3

Neil T. Martin
Keele University
E.A. Gaffan e T. Williams
University of Reading

Analisi funzionale sperimentale dei comportamenti problema: una metodologia che fa nascere molti dubbi

SOMMARIO

In questo studio si esaminò la validità convergente di un'analisi funzionale sperimentale attraverso il confronto di tre modalità distinte di interpretazione dei dati: due metodi provenienti dalla ricerca e il metodo del criterio Z, sviluppato dagli autori di questo articolo. Per valutare l'accordo fra queste tre forme di interpretazione, furono analizzati i dati risultanti dall'analisi funzionale sperimentale dei comportamenti problema di 27 persone con ritardo mentale, calcolando anche l'attendibilità test-retest dei tre metodi dopo periodi di 2 settimane, un mese e 3 mesi. I risultati suggeriscono che i metodi esaminati per interpretare le valutazioni e identificare la funzione dei comportamenti problema possono dare risultati diversi e che l'attendibilità test-retest delle analisi funzionali sperimentali è molto bassa. Le implicazioni di questi risultati sono discusse in riferimento alla pratica clinica.

Negli ultimi vent'anni vi è stata un'enorme diffusione dell'analisi funzionale come strumento per comprendere i comportamenti problema e intervenire su di essi (Carr, 1994; Durand e Crimmins, 1988; Hall e Oliver, 1992; Iwata et al., 1982; Bauman e Richman, 1982; Mace, 1994; Repp et al., 1990). L'analisi funzionale ha fornito dei metodi per verificare l'ipotesi secondo la quale i comportamenti problema vengono mantenuti da antecedenti e conseguenze specifici e sistematici, che si verificano nell'ambiente della persona.

Ci sono tre tecniche distinte di analisi funzionale:

1. la valutazione informativa/indiretta (colloquio clinico);
2. l'osservazione diretta degli antecedenti e delle conseguenze nell'ambiente naturale della persona;
3. la manipolazione sperimentale degli antecedenti e delle conseguenze che si suppongono responsabili del comportamenti problema, creando «situazioni analoghe» a quelle delle funzioni principali.

La funzione del comportamento problema viene definita sulla base delle risposte coerenti all'interno di una condizione specifica. Ad esempio, frequenze sistematicamente elevate di comportamenti problema in condizioni di scarsa attenzione (o nelle quali l'attenzione è contingente ai comportamenti problema) suggeriscono una funzione mantenuta dall'attenzione; diversamente, se le frequenze elevate di comportamenti problema si registrano in assenza di stimoli, ciò suggerisce una funzione di autostimolazione.

In letteratura sono state descritte varie modalità di valutazione funzionale sperimentale che si distinguono per la loro ampiezza — ad esempio le valutazioni estese e quelle brevi (Northup et al., 1991) — e per l'uso di contingenze (Carr e Durand, 1985; Iwata et al., 1982). Secondo Iwata e colleghi (1982), l'uso delle contingenze è necessario per rendere le condizioni sperimentali simili a quelle fornite nell'ambiente naturale. Iwata (1994) suggerisce inoltre la possibilità che, se durante la valutazione non vengono utilizzate contingenze, il comportamento in esame potrebbe estinguersi. Tuttavia, una critica che si potrebbe avanzare al metodo di Iwata e colleghi (1994), nel quale la valutazione può protrarsi finché i dati si stabilizzano (Iwata et al., 1994; Vollmer et al., 1995), è che il contatto prolungato con una contingenza può dare luogo a nuovi apprendimenti. Infatti, applicando una contingenza a una determinata probabile risposta è possibile che quella risposta venga da quel momento in poi controllata dalla contingenza (Mace, Lalli e Pinter-Lalli, 1991), anche se prima, nell'ambiente naturale della persona, non lo era.

Validità convergente della valutazione sperimentale

Sebbene il metodo di analisi funzionale sperimentale delle «situazioni analoghe» venga ampiamente utilizzato nella ricerca, finora sono stati condotti ben pochi studi per valutare in che misura i risultati forniti da questo tipo di valutazione siano congruenti con quelli ottenuti attraverso altri metodi (cioè se tutti identifichino la medesima funzione).

Durand e Crimmins (1988) rilevarono un accordo tra i risultati forniti dalle analisi sperimentali e dalle scale di valutazione, ma solo per quanto riguarda gli 8 soggetti selezionati. Lerman e Iwata (1993) riscontrarono che l'osservazione diretta era meno precisa delle analisi funzionali sperimentali perché non forniva dati sufficienti a identificare la funzione dei comportamenti problema di 5 dei 6 partecipanti al loro studio. Emerson e colleghi (1995) ottennero una buona concordanza tra analisi funzionale sperimentale e osservazione naturalistica diretta limitando il confronto a topografie per le quali entrambi i metodi erano in grado di identificare la funzione. La funzione veniva identificata quando si presentavano schemi di risposta chiari e omogenei, attribuibili a una condizione sperimentale (nel caso della valutazione analoga) oppure a specifici antecedenti o conseguenze (nel caso dell'osservazione diretta).

Toogood e Timlin (1996) hanno confrontato tutte e tre le metodologie di analisi funzionale valutando la funzione di 121 comportamenti problema manifestati da 20 persone con ritardo mentale grave. La metodologia sperimentale (analoga) utilizzata si basava su quella descritta da Iwata e colleghi (1982). Questi autori rilevarono che i tre metodi differivano in più aspetti: la loro probabilità di identificare la funzione del comportamento problema — il colloquio clinico (valutazione informativa/indiretta) individuò la funzione del 74% dei comportamenti problema considerati, l'osservazione diretta del 68%, mentre la valutazione sperimentale solo del 41% —, i tipi di comportamento ai quali le funzioni venivano attribuite con maggiore probabilità, i tipi di funzioni che venivano più probabilmente identificate, e la probabilità che venisse identificata una funzione multipla (ciò successe nel 73% dei casi utilizzando il colloquio clinico, nel 58% dei casi utilizzando una valutazione analoga e nel 37% dei casi utilizzando metodi descrittivi). Un completo accordo sull'identificazione della funzione, o sull'incapacità di individuarla, si ottenne soltanto per 5 dei 121 comportamenti considerati.

È sorprendente che, nello studio di Toogood e Timlin, il metodo sperimentale analogo non fu in grado di identificare la funzione del 50% dei comportamenti problema in esame, a fronte di una percentuale di insuccesso dell'11-15% per il colloquio clinico e il *Motivation Assessment Schedule* (Durand e Crimmins, 1998) e del 18% per l'analisi descrittiva. Questo dato è fortemente in contrasto con la percentuale di risultati positivi del 95% (su 152 casi) fornita da Iwata e colleghi (1994) riguardo alla loro metodologia sperimentale. Toogood e Timlin proposero alcune possibili spiegazioni di questa discrepanza, tra le quali le diverse caratteristiche dei campioni usati nei due studi, la differenza tra i comportamenti problema considerati (in quello di Iwata e colleghi si esaminava solo l'autolesionismo) e la differenza tra i criteri richiesti per l'identificazione della funzione.

