Diagnosing psychotic disorders: validity, reliability and applications of the Diagnostic Interview for Psychosis (DIP). Italian version

ALBERTO ROSSI,[1] VERA MORGAN,[2] FRANCESCO AMADDEO,[1] MARCO SANDRI,[1]
LAURA GRIGOLETTI,[1] FRANCESCA MAGGIONI,[1] ADELE FERRO,[1] ELENA RIGON,[1]
VALERIA DONISI,[1] VALERIA VAILATI VENTURI,[1] FABRIZIO GORIA,[3]
INGUNN SKRE,[4] MICHELE TANSELLA,[1] ASSEN JABLENSKY[3]

[1]*Department of Medicine and Public Health, Section of Psychiatry and Clinical Psychology, University of Verona, Verona (Italy)*
[2]*School of Psychiatry and Clinical Neurosciences. The University of Western Australia. Perth (Australia)*
[3]*Centre for Clinical Research in Neuropsychiatry & School of Psychiatry & Clinical Neurosciences,*
The University of Western Australia. Perth (Australia)
[4]*Department of Psychology, University of Tromsø. Tromsø, Norway*

SUMMARY. **Aims** – The Diagnostic Interview for Psychoses (DIP) is a comprehensive interview schedule for psychotic disorders, linked to the OPCRIT diagnostic algorithm, bridging the gap between fully structured, lay-administered schedules and semi-structured, psychiatrist-administered interviews. Here we describe the validity, reliability and applications of the Italian version of the DIP. **Methods** – The interview was translated into Italian and its content validity tested by back translation. Sixty patients, drawn from among those who contacted the South-Verona Community Mental Health Service, were included in the study. Each patient was first assessed independently by two raters, one of whom conducted the interview, while the other assumed the role of observer. Subsequently (median: 89 days), 44 of these patients were re-interviewed by a third rater, who made an independent assessment. Diagnostic validity was assessed in 18 cases, interviewed with the DIP and using the SCAN as 'gold standard'. **Results** – The mean duration of the interview was 37 minutes for the *inter-rater* interviews and 39 minutes for the *retest* interviews. Good to excellent *inter-rater reliability* was demonstrated for both ICD-10 and DSM-IV diagnoses, while in the *test-retest reliability* pairwise agreement was high for half of the items. Diagnostic validity was good, with twelve out of the 18 DIP-OPCRIT diagnoses (67%) matching the SCAN diagnosis. **Conclusions** – Overall, the results support the reliability and validity of the Italian translation of the DIP. The Italian version will be useful both in routine practice to establish standard reference diagnoses of psychosis and in the research field, where it can be used by academic researchers in clinical trials and epidemiological studies.
Declaration of Interest: None.

KEY WORDS: psychiatric diagnostic structured interviews, psychotic disorders, reliability study, Cohen's Kappa.

INTRODUCTION

In psychiatry, diagnostic validity can be defined as the extent to which a diagnostic category meets a consensus definition of a psychiatric disorder and is distinguishable from other disorders. Analogously, diagnostic reliability refers to the agreement between different clinicians applying common diagnostic decision rules. The validity and reliability of psychiatric diagnosis has improved with the use of operational criteria (Spitzer *et al.*, 1979; Fyer *et al.*, 1989; Sartorius *et al.*, 1993) and corresponding diagnostic structured interviews (Spitzer & Williams, 1988; Mannuzza *et al.*, 1989; Spitzer *et al.*, 1992; Balestrieri *et al.*, 2007).

The most frequently used clinical tools range from 'hard-wired', fully structured interview schedules designed for use by lay interviewers, such as the *Composite International Diagnostic interview for* ICD-

Address for correspondence: Dr. A. Rossi, Department of Medicine and Public Health, Section of Psychiatry and Clinical Psychology, Ospedale Policlinico, Piazzale L.A. Scuro 10, 37134 Verona (Italy).
Fax: +39-045-8027498
E-mail: alberto.rossi@univr.it

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

10 (CIDI) (World Health Organization, 1990), to comprehensive semi-structured interviews, such as the *Schedules for Clinical Assessment in Neuropsychiatry* (SCAN) (World Health Organization, 1992) and the *Structured Clinical Interview for DSM-IV* (SCID) (First *et al*., 1997) whose administration requires clinical experience and judgement. Comparisons of survey findings obtained by means of lay- and clinician-administered instruments have suggested that the two types of interview may result in discrepant diagnostic classification of cases, especially in the assessment of psychotic disorders (Brugha *et al*., 1999). There is thus a need for an instrument that allows clinicians to establish accurate diagnoses of psychosis using a structured interview schedule that ensures uniformity of use but which also allows clinical expertise and experience to be factored into the decision-making about reported symptoms.

To meet these aims, the *Diagnostic Interview for Psychoses* (DIP) was developed. The DIP is a comprehensive interview schedule that bridges the gap between highly structured lay interviews such as the CIDI and comprehensive schedules such as the SCAN. The DIP is a semi-structured clinical interview for recording symptoms and obtaining a diagnosis and is intended for use by interviewers with a clinical background (psychiatrists, mental health nurses, clinical psychologists, and allied disciplines). The DIP encompasses the following main domains (see Appendix I):

a) demographic data;
b) social functioning and disability;
c) symptoms, signs and past history items required for the diagnosis of psychotic disorders (diagnostic module, DIP-DM); and
d) patterns of service utilization and patient-perceived unmet need for services.

Where appropriate, the diagnostic module (DIP-DM), which takes 20-30 minute to complete, can be used alone (the CIDI, the SCAN and the SCID require from 1 to 2 hours to administer).

