L'attendibilità delle misure osservative in psicologia clinica dello sviluppo

Augusto Gnisci (Dipartimento di Psicologia, Seconda Università di Napoli)
Angiola Di Conza (Dipartimento di Psicologia, Seconda Università di Napoli)

L'osservazione sistematica del comportamento è una tecnica sempre più diffusa in psicologia clinica dello sviluppo. Tuttavia, le problematiche relative ad una codifica attendibile non sempre sono descritte in modo chiaro e sistematico. Questo contributo propone una rassegna che, senza trascurare le corrette indicazioni psicometriche, sistematizza le conoscenze tradizionali e su di esse innesta un quadro aggiornato e critico dello stato dell'arte sull'attendibilità dell'osservatore. Oltre alla trattazione dei concetti di riferimento (accordo, calibrazione, attendibilità, precisione, stabilità, accuratezza, ecc.), essa prende in considerazione i principali coefficienti proposti in letteratura – primo tra tutti il *K* di Cohen – mostrandone vantaggi e limiti.

1. Introduzione

Sebbene i metodi osservativi siano più diffusi nell'ambito della ricerca che della clinica, essi costituiscono il metodo elettivo nella psicologia dello sviluppo e si stanno sempre più diffondendo nella psicologia clinica, in particolare quando è finalizzata all'osservazione dello sviluppo del bambino, tipico e atipico. In concorso con strumenti più tradizionali, come l'intervista e i test, i metodi osservativi permettono un accesso più diretto al comportamento del bambino e ai suoi cambiamenti nel tempo, nonché la sua osservazione in setting terapeutici. La raccolta sistematica di dati basati sull'osservazione in contesti clinici permette una diagnosi più efficace e una valutazione più completa del trattamento e dei suoi esiti. Qualora si decida di effettuare un'osservazione sistematica del comportamento del bambino, associando sistematicamente alcuni comportamenti a determinate categorie sulla base di regole esplicite (Bakeman e Gottman, 1997), è necessario utilizzare delle griglie di codifica e (almeno) una coppia di osservatori che devono essere addestrati a tal fine. In questa fase si pone il problema dell'attendibilità dei dati che vengono raccolti tramite l'osservazione, in quanto l'osservatore diviene la principale fonte d'errore. Questo contributo intende focalizzarsi proprio sui problemi dell'attendibi-

lità dell'osservazione sistematica, proponendo una chiarificazione lessicale in un ambito talvolta oscuro da questo punto di vista, distinguendo i tipi fondamentali di attendibilità dell'osservatore e trattando i più importanti coefficienti di accordo. Come si vedrà, la trattazione si basa sulle nozioni classiche dell'attendibilità dell'osservatore ma intende fornire una valutazione aggiornata e critica della questione, sulla base dei contributi della letteratura più recenti e più innovativi.

2. Le basi dell'osservazione

L'osservazione diretta del comportamento è stata inizialmente considerata, per definizione, oggettiva, scevra da errore e intrinsecamente valida (cfr. Bryington, Palmer e Watkins, 2004; Watkins e Pacheco, 2000); tuttavia, questa convinzione non è durata a lungo: per quanto gli osservatori possano credere in buona fede nella correttezza e nell'oggettività della codifica che hanno effettuato, ci sono prove ripetute del fatto che, nel momento in cui si confronta la codifica prodotta da diversi osservatori, emergono molteplici differenze (Bakeman e Gottman, 1997). Per questo motivo, nella ricerca osservativa, coloro che eseguono la codifica – gli osservatori o codificatori – sono considerati la fonte principale di errore.

Quando ci si riferisce alla qualità dei dati derivanti dall'applicazione di un sistema di codifica, i ricercatori o gli studenti rimangono spesso confusi dall'uso di una pletora di termini diversi che si riferiscono a concetti a volte sfuggenti, quali accordo, affidabilità, attendibilità, calibrazione, precisione, accuratezza, validità, fedeltà, errore casuale ed errore sistematico, coerenza, stabilità, e dall'esistenza di un buon numero di indici statistici proposti per valutare l'attendibilità di un sistema di codifica (percentuale di accordo, proporzione di accordo, *K* di Cohen, coefficiente intraclasse, ecc.). A questo proposito, già nel 1987, Cone sottolineava la necessità di una chiarificazione del lessico nell'ambito della misurazione del comportamento. L'anno successivo Suen (1988) pubblicò un lavoro fondamentale che operò una serie di distinzioni alla base della discussione sull'attendibilità delle misure osservative. Ancora oggi molto spesso ricercatori e clinici dello sviluppo, quando usano i metodi osservativi, si pongono le stesse domande e hanno gli stessi dubbi: come fare una corretta osservazione? Quale concetto devo usare per riferirmi all'attendibilità nella mia ricerca? Qual è l'indice migliore per valutare l'attendibilità dei miei dati? Come calcolare l'attendibilità tra osservatori? Quali procedure devo eseguire per arrivare a delle conclusioni corrette? Questo contributo intende rispondere a molti di questi dubbi, cercando di fare ordine in que-

sta materia alquanto complessa e cercando di assumere una prospettiva legata all'approccio psicometrico tradizionale (per un approccio alternativo, quello rappresentazionale, v. John e Benet-Martínez, 2000).

3. Attendibilità e validità delle misure osservative e loro concetti di base: chiarificazione lessicale e concettuale

La codifica dei dati rappresenta l'esito di un processo messo in atto da un osservatore, un giudice, un codificatore, ecc.; per questo, anche se si possono individuare molte cause di errore, come differenze situazionali, modificazioni temporali, errori di campionamento, la fonte principale di errore è costituita dagli osservatori, ovvero dagli esseri umani a cui viene assegnato il compito di codificare il comportamento del bambino.

Secondo il più generale approccio psicometrico tradizionale, durante la codifica si possono verificare errori di due tipi, casuali o sistematici. Gli *errori casuali* sono dovuti ai problemi pratici nei quali si può incorrere durante la realizzazione della ricerca osservativa e, in particolare, nella fase di codifica (fatica, livello di attenzione, prontezza, fretta, ecc.). Ad esempio, può accadere che un osservatore sia molto stanco e, per questo motivo, tenda a rilevare solo le categorie comportamentali che si verificano con maggiore frequenza; oppure, al contrario, può succedere che l'osservatore sia particolarmente vigile e incline a notare soprattutto gli eventi più rari, corrispondenti ai codici meno usati. Si può immaginare che, se la codifica venisse eseguita infinite volte, gli errori casuali tenderebbero a compensarsi reciprocamente e perciò la loro sommatoria sarebbe nulla.

Il secondo tipo di errore, l'*errore sistematico*, si verifica quando, per qualsiasi ragione, un osservatore sistematicamente attribuisce un determinato evento osservato ad una categoria diversa da quella in cui esso rientra. Supponiamo che a tutti gli osservatori venga fornito un manuale relativo ad un sistema di codifica in cui la definizione di un determinato comportamento è sbagliata: tutti gli osservatori, adeguandosi a quella definizione erronea, forniranno una codifica sbagliata nella stessa direzione, ovvero viziata dallo stesso errore, per esempio, ipo-rappresentando il comportamento A e iper-rappresentando B. Nella fase successiva, quando si andranno ad analizzare i dati, potrebbe risultare che B è molto più frequente di A non perché questo è realmente accaduto, ma proprio a causa dell'errore sistematico.

Se applichiamo i concetti psicometrici all'osservazione, la *validità* di un sistema di codifica rappresenta il grado in cui esso misura realmente

ciò che si propone di misurare, mentre la sua *attendibilità* (fedeltà o affidabilità ne sono sinonimi) corrisponde al grado di accordo fra codifiche indipendenti dello stesso comportamento. Il rapporto tra attendibilità e validità può essere espresso come segue: l'attendibilità si riferisce alla coerenza interna al sistema di codifica (cioè si riferisce alla coerenza tra comportamenti rilevati dalle varie categorie del sistema di codifica utilizzato dai due osservatori), mentre la validità si riferisce alla capacità del sistema di codifica di riflettere realmente il processo che si realizza «là fuori», nel mondo «reale». Per chiarificare questo rapporto è possibile ricorrere a una metafora: se due testimoni forniscono la stessa versione dei fatti, essi concordano (alta attendibilità), ma questo non implica necessariamente che essi dicano la verità rispetto a quanto è realmente accaduto (bassa validità). Tuttavia, se essi non sono d'accordo (attendibilità bassa), il problema della validità non si pone perché non esiste una versione coerente da confrontare con quanto accaduto. Per questo l'attendibilità viene considerata una condizione necessaria per la validità.

Per comprendere meglio il concetto di attendibilità è utile riferirsi ad alcuni concetti di base, quali precisione, stabilità e accuratezza. La *precisione* riguarda il grado di coerenza con il quale l'osservatore associa eventi o oggetti a determinate categorie (sebbene essa non sia un concetto consensuale; v. Suen, 1988, nota 1). Se, in una sessione di codifica, un codificatore assegna sistematicamente un comportamento, ad esempio un sorriso del bambino, ad una specifica categoria A, definita «Comportamenti positivi», la risultante attività di codifica può dirsi precisa. La precisione, però, non implica l'accuratezza: infatti, un osservatore può essere preciso ma non accurato. Se, ad esempio, egli assegna sistematicamente il sorriso, anziché ad A, alla categoria B, corrispondente a «Comportamenti negativi», la risultante attività di codifica sarà precisa (in quanto coerente), ma inaccurata, in quanto caratterizzata dalla presenza di un errore sistematico. Infatti, l'*accuratezza* è il grado di corrispondenza tra le categorie usate e la realtà, cioè essa fa riferimento a quanto ciò che viene codificato dall'osservatore riflette i comportamenti o i processi realmente verificatisi durante l'interazione osservata. Il termine accuratezza, quindi, identifica un concetto che è alla base della validità, ma che ha a che fare anche con l'attendibilità. Un esempio favorirà la comprensione del senso attribuito a questo termine: se, applicando correttamente un sistema di codifica attendibile che rileva le dinamiche relazionali familiari, emerge un elevato livello di aggressività della madre nei confronti del bambino, a fronte di una situazione in cui, di fatto, elementi di aggressività sono assenti o poco frequenti, bisognerà concludere che il sistema di codifica utilizzato è semplicemente sbagliato, cioè *inaccurato*.

