulla variabile indipendente) (Ary et al. 2006). In pratica quando si mettono a confronto giocatori di Serie A con giocatori di Serie B non sono noti quali siano gli avvenimenti che hanno portato i giocatori a militare in quelle categorie, e pertanto non risulta possibile definire quali fattori possano avere influenzato le caratteristiche del gruppo che si sta studiando. Negli studi in cui si confrontano gruppi (definiti studi *cross-sectional*) ci sono altri fattori che possono influenzare il risultato come le abitudini e la quantità dell'allenamento. Prendiamo ad esempio il caso nel quale si voglia confrontare il risultato dei test tra una squadra di Serie A e una di Serie C. Mettiamo in questo caso che il preparatore atletico della squadra di serie A essendo convinto che i lavori di forza siano importanti ponga particolare enfasi allo sviluppo di questa caratteristica dedicandovi quindi molto tempo. Diversamente Il preparatore della squadra di Serie C essendo invece convinto che siano più importanti gli allenamenti aerobici svolgerà una grande mole di allenamento per lo sviluppo di questa caratteristica. Date le premesse risulta assai probabile che nel caso vengano effettuati dei test di forza e per la *fitness* aerobica sui giocatori delle due squadre i giocatori di Serie A risultino avere livelli di forza più elevata con livelli possibilmente inferiori o uguali nel comparto aerobico. Seguendo il paradigma della validazione di costrutto secondo la tecnica della differenza tra gruppi, potrei concludere che, per giocare ad alto livello nel calcio sia importante avere alti livelli di forza. In realtà la differenza riflette semplicemente differenti abitudini di allenamento e diverse preferenze dei rispettivi preparatori atletici. Lo stesso avviene se confronto amatori con professionisti, in quanto eventuali differenze nei test possono semplicemente riflettere il fatto che i professionisti si allenano il doppio rispetto agli amatori e non a differenze nei prerequisiti fisici necessari per eccellere nello sport. Un altro esempio di verifica della validità di costrutto con questa tecnica è il confronto tra giocatori con ruoli differenti. L'assunto di questa tipologia di confronto risiede nel fatto che, data la documentata qualità e quantità di attività fisica svolta dai giocatori in partita, sia possibile che questi siano anche caratterizzati da diverse capacità fisiche. In un recente studio questo ragionamento è stato applicato per determinare la validità del test a navetta con l'obiettivo di verificare se questo riflettesse l'abilità di svolgere attività ad alta intensità in partita (Impellizzeri et al. 2008a). Dato che i difensori centrali risultano tra coloro che corrono meno ad alta intensità nel corso di una partita, si sono confrontati i risultati nel test dei difensori centrali con quelli rilevati nei giocatori di altri ruoli. Come ipotizzato questi giocatori hanno mostrato valori nel test più bassi rispetto agli altri ruoli. Lo stesso metodo (differenza nota tra gruppi) può essere applicato non a un test, ma ad indicatori di costrutto che poi a loro volta vengono utilizzati come criterio per validare i test. Ad esempio, per verificare se la distanza percorsa ad alta intensità in partita fosse un parametro di *performance* fisica valido, il gruppo di ricerca danese ha confrontato la distanza corsa ad alta intensità in partita da giocatori di Serie A italiani con la prima divisione danese (di livello certamente più basso rispetto agli italiani) trovando che i nostri giocatori corrono di più ad alta intensità (Mohr et al. 2003). Da questo si è concluso che l'alta intensità in partita è importante per poter giocare ad alto livello. Questo ha costituito il modello teorico su cui si sono successivamente basate le validazioni dei test sviluppati per dare indicazioni sull'abilità dei giocatori di correre ad alta intensità. Come abbiamo detto, tuttavia, questo metodo (differenza nota tra gruppi) è influenzabile da molti fattori e di conseguenza non è, o non dovrebbe essere, l'unico modo per validare i test.

