Interrater Reliability of Thai Version of the Structured Clinical Interview for DSM-IV Axis II Personality Disorders (T-SCID II)

TinakonWongpakaran MD*, Nahathai Wongpakaran MD*,
Putipong Bookkamana BS, MSc**, Vudhichai Boonyanaruthee MD*,
Manee Pinyopornpanish MD*, Surinporn Likhitsathian MD*,
Sirijit Suttajit MD, MSc*, Usaree Srisutadsanavong MD*

** Department of Psychiatry, Faculty of Medicine, Chiang Mai University, Chiang Mai, Thailand*
*** Department of Statistics, Faculty of Sciences, Chiang Mai University, Chiang Mai, Thailand*

Objective: *Explore the interrater reliability of Thai version of Structured Clinical Interview for DSM-IV Personality Disorders (T-SCID-II).*

Material and Method: *Fifty-four psychiatric patients were assessed by seven psychiatrists. Each patient was interviewed independently and separately by two psychiatrists using T-SCID-II, with the second interview held within one to six weeks of the first.*

Results: *The Kappa value between the first and second raters with regard to the diagnosis of each personality disorder, ranged from 0.70 for Depressive Personality Disorder, to 0.90 for Obsessive-compulsive Personality Disorder, with a mean of 0.81 for all the personality disorders. The mean trait intraclass correlation coefficient score was 0.90 and the summed score was 0.83. The overall interrater reliability was shown to be good across all the studies.*

Conclusion: *Overall, Thai version of Structured Clinical Interview for DSM-IV Personality Disorders (T-SCID-II) showed between good and excellent reliability. Limitation of the present study and its generalizability was discussed.*

Keywords: *Personality, Disorder, Thai; SCID-II; Reliability*

J Med Assoc Thai 2012; 95 (2): 264-9
Full text. e-Journal: *http://www.jmat.mat.or.th*

Although a new criterion of personality disorders (PDs) is being developed and reformulated for the DSM-V in the near future[1], the diagnosis of PDs is still dependent on a clinically-rated method where an interrater agreement comes into play. The reliability of a given diagnosis thus remains subject to a variety of factors, including rater variance, patient variances, item contexts in the elicitation of targeted criterion signs and symptoms (measurement error), and the type of design (nested or crossed rating). Understanding and accounting for patient variance presents a complex and ongoing challenge. The Structured Clinical Interview for DSM-IV Personality Disorders or SCID-II[2] technique is considered a standardized, semi-structured measurement for diagnosing PDs is based on DSM-IV. In previous studies, it has demonstrated agreement measures from adequate to excellent. A recent study[3] found that most reliability values of the SCID II were excellent, lower than the values obtained by Maffei et al[4], but higher than those of Weertman et al[5]. These studies based reliability values on Cohen's kappa and the intraclass correlation coefficient (ICC). However, any comparison of results based on such different designs is, as is the case here, problematic. A standardized semi-structured interview instrument for use in clinical practice and for the research of personality disorders has yet to be adopted in Thailand. Therefore, the authors employed the Thai version of SCID-II to test it against a Thai clinical sample by using the created variance method proposed by Weertman et al[5].

Material and Method

The present study was approved by an independent ethics committee at the Faculty of Medicine, Chiang Mai University.

Correspondence to:
Wongpakaran T, Department of Psychiatry, Faculty of Medicine, Chiang Mai University, Chiang Mai 50200, Thailand.
Phone: 053-945-422, Fax: 053-945-426
E-mail: tchanob@med.cmu.ac.th

Instrument

The Structure Clinical Interview for DSM-IV Axis II Personality Disorders (SCID-II)[2] has semi-structured format covering 10 standard DSM-IV personality disorders, as well as, Personality Disorder Not Otherwise Specified, Depressive PD and Passive-Aggressive PD. As for Thai version of Structured Clinical Interview for DSM-IV Personality Disorders (T-SCID-II) was developed based on the translation and cultural adaptation method. The principal investigator translated the English version of the SCID-II into Thai and then a bilingual English professor, who had not been exposed to the instrument, wrote a backward translation in English. Finally, a consensual version was made by comparing the original and the backward translated versions. Furthermore, the authors' backward translation version was checked with the original and approved by the SCID II author (Michael B. First), before the final draft of Thai version was carried out.

