DiPAV, 2007, 20, 25-40

INDICATORI E COSTRUTTI IN PSICOMETRIA: VALIDITÀ DEI TEST

Riccardo Sartori *
Margherita Pasini **

Sommario

Molti autori hanno offerto il loro contributo all'argomento dibattuto della validità dei test psicologici. Il concetto di validità, ovvero il grado con cui uno strumento misura ciò che deve misurare, sembra essere semplice fintanto che non si consulta la letteratura sull'argomento. Questo contributo vuol essere una rassegna critica utile agli psicologi, sia orientati alla teoria che alla pratica.

Abstract

Many authors have offered their contributions to the controversial subject of test validity. Thus, validity, which can be defined as "the degree to which the test actually measures what it purports to measure", seems to be quite a simple idea until one looks at the literature on the subject. This paper is thought as a critical review which can be useful both to theoretically oriented and practically inclined psychologists.

1. Introduzione

Possiamo definire la psicometria come quella disciplina che, all'interno della psicologia, si occupa dei metodi utilizzati per misurare, con opportune trasformazioni quantitative, le differenze individuali nelle reazioni psicologiche di soggetti diversi o di uno stesso soggetto in momenti diversi.

La psicometria, quindi, rappresenta il campo d'indagine delle teorie e delle tecniche messe a punto per misurare in psicologia. Ma cosa si misura in

* Ricercatore in Psicometria, Dipartimento di Psicologia e Antropologia Culturale. E-mail: riccardo.sartori@univr.it.
** Professore Associato in Psicometria, Dipartimento di Psicologia e Antropologia Culturale. E-mail: margherita.pasini@univr.it.

psicologia? La risposta più generale, che si può dare a questa domanda, è che in psicologia si misurano caratteristiche psicologiche, come percezioni e tempi di reazione, o *costrutti*, come ad esempio conoscenze, abilità, atteggiamenti, tratti, ecc. Secondo la classica definizione fornita da Crocker e Algina (1986), il costrutto può essere definito come il prodotto di una fondata riflessione scientifica, un'idea sviluppata per permettere la categorizzazione e la descrizione di alcuni comportamenti direttamente osservabili. I costrutti sono, per definizione, non accessibili all'osservazione diretta, ma vengono inferiti o postulati sulla base dell'osservazione di opportuni indicatori (cfr. Sartori, 2005).

La psicometria riguarda due aspetti importanti della ricerca psicologica, vale a dire: sviluppo e perfezionamento dei modelli teorici per la misura; messa a punto di procedure e costruzione di strumenti atti a misurare i costrutti di interesse. Consapevoli, quindi, che la psicometria è un settore di indagine che non si esaurisce con lo studio del funzionamento dei test psicologici, il presente lavoro intende porsi come una rassegna critica su come i vari autori intendano la validità dei test e sulle differenti proposte avanzate per accertarla.

2. La validità

Secondo alcuni autori (cfr. Borsboom, Mellenbergh & van Heerden, 2004), quello di validità è un concetto semplice. Esso si riferisce alla questione se un test misura ciò che intende misurare. Vernon (1963) evidenzia che un test può essere valido solo per alcuni scopi. Kline (1998) ribatte che, pur apprezzando la puntualizzazione di Vernon, un test psicometrico veramente scientifico dovrebbe essere valido di per sé, ovverosia per tutti gli scopi per cui un test può essere legittimamente usato. Ad ogni modo, l'affermazione di Vernon solleva una questione interessante: mentre infatti l'attendibilità di un test può essere facilmente misurata, ad esempio attraverso l'Alpha di Cronbach (almeno per quanto riguarda un aspetto dell'attendibilità: quello della coerenza interna), l'accertamento della validità non si esaurisce con il calcolo di indici statistici[1].

[1] Per attendibilità di un test si intende il grado con cui il test, a parità di condizioni di somministrazione, restituisce misure stabili nel tempo, ovverosia indipendenti dall'errore di misura.

Il problema della validità dei test si è evoluto lungo tutto il ventesimo secolo: dalla questione se un test misura ciò che intende misurare (Kelley, 1927; Cattell, 1946), alla questione se le relazioni empiriche tra punteggi riproducano alcune ben definite relazioni teoriche (Cronbach & Meehl, 1955), alla questione se le interpretazioni e le azioni basate sui test siano giustificate non solo scientificamente ma anche socialmente ed eticamente (Messick, 1980, 1993, 1998).

