UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia dello Sviluppo e della Socializzazione (DPSS)

Corso di laurea in scienze psicologiche dello sviluppo, della personalità e delle relazioni interpersonali

Elaborato finale

Lo sviluppo di un test psicologico: una panoramica delle buone pratiche di costruzione e validazione

Developing a psychological test: an overview of good practice in construction and validation

Relatrice

Dott.ssa Tatiana Marci

Laureanda

Giulia Castellarin
(matricola 2045788**)**

Anno Accademico 2023-2024

INDICE

INTRODUZIONE

Nel panorama della ricerca psicologica, uno dei maggiori ambiti di interesse riguarda lo studio del comportamento umano e dei processi coinvolti nella sua manifestazione. I lavori in questo campo utilizzano frequentemente i costrutti latenti come oggetto delle proprie indagini: si tratta di caratteristiche psicologiche che non possono essere catturate in modo diretto, bensì solo attraverso degli indicatori direttamente osservabili che le rappresentano; questo aumenta la complessità della misurazione delle variabili latenti.

Per esaminare i costrutti latenti di interesse vengono utilizzati i test psicologici. Grazie a questi strumenti è possibile raccogliere informazioni sulla presenza di un costrutto in un soggetto mediante la misurazione dei suoi comportamenti osservabili.

I test psicologici devono essere in grado di catturare con precisione e in modo effettivo la caratteristica latente sottostante: la qualità dei risultati ottenuti, infatti, dipende dalla qualità delle misurazioni effettuate (Flake *et al.*, 2017). Affinché questo sia possibile, è necessario creare e utilizzare strumenti di misurazione che si dimostrano validi e affidabili. Questo compito è supportato dalla letteratura scientifica, che mette a disposizione diverse linee guida che aiutano il ricercatore nello sviluppo e nella scelta di misure valide.

Il tema della validità delle misure è oggi particolarmente rilevante a causa dell'utilizzo ormai diffuso di misure di dubbia validità: la letteratura evidenzia infatti una frequente carenza o trascuratezza di prove a supporto delle misure utilizzate nel contesto di ricerca (Flake *et al.*, 2017). La mancanza di questi dati, che esercitano un ruolo primario nel confermare la validità dei risultati, risulta spesso legata all'utilizzo di pratiche di misurazione discutibili (Flake & Fried, 2020). Questo ultimo aspetto si riscontra anche negli studi di replicabilità, che risultano influenzati dalla validità e dall'affidabilità delle misure utilizzate.

La creazione di una scala psicologica è una procedura ampia e complessa; la presenza di linee guida costituisce un punto di riferimento per il ricercatore nel compiere le scelte migliori per lo sviluppo del test.

Sulla base delle osservazioni riportate, che saranno affrontate durante il primo capitolo di questo elaborato, l'obiettivo di questa ricerca è quello di offrire, attraverso l'integrazione di diverse fonti presenti in letteratura (ad esempio, Boateng *et al.*, 2018;

Flake *et al.*, 2017; MacKenzie *et al.*, 2011), una panoramica delle buone pratiche nel processo di sviluppo e validazione di scale psicologiche valide e affidabili.

Sarà fornita una descrizione dei diversi passaggi necessari per la creazione di uno strumento di misura, che attraversa tutti i suoi stadi: partendo da un'adeguata definizione del dominio di costrutto e dalla generazione degli *item*, passando per l'identificazione e la valutazione del modello che rappresenta la struttura latente della scala e terminando con la verifica delle proprietà psicometriche dello strumento e la creazione di norme, viene così illustrato in tre fasi principali, suddivise in step più specifici, il percorso che porta allo sviluppo e alla valutazione di un test psicologico.

L'insieme di queste fasi, che costituiscono il secondo capitolo dell'elaborato, può fornire una guida utile per avere una prospettiva più chiara e comprensibile delle tappe da percorrere nel processo di realizzazione di un test, favorendo in questo modo una maggiore validità e affidabilità dei risultati che emergono dalle misurazioni.

CAPITOLO 1

BUONE PRATICHE NELLA MISURAZIONE PSICOLOGICA

Questo capitolo illustra brevemente il tema della misurazione in campo psicologico descrivendo la modalità attraverso cui i test utilizzati in questo ambito riescono a catturare le misure di una caratteristica psicologica latente. Sono inoltre riportate alcune conseguenze negative legate a una mancanza di prove a supporto dell'affidabilità e della validità degli strumenti che vengono creati e utilizzati nel contesto di ricerca. Viene infine fornita una breve panoramica introduttiva su quelle che possono essere definite delle buone pratiche di sviluppo e validazione di un test psicologico.

1. La misurazione in psicologia

La psicologia è una scienza empirica che, in quanto tale, fonda le sue teorie su fenomeni che sono innanzitutto riscontrabili nell'ambiente circostante. Nello specifico, il campo di interesse della psicologia è il comportamento umano: la ricerca in questo ambito si occupa di identificare e chiarire i processi che rendono possibile e che sono coinvolti nella manifestazione di un comportamento.

La creazione di una teoria psicologica parte dall'osservazione di un comportamento; sulla base di quello che osserva, il ricercatore formula un'ipotesi a cui, attraverso la conduzione di uno studio o di un esperimento, tenta di fornire una risposta. È necessario avere in mente in modo chiaro qual è la caratteristica psicologica che si intende indagare, in quanto essa costituisce il contenuto della misurazione che verrà effettuata.

La misurazione è un requisito fondamentale della scienza: senza di essa, infatti, non è possibile procedere a una verifica del fenomeno che si desidera studiare (Chiorri, 2023). In accordo con la definizione di Stevens (1946), con *misurazione* si intende il processo di attribuzione di un numero a un oggetto seguendo delle regole precise. Questo concetto appare di facile comprensione e risulta semplice da utilizzare quando si lavora con attributi o caratteristiche che possono essere misurate in modo diretto (ad esempio, la lunghezza): in questi casi, è possibile rilevare immediatamente il valore quantitativo della proprietà oggetto di esame, ottenendo una misura della stessa (Chiorri, 2023). In ambito psicologico, a differenza di altre discipline, questa operazione risulta tuttavia più

complicata, poiché la caratteristica che si vuole indagare non è sempre misurabile per via diretta: molti aspetti del comportamento umano sono infatti costituiti da costrutti latenti.

Un costrutto latente è un concetto teorico che non può essere catturato in modo diretto, in quanto non direttamente osservabile; per poter misurare un costrutto risulta necessario rilevare le sue manifestazioni osservabili, che prendono il nome di indicatori del costrutto (Chiorri, 2023). Una caratteristica psicologica latente, quindi, può essere catturata attraverso la misurazione dei comportamenti manifesti che si presuppongono essere gli indicatori direttamente osservabili del costrutto di interesse.

Per individuare gli indicatori specifici che consentono la misurazione di un costrutto latente è importante, innanzitutto, definire in modo preciso il costrutto che si intende esaminare. Questo implica che il ricercatore possieda a livello teorico un'idea chiara del costrutto che intende misurare e, successivamente, sappia identificare le manifestazioni comportamentali maggiormente rappresentative della variabile latente (Chiorri, 2023). La procedura di individuazione degli indicatori di un costrutto è conosciuta come operazionalizzazione: durante questo processo i concetti scientifici vengono legati a operazioni osservabili ed eseguibili da tutti, ottenendo un insieme di comportamenti direttamente osservabili che costituiscono il dominio di contenuto del costrutto (Chiorri, 2023).

Per misurare i suoi costrutti di interesse, la ricerca psicologica ha bisogno di utilizzare degli strumenti che catturino in modo preciso e accurato gli indicatori che rappresentano la caratteristica latente: questi strumenti sono i test psicologici.

Volendo riferirsi a una definizione specifica, un test psicologico può essere descritto come "una procedura sistematica per ottenere un campione di comportamento, rilevante per il funzionamento cognitivo o affettivo, e per assegnare a tale campione di comportamento un punteggio che sia confrontabile rispetto a dei valori standard di riferimento, in modo da poter compiere una valutazione" (Urbina, 2004, p. 1). Lo scopo di un test psicologico è quindi quello di raccogliere informazioni su un costrutto mediante la misurazione dei suoi indicatori direttamente osservabili.

L'utilizzo di un test psicologico che consente di misurare il costrutto latente di interesse richiede di considerare lo scopo per cui il test stesso viene usato: la distinzione principale è quella tra i test orientati al criterio, che hanno l'obiettivo di individuare gruppi di persone sulla base del punteggio ottenuto mediante la somministrazione dello strumento, e i test

orientati al costrutto, i quali intendono indagare o confermare a livello teorico nuovi aspetti del costrutto di interesse (Chiorri, 2023). La definizione dell'obiettivo dello strumento è essenziale per la fase di operazionalizzazione del costrutto, in quanto guida il processo di selezione degli indicatori osservabili della variabile latente.

All'interno di un test psicologico, gli indicatori scelti per rappresentare il costrutto sono presentati sotto forma di *item*; questo insieme di stimoli consente di catturare i comportamenti manifesti della caratteristica psicologica latente. Il compito di creazione degli *item* prevede che le prove proposte siano generate rispettando precise regole strutturali e linguistiche quali la chiarezza, la centralità rispetto al costrutto e la non offensività e inclusività di linguaggio (Chiorri, 2023); richiede inoltre una decisione rispetto al formato di risposta che verrà utilizzato dai soggetti a cui il test è indirizzato.

Nella selezione del formato di risposta agli *item* si considera una suddivisione dei test tra test di prestazione tipica e test di prestazione massima (Chiorri, 2023): mentre i test di prestazione tipica intendono raccogliere informazioni personali su atteggiamenti, stati d'animo e opinioni degli individui, i test di prestazione massima osservano invece la capacità dei soggetti di rispondere correttamente alle prove proposte; da questi due scenari deriva una diversa impostazione degli *item* e della loro modalità di risposta.

Nel caso dei test di prestazione tipica, gli stimoli sono normalmente costituiti da *item* dicotomici (che presentano cioè due opzioni generalmente opposte), risposte a scelta multipla forzata o scale di valutazione (un esempio molto conosciuto è quello della scala Likert) (Chiorri, 2023). L'intento di questi test è quello di verificare la "tipicità" dei comportamenti che riflettono il costrutto indagato, motivo per cui questi formati di risposta non prevedono la selezione della risposta corretta, bensì quella che sembra descrivere al meglio i comportamenti dei soggetti.

