Short-Interval Test-Retest Interrater Reliability of the Spanish Version of the Structured Clinical Interview for DSM-IV for Depressive Disorder*

Germán Eduardo Rueda-Jaimes[1]
Álvaro Andrés Navarro-Mancilla[2]
Paul Anthony Camacho López[3]
Jorge Augusto Franco López[4]
Mauricio Escobar Sánchez[5]

Abstract

Introduction: Psychometric data concerning structured clinical interviews are affected by age and population characteristics. The present study was designed to investigate the test-retest interrater reliability of a Spanish version of the Structured Clinical Interview for Major Depressive Disorder (MDD) in adolescent students from Colombia. *Methods*: All participants were interviewed by a psychiatrist with the section for current MDD from the Spanish SCID-I clinical version; after three to ten days another psychiatrist performed the same interview. Both adolescent and second interviewer were blind to the first interview results. The statistical significance of kappa was tested by calculation of z-score. Statistical tests were done in STATA 8.0. *Results*: 164 adolescents were interviewed; mean age was 15.3 years (SD 0.96); 125 (76.2%) were female. The test-retest levels of agreement between first rater and second rater for current MDD was Kappa coefficient 0.612 (CI 95% 0.457-0.765). *Conclusions*: The results of this study suggest that SCID-I is reliable not only in adults, as was previously demonstrated, but also in adolescents. New studies are necessary to test the reliability of the different modules of the SCID in adolescents.

Key words: Adolescent, major depressive disorder, mental health, psychological interview.

....

Título: Confiabilidad interevaluador prueba-reprueba en un intervalo corto de la versión en español de la Entrevista Clínica Estructurada para Trastorno Depresivo

Resumen

Introducción: Los datos psicométricos acerca de la entrevista clínica estructurada son modificados por la edad y las características de la población. El objetivo del presente estudio es establecer la confiabilidad interevaluador prueba-reprueba en un intervalo corto de la versión en español de la Entrevista Clínica Estructurada (SCID, por su sigla en inglés) para trastorno depresivo mayor (TDM) en estudiantes adolescentes. *Métodos*: Un psiquiatra entrevistó a todos los participantes con el módulo para TDM de la versión clínica en español de la SCID-I; tres a diez días después otro psiquiatra realizó la misma entrevista. El adolescente y el entrevistador fueron enmascarados respecto a los resultados de la primera entrevista. La significancia estadística de kappa fue analizada mediante el cálculo del resultado de z. Los análisis estadísticos fueron hechos en STATA 8.0. *Resultados*: Se entrevistaron a 164 adolescentes, cuyo promedio de edad era 15,3 años (DS: 0,96); 125 (76,2%) eran mujeres. El nivel de concordancia prueba-reprueba del primero y el segundo entrevistador para TDM actual mostró un coeficiente de kappa de 0,612 (IC 95%: 0,457-0,765). *Conclusiones*: Estos resultados sugieren que la SCID-I es confiable en adolescentes y en adultos, como se ha demostrado previamente. Son necesarios nuevos estudios para probar la confiabilidad de los diferentes módulos de la SCID en adolescentes.

Palabras clave: adolescente, trastorno depresivo mayor, salud mental, entrevista psicológica.

Introduction

The reliability of psychiatric diagnosis has been markedly enhanced through the use of standardized interview schedules (1). The Structured Clinical Interview Disorders (SCID) for DSM is a widely used semi-structured instrument to measure all DSM-IV disorders that has shown a relatively high reliability (2-4). The SCID has become a standard for assessing the major axis I disorders by clinically experienced raters.

Most investigators are interested in particular diagnostic classes and they do not need to assess diagnoses that are irrelevant to their research questions. For that reason, the SCID is organized into modules (5). In our case we are interested in the reliability of the MDD because it has a high prevalence in the general population of Colombia, particularly in adolescents.

The initial field studies with the DSM-III indicated that the reliability for MDD presents kappa coefficient 0.80 (2). Later studies confirm a high interrater reliability of the SCID in versions for various languages with kappa coefficient between 0.72-0.93 (2,3,6-8). When the test-retest interrater reliability was measured the kappa coefficient was 0.61 (7,9). Despite the extensive literature regarding reliability diagnoses of depressive disorder, neither in Spanish nor in Colombia did we find studies. Moreover, Hodges et al suggested that psychometric data concerning semi-structured clinical interviews are affected by age and characteristics of the population (10).

Several factors may have influenced the size of the reliability

Rev. Colomb. Psiquiat., vol. 37 / No. 4 / 2008

593

coefficients found in these studies. First, the majority of the studies had small samples. Second, studies of reliability using patients with significant pathology may achieve higher reliability than studies of patients with less severe pathology. Third, the methodology of the study (joint interview or test-retest) has great influence on the reliability coefficients. In joint interview studies only one source of unreliability is investigated (rater variance in criterion interpretation) whereas there are three sources of error in test-retest studies: rater variance in criterion interpretation, rater variance in the elicitation of information and patient variance across interviews (1,11).