Identificazione della funzione

Nel loro studio, Iwata e colleghi (1994) svilupparono alcune ipotesi sulla funzione dei comportamenti problema ricavate dall'esame dei punteggi Z, cioè il numero di deviazioni standard dal punteggio medio del totale delle risposte (frequenza percentuale) per ogni condizione analoga. Questi autori riportano che i dati furono interpretati da gruppi di 4-10 osservatori che dovevano raggiungere l'accordo sulle variabili coinvolte nel mantenimento del comportamento problema; ad esempio, punteggi Z elevati nella condizione «da solo» suggerivano che un comportamento problema poteva avere funzione di autostimolazione. Nello studio non viene spiegato come venisse raggiunto tale consenso e uno degli osservatori (il primo autore) era sempre presente in tutti i gruppi, dato che suscita perplessità circa l'effettiva obiettività del processo. Là dove esistono poche diffe-

renze generali tra le condizioni oppure dove alcune condizioni specifiche non forniscono dati sistematicamente superiori rispetto alle altre, identificare la funzione diventa difficile, se non impossibile. Qualche difficoltà di interpretazione può sorgere anche quando si effettua una serie di valutazioni analoghe distinte per un certo periodo di tempo su uno specifico comportamento, e le diverse valutazioni forniscono risultati differenti, che potrebbero essere dovuti a una modificazione del comportamento, alla presenza di più funzioni oppure semplicemente all'inadeguatezza della metodologia. Ciò evidenzia uno degli aspetti più critici e criticabili del metodo di interpretazione di Iwata e colleghi (1982), che si basa su risultati di periodi di valutazione potenzialmente lunghi. Questi autori, infatti, ritengono necessario protrarre le valutazioni fino a quando non emergano dai dati degli schemi di risposta chiari. Dato che Iwata e colleghi applicano delle contingenze nelle loro condizioni analoghe — fatta eccezione per la condizione «da solo» —, la possibilità che, con l'esposizione a contingenze specifiche, il soggetto apprenda nuovi comportamenti dovrebbe indurre alla cautela nell'interpretare la comparsa di differenziazioni tra le condizioni, soprattutto nel corso di lunghi periodi di tempo.

Dall'esame dei grafici presentati da Iwata e colleghi (1982) sembra che questi autori abbiano utilizzato un criterio per la differenziazione delle condizioni, sebbene esso non venga reso esplicito. Essi non riuscirono a identificare la funzione dei comportamenti autolesionistici di 3 dei 9 partecipanti al loro studio. Un'analisi più attenta dei grafici mostra che, nei casi in cui la funzione veniva individuata, i punteggi Z delle condizioni rilevanti erano molto vicini o superiori a una deviazione standard dalla media. Perciò, il criterio da loro utilizzato potrebbe essere quello secondo cui una condizione è diversa quando il punteggio Z è uguale o superiore a 1. Questo valore, tuttavia, è comunque arbitrario e rimane il problema di decidere quale valore dovrebbe avere il punteggio Z perché una condizione possa essere considerata differente. Un criterio per individuare la funzione dei comportamenti problema utilizzando i punteggi Z, quindi, è che Z sia significativamente superiore a 0; esso è il metodo del criterio Z.

Nel loro studio, Toogood e Timlin attribuivano le funzioni soltanto quando venivano soddisfatti alcuni criteri; in particolare, le differenze tra le condizioni dovevano essere coerenti con una data ipotesi, dovevano essere uniformi in due o più sessioni (su 4) e la frequenza media delle risposte nella condizione in cui essa era più elevata doveva essere superiore del 50% alla media generale di tutte le condizioni e tutte le sessioni. Benché Toogood e Timlin abbiano utilizzato criteri specifici per attribuire la funzione, tali criteri sono comunque arbitrari ed evidenziano il modo soggettivo e spesso personale in cui i dati vengono da loro interpretati.

Hagopian e colleghi (1997) si avvalsero di una commissione di esperti per sviluppare criteri strutturati per l'interpretazione dei dati. Perché una condizione analoga potesse essere considerata differente rispetto alle altre, almeno la

metà dei punteggi doveva collocarsi al di sopra di una linea di criterio. Tuttavia, se nella seconda fase della valutazione parte di questi dati scendeva al di sotto di tale linea, ciò veniva interpretato come tendenza decrescente e la condizione non era ritenuta differenziata. Esistono anche dei criteri per la differenziazione multipla delle condizioni: se la condizione «da solo» presenta i punteggi più elevati, le altre possono essere ignorate e può essere attribuita la funzione di autostimolazione/rinforzamento automatico; se risulta differenziata, insieme a un'altra condizione, e rispetto a questa mostra punteggi relativamente più bassi, può essere attribuita una funzione multipla (comprendente quella di autostimolazione/rinforzamento automatico); se ci sono tre condizioni differenziate e quella «da solo» non ha i punteggi relativamente più alti, essa dovrebbe essere ignorata e la funzione — multipla — dovrebbe essere attribuita sulla base delle altre due (in queste circostanze, secondo Hagopian e colleghi, punteggi elevati nella condizione «da solo» potrebbero essere dovuti semplicemente agli effetti residui della condizione precedente). La linea di criterio viene determinata in base al livello di risposta ottenuto nella condizione di controllo (nel caso dello studio di Iwata e colleghi si tratta della condizione di gioco) e fissata a circa una deviazione standard sopra la media di tutti i dati relativi a tale condizione. In una distribuzione normale, con 10 valori per ogni condizione, la linea di criterio si collocherebbe tra il secondo e il terzo valore più alto della condizione di controllo. Tale linea risulta ben diversa da quella definita con le procedure di Toogood e Timlin (1996), che si pone a un livello del 50% superiore alla media generale di tutte le condizioni e sessioni.

I criteri proposti da Hagopian e colleghi rappresentano un tentativo di standardizzare l'interpretazione dei dati forniti dalla valutazione analoga, ma implicano una serie di regole e criteri piuttosto complicati, alcuni dei quali rimangono comunque soggettivi: «Nei casi in cui una condizione soddisfi i criteri per differenziarsi dalle altre, ma più di uno dei valori è superiore *soltanto di poco* [il corsivo è originale] alla linea di criterio superiore, essa deve essere aumentata del 20%» (Hagopian et al., 1997, p. 325).

Si può sostenere inoltre che il fatto che questi criteri siano stati definiti da «una commissione di esperti» è una garanzia relativa, dal momento che tale commissione era composta da due sole persone, per quanto molto esperte nell'applicazione e nell'interpretazione della metodologia sperimentale.

Scopo del nostro studio era in primo luogo quello di esaminare la validità convergente di una valutazione sperimentale attraverso il confronto di tre metodi di interpretazione dei dati: quelli proposti da Toogood e Timlin (1996) e da Hagopian e colleghi (1997) e una procedura statistica basata sulle probabilità (il metodo del criterio Z). Inoltre, si valutò l'attendibilità test-retest di questi metodi confrontando i risultati ottenuti in momenti diversi. Per quanto è a nostra conoscenza, non sono stati finora condotti altri studi sull'attendibilità test-retest delle valutazioni analoghe.

Metodi

I dati necessari per questa analisi furono raccolti nel corso di una valutazione, durata 16 mesi, dell'esposizione a un ambiente multisensoriale composta da diverse fasi pre e post intervento (Martin, Gaffan e Williams, 1998). Non è stato rilevato alcun effetto dell'intervento, per cui i dati possono essere trattati come una sequenza omogenea.

Partecipanti

Parteciparono allo studio 27 persone adulte (18 maschi e 9 femmine) con ritardo mentale grave o gravissimo che presentavano comportamenti problema (comprese le stereotipie) e provenivano dai servizi locali di Dartford, Kent, in Gran Bretagna. L'età media era di 38 anni (gamma 22-61 anni), con un punteggio medio alla *Vineland Adaptive Behavior Scale* di 9 mesi (gamma 3-21 mesi). Per la partecipazione allo studio venne richiesto il consenso delle famiglie, alle quali fu inviata una lettera nella quale veniva descritto il tipo di ricerca e le procedure che sarebbero state utilizzate. I comportamenti problema (identificati attraverso colloqui preliminari con gli operatori) comprendevano varie topografie di autolesionismo, aggressività, aerofagia (ingestione compulsiva di aria) e comportamenti stereotipati prolungati, come dondolarsi.