Nei test di prestazione massima, trattandosi di prove che valutano la rapidità e l'accuratezza di risposta dei partecipanti, sono frequenti domande a risposta aperta e chiusa con *item* composti da prove di diversa tipologia, con l'obiettivo di catturare il grado di abilità posseduto dagli individui (Chiorri, 2023); sono spesso presenti anche formati di risposta dicotomici o a scelta multipla (Chiorri, 2023).

I comportamenti che vengono rilevati attraverso la somministrazione del test forniscono informazioni sulla presenza della variabile indagata in un soggetto: ciascuna risposta agli *item* di un test cattura, nello specifico, il grado in cui un individuo possiede

la caratteristica psicologica che si sta esaminando (Chiorri, 2023). A partire dalle risposte agli stimoli, per ottenere questa misura in forma quantitativa si ricorre allo *scaling*, una tecnica psicometrica che permette di assegnare un punteggio a ogni risposta degli *item*: ciascun valore descrive la quantità di costrutto posseduta da un individuo sulla base dei punteggi osservati nei suoi indicatori (Chiorri, 2023). Grazie allo *scaling* è possibile confrontare soggetti differenti che possiedono quantità diverse del costrutto considerato.

L'operazione di *scaling*, che consente in seguito di ottenere il punteggio finale di un test, richiede che il ricercatore consideri la relazione tra la caratteristica psicologica che intende misurare e i suoi indicatori: viene creato un modello di misurazione che descrive questo legame e che necessita di essere testato per verificare che gli stimoli siano effettivamente connessi al costrutto latente secondo quanto ipotizzato (Chiorri, 2023). In molti casi si fa riferimento alla Teoria Classica dei Test (*Classic Test Theory,* TCC), per la quale le variabili osservabili sono influenzate per la maggior parte dal costrutto latente e in minore quantità da altri fattori specifici per ciascuna variabile non direttamente riconducibili al costrutto; questi ultimi costituiscono l'errore di misurazione (Chiorri, 2023).

L'errore di misurazione, secondo la concezione della TCC, è una componente intrinseca del punteggio finale di un test: essa che si somma al punteggio vero che rappresenta la quantità effettiva della caratteristica psicologica misurata. Questa consapevolezza permette al ricercatore di adottare delle pratiche metodologiche per provare a contenere e ridurre l'errore di misurazione (Chiorri, 2023), in modo da ottenere dei punteggi che rispecchino il più possibile la reale presenza del costrutto.

Le misure raccolte mediante la somministrazione del test sono dei valori che non consentono di stabilire a priori se lo strumento misuri effettivamente quello che intende misurare, né forniscono una stima della precisione delle misure stesse: nonostante la selezione degli indicatori più rappresentativi del costrutto, la creazione di domande che riflettono il contenuto della variabile latente e la verifica del modello di misurazione scelto, per poter interpretare e usare correttamente i risultati di un test è necessario dimostrare che le misure ottenute sono valide e affidabili (Chiorri, 2023).

2. Criticità e conseguenze di misure non valide nel contesto di ricerca

Il tema della validità delle misure è oggi al centro dell'interesse della ricerca in diversi ambiti, tra cui quello psicologico, a causa dell'utilizzo di un largo numero di misure di dubbia validità (Flake *et al.*, 2017). L'utilizzo di misure di questo tipo ha diverse ripercussioni negative sui risultati di uno studio e sulle loro possibili applicazioni pratiche: come sottolineano Flake e collaboratori (2017), se un costrutto è studiato con misure qualitativamente scadenti, si riduce la possibilità di compiere affermazioni sul fenomeno indagato, dal momento che non si conosce quello che effettivamente viene misurato. La possibilità di verifica di un costrutto psicologico risiede allora nella validità della sua misurazione (Flake *et al.*, 2017).

Diverse sono le linee guida che sono state sviluppate per indirizzare e sostenere la creazione e la valutazione di test psicologici validi e affidabili; tra le più conosciute rientrano *The Standards of Educational and Psychological Testing (*American Educational Research Association, American Psychological Association, & National Council on Measurement in Education [AERA, APA, & NCME], 2014). Lo scopo di lavori come questo è quello di guidare il ricercatore nel processo di generazione e validazione di strumenti in grado di raccogliere le misure di costrutti latenti seguendo procedure che garantiscono la validità dei risultati ottenuti.

Nonostante sia stata sottolineata più volte nel corso del tempo l'importanza della validità delle misure, promossa anche dalla presenza diffusa delle linee guida come quella degli *Standards* (AERA, APA, & NCME, 2014), l'attuale contesto di ricerca in psicologia mostra una realtà diversa: accade frequentemente che gli strumenti che vengono utilizzati non presentino prove a supporto della loro validità (Flake *et al.*, 2017); un altro problema presente riguarda inoltre la somministrazione di strumenti la cui validità è stata confermata esclusivamente in alcune popolazioni, ma non in altre (Kane, 2013; Messick, 1995).

Situazioni come queste rispecchiano il modo in cui le prove a supporto della validità dei test psicologici vengono descritte e riportate all'interno degli studi che ne fanno uso. Questo aspetto è stato messo in luce da uno studio di Flake e colleghi (2017), che hanno sottoposto a revisione un campione rappresentativo di 35 articoli pubblicati sul *Journal of Personality and Social Psychology* (2014) con l'obiettivo di determinare la rigorosità delle metodologie psicometriche utilizzate dagli autori nelle loro ricerche.

I risultati rilevati evidenziano una marcata carenza o, talvolta, una totale assenza di prove di evidenza empirica e riferimenti chiari in merito alla validità e affidabilità degli strumenti usati: nello specifico, il 53% degli studi considerati fornisce prove di validità riportate in studi precedenti senza sottoporle a ulteriore verifica; il 19% di queste scale, inoltre, risulta sottoposto a modifiche o adattamenti specifici ai fini degli studi senza presentare alcuna conferma del mantenimento delle proprietà psicometriche della versione originale (Flake *et al.*, 2017). La percentuale rimanente degli strumenti utilizzati non presenta alcuna fonte o citazione di validità, motivo per cui si ritiene che queste scale psicologiche siano state create o adattate direttamente dagli autori (supposizione confermata dagli stessi autori nel 7% dei casi) (Flake *et al.*, 2017).

Per quanto riguarda l'affidabilità degli strumenti degli studi esaminati, la maggior parte di questi presenta come unica conferma la misura dell'Alpha di Cronbach, presente nel 73% delle ricerche (Flake *et al.*, 2017); talvolta, questo coefficiente costituisce l'unica evidenza psicometrica riportata.

Anche un lavoro di MacKenzie e collaboratori (2011) sottolinea come la maggior parte degli studi pubblicati non si dedichi ancora a una esaustiva validazione degli strumenti (Boudreau *et al.*, 2001) e che, anzi, nel tempo si abbia assistito a un decremento degli studi che riportano prove di validità (Scandura & Williams, 2020).

La mancanza di informazioni sulla validità delle misure utilizzate apre la strada ad alcune riflessioni in merito all'utilizzo di pratiche di questo tipo in campo scientifico e alle loro possibili ripercussioni negative sui risultati di ricerca. Secondo Flake e Fried (2020), l'assenza di prove di validità rientra all'interno delle cosiddette *Questionable Measurement Practices* (QMPs*)*; queste sono descritte dagli autori come "decisioni prese dai ricercatori che sollevano dubbi sulla validità dell'uso della misura utilizzata in uno studio e, di conseguenza, sulle conclusioni finali dello studio" (Flake & Fried., 2020, p. 458).

Le QMPs sollevano la necessità di affrontare il problema della frequente mancanza di trasparenza delle procedure metodologiche di sviluppo ed esecuzione delle ricerche (Flake & Fried, 2020) e, quindi, anche della creazione e validazione dei test psicologici. Le linee guida presenti in letteratura cercano di ostacolare l'applicazione delle QMPs promuovendo l'utilizzo di buone pratiche di misurazione.

Il ruolo delle linee guida è stato considerato anche di fronte alla crisi di replicabilità a cui si sta assistendo negli ultimi anni. La possibilità di poter replicare uno studio, che consente di ottenere gli stessi risultati dello studio originale a fronte di identiche condizioni procedurali utilizzando un nuovo campione di dati, è considerata un aspetto cruciale della ricerca in psicologia e della scienza in generale (Zwaan et al., 2018), in quanto garantisce di confermare i risultati già ottenuti in precedenza e aprire la strada a eventuali futuri obiettivi di ricerca.

Tra le cause che contribuiscono alla diffusione della recente crisi si riscontra la trascuratezza dei principi di base della misurazione (Lilienfeld & Strother, 2020), che riguardano la validità e l'affidabilità delle misure; anche Fried e Flake (2018) riconoscono il legame presente tra cattive pratiche di misurazione e la crisi di replicabilità in corso.

La conduzione di studi di replicabilità e il ruolo della replicabilità in campo scientifico, per quanto importanti, rischiano di perdere utilità se un test psicologico manca di validità o si rivela misurare costrutti diversi in campioni differenti (Fried & Flake, 2018). Quello che serve in primo luogo sono buone misure di validità: è solo rendendo primario il ruolo delle teorie di misurazione che gli studi di replicabilità possono progredire di conseguenza (Fried & Flake, 2018).

Indagando più profondamente sulla mancanza di evidenze di misure valide, MacKenzie e colleghi (2011) si interrogano sulle possibili cause che rendono il processo di sviluppo e validazione di una scala psicologica difficile da seguire in modo preciso e completo. Il problema non risiederebbe, secondo gli autori, nella carenza di articoli che si focalizzano sulle procedure di validazione: si ipotizza piuttosto che alcuni ricercatori non possiedano una conoscenza approfondita delle procedure stesse e non riescano dunque a comprendere pienamente le raccomandazioni suggerite (MacKenzie *et al.*, 2011).

Il lavoro di creazione e validazione di un test è molto ampio e talvolta può risultare complicato stabilire delle priorità e compiere delle scelte su quello che deve essere effettivamente svolto. Come sottolineano Flake e Fried (2020), i possibili approcci alla misurazione in psicologia sono molteplici e le decisioni che il ricercatore deve prendere sono numerose. L'utilizzo di linee guida garantisce un riferimento metodologico solido su cui poter fare affidamento nella selezione delle migliori procedure di sviluppo di uno strumento psicologico valido e affidabile.

3. Buone pratiche per lo sviluppo e la validazione delle scale psicologiche

Le difficoltà che il ricercatore incontra nelle procedure di sviluppo e validazione di scale psicologiche valide e affidabili sono reali; le conseguenze di pratiche di misurazione discutibili, che in queste circostanze possono apparire come una strada facilmente percorribile, si riflettono negativamente sui risultati di ricerca e sulle conclusioni finali degli studi condotti. La presenza di linee guida come gli *Standards (*AERA, APA, & NCME, 2014) promuove l'utilizzo di misure valide e affidabili, fornendo una spiegazione dettagliata e rigorosa dei criteri da utilizzare nello sviluppo di uno strumento di misurazione in campo psicologico.