Un articolo pubblicato da Messick nel 1995 argomenta che la validità nel testing psicologico debba produttivamente essere concettualizzata come un costrutto unico. Egli infatti nota che (Messick, 1995, 747), "*The essence of unified validity is that the appropriateness, meaningfulness and usefulness of score-based inferences are inseparable […] both meaning and values are integral to the concept of validity, and psychologists need a way of addressing both concerns in validation practice*". Al momento, tuttavia, per quanto riguarda la validità dei test, il modello più usato e diffuso (sebbene criticato) è il cosiddetto *Trinitarian Model*, che prevede l'assessment combinato delle validità di contenuto, criterio e costrutto.

Come accennato, quindi, il cosiddetto *Unitary Concept of Validity* postula che la validità sia una caratteristica unitaria di un test. Dal momento però che esiste più di un modo attraverso il quale verificare la validità di un test, il cosiddetto *Trinitarian Model of Validity* sottolinea proprio questo aspetto. A questo proposito Ammassari (1984, 149) fa notare che "la validità è una sola" per cui, come scrive Giampaglia (1990, 51) "sarebbe più opportuno riferirsi non a tipi diversi di validità, bensì a procedure diverse dirette a provare, in modo diretto o indiretto, quella che Zetterberg chiama validità interna" (cfr. Sartori, 2004).

Sempre Giampaglia (1990, 51) fa però anche notare che "le varie procedure risultano così diverse l'una dall'altra [...] che finiscono nella realtà per ridefinire il concetto stesso di validità in funzione degli aspetti che privilegiano".

2. Tipi di validità

Angoff (1988) scrive che la formula tradizionale per la validità dei test, prima del 1950, era la seguente: un test è valido se misura ciò che si propone di misurare (e.g., Kelley, 1927, 14); mentre McDonald (1999, 197) scrive

che "il punteggio di un test è valido se riflette il costrutto dei rispondenti per cui il test viene impiegato".

Cook e Campbell sostengono che la validità di un metodo o di uno strumento è la migliore approssimazione disponibile alla verità. "Un gruppo di indicatori", scrive Giampaglia (1990, 49), "deve innanzitutto essere valido perché possa costituire un efficace strumento di rilevazione". Kerlinger (1974, 457), Nunnally (1978, 86), Carmines e Zeller (1979, 12) definiscono la validità di un indicatore "la misura in cui esso rileva ciò che è destinato a rilevare". Anastasi (1993, 196) sostiene che "la validità di un test concerne ciò che viene misurato dal test e con quale precisione esso riesce ad effettuare tale misurazione". Dello stesso parere sono autori italiani come Pedrabissi e Santinello (1997) e Sanavio e Sica (1999). A questo proposito Bailey (1991, 84) scrive: "la definizione di validità è composta di due parti, che si riferiscono al fatto i) che lo strumento di misura stia effettivamente misurando il concetto in questione, e non un qualche altro concetto; ii) che il concetto venga misurato accuratamente"[2]. A questi due aspetti della validità, Nunnally (1978) ne aggiunge un terzo che si riferisce all'utilità del metodo e dello strumento utilizzato. Egli infatti definisce la validità di un test nei termini, anche, di quanto esso è scientificamente utile.

Borsbomm, Mellenberg e van Heerden (2004) affermano che molte delle considerazioni relative alla validità dei test, fatte durante il ventesimo secolo, risultano irrilevanti, complicano un concetto che è sostanzialmente semplice e falliscono nell'intento di offrire una visione semplice, chiara e operativa della teoria legata alla validità dei test. In particolare, essi scrivono: "*Validity is not complex […] It is a very basic concept*" (Borsboom, Mellenbergh & van Heerden, 2004, 1061).

È del resto vero che, sebbene la validità possa essere considerata un concetto sostanzialmente basilare e unitario, i sistemi attraverso cui accertarla sono diversi. Essi finiscono col ridefinire il concetto stesso di validità. Generalmente parlando si assume che esistano almeno quattro tipi di validità (tre nel *Trinitarian Model*): validità di facciata, validità di contenuto, validità di criterio (concorrente e predittiva), validità di costrutto (convergente e divergente o discriminante). Le prime due vengono testate soprattutto attraverso metodi qualitativi (anche se alcuni autori considerano la validità di contenuto

[2] Nella definizione di validità fornita da alcuni autori rientrerebbero anche i concetti di accuratezza, affidabilità e attendibilità (cfr. Bailey, 1991; Dipboye, Smith, & Howell, 1994). Altri autori, invece, sostengono la necessità di mantenere separati concetti come quelli di validità, attendibilità e sensibilità (cfr. Anastasi, 1993).

come un giudizio quantitativo sui test), le ultime due soprattutto attraverso metodi quantitativi. Diventa in questo modo possibile dividere questi quattro tipi in due gruppi ipotetici: tipi qualitativi di validità e tipi quantitativi di validità (Sartori & Pasini, 2007).

