

Inter-Rater Reliability of the Brazilian Version of the Paykel's Interview for Recent Life Events in a sub-sample of a population-based study.

Confiabilidade entre Avaliadores da Versão Brasileira da Entrevista de Eventos de Vida Recentes de Paykel em uma sub-amostra de um estudo de base populacional.

Confiabilidad entre Evaluadores de la Versión Brasileña de la Entrevista de Paykel para Eventos de Vida Recientes en una sub-muestra de un estudio basado en la población.

Tania C. Martinho¹

Gabriel Graça de Oliveira²

Paulo R. Menezes³

The present study was funded by Wellcome Trust (Inglaterra) and FAPESP (Brazil)

Conflicts of interest: None.

ABSTRACT: *Background* The Interview for Recent Life Events (IRLE) developed by Paykel is one of the most widely used instruments for assessing life events and its relation to the onset and maintenance of illness. This study aimed to assess the inter-rater reliability of the Brazilian version of this scale in the setting of a population-based case-control study of first episode psychosis in Brazil, considering independent and severe events, the most frequently analysed in psychiatric research. *Method* Fifteen first episode psychosis cases and 26 controls from the community were assessed simultaneously by two raters using the IRLE. Agreement was measured using percentage of agreement and the weighted kappa (k_w). *Results* There was excellent agreement for occurrence of events ($k_w=0.90$) and month of occurrence ($k_w=0.96$) and moderate agreement for independence ($k_w=0.51$) and objective negative impact ($k_w=0.68$). When considering only independent and severe

¹ Psychiatrist, University of São Paulo Medical School.

² Assistant Professor, Department of Clinical Medicine, University of Brasilia Medical School. University of Brasilia Medical School. Campus Darcy Ribeiro, Asa Norte, Brasília – DF, CEP 70910-900, Brazil. E-mail: gabrieloliveira@unb.br

³ Associate Professor, Department of Preventive Medicine, University of São Paulo Medical School.

events, very low values of weighted kappa, contrasting to high values for percentage of agreement, were found. *Conclusions* The results indicate satisfactory reliability, comparable to those found in international literature. The use of weighted kappa results on a more adequate estimate of reliability, but larger samples are needed to estimate agreement by category of events more precisely.

Key words: inter-rater reliability - Paykel - weighted kappa – life events

RESUMO: *Introdução* A Entrevista para Eventos de Vida Recentes (EEVR) de Paykel é um dos instrumentos mais utilizados para acessar os eventos de vida e suas relações com o início e manutenção da doença. O objetivo deste estudo foi avaliar a confiabilidade entre-avaliadores da versão brasileira desta entrevista no campo de um estudo caso-controle de base populacional de primeiros episódios psicóticos no Brasil, considerando-se eventos independentes e graves, os mais frequentemente analisados em pesquisas em Psiquiatria. *Método* Quinze casos de primeiro episódio psicótico e 26 controles da população geral foram avaliados simultaneamente por dois entrevistadores, usando a EEVR. A concordância foi avaliada através da porcentagem de concordância e do kappa ponderado (k_p). *Resultados* Houve excelente concordância para ocorrência de eventos ($k_w=0.90$) e mês de ocorrência ($k_w=0.96$), e concordância moderada para independência ($k_w=0.51$) e impacto objetivo negativo ($k_w=0.68$). Quando apenas os eventos independentes e graves foram considerados, observou-se valores muito baixos de kappa ponderado, contrastando com altos valores de porcentagem de concordância. *Conclusões* Os resultados indicam níveis satisfatórios de confiabilidade, comparáveis aos encontrados na literatura internacional. O uso do Kappa ponderado resulta em uma estimativa mais adequada da confiabilidade, porém amostras maiores são necessárias para estimar concordância por categoria de eventos.

Palavras-chave: confiabilidade entre avaliadores – Paykel – Kappa ponderado – eventos de vida.