L'insieme di queste pratiche raccomandate è riportato in uno studio di Flake e colleghi (2017), che evidenziano lo stretto legame tra la validità di uno studio e il processo di validazione del costrutto. Questa procedura accompagna lo sviluppo di una scala psicologica in tutte le fasi di costruzione di un test e ha l'obiettivo di fornire evidenza circa il fatto che i punteggi delle misure riflettono effettivamente il costrutto indagato, rafforzando così le conclusioni tratte dalle ricerche svolte (Flake *et al.*, 2017).

Gli *Standards* (AERA, APA, & NCME, 2014) individuano tre fasi da seguire nella generazione di uno strumento psicologico e nel processo di validazione del costrutto: la fase sostanziale, la fase strutturale e la fase esterna (Flake *et al.*, 2017). Il loro dispiegamento accompagna lo sviluppo del test in maniera sequenziale: ciascuna fase rappresenta infatti un prerequisito per la fase successiva. Questo significa, ad esempio, che i risultati della terza fase potrebbero non essere validi se non si ha la certezza che ad essere valide siano sia le teorie di riferimento, sia le proprietà psicometriche, che riguardano le due fasi precedenti della validazione del costrutto (Flake *et al.*, 2017).

La fase sostanziale prevede che venga eseguita un'analisi della letteratura presente in merito al tema di interesse, indagando gli aspetti teorici e gli aspetti di misura legati al costrutto per procedere quindi allo sviluppo degli *item* e alla valutazione della loro rilevanza (Flake *et al.*, 2017).

La fase strutturale implica lo svolgimento di alcune analisi di tipo quantitativo per esplorare le proprietà psicometriche della misura: queste consistono generalmente nell'analisi degli *item*, nell'analisi fattoriale e nei test di affidabilità e invarianza della misura (Flake *et al.*, 2017).

La fase esterna costituisce, infine, il momento di raccolta di prove sulla relazione del costrutto indagato con altri costrutti o sulla sua capacità predittiva rispetto ad altri criteri (Flake *et al.*, 2017).

Un altro lavoro di Boateng e collaboratori (2018) fornisce un *primer* in cui vengono illustrate delle buone pratiche per lo sviluppo delle scale che si occupano della misurazione di fenomeni psicologici complessi. L'obiettivo è quello di facilitare la creazione di nuove scale valide e affidabili e contribuire a migliorare quelle già esistenti utilizzate nel contesto di ricerca.

Anche questi autori, come nel lavoro di Flake e colleghi (2017), identificano una procedura contraddistinta da tre fasi principali, a loro volta suddivise in sottofasi più specifiche: lo sviluppo degli *item*, che inizia con l'identificazione del dominio di costrutto e si muove successivamente sulla creazione degli stimoli e sulla valutazione della validità interna; lo sviluppo della scala, che prevede che gli *item* siano somministrati a più riprese per selezionare e mantenere solo quelli più rappresentativi del costrutto e che, attraverso tecniche di analisi fattoriale, identifica la struttura latente di un insieme di *item* (Boateng *et al.*, 2018); la valutazione della scala, che indaga la dimensionalità dello strumento e in seguito si occupa di testare l'affidabilità e la validità esterna della scala creata.

Le fasi descritte in questi due lavori, parzialmente sovrapponibili, mettono in luce ancora una volta la complessità della misurazione dei costrutti latenti in ambito psicologico, dovuta alla loro impossibilità di essere catturati in modo diretto e che necessita pertanto di una massima attenzione in tutti i suoi aspetti procedurali.

Nel prossimo capitolo sarà svolto un lavoro di sintesi e integrazione di questi step con altri riferimenti dalla letteratura; ciascun passaggio sarà illustrato e approfondito con maggiore precisione.

CAPITOLO 2

GLI STEP PER LO SVILUPPO E LA VALIDAZIONE DI UN TEST PSICOLOGICO

L'obiettivo di questo capitolo è quello di offrire una panoramica lineare e ordinata delle procedure che portano allo sviluppo e, successivamente, alla valutazione di un test psicologico attraverso una descrizione di tutte le sue fasi. Questo percorso comincia con l'individuazione degli indicatori che rappresentano la variabile latente e termina con il processo di validazione dello strumento creato.

1. La prima fase: sviluppo degli *item*

Come precedentemente descritto, un test psicologico viene generalmente creato con lo scopo di indagare una caratteristica psicologica che, frequentemente, si manifesta attraverso degli indicatori comportamentali che riflettono il costrutto latente. Non si può creare una scala senza prima aver identificato con esattezza il dominio del costrutto che si vuole studiare.

Di conseguenza, il primo step per la creazione di un test riguarda l'articolazione del dominio o dei domini che costituiscono l'oggetto della misurazione (Boateng et al., 2018). Sulla base del dominio concettuale che si va a definire si può successivamente comporre un primo grande insieme di *item* (stimoli) in grado di catturare il costrutto latente.

Per assicurare la qualità degli stimoli generati è in seguito essenziale testare la validità di contenuto dello strumento, che riflette la rilevanza e la rappresentatività degli *item* rispetto alla variabile latente.

1.1. Identificazione del dominio di costrutto

Nel tentativo di fornire una definizione alla parola *costrutto*, Nunnally e Bernstein (1994, p. 85) lo descrivono come "qualcosa che gli scienziati costruiscono, cioè mettono insieme a partire dalla loro immaginazione". Il dominio di costrutto fa riferimento, nello specifico, all'attributo o al comportamento obiettivo di cattura dello studio. A livello pratico, questo si traduce in una procedura che mira a creare una conoscenza operativa

del fenomeno, stabilire i confini del dominio e definire chiaramente quello che rientra nel dominio del costrutto scelto e cosa, invece, non ne fa parte (Boateng *et al.*, 2018).

Il processo di definizione del costrutto passa per diverse fasi: queste riguardano la specificazione dello scopo della caratteristica psicologica e del suo dominio, la conferma dell'utilità dello strumento nella misura in cui intende catturare il costrutto scelto in quanto non ancora presente in ambito di letteratura o, se esistente, nel modo in cui si differenzia dagli altri strumenti, e una sua descrizione attraverso una definizione preliminare accurata e completa (Boateng et al., 2018; McCoach et al., 2018).

Nell'individuazione degli aspetti chiave del dominio di costrutto entrano in gioco sia la revisione della letteratura già presente, sia le interviste con persone professioniste e/o esperte nel campo, da cui è possibile ricavare preziose informazioni che delineano in una forma migliore le caratteristiche della variabile latente (MacKenzie *et al.*, 2011).

Il dominio del costrutto fa riferimento sia a quello che il costrutto intende rappresentare, sia al modo in cui differisce da altri costrutti simili. Nel dettaglio, il compito del ricercatore in questa fase riguarda la specificazione della natura del costrutto e del suo tema concettuale in una forma esclusiva, non ambigua e coerente con le ricerche già svolte (MacKenzie, 2003).

La natura del costrutto è composta da due componenti: il dominio concettuale a cui appartiene il costrutto e l'entità a cui esso si applica (MacKenzie *et al.*, 2011). Il primo riguarda le proprietà a cui il costrutto si riferisce, ovvero la sua tipologia (ad esempio, un pensiero, una sensazione, una caratteristica intrinseca…); la seconda identifica l'oggetto a cui questa proprietà si applica (ad esempio, una persona, una diade, un compito…).

La definizione del tema concettuale del costrutto, che avviene subito dopo, considera gli attributi e le caratteristiche necessarie e sufficienti affinché un elemento possa essere considerato un esemplare del costrutto (MacKenzie *et al.*, 2011). Idealmente, ciascun attributo o caratteristica del tema concettuale dovrebbe essere comune e, al tempo stesso, posseduto esclusivamente da tutti gli esemplari del costrutto. Il tema concettuale richiede di considerare, inoltre, la stabilità del costrutto al trascorrere del tempo e in situazioni e casi differenti, come evidenziano, ad esempio, Chaplin e collaboratori (1998) nella differenziazione tra i concetti di *tratto* e *stato*.

Nessuno di questi due compiti risulta di semplice svolgimento; è importante prestare molta attenzione a questa fase delicata che viene spesso trattata in modo solo superficiale

e che, come conseguenza, rischia di creare confusione attorno a quello che si intende misurare; questo, secondo MacKenzie (2003), può minacciare diversi aspetti della ricerca, tra cui la validità di costrutto e la validità interna.

La fase successiva è quella della dimensionalità del costrutto, aspetto che viene approfondito da MacKenzie e collaboratori (2011). Questa fase indaga il numero di dimensioni del costrutto e il modo in cui queste sono relazionate con la variabile latente.

Per stabilire la dimensionalità del costrutto bisogna considerare l'unicità di ogni dimensione e le conseguenze dell'eliminazione di una di queste (MacKenzie *et al.*, 2011): se le diverse sotto-dimensioni di un costrutto non presentano aspetti di unicità e l'eliminazione di una di queste non comporta un restringimento del dominio del costrutto, allora si tratta di un costrutto unidimensionale; se invece ogni sotto-dimensione rappresenta un aspetto di unicità del costrutto e la rimozione di una di queste produce una riduzione del dominio concettuale del costrutto, allora ci si trova di fronte a un costrutto multidimensionale.

Il lavoro con i costrutti multidimensionali introduce, oltre a una definizione attenta di ciascuna delle singole componenti del costrutto, un'ulteriore riflessione sul legame tra ciascuna sotto-dimensione e la variabile latente a cui si riferiscono (MacKenzie *et al.*, 2011), che può essere di tipo formativo o riflettivo: se le sotto-dimensioni sono viste come caratteristiche definenti il costrutto e se la modifica di una sola di queste componenti comporta un cambiamento a livello del costrutto principale, allora gli indicatori del costrutto sono definiti formativi; diversamente, se le sotto-dimensioni si presentano come manifestazioni di un costrutto, il costrutto sembra esistere anche in modo indipendente e un suo cambiamento causa una modifica di tutte le sue sotto-dimensioni, allora ci si trova di fronte a indicatori di tipo riflettivo.