Valutazioni

Si effettuò per otto volte, nell'arco di 64 settimane (16 mesi), una valutazione analoga sperimentale, precisamente nelle settimane 1, 4, 20, 24, 40, 44, 60 e 64 (si veda la tabella 1). Si ebbero così quattro coppie di valutazioni, distanziate tra di esse di 2 settimane e dalle coppie precedenti e successive di 16 settimane. Durante gli intervalli di 16 settimane, i partecipanti venivano esposti a un ambiente «terapeutico» multisensoriale o a una condizione di controllo ma, come accennato in precedenza, nessuna delle due condizioni produsse alcun effetto specifico sul loro livello di comportamenti problema (Martin et al., 1998). Ogni valutazione prevedeva la ripetizione per 4 volte di 4 condizioni della durata di 5 minuti, per un totale di 16 condizioni per valutazione con ogni partecipante. Le valutazioni si effettuavano in una stanza tranquilla all'interno del centro diurno oppure dell'istituto in cui era ospite il partecipante. Tutte le valutazioni vennero riprese con la videocamera per facilitare la successiva registrazione dei dati e per calcolare l'accordo tra osservatori.

Le quattro condizioni analoghe erano le seguenti:

1. *da solo:* la persona veniva lasciata da sola nella stanza (era presente soltanto l'operatore con la videocamera);

2. *attenzione contingente:* l'attenzione veniva fornita contingentemente solo in corrispondenza del comportamento problema specifico e consisteva in una richiesta verbale di smetterla, ad esempio: «Non fare così, John»;
3. *attenzione non contingente:* l'attenzione veniva fornita in modo continuo e non contingente per tutta la durata della sessione, sotto forma di una «chiacchierata» verbale (verbalizzazioni non contingenti che non contenevano alcun tipo di richiesta o domanda); questa corrispondeva alla condizione di controllo;
4. *richiesta:* al partecipante veniva continuamente richiesto di terminare qualche compito, come abbinare immagini o comporre un semplice puzzle (il compito di abbinare le immagini veniva utilizzato con i partecipanti che erano in grado di comporre il puzzle senza aiuti). Nel compiere le richieste utilizzavano una sequenza di aiuti graduati; all'inizio erano verbali (ad esempio, «Jim, sai mettere l'immagine dell'autobus al posto giusto?»); poi, se il partecipante non rispondeva entro 10 secondi, la richiesta veniva ripetuta con un aiuto gestuale. Se ancora il partecipante non rispondeva, veniva dato un aiuto fisico per eseguire il compito. A differenza della condizione di richiesta dello studio di Iwata e colleghi (1982), non furono applicate contingenze per nessuno dei comportamenti problema manifestati dal partecipante, allo scopo di eliminare l'eventualità di nuovi apprendimenti, ovvero di evitare che la risposta venisse determinata da quella contingenza (aspetto criticabile del metodo utilizzato da Iwata e colleghi, come spiegato nell'introduzione).

Si fece in modo che in tutte le valutazioni (nelle quali veniva ripetuta per 4 volte ognuna delle 4 condizioni analoghe) la sequenza delle condizioni fosse sempre diversa.

Accordo tra osservatori

Per ogni partecipante furono registrati fino a tre comportamenti problema e/o stereotipie e i controlli dell'accordo tra osservatori furono effettuati sul 20% dei

TABELLA 1
Distribuzione nel tempo e raggruppamento delle valutazioni

Periodo	*Valutazione**	*Periodo composito di valutazione*
Settimana 1	1	1
Settimana 4	2	
Settimana 20	3	2
Settimana 24	4	
Settimana 40	5	3
Settimana 44	6	
Settimana 60	7	4
Settimana 64	8	

* Quattro ripetizioni di quattro condizioni.

dati di osservazione dal primo autore e da un assistente addestrato in questa verifica, utilizzando dati che non facevano parte del campione selezionato per la verifica dell'attendibilità. Essa fu calcolata sommando gli accordi e i disaccordi sulla presenza di comportamenti problema all'interno di ogni intervallo di 5 secondi (Martin, Oliver e Hall, 1996). I valori di attendibilità media furono calcolati considerando tutte le registrazioni dei due osservatori, per tutte le variabili codificate. La percentuale di accordo globale superava il 97% in tutte le variabili e i valori erano compresi tra 0,74 e 0,77 (Cohen, 1960; 1968). Le somme dei dati per l'attendibilità (totale di accordi e disaccordi di tutte le registrazioni dei due osservatori) avevano un accordo percentuale generale compreso tra il 97,31% e il 99,96% e valori compresi tra 0,78 e 0,90.

Interpretazione dei dati

I dati delle otto valutazioni furono raggruppati in 4 periodi compositi di valutazione (si veda la tabella 1), ognuno dei quali comprendeva una coppia di valutazioni analoghe distanziate da due settimane senza intervento (valutazioni 1 e 2, 3 e 4, 5 e 6, 7 e 8). Ognuno di questi quattro periodi di valutazione avveniva a distanza di 16 settimane ed era composto da 8 ripetizioni di 4 condizioni analoghe (per un totale di 32 valori).

Per ogni condizione analoga dei 4 nuovi periodi di valutazione si calcolarono i punteggi Z relativi a tutti i comportamenti problema (comprese le stereotipie) per tutti i partecipanti. Dai dati di ognuna delle 32 condizioni furono calcolate anche le frequenze al minuto, rappresentate sequenzialmente nel grafico della figura 1. Si applicò un criterio che includeva nell'analisi soltanto i periodi di valutazione che contenevano più di tre valori (su 32) superiori a 0. Con questo criterio si ottennero 152 periodi di valutazione validi dei comportamenti problema di 25 partecipanti, usati poi per il confronto dei tre metodi.

Il metodo statistico proposto per l'interpretazione dei dati sperimentali (il metodo del criterio Z) è il seguente. La distribuzione nulla di Z ha una media di 0 e una deviazione standard di 1. Se i punteggi Z utilizzati per ogni valutazione sono tratti da *n* ripetizioni di ciascuna condizione, la deviazione standard attesa della media Z nell'ipotesi nulla è $DS_2 = DS_1/n$, dove DS_1 è 1 e DS_2 è l'errore standard della media Z.

Perché un punteggio Z sia significativamente superiore a 0 (per un test unidirezionale) la sua probabilità deve essere inferiore o pari a 0,5. Dalle condizioni analoghe C, si testa la significatività dei valori Z C-1 (l'ultimo punteggio Z viene determinato quando gli altri sono noti); perciò, la probabilità α deve essere adattata (correzione di Bonferroni) per evitare la possibilità di attribuire erroneamente la significatività (errore di tipo 1).

$$p_2 = p_1 / C - 1$$

P_1 rappresenta qui il valore α convenzionale (0,05) e p_2 rappresenta il nuovo α modificato. Per un test unidirezionale, il valore di Z che corrisponde a p_2 può essere trovato nella tabella corrispondente (Z_1), ma questo Z si verifica in una distribuzione con deviazioni standard uguali a DS_2; pertanto, il punteggio Z medio richiesto per la significatività è $Z_{crit} = (Z_1 \times DS_2)$.

Z_{crit} rappresenta qui la media Z di *n* ripetizioni significativamente superiori a 0. Il metodo del criterio Z per differenziare una condizione analoga e attribuire una funzione, perciò, implica che lo Z medio sia uguale o superiore a Z_{crit}. Occorre notare, tuttavia, che nella derivazione di questo valore si presupponeva che: primo, i dati grezzi (dati di frequenza percentuale dai quali vengono calcolati i punteggi Z) avessero una distribuzione normale; secondo, la DS globale tra tutte le ripetizioni fornisse una buona stima delle DS nelle singole ripetizioni (quelle che si sarebbero derivate da una ripetizione).

Si effettuò un confronto tra il metodo del criterio Z, una forma leggermente modificata del metodo di Hagopian e colleghi e quello di Toogood e Timlin, utilizzando i dati di tutti i partecipanti. La forma modificata del metodo di Hagopian e colleghi consisteva semplicemente nell'usare il loro criterio principale per la differenziazione delle condizioni, ovvero almeno metà dei valori della condizione dovevano essere al di sopra della linea di criterio. Ogni eventuale tendenza evidenziata nei dati veniva ignorata per la potenziale ambiguità dell'interpretazione; vale a dire, le tendenze crescenti potevano rispecchiare semplicemente l'acquisizione della funzione (una delle critiche avanzate all'uso delle contingenze nelle condizioni analoghe) e quelle decrescenti potevano riflettere l'estinzione della funzione precedente.