Infine, la *stabilità* nel tempo (o *attendibilità retest*) si riferisce al grado di correlazione tra codifiche dello stesso comportamento, condotte in momenti diversi. Viene applicata nel corso di sessioni di codifica realizzate in momenti successivi (per esempio a distanza di alcuni mesi) e garantisce che la perfomance di codifica dell'osservatore non degeneri, ovvero non si modifichi nel tempo in maniera sostanziale (Bakeman e Gottman, 1997).

Riassumendo, la precisione si riferisce alla coerenza della codifica all'interno della stessa sessione, la stabilità alla coerenza tra le sessioni. Precisione e stabilità riguardano esclusivamente l'attendibilità in quanto permettono, sebbene in modo diverso, di valutare la coerenza interna dell'attività di codifica, tra due osservatori o nel corso di differenti rilevazioni dello stesso fenomeno. Si è affermato che l'attendibilità è un concetto multidimensionale (Suen, 1988) perché esso prende forme diverse a seconda delle diverse applicazioni di ricerca, basandosi sui concetti di precisione, stabilità e anche accuratezza. A seconda di quale concetto essa implementa, si parla, rispettivamente, di attendibilità inter-osservatore, intra-osservatore o dell'osservatore (v. sotto).

Accanto ai concetti descritti, spesso, parlando dell'attendibilità dei sistemi di codifica, si ricorre all'uso del termine *calibrazione* tra gli osservatori. Essa si riferisce alla procedura con la quale due osservatori vengono portati a utilizzare in modo simile il sistema di codifica in modo che, al termine di questa procedura, venga raggiunto un alto livello di accordo tra loro. Sebbene essa vari in funzione della complessità dei sistemi di codifica implicati e dell'investimento di energie che un gruppo di ricerca intende dedicare ad una ricerca, essa si raggiunge col training degli osservatori sulla base del manuale di codifica, con ripetute sedute di prova e supervisione per assicurarsi che essi condividano le definizioni dei comportamenti rilevanti. Essa può implicare, ad un livello, la calibrazione tra i due osservatori, ad un livello migliore, la calibrazione di ciascun osservatore con un protocollo standard (v. sotto). Questa procedura consente di ridurre le differenze tra le codifiche dei dati osservativi eseguite da più osservatori, in quanto il materiale osservato e i dati rilevati tramite codifica non dovrebbero variare in funzione dell'osservatore (Bakeman e Gottman, 1997). Il vantaggio ottenuto dall'applicazione di procedure di calibrazione è *essenzialmente pratico*: una volta calibrati, due osservatori forniscono una codifica simile, se non identica, e perciò possono essere utilizzati in modo interscambiabile. Tuttavia, sebbene in letteratura ci sia qualche ambiguità, la calibrazione è solo una procedura, che se ben condotta, può portare ad una buona attendibilità, ma non ha un valore concettuale in sé: cioè, non è un concetto psicometrico che sta alla base dell'attendibilità, così come le procedure per costruire item riusciti

di una scala influiscono positivamente sull'attendibilità ma non ne sono il concetto psicometrico alla base.

Quando due codificatori sono stati calibrati tra loro, la codifica ottenuta è precisa, e perciò attendibile, ma non necessariamente valida. Una situazione di questo tipo (in cui l'attendibilità raggiunta è buona o ottima, mentre la validità è bassa) può verificarsi quando due codificatori vanno d'accordo nella loro attività di codifica solo perché condividono un modo di codificare errato rispetto a quello presentato nel manuale di codifica: ci sono molti motivi (ad esempio, un errore nel processo di addestramento degli osservatori all'uso del sistema di codifica) per cui si può determinare una condivisione da parte dei due osservatori di una visione del mondo distorta, che porta entrambi alla convinzione che un certo tipo di comportamento osservato rientri nella categoria sbagliata. Ne consegue che il loro accordo e, quindi, l'attendibilità del sistema, potrà anche essere alto, mentre la validità sarà bassa, in quanto la codifica non rispecchia il mondo reale.

Una volta precisati i concetti di base dell'attendibilità nella ricerca osservativa, abbiamo gli strumenti per fare alcune precisazioni. Per prima cosa, negli studi osservativi non è dato distinguere tra l'inattendibilità del sistema di codifica e del manuale d'uso (errata costruzione delle categorie, definizioni dei comportamenti ambigue, contraddittorie, non esclusive, ecc.) e inattendibilità dovuta all'osservatore (errore nella codifica dovuto al solo osservatore). In letteratura non esiste alcun riferimento a questo problema. In altre parole quella che noi chiamiamo mancanza di attendibilità in ambito osservativo e che attribuiamo all'osservatore è in realtà attribuibile ad una serie di errori che confonde le due categorie sopra indicate.

In secondo luogo, va ribadito che questo contributo intende basarsi su una letteratura psicometrica più ampia rispetto a quella solitamente riservata agli studi sull'attendibilità dell'osservazione, cioè vuol riferirsi alla tradizione psicometrica classica. Detto ciò, va sottolineato che rispetto ad altre forme di misurazione, in particolare quelle che si basano su dati «riportati», esistono delle differenze fondamentali. Il dato riportato, infatti, è un dato psicologico in se stesso nella misura in cui esso riporta in modo corretto la soggettività di un individuo, per esempio, la sua opinione, l'atteggiamento o il giudizio. L'attendibilità nell'ambito delle misure osservative, invece, coinvolge un codificatore che osserva la realtà e si riferisce perciò ad un dato psicologico «oggettivo» (per esempio, un bambino mostra un chiaro sintomo comportamentale). La conclusione che se ne trae è che, per le misure «riportate», gli indici basati sulla consistenza interna e, a seconda dei casi, sulla stabilità sono idonei a valutare l'attendibilità mentre l'accuratezza è alla base esclusi-

vamente della validità; per le misure osservative, invece, il concetto di accuratezza sta alla base dell'attendibilità, insieme a consistenza e stabilità, e anche della validità. Vedremo, infatti, che la presenza di un'attendibilità dell'osservatore, basata su una codifica considerata vera, si basa proprio sull'accuratezza.

4. Attendibilità dei sistemi di codifica

Accordo e attendibilità sono concetti diversi, sebbene collegati tra loro. Ma che differenza c'è? In realtà, nella letteratura di riferimento non si trova una risposta conclusiva. Bakeman e Gottman (1997) propongono la seguente distinzione: l'accordo si riferisce, come implica la parola, al grado in cui due osservatori concordano tra loro. Questo accordo non necessariamente è un accordo sostanziale o «vero», né contempla o previene le molteplici fonti d'errore che possono alterare la ricerca. L'attendibilità, invece, facendo riferimento a una ricca e articolata tradizione psicometrica, intende idealmente far fronte a tutte le possibili fonti d'errore presenti e non si limita al caso in cui la codifica sia eseguita da due o più osservatori (come vedremo tra poco per l'attendibilità intra-osservatore e dell'osservatore). Attendibilità è, quindi, un termine riferito ad un concetto più generale, fondato su una più solida tradizione psicometrica, che può essere definito come rapporto tra varianza vera e varianza osservata (ovvero, la varianza totale, composta da varianza vera e varianza d'errore). Qualsiasi sia la distinzione evocata dai due termini, comunque, molti studiosi ritengono che la presenza di accordo debba essere considerata una *conditio sine qua non* per poter ottenere una buona attendibilità (Bakeman e Gottman, 1997): in mancanza di un accordo empirico tra gli osservatori, l'attendibilità non può essere stabilita; tuttavia, la presenza di accordo non necessariamente si traduce in attendibilità.

In generale, l'attendibilità è definita come il grado in cui i dati sono esenti da errore di misura: minore è l'errore, maggiore è la coerenza dei dati (Suen, 1988). Facendo riferimento all'osservatore come fonte di errore, si possono distinguere almeno tre tipi di attendibilità (Berk, 1979; Martin e Bateson, 1986): un osservatore può non essere attendibile rispetto a un altro osservatore (*attendibilità inter-osservatore*), rispetto a se stesso (*attendibilità intra-osservatore*), o rispetto a un osservatore ideale, che si assume abbia codificato perfettamente (*attendibilità dell'osservatore*).

I prossimi paragrafi sono dedicati a definire ciascun tipo di attendibilità.

4.1. Attendibilità inter-osservatore

L'*attendibilità inter-osservatore* si basa sul concetto di *precisione* e corrisponde al grado in cui due osservatori opportunamente addestrati e indipendenti, producono risultati di codifica simili quando osservano la stessa interazione (Watkins e Pacheco, 2000), per esempio, un'interazione tra padre e figlio videoregistrata. Essa può essere interpretata come il grado in cui i due osservatori possono essere considerati intercambiabili e indica quanto i dati siano liberi da errore casuale e/o sistematico legato alla codifica eseguita dagli osservatori.