Nonostante sia utile definire la tipologia di relazione tra un costrutto e le sue sotto-dimensioni per la scelta di alcune misurazioni in fasi successive dello sviluppo del test, è importante ricordare che nessun costrutto è interamente formativo o riflettivo (MacKenzie *et al.*, 2011), in quanto questa connessione dipende dal contenuto di ogni dimensione e dal modo in cui il costrutto viene concettualizzato dal ricercatore.

1.2. Generazione degli *item*

In presenza di una chiara definizione del costrutto e della sua dimensionalità, è finalmente possibile procedere al passaggio successivo, che riguarda lo sviluppo degli *item*. L'obiettivo è quello di generare un insieme di stimoli che rappresentano in modo completo il dominio del costrutto cercando, allo stesso tempo, di non fuoriuscire dal quadro del dominio precedentemente individuato (MacKenzie *et al.*, 2011).

Boateng e colleghi (2018) ritengono che ci siano due modi per generare gli *item* di uno strumento e che l'ideale sia utilizzarli entrambi: si tratta dei metodi deduttivi e dei metodi induttivi. I primi si basano sulla descrizione del dominio e l'identificazione degli stimoli attraverso revisioni della letteratura e valutazioni di scale già esistenti; i metodi deduttivi per la creazione degli *item* partono invece dalle risposte degli individui durante procedure come interviste e *focus group*. Coniugando due linee di lavoro diverse tra loro, entrambe le soluzioni permettono di generare stimoli appropriati al costrutto in esame.

La forma degli *item* è importante tanto quanto il loro contenuto. Boateng e collaboratori (2018), così come MacKenzie e colleghi (2011), concordano sul fatto che gli stimoli devono essere concettualizzati nel modo più semplice, diretto e preciso possibile e senza risultare ambigui (Boateng *et al.*, 2018); questo aspetto deve essere tenuto in considerazione anche nel momento in cui si sceglie il formato di risposta degli *item*, insieme ai dati della letteratura sulla validità di questi ultimi. Va prestata attenzione anche all'inclusività del linguaggio e alla presenza di eventuali *bias*, ovvero di distorsioni cognitive, che dovrebbero essere risolti (Schinka *et al.*, 2013).

Nella scelta degli stimoli è necessario stabilire anche il formato di risposta per raccogliere i dati (Chiorri, 2023): questa decisione dipende prevalentemente dalla tipologia di test che si intende sviluppare e dal suo scopo. Le scale di prestazione tipica utilizzano modalità di risposta che riflettono la diversa "tipicità" dei comportamenti dei soggetti; i test di prestazione massima sono in gran parte costituiti da domande o prove a cui il soggetto prova a fornire la risposta corretta.

In generale, è raccomandabile sviluppare un numero di *item* anche piuttosto elevato, circa il doppio (Kline, 1993; Schinka *et al.*, 2013) rispetto a quelli che faranno parte della versione finale della scala: in questo modo è infatti possibile avere a disposizione una

maggiore quantità di opzioni nella scelta degli stimoli che si dimostrano più rilevanti nel catturare accuratamente il costrutto di interesse.

1.3. Validità di contenuto

Gli *item* della scala, una volta creati, sono esaminati mediante una procedura che permette di stabilire se ogni stimolo misura e riflette effettivamente il dominio della caratteristica psicologica oggetto di studio.

La validità di contenuto, secondo la definizione di Hinkin (1995, p. 968), si riferisce alla "adeguatezza con cui una misura valuta il dominio di interesse": grazie a questa verifica è possibile controllare che gli indicatori della scala si occupino di misurare esclusivamente il dominio di costrutto definito precedentemente, e dunque la rilevanza, la rappresentatività e la qualità tecnica degli *item* che compongono lo strumento (Boateng *et al.*, 2018; Chiorri, 2023). Questo significa che ogni stimolo preso singolarmente deve essere in grado di catturare un aspetto del dominio, e la totalità degli *item* deve risultare complessivamente rappresentativa dell'intero costrutto (MacKenzie *et al.*, 2011).

Per testare la validità di contenuto si ricorre spesso a valutazioni di persone che sono giudici esperti della materia oppure riflettono la popolazione *target* alla quale è indirizzata la somministrazione del test (Boateng *et al.*, 2018). In entrambi i casi, la raccomandazione è quella di selezionare persone che possiedono sufficienti conoscenze e abilità per valutare correttamente la corrispondenza tra gli *item* e la componente teorica da considerare (MacKenzie *et al*., 2011).

Le valutazioni effettuate da giudici esperti prendono in esame ciascun *item* singolarmente per stabilire se si dimostra sufficientemente rappresentativo del dominio. Le opinioni raccolte sono quantificate utilizzando procedure come la *Content Validity Ratio*, il *Content Validity Index* e il coefficiente "kappa" (*k*) di Cohen (Boateng *et al.*, 2018).

Un'altra possibilità che si può percorrere è la valutazione degli *item* da parte della popolazione *target* (Boateng *et al.*, 2018). Questa tipologia è particolarmente utile per la verifica della validità di facciata, che rientra nella validità di contenuto e viene definita come "il grado in cui i rispondenti giudicano che gli *item* di uno strumento di valutazione sono appropriati al costrutto *target* e agli obiettivi della valutazione" (Haynes *et al.*, 1995, p. 243). In parole più semplici, riguarda l'impressione del soggetto a cui la scala è

somministrata circa il test stesso, che "sembra" misurare il costrutto che intende catturare (Chiorri, 2023).

Idealmente, la disponibilità di entrambi i tipi di valutazione è da favorire per ottenere una migliore stima della validità di contenuto. Tuttavia, se le risorse disponibili sono limitate, Boateng e collaboratori (2018) raccomandano almeno l'utilizzo dei giudici esperti.

2. La seconda fase: sviluppo della scala e valutazione dell'affidabilità

Una volta che gli *item* sono stati creati, questi sono poi sottoposti a una prima fase di *pre-testing*. In seguito, gli stimoli sono utilizzati per la conduzione di uno studio preliminare: è il momento in cui gli *item* vengono somministrati a un campione che riflette quello reale, viene determinata la numerosità del campione per raccogliere dati utili allo sviluppo del test e alla valutazione delle proprietà psicometriche e si stabilisce la modalità di utilizzo dei dati che verranno utilizzati per la validazione della scala.

Si procede quindi all'analisi formale degli *item* attraverso tecniche di analisi statistica, che consentono di individuare gli stimoli più rilevanti per il dominio del costrutto in esame.

Attraverso la successiva procedura di analisi fattoriale è possibile individuare il numero dei fattori latenti correlati ai rispettivi *item* che li rappresentano.

Infine, i test di affidabilità verificano la consistenza e la continuità delle risposte fornite dai soggetti quando la misurazione viene ripetuta utilizzando la stessa scala psicologica.

2.1. *Pre-testing* degli *item*

Procedere a testare gli stimoli appena generati prima della somministrazione della scala definitiva consente di essere certi che gli *item* siano significativi per la popolazione per cui sono pensati (Boateng *et al.*, 2018). Questo step si articola in due parti: la prima riguarda l'indagine del grado in cui le domande riflettono il dominio di interesse; l'altra considera la capacità degli *item* di produrre, attraverso le risposte dei soggetti, misure valide (Fowler, 1995).

Lo strumento di *pre-testing* maggiormente utilizzato è costituito dalle interviste cognitive, mediante le quali gli stimoli che compongono la scala, ancora in stato di bozza, vengono presentati a campioni della popolazione *target*; si chiede quindi ai partecipanti

di verbalizzare il processo di scelta delle risposte agli *item* (Beatty & Willis, 2007), con lo scopo di analizzare i processi mentali sottostanti (Boateng *et al.*, 2018). In particolare, si può procedere a un'analisi di diversi tipi di informazioni, come considerazioni dei soggetti sul modo in cui sono state costruite le risposte, sull'interpretazione personale delle domande, su eventuali difficoltà riscontrate nel comprendere uno specifico *item* o nel fornire la risposta corrispondente.

La scelta dei campioni a cui somministrare gli stimoli in fase di valutazione dovrebbe portare, secondo Beatty e Willis (2007), alla selezione di persone che coprono una varietà di situazioni rilevanti per il costrutto indagato e che rispecchiano possibilmente una qualche misura di varietà a livello demografico. Per quanto riguarda la numerosità dei partecipanti, gli autori suggeriscono di effettuare alcuni cicli di domande composti da gruppi di 5-15 persone, procedendo progressivamente alla revisione delle eventuali problematiche presenti.

Grazie alla tecnica di *pre-testing* è quindi possibile eliminare dall'insieme degli *item* gli stimoli che non rientrano nel dominio di costrutto dello studio o che non sono formulati in modo adeguato; questi ultimi possono essere eventualmente modificati per migliorarne la forma e renderli così maggiormente accessibili ai partecipanti (Boateng *et al.*, 2018). L'obiettivo finale è quello di stabilire se gli *item* proposti sono in grado di generare l'informazione di interesse del ricercatore con la certezza che il soggetto comprenda correttamente quello che gli viene chiesto.

2.2. Studio pilota e selezione del campione

Studio pilota

Dopo una prima rimozione delle domande meno adatte per lo studio della variabile latente di interesse, si può proseguire con una prima somministrazione della scala, con l'intento di raccogliere un ampio numero di informazioni da utilizzare nelle fasi seguenti della valutazione dello strumento: gli stimoli, ancora non definitivi, vengono somministrati a un campione rappresentativo della popolazione *target* attraverso due possibili modalità.

Una possibilità di somministrazione comporta la raccolta delle risposte tramite carta e penna/matita (*Paper and Pen/Pencil Interviewing*, PAPI*)* (Boateng *et al.*, 2018). È una procedura economicamente abbastanza impegnativa se ad essere coinvolti sono campioni

numerosi e rischia inoltre di produrre errori di diverso tipo, ma permette di raccogliere dati in qualsiasi contesto.

L'altra modalità di raccolta dei dati prevede l'utilizzo di dispositivi elettronici (*Computer Assisted Personal Interviewing*, CAPI) (Boateng *et al.* 2018). Questa tipologia è utile per riunire i dati in modo veloce ed economico e si rivela uno strumento più inclusivo, in quanto gli stimoli proposti possono essere registrati e ascoltati, riducendo anche la possibile presenza di errori. Per queste motivazioni, Boateng e colleghi (2018) raccomandano la somministrazione attraverso questa seconda modalità laddove vi sia possibilità di scelta.

Selezione del campione

Un passaggio chiave nella creazione della scala riguarda la fase di campionamento, che tratta la selezione della numerosità del campione e della tipologia di dati da utilizzare per la validazione dello strumento. Entrambe queste decisioni sono importanti, perché la disponibilità di dati che ne deriva garantisce la successiva valutazione della scala (Boateng *et al.* 2018).