2.1 Tipi qualitativi di validità: validità di facciata e validità di contenuto

Validità di facciata. Un test possiede validità di facciata quando appare valido a personale non esperto (Anastasi, 1993). Cronbach (1984, 182) scrive: "un test che sembra rilevante alla persona non esperta viene detto possedere validità di facciata". La validità di facciata non discende da modelli teorici (Fink, 1995), il che l'ha resa a volte sospetta agli occhi dei ricercatori. Lo stesso Cronbach (1971) si dice convinto che quello di validità di facciata sia un concetto vago e soggettivo. Ingram (1977, 18) si riferisce ad essa come *fiducia superficiale* o *credenza popolare*. Alderson, Clapham e Wall (1995, 172) la considerano olistica piuttosto che analitica.

Per queste ragioni, il concetto di validità di facciata è diventato nel tempo controverso. Alcuni autori sostengono che esso mantenga una sua utilità (e.g., Roberts, 2000), altri che sia persino pericoloso continuare a considerarlo (e.g., Newfields, 2002). Roberts (2000) dichiara che la validità di facciata è una parte essenziale del processo di valutazione di uno strumento per tre ragioni principali:

1. dal momento che oggi esiste una gran quantità di strumenti che potrebbero essere utilizzati per la misurazione di uno stesso costrutto, il professionista che intendesse avvalersi di uno solo di essi può restringere le opzioni di scelta attraverso l'esplorazione delle componenti dei test e scegliere lo strumento più adeguato allo specifico contesto di somministrazione;
2. l'esplorazione del test può anche fornire informazioni utili circa alcune precauzioni che è meglio prendere per applicarlo in modo tale da non influire negativamente sulla validità e da rendere l'intera somministrazione quanto più possibile ricca di significato;
3. in generale, una buona indagine sulla validità di facciata può fornire evidenze, nei casi in cui manchi il tempo di documentarsi e le decisioni debbano essere prese velocemente, circa l'appropriatezza o meno di avvalersi di quel test in quel contesto.

Nonostante questi aspetti positivi della validità di facciata, vi sono alcune voci di dissenso. Hajipournezhad (2003) menziona quanto l'espressione validità di facciata sia ampiamente detestata tra i ricercatori che si occupano di testing psicologico. Citando Mosier (1947, 194) egli sottolinea che "il concetto è tanto più pericoloso quanto più legittima coloro che, per mancanza di tempo, risorse o competenze, si sentono esonerati dal dimostrare la validità (o la non validità) di un test in qualsiasi altro modo […] Questa nozione è anche pericolosamente gratificante per i costruttori inesperti di test". Trochim (2002) avverte che la validità di facciata è il modo più debole di provare a dimostrare la validità di costrutto. Lacity e Jansen (1994) descrivono la validità di facciata nei termini di appeal persuasivo e fanno notare che gli item di un test possono sembrare convincenti anche se in realtà mancano di validità interna.

Su questo argomento, Newfields (2002) scrive che:

1. la validità di facciata è un termine contraddittorio. Questioni che coinvolgono la superficie possono avere un valore cosmetico, non di validità. La validità dovrebbe coinvolgere fattori più profondi e seguire la logica della veridicità, della consistenza e della congruenza;
2. se vogliamo considerare il *testing* come una disciplina rigorosa, la validità di facciata ha poco spazio perché è sia ateoretica sia generalmente imprecisa. Il concetto di *face validity* (validità di facciata) si riferisce, sostanzialmente, a ciò che Buck (2001) chiama *faith validity* (validità di fiducia, ovvero: la credenza che un test funziona, senza alcuna evidenza empirica). L'evidenza empirica è un presupposto findamentale del *testing*. Dal momento che la validità di facciata si basa primariamente sui giudizi dei non esperti, questo concetto può risultare forse interessante in un'ottica di marketing, ma non dovrebbe affatto essere preso in considerazione dagli sviluppatori di test.