The structure of the DIP-DM follows the *Operational Criteria for Psychosis*, OPCRIT, version 3.31 (McGuffin *et al*., 1991; Williams *et al*., 1996) 90-item checklist. The OPCRIT is essentially a phenomenological checklist that can be rated from practically any source and which, through the allied computerized algorithm, generates diagnoses according to the criteria of ICD-10 (World Health Organization, 1993); DSM-III (American Psychiatric Association, 1980), DSM-III-R (American Psychiatric Association, 1987) and DSM-IV (American Psychiatric Association, 1994); the *Research Diagnostic Criteria* (Spitzer *et al*., 1978); the St Louis criteria (Feighner *et al*., 1972), the Carpenter *et al*. (1973) classification, the French Classification (Pull *et al*., 1987) and the Taylor Abrams criteria (Taylor & Abrams, 1978). OPCRIT also allows subtyping of psychotic disorders according to the typologies proposed by Crow (1980), Tsuang & Winokur (1974), and Farmer *et al*. (1983).

The diagnostic module (DIP-DM) consists of a series of interview questions and probes either written de novo or, where relevant, using wording from the SCAN, version 2.0 (Wing *et al*., 1990; 1998), to elicit the OPCRIT checklist items. Questions are formulated in such a way as to allow the interviewer to ask about present state, past year, or lifetime occurrence of symptoms. The items are ordered in a way allowing a natural progression to be followed, with symptoms being grouped into sections on depression, psychotic symptoms, and behavior and affect. The DIP also includes a section on drug and alcohol use, to allow rating of these factors. The full list of items and their ordering is detailed in Appendix I. Whilst essentially interview-based, the DIP also encourages use of other sources of information, where available. For example, information on pre-morbid functioning and family history of psychiatric illness can be augmented by interview with a family member, although this is not mandatory. Signs, such as affect, psychomotor behavior, or form and flow of speech, can be rated on the basis of observation during the interview, as well as using relevant information provided by informants or in clinical case-notes. Responses to the interview questions are eventually entered onto a computer database where the underlying OPCRIT algorithm generates diagnoses according to the above listed diagnostic classification systems.

The development of the DIP in its English version had been previously described (Castle *et al*., 2006). Here we describe the validity, reliability and applications of the Italian version of the DIP.

METHODS

Translation procedures

The Italian version of the DIP was developed in a process that started with the translation of the original English language version (Castle *et al*., 2006) by a translator whose native language was Italian. It was then back-translated by an experienced bilingual psychiatrist whose native language was Italian, but who was trained as a psychiatrist in Australia. Following the back-translation,

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

both the Italian version and the back-translation were reviewed and revised by the original authors and the translators. The final version was corrected for any inconsistencies and related problems detected by the research team, and a final agreement was reached on the content validity of the instrument (Prince, 2008).

Sampling for the reliability study

Reliability studies should be conducted in populations and settings where the distribution of factors that may influence the diagnostic process is similar to that in the populations and settings in which the assessment technique will ultimately be applied (Thompson & Walter, 1988). Accordingly, the subjects for this study were drawn from among persons contacting the *South-Verona Community Mental Health Service* (CMHS) (Tansella *et al*., 1998). Sixty consecutive patients were included, all aged 18 years or older. Subjects with dementia, mental retardation or language problems were excluded. All patients were interviewed only after written informed consent had been obtained, following full explanation of the purpose of the study by research staff, provision of details to each patient in writing, and making it clear that participation was entirely voluntary, that they could choose whether to participate or not, and that they could withdraw from the study at any time, without detriment to their clinical care.

Participants were compensated to the amount of 10 Euros for their time.

Data collection

Each patient was first assessed independently by two interviewers (*inter-rater reliability* test), one of whom conducted the interview, while the other assumed the role of observer. Within the agreed protocol, at the end of each section the observer was allowed to repeat any interview questions if there was disagreement with the interviewer about the use of skips between and within sections (the interview contains a number of built-in cut-offs and skips between and within sections to avoid redundant questioning when initial screening questions have indicated that psychopathology is unlikely to be present in that section - see appendix II). In this way, the source of variability was not constrained in this protocol, as the assessors could rate different responses during the same interview.

After the initial rating, a third interviewer made an independent assessment on forty-four of these patients using the DIP, blind to the result of the first interview (*test-retest reliability test*). Accordingly with Castle *et al*. (2006), we asked to the raters of the retest to do the second interview with intervals from 2 to 11 weeks from the first interview.

Raters and training

Three psychiatrists in training and four clinical psychologists conducted the interviews. None of them knew the patients prior to the interview, nor had examined the case notes. The seven raters were assigned to the roles of interviewer, observer and re-interviewer, according to a balanced randomised design. To ensure comparability between raters, all of them received a two weeks training sessions in the use of the DIP.

Data analysis

In the *inter-rater reliability* analysis, every DIP rated by the interviewer was compared with that rated by the observer, while in the *test-retest reliability* analysis, the interviews rated by the interviewers of the first and the second interview were compared. For each pair of interviews, all the 90 OPCRIT items and the ICD-10 and DSM-IV diagnoses were considered in the analysis. In both the *inter-rater* and *test-retest* comparisons, we used the o*verall pairwise agreement* (PAR), i.e. the ratio of the number of agreements between observers/raters to the total number of comparison made, and the *kappa statistic*, to measure the degree of agreement between two raters for each observation. The *kappa statistic* (Cohen, 1960), an index of diagnostic agreement commonly used in psychiatric research, is the statistic of choice for categorical data as it adjusts the observed agreement rate for the agreement due to chance. However, a disadvantage of *kappa* is that it is affected by the prevalence of the symptoms or the disorder, so that items that show high sensitivity and specificity may have low predictive accuracy if the prevalence of the symptom or the disorder is low (Feinstein & Cicchetti, 1990). Therefore, Cicchetti & Feinstein (1990) have proposed a simple way to solve the problem: the *kappa* index should always be accompanied by the observed proportions of ‘positive’ agreement, p_{pos} (i.e. agreement on the presence of the symptom) and ‘negative’ agreement, p_{neg} (i.e. agreement on the absence of the symptom).

Stata 10.0 (StataCorp, 2007) was used for data management, descriptive statistics, Cohen’s kappa estimation, and the assessment of positive and negative agreement.