Furono ignorati anche i criteri di Hagopian e colleghi per la differenziazione multipla delle condizioni, perché non si disponeva di valori sufficienti (solo 8 anziché 10) il che rendeva il metodo modificato meno sensibile alle differenze tra condizioni; in altre parole, se c'erano più condizioni differenziate veniva attribuita una funzione multipla. Lo stesso criterio principale fu applicato anche nell'uso del metodo di Toogood e Timlin, per cui ai fini del confronto l'unica differenza tra le metodologie delle due équipe di autori era la posizione della linea di criterio.

Le linee di criterio previste dal metodo sia di Hagopian e colleghi che di Toogood e Timlin furono tracciate sul grafico dei dati relativi alle frequenze per minuto, così che il numero di valori di ogni condizione analoga collocati al di sopra di essa potesse essere identificato attraverso l'esame visivo (si veda la figura 1). La linea di criterio 1 è quella derivata dal metodo di Hagopian e colleghi e fu stabilita in base al livello di risposta nella condizione di controllo (condizione di attenzione non contingente), collocandosi così a circa una DS sopra la media della condizione di controllo. Ipotizzando una distribuzione normale e 8 valori per condizione, la linea di criterio 1 fu tracciata tra il primo e il secondo valore più alto della condizione di controllo in quel periodo di valutazione.

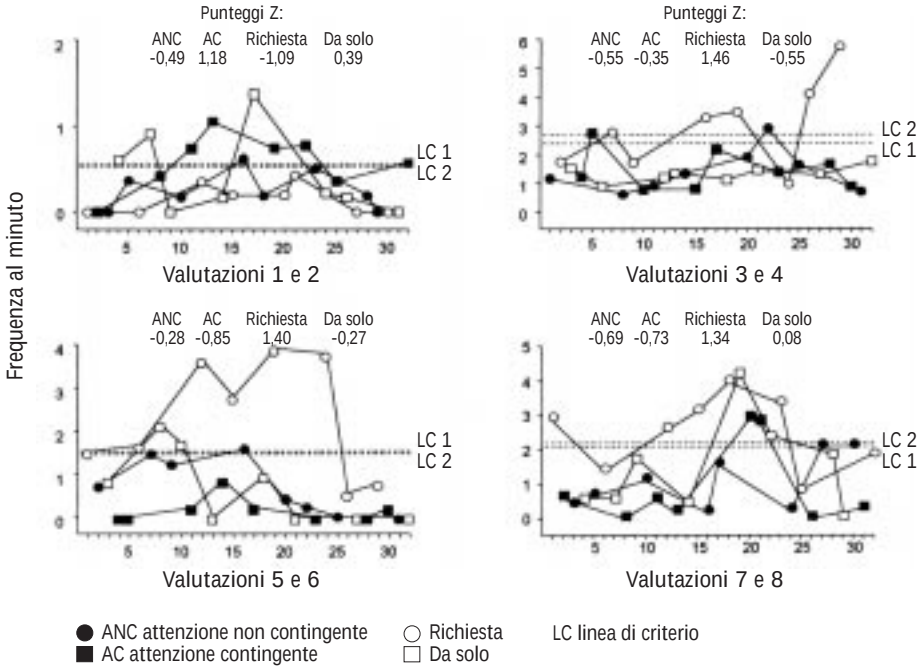

Fig. 1 Frequenza al minuto dei comportamenti problema e punteggi Z medi per condizione nei quattro periodi compositi di valutazione. I dati sono riferiti a uno dei partecipanti.

La linea di criterio 2 rappresenta quella del metodo di Toogood e Timlin e fu tracciata al livello del 50% superiore alla media generale di tutte le condizioni e sessioni del periodo di valutazione considerato.

Utilizzando la formula del metodo del criterio Z, con 8 ripetizioni di ogni condizione in ciascun periodo di valutazione si ottiene

$$DS_2 = DS_1/ \sqrt{n} = 1/ \sqrt{8} = 0{,}354$$

Dalle 4 condizioni analoghe si valuta la significatività di tre valori Z:

$$p_2 = p_1/ n - 1 = 0{,}05 / 3 = 0{,}017$$

Per un test unidirezionale, il valore di Z che corrisponde a p_2 è 2,12, ma questo Z si ha su una distribuzione con DS_2 uguale a 0,354; perciò, il punteggio Z medio richiesto per la significatività (p = 0,017, unidirezionale) è:

$$Z_{crit} = (Z_1 \times DS_2) = (2{,}12 \times 0{,}354) = 0{,}75$$

Se il criterio di uno qualsiasi dei metodi veniva soddisfatto da una qualsiasi delle condizioni analoghe, essa si considerava differenziata. Se la condizione di

attenzione contingente si differenziava, si attribuiva al comportamento problema una funzione mantenuta dall'attenzione; se si differenziava la condizione di richiesta, si attribuiva la funzione di fuga; se si differenziava quella «da solo», si attribuiva una funzione di autostimolazione/rinforzamento automatico; se in tutte le condizioni i valori al di sopra della linea di criterio (secondo il metodo sia di Hagopian e colleghi che di Toogood e Timlin) erano meno della metà, non si avevano condizioni differenziate e non si attribuiva alcuna funzione; infine, se il criterio per la differenziazione era soddisfatto da più condizioni, si attribuiva al comportamento una funzione multipla.

Risultati

Validità convergente

La figura 1 mostra alcuni esempi di grafici e i relativi punteggi Z medi di ognuno dei 4 periodi di valutazione dei comportamenti problema di un partecipante. In questo caso specifico, i tre metodi convergono nell'identificare la funzione all'interno di ciascuna valutazione, ma la funzione attribuita nelle quattro valutazioni non è la stessa. Il primo grafico della figura 1 (angolo in alto a sinistra) indica una funzione mantenuta dall'attenzione: 5 degli 8 valori nella condizione di attenzione contingente si collocano al di sopra delle linee di criterio e il punteggio Z medio di questa condizione — 1,18 — è superiore al criterio Z di 0,75. Tutti gli altri tre grafici suggeriscono una funzione di fuga: quattro o più valori della condizione di richiesta si collocano al di sopra di entrambe le linee di criterio e i punteggi Z medi di questa condizione (1,46, 1,40 e 1,34) sono superiori al criterio. Inoltre, le altre condizioni non forniscono un punteggio Z superiore al criterio in nessuna delle valutazioni. Come si possano creare queste incongruenze nell'attribuzione della funzione verrà discusso più avanti.

La differenza tra le linee di criterio C1 e C2 nel grafico presentato come esempio (si veda la figura 1) è relativamente contenuta. In due dei grafici, la linea di criterio 1 è maggiore della 2, mentre negli altri due grafici è il contrario. La differenza media tra le due linee di criterio era, in generale, di 3,71 (frequenza di comportamenti problema al minuto) e la differenza mediana di 0,81. Perciò, le differenze nel caso specifico di questa persona sono alquanto inferiori alla media ma, in generale, la linea di criterio 1 era superiore alla 2 nel 55% delle valutazioni, per cui nessuna delle due linee di criterio tendeva a essere superiore all'altra.

Sebbene la figura 1 mostri un caso nel quale le funzioni potevano essere attribuite, accadeva più spesso che nessuna delle condizioni raggiungesse il criterio necessario a questo scopo. La tabella 2 fornisce il numero di valutazioni nelle quali l'identificazione della funzione fu possibile o meno e il numero di quelle in cui i tre metodi concordarono nell'attribuirla. Per questo confronto si utilizzaro-

no 152 periodi di valutazione effettuati con 25 persone; tuttavia, il numero di funzioni spiegate (attribuite o altro) non raggiunge 152 né nel caso del metodo del criterio Z né in quello del metodo di Toogood e Timlin. Ciò è dovuto al fatto che in alcune occasioni la condizione di controllo risultò differenziata (attenzione non contingente) e non fu dunque possibile attribuire la funzione. Questo non succede quando si utilizza il metodo di Hagopian e colleghi, perché le condizioni si considerano differenziate solo in presenza di un numero sufficiente di valori che rientra nella gamma più alta della condizione di attenzione non contingente.

