UN APPROCCIO INNOVATIVO AL *TESTING* PSICOPATOLOGICO: TALEIA PARTE I: VALIDITÀ DI CONTENUTO E SCALE DI CONTROLLO

LUCIA BONCORI[1], ALESSANDRA DE CORO[1], GIOVANNI CUOMO[2] E FRANCO LUCCHESE[1]

[1] *Sapienza Università di Roma,* [2] *Sanità della Polizia di Stato*

Riassunto. L'articolo descrive il metodo usato per costruire TALEIA-400A, un questionario per la valutazione dei disturbi clinici e di personalità secondo le linee-guida di DSM-IV e ICD-10. TALEIA è un acronimo da *Test for AxiaL Evaluation and Interview for Clinical, Personnel, and Guidance Applications*. In particolare vengono discusse le scale di controllo sotto aspetti teorici e operativi. Alcuni metodi di costruzione, innovativi nella sostanza o nei dettagli, potrebbero stimolare una discussione metodologica: l'ancoraggio della scala L alla desiderabilità sociale definita empiricamente dai valori condivisi da un grande campione, le tecniche adottate per ridurre al minimo il peso del fattore linguistico-culturale, l'integrazione di DSM-IV e ICD-10 con altri sistemi nosologici e la consultazione confermatoria di operatori clinici qualificati.

1. INTRODUZIONE

Le critiche rivolte ai test psicologici negli anni Settanta erano in gran parte motivate dall'uso «politicamente scorretto» dei test, ma si riferivano anche ad alcune ipersemplificazioni del processo di costruzione, di validazione e di standardizzazione dei test. Per quel che riguarda l'applicazione dei questionari in campo psicopatologico, in Italia gli echi delle critiche statunitensi arrivarono in un contesto culturale e professionale in cui i test di personalità erano stati da poco introdotti, dopo il «blocco» imposto dalla posizione ideologica gentiliana («lo spirito non si misura»), e non erano ancora oggetto di insegnamento e di ricerca sistematica in ambito accademico. Tali critiche vennero quindi indebitamente generalizzate, senza suscitare reazioni costruttive adeguate. Di fatto, molti psicologi preferirono rinunciare all'uso dei test (o addirittura della diagnosi, ritenuta indebito etichettamento) oppure, paradossalmente, rifugiarsi nell'uso di tecniche proiettive, trascurando le pesantissime critiche che in campo internazionale ne avevano già escluso l'utilizzazione come strumento di misura. A partire dagli anni Ottanta, con il diffondersi anche in Italia di un ritorno dell'attenzione alla diagnosi come indispensabile strumento di verifica delle indicazioni terapeutiche e prognostiche, si tornò all'uso di test ideati in ambiente anglosassone prima delle critiche stesse, di

solito soltanto tradotti in italiano e dotati di parametri calcolati su piccoli e mal descritti campioni italiani, senza nessuna effettiva verifica della validità in ambiente culturale italiano. D'altra parte, la costruzione ex novo e la verifica della validità di un test sono lavori di grosso impegno, che richiedono finanziamenti, lavoro di gruppo e competenze interdisciplinari ad alto livello:. Il lavoro presentato in questo articolo è il risultato della felice coesistenza, per alcuni anni, di queste rare circostanze. Nel 1991 il Ministero della Difesa, dopo aver utilizzato per alcuni anni come strumento di selezione psicopatologica in sede di visita di leva lo storico MMPI, decise di emettere un bando europeo, limitato a istituzioni universitarie, per la costruzione di un test rispondente alla nosografia internazionale corrente (ICD dell'Oms e DSM dell'Apa) e specificamente validato sulla popolazione italiana. La Direzione Generale della Leva aveva preso questa decisione (e stanziato i relativi fondi) in seguito a ripetute sollecitazioni del proprio personale tecnico (psicologi, medici, psichiatri), che segnalava difficoltà di comprensione dei quesiti da parte dei giovani meno scolarizzati e larga diffusione di manuali e di corsi per «truccare» i risultati del test. Era anche segnalata una problematica forense: in caso di ricorso contro la decisione medico-legale, la diagnosi doveva essere formulata in base alla nosografia corrente dell'ICD (9ª edizione, 1991), ma il medico che la firmava poteva chiamare a sostegno di questa diagnosi solo i risultati di un test che, ideato nel 1939, si riferiva a categorie completamente diverse. La gara di appalto fu vinta dal progetto del Dipartimento di Psicologia dell'Università di Roma «La Sapienza», che impegnò nella creazione del nuovo test ventuno fra psicologi e psichiatri[1] a cui affiancò professionisti prevalentemente appartenenti a strutture pubbliche[2]. Il test commissionato dal bando avrebbe avuto per contratto il copyright del Ministero della Difesa. Tuttavia molti dei partecipanti al Comitato Tecnico-Scientifico (d'ora in poi: CTS) del Dipartimento di Psicologia pensavano che un nuovo strumento oggettivo per la valutazione psicopatologica fosse una necessità e potesse anche essere l'occasione per risolvere alcuni problemi professionali, quali ad esempio la connessione fra diagnosi e impostazione della terapia. Il lavoro venne quindi impostato in modo da poter successivamente produrre anche forme parallele da validare suc-

[1] Psicologi: L. Boncori, C. Candelori, A. De Coro, F. Lucchese, A. Orsini, F. Ortu, P. Renzi, G. T. Scalisi; Psichiatri e psicologi: V. Lingiardi, S. Nicole, coadiuvati dai professionisti di altre strutture elencati successivamente.

[2] Psicologi: I. Bigari (Comune di Roma), F. Bonaiuto (privato: selezione del personale), F. Borrelli (Selezione P.S.), A. Ciaglia (privato: famiglia), G. Infantino (ASL RMC), P. Salotti (ASL VT), G. M. Sferrazza (privato: psicoterapeuta), D. Solfaroli Camillocci (privato: famiglia), S. Tomassini (EI), U. Ungaro (Sanità della P.S.); Psichiatri: I. Baldacci, G. Buffardi.

cessivamente e da utilizzare al di fuori dell'applicazione per cui veniva costruita la prima forma del questionario. Questo articolo riassume rapidamente le fasi iniziali della costruzione del test, evidenziandone le caratteristiche metodologicamente innovative; la seconda parte del lavoro documenta invece la validità della versione disponibile per l'uso «civile» del test (per la versione riservata al Ministero della Difesa v. Boncori, Baiocco, Barruffi e Comignani, 2001a, 2001b). In ambedue gli articoli, dove non sono dati riferimenti bibliografici ad altre pubblicazioni su TALEIA, gli studi presentati sono originali.

2. PROGETTAZIONE DI UN INSIEME DI QUESTIONARI COME FORME PARALLELE

Già nel 1991 la metodica di costruzione dei test basata sulle banche di item era nota e largamente praticata nelle applicazioni ai test di profitto e al mondo del lavoro, anche se non era stata utilizzata per i test di personalità (Buckley-Sharp e Harris, 1970; Gorth, Allen e Grayson, 1971; Millman e Arter, 1984; Vale, 1986; Van der Linden e Eggen, 1986; Wood e Skurnik, 1969). Il riferimento obbligatorio a una specifica nosografia, come l'insieme di ICD-10 e DSM-IV (fra l'emissione del bando e l'inizio dei lavori erano state pubblicate queste versioni, tuttora in uso), rendeva possibile l'impostazione di una banca-dati riferita a un insieme unico di indicatori (i «criteri diagnostici») delle due nosografie, in riferimento a ciascuno dei quali furono creati più item, da assegnare a caso alle diverse forme del questionario. La disponibilità di numerosi docenti universitari a collaborare per la stesura di un consistente numero di quesiti fu un'altra circostanza favorevole, che portò alla costituzione di una banca-dati di circa 2.000 item (1.952, per l'esattezza), gestita informaticamente con una delle prime versioni del programma FileMaker[3], in vista della creazione di forme del questionario parallele per contenuto.

3. FORME E SCALE DEI QUESTIONARI

Le forme parallele dei questionari originate dalla banca dati di partenza sono presentate schematicamente nella tab. 1. Il Difesa Test (Boncori 1999a, 1999b) è stato ottenuto per semplice eliminazione di 100 quesiti dalla forma D3. Le forme pubblicate o in via di pubblicazione (Difesa Test, T2000A, T2000B) si compongono di item completamente diversi l'una dall'altra, anche se paralleli per contenuto

[3] L'ideazione e la strutturazione della banca dati si devono a P. Renzi.