EVIDENZE CONVERGENTI (CONVERGENT EVIDENCE)

Tra i vari metodi per fornire evidenze di validità il più usato e metodologicamente più forte è la *validità convergente*, attraverso la quale si va alla ricerca di una relazione tra il test e l'indicatore del costrutto in questione. Nell'ambito del modello teorico del calcio abbiamo detto che si assume che la *performance* fisica sia importante per il determinismo competitivo. In questo contesto l'attività ad alta intensità viene considerata un indicatore causale di *performance* fisica dato che tra questo e l'impegno fisico di gioco esiste una proporzionalità diretta. L'attività ad alta intensità, inoltre, sembra discriminare i giocatori di livello competitivo più alto (italiani vs danesi) rendendo questo parametro sia un valido indicatore del costrutto *performance fisica della partita*, sia una variabile rilevante per la performance calcistica. Quindi se voglio validare un test sviluppato per misurare o riflettere la capacità del giocatore di svolgere alta intensità in partita, dovrò utilizzare come criterio di validazione l'alta intensità misurata in partita o in simulazioni, qualora siano validate. Per determinare l'esistenza di evidenze convergenti di validità vengono calcolate le cosiddette correlazioni (Pearson o Spearman). Ad esempio, per validare sia il test *Yo Yo* sia il test a navetta sono state esaminate le correlazioni tra i risultati dei test e la distanza coperta ad alta intensità misurata in partita. Questi due test, infatti, sono stati sviluppati (o validati) come indicatori della capacità del giocatore di correre ad alta intensità. Le correlazioni significative, e superiori a 0,60 per il test a navetta e superiori a 0,70 per lo *Yo Yo* test hanno fornito evidenze convergenti sulla loro validità di costrutto (Bangsbø et al. 2008; Krustrup et al. 2003; Rampinini et al. 2007a). In pratica questi studi ci dicono che i due test sono validi indicatori dell'abilità del calciatori di svolgere attività ad alta intensità in partita, e che i meccanismi fisiologici coinvolti durante queste fasi della partita sono coinvolti in qualche misura anche durante l'esecuzione del test. In letteratura, ci sono molti studi che dimostrano che le soglie lattacide (o meglio qualsiasi punto della curva del lattato) sono correlate con la *performance* di *endurance* (dai 5000 m alla maratona) (Tokmakidis et al. 1998). A differenza di quanto sopra, in questo caso si parla di validità di criterio perché il confronto avviene tra il risultato del test e un criterio di riferimento (*gold standard*) e non con un indicatore del costrutto. Il riferimento in questione in questo caso è la *performance* stessa: tempo per correre una distanza. Purtroppo negli sport di squadra la *performance* non è così facilmente quantificabile, ed è per questo che si parla di costrutto e di conseguenza si applicano diversi metodi di validazione (o diverse definizioni). Questi metodi di validazione sono comunque un passaggio obbligato e vi sono molte altre tecniche di validazione più o meno appropriate secondo lo scopo del test (ad esempio la validità predittiva viene utilizzata quando un test viene sviluppato per predire un'altra misura). Inoltre, i vari tipi di validità spesso si sovrappongono e non sempre è possibile differenziarle. Tuttavia, anche se verificate e i risultati sono soddisfacenti, queste evidenze di validità risultano condizioni necessarie ma non sufficienti per validare un test. Come vedremo occorre verificare altri attributi come la ripetibilità e la responsività.

Ripetibilità

Un altro attributo importante di un test è la ripetibilità. Questo attributo dei test è molto complesso e ci vorrebbero interi articoli dedicati per affrontarlo in modo completo. Ricordiamo tuttavia che un test ripetibile non è necessariamente valido, ma un test non ripetibile non può essere valido. La riproducibilità nelle scienze dello sport viene classificata in due categorie (Atkinson, Nevill 1998): ripetibilità *relativa* e *assoluta*. In questo articolo per semplicità faremo riferiremo solo a quella assoluta (chiamata anche *agreement*) la quale risulta appropriata per test utilizzati in controlli longitudinali (cioè nel tempo). Quindi, anche se non specificato, il termine *ripetibilità* nel resto dell'articolo farà riferimento alla ripetibilità assoluta. La ripetibilità indica la consistenza di un test, cioè la sua abilità nel dare risultati simili quando il test viene ripetuto nelle stesse condizioni e sullo stesso soggetto. La ripetibilità deve essere considerata come strumento per valutare il *rumore* della misura. Questo rumore risulta in pratica determinato da variazioni casuali e/o sistematiche prodotte da fattori intrinseci ed estrinseci al test. Ci sono diversi metodi statistici per calcolare la ripetibilità. I più utilizzati e appropriati sono l'*Errore Standard della Misura* (*ESM*) ed i *limiti di confidenza* (95%) di Bland e Altman (Atkinson, Nevill 2000; Atkinson, Nevill 1998; Hopkins 2000). Ricordiamo che la cosiddetta e popolare correlazione *test-retest* è ormai abbandonata e sconsigliata da praticamente tutti gli statistici.