4.2. Attendibilità intra-osservatore

L'*attendibilità intra-osservatore,* il cui concetto di base è la *stabilità,* corrisponde al grado in cui un osservatore, che osserva la stessa interazione in condizioni identiche in momenti diversi, produce gli stessi risultati di codifica, realizzando così un buon livello di consistenza interna all'osservatore. A questo proposito esistono due posizioni diverse. La prima posizione sostiene che, siccome questo tipo di attendibilità implica che il medesimo osservatore codifichi ripetutamente lo stesso materiale, la valutazione dell'attendibilità intra-osservatore può essere viziata da problemi legati a stanchezza, noia, facilitazione, ecc. (Suen, 1988, nota 6); per calcolare questa attendibilità, dunque, si ricorre a due osservatori diversi che, però, vengono considerati come forme parallele di un singolo osservatore. Questa posizione, in teoria corretta, ha trovato molte critiche: alcuni sostengono che di fatto si sta misurando l'attendibilità inter-osservatori, altri che, operando in tal modo, non è data la possibilità di distinguere la parte di errore dovuto al decadimento nel tempo nella performance dell'osservatore da quella dovuta alla performance dei due diversi osservatori. Per questo, l'utilizzo dello stesso osservatore che codifica lo stesso identico materiale più volte nel tempo sembra più ragionevole; tuttavia, lo studio della stabilità deve prevedere un disegno di ricerca che minimizzi le fonti di errore dovute all'effetto delle prove ripetute.

Come indicato nella definizione a inizio paragrafo, la condizione per ottenere una stima corretta dell'attendibilità intra-osservatore è che l'osservatore valuti nel tempo lo stesso materiale, condizione resa possibile dalle moderne procedure di audio-registrazione. Tuttavia, in psicologia clinica dello sviluppo si può avere a che fare con uno stesso bambino che in sedute diverse si comporta diversamente, ciò che si valuta non necessariamente è un comportamento stabile perché una persona in età evolutiva o in un contesto clinico è soggetto per natura a processi di cambia-

mento o perché l'intervallo tra le due valutazioni è molto ampio. Inoltre, alcuni progetti di ricerca possono programmare, per scelta o per motivi legati ai vincoli del contesto in cui si effettua l'osservazione (per esempio, un tribunale minorile), di osservare in diretta il comportamento. Perciò, quando non si ha a disposizione una registrazione, la componente d'errore, che comprende l'errore dovuto all'osservatore, è inflazionata da una componente dovuta al possibile cambiamento del comportamento del bambino nel tempo. Se si utilizzano sessioni diverse con lo stesso bambino ci possiamo trovare di fronte a due situazioni: se si ottiene un'alta attendibilità intra-osservatore, si può concludere, con un po' di pragmatismo, che il problema non si è posto; tuttavia, se si ottiene una bassa attendibilità, la situazione rimane ambigua in quanto essa non può essere necessariamente attribuita al decadimento nel tempo della performance dell'osservatore. In questi casi non c'è una soluzione semplice. Una possibilità è ottenere delle registrazioni del bambino in questione, nel contesto in cui sarà effettuata la codifica in diretta; un'altra possibilità è campionare e registrare comportamenti simili di diversi bambini nello stesso contesto. Mentre l'attendibilità inter-osservatore può essere stabilita facendo osservare in diretta il comportamento del bambino nella stessa sessione, l'attendibilità intra-osservatore può essere stabilita facendo codificare all'osservatore il materiale registrato in due momenti diversi. Per esempio, alcuni gruppi di ricerca, come quello di Patterson (1982) che utilizzavano osservatori per una codifica *live*, hanno ottenuto delle registrazioni di comportamenti simili a quelli che gli osservatori avrebbero codificato dal vivo e preparato una codifica «corretta», sulla base della quale la codifica degli osservatori veniva valutata anche nel tempo.

4.3. Attendibilità dell'osservatore

Assumiamo che un ricercatore predisponga un flusso di codifica, definito protocollo standard, che rappresenti il prodotto della codifica eseguita da un osservatore ideale e infallibile (Bakeman e Gottman, 1997) o da un *master*, cioè un codificatore esperto o professionista (Suen, 1988). A volte, questa versione della codifica preparata da esperti e che si presume accurata viene detta «gold standard» (Bakeman e Quera, 2011, p. 61). Assumiamo che questo flusso di codifica sia perciò considerabile «vero» e che venga confrontato con il prodotto della codifica di uno o più osservatori normali. Secondo Bakeman e Gottman (1997), tramite questa procedura il ricercatore può: *a*) controllare che il codificatore esegua correttamente la codifica; *b*) calibrare i codificatori; *c*) ottenere una codifica che riflette il contenuto di ciò che è suo interesse

codificare. Ne consegue che l'*attendibilità dell'osservatore*, basata sul concetto di *accuratezza*, corrisponde al grado in cui l'osservatore concorda con quanto stabilito da un protocollo standard assunto come vero. Questa procedura permette di eliminare qualsiasi tipo di errore, purché il protocollo standard sia formulato correttamente. Proprio perché la codifica dell'osservatore viene confrontata con quella di un protocollo standard, alcuni assimilano questo tipo di attendibilità alla validità di criterio concorrente (Suen, 1988); tuttavia, dato che il protocollo è una misurazione, assunta come vera, dello stesso costrutto misurato da un osservatore, essa potrebbe essere equiparata più appropriatamente alla validità di costrutto[1].

Nel proseguo, a parte quando altrimenti specificato, assumeremo che l'applicazione di riferimento è quella tra due osservatori (inter-osservatori) perché essa è di fatto la più utilizzata. Tuttavia, se al posto del secondo osservatore si considera una valutazione dislocata nel tempo da parte del primo osservatore o un protocollo standard, le considerazioni che faremo a proposito dell'attendibilità tra osservatori possono essere correttamente applicate anche all'attendibilità intra-osservatore o dell'osservatore.

5. I coefficienti per l'attendibilità e il loro calcolo

Solitamente si possono utilizzare due tipi di indici per misurare l'attendibilità tra osservatori (cfr. Pedon e Gnisci, 2012). I coefficienti globali calcolano l'andamento congiunto generale tra due osservatori, che riguarda quante volte essi individuano i comportamenti A, B, e così via, indipendentemente dal fatto che quel comportamento sia stato identificato nello stesso istante. I coefficienti punto-per-punto, invece, calcolano l'accordo rispetto a ciascun accadimento.

5.1. I coefficienti globali

In genere, nello studio dell'attendibilità hanno molta importanza i coefficienti di correlazione. Esistono due tipi di coefficienti di correlazione, semplici e intraclasse.

[1] Il concetto di validità di criterio concorrente si riferisce al grado di associazione o corrispondenza tra la misurazione di un costrutto e le misurazioni di altri costrutti diversi e utilizzati come criteri di riferimento esterno, rilevati contemporaneamente; la validità di costrutto si riferisce al grado in cui una misurazione riflette accuratamente ciò che dice di misurare (Pedon e Gnisci, 2004).

I primi sono coefficienti in grado di rilevare le variazioni congiunte di due variabili, cioè se al crescere di una variabile *x* cresce (o decresce) una variabile *y*. Variano da +1 a –1. Quando il loro valore varia da 0 a +1, la correlazione tra le due variabili è positiva (cioè se i valori dell'una salgono, anche quelli dell'altra aumentano); se varia da 0 a –1, la correlazione è negativa (mentre i valori di una variabile salgono, quelli dell'altra diminuiscono); coefficienti vicino allo 0 riflettono l'assenza di correlazione tra le due variabili, ovvero la loro indipendenza; più il loro valore si avvicina a |1|, più le due variabili sono associate. Essi possono essere testati per la significatività, anche se nella valutazione dell'attendibilità è molto importante la valutazione dell'intensità della correlazione.

L'esempio più noto è il coefficiente di correlazione di Pearson, detto «prodotto-momento» e indicato con il simbolo *r*, che si usa quando, in una matrice casi per variabili, diverse sessioni interattive vengono poste in riga (come se fossero dei soggetti) e i due osservatori vengono collocati in colonna, riportando le frequenze rilevate di un determinato comportamento, per esempio A (come se fossero due variabili a rapporti). Se le due variabili correlano ampiamente, significa che i due osservatori vanno molto d'accordo. Tuttavia, un alto coefficiente di correlazione indica che i due osservatori sono d'accordo su quali sessioni sono meno e più frequenti, indipendentemente dal numero preciso di accadimenti che rilevano; cioè, sono d'accordo sull'ordinamento delle sessioni in base al comportamento di frequenza. La *r* di Pearson, perciò, può essere considerata un indice relativo perché non appaia (*match*) le misure, semplicemente le correla. Elevando la correlazione al quadrato otteniamo il coefficiente di determinazione, che indica la percentuale di varianza comune alle valutazioni dei due osservatori.

La seconda categoria di coefficienti globali è costituita dai *coefficienti di correlazione intraclasse* (*CCI*), che, differentemente dai primi, variano tra 0 e 1. Valori che si avvicinano a zero indicano che le variabili hanno poca varianza in comune e molta varianza d'errore, valori che si avvicinano a 1 indicano molta varianza comune e poca varianza d'errore. Proprio perché hanno queste caratteristiche, i coefficienti di correlazione intraclasse sono spesso preferiti ai primi: essi, infatti, sono idonei a esprimere il rapporto tra varianza vera e varianza osservata (proporzione di varianza), che, come abbiamo visto, è il concetto di base dei coefficienti di attendibilità.