RESUMEN: *Introducción* La Entrevista para Eventos de Vida Recientes (EEVR) de Paykel es una de las herramientas más utilizadas para acceder a los acontecimientos de la vida y sus relaciones con el inicio y el mantenimiento de la enfermedad. El objetivo de este estudio fue evaluar la confiabilidad entre los evaluadores de la versión brasileña de la entrevista en el campo de un estudio de casos y controles de base poblacional de los primeros episodios psicóticos en Brasil, considerando eventos independientes y graves, los más frecuentemente analizados en la investigación en Psiquiatria. *Método* Quince casos de primer episodio de psicosis y 26 controles de la población general se evaluaron simultáneamente por dos entrevistadores, utilizando EEVR. El acuerdo fue evaluado por el porcentaje de acuerdo y kappa ponderado (K_p). *Resultados* Hubo un acuerdo excelente para la ocurrencia de eventos ($KW = 0,90$) y el mes de ocurrencia ($KW = 0,96$) y moderado para la independencia ($kw = 0,51$) y el impacto objetivo negativo ($KW = 0,68$). Cuando sólo se consideraron eventos independientes y graves, hubo muy bajos valores de kappa ponderado, lo que contrasta con los valores de porcentaje de concordancia alta. *Conclusiones* Los resultados indican niveles satisfactorios de confiabilidad, comparables a los encontrados en la literatura. Usar lo Kappa ponderado permite una mejor estimación de la confiabilidad, pero se necesitan muestras más grandes para estimar el cumplimiento por categoría de eventos.

Palabras clave: confiabilidad entre evaluadores - Paykel - Kappa ponderado - los acontecimientos de vida.

INTRODUCTION

Several studies have assessed the relationship between life events and the onset and course of physical and psychiatric illness.⁽¹⁾ The Paykel's Interview for Recent Life Events (IRLE) has been one of the instruments extensively used to investigate the relationship between life events and psychiatric disorders or medical conditions.⁽²⁾ It has been translated into a number of languages and has been the object of many studies of inter-rater reliability.⁽³⁻⁶⁾ Some of these studies used percentages of agreement or Spearman's correlation coefficient in the analysis, which are measures of agreement that do not take the effects of chance into consideration.^(3,5) Other studies^(4,6) used weighted kappa in their analysis, one examining a sample from the general population⁽⁴⁾ and another examining a sample of psychiatric patients,⁽⁶⁾ but they did not analyse agreement for independent and severe events (ISE), the categories of events most relevant in psychiatric research.⁽²⁾ Moreover, the result of a reliability assessment of an instrument of measurement is not dependent only on its own characteristics but also depends on the environment where it is applied, characteristics of the population assessed and the way with which it is employed.⁽⁷⁾

This scale has been translated into Portuguese by Oliveira *et al.*,⁽⁸⁾ and its inter-rater reliability was assessed with a sample of psychiatric patients interviewed in a hospital setting.⁽⁶⁾ Therefore, results from that study might not apply directly to population-based epidemiological investigations, due to differences between samples and settings where the assessments took place. Also, neither that study nor previous ones examined inter-rater agreement for ISE. The "Schizophrenia and other Psychoses at unveiling and long-term outcome" study (*SaoPaulo*) is a large population-based case-control study of first episode psychosis in Brazil, and the IRLE was one of the instruments used to investigate factors involved in the aetiology of psychosis in that study. The aim of the present investigation was to assess the inter-rater reliability of the Brazilian Version of Paykel's IRLE in a sub-sample of participants of the *SaoPaulo* study, taking ISE into account.

METHODS

INSTRUMENT

The Paykel's IRLE is a 64-item semi-structured interview well defined and divided into 10 areas: work, education, finance, health, bereavement, migration, courtship and cohabitation, legal, family and social, and marital. The 64th question is unspecific and exists to record an event unclassifiable in the rest of the interview.⁽²⁾ Once the occurrence of an event is detected, further questions must be asked to ascertain the circumstances of the event. The scale was designed to cover the six months prior to the illness onset. The level of independence of the event from the illness and the negative objective impact, a judgement of the expected stressfulness of the event for an average person, are

also assessed. These are coded in a 5 point rate scale, from almost certainly independent to almost certainly dependent and from severe negative impact to no impact, respectively. Quantification of independence is crucial to reliability, since events that are consequences of the illness should not be considered.⁽²⁾