Parleremo qui di seguito solo dell'ESM. L'ESM andrebbe calcolato da un altro indice di ripetibilità chiamato *Intraclass Correlation Coefficient* (*ICC*). Tuttavia, per semplicità spieghiamo un altro metodo di calcolo dell'ESM, tecnicamente non ottimale, ma che fornisce un'indicazione grezza, ma utile della ripetibilità del test, il quale può essere calcolato facilmente usando *Microsoft Excel*. Assumiamo di avere dei giocatori, e di sottoporli per due volte allo stesso test a distanza di due giorni, avendo cura di effettuarli alla stessa ora del giorno e con i giocatori non affaticati da allenamenti intensi svolti nei giorni precedenti i test (per evitare che la fatica residua influisca sui risultati).

La tabella 2 mostra degli ipotetici risultati per cinque soggetti che effettuano due volte un test di salto: test 1 e 2. In una colonna calcoliamo la differenza tra il test 2 e 1 (test 2 – test 1). Si calcola poi la deviazione standard delle differenze e la si divide per la radice quadrata di 2 (ESM=SD/2). Dall'esempio in tabella 2 risulta che l'ESM è di 0,6 cm. Per comodità si esprime di solito la ripetibilità in percentuale. Nel nostro caso la media totale dei salti (test 1 e test 2) è 44,8 cm. L'ESM diventa quindi 1,3% (0,6/44,8 x 100). In genere con questa formula la ripetibilità risulta leggermente migliore di quella reale, perché la formula non tiene conto di un eventuale errore sistematico. In ogni caso per vedere se c'è un errore sistematico è sufficiente applicare il t-test. Anche il t-test è contenuto nelle funzioni di *Excel* o in altri *software* per computer *Mac*, come *Numbers*. Sempre nel nostro esempio le probabilità che la differenza sia casuale e non reale è di circa 50%; quindi non c'è errore sistematico.

La presenza di un errore sistematico costituisce un problema e una volta che se ne verifica la presenza (vedi sopra) è bene capirne l'origine, così da ripetere lo studio controllando i fattori che hanno potenzialmente causato l'errore sistematico. I motivi più frequenti sono l'*errata calibrazione* o *taratura degli strumenti*, la *presenza di fatica temporanea*, ma soprattutto l'*effetto familiarizzazione* (o effetto apprendimento). Quest'ultimo si verifica soprattutto quando i soggetti non conoscono il test e alla seconda esecuzione migliorano la *performance* attraverso l'ottimizzazione del gesto tecnico. Per capire quanto sia pericoloso un errore sistematico provate a pensare di aver effettuato un test di ingresso, un allenamento che in realtà è inefficace a cui fa seguito la somministrazione di un test di uscita, nel quale i soggetti sono migliorati semplicemente per l'effetto apprendimento e non perché siano migliorate realmente le loro capacità fisiche. In un caso simile l'allenatore concluderebbe erroneamente che il suo allenamento è efficace.