I dati evidenziano uno scarso accordo tra i metodi nell'identificare le funzioni di fuga e mantenuta dall'attenzione (là dove sono queste le funzioni identificate); tuttavia, la maggior parte dell'accordo si verifica nell'incapacità di individuare la funzione del comportamento problema.

La tabella 2 mostra che tutti e tre i metodi non riescono a identificare la funzione in molti dei periodi di valutazione: quello del criterio Z riesce nel 58% dei casi, quello di Hagopian e colleghi solo nel 32% e quello di Toogood e Timlin nel 45%. È interessante il fatto che Toogood e Timlin stessi, nell'analisi funzionale in cui applicano il loro metodo a 121 comportamenti, riferiscano una percentuale di successo del 41% (la percentuale di comportamenti per i quali riuscirono ad attribuire la funzione), un dato che collima con quello da noi riscontrato. Iwata e colleghi, dall'altro lato, affermano di essere riusciti a individuare la funzione nel 95,4% dei loro 152 casi attraverso il solo esame visivo dei grafici dei dati. Hagopian e colleghi (1997) non forniscono cifre riguardo alla proporzione di comportamenti spiegati, ma solo il grado di accordo tra osservatori. Nella tabella 2 è presentato anche un indice dell'accordo (k di Cohen) tra coppie di metodi; i valori suggeriscono una scarsa congruenza generale tra ogni coppia; l'accordo relativamente maggiore è quello tra i metodi di Hagopian e colleghi e Toogood e Timlin (0,54) e quello minore tra il metodo del criterio Z e quello di Hagopian e colleghi (0,42).

TABELLA 2

Numero e percentuali di valutazioni nelle quali i metodi identificarono la funzione e in cui concordano sul comportamento problema

	Mantenuto dall'attenzione	*Fuga*	*Automatico/ autostimolazione*	*Funzione multipla*	*Non identificata*	*K di Cohen*
Criterio Z (metodo 1)	24 (18%)	18 (14%)	33 (25%)	1 (1%)	56 (42%)	
Hagopian et al. (metodo 2)	15 (10%)	12 (8%)	6 (5%)	14 (9%)	105 (68%)	
Toogood e Timlin (metodo 3)	19 (13%)	13 (9%)	21 (15%)	11 (8%)	78 (55%)	
Accordi (metodi 1 e 2)	13	9	5	0	50	0,42
Accordi (metodi 1 e 3)	15	10	15	0	42	0,47
Accordi (metodi 2 e 3)	14	10	4	2	72	0,54
Accordi (metodi 1, 2 e 3)	12	8	3	0	41	

Attendibilità test-retest in 16 settimane (tutti e tre i metodi)

La stabilità nel corso del tempo dei tre metodi confrontati fu valutata verificando l'accordo sulle funzioni identificate nei 4 periodi compositi di valutazione: il primo con il secondo, il secondo con il terzo e il terzo con il quarto, ognuno dei quali distanziato di 16 settimane. Per esempio, se il metodo del criterio Z individuava una funzione di fuga nei periodi di valutazione 1 e 2, si registrava accordo sulla funzione identificata (si veda la tabella 3, prima riga). Se invece né la valutazione 1 né la 2 fornivano una funzione, si registrava accordo nell'incapacità di individuare la funzione (seconda riga). Se i due periodi di valutazione individuavano funzioni diverse, si registrava disaccordo (terza riga).

Gli stessi dati utilizzati per confrontare i tre metodi (si veda la tabella 2) furono poi usati anche in questa analisi, però solo quando entrambi i periodi di valutazione confrontati non erano stati esclusi dal criterio descritto sopra. Gli accordi sulla funzione tra periodi di valutazione, per tutti e tre i metodi, sono presentati nelle tabelle 3, 4 e 5. I dati di queste tabelle mostrano differenze notevoli fra i tre metodi. Il metodo del criterio Z è quello che presenta il maggior grado di accordo medio sulla funzione nelle diverse valutazioni (27%, a fronte del 12% del metodo di Toogood e Timlin e del 6% di quello di Hagopian e colleghi). Tuttavia, il metodo del criterio Z presenta l'accordo medio minore nei casi di incapacità di identificare la funzione e il livello più alto di disaccordo. Visto che i dati sull'attendibilità test-retest dei tre metodi nell'arco di 16 settimane erano contraddittori — e in generale indicavano valori bassi —, decidemmo di indagare ulteriormente in questa direzione utilizzando periodi più brevi. Ciò non fu possibile con il metodo di Hagopian e colleghi, per via dell'insufficienza

TABELLA 3

Numero di accordi e accordo percentuale sulle funzioni identificate con il metodo del criterio Z tra un periodo e quello successivo e accordo medio generale

Criterio Z	*Valutazioni 1 e 2*	*Valutazioni 2 e 3*	*Valutazioni 3 e 4*	*Media*
Accordi (funzione identificata)	8 (25%)	11 (32%)	8 (24%)	27%
Accordi (funzione non identificata)	5 (16%)	2 (6%)	2 (6%)	9%
Disaccordi	19 (59%)	21 (62%)	24 (70%)	64%

TABELLA 4

Numero di accordi e accordo percentuale sulle funzioni identificate con il metodo di Hagopian e colleghi tra un periodo e quello successivo e accordo medio generale

Hogopian et al.	*Valutazioni 1 e 2*	*Valutazioni 2 e 3*	*Valutazioni 3 e 4*	*Media*
Accordi (funzione identificata)	0 (0%)	2 (6%)	4 (12%)	6%
Accordi (funzione non identificata)	18 (56%)	18 (53%)	14 (41%)	50%
Disaccordi	14 (44%)	14 (41%)	16 (47%)	44%

TABELLA 5
Numero di accordi e accordo percentuale sulle funzioni identificate con il metodo di Toogood e Timlin tra un periodo e quello successivo e accordo medio generale

Toogood e Timlin	*Valutazioni 1 e 2*	*Valutazioni 2 e 3*	*Valutazioni 3 e 4*	*Media*
Accordi (funzione identificata)	4 (12%)	3 (9%)	5 (15%)	12%
Accordi (funzione non identificata)	12 (38%)	9 (26%)	9 (26%)	30%
Disaccordi	16 (50%)	22 (65%)	20 (59%)	58%

dei valori disponibili (quattro per l'analisi dopo un mese e due per l'analisi dopo una settimana).

Attendibilità dopo un mese e dopo una settimana (metodi del criterio Z e di Toogood e Timlin)

Per ogni condizione di ognuna delle 8 valutazioni iniziali (si veda la tabella 1) furono calcolati i punteggi Z relativi a tutti i comportamenti problema di tutti i partecipanti. Ogni punteggio Z, perciò, fu derivato dalle quattro repliche delle condizioni nel periodo di una settimana. Si calcolarono gli accordi sulla funzione attribuita, quelli parziali, quelli sull'incapacità di identificare la funzione e i disaccordi nelle quattro coppie di valutazioni consecutive (valutazioni 1 e 2, 3 e 4, 5 e 6, 7 e 8; le valutazioni consecutive venivano effettuate in un periodo di 4 settimane, per cui i dati possono essere considerati rappresentativi dell'attendibilità dopo un mese). Gli accordi parziali si avevano quando due valutazioni consecutive attribuivano funzioni multiple e c'era accordo soltanto su una o più di esse ma non su tutte. I dati per i quali non era possibile calcolare il punteggio Z furono esclusi dall'analisi.

Per quanto riguarda il metodo di Toogood e Timlin, ogni valutazione comprendeva quattro valori, corrispondenti a ognuna delle quattro repliche di una stessa condizione. Almeno due dei valori di ciascuna condizione analoga dovevano collocarsi al di sopra della linea di criterio, cioè del 50% al di sopra della media generale di tutte le condizioni e i valori. Gli accordi sulla funzione, gli accordi parziali, quelli nell'incapacità di identificarla e i disaccordi furono calcolati per le quattro coppie di valutazioni consecutive come descritto sopra.