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

RESULTS

Reliability study sample

As this study was conducted in a case register area, it was possible to obtain directly from the South-Verona Psychiatric Case Register patients' socio-demographic and clinical characteristics of the patients (Tansella *et al.*, 1998). The mean age for the 60 subjects included in the analysis was 52 years (SD = 14). Thirty-five of the subjects (58%) were female, 9 (15%) were married, 29 (48%) lived with a partner or family, 9 (15%) had a higher level of education, and 18 patients (30%) were employed. In this sample, there was a marked prevalence of schizophrenic (55%) and affective (27%) disorders; 10% had personality disorders and the rest were distributed among other diagnostic groups.

The median length of time between the first (*inter-rater*) and the second (*retest*) interview was 89 days (min 24, max 351, inter-quartile range 82.5). The mean duration of the interview was 37 minutes (SD = 18) in the *inter-rater* (range: 15-89) and 39 minutes (SD = 15) in the *retest* (range: 20-80) study.

Inter-rater reliability

Table I shows, for a selection from the 90 OPCRIT items in the diagnostic module, the overall PAR and kappa ranges with 95% confidence intervals (except for items where the positive response frequency was too low for calculating these indices). A full list of the items with their individual kappa/PAR reliability coefficients is available on request. The table also show the values of p_{pos} and p_{neg} for every single item. According to the standard benchmarks proposed by Landis & Koch (1977) (<0 no agreement; 0-0.19 poor agreement; 0.20-0.39 fair agreement; 0.40-0.59 moderate agreement; 0.60-0.79 good agreement; 0.80-1.00 excellent agreement) inter-rater reliability was very high. In terms of PAR, 88 items had a rate of 0.8-1.0. Using the kappa statistic, 83% (75) of the items achieved a *kappa* value of ≥0.6, i.e. good to excellent concordance, with 71% (64 items) in the >0.8 range. A *kappa* of <0.4 was obtained for 13% (12 items), of which five items resulted in a *kappa* of zero (these, and several other items with a low *kappa*, actually had attained a high PAR). The majority of these items contained dichotomous response categories (yes/no) where almost all of the responses fell into one category, causing the data to be skewed. Thus, the zero or low *kappa* in these instances could be attributed to the instability of *kappa* when the data distributions are skewed and small variations can cause large fluctuations in *kappa* values. Both p_{pos} and p_{neg} values were very high in almost all the items: 87/90 p_{pos} values and 71/90 p_{neg} values were 0.80 or greater (indicating high positive and negative agreement). We also calculated the median *inter-rater reliability* of the items eliciting positive symptoms of schizophrenia (e.g. hallucinations, delusions) and that of the items eliciting negative symptoms (e.g. deterioration from premorbid level of functioning). The results showed an excellent agreement (median kappa 0.90, inter-quartile range 0.09) for positive symptoms and a good agreement (median kappa 0.60, inter-quartile range 0.42) for negative symptoms (Figure 1). The difference between the median *kappa* of positive and negative symptoms was statistically significant (Wilcons rank-sun test, z=2.48, *P*=0.013).

The DIP software currently generates the ICD-10 and DSM-IV OPCRIT diagnostic subtypes detailed in Appendix III. We aggregated the diagnoses of the two classifications into broader diagnostic groupings (broad categories), identical for both ICD-10 and DSM-IV: schizophrenia, schizoaffective disorder, bipolar disorder, depressive disorder with/without psychotic features, and other psychosis. Table II summarizes *inter-rater* agreement on ICD-10 and DSM-IV diagnoses when two alternative groupings of the diagnostic rubrics were used:

a) the OPCRIT diagnostic subtypes (narrow diagnostic categories); and
b) the broad diagnostic categories of the two classifications.

At the level of detailed, narrow diagnostic breakdown, the *inter-rater reliability* results for both ICD-10 and DSM-IV diagnoses showed good agreement beyond chance according to PAR (0.87 for ICD-10 and 0.78 for DSM-IV). Using *kappa*, agreement was high both for ICD-10 diagnoses (*kappa*=0.84) and for DSM-IV diagnoses (*kappa*=0.74). Analysis of discrepancies between the ICD-10 and DSM-IV results on a case by case basis indicated as the main reason the sensitivity of the OPCRIT algorithm to small differences between raters in the coding of items used to allocate a case to a specific DSM-IV diagnostic category. In the light of this, additional analysis was conducted based on the broad categories. In the aggregated format, the reliability was similar for ICD-10 (kappa=0.85) and DSM-IV diagnostic categories (kappa=0.74).

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

Table I – *DIP-Italian version inter-raters and test-retest reliability – selected items*[1]