	A	B	C	D	
1		**Test 1**	**Test 2**	**Diff. T2 – T1**	< Formula in Excel
2	Soggetto 1	40,5	41,5	1,0	=C2-B2
3	Soggetto 2	45,6	44,5	-1,1	=C3-B3
4	Soggetto 3	38,2	39,1	0,9	=C4-B4
5	Soggetto 4	50,4	50,6	0,2	=C5-B5
6	Soggetto 5	48,5	48,9	0,4	=C6-B6
7	Media	44,6	44,9	0,3	=MEDIA(E2:E7)
8	DS	5,2	4,8	0,8	=DEV.ST(E2:E7)
9			ESM=	0,6	=D8/RADQ(2)
		Sistematicità errore (t test)		0,49	=TEST.T(B2:B6,C2:C6,2,1)

NB: Ai fini esemplificativi abbiamo presentato solo cinque soggetti. Per verificare la ripetibilità occorrono da venti a trenta soggetti. Tra dieci e venti i risultati sono meno generalizzabili. Sotto i dieci risultati non sono attendibili.

Tabella 2 – Esempio di come calcolare la ripetibilità (Errore standard della misura: ESM) utilizzando le formule di Microsoft Excel su due ipotetiche sessioni di test di salto (cm)

Dalla ripetibilità si può calcolare qual è il cambiamento minimo a livello individuale che può essere interpretato come reale e non dovuto all'errore della misura. In genere un cambiamento individuale pari all'ESM indica che la probabilità che il test sia cambiato è di circa l'80%, cioè solo una possibilità su cinque che il cambiamento sia dovuto all'errore della misura. La probabilità di cambiamento dovrebbe essere superiore all'80% (ideale 95%) prima di interpretare con una certa confidenza che il test sia migliorato realmente. Al di sotto del 75% il rischio di errore è solitamente considerato non accettabile.

Detto questo, qual è il valore di ripetibilità accettabile? Per quanto si tenda a ritenere che la ripetibilità sia accettabile quando l'ESM ha valori bassi (in genere inferiore a 5-10%), questo approccio è pericoloso e statisticamente non corretto. La ripetibilità è accettabile solo in base alla sua responsività (*responsiveness*) o alla sensitività del test ai cambiamenti.

Responsività (Responsiveness)

Se si interpreta la ripetibilità come il rumore di una misura, questo rumore sarà accettabile solo se più basso del segnale, dove il segnale è il cambiamento che un test ha in conseguenza ad un intervento (nutrizionale, allenamento, etc.). Per fare un esempio supponiamo di aver inventato un nuovo test di agilità la cui ripetibilità è del 3%. A prima vista la ripetibilità sembrerebbe ottima, ma ipotizziamo di aver verificato che un allenamento di due mesi per lo sviluppo dell'agilità determini miglioramenti del test pari al

2%. Dato che in questo caso il rumore (ESM) risulta superiore al segnale (2%), si dice che il test è troppo rumoroso e quindi poco sensibile ai cambiamenti a livello individuale. Questa proprietà del test è chiamata *responsività interna*, ovvero l'abilità del test di rilevare cambiamenti. Semplificando si può indicare la responsività interna come il rapporto *rumore : segnale*. Questo modo di interpretare la ripetibilità di un test è molto importante. Per anni si è ritenuto che i test ad esaurimento non fossero abbastanza ripetibili da poter essere utilizzati, ad esempio, nel ciclismo. Infatti la loro ripetibilità risulta oscillare tra il 20-25% per test di lunga durata (vicini o superiori all'ora). Al contrario si riteneva che i cosiddetti *time trial* (prove a tempo o distanza fissa) fossero più appropriati avendo una ripetibilità inferiore al 3-5%.