Esistono tanti coefficienti di correlazione intraclasse diversi (Shrout e Fleiss, 1979; McGraw e Wong, 1996). Tuttavia, essi possono essere ricondotti a due categorie. I *CCI relativi* e i *CCI assoluti* esprimono, rispettivamente, il grado di consistenza relativo *versus* assoluto tra i punteggi (Bakeman e Quera, 2011). In altre parole, i primi esprimono quanto

due variabili variano in modo congiunto (al variare dell'una varia l'altra, come i coefficienti di correlazione sopra descritti), i secondi ci informano su quanto i due punteggi sono uguali. Per esempio, un alto coefficiente assoluto indica che il punteggio di un osservatore è simile al punteggio di un altro osservatore, un alto coefficiente relativo significa che sessioni che hanno una frequenza alta, media o bassa per un osservatore, ne hanno una alta, media e bassa anche per il secondo osservatore. In questo caso le sessioni sono ordinate nello stesso modo dai due osservatori, ma esse non hanno necessariamente la stessa frequenza. Per questo, un alto coefficiente assoluto implica un alto coefficiente relativo, ma non il contrario. Se utilizzati correttamente, i *CCI assoluti* permettono di controllare sia l'errore casuale (come i coefficienti di correlazione semplici e relativi), sia l'errore sistematico.

I *CCI* hanno l'ulteriore vantaggio di essere gli indici di attendibilità scelti dalla *teoria della generalizzabilità* (Cronbach *et al.*, 1972), secondo cui gli strumenti di misura devono fare il lavoro per cui sono stati costruiti, cioè discriminare tra aspetti rilevanti per il costrutto (per esempio, tra i comportamenti degli individui osservati) e non discriminare tra aspetti irrilevanti (per esempio, due forme parallele del test o i due osservatori).

Basandosi su un approccio legato all'analisi della varianza, i *CCI* fanno proprio questo: valutano la grandezza della variabilità della fonte individuata rispetto alla varianza d'errore. Se la varianza dovuta agli aspetti irrilevanti, per esempio i due osservatori, è bassa rispetto alle altre fonti di varianza, allora è dimostrato che essi sono realmente sostituibili. I risultati ottenuti con una delle due forme sono generalizzabili. Per questo i *CCI* sono detti anche *coefficienti di generalizzabilità*.

5.2. Cenni di teoria della generalizzabilità applicata all'osservazione

La *teoria della generalizzabilità* rappresenta un approccio integrato che consente di isolare e distinguere tre diverse fonti di errore, per mezzo di procedure di analisi della varianza (Bakeman e Quera, 2011; Cronbach *et al.*, 1972; Gnisci e Bakeman, 2000; Gnisci, Maricchiolo e Bonaiuto, 2013). Queste fonti sono: la varianza dovuta ai soggetti che vengono osservati (varianza vera); la varianza dovuta all'osservatore (errore sistematico o bias); la varianza legata all'errore casuale (errore casuale intra-osservatore) (Suen, 1988). Da queste fonti di varianza si possono derivare diversi *coefficienti di correlazione intra-classe*, che consentono di valutare se il sistema di codifica adottato svolge effettivamente «il lavoro per il quale è stato pensato», ovvero se è in grado di discriminare tra aspetti rilevanti dello studio (per esempio, tra i soggetti osservati) e di

non discriminare tra caratteristiche irrilevanti dello studio (come i codificatori). Questo approccio mira ad estendere il concetto di attendibilità così da assottigliare i confini tra attendibilità e validità (le specifiche procedure di implementazione e calcolo sono descritte in: Berk, 1976; Brennan, 1983; Hartmann, 1982; Suen, 1988; in Bakeman e Gottman, 1997, e Bakeman e Quera, 2011, è possibile trovare anche le procedure idonee applicabili a dati sequenziali).

5.3. I coefficienti punto-per-punto

Essi sono indicatori del grado in cui gli osservatori codificano nello stesso modo ogni unità di base dell'osservazione (eventi, intervalli o tempi). Per comprendere gli indici che permettono di misurare adeguatamente l'attendibilità punto-per-punto è necessario comprendere cos'è e come è formata una matrice di confusione, spiegata nel prossimo paragrafo, riprendendo un esempio derivato dalla letteratura.

5.4. La matrice di confusione o matrice del *K*

L'esempio riportato di seguito riprende, semplificandoli, i sistemi di codifica degli episodi di negoziazione del gioco tra bambini di età compresa tra 6 e 10 anni, proposto da Bearison *et al.* (2001). Oggetto dell'osservazione sono le attività svolte dai bambini nel corso dell'episodio che conduce a strutturare le regole di un gioco da fare insieme. I sistemi di codifica adottati riguardano l'argomento della negoziazione (che prevede le categorie: regole di gioco e scopo del gioco); l'esito (non risolto, consenso passivo, accettazione attiva, compromesso comune); la funzione dei turni (proposta iniziale, controproposta, disaccordo o accordo); la giustificazione (nessuna, fattuale, assunzione di prospettiva).

A scopo esemplificativo per il calcolo della percentuale di accordo tra gli osservatori e del *K* di Cohen, consideriamo le categorie relative all'esito dell'episodio come oggetto di interesse per la codifica. Due osservatori O_1 e O_2 eseguono la codifica, applicando il sistema descritto precedentemente, ovvero composto dalle quattro categorie corrispondenti a «esito non risolto» (i bambini non trovano un accordo sulle regole e lo scopo del gioco), «consenso passivo» (uno dei due bambini accetta le regole e lo scopo proposti dall'altro, senza avanzare proposte a sua volta), «accettazione attiva» (un bambino propone regole e scopo, l'altro accetta, intervenendo attivamente nella ristrutturazione della proposta del primo) e «compromesso comune» (i due bambini si impegnano in maniera

TAB. 1. *Matrice di confusione su dati simulati con quattro codici relativi ai possibili esiti di un episodio di negoziazione del gioco, codificati da due osservatori ipotetici (O_1 e O_2) per il calcolo delle percentuali di accordo e del* K *di Cohen*

		O_2				
		Non risolto	Consenso passivo	Accettazione attiva	Compromesso comune	Tot.
O_1	Non risolto	35	0	0	1	36
	Consenso passivo	5	46	1	0	52
	Accettazione attiva	0	7	33	2	42
	Compromesso comune	1	0	0	31	32
	Tot.	41	53	34	34	162

collaborativa in un percorso di co-costruzione delle regole e dello scopo del gioco).

Per calcolare l'accordo, è necessario costruire una tabella, definita *matrice di confusione* o *di accordo* o *del K*, relativa alla codifica empirica eseguita dagli osservatori, ponendo l'esito della codifica di uno dei due osservatori in riga (O_1) e l'esito della codifica dell'altro osservatore in colonna (O_2). La tabella risultante è composta da quattro righe e quattro colonne corrispondenti alle quattro categorie di esito possibile (mostrata in tab. 1).

La prima cella in alto a sinistra indica che per 35 volte entrambi gli osservatori codificano l'esito osservato come non risolto (accordo), la cella sottostante indica che in 5 occasioni O_1 osserva un esito che codifica come «consenso passivo», mentre O_2 codifica quello stesso episodio come «non risolto» (disaccordo). Le frequenze di accordo sono disposte lungo la diagonale maggiore e la loro somma costituisce l'accordo totale (145); mentre le frequenze di disaccordo sono identificate nelle celle fuori dalla diagonale maggiore e la loro somma rappresenta il disaccordo totale (17). I totali marginali di riga e di colonna corrispondono al totale delle osservazioni che ciascun osservatore (rispettivamente O_1 e O_2) codifica in ciascuna delle categorie considerate. Il totale generale (162) corrisponde, infine, al totale delle osservazioni codificate.

La matrice di confusione rappresenta il punto di partenza per il calcolo degli indici di accordo e per l'identificazione di eventuali fonti di assenza di calibrazione tra gli osservatori.

5.5. La percentuale di accordo

Di fronte ai dati riportati in una matrice di confusione, la soluzione più diffusa e più semplice per calcolare l'attendibilità inter-osservatore è

il ricorso alla *percentuale di accordo* (sebbene ormai quasi tutti gli studiosi, come vedremo, sostengano che essa non possa corrispondere ad un vero e proprio indice di attendibilità). La percentuale di accordo è il rapporto tra numero di accordi tra gli osservatori e somma di accordi e disaccordi, il tutto moltiplicato per 100 (quando è espressa su una scala 0-1 viene detta *proporzione di accordo*). La procedura per il suo calcolo e l'applicazione all'esempio descritto in tabella 1 è riportata di seguito.

$$\text{Percentuale di accordo} = \frac{\sum_{i=j=1}^{k} x_{ij}}{N} = \frac{145}{162} = .89$$

Sebbene questo indice abbia il vantaggio di essere intuitivo e facile da calcolare, ha due difetti che non possono essere eliminati. Il primo è che la percentuale di accordo risulta gonfiata, rispetto al valore di accordo reale, in quanto non viene corretta per il cosiddetto *accordo dovuto al caso*. Infatti, se noi assegnassimo a due osservatori indipendenti il compito di generare a caso una sequenza di codici appartenenti ad un sistema di codifica, le loro codifiche mostreranno comunque un certo livello di accordo, un accordo dovuto al caso. Secondo Cohen (1960), l'accordo dovuto al caso può essere calcolato tenendo conto delle frequenze di cella attese in caso di verità dell'ipotesi nulla (proprio come nel caso del chi quadro), la quale postula l'indipendenza tra gli osservatori: in effetti, assunto che i due osservatori abbiamo esercitato un'attività di codifica che rispetta i valori marginali di riga e di colonna, queste frequenze attese corrispondono esattamente a quelle che si avrebbero se i due osservatori fossero indipendenti. Una volta calcolato l'accordo dovuto al caso, bisogna sottrarlo all'accordo osservato per aver un buon indice di attendibilità.