	Inter-rater reliability (No. of cases: 60) (No. of reters: 7)						Test-retest reliability (No. of cases: 44) (No. of raters: 7)					
	p_{po}[2]	p_{neg}[2]	Overall pairwise agreement	Kappa	Kappa 95% CI	Level of agreement[3]	p_{pos}[2]	p_{neg}[2]	Overall pairwise agreement	Kappa	Kappa 95% CI	Level of agreement[3]
Age of onset	N.C.	N.C.	0.92	0.91	(0.84-0.98)	Excellent	N.C.	N.C.	0.32	0.30	(0.18-0.46)	Fair
Mode of onset	N.C.	N.C.	0.92	0.90	(0.77-0.98)	Excellent	N.C.	1.00	0.39	0.18	(-0.01-0.37)	Poor
Psychosocial stressor prior to first episode	1.00	1.00	1.00	1.00	(N.C.)	–[4]	0.50	0.77	0.69	0.27	(-0.03-0.57)	Fair
Pre-morbid personality disorder	0.95	0.54	0.92	0.50	(0.12-0.88)	Moderate	0.90	0.36	0.84	0.27	(-0.12-0.67)	Fair
Dysphoria	1.00	1.00	0.97	0.94	(0.84-1.00)	Excellent	0.61	0.83	0.66	0.36	(0.16-0.60)	Fair
Suicidal ideation	0.99	0.98	0.98	0.97	(0.89-1.00)	Excellent	0.83	0.80	0.75	0.59	(0.40-0.78)	Moderate
Loss of energy	0.97	0.96	0.97	0.94	(0.83-1.00)	Excellent	0.70	0.78	0.68	0.45	(0.22-0.65)	Moderate
Diminished libido	0.97	0.96	0.93	0.87	(0.72-0.97)	Excellent	0.74	0.64	0.68	0.39	(0.13-0.66)	Fair
Early morning waking	1.00	1.00	1.00	1.00	(N.C.)	–[4]	0.78	0.21	0.66	0.06	(-0.17-0.38)	Poor
Excessive sleep	0.97	0.86	0.95	0.83	(0.58-1.00)	Excellent	0.84	0.27	0.73	0.07	(-0.16-0.37)	Poor
Delusions of guilt	0.97	0.84	0.95	0.81	(0.57-1.00)	Excellent	0.91	0.22	0.84	0.20	(0.00-0.65)	Fair
Elevated mood	1.00	0.92	0.98	0.92	(0.69-1.00)	Excellent	0.91	0.36	0.84	0.31	(-0.04-0.71)	Fair
Thoughts racing	1.00	1.00	0.98	0.94	(0.82-1.00)	Excellent	0.89	0.43	0.82	0.34	(-0.04-0.70)	Fair
Excessive activity	1.00	1.00	0.98	0.92	(0.53-1.00)	Excellent	0.90	0.53	0.84	0.44	(-0.16-1.00)	Moderate
Increased sociability	1.00	1.00	0.98	0.91	(0.73-1.00)	Excellent	0.92	0.25	0.86	0.20	(-0.06-0.66)	Fair
Non-affective hallucinations in any modality	0.96	0.97	0.90	0.84	(0.71-0.94)	Excellent	0.88	0.89	0.77	0.63	(0.45-0.83)	Good
Third person auditory hallucinations	0.99	0.96	0.95	0.87	(0.69-1.00)	Excellent	0.87	0.53	0.75	0.32	(0.07-0.61)	Fair
Thought insertion	1.00	0.93	0.95	0.78	(0.55-1.00)	Good	0.91	0.36	0.84	0.30	(-0.06-0.79)	Fair
Thought broadcast	0.98	0.95	0.97	0.90	(0.73-1.00)	Excellent	0.93	0.70	0.82	0.45	(0.23-0.75)	Moderate
Thought withdrawal	1.00	0.91	0.97	0.80	(0.55-1.00)	Excellent	0.92	0.25	0.84	0.07	(-0.07-0.31)	Poor
Thought echo	1.00	1.00	1.00	1.00	(N.C.)	–[4]	0.91	0.46	0.82	0.31	(-0.03-0.67)	Fair
Delusions of passivity	0.98	0.96	0.97	0.90	(0.80-1.00)	Excellent	0.87	0.47	0.77	0.31	(0.01-0.65)	Fair
Persecutory delusions	0.97	0.95	0.97	0.94	(0.81-1.00)	Excellent	0.73	0.54	0.57	0.18	(-0.02-0.41)	Poor
Grandiose delusions	1.00	1.00	1.00	1.00	(N.C.)	–[4]	0.97	0.80	0.93	0.66	(0.05-1.00)	Poor
Bizarre delusions	0.98	0.87	0.96	0.86	(0.58-1.00)	Excellent	0.89	0.33	0.82	0.26	(-0.05-0.65)	Fair
Lack of insight	0.94	0.72	0.90	0.67	(0.42-0.91)	Good	0.79	0.48	0.70	0.28	(-0.02-0.58)	Fair
Persecutory delusions and hallucinations	0.87	0.50	0.95	0.87	(0.74-1.00)	Excellent	0.90	0.63	0.77	0.37	(0.13-0.67)	Fair
Lifetime diagnosis of alcohol abuse/dependence	0.96	0.91	0.73	0.58	(0.37-0.74)	Moderate	0.90	0.75	0.86	0.66	(0.41-0.90)	Good
Lifetime diagnosis of cannabis abuse/dependence	0.99	0.95	0.98	0.94	(0.83-1.00)	Excellent	0.94	0.78	0.91	0.72	(0.47-0.98)	Good
Course of the disorder	N.C.	1.00	0.73	0.58	(0.37-0.74)	Moderate	N.C.	1.00	0.48	0.25	(0.04-0.45)	Fair
Bizarre behaviour	0.95	0.25	0.90	0.20	(-0.22-0.61)	Fair	0.96	0.40	0.93	0.38	(-0.15-0.91)	Fair
Blunt affect	0.92	0.43	0.85	0.30	(-0.02-0.65)	Fair	0.90	0.20	0.79	0.02	(-0.09-0.28)	Poor

[1] Based on lifetime ratings, where applicable.

[2] To calculate this index, where necessary the rating of the corresponding items were dichotomized: presence (0), absence (1).

[3] See Landis & Koch (1977).

[4] Kappa reflects skewed data due to a dichotomous response category with almost all of the responses in the one category.

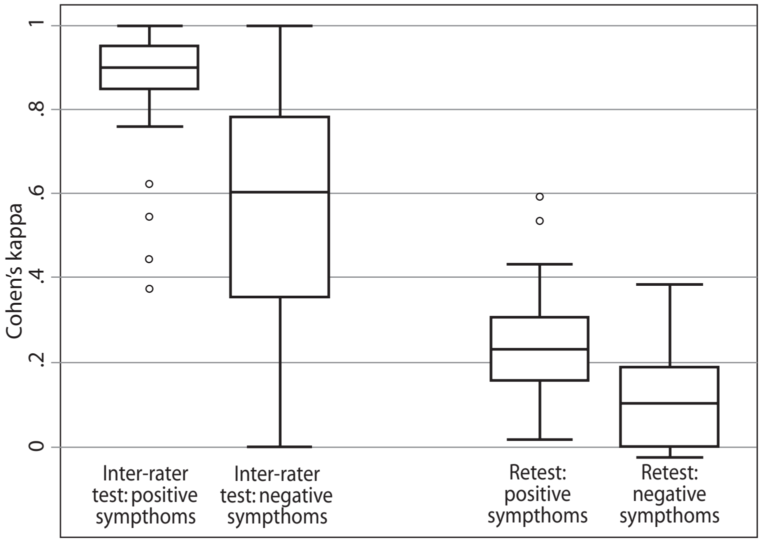

Figure 1 - *Comparison between the median inter-rater and test-retest reliability of the items eliciting positive symptoms of schizophrenia with that of the items eliciting negative symptoms.*