TAB. 1. *Forme parallele del questionario TALEIA*

Forme parallele	D 3	Difesa test	T2000-A	T2000-B	T2001-A	T2001-B
N. quesiti	400	300	400	400	300	300
Scale «cliniche» (tutte riferite a ICD-10 eDSM-IVR)	Asse I	Asse I	Asse I	Asse I	Asse I	Asse I
	1. S	1. S	1. S	1. S	1. S	1. S
	2. D	2. D	2. D	2. D	2. D	2. D
	3. M	3. M	3. M	3. M	3. M	3. M
	4. AA	4. AA	4. AA	4. AA	4. AA	4. AA
	5. FO	5. FO	5. FO	5. FO	5. FO	5. FO
	6. SOC	6. SOC	6. SOC	6. SOC	6. SOC	6. SOC
	7. AG	7. AG	7. AG	7. AG	7. AG	7. AG
	8. AL	8. AL	8. AL	8. AL	8. AL	8. AL
	Asse II	Asse II	Asse II	Asse II	Asse II	Asse II
	9. PP	9. PP	9. PP	9. PP	9. PP	9. PP
			10. PAS	10. PAS	10. PAS	10. PAS
			11. PB	11. PB	11. PB	11. PB
	10. PAS	10. PAS	12. PAS	12. PAS	10. PAS	10. PAS
	11. PB	11. PB	13. PB	13. PB	11. PB	11. PB
	12. PI	12. PI	14. PI	14. PI	12. PI	12. PI
	13. PN	13. PN	15. PN	15. PN	13. PN	13. PN
	14. PEV	14. PEV	16. PEV	16. PEV	14. PEV	14. PEV
	15. PD	15. PD	17. PD	17. PD	15. PD	15. PD
	16. POC	16. POC	18. POC	18. POC	16. POC	16. POC
Scale «di controllo»	17. L	17. L	18. L	18. L	17. L	17. L
	18. F	18. F	19. F	19. F	18. F	18. F
	19. INC	19. INC	20. INC	20. INC	19. INC	19. INC

Legenda: S = Schizofrenia; D = Depressione; M = Ipomania e mania; AA = Ansia acuta (attacchi di panico); FO = Fobiche, sindromi; SOC = Sindrome ossessivo-compulsiva; AG = Ansia generalizzata; AL = Disturbi da alterato comportamento alimentare; PP = Disturbo di personalità paranoide; PSK = Disturbo di personalità schizoide; PSKT = Disturbo di personalità schizotipico; PAS = Disturbo di personalità antisociale; PB = Disturbo di personalità borderline; PI = Disturbo di personalità istrionico; PN = Disturbo di personalità narcisistico; PEV = Disturbo di personalità evitante (ansioso); PD = Disturbo di personalità dipendente; POC = Disturbo di personalità anancastico (ossessivo-compulsivo); L = Tendenza a presentare un'immagine favorevole di sé (Desiderabilità sociale); F = Tendenza a presentare un'immagine sfavorevole di sé (Risposte infrequenti); INC = Incongruità variabile fra le risposte.

in quanto riferiti agli stessi indicatori e criteri diagnostici. La denominazione del test è un acronimo da *Test for AxiaL Evaluation and Interview for Clinical, Personnel, and Guidance Applications*. La prima forma disponibile per l'uso di operatori esterni al Ministero della Difesa è denominata «TALEIA-400-A», dove il suffisso sta per «forma A, 400 item» (Boncori, 2007a).

3.1. *Le scale «cliniche»*

I disturbi da assumere come riferimento per le scale «cliniche» (Asse I e Asse II) sono stati selezionati in base ai seguenti criteri: 1) presenza con frequenza non irrilevante nella popolazione in età 17-65 anni (fascia d'età indicata dal bando del Ministero della Difesa); 2) possibilità realistica di dare una valutazione dimensionale dei disturbi mediante le risposte a un questionario. La metodologia seguita per definire quali e quante scale inserire nel questionario è descritta più particolareggiatamente nel paragrafo sulla validità di contenuto. Per ogni scala, sono stati classificati a parte i criteri diagnostici che era preferibile sottoporre ad accertamento mediante colloquio. Il colloquio è stato sempre considerato come componente integrativa necessaria ai fini della diagnosi e/o della decisione medico-legale. Date le situazioni applicative a cui il test era destinato (selezione del personale, orientamento, utilizzazione in strutture sanitarie pubbliche), il colloquio doveva essere impostato come un'intervista strutturata, accompagnata da sussidi per la notazione e la classificazione delle risposte, tale da poter essere appresa e gestita con un minimo di tempo e d'impegno. Per queste caratteristiche, il colloquio integrato con i test TALEIA si differenzia nettamente da strumenti come il colloquio OPD (*Operationalisierte Psychodynamische Diagnostik*: Gruppo OPD, 2002) o la SWAP-200 (*Shedler-Westen Assessment Procedure*: Westen, Shedler e Lingiardi, 2003), che si collocano a un livello di specializzazione clinica più sofisticato. Per consentire agli operatori di fare il colloquio immediatamente dopo il test, ma dopo averne conosciuto i risultati e basandosi su quesiti integrativi «personalizzati» rispetto al singolo profilo, è stato necessario mettere a punto un software dedicato, semplice da usare ma sofisticato nella struttura, che consentisse da un lato l'immissione rapidissima delle risposte (le persone possono rispondere al computer, oppure su fogli di risposta – stampati dal computer stesso – che possono essere scannerizzati e acquisiti) e dall'altro la restituzione immediata dei risultati, in un foglio che contenesse, oltre al solito profilo grafico-numerico, anche i quesiti da porre a quel particolare soggetto. Nel caso di grandi numeri si può utilizzare il foglio di risposta per il lettore ottico e in questo caso la velocità di acquisizione dei risultati è condizionata dal numero delle macchine e degli operatori disponibili, ma consente comunque di condurre tutti i colloqui entro poche ore.

3.2. *Le scale di controllo*

La realizzazione delle scale di controllo è stata oggetto di particolare studio, in quanto gli Autori le ritengono non solo di fondamen-

tale importanza nelle applicazioni concorsuali e periziali, ma utili, quando sono presenti alcune tipologie di disturbi, anche nelle applicazioni cliniche e orientative. Un breve *excursus* storico è necessario per motivare le innovazioni introdotte a questo riguardo e per evidenziare le differenze di contenuto rispetto ad analoghe scale di altri test (per esempio MMPI-2 ed EPQ).

La maggior parte dei questionari di personalità contemporanei includono «scale di controllo», come sussidio per controllare l'effetto di distorsioni volontarie o involontarie nelle risposte, differenziandosi in questo dai test proiettivi, che lasciano la decisione sull'autenticità delle risposte al singolo operatore in fase di somministrazione o interpretazione.

Tra il 1950 e il 1965 vennero compiuti molti studi sugli stili di risposta, i cui risultati vennero sintetizzati come «il grande mito dello stile di risposta» da un'estensiva rassegna critica (Rorer, 1965). Sotto l'aspetto pratico, i metodologi accreditati inclusero in blocco gli stili di risposta tra le possibili fonti di invalidità del test, da evitare adottando una molteplicità di accorgimenti in varie fasi della costruzione del test (Nunnally, 1978). Venne anche proposto di distinguere fra «distorsioni nelle risposte» (*response bias*), artefatti della misurazione che alterano i valori medi delle risposte date da gruppi di persone e che non necessariamente dipendono da differenze individuali stabili, e «stili di risposta», che riguardano differenze individuali (Nunnally, 1978). Fra gli stili di risposta, Nunnally discute esplicitamente la desiderabilità sociale (a cui dedica la maggiore attenzione), la tendenza a indovinare quando si è in dubbio, la tendenza a rispondere «vero» (anche nei test cognitivi) senza una reale convinzione, la tendenza a dare/non dare risposte estreme, la tendenza a dare risposte poco comuni, socialmente devianti. Le scale di desiderabilità sociale utilizzate negli studi condotti dopo il 1970 sulle distorsioni classificabili nell'insieme come «faking good» hanno utilizzato scale costruite su modelli diversi: al modello originario di Edwards (Edwards, 1957) e a quello molto diffuso di Marlowe e Crowne (Crowne e Marlowe, 1960) se ne sono aggiunti vari altri, presi in considerazione da studi meta-analitici (Baer, Wetter e Berry, 1992), interessanti nonostante le difficoltà poste dal confronto di strumenti originati da modelli teorici diversi. Comunque, nel costruire le scale di controllo per TALEIA, è stata assunta come ragionevole la sintesi degli studi proposta nel manuale di Nunnally e Bernstein (1994, p. 383), secondo cui si possono distinguere quattro tipi di desiderabilità sociale:

1. specifica rispetto alle situazioni (distorsioni);
2. generalizzabile da una situazione all'altra, in quanto prodotto secondario di strategie consce (stili);
3. generalizzabile da una situazione all'altra, in quanto manifestazione inconscia di un tratto di personalità più ampio;
4. non meritevole di considerazione.