Raters

The raters consisted of four fixed pairs of psychiatrists who had had at least five years' experience in clinical practice (ranging from 5 to 20 years). The fixed pairing was based on the rater's years of experiences; only one rater was paired with two independent raters.

Participants

Fifty-four psychiatric patients, who were not in a condition to disturb the interviewing process or diminish the cognitive functions of the respondents, were recruited from in-patient and out-patient units of Maharaj Nakorn Chiang Mai Hospital, a university hospital. These patients met the inclusion criteria-they were at least 18 years old, able to read, write and communicate with the investigators and were available for both the one-hour interviews. They were excluded if their IQ score was tested as below 76 or if they had one or more of the following conditions: organic mental disorder, schizophrenia, delusional disorder, schizoaffective disorder or other Axis I psychotic disorders; bipolar disorder (active mania), major depressive disorder (non-remitting), intoxication or a withdrawal state in terms of a substance, or if they were unable to complete the data collection process.

Procedure

Participants were recruited proportional to the number of the patients who attended the psychiatry, *i.e.* 10 inpatients and 44 outpatients. Randomization was applied by using random numbers in the selection of participants for both groups. After signing an informed consent form, the participants were given an IQ test. Then, the Thai version of the MINI[8,9] was administered to diagnose for Axis I disorders. Participants who met the eligible criteria were then randomly assigned to go through two interviews with the SCID II. The interviews were given by two psychiatrists and both interviews occurred within a period of one to six weeks of each other. Neither the second interviewer nor the patient was aware of the results of the first interview. All interview questions were applied, without using a personality-screening questionnaire, in order to avoid unnaturally exaggerating the reliability[10].

Data analysis

Cohen's Kappa was used to analyze whether there was agreement of diagnoses between the first and the second raters and as to whether PD was present or absent. The result for each tested patient was rated as: (1) absent or false, (2) sub-threshold, or (3) threshold or true. Kappa was calculated when at least 10% of the cases were diagnosed with any PDs or traits by both raters. Since, as suggested by Zimmerman[11], interrater reliability is affected by the illness base rate, kappa values were calculated only if at least 5 participants were diagnosed as having any particular PD. The intraclass correlation coefficient (ICC) was calculated to find dimensional PDs where there was an agreement between the trait score and the sum score. A p-value of less than 0.05 was set for statistical significance.

Results

Among the 54 participants (28 males, 26 females), the mean age was 39.37 ± 13.10 years and the age range 18 to 67. The three most common Axis I diagnoses were substance related disorders (27.8%), anxiety disorders (24.1%) and major depressive disorders (20.4%). There was no gender difference in Axis I and Axis II diagnoses. However, the number of Axis I diagnoses was significantly correlated with the number of PDs ($r = 0.54$, $p < 0.0001$).

Obsessive-compulsive PD and avoidant PD were the first and second most common diagnoses, whilst PDs detected in less than five cases by the raters were not calculated for Kappa. The Kappa value between the first and second raters with regard to the diagnosis of each PD ranged from 0.70 for Depressive

PD to 0.90 for Obsessive-Compulsive PD. The mean Kappa of all PDs was 0.81, indicating excellent interrater reliability. The ICC value ranged from 0.77 for histrionic and Schizoid PDs to 0.92 for Obsessive-Compulsive PD by trait score, and from 0.67 for Histrionic PD to 0.90 for Borderline and Antisocial PDs using the sum score (Table 1). The mean ICC for the trait score was 0.90 and was 0.83 for the sum score. Table 2 compares Kappa and ICC values between the present study and previous studies using DSM-III-R and DSM-IV Personality Disorders. Overall, interrater reliability is shown to be good to excellent across studies.

The mean duration of interviews was 60.7 ± 21.6 minutes (18-120). The mean duration of the first and second interviews was not significantly different nor was there significant time differences based on the interviewers' years of experience. No relationship was found between IQ score and the number of PDs, ($p = 0.94$).

Discussion

The overall reliability of the Thai version of the SCID II is promising, as the Kappa value ranged from 0.70 for Depressive PD to 0.90 for Obsessive-compulsive PD. According to Fleiss, Landis and Koch[12], a kappa coefficient in the range of 0.75 and higher is evidence of excellent agreement, while a coefficient between 0.40 and 0.75 characterizes fair to good agreement. Moreover, Avoidant PD, Obsessive-Compulsive PD, Borderline PD and Antisocial PD are considered to reflect good dependability ($K > 0.80$)[13]. The present results, when compared to a previous study by Weertman et al[5], generally yielded higher ICC coefficients. The possible explanation why the correlation between raters was rather high in the present study may be attributed to the fact that there was homogeneity among the raters' backgrounds-all the raters are psychiatrists familiar with using the DSM system to interview patients for Axis II-with or without a structured interview.