Generalmente parlando, il concetto di validità di facciata presenta i seguenti problemi:

1. la validità di facciata è orientata al prodotto. Le persone deputate alla valutazione del test lo considerano solo *dopo* che esso è stato costruito. Esse non hanno alcuna possibilità di familiarizzare con gli assunti teorici sottostanti alla costruzione del test;
2. le persone che valutano il test tramite la sua esplorazione superficiale hanno la tendenza a dire che esso è valido;
3. un'altra tendenza è quella di valutare la validità di facciata per mezzo di operazioni qualitative, non quantitative. In altre parole, una valutazione

della validità di facciata tramite *rating scales* o interviste strutturate è solo raramente praticata;

4. la validità di facciata, così come viene normalmente praticata, implica un esame superficiale del contenuto del test. Questo non lascia spazio ad alcuna analisi in profondità;
5. la validità di facciata è un tipo di giudizio che non segue i principi del giudizio in generale. Gli sviluppatori di test chiedono generalmente ad uno, due o pochi osservatori esterni di validare il test dal punto di vista della facciata. Inoltre la maggior parte delle procedure che indagano la validità di facciata non si serve di analisi statistiche di alcun genere;
6. il sesto punto riguarda la tendenza di molti costruttori di test ad usare la validità di facciata come unico criterio di validità del test.

Roberts (2000), tuttavia, sottolinea che non dobbiamo seguire questo tipo di concettualizzazione. Egli sostiene che possiamo avvalerci della validità di facciata in un altro modo, ovvero tramite informatori esperti, e non solo come esame superficiale. Questo modo di procedere, sostiene l'autore, potrebbe trasformare la validità di facciata in una misura più efficace, dato che essa diventerebbe più simile al concetto di validità di contenuto. Essa può essere pensata riferirsi al grado con cui i rispondenti o gli utenti giudicano che gli item di un test siano appropriati al tipo di costrutto e agli obiettivi dell'*assessment* (Allen & Yen, 1979; Nevo, 1985). Inoltre, ci sono dei ricercatori che sostengono l'applicazione della validità di facciata perché aumenta la validità di costrutto attraverso l'accettazione da parte del rispondente della procedura testistica (Alderson, Clapham & Wall, 1995; Davies, Brown & Elder, 1999). Ad esempio, la validità di facciata ha ottenuto un nuovo status all'interno del testing sulle abilità linguistiche. In questo contesto, la favorevolezza nei confronti del concetto di validità di facciata è dovuta alla definizione *real-life* della competenza verbale (Carroll, 1985).

Validità di contenuto. La validità di contenuto si riferisce all'adeguatezza con cui gli item di un test rappresentano l'area di contenuto da misurare. Essa è primariamente una questione che riguarda i test di profitto, attitudine e abilità. La validità di contenuto si concentra sul campione di item di un test, che dovrebbero essere rappresentativi dei più importanti contenuti, abilità o comportamenti del dominio di interesse (Haynes, Richard, & Kubany, 1995). Essa è diversa dalla validità di facciata, ma quest'ultima può essere considerata, a volte, parte della prima.

Molte definizioni di validità di contenuto sono state pubblicate (e.g. APA Standards, 1985; Anastasi, 1993; Suen, 1990; Messick, 1993; Nunnally & Bernstein, 1994; Walsh, 1995; McDonald, 1999). Tuttavia, alcuni autori rifiutano il concetto di validità di contenuto come categoria di validità (Messick, 1993), e altri suggeriscono che sia più accurato considerare la validità di contenuto come il processo di operazionalizzazione di un costrutto (e.g. Guion, 1977).

Un test possiede validità di contenuto se i contenuti degli item sono indicatori del costrutto da misurare (McDonald, 1999, 457). La validità di contenuto si basa sul grado con cui una misura riflette lo specifico dominio di contenuto (Carmines & Zeller, 1979, 20). La validità di contenuto può essere illustrata attraverso il seguente esempio: alcuni ricercatori sono interessati a studiare l'apprendimento grammaticale di una lingua e, conseguentemente, a creare uno strumento che testi il costrutto "abilità grammaticale". Se i ricercatori testassero solo il dominio dei verbi, il loro strumento non avrebbe validità di contenuto, perché escluderebbe altri domini del costrutto. Come si vede è relativamente semplice stabilire la validità di contenuto per test di profitto, attitudine o abilità, ma il processo diviene via via più complesso mano a mano che ci si sposta sul piano di costrutti psicologici più astratti, come ad esempio l'autostima.

Proprio per quest'ultimo aspetto legato alla difficoltà di operazionalizzare costrutti psicologici complessi, molti autori (Guion, 1978; Mitchell, 1986; Groth-Marnat, 1990; Tallent, 1992; Messick, 1993) hanno messo in discussione la rilevanza della validità di contenuto per il testing psicologico. Inoltre, non vi è un accordo sostanziale tra gli psicometristi per quanto riguarda le procedure da adottare per monitorare la validità di contenuto di un test.