The present study was designed to investigate the test-retest interrater reliability of a Spanish version of the structured clinical interview for depressive disorder in community samples of adolescents.

Methods

Instruments

SCID Axis I. The section for current major depressive disorder from the Spanish SCID-I clinical version was used in our study (12,13). The SCID-I is organized so that DSM-IV criteria may be systematically examined for each disorder. The criteria are embedded directly in the SCID, and the sequence of questions is approximated to the DSM-IV decision trees. The section for depressive disorder begins with two essential criteria

and continues with questions and prompts to determine whether an individual meets the requirements for a sufficient number of additional symptoms to warrant the diagnosis. Since we evaluated adolescents a question regarding irritability was added in case the answer for depressive mood was negative. When it was a depressive or irritable mood the first criteria was accepted as positive, just as the DSM-IV suggests in adolescents (14).

Sample

A sample size was calculated to determine the test-retest interrater reliability with a minimum and maximum kappa of 0.5 and 0.75, keeping in mind a significance level of 5% and a power of 80% with an expected prevalence of current MDD in the sample of 10%. The size of the obtained sample was 166 adolescents for evaluation.

Interviewers

The interviewers were one child psychiatry, 2 general psychiatrists and one third-year psychiatry resident. The interviewers were trained before starting the collection of data. The training program included lectures on DSM-IV, lectures on the use of the SCID, role-playing in SCID interviews and actual SCID interviews with psychiatric patients. Training was considered complete when the main investigator, after observing an interview, certified that the interviewer was sufficiently

prepared to begin to interview subjects independently (13,14).

Procedure

Review Boards from Universidad Autónoma de Bucaramanga approved this study. After being informed about the objectives and the minimum risk to interviewes, parents and adolescents gave their informed consent (15). First, all participants were interviewed by a psychiatrist; after seven to ten days another psychiatrist made the same interview. The adolescent and second interviewer were blind to the first interview results. Adolescents who met criteria for major depressive disorder or other depressive disorder were remitted to medical services.

Statistics

Diagnoses were treated as dichotomous categories, either present or absent. Cohen's kappa was used to calculate the test-retest interrater reliability. The statistical significance of kappa was tested by calculation of z-score (16,17). According to Fleiss, Kappa values greater than or equal to 0.75 represent excellent agreement beyond chance, values bellow 0.4 or so may be taken to represent poor agreement beyond chance, and values between 0.4 and 0.75 may be taken to represent fair to good agreement beyond chance (18). All statistical tests were done in STATA 8.0 (19). Significant difference was accepted when the probability of mistake was inferior to 5% (p<0.05).

Results

We interviewed one hundred and sixty-four adolescents from 13 to 17 years-old; mean age was 15.3 years (SD 0.96); 125 (76.2%) were female; formal schooling ranged from 8 to 11 years; and socioeconomic status was low in 46 (28.4%), middle in 115 (70.9%), and high in 1 (0.62%). The prevalence of current major depressive episode was 20.7% according to the first rater and 18.2% according to the second rater (p=0.806).

The test-retest levels of agreement between first rater and second rater for current major depressive disorder are shown in Table 1.

Discussion

This investigation shows that the level of test-retest interrater reliability of a Spanish version of the SCID for current MDD in adolescents is good. This is consistent with the results of two studies that show good interrater reliability of the SCID using a short interval test-retest method (7,9). However, Kappa

Table 1. The Test-Retest Levels of Agreement between First Rater and Second Rater for Current Major Depressive Episode

First rater	Second rater	
	Depression	No depression
Depression	22	12
No depression	8	122

Simple agreement rates 87.8%.
Kappa coefficient: 0.612 (IC 95%: 0.457-0.765).

Rev. Colomb. Psiquiat., vol. 37 / No. 4 / 2008

595

coefficient is lower than in studies were the joint interview method is used (2,3,6-8). This discrepancy is not surprising since in joint interview studies only one source of unreliability is investigated whereas there are three sources of error in test-retest studies: rater variance in criterion interpretation, rater variance in the elicitation of information and patient variance across interviews (1,11).

We have found that community samples of adolescents interviewed with the SCID by psychiatrists show reliability similar to patients with severe pathology (7,9). The use of a non-clinical population represents a strength of this study, since disorders in community samples are milder and therefore the reliability could be minor (20).

There was a small trend for lower rates of current major depressive disorder reported in the second interview. Fenig et al. suggested that lower reporting of disorder in the second interview may be due to subject confusion and speculation regarding the purpose of the second interview, desire to create a more favorable impression in the second assessment, and even the therapeutic effects of the first interview (21). In our case, that difference was not significant and did not alter the reliability.