The PD that tends to be less agreement both among psychiatrists in the present study is Dependent PD (even though it is not calculated for Kappa value because the value of the first rating is less than five). Such a low level of agreement with regard to Dependent PD was also found by Osone and Takahashi[7], Dreesen et al[10] and Weertman[5] (Table 2). It is important to note that all studies as well as the present one were from non-English speaking countries. However, it is difficult to conclude if this result can be attributed to the translation method used or culture difference factors. In addition, it can be noted that differences were found between Weertman's study[5] and the present study with regard to Obsessive-Compulsive PD and Borderline PD, whereby the former study found only a fair-to-good level of agreement while the present study found excellent agreement. Another interesting point

Table 1. Kappa values and ICC

Personality disorders	No. of diagnoses (%)		% observed agreement			Trait score		Sum score	
	1st rater	2nd rater		Kappa[+]	95% CI	ICC	95% CI	ICC	95% CI
Avoidant	11 (24.49)	11 (20.4)	96	0.89	0.81-0.93	0.91	0.84-0.94	0.84	0.72-0.91
Dependent	3 (5.6)	5 (9.3)	100	-		0.87	0.78-0.92	0.88	0.79-0.93
Obsessive-compulsive	14 (25.9)	14 (25.9)	96	0.90	0.84-0.94	0.92	0.83-0.94	0.87	0.78-0.93
Passive-aggressive	8 (14.8)	7 (13.0)	94	0.77	0.63-0.86	0.88	0.79-0.93	0.86	0.76-0.92
Depressive	5 (9.3)	6 (11.1)	94	0.70	0.53-0.81	0.87	0.77-0.92	0.79	0.66-0.87
Paranoid	9 (16.7)	6 (11.1)	94	0.78	0.65-0.87	0.86	0.75-0.92	0.82	0.68-0.90
Schizotypal	2 (3.7)	2 (3.7)	100	-		0.80	0.65-0.88	0.84	0.71-0.90
Schizoid	7 (13)	8 (14.8)	94	0.77	0.63-0.86	0.77	0.61-0.87	0.78	0.64-0.88
Histrionic	1 (1.9)	2 (3.7)	98	-		0.77	0.60-0.87	0.67	0.42-0.81
Narcissistic	1 (1.9)	1 (1.9)	100	-		0.75	0.56-0.85	0.84	0.71-0.91
Borderline	7 (13.0)	7 (13.0)	96	0.84	0.73-0.90	0.90	0.82-0.94	0.90	0.83-0.94
Antisocial	5 (9.3)	5 (9.3)	96	0.78	0.65-0.87	0.87	0.78-0.93	0.90	0.83-0.94

[+] Kappa coefficients are only presented for those diagnostic categories where this value is = 5

Table 2. A comparison of characteristics, Kappa values and ICC values among studies

	*Dreesen et al.[10]	Weertman et al.[5]	The present study	Lobbestael et al.
Design	Nested rating	Nested rating	Nested rating	Join-reliability+
Interval for retesting	1-6 week	1-6 week	1-6 week	n/a
Raters		10 raters 10 CBT therapists (or in training)	4 pairs, fixed 7 psychiatrists	16 first, and 14 second raters PhD-level psychologists (or in training)
Avoidant	0.80	0.82	0.91	0.89
Dependent	0.49	0.20	0.87	0.90
Obsessive-compulsive	0.75	0.63	0.92	0.87
Passive-aggressive	0.62	-	0.88	0.85
Depressive	++	0.71	0.87	0.94
Paranoid	0.66	-	0.86	0.85
Schizotypal	0.59	-	0.80	0.62
Schizoid	-	-	0.77	0.76
Histrionic	0.24	-	0.77	0.75
Narcissistic	-	-	0.75	0.67
Borderline	0.72	0.70	0.90	0.93
Antisocial	0.75	0.88	0.87	0.78