Per valutare l'attendibilità dopo una settimana, furono calcolati i punteggi Z per ogni condizione analoga relativamente alle prime e ultime due ripetizioni della stessa condizione nell'arco di una settimana. Come nell'analisi precedente, furono calcolati gli accordi sulla funzione, gli accordi parziali, quelli nell'incapacità di identificarla e i disaccordi tra queste coppie di valutazioni consecutive effettuate nell'arco di una settimana (otto coppie per partecipante per comportamento problema; anche in questo caso parte dei dati fu esclusa per l'impossibilità di calcolare il punteggio Z).

Riguardo al metodo di Toogood e Timlin, in questo caso ogni valutazione forniva soltanto due valori per ognuna delle quattro ripetizioni della stessa condizione e, perché una condizione si differenziasse, almeno uno dei valori di una qualsiasi delle condizioni doveva porsi al di sopra della linea di criterio.

È importante notare che i punteggi Z di criterio utilizzati per attribuire la funzione ai fini dell'attendibilità dopo una settimana e dopo un mese erano diversi da quelli usati prima, dove la DS di 0,354 era stata derivata da otto ripetizioni. Utilizzando la formula $DS_2 = DS_1/n$, la DS_2 dopo un mese diventa, con 4 ripetizioni, 0,5 mentre la DS_2 dopo una settimana, con due ripetizioni, diventa 0,707. I nuovi punteggi Z di criterio adattati vengono derivati, come in precedenza, moltiplicando DS_2 per 2,12:

$$Z_{crit} \text{ (dopo un mese)} = 1{,}06$$
$$Z_{crit} \text{ (dopo una settimana)} = 1{,}50$$

I dati sull'accordo per le analisi dopo un mese e dopo una settimana sono presentati nelle tabelle 6 e 7.

I dati presentati nella tabella 6, relativi al metodo del criterio Z, mostrano che, man mano che il periodo tra valutazioni si riduce, vi sono meno disaccordi, il numero di comportamenti per i quali si riesce a identificare la funzione si riduce e che, parallelamente, si ha un aumento del numero di comportamenti per i quali vi è accordo nell'incapacità di attribuire la funzione. Sebbene nell'analisi

TABELLA 6

Attendibilità dopo un mese e dopo una settimana del metodo del criterio Z: numero e percentuale di accordi sulla funzione

Criterio Z	*Accordi (funzione identificata)*	*Accordi (funzione non identificata)*	*Accordi parziali*	*Disaccordi*
Attendibilità dopo un mese	39 (25%)	41 (26%)	0 (0%)	76 (49%)
Attendibilità dopo una settimana	7 (3%)	224 (86%)	0 (0%)	30 (11%)

TABELLA 7

Attendibilità dopo un mese e dopo una settimana del metodo di Toogood e Timlin: numero e percentuale di accordi sulla funzione

Toogood e Timlin	*Accordi (funzione identificata)*	*Accordi (funzione non identificata)*	*Accordi parziali*	*Disaccordi*
Attendibilità dopo un mese	25 (16%)	37 (23%)	13 (8%)	83 (53%)
Attendibilità dopo una settimana	65 (25%)	16 (6%)	50 (19%)	131 (50%)

dopo un mese vi siano più disaccordi, in termini di capacità di identificare la funzione i dati di questo esame sembrano fornire i risultati migliori.

I dati presentati nella tabella 7, relativi al metodo di Toogood e Timlin, mostrano risultati per certi aspetti diversi. Man mano che il periodo tra valutazioni si riduce, ci sono più accordi sulle funzioni attribuite e contemporaneamente ce ne sono meno nell'incapacità di identificarle.

Discussione

Validità convergente: confronto dei metodi di interpretazione

In questo studio furono confrontati tre diversi metodi per l'interpretazione dei dati per esaminare la validità convergente di una valutazione sperimentale. Tutti i metodi risultarono allo stesso modo inefficaci nell'identificazione della funzione dei comportamenti problema.

Un'incapacità così diffusa nell'individuare la funzione potrebbe essere dovuta a difetti intrinseci dei metodi di interpretazione. Per esempio, i metodi di discriminazione delle funzioni potrebbero non essere sensibili verso i comportamenti con bassa frequenza. Molti dei comportamenti manifestati dai partecipanti a questo studio si presentavano con una frequenza relativamente ridotta (la percentuale media di emissione di comportamenti problema da parte di tutti i partecipanti in tutte le sessioni era del 5,2%, con una mediana di 1,3%), mentre i partecipanti allo studio di Iwata e colleghi (1982) mostravano comportamenti autolesionistici in una percentuale media di intervalli pari al 30,8% (mediana = 17,5%). È anche possibile che il metodo delle condizioni analoghe di per se stesso non rilevi i comportamenti a bassa frequenza. Quando la frequenza del comportamento considerato è bassa, se esso si presenta in più condizioni è più difficile differenziarle. Ciò potrebbe essere dovuto a un effetto di novità (ad esempio, un effetto prodotto semplicemente dalla presenza dello sperimentatore e dell'operatore della videocamera e dall'interazione con lo sperimentatore) o di estinzione (il comportamento continua anche nella condizione successiva nonostante l'assenza degli antecedenti e/o delle conseguenze che lo mantengono); oppure, potrebbe essere il risultato di contingenze di nuova acquisizione dopo la reiterata esposizione a ripetute condizioni analoghe specifiche (una critica avanzata alla metodologia utilizzata da Iwata et al., 1982; 1990; 1994).

Un'altra possibile spiegazione dell'incapacità, in molti casi, dei metodi di interpretazione descritti nell'attribuire la funzione ai comportamenti problema potrebbe essere il fatto che i dati contengono troppe interferenze (fluttuazione casuale). Un modo per verificare questa ipotesi sarebbe quello di vedere se perlomeno i risultati di uno dei metodi siano relativamente stabili nel corso del tempo. Per esempio, se il repertorio comportamentale di una persona è limitato, i cambia-

menti nelle contingenze ambientali potrebbero far sì che uno stesso comportamento venga emesso per ragioni molto diverse.

Per quanto riguarda i vantaggi e gli svantaggi relativi dell'uso dell'uno anziché dell'altro tra i metodi descritti, quello del criterio Z appare preferibile per una serie di ragioni: esso, in generale, rispetto agli altri metodi, mostra una maggiore capacità di fornire risultati in termini di attribuzione della funzione; indica la significatività statistica della differenziazione delle condizioni; tiene in considerazione la variabilità delle frequenze utilizzando i punteggi Z; basandosi su un sistema matematico, elimina in parte l'arbitrarietà nell'identificazione delle funzioni.

Attendibilità test-retest nell'arco di 16 settimane

L'attendibilità test-retest nell'arco di 16 settimane di tutti e tre i metodi risultò scarsa. È possibile che l'intervallo di tempo tra i periodi di valutazione utilizzato in questo studio fosse troppo lungo e che nel frattempo si verificassero dei cambiamenti nella funzione (forse a causa di fattori ambientali come cambiamenti nell'équipe o nella routine quotidiana). Se questa ipotesi fosse corretta, avrebbe naturalmente implicazioni importanti nel caso in cui si volesse sviluppare un intervento sulla base di analisi funzionali eseguite tempo addietro. Nell'arco delle 16 settimane, il metodo del criterio Z fornì la percentuale maggiore di accordo medio (tra tutte le valutazioni) sulla funzione attribuita; tuttavia, evidenziò anche la proporzione più bassa di accordo medio nell'incapacità di attribuire la funzione e la proporzione più alta di disaccordo medio.