I giudizi degli esperti è il metodo di elezione per determinare se un test possiede validità di contenuto e l'accordo tra giudici, misurato con indici diversi, il sistema attraverso il quale decidere se un test presenta validità di contenuto oppure no. Evidenze di validità di contenuto si ottengono chiedendo a persone informate di guardare gli item del test ed emettere giudizi relativi all'appropriatezza di ciascun item e alla completezza della copertura del dominio. In questo caso, tuttavia, diversamente dal caso della validità di facciata, i giudici valutano gli item prima che il test venga approntato.

Quando la validità di facciata (gli item del test appaiono validi e significativi?) e la validità di contenuto (gli item del test sono un campione rappresentativo dell'area di contenuto da misurare?) sono state monitorare, è possibile passare ad altri tipi di validità, qui chiamate quantitative.

2.2 Tipi quantitativi di validità: validità di criterio e validità di costrutto

Validità di criterio. Gli item di un test dovrebbero presentare correlazioni elevate con gli item di altri test che misurano lo stesso costrutto. La validità di criterio si basa proprio sul calcolo di coefficienti di correlazione tra un test e altri, o tra un test e un criterio esterno. Viene usata per dimostrare l'accuratezza di una misura comparandola con un'altra misura che ha già evidenziato la sua validità. Essa misura anche il grado della relazione tra il punteggio di un test e un criterio esterno nel caso di validità concorrente, validità predittiva e validità di costrutto (McDonald, 1999).

Il principale problema con questo sistema di validazione è che tende a creare sistemi chiusi (Sternberg, 2000), ovvero sistemi circolari che tendono a confermare la bontà di un test a dispetto della sua reale validità (Sternberg, 1997). Con le parole di Sternberg (2000, 160): "*In validations that involve correlating new tests with existing tests, we 'reward' test authors and publishers to the extent that their new products are like old products, in regard both to the strengths and the weaknesses of the old tests. The more a new test departs from the old tests – even if for the better – the worse the test looks*".

Sebbene ci siano alcuni problemi teorici legati all'uso dei coefficienti di correlazione tra test per monitorarne la validità, la validità di criterio è ancora considerata un aspetto importante della validità di un test. Un test possiede validità di criterio se mostra correlazioni elevate con test che misurano lo stesso costrutto o un costrutto opposto. Nel primo caso (stesso costrutto) la correlazione deve essere alta e positiva. Nel secondo (costrutto opposto) la correlazione deve essere alta e negativa[3].

Esistono sostanzialmente due tipi di validità di criterio: la validità concorrente e la validità predittiva. Nel primo caso, costrutto e criterio sono misurati contestualmente. Nel secondo caso, il costrutto viene misurato prima e il criterio in qualche momento futuro.

Facciamo un tipico esempio, quello di un test di intelligenza. Un test di intelligenza valido dovrebbe avere una correlazione alta con altri test di intelligenza. Dovrebbe anche correlare con comportamenti che si ritiene richieda-

[3] La misura della validità di criterio tramite l'utilizzo di test-criterio già validati è un'operazione che ritroviamo, nell'ambito della validità di costrutto, nel caso di validità convergente/divergente. Ciò che distingue quest'ultimo tipo di validità dalla validtà di criterio è la possibilità che il criterio non sia semplicemente un altro test, ma, ad esempio, la misura di una performance.

no intelligenza, come andare bene a scuola. Se il criterio di un test di intelligenza è se correla con il profitto scolastico di un bambino nel periodo in cui il test viene somministrato, esso viene chiamato validità concorrente. Se il criterio di un test di intelligenza è quanto bene il test è in grado di predire il futuro rendimento di un bambino, come il fatto che arriverà fino alla laurea, allora viene chiamato validità predittiva.

Validità di costrutto. Può essere considerata il più importante tipo di validità. Si riferisce all'idea che un test debba misurare il costrutto per cui è stato costruito, non un altro costrutto. In questo senso costituisce la validità *tout-court*. La validità di costrutto si riferisce al grado in cui le inferenze possono legittimamente essere tratte dai punteggi di un test. Come il concetto di validità esterna, la validità di costrutto ha a che fare con la generalizzazione. Ma mentre la validità esterna implica la generalizzazione da un certo contesto di studio ad altri individui, luoghi o tempi, la validità di costrutto riguarda la generalizzazione da alcuni punteggi al concetto di costrutto o tratto latente. Un test possiede validità di costrutto se misura accuratamente un costrutto teorico non osservabile.