Test-retest reliability

Table I presents the same indexes also for the comparison between the first and the second (*test-retest*) interview. Although the *inter-rater* study found that 71% (64 items) had a kappa score above 80 (indicating excellent agreement), in the *test-retest reliability* analysis only four of the 90 kappa values were above 80. For fourteen items (15%) the *kappa* indicated "good agreement" and for another fourteen "moderate agreement". For the remaining items, the *kappa* values were in the range from -0.02 to 0.39 (i.e. from "no agreement" to "fair agreement"). However, it should be considered that, for the most part of the items, a likely major source of variability in the ratings scales is change in the patients' responses to the same interview questions. As the assessor must make a decision whether a symptom is absent or present and, if present, for what duration (e.g. at least four days, at least one week, or at least two weeks), real changes in the actual duration of symptoms may have occurred, or the patients' recall may have been affected, given the long period of time between the *test* and *retest* interviews. We therefore re-calculated the *kappas* of the *test-retest* after having dichotomized the ratings of all the items into 'present' or 'absent', disregarding the duration estimate. As a result, we found that 20 items achieved a kappa value of ≥0.6, i.e. good to excellent concordance and 21 obtained a kappa between 0.6 and 0.4, indicating moderate agreement. In addition, when analysed for presence/absence of the item using PAR, agreement was high (0.80-1.00) for half of the items and the p_{pos} values were similarly high: 68/90 of them were 0.80 or superior, indicating very high positive agreement. The comparison between the median *test-retest reliability* of the items eliciting positive symptoms of schizophrenia with that of the items eliciting negative symptoms, showed a fair agreement (median *kappa* 0.25, inter-quartile range 0.17) for positive symptoms and a poor agreement (median *kappa* 0.10, inter-quartile range 0.18) for negative symptoms (Figure 1). The difference between the median *kappa* of positive and negative symptoms, however, was not statistically significant (z=1.93, *P*=0.053).

In terms of *kappa*, agreement for ICD-10 diagnoses was fair both at the level of narrow (*kappa*=0.21) and broad diagnostic categories (*kappa*=0.22) but for DSM-IV diagnoses it improved from fair (*kappa*=0.24) to moderate (*kappa*=0.40) when broad disorder categories were used (Table II).

Diagnostic validity

An assessment of the validity of the DIP generated diagnosis was possible by comparing diagnoses for 18 cases that had been assessed using both the DIP interview

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

Table II – *DIP-Italian version inter-raters and test-retest reliability – diagnosis.*

	Inter-rater reliability Narrow categories (No. of cases: 60) (No. of raters: 7)				**Inter-rater reliability Broad categories (No. of cases: 60) (No. of raters: 7)**				**Test-retest reliability Narrow categories (No. of cases: 44) (No. of raters: 7)**				**Test-retest reliability Broad categories (No. of cases: 44) (No. of raters: 7)**			
	Overall pairwise agreement	K	K 95% CI	Level of agreement	Overall pairwise agreement	K	K 95% CI	Level of agreement	Overall pairwise agreement	K	K 95% CI	Level of agreement	Overall pairwise agreement	K	K 95% CI	Level of agreement
ICD-10	0.87	0.84	0.73-0.96	Excellent	0.89	0.85	0.71-0.97	Excellent	0.34	0.21	0.08-0.36	Fair	0.44	0.22	0.01-0.45	Fair
DSM-IV	0.78	0.74	0.59-0.84	Good	0.80	0.74	0.57-0.88	Good	0.36	0.24	0.09-0.41	Fair	0.54	0.40	0.18-0.64	Moderate

and a comprehensive SCAN interview, with the SCAN interview as the 'gold standard'. The level of agreement between SCAN-generated diagnoses and DIP-generated diagnoses was good, with twelve out of the 18 DIP diagnoses (67%) matching the SCAN diagnosis.

Applications of the DIP

The DIP was originally designed for the study of psychoses in the context of the Australian National Survey of Mental Health and Wellbeing, which established point and one-year prevalence rates for psychotic disorders in geographically defined catchment areas across Australia. The findings of the study have been reported in detail elsewhere (Jablensky *et al.*, 1999; 2000).

As an illustration of the 'polydiagnostic' applications of the DIP, Figure 2 shows the diagnostic distribution, in terms of OPCRIT-generated ICD-10 and DSM-IV diagnoses, for the all set of 60 interviews rated by the interviewers of the *inter-rater reliability* test. The overall agreement between the two diagnostic systems (overlapping assignment of cases to the same diagnostic category) was high for schizophrenia and schizoaffective disorder (where the agreement was perfect) but rather low for the other diagnostic categories, which is not surprising since the DIP was specifically designed for the diagnostic assessment of psychotic disorders. Our finding, moreover, are consistent with that of Castle *et al.* (2006) who found a similar difference in the agreement between the DSM-IV and the ICD-10 diagnostic categories, using a much larger, Australian sample.

Another application of the DIP-OPCRIT diagnostic module is the generation of individual and group symptom profile plots. Figure 3 shows such plots for lifetime and present state (including the 4 weeks prior to interview) frequency of symptoms for the all set of 60 interviews rated by the interviewers of the *inter-rater* reliability test. When combined with a 'polydiagnostic' classification, such plots provide a convenient visualization of both the similarities and differences between alternative sets of diagnostic criteria in terms of actual symptomatology.