Gli autori di TALEIA hanno ritenuto di poter tenere sotto controllo la desiderabilità sociale del tipo 1 disponendo norme psicometriche specifiche per diverse situazioni di testing (il che è una novità in campo internazionale), di rinunciare a distinguere fra le tipologie 2 e 3, considerando che in base all'esperienza di numerosi colloqui condotti nell'ambito di attività sia professionali sia di ricerca dai membri del CTS in moltissimi casi la componente conscia e inconscia sono copresenti e che ai fini pratici è poco utile distinguere l'una dall'altra, e di lasciare che la componente di distorsione residua 4 fosse controllata nell'ambito degli studi sull'attendibilità, come porzione della «varianza erratica».

Quanto al modello da seguire per la costruzione di una «scala L», sono state recepite le critiche di Crowne e Marlowe alla scala L del MMPI (Crowne e Marlowe, 1960). Tenendo conto però anche di altri studi che hanno evidenziato come le distorsioni possano essere diverse in situazioni diverse (Paulhus e Reid, 1991; McFarland e Ryan, 2000; Baer e Miller, 2002) e che possono essere connesse con valori culturali ed etici e, più in genere, con modelli culturali (Novaga e Pedon, 1977), si è ritenuto di classificare i contenuti dei quesiti presenti nella banca-dati in riferimento a valori. I valori di più ampia condivisione culturale sono stati identificati in base alle risposte di campioni non-clinici, considerando operativamente che i valori più largamente condivisi fossero associati agli item caratterizzati da una più elevata frequenza percentuale di risposte in senso affermativo nei campioni di soggetti «normali». La lista così risultante è stata usata come base per la costruzione della «scala L» (Boncori, 2007b). Così facendo, la funzione che nelle scale costruite con la metodologia di Thurstone era demandata a un «gruppo di giudici», è stata attribuita a un campione nazionale composto da numerosi soggetti, presumibilmente più rappresentativo della popolazione per cui il test veniva costruito. La validità della procedura seguita si basa anche sul fatto che i quesiti inclusi nella scala L appartengono a una banca dati che comprende non solo item relativi alle scale dei disturbi mentali, ma anche quesiti relativi alle scale dei disturbi di personalità: queste ultime includono, a loro volta, numerosi indicatori che non risultano per sé dichiaratamente patologici quando sono presenti a livelli medi o bassi. Una scala di «desiderabilità sociale» basata sui valori consente di ovviare alla critica di identificare la desiderabilità sociale con l'assenza di patologia, rivolta da Crowne e Marlowe alla scala L del MMPI.

Come per le scale cliniche di TALEIA, anche per la scala L la risposta alfa (o risposta sintomatica) è collocata quasi con uguali proporzioni nella direzione «vero» e nella direzione «falso». Questo ci ha consentito di evitare la costruzione di una ulteriore scala di controllo, come la TRIN del MMPI-2, che corregga la distorsione imputabile

all'acquiescenza, implicita nel fatto che la scala L del MMPI-2 contiene esclusivamente item a cui la risposta alfa è «Vero».

La componente di distorsione «faking bad», che include tutti i «tentativi consci o inconsci di produrre schemi di risposta che raffigurano una sintomatologia esagerata rispetto a quella effettivamente vissuta da chi risponde» (Franke, 2002), con effetti che appaiono come il tentativo di dare un'impressione negativa irrealistica, ha ricevuto finora minore attenzione dalla ricerca. Tuttavia, l'incremento dell'uso dei test nelle applicazioni forensi sta incoraggiando gli studi su questa dimensione (cfr. Wood, Garb, Lilienfeld e Nezworski, 2002). La scala F dei test TALEIA, analogamente alla scala F di MMPI e MMMPI-2, si basa sul presupposto che chi afferma di avere numerosi sintomi la cui frequenza nella popolazione è rara, presumibilmente sta simulando. Il riferimento alla «frequenza» ha dato, d'altra parte, il nome alle scale di questo tipo (F). Le scale F di tutti i questionari TALEIA sono state costruite a partire dall'individuazione delle risposte a contenuto psicopatologico scelte dal 10% o meno di un campione di soggetti non clinici con N > 1.000 (circa 1.500 per la forma D3, N = 1146 per TALEIA-400A e N > 1000 per TALEIA-400B).

I quesiti così caratterizzati si sono però sempre rivelati troppo numerosi (N > 100), dato l'elevato numero di scale. Per ridurne la numerosità, in tutte le forme del questionario si è usato l'accorgimento di non includere nella scala F quesiti che potessero essere inseriti, alternativamente, in un'altra scala di controllo (L o INC). La scala F del Difesa Test, che doveva essere utilizzato per la visita di Leva e quindi avrebbe dovuto filtrare numerosi simulatori di patologie, non ebbe ulteriori ritocchi, anche se la numerosità degli item inclusi era notevole (N = 147). Per le forme TALEIA-400A e TALEIA-400B, destinate a una maggiore varietà di utilizzazioni, vennero anche esclusi dalla scala F i quesiti caratterizzati da valori H >1, cioè da una maggior incertezza dei soggetti nello scegliere fra le quattro alternative di risposta. I quesiti inclusi nella scala F sono presenti, a intervalli casuali, lungo tutta l'estensione dei test e non escludono, come nella scala F dell'MMPI, l'ultima porzione del questionario: d'altra parte, anche l'MMPI-2 ha aggiunto una scala Fb, riferita alla seconda parte del questionario. I sintomi inclusi nella scala F dei TALEIA sono attinenti ad una grande varietà di psicopatologie.

Più recente è l'introduzione nei questionari di personalità di una scala basata sulla coerenza/incoerenza fra risposte date a domande simili (Gough, 1987; Hathaway e McKinley, 2005), che dovrebbe misurare la credibilità d'insieme delle risposte, dipendente dalla comprensione del testo e/o dall'attenzione al compito.

La scala INC (da «incoerenze»), presente in tutte le forme di TALEIA, è stata costruita (separatamente per ogni forma) partendo da

dati empirici – i coefficienti di correlazione tra i quesiti – vagliati criticamente e successivamente selezionati applicando criteri di contenuto. In concreto, con riferimento a campioni di studenti (descritti in Boncori, 2007b) con livello di età medio intorno ai 19 anni e scolarità dai 12 ai 14 anni (che quindi si supponeva padroneggiassero a un buon livello la lingua materna) e in situazione di anonimato o di orientamento (situazioni in cui la motivazione alla distorsione è minima), è stata calcolata una matrice di intercorrelazione (r) fra tutti i quesiti e sono state individuate le coppie di quesiti correlate fra loro a livello $r => |0,5|$. Per le coppie di item così individuate sono stati trascritti gli indicatori e le scale attinenti. Le coppie che avevano in comune una di queste due classificazioni di contenuto sono state incluse nella scala INC. Per l'assegnazione del punteggio grezzo, è stato ipotizzato che la differenza fra la risposta (da «Mai» a «Sempre») data all'uno e all'altro elemento della coppia fosse una misura di «Incoerenza», utile per individuare chi aveva risposto senza molta attenzione, o senza comprendere bene il significato dei quesiti. Per ogni coppia di quesiti inclusa nella scala è stato quindi assegnato come «punteggio grezzo» il valore assoluto della differenza fra i due valori di scala attribuiti alle risposte: per esempio se a uno dei due quesiti era stata data la risposta «Mai» (valore di scala = 1) e all'altro «Sempre» (valore di scala = 5) il punteggio attribuito era 4.

Sulle innovazioni introdotte sarebbe interessante che si aprisse un dibattito metodologico, desiderio che gli Autori hanno espresso, riguardo a questa e ad altre caratteristiche innovative di TALEIA, anche in sede internazionale (Boncori, De Coro, Laganà, Nicole e Renzi, 2009).