Il campione scelto dovrebbe potenzialmente essere in grado di riflettere il *range* della popolazione per cui la scala è stata progettata (Boateng *et al.* 2018; MacKenzie *et al.*, 2011): ad esempio, Clark e Watson (1995) evidenziano l'utilità di usare un campione di pazienti anziché un campione generico nel caso in cui il test sia sviluppato per un utilizzo in contesto clinico. È inoltre importante definire dei criteri di inclusione ed esclusione nel campione, in quanto l'attendibilità delle risposte è influenzata dalle persone selezionate (Chiorri, 2023).

La numerosità specifica del campione è ancora oggetto di dibattito e le opinioni tra i vari autori sono discordanti tra loro. In generale, si raccomanda un minimo di dieci rispondenti per ciascun *item*, o una numerosità totale compresa almeno tra 200-300 persone (Boateng *et al.*, 2018). Tuttavia, è importante tenere in considerazione che, sebbene l'utilizzo di un grande campione sia sempre la scelta migliore, nel contesto pratico la disponibilità di risorse non è sempre compatibile con numerosità elevate.

La scelta della tipologia di dati da utilizzare impatta su diversi aspetti dello sviluppo del test. Il requisito minimo richiesto è che le risposte siano raccolte durante almeno un momento specifico nel tempo (*cross-sectional*) (Boateng *et al.*, 2018). Tuttavia, per essere

in grado di testare efficacemente l'affidabilità e la validità della scala, si rivela necessario un ulteriore insieme di dati di un campione indipendente oppure una nuova misurazione effettuata con lo stesso strumento in un secondo momento (*longitudinal*). In particolare, la valutazione longitudinale è importante in presenza di costrutti che si ipotizza non essere stabili nel tempo (MacKenzie *et al.*, 2011).

2.3. Analisi formale degli *item*

A questo punto dello sviluppo della scala, l'iniziale insieme di *item* generato è stato in parte diminuito grazie alle prime somministrazioni e alle valutazioni compiute. La forma finale del test, però, si ottiene anche grazie a specifiche procedure di analisi di riduzione degli *item*: attraverso questo procedimento, i ricercatori sono in grado di selezionare solo gli stimoli che dimostrano di essere appropriati dal punto di vista psicometrico, rimuovendo quelli meno relazionati al costrutto indagato (Boateng *et al.*, 2018).

Prima di procedere definitivamente all'eliminazione di un *item* che presenta una bassa adeguatezza statistica, tuttavia, è importante riconoscere il ruolo dello stimolo nella sua capacità di catturare il costrutto latente indagato: ad esempio, se l'*item* possiede un alto grado di validità di contenuto, è bene provare a risolvere le caratteristiche psicometriche inadatte che presenta (Chiorri, 2023).

La scelta della modalità specifica di analisi degli *item* viene compiuta sulla base della tipologia di test che si sta considerando: i test di prestazione massima utilizzano spesso come riferimenti l'indice di difficoltà e l'indice di discriminatività; i test di prestazione tipica considerano generalmente il numero di casi mancanti, la forma della distribuzione e le statistiche descrittive (Chiorri, 2023).

Indice di difficoltà dell'*item*

L'indice di difficoltà dell'*item* (Boateng *et al.*, 2018) ha lo scopo di determinare la proporzione di risposte corrette a uno stimolo. Un punteggio alto dell'indice sancisce la semplicità di risposta a una specifico *item*, indicando che la maggior parte delle persone è in grado di rispondere correttamente; viceversa, a un punteggio basso dell'indice di difficoltà corrisponde un *item* con un livello di difficoltà maggiore (Boateng *et al.*, 2018).

L'utilità di questo indice è dovuta al fatto che, sulla base dell'obiettivo del ricercatore, è possibile creare appositamente degli *item* con livello di difficoltà variabile che

permettono di distinguere diversi livelli di abilità posseduti da un individuo (Chiorri, 2023). L'indice di difficoltà permette anche di sviluppare *item* specifici per campioni o popolazioni selezionate (Hambleton & Jones, 1993).

La difficoltà degli stimoli, secondo Chiorri (2023), dovrebbe essere stabilita anche tenendo conto del numero di opzioni di risposta disponibili, della correlazione tra gli *item* e dallo scopo specifico dello strumento.

Indice di discriminatività

Un altro indice conosciuto in campo psicologico è l'indice di discriminatività, che identifica il grado in cui un *item* è in grado di differenziare soggetti che possiedono una diversa quantità del costrutto indagato. Grazie a questa misura emerge la differenza delle risposte tra un gruppo con punteggi bassi, in cui la probabilità di rispondere correttamente allo stimolo è bassa e che possiede quindi una quantità minore di costrutto, e un gruppo con punteggi alti, in cui la caratteristica indagata è maggiormente presente (Boateng *et al.*, 2018; Chiorri, 2023). Gli *item* con un alto indice di discriminatività sono in grado di compiere meglio questa differenziazione (DeMars, 2010).

L'utilizzo di un indice di discriminatività permette di identificare *item* in grado di discriminare positivamente, ovvero distinguere correttamente tra persone con alta o bassa presenza del costrutto; *item* che discriminano negativamente e che risultano pertanto progettati in modo non adeguato, in quanto le risposte danno esiti contrari a quelli corretti; *item* che non discriminano. Mentre le domande che operano una discriminazione adeguata possono generalmente continuare a far parte dell'insieme degli *item*, negli altri due casi è importante procedere a una modifica o completa cancellazione degli stessi (Brennan, 1973).

Casi mancanti, forma della distribuzione e statistiche descrittive

In situazioni in cui la risposta a uno stimolo non viene fornita o non rientra nelle possibili opzioni, ci si trova di fronte a casi mancanti. Le tipologie di casi mancanti possono essere di diversa natura (Acock, 2005): il soggetto può rifiutarsi di fornire una risposta, può essere indeciso sulla stessa e per questo motivo scegliere di non rispondere o può dimenticare accidentalmente di fornire una risposta; a volte, invece, la risposta è stata data, ma questa informazione non risulta presente all'interno del database completo.

Comprendere la causa di assenza di un dato permette di stabilire come gestire questa mancanza: generalmente, si prova a procedere attraverso tecniche che riescono a fronteggiare questo problema; laddove questo non è possibile, la rimozione della domanda è consigliata e preferibile (Boateng *et al*., 2018).

La forma della distribuzione riguarda la variabilità delle risposte ottenute per ogni *item*. Nella maggior parte dei casi, quando si lavora con le caratteristiche psicologiche ci si aspetta che l'insieme delle risposte assuma una distribuzione definita normale (Chiorri, 2023). Laddove la distribuzione possieda una forma diversa, è importante considerare il contenuto e la forma dello stimolo, le caratteristiche del campione di rispondenti e la relazione tra i due (Chiorri, 2023).

Nell'osservazione delle statistiche descrittive si tiene anche conto dei punteggi minimo e massimo di un *item*, che dovrebbero risultare scelti come risposta almeno una volta, e di alcuni indici di tendenza centrale come la mediana e la media, che identificano la probabilità di selezione delle risposte nei test di prestazione tipica.

2.4. Assegnazione del punteggio agli *item*

Le risposte agli *item* che compongono la scala psicologica riflettono la presenza del costrutto nei soggetti. Per determinare l'esatta quantità della variabile latente, però, gli stimoli devono essere sottoposti a una procedura di *scaling*, attraverso la quale ogni risposta viene quantificata attraverso l'assegnazione di un punteggio. Il punteggio finale del test è generalmente dato dalla somma dei singoli punteggi di ogni *item* (Chiorri, 2023).

Nell'assegnazione dei punteggi, lo *scaling* si occupa anche di individuare la relazione tra il costrutto latente e i suoi indicatori. La correlazione tra i punteggi degli *item* è generalmente spiegata da un fattore generale (la variabile latente) e da fattori specifici che influenzano ciascuna variabile osservata (Chiorri, 2023).

Grazie allo *scaling* è possibile utilizzare i punteggi ottenuti per successive analisi dello strumento (Boateng *et al.*, 2018).

2.5. Analisi fattoriale esplorativa

Con il termine *analisi fattoriale* si descrive un insieme di tecniche psicometriche che hanno l'obiettivo di identificare il numero minimo di dimensioni latenti (fattori) in grado di descrivere le relazioni che si osservano fra i punteggi degli *item* (Chiorri, 2023).

Attraverso l'analisi fattoriale è possibile scoprire se gli stimoli del test creato misurano altre caratteristiche psicologiche latenti oltre a quella di interesse del ricercatore. La presenza di ulteriori fattori può essere infatti in grado di spiegare la correlazione residua presente fra gli *item* (Chiorri, 2023).

Questa procedura è utile nello sviluppo e nella validazione di una scala psicologica, in quanto aiuta a determinare la struttura interna del test e a confermare che gli *item* misurano effettivamente i costrutti previsti (Flora & Flake, 2017), oltre a contribuire, eventualmente, a un'ulteriore scrematura degli stimoli (Boateng *et al.*, 2018).

All'interno degli *item* di un test si nota generalmente che gli stimoli che presentano un contenuto simile tendono a possedere punteggi che correlano maggiormente tra loro (Chiorri, 2023). In base al modo in cui gli *item* formano queste correlazioni, è possibile ricondurre un grande numero di variabili a un insieme ridotto di fattori, o domini, che meglio riflettono certi gruppi di *item*.

Il processo di estrazione dei fattori prende il nome di analisi fattoriale esplorativa (*Exploratory Factor* Analysis, EFA): l'obiettivo della EFA è quello di individuare le dimensioni latenti in grado di spiegare le correlazioni tra le variabili osservate di un test (Chiorri, 2023).

La EFA permette di testare diversi modelli e vedere quali risultano essere maggiormente adeguati agli obiettivi del ricercatore e alla scala costruita. Tuttavia, in questa fase i modelli propongono solamente una struttura ipotetica della scala, che necessita pertanto di essere ulteriormente indagata e confermata (Boateng *et al.*, 2018).

2.6. Test di affidabilità

Le misure di affidabilità, o attendibilità, (Boateng *et al.*, 2018) sono essenziali nel processo di valutazione di un test, poiché rappresentano un indice che esplicita la precisione dello strumento nel misurare una specifica variabile psicologica (Chiorri, 2023). Nello specifico, l'affidabilità indica la stabilità esibita da una misurazione quando questa viene ripetuta nelle stesse condizioni (Porta *et al.*, 2014). L'indice di affidabilità è inversamente proporzionale all'errore di misurazione, che è intrinseco alla misurazione stessa (Chiorri, 2023).