Kappa coefficients are only presented for those diagnostic categories where this value is = 5
* Using SCIDII- DSM-II-R
+ The 2ndrater used video tapes of the first rater as data
++ Does not exist in DSM-IV

is that the PDs whose assessment requires behavioral observation, for example, Histrionic PD, Schizoid PD, Schizotypal PD and Narcissistic PD, were found to have rather low frequency rates and low observed agreement. Not surprisingly, it appears that behavioral observation causes more variances in raters' interpretations than the simple use of restricted questions, especially so when the interview is conducted in an independent manner and by different raters, as in the present study. The present study's results showed only a fair level of agreement for those PDs in which criterion data was obtained by behavioral observation and that Histrionic PD yields the lowest agreement (ICC = 0.67), a finding supported by Dreesen et al (ICC = 0.24)[10] and Lobbestael et al[3]. It is hoped that the new criteria of the Axis II personality diagnoses (DSM-V) now being developed, will incorporate measures to address these problems.

The mean duration of interview in the present study was approximately one hour, as compared to 1.5 to 2.5 hours in a prior study by Arntz et al[14] and there was no significant difference found in time spent interviewing between the first and the second group of raters (p = 0.902). In addition, the interviewers' length of experience was not associated with the number of PDs elicited by each rater, as the differences were not found to be significant within the group of psychiatrist raters, but rather between psychiatrist and non-psychiatrist raters.

Conclusion

In summary, the T-SCID-II showed from good to excellent interrater reliability and the level of reliability was comparable to that produced in another recent study by Lobbestael et al[3], in which the video of a first group of raters was used for assessment purposes by a second group. However, the T-SCID-II has only been employed by psychiatrists, not by other mental health professionals; therefore, future studies involving other mental health personnel should be carried out to prove its general applicability. In addition, a practical method of using screening tool of the SCID II personality questionnaire before conducting a full interview should be further investigated in order to find its predictive value.

Acknowledgement

The authors wish to thank the Faculty of Medicine, Chiang Mai University, for granting funds for this study. TW and NW translated SCID-II into the

Thai language, conceived the study, designed and developed the proposal, interviewed participants, and wrote the manuscript. PB assisted with proposal writing and the plan for statistical analysis. All except PB participated in the interviews using the T-SCID-II. VB assisted in writing the manuscript. TW and PB performed the statistical analysis. All authors read and approved the final draft of the manuscript.

Potential conflicts of interest

None.

References

1. American Psychiatric Association DSM-5 Development [Internet]. 2010 [cited 2010 Mar 15]. Available from: http://www.dsm5.org/Pages/Default.aspx.
2. First MB, Gibbon M, Spitzer RL, Williams JBW, Benjamin LS. Structured clinical interview for DSM-IV axis II personality disorder (SCID-II). Washington, DC: American Psychiatric Press; 1997.
3. Lobbestael J, Leurgans M, Arntz A. Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I Disorders (SCID I) and Axis II Disorders (SCID II). Clin Psychol Psychother 2011; 18: 75-9.
4. Maffei C, Fossati A, Agostoni I, Barraco A, Bagnato M, Deborah D, et al. Interrater reliability and internal consistency of the structured clinical interview for DSM-IV axis II personality disorders (SCID-II), version 2.0. J Pers Disord 1997; 11: 279-84.
5. Weertman A, Arntz A, Dreessen L, van Velzen C, Vertommen S. Short-interval test-retest interrater reliability of the Dutch version of the Structured Clinical Interview for DSM-IV personality disorders (SCID-II). J Pers Disord 2003; 17: 562-7.
6. Farmer RF, Chapman AL. Evaluation of DSM-IV personality disorder criteria as assessed by the structured clinical interview for DSM-IV personality disorders. Compr Psychiatry 2002; 43: 285-300.
7. Osone A, Takahashi S. Twelve month test-retest reliability of a Japanese version of the Structured Clinical Interview for DSM-IV Personality Disorders. Psychiatry Clin Neurosci 2003; 57: 532-8.
8. Kittirattanapaiboon P, Khamwongpin M. The Validity of the Mini International Neuropsychiatric Interview (M.I.N.I.)-Thai Version. J Ment Health Thai 2005; 13: 126-36.
9. Sheehan DV, Lecrubier Y, Sheehan KH, Amorim P, Janavs J, Weiller E, et al. The Mini-International Neuropsychiatric Interview (M.I.N.I.): the development and validation of a structured diagnostic psychiatric interview for DSM-IV and ICD-10. J Clin Psychiatry 1998; 59 (Suppl 20): 22-33.
10. Dreessen L, Arntz A. Short-interval test-retest interrater reliability of the Structured Clinical Interview for DSM-III-R personality disorders (SCID-II) in outpatients. J Pers Disord 1998; 12: 138-48.
11. Zimmerman M. Diagnosing personality disorders. A review of issues and research methods. Arch Gen Psychiatry 1994; 51: 225-45.
12. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977; 33: 159-74.
13. Hoyt WT, Melby JN. Dependability of measurement in counseling psychology: an introduction to generalizability theory. Couns Psychol 1999; 27: 325-52.
14. Arntz A, van Beijsterveldt B, Hoekstra R, Hofman A, Eussen M, Sallaerts S. The interrater reliability of a Dutch version of the Structured Clinical Interview for DSM-III-R Personality Disorders. Acta Psychiatr Scand 1992; 85: 394-400.