This is the first study showing the reliability of the SCID for major depressive episode in adolescents and paves the way for the testing of other modules of the SCID in adolescents or other specific populations.

Taken together, the results of this study suggest that the SCID is reliable not only in adults, as was previously demonstrated, but also in adolescents. New studies are necessary to prove the reliability of the different modules of the SCID in adolescents.

Acknowledgements

References

1. Vacc NA, Juhnke GA. The use of structured clinical interviews for assessment in counseling. J Couns Dev. 1997;75(6):470-80.
2. Spitzer RL, Forman JBW, Nee J. DSM-III field trials: I. Inicial interrater diagnostic reliability. Am J Psychiatry. 1979;136(6):815-17.
3. Skre I, Onstad S, Torgersen S, Kringlen E. High interrater reliability for the Structured Clinical Interview for DSM-III-R Axis I (SCID-I). Acta Psychiatr Scand. 1991;84(2):167-73.
4. 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 Personal Disord. 2003;17(6):562-7.
5. Spitzer RL, Williams JBW, Gibbon M, First MB. The Structured Clinical Interview for DSM-III-R (SCID). Arch Gen Psychiatry. 1992;49(8):624-9.
6. Riskind JH, Beck AT, Berchick RJ, Brown G, Steer RA. Reliability of DSM-III diagnoses for Major Depression and Generalized Anxiety Disorder using the Structured Clinical Interview for DSM-III. Arch Gen Psychiatry. 1987;44(9):817-20.
7. Zanarini MC, Skodol AE, Bender D, Dolan R, Sanoslow C, Schaefer E, et al. The Collaborative Longitudinal

Personality Disorders Study: reliability of axis I and II diagnoses. J Personal Disord. 2000;14(4):291-9.

8. Del-Ben CM, Vilela JA, Crippa JA, Hallak JE, Labate C, Zuardi A. Reliability of the Structured Clinical Interview for DSM-IV Clinical Version translated into Portuguese. Rev Bras Psiquiatr. 2001;23(3):156-9.

9. Williams JBW, Gibbon M, First MB, Spitzer RL, Davies M, Borus J, et al. The Structured Clinical Interview for DSM-III-R (SCID): II. Multi-site test-retest reliability. Arch Gen Psychiatry. 1992;49(8):630-6.

10. Hodges K, Zeman J. Interviewing. In: Handbook of child and adolescent assessment. Needham Heights: T. Ollendick and M. Hersen; 1993. p. 65-81.

11. Paget KD. The structured assessment interview: a psychometric review. J Sch Psychol. 1984;22:415-27.

12. First MB, Spitzer RL, Gibbon M, Williams JBM. Structured Clinical Interview for DSM-IV Axis I Disorders. New York: New York State Psychiatric Institute, Biometrics Research Department; 1996.

13. First MB, Spitzer RL, Gibbon M, Williams JBW. Structured Clinical Interview for DSM-IV Axis I Diagnosis-clinical version [In Spanish]. Barcelona: Masson; 1999.

14. American Psychiatric Association. Diagnostic and statistical manual of mental disorders. 4th ed. Washington: American Psychiatric Association; 1994.

15. Resolution 8430 de 1993, for establishing scientific, technical, and administrative norms for health investigation [in Spanish]. Santafé de Bogotá: Ministerio de Salud de Colombia; 1993.

16. Fleiss JL, Nee JCM, Landis JR. Large sample variance of kappa in the case of different sets of raters. Psychol Bull. 1979;86(5):974-7.

17. Growe WM, Andreassen NC, McDonald-Scott P, Keller MB, Shapiro RW. Reliability studies of Psychiatric diagnosis. Arch Gen Psychiatry. 1981;38(4):408-13.

18. Fleiss JL. The measurement of interrater agreement. In: Statistical methods for rates and proportions. 2nd ed. New York: John Wiley; 1981. p. 212-36.

19. STATA for Windows 8.0. Stata Corporation, College Station, Texas; 2003.

20. Bromet EJ, Dunn LO, Connell MM, Dew M, Schulberg HC. Lon-term reliability of diagnosing lifetime major depression in a community sample. Arch Gen Psychiatry. 1986;43(5):435-40.

21. Fenig S, Levav I, Kohn R, Yelin N. Telephone vs face-to-face interviewing in a community psychiatric survey. Am J Public Health. 1993;83(6):896-8.

Correspondence
Germán Eduardo Rueda-Jaimes
Facultad de Medicina
Universidad Autónoma de Bucaramanga
Campus El Bosque
Calle 157 Nº 19-55 (Cañaveral Parque)
Bucaramanga, Colombia
gredu@unab.edu.co

Rev. Colomb. Psiquiat., vol. 37 / No. 4 / 2008

597