L'attendibilità di una scala psicologica considera le possibili fonti di errore di un test, che possono riguardare lo strumento stesso, la persona che lo somministra, il contesto di

svolgimento e i soggetti rispondenti (Chiorri, 2023). Alcuni fattori che influenzano l'affidabilità sono, ad esempio, la lunghezza del test e l'intervallo di tempo tra due somministrazioni della scala.

Nella valutazione dell'affidabilità di un test si considerano spesso il grado di accordo tra gli osservatori, la stabilità nel tempo e la coerenza interna degli stimoli. Tra i metodi statistici che si occupano di indagare l'affidabilità, i più conosciuti e utilizzati sono l'alpha di Cronbach (Cronbach, 1951), insieme alla correlazione *inter-item* e *item*-totale, e l'affidabilità *test-retest* (Raykov & Marcoulides, 2011).

Alpha di Cronbach

Il coefficiente alpha di Cronbach è uno degli indici che valutano la coerenza interna di una scala: questo significa che a essere indagato è il modo in cui gli *item* che la compongono co-variano insieme (Cronbach, 1951) in modo stabile. Se gli stimoli di un test sono altamente correlati tra loro, ci si aspetta infatti che misurino lo stesso costrutto (Tavakol & Dennick, 2011). Variazioni nel costrutto portano a una variazione nelle risposte, che cambiano nella stessa direzione: se una persona ottiene un punteggio finale alto, probabilmente i suoi *item* presentano tutti un punteggio alto; allo stesso modo, risposte con punteggi bassi dovrebbero produrre un basso punteggio totale (Chiorri, 2023).

L'alpha di Cronbach richiede normalmente la presenza di un costrutto unidimensionale per poter fornire una stima corretta della coerenza interna di una scala (Tavalok & Dennick, 2011): in caso di multidimensionalità del costrutto latente, infatti, il rischio è quello di una sottostima dell'affidabilità del test a causa della minore correlazione degli *item*.

Un errore che talvolta viene commesso nelle pratiche di misurazione è l'utilizzo dell'alpha di Cronbach come misura di omogeneità (Tavakol & Dennick, 2011). Quest'ultima fa riferimento all'unidimensionalità del costrutto, che viene generalmente indagata da procedure di analisi fattoriale e dimensionalità. Tavakol e Dennick (2011) evidenziano come la consistenza interna, che specifica invece la correlazione tra gli *item* all'interno di un test, sia una caratteristica necessaria, ma non sufficiente per stabilire l'omogeneità di una scala. Gli autori notano inoltre che talvolta il coefficiente alpha di Cronbach, in presenza di un costrutto multidimensionale, non mostra un valore inferiore

(Tavakol & Dennick, 2011). Sulla base di questa considerazione, pertanto, l'alpha di Cronbach non potrebbe comunque essere in grado di valutare l'unidimensionalità del costrutto.

Un'altra caratteristica di questo coefficiente riguarda la sensibilità alla correlazione degli *item* all'interno di un test (Tavakol & Dennick, 2011): una numerosità minima degli stimoli comporta un valore più basso dell'alpha di Cronbach; attraverso l'aggiunta di ulteriori domande alla scala è possibile aumentare le correlazioni tra gli *item* e, di conseguenza, anche il valore del coefficiente. È altrettanto vero, però, che in questo modo risulta molto semplice gonfiare artificialmente il risultato. Un valore alto del coefficiente, dunque, non garantisce necessariamente un elevato valore di consistenza interna.

L'utilizzo del coefficiente alpha di Cronbach nella valutazione dell'affidabilità di una scala psicologica è quindi uno strumento utile per valutare il grado di associazione tra gli *item* che la compongono, ma dovrebbe essere accompagnato da altri approcci di affidabilità che appoggiano e sostengono il valore riscontrato da questa misura.

Correlazione *inter-item* e *item*-totale

Nella valutazione della coerenza interna di uno strumento si utilizzano anche altri indici che si riferiscono ai singoli *item*, come la correlazione *inter-item* e la correlazione *item*-totale corretta.

La correlazione *inter-item* analizza il grado in cui i punteggi di una specifica domanda dell'insieme di *item* sono connessi ai punteggi di tutti gli altri *item* del test; questa correlazione considera anche la misura in cui gli *item* valutano lo stesso contenuto e permette quindi di identificare domande ridondanti (Piedmont, 2014).

La correlazione *item*-totale considera la relazione tra ciascun *item* e il punteggio totale degli *item* della scala (Boateng *et al.*, 2018). In altre parole, misura il grado in cui ogni *item* riflette l'intera scala psicologica. A essere utilizzata è soprattutto la versione corretta di questo indice, in quanto esclude dall'insieme degli *item* di confronto la domanda con cui si sta calcolando la correlazione (DeVellis, 1991).

Affidabilità test-*retest*

L'affidabilità test-*retest* è un altro tipo di indice di affidabilità (Boateng *et al.*, 2018), che permette di stabilire il grado in cui il punteggio di un individuo a un test è coerente nel corso del tempo; per individuarlo, la scala psicologica viene somministrata due volte e in seguito si procede a calcolare la correlazione tra le due misure ottenute (Chiorri, 2023). Grazie a questa stima è possibile osservare la stabilità di uno strumento nel corso di diverse misurazioni effettuate su uno stesso soggetto in momenti diversi e possedere la certezza che eventuali differenze di punteggio sono dovute a un cambiamento nell'individuo e non a uno strumento di misurazione non adeguato (Aldridge *et al.*, 2017).

La valutazione dell'affidabilità *test-retest* è soggetta a influenze di tipo temporale (Chiorri, 2010): affinché una misurazione possa essere replicata adeguatamente, deve trascorrere un giusto intervallo di tempo, che non deve essere né troppo breve, perché le persone potrebbero fornire le risposte sulla base dei ricordi della prima somministrazione, né troppo ampio, in quanto potrebbero verificarsi cambiamenti nei risultati dovuti a cause situazionali o personali.

Un altro punto importante che riguarda l'affidabilità *test-retest* è costituito dalla tipologia di costrutto su cui viene calcolata: l'utilizzo di questa misura si presta a caratteristiche psicologiche stabili che difficilmente sono soggette a modifiche nel corso del tempo (Chiorri, 2023).

3. Terza fase: validità esterna, invarianza di misura e sviluppo di norme

L'affidabilità di un test non garantisce la validità dello strumento; la scala psicologica, che ha raggiunto a questo punto una versione definitiva, deve essere testata per verificarne la validità. La validità di un test, che inizia già nelle prime fasi di sviluppo dello strumento, riguarda, in questa fase finale di validazione, la validità esterna: questa prevede il confronto con altre misure e costrutti simili a quello indagato e ha come obiettivo quello di confermare che il test misuri la caratteristica psicologica per cui è stato sviluppato.

Il processo di validazione di un test considera anche l'invarianza di misura in presenza di campioni o contesti alternativi a quelli originariamente pensati per lo strumento.

Infine, un test ha bisogno di un insieme di norme che consentono l'interpretazione dei punteggi ottenuti.

3.1. Test di validità esterna

La validità indica che un test misura effettivamente il costrutto per cui è stato progettato e che si propone di valutare (Raykov & Marcoulides, 2011). Si riferisce, nello specifico, alla misura in cui i punteggi ottenuti da un test sono correlati con le misure del fenomeno che la scala intende misurare; grazie a questa valutazione, è possibile stabilire se gli indicatori del test riescono a catturare accuratamente la caratteristica psicologica a cui si riferiscono (MacKenzie *et al.*, 2011).

Un test valido, secondo Chiorri (2023), è un test che è capace di prevedere, in qualche misura, il comportamento osservabile dei soggetti e che permette di compiere inferenze appropriate sulla base dei punteggi ottenuti.

La valutazione della validità comincia in realtà con l'inizio stesso della creazione di una scala psicologica, partendo da una chiara definizione del costrutto, passando per la validità di contenuto e terminando con la validità esterna dello strumento (Boateng *et al.*, 2018).

La validità esterna, nello specifico, fa riferimento alla "misura in cui le inferenze tratte dal campione di un dato studio si applicano a una popolazione più ampia o ad altre popolazioni *target*" (Findley *et al.*, 2021, p. 366). Ad essere testati sono due tipi di validità in particolare: la validità di criterio e la validità di costrutto.

Validità di criterio

La validità di criterio illustra il grado in cui esiste un'associazione tra il punteggio dato da una scala e il risultato di un'altra misura di particolare rilevanza a essa correlata, a cui ci si riferisce con il nome di criterio (Raykov & Marcoulides, 2011). La validità di criterio si scompone ulteriormente in due dimensioni: la validità concorrente e la validità predittiva.

La validità concorrente considera una misura o un parametro esterno che vengono raccolti nel momento stesso della somministrazione della scala o a distanza di un breve intervallo di tempo e osserva la forza di correlazione tra questa misura e il punteggio ottenuto dal test (Raykov & Marcoulides, 2011).

Un problema che talvolta si riscontra nell'indagare la validità concorrente risiede nella mancanza di un adeguato criterio di confronto al momento della somministrazione,

motivo che spesso spinge i ricercatori all'omissione di questa tipologia di validità (Boateng *et al.*, 2018).

La validità predittiva considera il confronto con un criterio esterno che viene rilevato a distanza di tempo e verifica il modo in cui la misura ottenuta con il test è in grado di prevedere un altro risultato successivo raccolto con uno strumento diverso, a cui la prima misura risulta essere correlata (Fowler, 1995).

In entrambi i casi, in seguito alla raccolta delle due misure si procede ad analizzare il valore della correlazione tra le due misurazioni compiute (Boateng *et al.*, 2018). Per la valutazione della validità di criterio si possono utilizzare, ad esempio, gli indici di correlazione (come l'indice di Pearson) (Chiorri, 2023).

La scelta del criterio da usare richiede che questo parametro esterno sia appropriato sia sul piano teorico sia sul piano della sua misurazione: nel primo caso, il ricercatore deve considerare lo scopo del test in relazione al criterio da selezionare; nel secondo caso, bisogna essere certi che il criterio sia stato rilevato in modo accurato (Chiorri, 2023).

Validità di costrutto

La validità di costrutto costituisce, secondo alcuni ricercatori, una delle misure più importanti per testare la validità di una scala psicologica, al punto da ritenerla una stima di validità generale. Con la validità di costrutto si cattura la capacità di uno strumento di misurare effettivamente la caratteristica psicologica che vuole indagare rispetto alla teoria del costrutto che misura (Raykov & Marcoulides, 2011; Chiorri, 2023). La validità di costrutto considera l'appropriatezza delle inferenze su quello che misura il test sulla base dei punteggi ottenuti (Chiorri, 2023).