Attendibilità dopo un mese e dopo una settimana

Questi risultati sembrano evidenziare una differenza tra i due metodi considerati (del criterio Z e di Toogood e Timlin), nonostante si riscontrino delle somiglianze nelle analisi dopo un mese (queste potrebbero essere stime più attendibili perché coinvolgono una maggiore quantità di dati). L'interpretazione della funzione attraverso il metodo del criterio Z appare più precisa nei periodi brevi; ciò può essere dovuto al fatto che il punteggio Z di criterio aumenta quando si utilizza un numero minore di ripetizioni delle condizioni, per cui i punteggi Z medi (e, di conseguenza, i dati grezzi di ciascuna condizione) devono presentare un grado di differenziazione maggiore per essere significativi. Il metodo di Toogood e Timlin, dall'altro lato, risulta meno preciso quando si usano poche ripetizioni, perché per la differenziazione di una condizione richiede che solo uno dei due valori dell'analisi su una settimana sia uguale o superiore alla linea di criterio.

I risultati, perciò, potrebbero suggerire che entrambi i metodi sono più accurati se applicati per un periodo ottimale di tempo.

L'attendibilità test-retest generalmente scarsa potrebbe essere dovuta a una serie di ragioni: nel corso del tempo i metodi potrebbero non essere più attendibili oppure le funzioni di specifici comportamenti potrebbero cambiare; in alternativa, i dati potrebbero variare troppo — sia all'interno di una stessa valutazione che da una valutazione all'altra — per essere in qualche modo utili. Le conclusioni che possiamo trarre dall'analisi dell'attendibilità sono che il metodo del criterio Z potrebbe essere quello più utile per l'interpretazione e che può esistere una durata ottimale di effettuazione delle valutazioni.

Le analisi dell'attendibilità test-retest dopo un mese confrontavano due valutazioni, ognuna delle quali comprendeva quattro ripetizioni di quattro condizioni; i dati suggeriscono che questa valutazione forniva le informazioni più utili e dei risultati confrontabili con quelli della valutazione successiva. Perciò, la durata ottimale di una valutazione potrebbe essere quattro ripetizioni di ogni condizione. Una simile conclusione porterebbe a privilegiare l'uso di valutazioni alquanto brevi.

Conclusioni

L'interpretazione di dati derivati da valutazioni sperimentali, in termini di identificazione delle possibili funzioni del comportamento in esame, appare ampiamente contraddittoria, dato l'uso, da parte dei vari ricercatori, di modalità interpretative e metodologie differenti.

Applicando le procedure standard (si veda Iwata et al., 1982; 1990) un assessment sperimentale comprende la valutazione ripetuta di una persona con tre o quattro possibili situazioni analoghe con specifiche contingenze fino a quando non si evidenzia quale di esse determina il livello più elevato di comportamenti problema. Alcuni sostenitori dell'uso dell'assessment sperimentale (ad esempio Iwata e colleghi) proseguono la valutazione fino a quando i dati si stabilizzano. Ciò può comportare due problemi: primo, l'esposizione ripetuta alle contingenze può determinare nuovi apprendimenti, soprattutto quando vengono utilizzate nelle condizioni analoghe; secondo, la funzione effettiva del comportamento problema potrebbe cambiare nel corso del tempo.

In questo studio, eliminammo il primo problema evitando l'uso di contingenze nelle condizioni analoghe — fatta eccezione per l'attenzione contingente — ed effettuando un numero limitato di ripetizioni all'interno del periodo di valutazione (solo 4 ripetizioni per valutazione). Il secondo è un problema che non può essere risolto direttamente e che potrebbe precludere il ricorso a valutazioni ampie e prolungate. Per esempio, se una valutazione sperimentale si protrae per un periodo di tempo (settimane o mesi) c'è l'effettiva possibilità che la funzione del comportamento cambi, magari a causa di eventi ambientali: cambiamenti di ambiente, di personale, ecc. Questa probabilità può aumentare se la persona

viene esposta ripetutamente a nuove contingenze che creano un particolare paradigma di rinforzamento (Skinner, 1948; Vollmer, Marcus e LeBlanc, 1994). In alternativa, se il comportamento ha più funzioni, le valutazioni prolungate mostreranno un grado elevato di variabilità nelle diverse condizioni, che potrebbero condurre, nelle varie fasi della valutazione, a conclusioni diverse.

Riassumendo, questo studio sull'uso della metodologia sperimentale intendeva rispondere a due principali domande:

1. l'uso di metodi diversi per interpretare i risultati di una valutazione analoga porta a identificare la stessa funzione?
2. l'uso della valutazione analoga fornisce sempre gli stessi risultati?

Dai dati ottenuti, la risposta a entrambe le domande sembra essere negativa. Dato il numero di partecipanti, la gamma di comportamenti e il periodo abbastanza lungo in cui furono effettuate le valutazioni, questo risultato solleva questioni importanti sull'utilità delle analisi funzionali sperimentali nella ricerca applicata e nella pratica clinica. È interessante notare che in una recente rassegna sull'uso dell'analisi funzionale nella valutazione dei comportamenti (Cone, 1997) non veniva fatto cenno né alla validità e attendibilità delle valutazioni analoghe né alla questione della significatività applicata/clinica.

Non è possibile pervenire a conclusioni definite riguardo ai risultati insoddisfacenti di questo studio, per via della varietà di modi in cui le valutazioni analoghe sono state effettuate. La metodologia sperimentale utilizzata nella nostra ricerca è solo una delle tante descritte in letteratura. Per evidenziare le differenze e gli aspetti in comune tra le varie metodologie, ne presentiamo alcuni esempi nella tabella 8, che illustra alcune differenze che potrebbero essere importanti tra la metodologia sperimentale utilizzata in questo studio e altre descritte in letteratura. Gli esempi forniti sono soltanto alcuni dei tanti possibili, scelti per mostrare il grado di variabilità nella ricerca. Le differenze principali sembrano essere quelle relative all'uso di un campione selezionato (cioè composto di persone che fruiscono di servizi specifici proprio per la frequenza e/o in-

TABELLA 8
Differenze tra varie applicazioni della metodologia sperimentale

	Carr e Durand (1985)	*Iwata et al. (1982)*	*Martin, Gaffan e Williams (questo studio)*	*Northup et al. (1991)*	*Toogood e Timlin (1995)*
Dimensioni del campione	4	9	27	3	20
Campione selezionato	Sì	Sì	No	Sì	No
Numero di condizioni	3	3-4	4	2-4	5
Durata delle condizioni (minuti)	10	15	5	5-10	5-10
Numero di ripetizioni	6-18	6-14	32	1	4
Ordine delle condizioni	Fisso	Bilanciato	Bilanciato	Fisso	Bilanciato
Uso di contingenze	No	Sì	No*	Sì	Sì

* Eccetto l'attenzione contingente

tensità dei loro comportamenti problema), la durata delle condizioni e l'uso di contingenze nelle condizioni sperimentali. Un articolo recente di Smith e Iwata (1997) suggerisce che l'identificazione della funzione di un comportamento problema non è possibile senza la manipolazione sperimentale delle variabili importanti; di qui l'uso di contingenze per i comportamenti problema nelle condizioni sperimentali dello studio di Iwata e colleghi. Anche il numero di ripetizioni delle condizioni varia, ma non è sempre possibile stabilire, sulla base degli studi pubblicati, il periodo di tempo nel corso del quale tali ripetizioni si distribuiscono.

La risposta negativa alle domande indicate sopra e il gran numero di varianti nella metodologia sperimentale portano a chiedersi se la differenza nelle risposte possa derivare dall'uso di metodologie diverse. La ricerca futura sulla validità delle valutazioni sperimentali potrebbe mantenere costanti alcune variabili potenzialmente importanti (sintetizzate nella tabella 8) e contemporaneamente manipolare sistematicamente singole componenti. Alcune domande importanti alle quali la ricerca futura potrebbe cercare risposta sono le seguenti:

– le valutazioni sperimentali sono adatte più ai comportamenti frequenti che non a quelli a bassa emissione?
– quali caratteristiche dovrebbero avere le singole condizioni sperimentali e quante volte dovrebbero essere ripetute?
– è necessario che tutte le condizioni sperimentali applichino delle contingenze all'emissione dei comportamenti problema in esame?