La validità di costrutto è stata definita dagli Standards APA come il grado con cui un individuo possiede un tratto ipotetico o costrutto che si presume si rifletta nella performance al test. L'idea di validare un costrutto è stata sviluppata principalmente in un lavoro pubblicato nel 1955 da Cronbach e Meehl. Generalmente parlando, un test possiede validità di costrutto se le evidenze suggeriscono che esso misuri il costrutto per cui il test è stato costruito. Come si vede la validità di costrutto coincide con il concetto stesso di validità. Essa è l'unica che viene considerata nell'Unitary Concept of Validity. Raccogliere evidenze di validità di costrutto implica fare delle ipotesi e raccogliere informazioni usando metodi diversi. Alcuni di questi metodi includono la Validazione Convergente/Divergente, l'Analisi Fattoriale e l'applicazione dei Modelli di Equazioni Strutturali, e, infine, i Modelli IRT (Item Response Theory) di cui il più famoso è il Modello di Rasch.

Validazione Convergente/Divergente. Il concetto di validità convergente si riferisce al grado con cui un test misura un certo costrutto in rapporto al grado con cui misura altri costrutti. Attualmente non c'è alcun accordo su come questa verifica debba essere portata a termine, ma generalmente parlando si può dire che un test possiede un'alta validità convergente se presenta una correlazione elevata con un altro test che misura lo stesso identico costrutto. Per contrasto, un test possiede validità divergente (o discriminante) se presenta un bassa correlazione con un test che misura un costrutto diverso.

Quest'ultima espressione – validità divergente (o discriminante) – è stata introdotta per descrivere il requisito intuitivo che il punteggio a un test non debba presentare correlazioni elevate con costrutti che non c'entrano nulla con quello che il test intende misurare.

La Matrice Multitratto-Multimetodo (Campbell & Fiske, 1959) si presenta come uno dei metodi più rigorosi, ma meno utilizzati, data la sua complessità di attuazione, per monitorare, contemporaneamente, la validità convergente e divergente. Gli autori suggeriscono che:

1. il coefficiente di correlazione (generalmente la r di Pearson) tra due misure massimamente simili dello stesso costrutto misuri l'attendibilità;
2. il coefficiente di correlazione tra due misure massimamente dissimili dello stesso costrutto fornisca evidenze nella direzione della validità convergente;
3. la differenza tra il coefficiente di validità convergente e la correlazione tra due misure massimamente dissimili di due costrutti differenti fornisca evidenze nella direzione della validità divergente.

Analisi Fattoriale e Modelli di Equazioni Strutturali. L'Analisi Fattoriale e i Modelli di Equazioni Strutturali sono tecniche di analisi multivariata che comprendono al loro interno altri metodi di analisi, come ad esempio:

1. *modelli Causali* o *Path Analysis*, per l'esplorazione di relazioni causali tra variabili mediante una rappresentazione grafica e un sistema di equazioni lineari;
2. *modelli di Regressione*, per il monitoraggio del rapporto indicatori-costrutto, quando i primi vengono trattati come formativi dello stesso e non semplicemente riflettivi[4].

L'Analisi Fattoriale Esplorativa viene utilizzata per monitorare come gli item di un test tendono a raggrupparsi in base alle correlazioni tra di loro. Essa viene condotta senza porre limiti al numero di raggruppamenti e nell'ottica di vedere (esplorare) cosa emerge dai dati, senza farsi guidare da un modello teorico. Viceversa, l'Analisi Fattoriale Confermativa viene con-

[4] È possibile distinguere fondamentalmente due tipi di indicatori o *item* (Ercolani & Perugini, 1997; Pedon & Gnisi, 2004): indicatori riflettivi e indicatori formativi. Nel caso di indicatori riflettivi, gli indicatori *riflettono* il costrutto psicologico sottostante. Essi vengono visti e trattati, di solito, come manifestazione empirica di detto costrutto, come la sua parte osservabile, e, anche, come conseguenza dell'esistenza e della presenza del costrutto medesimo. Nel caso di indicatori formativi, si può dire che essi *formano* il costrutto psicologico sottostante. In questo caso, essi sono ipotizzati formare, influenzare, determinare o addirittura causare il costrutto psicologico non osservabile.

dotta a partire da un modello teorico. Gli item vengono forzati ad appartenere a certi raggruppamenti e non ad altri. Gli indici di *fit* (adattamento) ci danno informazioni relativamente alla bontà del modello, data quella specifica matrice di dati.

Modelli di Rasch. È una famiglia di modelli matematici che studia la probabilità di dare una certa risposta a un test sulla base della differenza tra un tratto individuale e la difficoltà degli item. Il Modello di Rasch fornisce una cornice teorico-matematica con cui gli sviluppatori di test possono confrontare i loro dati e si basa sull'idea che una misurazione efficace implichi l'esame di un unico attributo alla volta (unidimensionalità). L'importanza dell'applicazione di questi modelli sugli item di un test risiede proprio nella possibilità di monitorarne l'unidimensionalità.