Tuttavia è stato dimostrato che i cambiamenti del test di esaurimento a seguito di intervento sono superiori al 30% mentre i *time trial* cambiano solo del 3-5%. Essendo il rapporto *rumore : segnale* simile, si evince che entrambe i test risultano utilizzabili e che, al contrario di quanto prima ritenuto, il 20-25% di ripetibilità rilevato nel test ad esaurimento risulta quindi accettabile. Ovviamente tanto più un test è ripetibile tanto più sarà in grado di rilevare piccoli cambiamenti. Tuttavia è necessario anche che il test sia in grado di cambiare in conseguenza ad un intervento, cioè è necessario che sia *sensibile*. Ricordiamo che quanto detto finora si applica all'interpretazione dei risultati dei test a livello *individuale*, e non sul gruppo. In quest'ultimo caso il fatto che il cambiamento sia reale o no è facilmente calcolabile con i tradizionali metodi statistici. Un altro elemento importante, molto trascurato nella validazione dei test, è la cosiddetta *responsività esterna*, anche chiamata *validità longitudinale*. Questo attributo ci indica l'abilità di un test di riflettere cambiamenti nel costrutto o nel criterio di riferimento. Se un preparatore rileva un miglioramento nello *Yo Yo Test* sui suoi giocatori, ma questo cambiamento non riflette un potenziale cambiamento nell'abilità del giocatore di svolgere alta intensità in partita, quale utilità avrebbe il test? Probabilmente poca. È tuttavia sorprendente constatare in quanti pochi test sia stata verificata la responsività esterna. I test vengono di solito usati per verificare l'effetto dell'allenamento sulle determinanti della prestazione. In pratica si misurano i cambiamenti in quei fattori fisiologici che si suppone possano influenzare la prestazione. Ci si aspetta quindi che i miglioramenti delle caratteristiche misurate con i test influenzino in modo positivo la prestazione: in altre parole ci si aspetta che la prestazione migliori. Dato che è questo ciò che giustamente si aspettano i preparatori ed allenatori, ne consegue che questa abilità del test debba venire verificata e che questo attributo sia essenziale nel determinare la validità di un test per controlli nel tempo.

Interpretabilità

Per poter utilizzare un test occorre essere in grado di interpretarne i risultati. Come si può già intuire per interpretare correttamente i risultati di un test occorre che gli attributi finora esposti siano stati esaminati. Il primo elemento da considerare è la ripetibilità. Questo ci permette di capire se i cambiamento nel test che rileviamo sul nostro atleta siano dovuti all'errore o che ci sia una accettabile probabilità che il cambiamento sia reale. In una revisione della letteratura scritta dal famoso statistico Will Hopkins (Hopkins et al. 2001) sono presentati i dati di ripetibilità dei test più comuni utilizzati nello sport, e per questo ne suggeriamo la lettura. Questi dati di ripetibilità possono inoltre essere confrontati con i risultati degli studi in cui i test sono stati usati per verificare il risultato di particolari interventi. Se il valore di ripetibilità è inferiore al cambiamento rilevabile a seguito di un allenamento, o qualsiasi altro intervento (cioè *rumore < segnale*), il test possiede abbastanza sensitività per poter essere utilizzato. Per il test a navetta, ad esempio, i cambiamenti che si rilevano sulla squadra, per quello che riguarda il

tempo medio impiegato per correre i 6 sprint, è dell'ordine del 3% a fronte di una ripetibilità dell'1% (Bravo et al. 2007; Impellizzeri et al. 2008a).
Data la sua portata applicativa vale la pena di introdurre il concetto di *cambiamento minimo importante* (o cambiamento significativo da un punto di vista pratico). Questo si può tradurre con la domanda: qual è il cambiamento minimo che può essere considerato significativo non da un punto di vista statistico, ma pratico/fisiologico?
Nell'ambito dello sport il modo migliore per determinare questo cambiamento minimo sarebbe attraverso la responsività esterna, cioè verificando qual è il cambiamento nel test che si traduce in un'accettabile probabilità di un cambiamento del costrutto o misura di riferimento. In altre parole, quale deve essere il cambiamento minimo nello *Yo Yo* o nel test a navetta perché questo si traduca in un cambiamento dell'abilità di effettuare attività ad alta intensità? Non sempre questo è calcolabile in relazione alla prestazione (anche se è il metodo migliore) e di conseguenza Will Hopkins ha proposto un altro metodo, basato sulla probabilità che un cambiamento determini una variazione all'interno del gruppo (Hopkins et al. 1999). In altre parole, è il cambiamento minimo che consente a un atleta di diventare migliore o peggiore rispetto ad un altro. Per quanto riguarda il test a navetta questo cambiamento è dello 0,5% (Impellizzeri et al. 2008a). Essendo la ripetibilità intorno all'1% ne consegue che il test è poco sensibile a cambiamenti individuali piccoli, ma importanti. Per far ulteriormente capire quanto sia importante conoscere la ripetibilità di un test, e i cambiamenti che questo deve essere in grado di identificare facciamo ancora riferimento al test a navetta per il quale è risultato che il decremento nel tempo di percorrenza degli sprint è il parametro meno ripetibile (circa 30%) e meno sensibile. Pur essendo il decremento il parametro più utilizzato dai preparatori nell'analisi dei risultati del test a navetta, questo parametro non dovrebbe essere utilizzato, o come minimo dovrebbe essere interpretato con cautela. La comunità scientifica internazionale ha più volte confermato la poca ripetibilità del decremento della prestazione di *sprint* (chiamato anche impropriamente *indice di fatica*) consigliando di conseguenza l'utilizzo di altri parametri per la valutazione dell'abilità di ripetere sprint dei soggetti (Oliver 2009).
Per consentire una migliore interpretabilità, infine, è utile avere dati di riferimento così da conoscere come il nostro giocatore si posiziona all'interno della popolazione.