Il secondo difetto della percentuale di accordo è che essa dipende dalla frequenza del comportamento osservato e, cioè, dalle distribuzioni marginali della matrice di confusione (Bryington, Palmer e Watkins, 2004; Towstopiat, 1984). Assumendo una matrice di confusione con due sole categorie (2 x 2), quando la probabilità marginale di occorrenza di una categoria di comportamento si avvicina a 1 o a 0 (per esempio, quando la percentuale di una categoria osservata è inferiore a .20 o superiore a .80), c'è una maggiore probabilità che la percentuale di accordo sia gonfiata impropriamente rispetto a quando la probabilità marginale di quella categoria sia equiprobabile (cioè si avvicini a .50).

Questi due problemi sono collegati e hanno importanti conseguenze. Dato che la grandezza della percentuale di accordo può essere aumentata indebitamente dall'accordo dovuto al caso, che, a sua volta, dipende

dalla distribuzione marginale dei comportamenti, non ha senso fornire una soglia della percentuale di accordo sopra la quale si può dire che l'indice è accettabile, né possono essere paragonate percentuali di accordo provenienti da studi diversi, che hanno ragionevolmente una diversa probabilità marginale (Nussbeck, 2005). In più, dato che il valore di accordo osservato, posto sia al numeratore sia al denominatore nella formula per il calcolo della percentuale di accordo, contiene in sé l'errore dovuto al caso, il numeratore non fornisce un indice di varianza vera, né il denominatore un indice di varianza totale. Di conseguenza, poiché un indice tradizionale di attendibilità si ottiene a partire dal rapporto tra varianza vera e varianza totale, la percentuale di accordo non può essere considerata ad alcun titolo un indice di attendibilità. Per questi motivi, nonostante sia probabilmente l'indice di accordo più utilizzato, molti autori sostengono che la percentuale di accordo non dovrebbe essere più utilizzata e al suo posto, invece, dovrebbe essere utilizzato il *K* di Cohen[2].

6. Il *K* di Cohen

Il *K* di Cohen (1960) è un indice per il calcolo dell'attendibilità tra gli osservatori che ha il notevole vantaggio di correggere l'indice di accordo per l'accordo dovuto al caso. Esso recepisce il suggerimento di Cohen (1960) di sottrarre all'accordo osservato l'accordo dovuto al caso. Le procedure per il calcolo, basate sulla stessa matrice di accordo, sono schematizzate di seguito.

Per procedere al calcolo del *K* di Cohen è necessario calcolare l'accordo osservato, il disaccordo osservato, l'accordo dovuto al caso, l'accordo vero e, infine, calcolare il rapporto tra la differenza tra accordo osservato e accordo dovuto al caso e questa stessa differenza più il disaccordo. Le procedure di calcolo relative alla matrice di confusione presentata in esempio (v. tab. 1) sono sviluppate di seguito:

$$\text{Accordo osservato } (A_0) = \frac{\sum_{i=j=1}^{k} x_{ij}}{N} = \frac{145}{162} = .89$$

[2] Quando la probabilità marginale di una determinata categoria è inferiore a .20 o supera .80, possono essere utilizzati i cosiddetti indici di accordo basati sull'occorrenza (o sulla non-occorrenza). Essi sono simili alla percentuale di accordo, ma sono calcolati tenendo conto esclusivamente dell'occorrenza o della non-occorrenza di una singola categoria (per approfondimenti su questi indici, v. Bryington *et al.*, 2004; Nussbeck, 2005; Watkins e Pacheco, 2000). Tuttavia, l'accordo basato sull'occorrenza/non-occorrenza, sebbene limiti i problemi legati all'accordo dovuto al caso, non li elimina del tutto (Suen e Ary, 1989).

L'accordo osservato è dato, dunque, dal rapporto tra la sommatoria di tutte le celle poste lungo la diagonale maggiore, che corrispondono alle osservazioni rispetto alle quali gli osservatori concordano (per cui $i = j$) e il totale delle osservazioni effettuate.

$$\text{Disaccordo osservato } (D) = \frac{\sum_{i \neq j=1}^{k} x_{ij}}{N} = \frac{17}{162} = .10$$

Il disaccordo osservato risulta, al contrario, dal rapporto tra la sommatoria delle frequenze collocate nelle celle fuori dalla diagonale, corrispondenti alle codifiche discordanti prodotte dai due osservatori (per cui $i \neq j$) e il totale delle osservazioni.

$$\text{Accordo dovuto al caso (AC)} = \frac{\sum_{j=1}^{k} x_{i+} x_{+i}}{N^2} = \frac{41 \times 36 + 53 \times 52 + 34 \times 42 + 34 \times 32}{162^2} = \frac{6748}{26244} = .26$$

L'accordo dovuto al caso si calcola come sommatoria del prodotto del totale delle frequenze riportati in ciascuna riga (x_{i+}) e il totale delle frequenze riportate in ciascuna colonna (x_{+i}), diviso il totale delle osservazioni elevato al quadrato.

$$\text{Accordo vero } (Av) = A_O - A_C = .89 - 26 = .63$$

L'accordo vero è dato dall'accordo osservato, depurato dell'accordo dovuto al caso (ovvero, dalla differenza tra questi due valori). Ottenuti questi indici è possibile calcolare il valore di *K*, sostituendo nella formula, riportata di seguito, gli indici calcolati di accordo osservato, accordo dovuto al caso e disaccordo.

$$K = \frac{A_O - A_C}{A_O - A_C + D} = \frac{A_V}{1 - A_C} = \frac{.63}{1 - .26} = .85$$

Poiché la differenza tra accordo osservato e accordo dovuto al caso corrisponde, come sopra riportato, all'accordo reale e la somma di accordo osservato e disaccordo è sempre uguale a 1, la formula classica per il computo dell'indice *K* può essere riscritta come: accordo vero (A_V) diviso 1 meno accordo dovuto al caso.

Dato che la relazione matematica tra percentuale di accordo e *K* è stata dimostrata, esistono delle tavole di conversione che permettono di

trasformare i valori della percentuale di accordo del proprio studio in valori di *K* (Suen, Ary e Ary, 1986).

Sebbene si sia sostenuto che il *K* vari tra 0 e 1, in realtà esso varia da –1 a +1, con valore positivo, negativo e nullo che indica che gli osservatori sono in accordo, rispettivamente, più, meno o pari a quanto atteso per il caso: se il valore è nullo, i due osservatori vanno d'accordo come due persone che eseguano una codifica a caso, quando è negativo, i due osservatori cominciano ad essere sistematicamente in disaccordo. Quando il valore del *K* è positivo, dunque, i due osservatori vanno d'accordo indipendentemente dall'accordo dovuto al caso. Questo potrebbe indurre a pensare che un indice positivo sia sufficiente per stabilire l'accordo. Invece, un accordo sostanziale è dato da un valore del *K* non solo positivo ma anche grande, il che solleva il problema di stabilire una soglia adeguata per la sua interpretazione (v. sotto).

6.1. Il *K* di Cohen: una valutazione critica aggiornata

Secondo i suoi sostenitori (Bakeman e Quera, 2011), questo indice ha molti pregi, dato che elimina l'accordo dovuto al caso dall'accordo osservato, e l'accordo dovuto al caso non gonfia indebitamente l'indice come, invece, accade per la percentuale di accordo. Per questi motivi, fino a due decenni fa, si sosteneva (per esempio, Sun e Ary, 1989) che questi suoi pregi consentissero il confronto tra studi e condizioni differenti, ma oggi tale credenza si è dimostrata erronea. I paragrafi successivi mirano a delineare in maniera oggettiva le caratteristiche positive e le limitazioni attualmente attribuite a questo indice.

6.2. I vantaggi del *K*

Il *K* ha notevoli vantaggi, che gli sono stati riconosciuti fin dalla sua nascita e che, oltre all'eliminazione della componente casuale (come vedremo, questo non viene considerato un pregio per alcuni), sono: 1) possibilità di pesare il *K* per codici ordinali (*k pesato*); 2) possibilità di applicazione all'intero sistema di categorie, al singolo codice o all'unità di codifica; 3) generalizzazione a più di due osservatori; 4) valutazione del *K* tramite la significatività; 5) valutazione del *K* tramite la scala di riferimento e il buon senso.

6.3. Il *K* pesato

Il *K* pesato è una versione modificata del *K* di Cohen (1968), che tiene conto della gravità dell'errore commesso dagli osservatori, appli-

cabile quando le categorie utilizzate sono ordinabili, ovvero quando alcune categorie che compongono un sistema di codifica sono più vicine rispetto ad altre. Si ricorre a questo indice nei casi in cui ci sia motivo di ritenere che alcune discrepanze nel processo di codifica siano più rilevanti (o gravi) rispetto ad altre. A vari tipi di disaccordo può essere assegnato un «peso» maggiore o minore (per esempio, verrà assegnato peso «0» agli accordi, «1» ai disaccordi meno importanti, «2» a quelli più gravi, ecc.), che, al momento del computo dell'indice di attendibilità, permetterà di assegnare una maggiore o minore rilevanza al disaccordo tra gli osservatori. Una volta assegnato un peso ai possibili disaccordi, il *K pesato* si calcola applicando la seguente formula:

$$K_w = 1 - \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}x_{ij}}{\sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}m_{ij}}$$

Nella formula sopra riportata K_w indica il *K pesato* (*weighted*), *w* indica il peso assegnato ai disaccordi, *x* gli accordi osservati e *m* gli accordi attesi per il caso.