4. VALIDITÀ DI CONTENUTO

La validità di contenuto negli anni Cinquanta era considerata requisito fondamentale per i test di profitto, mentre per i test di personalità si riteneva rilevante soprattutto la verifica della validità empirica basata sui «gruppi contrastanti». Con il passare degli anni si ebbe modo di osservare che il metodo empirico – basato su ipersemplificazioni sia riguardo alla casualità dei campioni sia riguardo alla validità delle misure-criterio utilizzate per definire il «contrasto» fra i campioni – raramente produceva conferme in studi replicati. Conseguentemente, già gli standard per la costruzione dei test pubblicati dall'American Psychological Association (APA) nel 1985 davano la validità di costrutto come elemento centrale e affermavano che «There is often no sharp distinction between test content and test construct» (AA.VV., 1985, p. 11). Nel nostro caso, la definizione dei contenuti

era imposta dal bando stesso di concorso: il test doveva produrre misure univocamente riconducibili a DSM e ICD[4]. Queste nosografie – peraltro descrittive – vanno considerate il principale fondamento teorico dello strumento, anche se sono state integrate con gli apporti descritti nei paragrafi successivi ed è stata abbandonata la criticatissima valutazione categoriale a favore della valutazione dimensionale, comune peraltro alla quasi totalità dei test di personalità in uso. Il carattere metodologico di questo contributo non ci consente peraltro una discussione dei contenuti delle scale sotto l'aspetto delle teorie psicopatologiche soggiacenti. In quanto componente della validità di costrutto, la validità di contenuto dei questionari TALEIA è stata ridefinita operativamente sotto più aspetti, ognuno dei quali è stato sottoposto a verifiche con disegni di ricerca appropriati. Le verifiche riguardanti l'individuazione dei disturbi da assumere come base per le scale (punto a) e l'individuazione dei criteri diagnostici da assumere come base per la formulazione degli item (punto b) sono state compiute, come si è detto, consultando anche esperti esterni al CTS e ottenendo i risultati già riferiti.

4.1. *Individuazione dei disturbi da assumere come base per le scale*

Una ricognizione di tutti i disturbi descritti nelle due nosografie di riferimento portò a identificare 468 categorie nosografiche: 16 menzionate solo nel DSM-IV, dieci solo nell'ICD-10 e tutte le altre comuni ad ambedue i repertori, anche se con denominazione non sempre identica. L'elenco completo divenne la base per un questionario da sottoporre ad esperti: accanto ad ogni categoria elencata era disposta una griglia a cinque livelli, su cui ogni esperto doveva segnare l'importanza da lei/lui attribuita a quella voce rispetto alle finalità del test, così come inizialmente definita nella lettera d'invito del Ministero della Difesa[5] e rispetto alle finalità che uno strumento analogo poteva avere in applicazioni «civili» nel campo sanitario, lavorativo, educativo. Vennero consultati come esperti i nove membri del CTS appartenenti al Dipartimento di Psicologia e un piccolo campione di psicologi e psichiatri (N = 29) operanti in strutture pubbliche. Successivamente, le categorie nosografiche vennero raggruppate in base a criteri clinici (in primo luogo: il codice attribuito da DSM-IV e ICD-

[4] Nel contratto venivano menzionate le forme ICD-9 e DSM-IIIR, che vennero sostituite da ICD-10 e DSM-IV non appena vennero pubblicati.

[5] «Un test di personalità idoneo a rendere possibile uno screening sui giovani chiamati a visita di Leva, onde poter individuare i soggetti psicolabili, potenzialmente a rischio della comunità militare» (Lettera-invito agli Enti, in data 15.4.1991).

10), riducendole a 64. Le categorie a cui tutti e tre i gruppi di esperti avevano attribuito importanza superiore al livello medio teorico dei punteggi erano 37: ancora troppo numerose per essere tutte assunte come base per scale di un questionario. Le discussioni collegiali all'interno del CTS portarono a escludere: a) disturbi che producono nel soggetto condizioni di difficoltà tali da compromettere la sua capacità di rispondere a un questionario: es. la Schizofrenia catatonica; b) disturbi che un questionario potrebbe valutare meno validamente rispetto a strumenti quali l'osservazione sistematica del comportamento o i test cognitivi (es. Ritardo mentale, Disturbi del linguaggio e della comunicazione); c) disturbi che hanno tipicamente il loro esordio in età infantile e precludono l'accesso ad attività lavorative o a studi superiori o evolvono in psicopatologie dell'età adulta caratterizzate in modo diverso dalle corrispondenti patologie infantili (es. Disturbo autistico e altri disturbi dello sviluppo); d) disturbi correlati con l'uso di sostanze o con l'astinenza da sostanze[6]; e) disturbi che non possono essere correttamente valutati senza il simultaneo ricorso a un esame fisico-medico (es. Disturbi di somatizzazione, conversione, ecc.); f) disturbi attinenti ad «aree sensibili», sulle quali la normativa riguardante la *privacy* vieta di indagare al di fuori di un'esplicita richiesta del soggetto a fini terapeutici o di consulenza personale (es. disturbi del funzionamento sessuale).

4.2. *Individuazione di criteri diagnostici da assumere come base per la formulazione degli item*

L'universo dei contenuti da sottoporre ad esame è stato identificato con l'insieme dei criteri diagnostici presenti nelle due nosografie di riferimento per i disturbi da includere nelle scale. Per assicurarci che questo insieme di indicatori fosse valido e attendibile anche nelle specifiche condizioni socioculturali italiane e per la fascia d'età assunta come obiettivo, gli indicatori sono stati raggruppati in funzione di 15 scale «possibili»[7] e per ciascun indicatore è stato chiesto a un

[6] I disturbi connessi con l'uso di sostanze non vennero inseriti sia perché, nel loro insieme, i criteri diagnostici segnalati sono sovrapponibili con quelli attinenti ad altri disturbi psicopatologici o di personalità, senza che sia possibile, in assenza di esami clinici, discriminare correttamente i fattori all'origine delle manifestazioni comportamentali rilevate dai questionari, sia perché sembrava non più attuale distinguere tra effetti prodotti da sostanze diverse, in quanto già ai tempi della progettazione del questionario la maggior parte dei consumatori di sostanze faceva uso di più sostanze e che entravano continuamente sul mercato nuove sostanze chimiche non considerate nelle nosografie.

[7] Le scale provvisorie erano: 1. AA Ansia acuta (attacco di panico) e disturbo di personalità ansioso; 2. AG Ansia generalizzata e disturbo di personalità ansioso; 3. B Disturbo di Personalità Borderline; 4. D Depressione e distimia; 5. IS Disturbi

campione di esperti (27 psichiatri e psicologi clinici, prevalentemente professionisti operanti in strutture pubbliche e con popolazioni affini a quelle a cui i questionari erano destinati)[8] di valutare se riteneva l'indicatore Utile e specifico; Utile, ma aspecifico; Raramente utile o pensava che fosse meglio non tenerne conto. I dati così raccolti hanno evidenziato che gli esperti consideravano «Utile e specifico» il 72,46% degli indicatori elencati, «Utile, ma aspecifico» il 20,35% «Raramente utile» il 4,96%. Gli indicatori riguardo ai quali gli esperti hanno risposto «Meglio non tenerne conto» riguardavano problematiche da loro ritenute secondarie (non però irrilevanti) nel quadro clinico generale, ed erano in tutto il 2,23%. Le percentuali di accordo tra operatori nel considerare utili gli indicatori elencati andava dal 100% («Disturbo di personalità istrionico»: un disturbo osservato e studiato da secoli) a un minimo del 79% («Sindromi fobiche»). Gli esperti aggiunsero anche, per alcune patologie, un limitato numero di altri indicatori, supportati dalla letteratura scientifica e/o dalla loro esperienza clinica, ma non recepiti dalle nosografie prese da noi come riferimento (sono riportate in Boncori, 2007b). Gli indicatori aggiuntivi riguardavano tutti i disturbi sottoposti a valutazione, ad eccezione della sindrome ossessivo-compulsiva. Ogni aggiunta era menzionata da un solo operatore e, nell'insieme, la proporzione delle aggiunte proposte era minore di quella segnalata in altri studi (Westen e Arkowitz-Westen, 1998), che però avevano contattato un campione più ampio di operatori. Molte proposte erano chiarificazioni o riformulazioni di

dell'identità sessuale; 6. POC Disturbo di personalità anancastico (ossessivo-compulsivo); 7. PAS Disturbo di personalità antisociale; 8. PD Disturbo di personalità dipendente; 9. PI Disturbo di personalità istrionico; 10. PN Disturbo di personalità narcisistico; 11. PP Disturbo di personalità paranoide; 12. F Fobiche, sindromi; 13. M Ipomania e mania; 14. S Schizofrenia, sindrome schizotipica e sindromi deliranti (incl. Disturbi dissociativi); 15. OC Sindrome ossessivo-compulsiva.