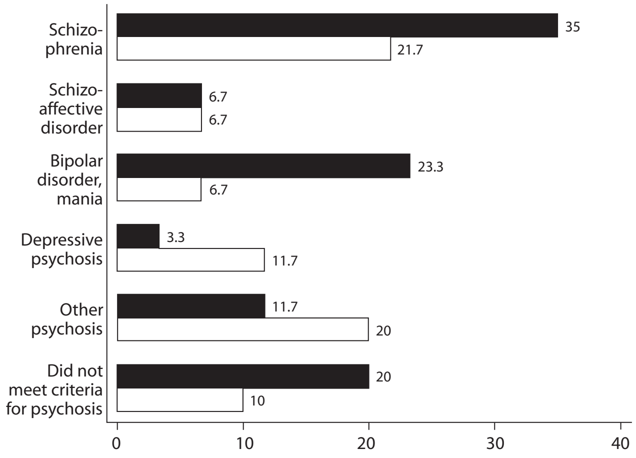

Figure 2 - *Diagnostic distribution of the 60 interview rated by the interviewers of the inter-rater reliability test, according to the DIP diagnostic algorithm, by diagnostic classification system. ICD-10; DSM-IV.*

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

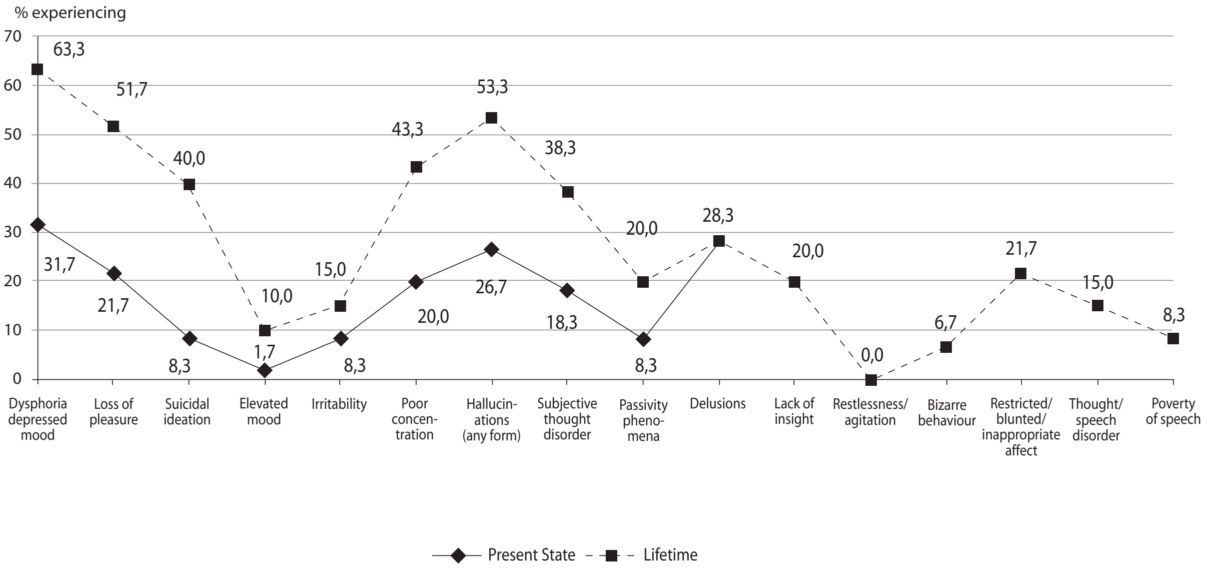

Figure 3 - *Present state and lifetime symptom profiles (selected DIP items) of the 60 interviews rated by the interviewers of the inter-rater reliability test.*

DISCUSSION

Psychiatric diagnostic interviews are widely used in clinical settings and in research studies, and there is a need to rely on sound psychometric properties of these instruments. This is the first study conducted on the Italian version of the DIP. The instrument proved to be relatively brief to administer (37-39 minutes) but sufficiently detailed to allow a fine-tuned analysis of symptoms, as well as the distinction of current and lifetime symptom profiles. Furthermore, it allows the user to establish diagnoses and subtype classification according to several sets of commonly used criteria. This feature is very important given the fact that there is still a lack of consensus about the optimal classification of psychotic disorders (Kendell & Jablensky, 2003; Castle & Jablensky, 2005).

The results of the analyses reported here indicated an inter-rater reliability similar or even superior to that reported by the previous study (Castle *et al*., 2006). The high level of agreement when two independent assessors rated the same DIP interview suggests that the instrument is robust to differences in the clinical background of users. For this reason, the DIP can be employed by interviewers with varying professional background (psychiatrists, clinical psychologists and mental health nurses). It was found that reliability was better for positive than for negative symptoms. This needs to be born in mind when training for symptom recognition as it may be easier to rate positive symptoms such as hallucinations and delusions than to detect more subtle signs and symptoms associated with the negative syndrome.

The validity data presented here are based on a relatively small samples but do give an indication of the robustness of the diagnostic module of the DIP. It has potential applications as a diagnostic tool for psychiatric hospital admissions and outpatient clinic evaluations. In the emerging competitive healthcare environment, structured diagnostic instruments such as the DIP can be used by providers (hospitals, outpatient care clinics) and government agencies to negotiate mental health contracts. Databases can be generated from the results of the DIP to assist hospital in calculating costs and negotiating payments. The Italian version of the DIP should be useful both in the clinical assessment of patients in Italy and in the assessment of Italian-speaking patients living in other countries (including immigrants with poor local language skills). In the research field, use of the DIP will provide a standardised diagnostic assessment of Italian patient samples included in international multicentre clinical trials, epidemiological or genetic studies.

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

These conclusions must be tempered by the relatively low kappa values found when two different interviewers rated the same respondent over time. The main reasons for such discrepancies was the long period of time between the two interviews (median: 89 days), during which the actual duration of symptoms, or the patient's recall of it, may have changed. One way of avoiding such discrepancies is to use, when appropriate, a more global judgement whether the symptom was present or absent, rather than attempt fine-tuned estimations of duration. Further research should aim to clarify the reliability of the Italian version of this interview across two separate interviews by using a shorter time interval between the interviews.