Il concetto di validità di costrutto è legato al processo di validazione del costrutto: questa seconda procedura, come sottolineano Flake e colleghi (2017), è costituita da un processo "aperto" e in continuo stato di valutazione che coinvolge tutte le fasi di sviluppo di una scala psicologica. La valutazione specifica della validità di costrutto rientra nell'ultima parte di analisi di un test e permette di stabilire se il lavoro svolto sulla scala nelle fasi precedenti è stato eseguito in modo corretto (Flake *et al.*, 2017).

Per testare la validità di costrutto si operano dei confronti con altri criteri e costrutti relazionati a quello indagato e si osserva l'associazione con la scala in fase di validazione

(Raykov & Marcoulides, 2011). Attraverso questi confronti si esaminano due diverse componenti della validità di costrutto: la validità convergente e la validità discriminante.

Mediante la valutazione della validità convergente si verifica il grado di associazione dei risultati di un test con le misure di altri costrutti che si ipotizza essere parte dello stesso quadro teorico della caratteristica psicologica di interesse, e dunque correlate con la scala valutata (Raykov & Marcoulides, 2011). In base alla forza della connessione tra queste misure, secondo Churchill (1979), è possibile capire se la scala creata mostra un'associazione elevata con altre variabili in grado di catturare lo stesso costrutto del test principale: se questo non accade, la validità convergente non è adeguata.

La validità discriminante garantisce il grado di esclusività di una misura rispetto ad altri costrutti (Churchill, 1979): attraverso questo tipo di validità si può confermare che il punteggio ottenuto dal test, se confrontato con valori di altre manifestazioni comportamentali che fanno riferimento a costrutti diversi, mostra effettivamente una bassa associazione con questi (Raykov & Marcoulides, 2011). Correlazioni troppo elevate risultano problematiche, in quanto indice di somiglianza con il costrutto indagato.

Per esaminare la validità di costrutto si utilizzano principalmente metodi di analisi fattoriale confermativa (*Confirmatory Factor Analysis*, CFA). La CFA è una tecnica statistica che, a differenza della EFA, si occupa di valutare modelli fattoriali già ipotizzati in precedenza; questa pratica, quindi, verifica la relazione individuata tra i fattori latenti e le variabili osservate (Chiorri, 2023).

Tra le tecniche di analisi fattoriale confermativa rientrano alcuni modelli statistici multivariati come i modelli di equazioni strutturali (*Structural Equation Modeling*, SEM), che permettono di valutare simultaneamente le relazioni tra il test oggetto di studio e una molteplicità di altri test o altre variabili ritenute rilevanti dal punto di visto teorico, e i modelli di Rasch, i quali permettono di studiare in profondità le proprietà dei singoli *item* di un test (Chiorri, 2023).

3.2. Invarianza di misura

Attraverso l'invarianza di misura è possibile stabilire il grado in cui le proprietà psicometriche degli indicatori osservati sono generalizzabili a gruppi diversi nel corso del tempo o in un differente contesto (Sideridis *et al*., 2015).

Qualora il ricercatore decidesse di utilizzare un test in campioni o in contesti diversi da quelli per cui lo strumento è stato progettato, Luong e Flake (2023) enfatizzano l'importanza di valutare l'invarianza delle misure per verificare l'adeguatezza dello strumento in queste possibili nuove situazioni. Le procedure statistiche che si possono usare in questi casi rientrano nelle analisi fattoriali di tipo confermativo (Luong & Flake, 2023).

3.3. Sviluppo di norme

L'ultima fase di sviluppo e validazione di un test consiste nello sviluppo di norme che facilitino l'interpretazione dei punteggi ottenuti dalla misurazione con la scala (MacKenzie *et al.*, 2011). Spector (1992) evidenzia l'importanza di possedere una guida per l'interpretazione dei risultati: il significato di un punteggio, infatti, si può ritrovare solo se questo è relazionato a una cornice contestuale di riferimento.

Queste norme sono costituite dalle statistiche descrittive dei punteggi del test che vengono calcolate su uno o più campioni normativi (Chiorri, 2023). Oltre all'individuazione di campioni rappresentativi, le norme dovrebbero considerare valori come medie e deviazioni standard dei punteggi, nonché la forma della loro distribuzione (MacKenzie *et al.*, 2011). Il processo di raccolta di questi dati e la selezione di gruppi adeguati, tuttavia, non risulta semplice, e a questo si aggiunge la consapevolezza che le norme possono essere soggette a cambiamenti nel corso del tempo (MacKenzie *et al.*, 2011).

Un'ulteriore questione che richiede particolare attenzione riguarda la numerosità dei campioni normativi: il limite minimo di numerosità campionaria richiesto viene stabilito sulla base della numerosità della popolazione per cui il ricercatore intende sviluppare le norme della scala (MacKenzie *et al.*, 2011).

Nonostante queste difficoltà, MacKenzie e colleghi (2011) sottolineano l'importanza di possedere informazioni sulla distribuzione dei punteggi in campioni diversi e raccomandano ai ricercatori di dedicare uno spazio allo sviluppo di queste norme.

DISCUSSIONE E CONCLUSIONI

Questo elaborato ha cercato di fornire delle linee guida basate sulla letteratura su come avvenga lo sviluppo di un test psicologico.

La prima parte si è focalizzata sul processo di misurazione in psicologia e sull'impatto che misure poco valide possono avere nel contesto di ricerca a livello applicativo. Nello specifico, è stata sottolineata in primo luogo la difficoltà legata alla misurazione di costrutti latenti come oggetto di studio, fattore che complica, riprendendo la definizione di Stevens (1964, p. 677), la procedura di assegnazione di un numero a un oggetto. Una caratteristica psicologica può essere rilevata grazie a strumenti in grado di catturare le sue manifestazioni comportamentali direttamente osservabili: i test psicologici assolvono questo compito raccogliendo informazioni sui costrutti che si occupano di misurare.

Un punto da considerare con attenzione riguarda la fase di operazionalizzazione del costrutto: l'individuazione dei comportamenti che riflettono il costrutto garantisce la possibilità di misurare la variabile latente in modo oggettivo. Come sottolinea Chiorri (2023), senza questi indicatori osservabili non sarebbe infatti possibile confrontare la presenza del costrutto in soggetti diversi.

Affinché i dati ottenuti dalle misurazioni siano riconosciuti come validi, è necessario verificare la validità degli strumenti di misurazione stessi (Flake *et al.*, 2017). È stata messa in luce la grande lacuna presente nell'attuale contesto di ricerca in psicologia in merito all'adeguatezza delle misure utilizzate, come hanno riportato diversi studi (Flake *et al.*, 2017; MacKenzie *et al.*, 2011) che mostrano situazioni di incompletezza o di totale omissione di prove a supporto dell'affidabilità e della validità di una scala.

È stata inoltre rivolta l'attenzione al problema della "crisi di replicabilità", che interessa molti ambiti di ricerca e che convoglia anche nelle pratiche di misurazione discutibili riportate da Flake e Fried (2020), le quali includono la mancanza delle prove di validità di misura discusse precedentemente.

Secondo quando evidenziato, al fine di ottenere misure valide è importante dedicarsi alla validazione degli strumenti che si utilizzano fornendo evidenze empiriche a supporto. La crisi di replicabilità in corso sta aumentando la consapevolezza della necessità di migliorare le pratiche metodologiche anche attraverso la promozione di maggiore trasparenza rispetto alla validità delle scale psicologiche utilizzate.

Questa esposizione dell'attuale contesto di misurazione in psicologia ha voluto essere una premessa al tema centrale dell'elaborato, costituito da una panoramica delle buone pratiche per lo sviluppo e la validazione di una scala psicologica. Il lavoro ha richiesto l'integrazione di diversi riferimenti presenti in letteratura (ad esempio, Boateng *et al.*; Flake *et al.*, 2017; MacKenzie *et al.*, 2011).

In questa seconda parte è stata fornita una descrizione delle tre fasi principali che sono alla base della costruzione di un test: lo sviluppo degli *item*, la creazione della scala e la sua affidabilità e la valutazione della validità esterna dello strumento. Ognuna di queste è stata affrontata nel dettaglio attraverso una suddivisione in ulteriori passaggi specifici, ciascuno dei quali svolge un compito preciso all'interno del processo di sviluppo di un test psicologico.

Quello che è emerso dalla prima fase è, in particolare, il ruolo chiave svolto da un'adeguata individuazione del dominio di costrutto: per misurare un costrutto latente è importante fornire dal principio una definizione chiara e completa di quello che si intende procedere a misurare. Questo step non dovrebbe mai essere lasciato da parte: errori in questo stadio iniziale hanno conseguenze su tutto il successivo sviluppo del test (MacKenzie *et al.*, 2003; Flake *et al.*, 2017), incidendo negativamente sul contenuto degli *item*, e dunque sulla validità interna dello strumento, e creando confusione in merito al legame con altri costrutti correlati, influenzando così anche il modello della struttura latente della scala. La letteratura attualmente disponibile non sembra tuttavia fornire ancora un quadro chiaro delle modalità per svolgere al meglio questa procedura (MacKenzie *et al*., 2011).

Anche la validità di contenuto ha un'importanza significativa nella costruzione di un test. La mancanza di prove a supporto di questo tipo di validità si può riscontrare nella presenza di elementi irrilevanti rispetto al costrutto indagato o nell'incapacità dello strumento di catturare tutte le sfaccettature della variabile latente. Questo si riflette in seguito sui risultati ottenuti, che non rappresentano l'effettiva presenza del costrutto nei soggetti e rischiano di avere conseguenze negative in un contesto applicativo. Un esempio a supporto di questo riguarda l'utilizzo dei test nel contesto diagnostico: in queste situazioni è necessario che lo strumento fornisca una misura il più possibile precisa, perché da questa possono dipendere alcune scelte di intervento.

Un'ulteriore considerazione può essere fatta sulla selezione del formato di risposta più appropriato a catturare il costrutto latente. Ad esempio, il formato dicotomico è una modalità di risposta semplice e immediata, ma non si presta a situazioni che intendono catturare sfumature di un comportamento come nel caso di scale di atteggiamento, in quanto richiede un giudizio assoluto. La scala Likert richiede invece di considerare il numero di punti che la costituiscono anche sulla base dei soggetti a cui lo strumento è indirizzato: i bambini, rispetto agli adulti, potrebbero non essere in grado di cogliere differenze minime tra le possibili risposte (Mellor & Moore, 2014).