Fattibilità

Un test deve essere fattibile e le risorse, così come l'impegno richiesto, devono essere commisurate all'importanza delle informazioni che ne derivano. In letteratura esiste un test (LIST) che simula 90 min di gioco prevedendo, nella sua parte finale, una fase che fornisce informazioni circa la capacità del giocatore di effettuare sforzi intermittenti fino ad esaurimento (Nicholas et al. 2000). Questo test è stato sviluppato per effettuare simulazioni di gioco a scopo di studio, ma è altresì vero che nessuno ne vieterebbe il suo uso per la valutazione funzionale di un giocatore. Tuttavia, è intuitivo capire che un test simile che richiede ai giocatori (uno alla volta) di effettuare 90 min di simulazione è di difficile realizzazione. Un test simile è irrealizzabile in una situazione reale e, quindi, non è proponibile per la valutazione di *routine* anche se fosse il più valido test a disposizione. Uno dei motivi di successo del test Mognoni, ad esempio, è la sua relativa semplicità e il fatto che non è richiesto uno sforzo massimale rendendolo ben accetto dai giocatori (Impellizzeri et al. 2004a; Impellizzeri et al. 2005; Sirtori et al. 1993). Con una buona organizzazione in un'ora si possono valutare almeno venti giocatori.

Step	Attributo	Yo Yo Intermittent Recovery Test	Test a Navetta di Capanna
1	Modello teorico	La performance fisica è importante per la prestazione calcistica. L'attività ad alta intensità è il miglior indicatore di performance fisica in partita	La performance fisica è importante per la prestazione calcistica. L'attività ad alta intensità è il miglior indicatore di performance fisica in partita. L'attività ad alta intensità contiene la distanza percorsa durante le fasi di sprint ripetuti
2	Validità	Il risultato dello Yo Yo test è correlato con l'attività ad alta intensità (r>0,70), è differente a secondo dei ruoli, è maggiore nei giocatori di più alto livello, varia maggiormente durante il campionato	Il risultato nel test navetta è correlato con la distanza percorsa ad alta intensità (0,60<r<0,65) e differenzia tra ruoli. Distingue tra professionisti ed amatori, ma non tra professionisti di livello competitivo differente (ad esempio Serie A o *Premier League* rispetto a Serie C)
3	Ripetibilità	La ripetibilità (coefficiente di variazione) è tra il 5 e 8%	La ripetibilità (errore standard della misura) del tempo medio impiegato per effettuare i 6 sprint è di 0,8-0,9%, mentre per il miglior sprint è 0,9-1,2%. La ripetibilità del decremento è di circa 30%
4	Responsività	Responsività interna: i cambiamenti a seguito di intervento (allenamento) vanno in genere dal 12 al 50%, indicando un rapporto rumore: segnale di 1:2 ad 1:5. *Responsività esterna:* non calcolata sui calciatori. Negli arbitri la correlazione tra i cambiamenti nel test ed i cambiamenti di alta intensità misurata in partita è di r=0,77	Responsività interna: i cambiamenti in pre-campionato e a seguito di specifici allenamenti sono di circa il 3%, indicando un rapporto rumore : segnale di 1:3. *Responsività esterna:* non calcolata.
5	Interpretabilità	Poca sensibilità a cambiamenti piccoli, ma importanti (1%) a livello individuale. Dati di riferimento (Bangsbo, Sport Medicine, 2008)	Poca sensibilità a cambiamenti piccoli, ma importanti (0,5%) a livello individuale. Non ci sono dati di riferimento (i risultati sono molto dipendenti dalla modalità di misura (manuale, fotocellule) e superficie del test