6.4. Applicazione all'unità di codifica e a ogni singolo codice

In molti ambiti di ricerca non-osservativi, quanto finora descritto sull'attendibilità dei codici sarebbe sufficiente. Invece, nell'ambito degli studi osservativi, ciascun atto di codifica dell'osservatore richiede di fatto due distinti atti di codifica, cosa di cui spesso l'osservatore stesso non è consapevole. Riprendendo l'esempio riportato in un paragrafo precedente relativo all'esito di un episodio di negoziazione del gioco tra bambini, considerando come oggetto della codifica la «funzione attribuita al cambio di turno» durante un episodio di negoziazione, prima di domandarsi in quale categoria rientra l'evento osservato (per esempio, la funzione del turno di parola osservato è di formulare una controproposta?), è necessario porsi una domanda più basilare: l'evento osservato è un turno? La questione non riguarda la natura o il tipo di turno, ma il fatto che esso si sia verificato o meno.

Questa è la ragione per cui le ricerche basate sull'osservazione a volte riportano due indici di accordo, uno dei due è un indice di accordo relativo al verificarsi dell'evento (nell'esempio precedente, il turno). Questo tipo di accordo viene definito *accordo relativo all'unità di codifica* (Bakeman e Gottman, 1997).

Uno studio di attendibilità sull'unità di codifica è piuttosto semplice da eseguire nei casi in cui si usa il tempo per fare le misurazioni (Stati ed Eventi Temporali; Gnisci e Bakeman, 2000; Gnisci, Bakeman e Maricchiolo, 2013) oppure nei casi in cui si ricorre alle trascrizioni di un discorso o di una conversazione. Infatti, il ricorso ad unità di tempo consente la segmentazione precisa del flusso di interazioni, semplificando il riconoscimento dell'unità di codifica e rendendola uguale per tutti gli osservatori; il ricorso alle trascrizioni permette, a sua volta, di incorporare i codici direttamente nei trascritti, ancorando i codici a punti specifici, il che permette di verificare se i due osservatori hanno posto il codice nello stesso punto del trascritto. Ovviamente, sia quando si ricorre ad unità temporali sia quando si fa uso di trascritti, l'accordo sull'unità di codifica può tollerare alcune sfasature minime tra i codificatori, che si assume siano casuali o dovute a errori di approssimazione: per esempio, se un osservatore codifica un comportamento che inizia a 00:04 e finisce a 00:09, mentre l'altro osservatore codifica lo stesso comportamento che inizia a 00:05 e finisce a 00:10, è molto probabile che ai fini dell'accordo sull'unità di codifica entrambi abbiano visto lo stesso comportamento con tempi di inizio e fine simili.

In altre situazioni e per particolari sistemi di codifica, non è facile confrontare i modi in cui i due osservatori hanno diviso l'interazione in unità, perché non ci sono punti di riferimento. In genere questo accade quando si usano Eventi o Intervalli. Un esempio è dato dai turni di conversazione. È relativamente semplice verificare che entrambi i codificatori abbiano identificato correttamente un turno, o che un codificatore abbia identificato un turno non identificato dall'altro (e viceversa), ma è impossibile stabilire se entrambi, nello stesso momento, abbiano ignorato l'accadimento di un turno.

Per questo problema i ricercatori hanno proposto soluzioni pratiche ma non del tutto soddisfacenti, come fornire una percentuale di accordo relativa all'unità (oltre all'indice di attendibilità relativo alle categorie), considerare l'inizio e la fine dell'evento, usare l'intervallo tra parole adiacenti come confini potenziali tra i turni (Bakeman e Gottman, 1997), oppure far utilizzare i tempi agli osservatori solo nello studio di attendibilità. Recentemente, sono stati proposti e si sono dimostrati efficaci metodi più avanzati per l'allineamento delle unità di codifica di più osservatori, derivanti da simulazioni basate su algoritmi, come il *kappa dell'allineamento tra osservatori* (per cui v. Bakeman e Quera, 2011; Bakeman, Quera e Gnisci, 2009; Gnisci, Bakeman e Quera, 2008; Quera, 2008; Quera, Bakeman e Gnisci, 2007).

Esiste un'applicazione importante del *K* di Cohen. Se una matrice di confusione *n* x *n* viene ridotta a *n* matrici 2 x 2, si può calcolare un *K* per ciascuna tabella; si avranno perciò tanti *K* quanti sono i codici utilizzati.

Insieme all'osservazione qualitativa della matrice del *K*, ciascuno di questi *K* può essere utilizzato per individuare le categorie su cui i codificatori sono andati meno d'accordo ed, eventualmente, proporre dei correttivi. È dimostrato, infatti, che il *K* del sistema di codifica in generale è una media ponderata dei singoli *K* (Bakeman e Quera, 2011; per categorie ordinali v. Warrens, 2011).

6.5. Applicazione a più osservatori

Esiste una versione del *K* di Cohen, generalmente definita *K di Fleiss-Cohen*, applicabile per valutare l'accordo tra più di due osservatori. Essa è stata proposta da Fleiss (1981; Gwet, 2010) e rappresenta una generalizzazione del *K* di Cohen a più osservatori. Si basa sul calcolo della media delle concordanze stimate per ciascuna categoria, poste in rapporto alla proporzione di accordi attesi per caso[3].

6.6. Valutazione tramite la significatività

Dato che la distribuzione campionaria del *K* è nota (Fleiss, Cohen e Everitt, 1969), il suo errore standard e il suo intervallo di confidenza possono essere calcolati e il valore assunto dall'indice può essere testato per la significatività statistica. Tuttavia, essa si rivela insufficiente per stabilire la presenza di un vero accordo tra osservatori, dato che dipende dal numero di soggetti osservati o di frequenze della matrice di confusione. Perciò, un valore del coefficiente di attendibilità molto basso (per esempio, .20), date adeguate condizioni, può risultare significativamente diverso da zero così come un coefficiente alto (.80) può risultare non significativamente diverso da zero. Per questo gli indici di attendibilità raramente vengono testati per la significatività; di solito, invece, vengono comparati alla scala di riferimento, che in genere varia da 0 a 1.

6.7. Valutazione tramite la scala di riferimento e il buon senso

Dato che è sconsigliato far riferimento alla significatività del *K* di Cohen, è necessario interpretarlo sulla base della sua grandezza. Perciò sembrerebbe molto semplice interpretare la sua intensità in modo simile ai voti

[3] Brevi cenni introduttivi sul calcolo del *K* tra più osservatori (Fleiss' Kappa) e un programma per il calcolo online del relativo indice sono disponibili alla pagina web: http://www.statto-ols.net/CohenKappa_Exp.php.

scolastici da 0 a 10 (solo in scala da 0 a 1). I primi a fornire dei criteri di interpretazione del *K* di Cohen sono stati Landis e Koch (1977), che sostenevano che valori tra. 41 e .60 esprimono un accordo «moderato», tra .61 e .80 «sostanziale», e tra .81 e 1 «praticamente perfetto». Fleiss (1981) sosteneva che valori tra .40 e .60 esprimono accordo «medio», tra .60 e .75 «buono» e maggiori di .75 «eccellente». Bakeman e Gottman (1997), comunque, suggeriscono di considerare problematici valori inferiori a .70.

Tuttavia, questo approccio è semplicistico. Secondo alcuni, l'ampiezza del *K* non dipende solo dall'accordo, ma anche da altri aspetti (distribuzione reale dei comportamenti codificati, similitudine delle distribuzioni marginali tra osservatori, numero di codici). Per questo motivo, a differenza di come appaiono in prima battuta, tutte le interpretazioni del *K* sulla base della scala 0-1 divengono indicative e non assolute. Perciò, i criteri di riferimento forniti sopra sono da considerarsi delle regole euristiche, che non hanno una base propriamente scientifica o logica, e sono basate principalmente sull'esperienza dei propositori.

6.8. Le critiche al *K* di Cohen

Le critiche al *K* di Cohen sono sostanzialmente due, ma hanno delle profonde conseguenze. Esse riguardano la sostenibilità del modello di accordo dovuto al caso e la relatività della sua valutazione.

6.9. Il modello dell'accordo dovuto al caso è sostenibile?

Sebbene quello del *K* di Cohen sia considerato l'approccio più promettente proprio per i motivi descritti, alcuni studiosi sostengono che il fatto di valutare correttamente se l'accordo è maggiore o minore di quello dovuto al caso, sia insufficiente per candidare il *K* a essere un indice adeguato di accordo (Gwet, 2008). Secondo questa prospettiva critica, quando gli osservatori effettuano una codifica del comportamento, essi non codificano a caso, quindi la correzione apportata dal *K* è inadeguata; è necessario, al contrario, che si costruisca un modello basato su come gli osservatori prendono le decisioni di codifica e si proponga una correzione adatta a questo caso specifico (Uebersax, 1987). L'assunzione di indipendenza su cui si basa la correzione dovuta al caso nel *K*, perciò, non sarebbe in alcun modo legittima. In assenza di un modello esplicito su come gli osservatori prendono le decisioni quando codificano, non si può sapere come il caso influenzi le decisioni degli osservatori e quale correzione possa essere adeguata.