SUBJECTS

The source of participants for the present study was a large population-based case-control study of first episode psychosis, the *SaoPaulo* study. Detailed procedures for inclusion of cases and controls in the *SaoPaulo* study can be found elsewhere.⁽⁹⁾ Briefly, individuals aged between 18 and 64 years, who had had a first contact with any mental health service, public or private, due to a psychotic episode (DSM IV criteria for either mania, schizophrenia, other psychoses from the spectrum of schizophrenia or psychotic depression), and who lived in a defined geographic area of the city of São Paulo were included. Controls were subjects without psychosis from the general population, matched by neighbourhood, age and sex. Participants of the present study were systematically selected among cases and controls who had been included in the main study. The *SaoPaulo* study was approved by the ethical committees of all institutions involved and by the National Ethics Committee (CONEP), and all participants signed informed consent forms.

PROCEDURES

The inter-rater reliability was assessed in a study with a cross-sectional design. The interviews were alternately conducted by one of the two previously trained interviewers, a psychiatrist and a medical student (MGO and TCM). During the interview, both investigators recorded the information obtained and made their ratings independently. Ratings were compared only at the end of the study. Interviews were applied mostly at subjects' homes.

ANALYSIS

Inter-rater agreement for occurrence of the event, month of occurrence, independence and objective negative impact were assessed using percentages of agreement and weighted kappa,⁽¹⁰⁾ with 95% confidence intervals. Landis and Koch established parameters to the judgement of the degree of agreement, according to which kappa coefficient ranging from 0.00 to 0.20 may be interpreted as poor agreement, from 0.21 to 0.40 as fair agreement, from 0.41 to 0.60 as moderate agreement, from 0.61 to 0.80 as substantial agreement and from 0.81 to 0.99 as almost perfect agreement.⁽¹¹⁾ Kappa coefficient equal to 1.00 means perfect agreement between raters.

RESULTS

Forty one subjects were interviewed, of whom fifteen were cases and 26 were controls.

Their ages ranged from 19 to 53 years, and 25 (61%) were male. A total of 84 events was registered by rater 1 (TCM) and 89 by rater 2 (MGO). Eleven events were recorded by only one of the two raters. The number of events per subject ranged from 0 to 6, 6 subjects had no events and 29 had from one to three events. Percentage agreement on number of events was 97%, with a weighted kappa of 0.90. The most frequent events were rated on family and social and work items (for instance, serious illness or death of a relative or friend, and theft of significant sentimental or material value). Weighted kappas for categories of events ranged from 0.77 to 1.00. Inter-rater agreement for month of occurrence was almost perfect, and substantial for negative objective impact, whereas the weighted kappa for independence showed only moderate agreement (table 1). When considering only ISE, percentages of agreement were very high for every category of event, ranging from 92.7% to 100%, but the weighted kappas were very low or showed even no agreement in 6 out of 10 categories of events. When the frequency of ISE was very low the weighted kappa was also low. The estimates of the weighted kappas were not very precise (Table 1).

TABLE 1: Percentages of agreement and weighted kappas for occurrence of events, month of occurrence, independence and negative impact (n = 41)

Classification	Percentage of agreement	Weighted Kappa	95%CI*
Occurrence of event	97.1%	0.90	0.72-1.00
Month of occurrence	98.3%	0.96	0.82-1.00
Independence	77.1%	0.51	0.36-0.66
Negative impact	91.8%	0.68	0.54-0.83

*95% confidence interval.

DISCUSSION

According to Dunn,⁽⁹⁾ the study of the reliability of an instrument must be carried out on a sample from the same population and environment where it will be utilized routinely. This was the case of the present study. Comparing the present study with previous work by Oliveira *et al.*,⁽⁶⁾ percentages of agreement concerning the specific occurrence of the event, month of occurrence, independence and objective negative impact were quite similar, differing only for objective negative impact (97%, 98%, 77%, 92% vs. 100%, 97%, 72% and 70%, respectively). However, percentage agreement are not adequate to assess reliability because they do not take into account the probability of agreement by chance alone. In fact it is very easy to agree on the non-occurrence of an event, especially if it is rare. From a list of 64 events, the average was only 2 events per person. This led to a sequence of negative answers where the probability of both raters marking non-occurrence of events is very high, without meaning a high level of agreement. Therefore, the Kappa coefficient is much more adequate for assessing reliability between raters than percentage agreement.