Naturalmente, è possibile che molte delle variabili menzionate sopra siano interdipendenti e che non si possa effettuare una valutazione sperimentale standardizzata su qualsiasi comportamento. Potrebbe essere più utile, e più corretto, utilizzare una piccola serie di ipotesi dirette a valutare le condizioni analoghe (in modo simile al disegno sperimentale di Northup et al., 1991, dove solo una serie di condizioni sperimentali era seguita da una fase di inversione delle contingenze, usando altre tre condizioni per validare i risultati ottenuti) e sviluppare tali condizioni tenendo conto soltanto delle caratteristiche individuali e peculiari del comportamento problema specifico. Ciò appare più congruo in termini sia di praticità che di accuratezza della valutazione; nel lavoro clinico individuale, più che nella ricerca, è auspicabile che la valutazione tenti di verificare il maggior numero possibile di ipotesi (Carr, Yarbrough e Langdon, 1997). Vi sono casi di valutazioni sperimentali, utilizzate nel contesto della ricerca, che comprendevano condizioni specifiche dirette a verificare ipotesi molto inconsuete — ad esempio, fuga dal rumore dell'ambiente, dalle visite mediche (Iwata et al., 1994) — ma si tratta più di eccezioni che della regola. La ricerca futura potrebbe anche valutare l'ipotesi che i risultati ottenuti da un'analisi funzionale utilizzando valutazioni analoghe brevi siano altrettanto efficaci delle valutazioni più lunghe nell'identificare la funzione dei comportamenti problema. I parametri precisi che stabiliscono sia il breve che il lungo termine devono essere definiti e specificati.

Probabilmente, la migliore indicazione della validità dei risultati che si otterranno in studi futuri sull'uso delle valutazioni sperimentali è quella fornita dalla valutazione dei risultati dopo l'intervento (cosa che non fu possibile in questo studio). Un intervento può essere infatti considerato efficace, e la validità dell'analisi funzionale realizzata può essere dimostrata, quando si verifica una riduzione significativa della frequenza e/o durata dei comportamenti problema.

TITOLO ORIGINALE

Experimental functional analyses for challenging behavior: A study of validity and reliability. Tratto da «Research in Developmental Disabilities», vol. 20, n. 2, 1999. © Elsevier Science Ltd. Pubblicato con il permesso dell'Editore. Traduzione italiana di Serena Banal.

Bibliografia

Carr E.G. (1994), *Emerging themes in the functional analysis of problem behaviour*, «Journal of Applied Behavior Analysis», vol. 27, pp. 393-399.

Carr E.G. e Durand V.M. (1985), *Reducing behavior problems through functional communication training*, «Journal of Applied Behavior Analysis», vol. 18, pp. 111-126.

Carr E.G., Yarbrough S.C. e Langdon N.A. (1997), *Effects of idiosyncratic stimulus variables on functional analysis outcomes*, «Journal of Applied Behavior Analysis», vol. 30, pp. 673-685.

Cohen J. (1960), *A coefficient of agreement for nominal scales*, «Educational and Psychological Measurement», vol. 20, pp. 37-46.

Cohen J. (1968), *Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit*, «Psychological Bulletin», vol. 70, pp. 213-220.

Cone J.D. (1997), *Issues in functional analysis in behavioural assessment*, «Behaviour Research and Therapy», vol. 35, pp. 259-275.

Durand V.M. e Crimmins D.B. (1988), *Identifying the variables maintaining self-injurious behavior*, «Journal of Autism and Developmental Disorders», vol. 18, pp. 99-117.

Emerson E., Thompson S., Reeves D., Henderson D. e Robertson J. (1995), *Descriptive analysis of multiple response topographies of challenging behaviour across two settings*, «Research in Developmental Disabilities», vol. 16, pp. 301-329.

Hagopian L.P., Fisher W.W., Thompson R.H., Owen-DeSchryver J., Iwata B.A. e Wacker D.P. (1997), *Toward the development of structured criteria for interpretation of functional analysis data*, «Journal of Applied Behavior Analysis», vol. 30, pp. 313-326.

Hall S. e Oliver C. (1992), *Differential effects of severe self-injurious behaviour on the behaviour of others*, «Behavioural Psychotherapy», vol. 20, pp. 355-365.

Iwata B.A. (1994), *Functional analysis methodology: Some closing comments*, «Journal of Applied Behavior Analysis», vol. 27, pp. 413-418.

Iwata B.A., Dorsey M.F., Slifer K.J., Bauman K.E. e Richman G.S. (1982), *Toward a functional analysis of self-injury*, «Analysis and Intervention in Developmental Disabilities», vol. 2, pp. 3-20.

Iwata B.A., Pace G.M., Dorsey M.F., Zarcone J.R., Vollmer T.R., Smith R.G., Rodgers T.A., Lerman D.C., Shore B.A., Mazaleski J.L., Goh H., Cowdery G.E., Kalsher M.J., McCosh K.C. e Willis K.D. (1994), *The functions of self-injurious behavior: An experimental-epidemiological analysis*, «Journal of Applied Behavior Analysis», vol. 27, pp. 215-240.

Iwata B.A., Pace G.M., Kalsher M.J., Cowdery G.E. e Cataldo M.F. (1990), *Experimental analysis and extinction of self-injurious escape behavior*, «Journal of Applied Behavior Analysis», vol. 23, pp. 11-27.

Lerman D.C. e Iwata B.A. (1993), *Descriptive and experimental analyses of variables maintaining self-injurious behavior*, «Journal of Applied Behavior Analysis», vol. 26, pp. 293-319.

Mace F.C. (1994), *The significance and future of functional analysis methodologies*, «Journal of Applied Behavior Analysis», vol. 27, pp. 385-392.

Mace F.C., Lalli J.S. e Pinter-Lalli E. (1991), *Functional analysis and treatment of aberrant behavior*, «Research in Developmental Disabilities», vol. 12, pp. 155-180.

Martin N.T., Gaffan E.A. e Williams T. (1998), *Behavioural effects of long-term multi-sensory stimulation*, «British Journal of Clinical Psychology, vol. 37, pp. 69-82.

Martin N.T., Oliver C. e Hall S. (1996), *ObsWin: Observational data collection and analysis for Windows™ - Version 2*, Birmingham, UK, University of Birmingham.

Northup J., Wacker D., Sasso G., Steege M., Cigrand K., Cook J. e DeRaad A. (1991), *A brief functional analysis of aggressive and alternative bahavior in an outclinic setting*, «Journal of Applied Behavior Analysis», vol. 24, pp. 509-522.

Repp A.C., Singh N.N., Olinger E. e Olsen D.R. (1990), *The use of functional analyses to test causes of self-injurious behaviour: Rationale, current status and future directions*, «Journal of Mental Deficiency Research», vol. 34, pp. 95-105.

Skinner B.F. (1948), *«Superstition» in the pigeon*, «Journal of Experimental Psychology», vol. 38, pp. 168-172.

Smith R.G. e Iwata B.A. (1997), *Antecedent influences on behaviour disorders*, «Journal of Applied Behavior Analysis», vol. 30, pp. 343-375.

Toogood S. e Timlin K. (1996), *The functional assessment of challenging behaviour: A comparison of informant-based, experimental and descriptive methods*, «Journal of Applied Research in Intellectual Disabilities», vol. 9, pp. 206-222.

Vollmer T.R., Marcus B.A. e LeBlanc L. (1994), *Treatment of self-injury and hand mouthing following inconclusive functional analyses*, «Journal of Applied Behavior Analysis», vol. 27, pp. 331-344.

Vollmer T.R., Marcus B.A., Ringdahl J.E. e Roane H.S. (1995), *Progressing from brief assessments to extended experimental analyses in the evaluation of aberrant behavior*, «Journal of Applied Behavior Analysis», vol. 28, pp. 561-576.