[8] Ringraziamo per la collaborazione il dott. F. Amore (psicologo dirigente presso le Ferrovie dello Stato), il dott. I. Ardizzone (Univ. di Roma «La Sapienza»), il col. Arduino (servizio psicologico dell'Aeronautica Militare), il dott. Bartolomei (Univ. di Roma «La Sapienza»), il dott. F. Borrelli (psicologo nella Polizia di Stato), il prof. S. Di Nuovo (Univ. di Catania), il dott. F. Giordano (Univ. di Roma «La Sapienza»), il dott. Grassi (psichiatra SERT nel SSN di Roma), il prof. G. Jervis (Univ. di Roma «La Sapienza»), il dott. L. Lucchetti (neuropsichiatra nella Polizia di Stato), la prof. M. Malagoli Togliatti (Univ. di Roma «La Sapienza»), il dott. M. Maurilio (Dirigente CSM nel SSN di Roma), il gen. Morelli (neuropsichiatra nel S.S. dell'Aeronautica Militare), il prof. A. Pazzagli (Università di Firenze), il cap. D. Panico (psicologo nell'Arma dei Carabinieri), la dott. M. Piscopo (Dirigente CSM nel SSN di Roma), il dott. Rocchi (psicologo DSM nel SSN di Roma), il col. C. Santinelli (neuropsichiatra dirigente la sezione di Psicologia Applicata di Levadife), il contrammiraglio Tomaselli (Ispettorato di Sanità della Marina Militare), il col. S. Tomassini (psicologo dirigente la sezione di Psicologia Applicata di Levadife), il dott. U. Ungaro (psicologo nella Polizia di Stato), il dott. Vacchini (psicologo DSM nel SSN di Roma), la dott. M. Vella (Dirigente CSM nel SSN di Roma), quattro docenti e ricercatori di una Università statale di Napoli che hanno preferito contribuire in forma anonima.

indicatori già inseriti nella lista e un buon numero erano in realtà riferimenti a comorbilità, e quindi già presenti in riferimento ad altre scale. Conseguentemente, il CTS non ritenne di apportare aggiunte al *corpus* degli indicatori desunti dai soli criteri diagnostici di ICD-10 e DSM-IV, ma di tenere presenti nella fase di formulazione degli item i dati acquisiti. Analogamente, nella fase di costruzione degli item vennero tenute presenti altre fonti di tipo clinico e psicodinamico, fra le quali occupavano un ruolo importante gli studi sui disturbi di personalità attinenti a diversi modelli teorici (si veda Clarkin e Lenzenweger, 1996; Livesley, Jang e Vernon, 1998), come pure le critiche sul problema dell'eccessiva comorbilità evidenziata dal sistema DSM (Oldham, Skodol, Kellman, Hyler, Doidge, Rosnick e Gallaher, 1995; Stuart, Pfohl, Battaglia, Bellodi, Grove e Cadoret, 1998) e le nuove proposte psicodiagnostiche quali la *Shedler-Westen Assessment Procedure* (SWAP; Shedler e Westen, 1998) e l'Asse «Struttura» dell'OPD (Gruppo OPD, 2002).

La consultazione dei 27 esperti evidenziò un problema metodologico di grande rilevanza. Soltanto due disturbi (Ansia generalizzata e Depressione e distimia) vennero valutati da tutti gli esperti intervistati. Riguardo ai rimanenti 13 disturbi, un numero maggiore o minore di esperti preferì non rispondere, dichiarando di aver incontrato un numero troppo ridotto di pazienti con quei disturbi: la frequenza più bassa di risposte è stata raccolta da Disturbo di personalità dipendente (18 esperti) e Disturbo di personalità anancastico o ossessivo-compulsivo (19 esperti); in media, valutarono i disturbi dell'Asse I 24 esperti su 27 e i disturbi di personalità 20 esperti su 27. Il problema metodologico posto da questa situazione è che per verificare la validità di uno strumento psicodiagnostico nuovo si può far riferimento a strumenti preesistenti oppure alle valutazioni e alle diagnosi di operatori esperti. Se gli operatori esperti manifestano incertezza non solo nell'applicazione dei criteri a singoli pazienti, ma nella definizione stessa della patologia è intuitivo che il questionario o il test da validare non può ragionevolmente riferirsi al criterio «esperti». Si dovrà invece cercar di assumere nell'ambito della validazione di costrutto basi di riferimento il più ampie e sicure possibile e, comunque, affidarsi a riscontri ripetuti nel tempo via via che anche negli operatori diventa più precisa la definizione dei criteri diagnostici e la loro identificazione clinica. Pensare ad una «validazione psicometrica» disgiunta dalla maturazione della concettualizzazione clinica sulle psicopatologie è un'illusione, quasi quanto lo è il sogno di fare una buona diagnosi senza l'aiuto di strumenti, fidando solo nel proprio intuito, che nessuno ha mai sottoposto a una verifica scientifica di validità. Di fatto, l'impostazione da noi seguita nella costruzione e nella verifica della validità di TALEIA ha seguito la tendenza prevalente attual-

mente nella letteratura clinica internazionale: superata la contrapposizione frontale tra predizione «attuariale» e «clinica» (una discussione fondamentale è in Wiggins, 1973, con conclusioni demolitive riguardo alla validità predittiva della diagnosi clinica), ma prendendo atto della scarsa attendibilità delle diagnosi basate sul colloquio clinico tradizionale, con indici di accordo intorno a $k = 0.25$ per i disturbi di personalità (Perry, 1992), mira a mettere a punto sistemi integrati, che utilizzano interviste strutturate o semistrutturate, che possono tener conto anche dei risultati di test.

In seguito all'analisi in indicatori, si decise di raggruppare all'interno di una stessa scala alcune categorie nosologiche caratterizzate da numerosi indicatori comuni e attinenti a disturbi di rara occorrenza. In particolare: a) nella scala S (Schizofrenia) vennero inclusi disturbi di contatto con la realtà (Delirium e Disturbi dissociativi), comunemente riconosciuti come appartenenti allo «spettro schizofrenico»; b) nelle scale riguardanti i Disturbi di personalità Borderline e Antisociale furono inclusi i Disturbi del controllo degli impulsi e il Comportamento antisociale dell'adulto e del bambino; c) i comportamenti di non collaborazione (Non collaborazione al trattamento, Simulazione) furono considerati nell'ambito delle scale di controllo; d) furono inclusi nella scala D anche gli indicatori attinenti alla Distimia, rinominando la scala in «Depressione e Distimia». Si progettò la predisposizione di sussidi che, utilizzando anche informazioni raccolte in fase di colloquio, consentissero una valida diagnosi differenziale tra le categorie accorpate nelle scale.

4.3. *Progettazione della scala di risposta appropriata*

Il nostro gruppo ha dedicato grande attenzione al problema della possibile distorsione delle risposte in direzione di stereotipi positivi (es. il profilo del «dipendente ideale» che un'azienda vorrebbe assumere o del «genitore ideale» a cui affidare un bambino) o di stereotipi negativi (es. il giovane a cui non si può mettere un'arma in mano, o la persona che nell'incidente ha subito un danno psichico), ritenendo fondamentale affrontare il problema in radice, nella predisposizione di stimoli che per le loro caratteristiche sia di contenuto sia formali suscitassero prevalentemente risposte attinenti alle realtà individuali anziché a stereotipi sociali. Dagli studi di psicologia sociale sappiamo che «gli stereotipi consistono in una serie di generalizzazioni diventate patrimonio degli individui: essi sono in gran parte derivati (o costituiscono uno dei casi) del processo cognitivo generale della categorizzazione. La funzione principale di questo processo consiste nel semplificare e nel sistematizzare, a fini di un adattamento cogni-