6.10. I limiti del *K*

Sebbene si sostenesse il contrario, alcuni autori hanno messo in luce da tempo che molte delle caratteristiche negative attribuite alla percentuale di accordo non erano risolte dal *K* di Cohen. Alcuni studiosi sostengono che l'ampiezza del *K* dipende non solo dall'accordo, ma anche da altri fattori, come l'equiprobabilità delle categorie e la similitudine nelle distribuzioni marginali dei due osservatori, nonché il numero di categorie utilizzate. Perciò, ogni sua valutazione rispetto alla scala è relativa e non possono essere fatte comparazioni tra indici tratti da studi diversi. Questi, infatti, se anche utilizzano lo stesso numero di categorie, difficilmente avranno le stesse probabilità marginali.

6.11. Da cosa dipende il *K*?

In uno studio fondamentale, Sim e Wright (2005; v. anche Bruckner e Yoder, 2006; Simon, 2006) hanno finalmente dimostrato che il *K* di Cohen dipende, oltre che dall'accuratezza degli osservatori e, in negativo, dalla loro mancanza di indipendenza, anche da tre ulteriori caratteristiche: 1) dalla *somiglianza o differenza tra le frequenze o probabilità marginali* dei due osservatori, cioè dal cosiddetto *errore sistematico* degli osservatori; 2) dal *numero di categorie* di un sistema di codifica; 3) dalla *prevalenza* delle categorie, cioè da come le categorie sono distribuite nella popolazione di riferimento.

Quando c'è un grande errore sistematico dell'osservatore, cioè le probabilità marginali sono molto diverse, è dimostrato che, anche se gli osservatori vanno sempre d'accordo (cioè vanno d'accordo tutte le volte che possono, stante quei valori marginali), il *K* non può raggiungere 1.

Una soluzione è calcolare il *K massimo* (Umesh, Peterson e Sauber, 1989), cioè il massimo valore che il *K* può assumere stante quelle distribuzioni marginali. Esso viene definito in questo modo:

$$K_{MAX} = \frac{P_{MAX} - P_C}{1 - P_C} \text{ in cui, } P_{MAX} = \sum_{i=1}^{n} min(p_{+i}, p_{i+})$$

cioè una funzione che somma la minima tra le probabilità marginali di riga (p_{i+}) o di colonna (p_{+i}) per ciascuna categoria di comportamento (i), n è il numero dei codici, e P_C è la probabilità dell'accordo dovuto al caso.

Tuttavia, è necessario evitare la tentazione di calcolare un nuovo *K* dividendo il *K* ottenuto per il *K massimo*, per fare in modo che esso possa variare tra 0 e 1: questo sarebbe solo un espediente per far aumentare indebitamente il *K* (Bakeman e Quera, 2011).

Per avere delle ulteriori indicazioni, Bakeman, Quera, McArthur e Robinson (1997) hanno fatto uno *studio di simulazione*, in cui hanno calcolato i valori del *K* in funzione del *numero delle categorie*, della loro *prevalenza* (equiprobabile, poco e molto variabile) e dell'*accuratezza* degli osservatori (assumendo che siano ugualmente accurati). Oltre a una serie di altre indicazioni, i risultati hanno mostrato che: *a*) il *K* aumenta all'aumentare del numero di codici, ma oltre i 5-7 codici l'incremento è piccolo e trascurabile; *b*) il *K* varia al variare della prevalenza, specie sotto i 5 codici: più la prevalenza è equiprobabile, maggiore è il valore del *K*; dopo i 5 codici il *K* praticamente non dipende dalla prevalenza e tende a stabilizzarsi; *c*) sotto i 5 codici, infine, valori bassi del *K* predicono alti valori di accuratezza. Per esempio, un *K* di .20 può predire un'accuratezza dell'80%! Questo conferma ciò che è stato più volte sostenuto, e cioè che il *K* è un indice di accordo conservativo (Strijbos *et al*., 2006). Perciò, al momento della sua interpretazione, anche valori sufficienti fornirebbero indizi della presenza di un buon accordo.

7. Alternative al *K* di Cohen: l'α di Krippendorff o la modellistica log-lineare

Esistono molte alternative al *K* di Cohen (cfr. von Eye e von Eye, 2005; Warrens, 2010) – la *S* di Bennett (Bennett, Alpert e Goldstein, 1954), la π di Scott (1955), il *K* di Fleiss (1971), l'α di Cronbach (1951) –, tuttavia quella più credibile sembra essere l'α di Krippendorff (2004; Hayes e Krippendorf, 2007). Piuttosto che di un singolo coefficiente di attendibilità, si tratta di una famiglia di coefficienti molto generale, che si applica a diverse situazioni di ricerca. Si usa indipendentemente dal numero di osservatori (cioè, a più di due), dai livelli della misurazione (può essere applicato a scale nominali, ordinali e metriche, che divengono perciò comparabili), dal numero delle categorie e delle frequenze rilevate da ciascun osservatore, dalle dimensioni del campione, dalla eventuale presenza di dati mancanti (per un'applicazione v. Hayes e Krippendorff, 2007). La sua distribuzione è calcolabile tramite tecniche di simulazione. Va infine notato che ora esistono molte macro (in SPSS, SAS e R)[4] che consentono il calcolo di questa statistica, caso alquanto difficile negli anni passati.

[4] Per scaricare le macro per SPSS e SAS v. http://www.afhayes.com/spss-sas-and-mplus-macros-and-code.html, per R v. http://www.R-project.org o http://cran.r-project.org/web/packages/irr/irr.pdf.

Un'altra possibilità verso cui la letteratura si sta dirigendo è l'uso di una modellistica log-lineare (Nussbeck, 2005) basata su tabelle in cui ciascuna variabile è costituita da un osservatore (come fossero metodi diversi), i livelli delle variabili sono costituiti dalle categorie di codifica e in ogni cella si distribuiscono gli accordi-disaccordi tra i due osservatori (cioè, una matrice del *K* ma con la possibilità di inserire più osservatori). Un'analisi dei modelli log-lineari e una conseguente analisi dei pattern associativi tramite residui standard o corretti (Gnisci e Pedon, 2011), indica se il modello che si adatta meglio ai dati suggerisce accordo tra osservatori. In genere, queste analisi vengono condotte utilizzando il modello di «quasi indipendenza», il quale possiede un parametro particolare che identifica un'associazione legata alle celle della diagonale (cioè quelle in cui si verifica l'accordo). Questa soluzione ha il pregio di basarsi su una solida tradizione statistica per le variabili qualitative.

8. Conclusioni

Quando uno psicologo clinico o un ricercatore intendono osservare in maniera sistematica il comportamento di un bambino o l'interazione del bambino con figure significative o in setting terapeutici, si trovano nella necessità di utilizzare dei codificatori. Spesso essi si trovano disorientati di fronte a problemi che sorgono dall'attendibilità della codifica dei codificatori. L'utilità delle informazioni presentate in questa rassegna per chi fa ricerca è diretto. Tuttavia, tali informazioni sono importanti anche per chi opera in psicologia clinica dello sviluppo. Innanzitutto, qualora l'osservazione del bambino sia, insieme ad altri strumenti di *assessment* psicologico, alla base della formulazione della diagnosi, esse permettono di scegliere gli strumenti più idonei (dalle griglie di codifica ai coefficienti di attendibilità), migliorando l'accuratezza diagnostica, e di valutare con più consapevolezza e capacità critiche i risultati dell'osservazione stessa. L'osservazione è anche uno strumento per la valutazione del cambiamento del bambino e della famiglia in psicoterapia. Perciò, un corretto uso dell'osservazione e delle tematiche relative al tema dell'attendibilità si rileva utile anche nel corso o alla fine del percorso psicoterapeutico.

In questo contributo sono stati presentati alcuni tra i concetti e le teorie tradizionali e più aggiornate riguardanti l'attendibilità delle misure osservative. Ci si è soffermati sulla differenza tra accordo tra osservatori, calibrazione e attendibilità, sul legame tra l'attendibilità intra-osservatore, inter-osservatori e dell'osservatore e i concetti di precisione, stabilità e accuratezza, sui diversi indici e procedure di verifica dell'at-

tendibilità (*K* di Cohen, *K pesato*, *ICC*, ecc.) e sui problemi della loro interpretazione, sui loro vantaggi e limiti. Riguardo quest'ultimo punto, sono state presentate due classi di indici, attualmente disponibili: i coefficienti globali che rendono conto dell'andamento congiunto generale tra due osservatori e i coefficienti punto-per-punto che, invece, riguardano l'accordo rispetto a ciascun accadimento.

Dall'analisi della letteratura presentata e dello stato attuale dell'arte emerge che la percentuale di accordo è l'indice meno consigliato – sebbene esso continui ad essere utilizzato – mentre il *K* di Cohen è l'indice di attendibilità suggerito dai ricercatori. Esso ha notevoli vantaggi: corregge per il caso, può essere pesato se le categorie sono ordinali, può essere applicato al sistema di codifica, alla singola categoria e all'unità di codifica, può essere generalizzato al caso di più di due osservatori, valutato in base alla significatività e alla scala di riferimento. Tuttavia, negli ultimi anni si sono venuti chiarendo anche i suoi aspetti critici: il modello su cui è fondato – quello dell'accordo dovuto al caso – non è un modello realistico, che riflette ciò che gli osservatori fanno realmente; inoltre, esso dipende da alcuni aspetti – il numero di categorie del sistema, la differenza nelle probabilità marginali dei due osservatori e la prevalenza del fenomeno – che ne rendono i valori non facilmente valutabili e confrontabili, anche se il fatto che esso sia un indice conservativo può fornire comunque delle maggiori garanzie. Allo stato attuale, gli studiosi focalizzati sugli indici di attendibilità sono impegnati in due diverse direzioni, migliorare le caratteristiche del *K* tenendo conto delle limitazioni ad esso attribuite (e.g., Bakeman *et al*., 1997; Bakeman e Quera, 2011) oppure proporre nuovi indici che non risentano dei limiti del *K* (e.g., Gwet, 2002, 2010), come nel caso dell'α di Krippendorff. Altri ricercatori si sono invece diretti verso la modellistica log-lineare per identificare le associazioni nella codifica tra osservatori.