Perhaps as important as any of these detailed considerations about reliability is the overall impression the DIP makes on the clinicians using it and the patients to whom it is administered. The three psychiatrists and the four psychologists involved in this study all came to regard the interview as an effective and valuable instrument. Standardization is achieved without making the interview cumbersome and stilted, most of the questions and the probes are well phrased and a reasonable balance is struck between the requirements of a standardised instrument and the clinician's wish to be flexible to alter the order or form of his questions if the occasion demands that. There are, however, a number of ways in which the interview can be further improved. The low test-retest reliability makes it clear that some probes are unnecessary detailed. This is a point which could be corrected in later editions. Arrangements already exist for such updates to be made at regular intervals after consultations with the main users of the interview.

Acknowledgment. We acknowledge Peter McGuffin and Anne Farmer, originators of the initial OPCRIIT checklist and diagnostic algorithm.

APPENDIX I

- DIP Part 1: Demography and social functioning module
- Contains 49 items under the following item headings:
- General information including sociodemographic data
- Children, carer role
- Education
- Accommodation
- Household and participation in household activities
- Socializing ; social withdrawal
- Confiding relationships, intimacy, sex life
- Work, housework, studying
- Finances
- Activities of daily living and self-care
- Interests

DIP Part 2: Diagnostic module
Contains 94 items under the following item headings:

- General items
- Pre-morbid characteristics and onset
- Family history
- Depression
- Mania
- Hallucinations
- Subjective thought disorder
- Delusions
- Insight
- Response to medication
- General ratings on psychotic symptoms
- Substance use: alcohol; non-medical use of drugs; tobacco and caffeine
- Alcohol and drug abuse and dependence
- Duration and course
- Behaviour
- Affect
- Speech

DIP Part 3: Service utilization module
Contains 40 items under the following item headings:

- In-patient treatment
- Care received from emergency/casualty department
- Treatment in the community (out-patient clinic/community mental health clinic)
- Other health professionals seen and services received
- Rehabilitation or day programme
- Health and welfare and voluntary agencies
- Guardianship, carers
- Medication and the perceived benefits
- Impairment due to side-effects of medication
- Self-harm
- Offending behaviour
- Satisfaction with life
- Unmet need for services
- Social and Occupational Functioning Assessment Scale (SOFAS)

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

APPENDIX II

41. Irritable mood (OPCRIT 36)
Irritable mood (SCAN 10.002)

- **I now want to ask whether you have ever felt very irritable or excessively annoyed with others, such that you lost your temper often?**

- Have other people commented on that or said you were much too impatient?
- How long did you feel like that?

0 = Not present
1 = Present for at least four days
2 = Present for at least one week
3 = Present for at least two weeks OR if lasted < one week the respondent was hospitalised for affective disorder

Respondent's mood is predominantly irritable
The respondent describes his/her mood as easily aroused aggressiveness
It is a:
- Pervasive mood of anger; or
- Impatience; or
- Over-readiness to respond to minor annoyances; or
- Being on a short fuse,

Which lasts for at least one week. The person may recognise the response as excessive, out of proportion to the circumstance and difficult to control. It may also be an unpleasant experience.

Do not rate:
An occasional period of irritability or loss of one's temper.

41. Irritable mood

□ PS □ PY □ LT

Present state, past year and lifetime coding

If NO to both Elevated mood (Item 40) and Irritable mood (Item 41):

Skip to Item 49: Hallucinations

APPENDIX III

ICD-10 OPCRIT diagnostic subtypes generated by the DIP software:

- mild depression disorder,
- moderate depression disorder,
- moderate depression disorder with somatic syndrome,
- severe depression disorder,
- severe depression with psychotic symptoms syndrome,
- hypomania,
- mania, mania with psychosis,
- bipolar affective disorder,
- schizophrenia,
- schizoaffective disorder manic type,
- schizoaffective disorder depressed type,
- schizoaffective disorder bipolar type,
- delusional disorder,
- other non-organic psychotic disorders.

DSM-IV OPCRIT diagnostic subtypes generated by the DIP software:

- major depressive disorder,
- major depressive disorder moderate,
- major depressive disorder severe,
- major depressive disorder with psychosis,
- hippomanic episode, manic episode,
- manic episode with psychosis,
- schizophrenia,
- schizophreniform disorder,
- schizophreniform disorder depressed type,

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press

- schizophreniform disorder bipolar type,
- delusional disorder,psychotic disorder not otherwise specified (atypical psychosis),
- bipolar I disorder,
- bipolar II disorder.

REFERENCES

American Psychiatric Association (1980). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed). American Psychiatric Association: Washington, DC.

American Psychiatric Association (1987). *Diagnostic and Statistical Manual of Mental Disorders* (3rd ed., revised). American Psychiatric Association: Washington, DC.

American Psychiatric Association (1994). *Diagnostic and Statistical Manual of Mental Disorders* (4rd ed.). American Psychiatric Association: Washington, DC.

Balestrieri M., Baldacci S., Bellomo A., Bellantuono C., Conti L., Perugi G., Nardini M., Borbotti M. & Viegi G. (2007) Clinical vs. structured interview on anxiety and affective disorders by primary care physicians. understanding diagnostic discordance. *Epidemiologia e Psichiatria Sociale* 16, 144-151.

Brugha T.S., Nienhuis F., Bagchi D., Smith J. & Meltzer H. (1999). The survey form of SCAN: the feasibility of using experienced lay survey interviewers to administer a semi-structured systematic clinical assessment of psychotic and non-psychotic disorders. *Psychological Medicine* 29, 703-711.

Carpenter W.T., Strauss J.S. & Bartko J.J. (1973). Flexible system for the diagnosis of schizophrenia: a report from the WHO Pilot Study of Schizophrenia. *Science* 182, 1275-1278.

Castle D.J. & Jablensky A. (2005). Diagnosis and classification in psychiatry. In *Core Psychiatry* (2nd ed) (ed. P.Wright, M. Phelanand, J. Stern), pp. 507-515. W.B. Saunders: London.

Castle D.J., Jablensky A., McGrath J.J., Carr V., Morgan V., Waterreus A., Valuri G., Stain H., McGuffin P. & Farmer A. (2006). The diagnostic interview for psychoses (DIP): development, reliability and applications. *Psychological Medicine* 36, 69-80.