tivo e comportamentale, l'abbondanza e la complessità dell'informazione che l'organismo umano riceve dal suo ambiente» (Tajfel, 1995, pp. 238-239). Per quel che riguarda le situazioni più comuni di utilizzazione di un test psicopatologico è anche prevedibile che siano diffusi stereotipi sociali, cioè immagini mentali semplificate al massimo riguardanti categorie di persone (es. ruoli professionali: il bancario, l'agente di polizia, ecc., oppure ruoli sociali: il buon genitore, la vittima innocente, il «normale», lo squilibrato...) e condivise al massimo da grandi masse di persone (Palmonari, Cavazza e Rubini, 2002, p. 260). Nel caso dei ruoli professionali, la diffusione e la condivisione di stereotipi «positivi» sono anche supportate da libri di testo e corsi di preparazione per i concorsi, nel caso delle perizie sono supportate dall'intervento di «consulenti», la cui efficacia ai fini di indurre distorsioni è stata documentata (Ziskin e Faust, 1988). Sembrò al CTS che le alternative dicotomiche del tipo Vero/Falso facilitassero risposte riferite alle categorie semplici che definiscono gli stereotipi e fossero potenzialmente più dipendenti da dinamiche di desiderabilità sociale, mentre una scala di risposta temporale più articolata, di supporto a «stem» descrittivi di comportamenti collocati in situazioni specifiche potesse stimolare maggiormente risposte autentiche, riferite a episodi vissuti. Si decise per l'adozione di una scala di tipo Likert a cinque livelli descrittivi della frequenza con cui il soggetto ritiene che ricorra il comportamento menzionato nel quesito. Per evitare risposte su una categoria centrale, di dubbia interpretazione in psicopatologia, si è lasciato senza definizione il livello centrale, interpretabile come «Non saprei», che poi è stato eliminato dal foglio di risposta. Nella forma attualmente usata abbiamo quindi la scala «Mai – Qualche volta – Spesso – Sempre». Gli avverbi da usare per le alternative sono stati scelti fra quelli che nei dizionari di frequenza della lingua italiana appaiono come i più comuni fra quelli che esprimono frequenza temporale (Bortolini, Tagliavini e Zampolli, 1971; De Mauro, Mancini, Vedovelli e Voghera, 1993). Nelle istruzioni orali si dice che «Mai» e «Sempre» possono essere interpretati come «Quasi mai» e «Quasi sempre». La decisione di adottare quattro livelli è stata corroborata dopo aver osservato, nelle prime somministrazioni sperimentali individuali o a piccoli gruppi, che la presenza di quattro alternative anziché di due non solo rendeva più evidenti le risposte «poco autentiche», di solito rapidissime e riferite a uno dei due estremi, ma rendeva più rapido il procedimento di risposta chi era restio a dare risposte assolute (si pensi a persone con disturbi ossessivo-compulsivi o ansiosi, ma anche soltanto a individui resi più ansiosi dalla situazione di esame) e induceva anche a una maggior ponderazione nel rispondere chi aveva problemi cognitivi opposti, come le persone con disturbi ipomaniacali o maniacali.

4.4. *Formulazione di quesiti ad elevato livello di comprensibilità linguistica*

I contenuti dei quesiti, ovviamente, sono stati riferiti agli indicatori desunti dai «criteri diagnostici», predisponendo almeno tre quesiti per ciascun indicatore, in modo da poter includere uno o più quesiti per indicatore in ciascuna delle tre forme parallele principali. Tutti i quesiti sono originali. La stesura iniziale dei quesiti è stata compiuta da 28 autori, prevalentemente psicologi o psichiatri professionisti con lunga esperienza professionale[9].

Anche nella scrittura degli item è stato tenuto presente il problema del riferimento a stereotipi e della desiderabilità sociale. Si è quindi preferito evitare quesiti autodescrittivi (es. come quelli presenti nelle scale psichiatriche e nelle interviste strutturate) e si dettero indicazioni per riferire sistematicamente gli indicatori a situazioni specifiche e concrete, individuate nell'ambito della famiglia, dei rapporti sociali e dell'ambiente di lavoro o scolastico, così da andare verso la costruzione di quel che Cattell definiva uno «strumento T» o test, contrapposto agli «strumenti Q» o questionari autobiografici. D'altra parte, è anche stato segnalato nella letteratura recente che i quesiti autodescrittivi, indipendentemente dal riferimento o meno a stereotipi, raccolgono risposte poco valide da pazienti che, come parte della propria sintomatologia, hanno scarsa introspezione o difficoltà a riferire sulle proprie emozioni (cfr. Westen e Shedler, 1999, 2000).

Sotto l'aspetto linguistico i quesiti vennero costruiti secondo le norme indicate nella letteratura psicometrica internazionale per ridurre al minimo gli effetti di fattori quali la modesta acculturazione, l'acquiescenza, la desiderabilità sociale, ecc. (Boncori, 1993, 2006). Il punto chiave delle norme date è riconducibile all'esigenza di controllare al massimo la rappresentatività dei processi impiegati dai soggetti per arrivare a una risposta (Rennon, 1956, cit. da Messick, 1980), il

[9] Membri del Dipartimento di Psicologia: L. Boncori (109 quesiti), A. De Coro (385 q.), S. Nicole (19 q.), A. Orsini (50 q.), F. Ortu (40 q.), T. G. Scalisi (101 q.), che hanno contribuito per un totale di 704 quesiti; operatori psicologi/psichiatri con oltre dieci anni di esperienza professionale: I. Bigari (61 q.), G. Buffardi (26 q.), G. Infantino (163 q.), P. Salotti (61 q.), G. M. Sferrazza (26 q.), D. Solfaroli Camillocci (61 q.), S. Tomassini (94 q.), che hanno contribuito con un totale di 492 quesiti; giovani psicologi professionisti: A. Barruffi (12 q.), F. Bonaiuto (100 q.), A. Ciaglia (98 q.), A. Costa (12 q.), M. Falocco (12 q.), C. Fusciello (82 q.), S. Magnone (62 q.), A. Martucci (62 q.), che hanno contribuito con un totale di 440 quesiti; laureati in psicologia: S. Alcini (53 q.), E. Caponera (53 q.), M. G. Castiglione (5 q.), M. T. Fiorentino (26 q.), L. Giannini (89 q.), L. Lovisetto (26 q.), A. Tomassini (62 q.), che hanno contribuito con un totale di 314 quesiti. I giovani psicologi e i laureati hanno scritto quesiti esclusivamente su tematiche attinenti alla propria tesi di laurea e hanno lavorato sotto la diretta supervisione della prof. Boncori, che ha inserito nella banca dati solo i quesiti da lei ritenuti formulati a livello professionale.

che può essere tradotto nella norma empirica data da L.J. Cronbach ai costruttori di test: «la norma più generale per assicurare la validità di contenuto è questa: *nessuna difficoltà non pertinente*» (Cronbach, 1984).

Prima della stesura dei quesiti è stata distribuito a tutti i potenziali autori dei quesiti un testo – previamente discusso e approvato in una riunione collegiale allargata – in cui venivano sintetizzate le raccomandazioni raccolte dalla letteratura psicometrica e clinica internazionale (Ward, 1981) riguardo alla stesura dei quesiti da includere in questionari di personalità (Boncori, 2007b). Le norme chiedevano, sostanzialmente, di scrivere «stem» (indifferentemente in forma affermativa o interrogativa) che fossero molto brevi (non più di una riga), usando vocaboli di significato non ambiguo e frequenti nella lingua parlata[10], strutturando le frasi come proposizioni semplici e senza subordinate, con gli elementi disposti in ordine diretto (soggetto – predicato – complementi), limitando al massimo l'uso della negazione e in particolare evitando l'uso del «non», da sostituire, quando possibile, con parole come «niente», «nessuno», «pochissimo» che consentissero di evitare l'ambiguità semantica connessa con la presenza delle doppie negazioni. Ovviamente i quesiti dovevano essere indipendenti l'uno dall'altro, in modo da poter essere cambiati di posto facilmente, e dovevano essere congruenti con le alternative di risposta (ossia: lo stem non doveva mai contenere parole come *mai, qualche volta, spesso, sempre*).

Per controllare l'effetto della tendenza a rispondere «sì» ai quesiti indipendentemente dal loro contenuto («acquiescenza») venne data anche la norma (Nunnally, 1978, p. 669) che le risposte alfa si collocassero per metà dei quesiti nel polo positivo (Sempre) e per metà nel polo negativo (Mai), pur sapendo che anche in questo caso i risultati possono essere di dubbio valore (cfr. Knowles, 1963). D'altra parte, introdurre correzioni per escludere più radicalmente l'influsso dell'acquiescenza, che è correlata con l'ansietà, comporterebbe distorsioni maggiori (Cattell, Eber e Tatsuoka, 1970, pp. 52-53).