9. Riferimenti bibliografici

Bakeman, R., Gottman, J.M. (1997). *Observing interaction. An introduction to sequential analysis*. 2nd Edition. New York: Cambridge University Press.

Bakeman, R., Quera, V. (2011). *Sequential analysis and observational methods for the behavioral sciences*. New York: Cambridge University Press.

Bakeman, R., Quera, V., Gnisci, A. (2009). Observer agreement for timed-event sequential data: A comparison of time-based and event-based algorithms. *Behavior Research Methods*, 41, 137-147.

Bakeman, R., Quera, V., McArthur, D., Robinson, B.F. (1997). Detecting sequential patterns and determining their reliability with fallible observers. *Psychological Methods*, 2, 357-370.

Bearison, D.J., Dorval, B., LeBlanc, G., Sadow, A., Plesa, D. (2001). *Collaborative cognition: Children negotiating ways of knowing*. Westport, CT: Ablex.

Bennett, E.M., Alpert, R., Goldstein, A.C. (1954). Communications through limited response questioning. *Public Opinion Quarterly*, 18, 303-308.

Berk, R.A. (1979). *Generalizability of behavioral observation: A clarification of interobserver agreement and interobserver reliability*. Cambridge: Cambridge University Press.

Brennan, R.L. (1983). *Elements of generalizability theory*. Iowa City: The American College Testing Program.

Bruckner, C.T., Yoder, P. (2006). Interpreting kappa in observational research: Baserate matters. *American Journal of Mental Retardation*, 111, 433-441.

Bryington, A.A., Palmer, D.J., Watkins, M.W. (2004). The estimation of interobserver agreement in behavioral assessment. *Journal of Early and Intensive Behavior Intervention*, 1, 115-119.

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

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

Cone, J.D. (1987). Behavioral assessment: Some things old, some things new, some things borrowed? *Behavioral Assessment*, 9, 1-4.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrica*, 16, 297-334.

Cronbach, L.J., Gleser, G.C., Nanda, H., Rajaratnam, N. (1972). *The dependability of behavioral measurement: Theory of generalizability for scores and profiles*. New York: Wiley.

Fleiss, J.L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76, 378-382.

Fleiss, J.L. (1981). *Statistical methods for rates and proportions*. New York: Wiley.

Fleiss, J.L., Cohen, J., Everitt, B.S. (1969). Large sample standard errors of kappa and weighted kappa. *Psychological Bulletin*, 72, 323-327.

Gnisci, A., Bakeman, R. (2000). *L'osservazione e l'analisi sequenziale dell'interazione*. Milano: LED.

Gnisci, A., Bakeman R., Quera, V. (2008). Blending qualitative and quantitative analyses in observing interaction. *International Journal of Multiple Research Approaches*, 2, 15-30.

Gnisci, A., Bakeman, R., Maricchiolo, F. (2013). Sequential notation and analysis for bodily forms of communication. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill e S. Teßendorf (a cura di), *Body, language, communication: An international handbook on multimodality in human interaction*. Vol. I. Berlin-New York: Mouton de Gruyter, 892-903.

Gnisci, A., Maricchiolo, F., Bonaiuto, M. (2013). Reliability and validity of coding systems for bodily forms of communication. In C. Müller, A. Cienki, E. Fricke, S.H. Ladewig, D. McNeill e S. Teßendorf (a cura di), *Body, language, communication: An international handbook on multimodality in human interaction*. Vol. I. Berlin-New York: Mouton de Gruyter, 879-892.

Gnisci, A., Pedon, A. (2011). *La ricerca nelle scienze sociali con i log-lineari*. Roma: Armando Editore.

Gwet, K.L. (2002). Kappa statistic is not satisfactory for assessing the extent of agreement between raters. *Statistical Methods for Inter-Rater Reliability Assessment*, 1, 1-5.

Gwet, K.L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61, 29-48.

Gwet, K.L. (2010). *Handbook of inter-rater reliability*. 2nd Edition. Gaithersburg: Advanced Analytics, LLC.

Hartmann, D.P. (1982). Assessing the dependability of observational data. In D.P. Hartmann (a cura di), *New directions for the methodology of behavioral sciences: Using observers to study behavior*. San Francisco: Jossey-Bass, 51-65.

Hayes, A.F., Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1, 77-89.

John, O.P., Benet-Martínez, V. (2000). Measurement: Reliability, construct validation, and scale construction. In H.T. Reis e C.H. Judd (a cura di), *Handbook of research methods in social and personality psychology*. Cambridge: Cambridge University Press, 339-369.

Krippendorff, K. (2004): Measuring the reliability of qualitative text analysis data. *Quality & Quantity*, 38, 787-800.

Landis, J.R., Koch, G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33, 159-174.

Martin, P., Bateson, P. (1986). *Measuring Behavior: An introductory guide*. Cambridge: Cambridge University Press.

McGraw, K.O., Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, 1, 30-46.

Nussbeck, F.W. (2005). Assessing multimethod association with categorical variables. In M. Eid e E. Diener (a cura di), *Handbook of multimethod assessment in psychology*. Washington, D.C.: American Psychological Association, 231-247.

Patterson, G.R. (1982). *Coercive family process*. Eugene, OR: Castalia Press.

Pedon, A., Gnisci, A. (2004). *Metodologia della ricerca psicologica*. Bologna: Il Mulino.

Pedon, A., Gnisci, A. (2012). *Manuale di psicodiagnostica*. Firenze: Le Lettere.

Quera, V. (2008). RAP: A computer program for exploring similarities in behavior sequences using random projections. *Behavior Research Methods*, 40, 21-32.

Quera, V., Bakeman, R., Gnisci, A. (2007). Observer agreement for event sequences: Methods and software for sequence alignment and reliability estimates. *Behaviour Research Methods*, 39, 39-49.

Scott, W.A. (1955). Reliability of content analysis: The case of nominal scale coding. *Public Opinion Quarterly*, 19, 321-325.

Shrout, P.E., Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86, 420-428.

Sim, J., Wright, C.C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85, 257-268.

Simon, P. (2006). Including omission mistakes in the calculation of Cohen's kappa and analysis of the coefficient's paradox feature. *Educational and Psychological Measurement*, 66, 765-777.

Strijbos, J.W., Martens, R.L., Prins, F.J., Jochems, W.M.G. (2006). Content analysis: What are they talking about? *Computers and Education*, 46, 29-48.

Suen, H.K. (1988). Agreement, reliability, accuracy and validity: Toward a clarification. *Behavioral Assessment*, 10, 343-366.

Suen, H.K., Ary, D. (1989). *Analyzing quantitative behavioral observation data*. N.J.: Lawrence Erlbaum Associates Hillsdale.
Suen, H.K., Ary, D., Ary, R. (1986). A note on the relationship among eight indices of interobserver agreement. *Behavioral Assessment*, 8, 301-303.
Towstopiat, O. (1984). A review of reliability procedures for measuring observer agreement. *Contemporary Educational Psychology*, 9, 333-352.
Uebersax, J.S. (1987). Diversity of decision-making models and the measurement of interrater agreement. *Psychological Bulletin*, 101, 140-146.
Umesh, U.N., Peterson, R.A., Sauber, M.H. (1989). Interjudge agreement and the maximum value of kappa. *Educational and Psychological Measurement*, 49, 835-850.
von Eye, A., von Eye, M. (2005). Can one use Cohen's kappa to examine disagreement? *Methodology*, 1, 129-142.
Warrens, M.J. (2010). Inequalities between kappa and kappa-like statistics for k x k tables. *Psychometrika*, 75, 176-185.
Warrens, M.J. (2011). Cohen's linearly weighed kappa is a weighted average of 2 x 2 kappas. *Psychometrika*, 76, 471-486.
Watkins, M.W., Pacheco, M.E. (2000). Interobserver agreement in behavioral research: Importance and calculation. *Journal of Behavioral Education*, 10, 205-212.

[*Ricevuto il 30 gennaio 2013*]
[*Accettato il 04 marzo 2015*]

Reliability of observational measures in development clinical psychology

Summary. Systematic observation of behavior is a technique widespread in development clinical psychology. However, the issues concerning a reliable coding are barely described in a clear and systematic way. This contribution proposes a review that, without neglecting the corrected psychometrics bases, organizes the traditional knowledge and on it built an up-to-date and critical picture of the state of art on the observer reliability. Along with the reference concepts (agreement, calibration, reliability, precision, stability, accuracy, etc.), it treats the main coefficients proposed in the literature – first of all the Cohen's *K* – and show their advantages and limitations.

Keywords: Systematic observation, observer agreement, reliability, Cohen's *K*, reliability coefficients.

Per corrispondenza: Augusto Gnisci, Dipartimento di Psicologia, Seconda Università di Napoli, Viale Ellittico 31, 01100 Caserta. E-mail: augusto.gnisci@unina2.it