Cicchetti D. & Feinstein A. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology* 6, 551-558.

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

Crow T.J. (1980). Molecular pathology of schizophrenia: more than one disease process? *British Medical Journal* 280, 66-68.

Farmer A.E., McGuffin P. & Spitznagel E.L. (1983). Heterogeneity in schizophrenia: a cluster-analytic approach. *Psychiatry Research* 8, 1-12.

Feighner J. P., Robins E., Guze, S. B., Woodruff R. & Munoz R. (1972). Diagnostic criteria for use in psychiatric research. *Archives of General Psychiatry* 26, 57-61.

Feinstein A. & Cicchetti D. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology* 6, 543-549.

First M., Spitzer R., Gibbon M. & Williams J. (1997). *Structured Clinical Interview for DSM-IV Axis I Disorders (SCID)*. American Psychiatric Press: Washington DC.

Fyer A., Mannuzza S., Martin L., Gallops M.S., Endicott J., Schleyer B., Gorman J.M., Liebowitz M.R. & Klein D.F. (1989). Reliability of anxiety assessment. II. Symptom agreement. *Archives of General Psychiatry* 46, 1102-1110.

Jablensky A., McGrath, J., Herrman, H., Castle, D., Gureje, O., Morgan, V. & Korten, A. (1999). People living with psychotic illness : An Australian study 1997-98. In *National Survey of Mental Health and Wellbeing Report 4*. Commonwealth Department of Health and Aged Care: Canberra.

Jablensky A., McGrath J., Herrman H., Castle D., Gureje O., Evans M., Carr V., Morgan V., Korten, A. & Harvey C. (2000). Psychotic disorders in urban areas: an overview of the Study on Low Prevalence Disorders. *Australian and New Zealand Journal of Psychiatry* 34, 221-236.

Kendell R. & Jablensky A. (2003). Distinguishing between the validity and utility of psychiatric diagnoses. *American Journal of Psychiatry* 160, 4-12.

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

Mannuzza S., Fyer A., Martin L., Gallops M.S., Endicott J., Gorman J., Liebowitz M.R. & Klein D.F. (1989). Reliability of anxiety assessment. I. Diagnostic agreement. *Archives of General Psychiatry* 46, 1093-1101.

McGuffin P., Farmer A. & Harvey I. (1991). A polydiagnostic application of operational criteria in studies of psychotic illness. Development and reliability of the OPCRIT system. *Archives of General Psychiatry* 48, 764-770.

Prince M. (2008) Measurement validity in cross-cultural comparative research. *Epidemiologia e Psichiatria Sociale* 17, 211-220.

Pull M.C., Pull C.B. & Pichot P. (1987). Empirical French criteria for psychoses. II. Consensus of French psychiatrists and provisional definitions. *Encephale* 13, 53-57.

Sartorius N., Kaelber C., Cooper J., Roper M.T., Rae D.S., Gulbinat W., Ustün T.B. & Regier D.A. (1993). Progress toward achieving a common language in psychiatry. Results from the field trial of the clinical guidelines accompanying the WHO classification of mental and behavioral disorders in ICD-10. *Archives of General Psychiatry* 50, 115-124.

Spitzer R. & Williams J. (1988). Revised diagnostic criteria and a new structured interview for diagnosing anxiety disorders. *Journal of Psychiatric Research* 22, Suppl. 2, 55-85.

Spitzer R. L., Endicott J. & Robins E. (1978). Research diagnostic criteria: rationale and reliability. *Archives of General Psychiatry* 35, 773-782.

Spitzer R.L., Forman J. & Nee J. (1979). DSM-III field trials: I. Initial interrater diagnostic reliability. *American Journal of Psychiatry* 136, 815-817.

Spitzer R., Williams J., Gibbon M. & First M.B. (1992). The structured clinical interview for DSM-III-R, I. History, rationale and description. *Archives of General Psychiatry* 49, 624-629.

StataCorp. (2007). *Stata Statistical Software: Release 10*. StataCorp LP: College Station, TX.

Tansella M., Amaddeo F. & Burti L. (1998). Community based mental care in Verona, Italy. In *Mental Health in Our Future Cities* (ed. D. Goldberg and G. Thornicroft), pp. 239-262. Psychological Press: Hove.

Taylor M.A. & Abrams R. (1978). The prevalence of schizophrenia: a reassessment using modern diagnostic criteria. *American Journal of Psychiatry* 135, 945-948.

Thompson W.D. & Walter S.D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology* 41, 949-958.

Tsuang M.T. & Winokur G. (1974). Criteria for subtyping schizophrenia. Clinical differentiation of hebephrenic and paranoid schizophrenia. *Archives of General Psychiatry* 31, 43-47.

Williams J., Farmer A.E., Ackenheil M., Kaufmann C.A., McGuffin P. & the OPCRIT Reliability Research Group (1996). A multicentre inter-rater reliability study using the OPCRIT computerized diagnostic system. *Psychological Medicine* 26, 775-783.

Wing J.K., Babor T., Brugha T., Burke J., Cooper J. E., Giel R., Jablensky A., Regier D. & Sartorius N. (1990). SCAN: Schedules for Clinical Assessment in Neuropsychiatry. *Archives of General Psychiatry* 47, 589-593.

Wing J.K., Sartorius N. & Üstün T. B. (Eds.) (1998) Diagnosis and Clinical Measurement in Psychiatry. A *Reference Manual for SCAN/PSE-10*. Cambridge University Press: Cambridge.

World Health Organization (1990). *The Composite International Diagnostic Interview (CIDI). Authorised Core Version 1.0*. World Health Organization: Geneva.

World Health Organization (1992) *Schedules for Clinical Assessment in Neuropsychiatry* (SCAN). World Health Organization: Geneva.

World Health Organization (1993). *The ICD-10 Classification of Mental and Behavioural Disorders. Diagnostic Criteria for Research*. World Health Organization: Geneva.

https://doi.org/10.1017/S1121189X00001585 Published online by Cambridge University Press