3. Conclusioni

Quando gli psicologi si trovano nella condizione di valutare la validità di un test che già esiste o che stanno essi stessi costruendo, essi possono passare attraverso una serie di operazioni progressive, tutte da condurre con cautela e circospezione, ma tutte possibilmente utili se riferite a un fine specifico e adeguato al metodo.

Prima di tutto essi posso *guardare* ciascun item del test e valutarlo sulla base di ciò che esso *sembra* misurare. Questa operazione può essere sensata nel caso, ad esempio, di test di profitto, mentre deve quanto meno essere giustificata nel caso, ad esempio, di test di personalità. Poi possono chiedere ad altre persone (esperte o no) di guardare gli stessi item e valutarli sulla base di alcuni criteri (chiarezza, pertinenza, esaustività, ecc.). In questi modi si raccolgono informazioni sulla validità di facciata e sulla validità di contenuto.

In seconda battuta, e per raccogliere informazioni più squisitamente quantitative, essi possono correlare ciascun item del test con ciascun item di altri test. Se un nuovo test, costruito per misurare un certo costrutto, mostra alte correlazioni con un test che già esiste e che misura lo stesso costrutto, operazionalizzato, possibilmente, nello stesso modo, questa può costituire un'evidenza che il nuovo test misura ciò che deve. Il vantaggio potrebbe essere che il nuovo test è più corto. Se poi il nuovo test mostra anche di correlarsi con un opportuno criterio esterno, anche questa diventa un'evidenza della sua bontà psicometrica. In questo modo si raccolgono

informazioni relative alla validità tramite criterio. Con la classica Analisi Fattoriale, invece, si possono avere indicazioni circa la validità di costrutto.

Alla fine di questo excursus, ci sia consentito fare qualche considerazione. Se è vero che la validità di un test si riferisce probabilmente ad un unico concetto (ovvero se il test fa ciò che deve), è pur vero che esistono strade diverse che si possono percorrere per accertarla. Il seguire strade diverse conduce a fare verifiche diverse, le quali, se tutte raccolte nel manuale, che sempre deve accompagnare un test validato, offrono all'utilizzatore la possibilità di capire nello specifico il funzionamento del test in diverse situazioni e da diverse angolature.

Bibliografia

Alderson, J. C., Clapham, C., Wall, D. (1995). *Language test construction and evaluation.* Cambridge: University Press.

Allen, M., Yen, W. M. (1979). *Introduction to measurement theory*. Monterey, CA: Brooks/Cole.

Ammassari, P. (1984). Validità e costruzione delle variabili: elementi per una riflessione. *Sociologia e Ricerca Sociale, V (13)*, 141-156.

Anastasi, A. (1993). *Psychological testing*. New York: Macmillan.

Anastasi, A., Urbina, S. (1997). *Psychological testing* (7th ed.). Upper Saddle River, NJ: Prentice Hall.

Angoff, W. H. (1988). Validity: an evolving concept. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 19-32). Hillsdale, NJ: Lawrence Erlbaum Associates.

Bachman, L. F (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Bailey, K. D. (1991). *Metodi della ricerca sociale*. Bologna: il Mulino.

Borsboom, D., Mellenbergh, G. J., Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111 (4)*, 1061-1071.

Buck, G. (2001). Validities. Message posted on L-TESL Online Forum. Available: http://f05n16.cac.psu.edu/archives/ltest-l.html.

Campbell, D. T., Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81-104.

Carmines, E. G., Zeller, R. A. (1979). *Reliability and validity assessment*. Beverly Hills: Sage.

Carroll, B. J., (1985). Second language performance testing for university and professional contexts. In P.C. Hauptman, R. LeBlanc, & M.B. Wesche (eds.), *Second language performance testing.* Ottawa: University of Ottawa Press.

Cattell, R. B. (1946). *Description and measurement of personality*. New York: World Book Company.

Crocker, L., Algina, J. (1986). *Introduction to Classical and Modem Test Theory*. New York: Holt, Rinehart and Winston.

Cronbach, L. J. (1971). Validity. In R. L. Thorndike (Eds.), *Educational Measurement* (pp. 443-597). Washington, D. C.: American Council on Education.

Cronbach, L. J. (1984). *Essentials of psychological testing. Fourth edition*. New York: Harper and Row.

Cronbach, L. J., Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bullettin, 52*, 281-302.

Davies, A., Brown, A., Elder, C. (1999). *Dictionary of language testing.* Cambridge University Press.

Dipboye, R. L., Smith, C. S., Howell, W. C. (1994). *Understanding Industrial and Organizational Psychology: An Integrated Approach*. Harcourt Brace College Publishers.