Tabella 3 – Attributi che devono essere verificati durante il processo di validazione di un test ed esempio di verifica per i test Yo Yo e a navetta

Anche lo *Yo-Yo Intermittent Recovery Test* possiede una buona fattibilità. Infatti con esso è possibile valutare contemporaneamente molti giocatori risparmiando così una considerevole quantità di tempo.

Stato dell'arte sui test nel calcio

Nel calcio non ci sono test sui quali siano stati verificati tutti gli attributi e lo stesso vale per il modello teorico che non è ancora ben sviluppato e validato. Solo due test presentano qualche evidenza di validità (tabella 3) di cui solo uno forte: il *test a navetta di Capanna* e lo *Yo Yo Intermittent Recovery Test* di Bangsbø (Bangsbø et al. 2008; Castagna et al. 2006; Impellizzeri et al. 2008a; Krustrup et al. 2003; Krustrup et al. 2006; Rampinini et al. 2007a). Il test a navetta di Capanna ha qualche evidenza di validità essendo correlato con l'attività ad alta intensità e la distanza coperta sprintando in partita. Inoltre i suoi valori si sono dimostrati ruolo dipendenti e, quindi, in grado di riflettere le diverse richieste fisiche di ciascun comparto di gioco. La responsività interna del test di Capanna è adeguata a livello di gruppo, tuttavia a livello individuale, in relazione a cambiamenti importanti (che sono dell'ordine del 0,5% per il tempo medio) questa risulta moderata. Questo sta ad indicare che il test di Capanna risulta poco sensibile a piccoli, ma pur importanti cambiamenti a livello individuale. Purtroppo la responsitività esterna del test di Capanna non è stata mai esaminata. Quindi questo test a navetta presenta qualche evidenza di validità, ma andrebbe usato ed interpretato con cautela fino a che non se ne esaminino tutte le proprietà.

Lo *Yo-Yo Intermittent Recovery Test* di Bangsbø è il test che, piaccia o no, sicuramente presenta le maggiori evidenze di validità in quanto risulta correlato all'attività svolta ad alta intensità in partita, la prestazione è differente secondo i ruoli, possiede una ottima ripetibilità e una buona responsività interna (segnale > 3 volte il rumore), ed è ben interpretabile dato che ormai esistono molti dati in letteratura essendo il test più utilizzato al mondo. Anche per lo *Yo-Yo Intermittent Recovery Test*, però, non è stata verificata la responsività esterna nei giocatori. Tuttavia questa caratteristica è stata calcolata negli arbitri mostrando buone correlazioni (r = 0,77) tra i cambiamenti nel test e le variazioni nell'alta intensità effettuata in partita dopo allenamento intermittente (Bangsbø et al. 2008). Questo fa ben sperare in una possibile e auspicabile verifica sui giocatori. Ma se i cambiamenti nel test non dovessero risultare correlati con i cambiamenti dell'abilità dei giocatori di correre ad alta intensità la sua validità ne risulterebbe automaticamente compromessa. Come per il test a navetta la sensibilità a cambiamenti piccoli ma importanti, sembra essere scarsa, dato che il cambiamento minimo importante calcolato con il metodo di Hopkins (Hopkins et al. 1999) e usando i dati presentati in una recente *review* da Bangsbø et al. (2008) è di circa 1%. Quindi ben al di sotto della ripetibilità del test (5-8%).