La seconda fase della costruzione di un test ha illustrato il modo in cui si raggiunge una versione ridotta degli *item* più adatti, l'analisi del modello fattoriale e la stima dell'affidabilità della scala.

Tra le misure che stimano l'affidabilità di una scala, una delle più diffuse è quella dell'alpha di Cronbach. Questo indice, che valuta la coerenza interna di una scala, risente di molti limiti, tra cui la possibilità di essere utilizzato solo in presenza di costrutti unidimensionali e la sensibilità al numero di *item* che costituiscono la scala (Tavalok & Dennick, 2011); inoltre, questo coefficiente assume che ogni stimolo contribuisca nella stessa misura alla determinazione del punteggio finale del test, ipotesi raramente soddisfatta nel contesto pratico di misurazione (Chiorri, 2023).

La terza fase di sviluppo di un test ha evidenziato l'importanza della valutazione della validità esterna della scala psicologica creata, che conferma che lo strumento misura effettivamente il costrutto per cui è stato creato. La validità di costrutto, in particolare, identifica la capacità della scala di misurare il costrutto latente secondo quanto ipotizzato dal ricercatore nelle sue teorie e rappresenta una stima della validità generale del test. Il confronto che viene effettuato in questa fase con altre misure esterne correlate al costrutto richiede di prestare particolare attenzione alla scelta di un criterio di paragone che garantisca una buona accuratezza delle conclusioni che si possono trarre dal punteggio di correlazione delle due misure.

In generale, il lavoro con i costrutti psicologici richiede di considerare anche l'evoluzione degli strumenti di misurazione nel corso del tempo. Le scale e i test possono infatti avere bisogno di aggiornamenti o modifiche per continuare a garantire l'affidabilità e la validità delle misure in contesti e popolazioni diverse. La validazione di una scala

psicologica, come ricordano Flake e colleghi (2017), è una procedura sempre aperta che raccoglie continuamente nuove prove di evidenza della sua validità.

Per concludere, quindi, questo elaborato è stato scritto con l'obiettivo di delineare le buone pratiche che accompagnano lo sviluppo e la validazione di un test psicologico. Le procedure che seguono la creazione di una scala sono numerose: questo lavoro ne fornisce una descrizione generale e ordinata, ma certamente non completa. Inoltre, il processo di generazione e valutazione di un test è dispendioso e molto impegnativo e nello scenario pratico risulta difficile riuscire a svolgere rigorosamente tutte le fasi previste. Laddove vi siano difficoltà nel seguire una procedura completa di tutte le fasi di creazione di una scala, si dovrebbe compiere una valutazione che permetta di identificare i passaggi fondamentali da seguire sulla base dell'obiettivo dello studio e delle risorse disponibili (Boateng *et al.*, 2018).

Nella pratica psicologica, la misurazione è un requisito essenziale per poter rilevare e indagare i costrutti latenti oggetto di studio. Per garantire la qualità delle misurazioni effettuate è necessario confermare la validità dei dati raccolti: delle buone pratiche di misurazione richiedono infatti una presenza esaustiva delle informazioni sullo sviluppo e sulla validazione dei test utilizzati, con strumenti che dimostrano di essere in grado di catturare il costrutto di interesse in modo valido e affidabile.

L'importanza che rivestono le buone pratiche di misurazione nella ricerca in psicologia si configura dunque come un aspetto di rilevanza centrale che deve continuare a essere studiato e approfondito.

BIBLIOGRAFIA

Acock, A. C. (2005). Working with missing values. *Journal of Marriage and family*, *67*(4), 1012-1028. https://doi.org/10.1111/j.1741-3737.2005.00191.x

Aldridge, V. K., Dovey, T. M., & Wade, A. (2017). Assessing test-retest reliability of psychological measures: Persistent methodological problems. *European Psychologist*, 22(4), 207. https://doi.org/10.1027/1016-9040/a000298

Beatty, P. C., & Willis, G. B. (2007). Research synthesis: The practice of cognitive interviewing. *Public opinion quarterly*, 71(2), 287-311. https://doi.org/10.1093/poq/nfm006

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: a primer. *Frontiers in public health*, 6, 149. https://doi.org/10.3389/fpubh.2018.00149

*Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. *MIS quarterly*, 1-16. https://doi.org/10.2307/3250956

*Brennan, R. L. (1972). A generalized upper-lower item discrimination index. *Educational and Psychological Measurement*, 32(2), 289-303. https://doi.org/10.1177/001316447203200206

Chaplin, W. F., John, O. P., & Goldberg, L. R. (1988). Conceptions of states and traits: Dimensional attributes with ideals as prototypes. *Journal of Personality and Social Psychology*, 54(4), 541. https://psycnet.apa.org/doi/10.1037/0022-3514.54.4.541

Chiorri, C. (2023). *Teoria e tecnica psicometrica: costruire un test psicologico*. McGraw-Hill.

Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, 16(1), 64-73. https://doi.org/10.1177/002224377901600110

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, 7(3), 309–319. https://psycnet.apa.org/doi/10.1037/1040-3590.7.3.309

Cronbach, L.J. Coefficient alpha and the internal structure of tests. *Psychometrika* 16, 297–334 (1951). https://doi.org/10.1007/BF02310555

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, 52(4), 281. https://psycnet.apa.org/doi/10.1037/h0040957

DeMars, C. (2010). *Item response theory*. Oxford University Press.

*DeVellis, R. F. (1991). *Scale development: Theory and applications*. Sage publications.

Findley, M. G., Kikuta, K., & Denly, M. (2021). External validity. *Annual Review of Political Science*, 24(1), 365-393. https://doi.org/10.1146/annurev-polisci-041719-102556

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, 3(4), 456-465. https://doi.org/10.1177/2515245920952393

Flake, J. K., Pek, J., & Hehman, E. (2017). Construct validation in social and personality research: Current practice and recommendations. *Social Psychological and Personality Science*, 8(4), 370-378. https://doi.org/10.1177/1948550617693063

Flora, D. B., & Flake, J. K. (2017). The purpose and practice of exploratory and confirmatory factor analysis in psychological research: Decisions for scale development and validation. *Canadian Journal of Behavioural Science / Revue canadienne des sciences du comportement*, 49(2), 78–88. https://doi.org/10.1037/cbs0000069

*Fowler, F. J. (1995). *Improving survey questions: Design and evaluation*. Sage.

Fried, E. I., & Flake, J. K. (2018). Measurement matters. *APS Observer*, 31.

Hambleton, R. K., & Jones, R. W. (1993). An NCME Instructional Module on: Comparison of Classical Test Theory and Item Response Theory and Their Applications to Test Development. *Educational Measurement: Issues and Practice*, 12(3), 38-47. https://doi.org/10.1111/j.1745-3992.1993.tb00543.x

Haynes, S. N., Richard, D., & Kubany, E. S. (1995). Content validity in psychological assessment: A functional approach to concepts and methods. *Psychological assessment*, 7(3), 238. https://psycnet.apa.org/doi/10.1037/1040-3590.7.3.238

Hinkin, T. R. (1995). A review of scale development practices in the study of organizations. *Journal of management*, 21(5), 967-988. https://doi.org/10.1177/014920639502100509

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of educational measurement*, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

*Kline, P. (2013). *Handbook of psychological testing*. Routledge.

Luong, R., & Flake, J. K. (2023). Measurement invariance testing using confirmatory factor analysis and alignment optimization: A tutorial for transparent analysis planning and reporting. *Psychological Methods, 28*(4), 905–924. https://doi.org/10.1037/met0000441

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the academy of marketing science*, 31(3), 323-326. https://doi.org/10.1177/0092070303031003011

MacKenzie, S. B., Podsakoff, P. M., & Jarvis, C. B. (2005). The problem of measurement model misspecification in behavioral and organizational research and some recommended solutions. *Journal of applied psychology*, 90(4), 710. https://psycnet.apa.org/doi/10.1037/0021-9010.90.4.710

MacKenzie, S. B., Podsakoff, P. M., & Podsakoff, N. P. (2011). Construct measurement and validation procedures in MIS and behavioral research: Integrating new and existing techniques. *MIS quarterly*, 293-334. https://doi.org/10.2307/23044045

*McCoach, D. B., Gable, R. K., & Madura, J. P. (2013). *Instrument development in the affective domain* (Vol. 10, pp. 978-1). New York, NY: Springer.

Mellor, D., & Moore, K. A. (2014). The use of Likert scales with children. *Journal of pediatric psychology*, *39*(3), 369-379. https://doi.org/10.1093/jpepsy/jst079

*Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American psychologist*, 50(9), 741. https://psycnet.apa.org/doi/10.1037/0003-066X.50.9.741

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

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. https://doi.org/10.1126/science.aac4716

Piedmont, R.L. (2014). Inter-item Correlations. In: Michalos, A.C. (eds) *Encyclopedia of Quality of Life and Well-Being Research*. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-0753-5_1493

Porta, M. S., Greenland, S., Hernán, M., dos Santos Silva, I., & Last, J. M. (Eds.). (2014). *A dictionary of epidemiology*. Oxford University Press, USA.

*Raykov, T., & Marcoulides, G. A. (2011). *Introduction to psychometric theory*. Routledge.

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of personality assessment*, 92(6), 544-559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Morizot, J., & Hays, R. D. (2007). The role of the bifactor model in resolving dimensionality issues in health outcomes measures. *Quality of Life Research*, 16, 19-31. https://doi.org/10.1007/s11136-007-9183-7

*Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management journal*, 43(6), 1248-1264. https://doi.org/10.5465/1556348

*Schinka, J. A., Velicer, W. F., & Weiner, I. B. (2013). *Handbook of psychology: Research methods in psychology*, Vol. 2. John Wiley & Sons, Inc.

Sideridis, G. D., Tsaousis, I., & Al-harbi, K. A. (2015). Multi-Population Invariance With Dichotomous Measures: Combining Multi-Group and MIMIC Methodologies in Evaluating the General Aptitude Test in the Arabic Language. *Journal of Psychoeducational Assessment*, 33(6), 568-584. https://doi.org/10.1177/0734282914567871

Spector, P. E. (1992). *Summated rating scale construction: An introduction* (Vol. 82). Sage.

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *International journal of medical education*, 2, 53. https://doi.org/10.5116%2Fijme.4dfb.8dfd

Urbina, S. (2014). *Essentials of psychological testing*. John Wiley & Sons.

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences*, 41, e120. https://doi.org/10.1017/S0140525X17001972

*opere non consultate direttamente