Nella formulazione dei quesiti è stato anche raccomandato di evitare elementi – contenutistici o formali – che potevano mettere in situazione di svantaggio gli appartenenti ad una minoranza. In particolare sono stati sconsigliati i riferimenti a convinzioni religiose, a consuetudini locali o etniche, a preferenze sessuali non strettamente attinenti alle variabili da valutare.

[10] Successivamente venne fatto un controllo assumendo come criterio il Lessico di frequenza dell'italiano parlato (De Mauro *et al.*, 1993) e sostituendo i vocaboli di uso meno frequente.

Gli effetti sulla comprensibilità sono stati verificati con due studi distinti, diversi per disegno di ricerca e per caratterizzazione dei campioni su cui sono stati compiuti.

Il primo studio si proponeva di identificare difficoltà di risposta dovute, rispettivamente, a problemi emotivi, linguistici, cognitivi. Per controllare la componente cognitiva (Stone, Stone e Gueutal, 1990), l'intera banca dati dei quesiti venne somministrata insieme a un test d'intelligenza generale (Otis Quick-scoring Gamma: Otis, 1962) e a un test che valutava intelligenza verbale e capacità di ragionamento spaziale (L-S: Boncori, 1995): i quesiti le cui risposte sarebbero state trovate significativamente correlate con i risultati del test d'intelligenza generale e/o con il test d'intelligenza verbale sarebbero stati considerati misure «spurie» dei disturbi ed escluse dalla banca dati. Per identificare i quesiti che ponevano in difficoltà le persone, ritardandone le risposte venne progettata una ricerca qualitativa, che affidava la discriminazione fra difficoltà di tipo emotivo e difficoltà di tipo linguistico a somministratori psicologi professionalmente preparati. Ambedue gli studi vennero condotti su un campione nazionale di giovani, di età media 17.6 anni; 330 M e 80 F, residenti in cinque diverse città del Nord, del Centro e del Sud: gli estremi geografici erano Pordenone e Palmi. I giovani frequentavano il 2° anno in corsi di formazione (prevalentemente) o di studio in aree culturalmente e/o linguisticamente marginali rispetto agli standard nazionali. La loro età media dice chiaramente che molti di loro erano «in ritardo» rispetto all'età di uno studente che frequenti il 10° anno di scuola senza avere mai ripetuto. Si trattava quindi di un campione di livello culturale molto più modesto rispetto ai campioni studenteschi universitari o del 4°-5° anno di scuola secondaria superiore a cui si riferiscono gli studi sulla validità di altri test. I vocaboli o le strutture sintattiche che risultarono mal compresi nei gruppi condotti da almeno due diversi somministratori vennero sostituiti con le forme alternative proposte (e già verificate sul campo) dai somministratori stessi. Una delle modifiche introdotte riguardò un'alternativa di risposta: «Talvolta» venne sostituito da «Qualche volta», più comprensibile anche se più lungo. Nessun quesito risultò avere un coefficiente di correlazione $r => |.30|$ con i punteggi del test di ragionamento spaziale, mentre una decina di item superarono questa soglia in riferimento al test d'intelligenza verbale e vennero esclusi dalla banca dati.

Lo studio successivo venne compiuto utilizzando una campionatura degli item più ristretta («Difesa Test») e una campionatura di soggetti più ampia. Questo secondo campione era costituito da $N = 4.070$ giovani di Leva, esaminati in 18 diverse città sede di distretto di Leva, di età media 17.6 anni come il campione precedente, ma di livello culturale mediamente più modesto. Il Difesa Test includeva 300 item

che costituivano un campione casuale stratificato (per indicatori) della banca dati, a cui era stato aggiunto un ultimo item: «È risultato comprensibile il presente questionario?». Su una scala da 1 (= Mai) a 5 (= Sempre) i valori medi delle risposte andavano da un minimo di 2.7 a un massimo di 4. L'indicatore implicito di comprensibilità costituito dalla percentuale di risposte omesse o segnate in modo poco chiaro era inferiore all'1%.

5. CONCLUSIONI

Il questionario presentato in questo articolo si caratterizza come uno strumento per la diagnosi psicopatologica in riferimento alle nosografie internazionali correnti (DSM-IV e ICD-10), ed ha utilizzato in fase di predisposizione degli item anche contributi clinici ad esse esterni e i risultati delle ricerche empiriche apparsi nella letteratura internazionale fino al 2007. Fin dalla definizione dei contenuti da includere, si è ritenuto che l'uso di tale strumento debba considerare il colloquio come parte integrante del processo diagnostico. La validità di contenuto non è stata solo definita a priori, ma anche sottoposta a una serie di controlli lungo tutto l'iter di costruzione. In particolare, è stato verificato l'accordo di esperti (anche esterni rispetto a quelli che partecipavano al Comitato Tecnico-scientifico) e l'indipendenza delle risposte agli item dall'intelligenza verbale e dall'appartenenza a categorie socio-culturali di livello medio-alto. L'articolo presenta solo una parte dell'ampia serie di ricerche svolte, alcune delle quali sono pubblicate altrove (Boncori, 2007b, Boncori *et al.*, 2001a, 2001b; Boncori *et al.*, 2009). Nella seconda parte del lavoro viene dato un resoconto delle ricerche compiute per verificare attendibilità e validità diagnostica dello strumento.

BIBLIOGRAFIA

AA.VV. (1985). *Standards for educational and psychological tests*. Washington DC: American Educational Research Association, American Psychological Association, National Council on Measurement in Education.

American Psychiatric Association (1994). *Diagnostic and statistical manual of mental disorders, fourth edition (DSM-IV)*. Washington, DC: APA (trad. it. *DSM-IV. Manuale diagnostico e statistico dei disturbi mentali*. Milano: Masson, 1995).

Baer R.A., Miller J. (2002). Underreporting of psychopathology on the MMPI-2: A meta-analytic review. *Psychological Assessment*, *14*, 16-26.

Baer R.A., Wetter A.W., Berry D.T.R. (1992). Detection of underreporting of psychopathology on the MMPI: A meta-analysis. *Clinical Psychology Review*, *12*, 509-525.

BONCORI L. (1993). *Teoria e tecniche dei test*. Torino: Boringhieri.
BONCORI L. (1995). *Test L-S*. Roma: CRISP.
BONCORI L. (1999a). Un nuovo questionario psicopatologico per le Forze Armate. *Atti del 1° Simposio nazionale di Psicologia militare (Civitavecchia, 12-13 novembre 1997)*. Roma: Ministero della Difesa – Direzione Generale della Leva, pp. 119-131.
BONCORI L. (a cura di) (1999b). *Il nuovo questionario per le Forze Armate «Difesa Test» (DT) – Manuale d'uso per gli operatori*. Roma: Ministero della Difesa.
BONCORI L. (2006). *I test in psicologia: Fondamenti teorici e applicazioni*. Bologna: Il Mulino.
BONCORI L. (a cura di) (2007a). TALEIA-400A: *Test for AxiaL Evaluation and Interview for Clinical, Personnel, and Guidance Applications*. Trento: Erickson.
BONCORI L. (2007b). *TALEIA-400A: Test for AxiaL Evaluation and Interview for Clinical, Personnel, and Guidance Applications – Manuale*. Trento: Erickson.
BONCORI L., BAIOCCO R., BARRUFFI A., COMIGNANI E. (2001a). The «Alexithymiaz Scale» in a new questionnaire for the Army. *Atti del 35° Simposio internazionale di Psicologia militare applicata (Firenze, 24-28 maggio 1999)*. Roma: Ministero della Difesa – Direzione Generale della Leva, pp. 173-183.
BONCORI L., BAIOCCO R., BARRUFFI A., COMIGNANI E. (2001b). The «Schizophrenia Scale» in a new questionnaire for the Army. *Atti del 35° Simposio internazionale di Psicologia militare applicata (Firenze, 24-28 maggio 1999)*. Roma: Ministero della Difesa – Direzione Generale della Leva, pp. 185-194.
BONCORI L., DE CORO A., LAGANÀ L., NICOLE S., RENZI P. (2009). Psychopathological assessment in clinical and workplace applications: An Italian measure. *International Journal of Testing*.
BORTOLINI U., TAGLIAVINI C., ZAMPOLLI A. (1971). *Lessico di frequenza della lingua italiana contemporanea.* Milano: Garzanti.
BUCKLEY-SHARP M.D., HARRIS F.T.C. (1970). The banking of multiple-choice questions. *British Journal of Medical Education*, *4*, 42-52.
CATTELL R.B., EBER H.W., TATSUOKA M.M. (1970). *Handbook for the Sixteen Personality Factor Questionnaire (16 PF)*. Champaign, IL: IPAT.
CLARKIN J.F., LENZENWEGER M.F. (1996). *Theories of personality disorder*. New York: Guilford Press (trad. it. *I disturbi di personalità. Le cinque principali teorie*. Milano: Raffaello Cortina, 1997).
CRONBACH L.J. (1984). *Essentials of psychological testing* (4th ed.). New York: Harper.
CROWNE D.P., MARLOWE D. (1960). A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology*, *24*, 349-354.
DE MAURO T., MANCINI F., VEDOVELLI M., VOGHERA M. (1993). *Lessico di frequenza dell'italiano parlato.* Milano: ETAS Libri.
EDWARDS A.L. (1957). *The social desirability variable in personality assessment and research*. New York: Dryden.
FRANKE G.H. (2002). *Faking bad in personality inventories: Consequences for the clinical context*. *Psychologische Beiträge*, *44*, 50-61.
GORTH W.P., ALLEN D.W., GRAYSON A. (1971). Computer programs for test objective and item banking. *Educational and Psychological Measurement*, *31*, 245-250.
GOUGH H.G. (1987). *CPI – California Psychological Inventory*. Palo Alto, CA: Consulting Psychologists Press.