Fink, A. (1995). *How to ask survey questions*. Thousand Oaks, CA: Sage.

Giampaglia, G. (1990). *Lo scaling unidimensionale nella ricerca sociale*. Napoli: Liguori.

Groth-Marnat, G. (1990). *Handbook of psychological assessment (2nd ed.)*. New York: Wiley.

Guion, R. M. (1977). Content validity. The source of my discontent. *Applied Psychological Measurement, 1*, 1-10.

Guion, R. M. (1978). Content validity in moderation. *Personal Psychology*, *31*, 205-213.

Hajipournezhad, G. (2000). *An Approach to the Validation of Judgments in Language Testing*. [unpublished manuscript]

Hajipournezhad, G., (2002). Which one speaks louder in language testing, actions or words? *Proceedings of the Nov. 23-25, 2001 JALT Conference* in Kita Kyushu, Japan. Tokyo: JALT.

Haynes, S. N., Richard, D. C. S., Kubany, E. S. (1995). Content validity in psychological assessment: a functional approach to concepts and methods. *Psychological Assessment*, *7*, 238-247.

Kelley, T. L. (1927). *Interpretation of educational measurement*. New York: Macmillian.

Kerlinger, F. N. (1974). *Foundations of behavioural research*. New York: Holt, Rinehart and Winston.

Kline, P. (1998). *The New Psychometrics*. London: Routledge.

Lacity, M., Jansen, M. A. (1994). Understanding qualitative data: A framework of text analysis methods. *Journal of Management Information Systems*, *11*, 137-166.

Lennon, R. T. (1956). Assumption underlying the use of content validity. *Educational and Psychological Measurement, 16*, 294-304.

McDonald, R. P. (1999). *Test theory*. Mahwah, NJ: Lawrence Erlbaum Associates.

Messick, S. (1980). Test validity and the ethics of measurement. *American Psychologist, 30*, 955-966.

Messick, S. (1993). Validity. In R. L. Linn (Eds.), *Educational measurement (2nd ed.)* (pp. 13-104). Phoenix: American Council on Education and Oryx Press.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741-749.

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*, 35-44.

Mitchell, J. V. (1986). Measurement in the larger context: Critical current issues. *Professional Psychology: Research and Practice*, *17*, 544-550.

Mosier, C. I. (1947). A critical examination of the concepts of face validity. *Educational & Psychological Measurement*, *7*, 191-205.

Nevo, B. (1985). Face validity revisited. *Journal of Educational Meaurement*, *22*, 287-293.

Newfields, T. (2002). Challenging the notion of face validity. *SHIKEN: The JALT Testing & Evaluation SIG Newsletter*, 6 (3), 19. Available: http://www.jalt.org/test/new_2.htm.

Nunnally, J. C. (1978). *Psychometric theory*. New York: McGraw-Hill.

Nunnally, J. C., Bernstein, I. H. (1994). *Psychometric theory (3rd ed.)*. New York: McGraw-Hill.

Roberts, D. M. (2000). Face Validity: Is There a Place for This in Measurement? *SHIKEN: The JALT Testing & Evaluation SIG Newsletter*, 4 (2), 5. Available: http://www.jalt.org/test/rob_1.htm.

Sartori R. (2004). Validità, attendibilità, sensibilità e sensatezza dei metodi e delle misure in psicologia, *Quaderni DiPAV*, *9*, 147-165.

Sartori R. (2005). Le caratteristiche psicologiche esistono? Per una filosofia della psicometria. *Giornale Italiano di Psicologia*, *2*, 425-435.

Sartori R., Pasini M. (2007). Quality and quantity in test validity: how can we be sure that psychological tests measure what they have to? *Quality and Quantity*, *International Journal of Methodology*, *41*, 3, 359-374.

Standards for educational and psychological testing (1985). Washington, DC: American Psychological Association.

Sternberg, R. J. (1997). *Successful intelligence*. New York: Plume.

Sternberg, R. J. (2000). Implicit theories of intelligence as exemplar stories of success. *Psychology, Public Policy, and Law*, *6 (1)*, 159-167.

Suen, H. K. (1990). *Principles of test theories*. Hillsdale, NJ: Erlbaum.

Tallent, N. (1992). *The practice of psychological assessment*. Englewood Cliffs, NJ: Prenctice-Hall.

Trochim, W. (2002). *Measurement Validity Types*. [Online] Available: http://trochim.omni.cornell.edu/kb/.

Vernon, P. E. (1963). *Personality Assessment*. London: Metheun.

Walsh, W. B. (1995). *Tests and assessment*. New York: Prentice-Hall.