Alla luce di quanto fin qui illustrato il quadro potrebbe sembrare troppo disfattista, ma di fatto rispondente alla realtà delle cose. Purtroppo si assiste nel calcio a un inconsapevole proliferare di test e nessuno di questi test risulta validato né su riviste scientifiche internazionali né su riviste nazionali, anche divulgative. Nel migliore dei casi si assiste alla proposta di test la cui validazione si ferma molto spesso alle sole evidenze logiche (*face validity*) e solamente in qualche raro caso si spinge fino all'analisi delle differenze tra gruppi di livello competitivo diverso. Per evitare le problematiche sopra esposte sarebbe di sicura efficacia rinunciare all'introduzione di nuovi test per meglio concentrarsi sulla determinazione della validità dei test già esistenti. E quindi sviluppare nuovi test solo qualora ve ne sia veramente la necessità. Nello sviluppo di nuovi test è comunque necessario seguire i criteri scientifici brevemente descritti in questo articolo.

Conclusione

In questo articolo abbiamo tentato di fornire i principi che stanno alla base dello sviluppo e della validazione dei test. Queste basi teoriche non si applicano solo ai test propriamente detti, ma in generale alle misure come, ad esempio, nella quantificazione della percezione dello sforzo per la determinazione del carico di allenamento.

Nello studio che ha proposto questo metodo nel calcio, che risale a cinque anni fa, la cosiddetta session-RPE è stata validata verificando la validità di costrutto convergente, cioè esaminando le correlazioni tra metodi di quantificazione del carico di allenamento basati su frequenza cardiaca e quello basato sulla scala di Borg 0-10 (Impellizzeri et al. 2004b). La frequenza cardiaca fu usata come indicatore di intensità della sessione di allenamento, dove quest'ultima costituiva il costrutto che si voleva misurare. In seguito è stata fatta la stessa cosa utilizzando la combinazione frequenza cardiaca e lattato ematico come indicatori del costrutto (Coutts et al. 2007). Questo metodo si sta oggi diffondendo, ma sembra che spesso venga applicato male, ad esempio sostituendo o modificando le scale di percezione rispetto a quanto proposto (Impellizzeri et al. 2004b). Anche in questo caso, il metodo va applicato così come è stato validato e non apportando variazioni personali. A questo argomento verrà probabilmente dedicato un articolo a parte data la diffusione che sta avendo l'uso della percezione dello sforzo, purtroppo concomitante ad una superficiale applicazione del metodo, ad una scarsa conoscenza di come si usano le scale di percezione, e alla luce degli studi ancora in corso che potrebbero modificare la metodologia.

La complessità dell'argomento (validazione dei test e delle misure) è tale che non è possibile esaurire in modo soddisfacente tutti gli aspetti in un solo articolo. Ad esempio non abbiamo affrontato il tema dei test utilizzati per ricavare i ritmi di allenamento, invece che per monitorare i cambiamenti. Anche in questo caso, ovviamente, il test andrebbe validato per questa finalità specifica. Come abbiamo accennato, infatti, un test può essere valido per uno scopo e non per un altro. Non abbiamo neanche discusso la validità dei test predittivi come il test di Leger, erroneamente utilizzato per stimare il massimo consumo di ossigeno a livello individuale (la stima è appena accettabile a livello di gruppo ma in nessun modo a livello individuale). Il messaggio di questo articolo è, quindi, di cautela sia nello sviluppo sia nell'utilizzo di test di cui non vengono fornite le prove di validità. Il lettore si sarà anche reso conto che l'ambito della valutazione funzionale e dei test è complesso, ma regolamentato da rigidi percorsi scientifici. Questo non significa che gli altri popolari test da campo, qui non citati, non siano validi in assoluto, ma semplicemente che per lo meno la loro validità non è stata ancora dimostrata. Fino a prova contraria, per quanto autorevole, non ci si può "fidare" semplicemente della parola dell'ideatore. Come diceva il famoso William Edwards Deming: "in Dio crediamo, tutti gli altri ci mostrino i dati!"

La bibliografia del presente articolo può essere consultata sul sito www.calzetti-mariucci.it e sul sito della Scuola dello Sport http://scuoladellosport.coni.it

Indirizzo dell'Autore:
franco.impellizzeri@kws.ch