GRUPPO OPD (2002). *Diagnosi psicodinamica operazionalizzata*. Milano: Masson.
HATHAWAY S.R., MCKINLEY J.C. (2005). *MMPI-2 Minnesota Multiphasic Personality Inventory – 2™. Manuale*. Firenze: Organizzazioni Speciali.
KNOWLES J.B. (1963). Acquiescence response set and the questionnaire measurement of personality. *British Journal of social and clinical Psychology*, *2*, 131-137.
LIVESLEY W.J., JANG K.L, VERNON P.A. (1998). Phenotypic and genetic structure of traits delineating personality disorder. *Archives of General Psychiatry*, *55*, 941-947.
MCFARLAND L.A., RYAN A.M. (2000). Variance in faking across noncognitive measures. *Journal of Applied Psychology*, *85*, 812-821.
MESSICK S. (1980). Test validity and the ethics of assessment. *American Psychologist*, *35*, 1012-1027.
MILLMAN J., ARTER J.A. (1984). Issues in item banking. *Journal of Educational Measurement*, *21*, 315-330.
NOVAGA M., PEDON A. (1977). *Contributo allo studio della personalità: il 16 P.F. test di Cattell*. Firenze: Organizzazioni Speciali.
NUNNALLY J.C. (1978). *Psychometric theory* (2nd ed.). New York: McGraw-Hill.
NUNNALLY J.C., BERNSTEIN I.H. (1994). *Psychometric theory* (3nd ed.). New York: McGraw-Hill.
OLDHAM J.M., SKODOL A.E., KELLMAN H.D., HYLER S.E., DOIDGE N., ROSNICK L., GALLAHER P.E. (1995) *Comorbidity of axis I and axis II disorders*. *American Journal of Psychiatry*, *152*, 571-578.
OTIS A.S. (1962). *Otis Quick-Scoring Mental Ability Tests.* Adattamento italiano a cura di L. Calonghi e G. Baronchelli. Firenze: Organizzazioni Speciali.
PALMONARI A., CAVAZZA N., RUBINI M. (2002). *Psicologia sociale*. Bologna, Il Mulino.
PAULHUS D.L., REID D.B. (1991). Enhancement and denial in socially desirable responding. *Journal of personality and Social Psychology*, *60*, 307-317.
PERRY J.C. (1992). Problems and considerations in the valid assessment of personality disorders. *American Journal of Psychiatry*, *149*, 1645-1653.
RORER L.G. (1965). The great response style myth. *Psychological Bulletin*, *63*, 129-156.
SHEDLER J., WESTEN D. (1998). Refining the measurement of Axis II: A Q-sort procedure for assessing personality pathology. *Assessment*, *5*, 335-355.
STONE E.F., STONE D.L., GUEUTAL H.G. (1990). Influence of cognitive ability on responses to questionnaire measures: Measurement precision and missing response problems. *Journal of Applied Psychology*, *75*, 418-427.
STUART S., PFOHL B., BATTAGLIA M., BELLODI, L., GROVE W., CADORET R. (1998). The co-occurrence of DSM-III-R personality disorders. *Journal of Personality Disorders*, *12*, 302-315.
TAJFEL H. (1995). *Gruppi umani e categorie sociali* (2ª ed.) Bologna: Il Mulino.
VALE C.D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*, 333-344.
VAN DER LINDEN W., EGGEN T.J.H.M. (1986). An empirical bayesian approach to item banking. *Applied Psychological Measurement*, *10*, 345-354.
WARD C. (1981). *Preparing and using objective questions*. Cheltenham: Stanley Thornes.
WESTEN D., ARKOWITZ-WESTEN L. (1998). *Limitations of Axis II in diagnosing personality pathology in clinical practice*. *American Journal of Psychiatry*, *155*, 1767-1771.

WESTEN D., SHEDLER J. (1999). Revising and assessing axis II, part I: Developing a clinically and empirically valid assessment method. *American Journal of Psychiatry*, *156*, 258-272.
WESTEN D., SHEDLER J. (2000). A prototype matching approach to diagnosing personality disorders: Toward DSM-V. *Journal of Personality Disorders*, *14*, 109-126.
WESTEN D., SHEDLER J., LINGIARDI V. (2003). *La valutazione della personalità con la SWAP-200*. Milano: Raffaello Cortina.
WIGGINS J.S. (1973). *Personality and prediction. Principles of personality assessment*. Reading, MA: Addison-Wesley.
WOOD J.M., GARB H.N., LILIENFELD S.O., NEZWORSKI M.T. (2002). Clinical assessment. *Annual Review of Psychology*, *53*, 519-543.
WOOD R., SKURNIK L.S. (1969). *Item banking*. London: National Foundation for Educational Research in England and Wales.
WORLD HEALTH ORGANIZATION (1992). *The ICD-10 Classification of mental disorders and behavioral disorders: Clinical descriptions and diagnostic guidelines*. Geneva: WHO (trad. it. *ICD-10. Decima revisione della classificazione internazionale delle sindromi e dei disturbi psichici e comportamentali: descrizioni cliniche e direttive diagnostiche*. Milano: Masson, 1992).
WORLD HEALTH ORGANIZATION (1993). *The ICD-10 classification of mental and behavioral disorders: Diagnostic criteria for research*. Geneva: WHO (trad. it. *ICD-10. Decima revisione della classificazione internazionale delle sindromi e dei disturbi psichici e comportamentali: criteri diagnostici per la ricerca*. Milano: Masson, 1994).
ZISKIN J., FAUST D. (1988). *Coping with psychiatric and psychological testimony* (4th ed.). Marina del Rey, CA: Law & Psychology Press.

[*Ricevuto il 6 febbraio 2009*]
[*Accettato il 2 agosto 2009*]

Innovation in psychopathological testing: TALEIA. Part I: Content validity and validity scales

Summary. Studies content validity of TALEIA (*Test for AxiaL Evaluation and Interview for Clinical, Personnel, and Guidance Applications*), focusing on construction methods. Evidence is given for expert agreement about relevance to nosographies (DSM-IV & ICD-10), psychometric grounds underlying the validity scales, independence from linguistic competence and verbal intelligence. Nosographic categories were reduced to the ones of above average importance in the experts' (N = 29) assessment. Diagnostic indicators were evaluated by a different panel (N = 27) mostly as «useful» (93%) or «seldom useful» (5%), agreement varying between 100% and 79%. Correlations with intelligence test showed significant only for ten items, therefore excluded from the data bank. The test readability was verified in a quasi-clinical setting (national sample of small groups, N = 410), and on a large national sample (N = 4070) of draftees. The few words «linguistically difficult» in two or more samples were substituted. In the large sample less than 1% of Ss. omitted or not clearly answered one or more items. TALEIA appears as a potentially sound instrument to be further investigated.

Keywords: ???

La corrispondenza va inviata a Lucia Boncori, Dipartimento di Psicologia, Sapienza Università di Roma, Via dei Marsi 78, 00185 Roma, e-mail: lucia.boncori@uniroma1.it