ความน่าเชื่อถือระหว่างผู้ประเมินค่า (interrater reliability) เครื่องมือแบบสัมภาษณ์ตามรูปแบบที่กำหนดบุคลิกภาพแปรปรวนตาม DSM-IV ฉบับภาษาไทย (T-SCID-II)

ทินกร วงศ์ปการันย์, ณหทัย วงศ์ปการันย์, พุฒิพงษ์ พุกกะมาน, วุฒิชัย บุณยนฤธี, มณี ภิญโญพรพาณิชย์, สุรินทร์พร ลิขิตเสถียร, ศิริจิต สุทธจิตต์, อุสรี ศรีสุทัศนวงษ์

***วัตถุประสงค์**: การศึกษาครั้งนี้เป็นการหาความน่าเชื่อถือระหว่างผู้ประเมินค่า (interrater relibility) เครื่องมือแบบสัมภาษณ์ตามรูปแบบที่กำหนดบุคลิกภาพแปรปรวนตาม DSM-IV ฉบับภาษาไทย*

***วัสดุและวิธีการ**: ผู้ป่วยจิตเวชจำนวน 54 ราย ได้รับการสัมภาษณ์โดยจิตแพทย์ 7 คน ซึ่งจับคู่สัมภาษณ์ผู้ป่วยแต่ละคนด้วยใช้แบบสัมภาษณ์ฉบับภาษาไทยโดยระยะเวลาในการสัมภาษณ์ระหว่างจิตแพทย์คนที่ 1 และคนที่ 2 ห่างกันประมาณ 1 ถึง 6 สัปดาห์*

***ผลการศึกษา**: ค่าความสอดคล้องในการวินิจฉัย (Kappa) ของผู้ประเมินค่าทั้งสองคนเฉลี่ยอยู่ที่ 0.81 ซึ่งถือว่าอยู่ในเกณฑ์ดีมาก โดยมีค่าความสอดคล้องในการวินิจฉัยตั้งแต่ 0.70 ในการวินิจฉัยบุคลิกภาพแปรปรวนแบบซึมเศร้า (Depressive personality disorder) ถึงค่า 0.90 ในการวินิจฉัยบุคลิกภาพแปรปรวนแบบย้ำคิดย้ำทำ (Obsessive-Compulsive personality disorder) ส่วนค่าสัมประสิทธิ์ความสัมพันธ์ระหว่างคะแนนคุณสมบัติของบุคลิกภาพแปรปรวนแต่ละชนิด (trait score) เฉลี่ยเท่ากับ 0.90 ในขณะที่ค่าสัมประสิทธิ์ความสัมพันธ์ระหว่างคะแนนรวมบุคลิกภาพแปรปรวนแต่ละชนิด (sum score) เฉลี่ยเท่ากับ 0.83 โดยภาพรวมความน่าเชื่อถืออยู่ในเกณฑ์ดี*

***สรุป**: เครื่องมือแบบสัมภาษณ์ตามรูปแบบที่กำหนดบุคลิกภาพแปรปรวนตาม DSM-IV ฉบับภาษาไทย มีความน่าเชื่อถือได้ในระดับดีถึงดีเยี่ยม ได้มีการอภิปรายถึงข้อจำกัดของการศึกษาและการนำไปใช้ขยายผล*