Cooke and Oliveira *et al.* also used weighted kappa in their analyses.^(4,6) Two distinct pairs of interviewers took part in Cooke's study, who obtained weighted kappas of 0.82 and 0.62 for independence and of 0.64 and 0.58 for objective negative impact.⁽⁴⁾ Oliveira *et al.* obtained weighted kappas of 0.66 and 0.75 for independence and impact, respectively;⁽⁶⁾ corresponding results in the present study were 0.51 and 0.68, respectively. Considering the precision of our estimates, the agreements obtained for occurrence of events, independence and impact in the present study may be considered as good as those found in previous studies. We tried to analyse if agreement would also be good for each category of events if we considered only independent or almost certainly independent and severe events, but the frequency of positive events became very low and, as a consequence, kappa coefficients were also very low.

The psychiatrist rated a greater number of events than the medical student. This may mean that experienced and trained mental professionals have a more accurate perception of the importance of events, suggesting that the IRLE should be applied by trained mental health professionals or more intensive training would be necessary when lay interviewers are enrolled for carrying out this assessment.

CONCLUSION

Our results indicate that the Brazilian Version of the Paykel's IRLE presented satisfactory reliability, considering both percentages of agreement and weighted kappas, with results similar to those reported in the international literature. The use of weighted kappa results in more adequate estimates of reliability, but low frequency of ISE and the small sample size represented limitations of the study, especially regarding the precision of the estimates of reliability. Since events with specific characteristics are of interest in studies assessing the relationship between life events and psychiatric disorders, further studies with larger sample sizes are needed to produce estimates of agreement for categories of events.

Acknowledgements: TPCM and GGO had fellowships from FAPESP, Brazil (proc. no. 01/13921-0 and 01/13923-2, respectively). PRM is partly funded by the CNPq, Brazil. The main project was funded by the Wellcome Trust. The authors thank all interviewees that accepted participating in this study.

REFERENCES

1. Paykel ES. The evolution of life events research in psychiatry. *J Affect Disord.* 2001; 62: 141-149.
2. Paykel ES. The interview for recent life events. *Psychol Med.* 1997; 27: 301-310.
3. Paykel ES. Methodological aspects of life events research. *J Psychosom Res.* 1983; 27: 341-352.
4. Cooke DJ. The reliability of a brief life event interview. *J Psychosom Res.* 1985; 29: 361-365.
5. Baratta S, Colorio C, Zimmerman-Tansella C. Inter-rater reliability of the Italian version of the Paykel Scale of Stressful Life Events. *J Affect Disord.* 1985; 8: 279-282.
6. Oliveira GG, Fonseca PP, Del Porto JA. Inter-rater reliability of the Brazilian version of the Paykel scale of stressful life events. *Psychopathology.* 2003; 36: 320-323.
7. Dunn G. Design and analysis of reliability studies. In: Hodder, Stoughton editors. *The statistical evaluation of measurements errors.* New York: Oxford University Press; 1989.
8. Oliveira GG, Pinto Fonseca P, Del Porto JA. Versão brasileira da entrevista de Paykel para eventos de vida recentes. *Rev Bras Terap Comport Cogn.* 2002; 4: 47-61.
9. Ayres AM, Busatto G, Menezes PR, Schaufelberger MS, Coutinho L, Murray RM, McGuire PK, Rushe T, Scazufca M. Cognitive deficits in first-episode psychosis: A population-based study in São Paulo, Brazil. *Schizophr Res* 2007; 90: 338-343.
10. Cohen J. Weighted kappa: nominal scale agreement with provision for scale disagreement or partial credit. *Psychol Bull.* 1968; 70: 213-220.
11. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics.* 1977; 33: 159-174.

Artigo apresentado em 24/05/16

Artigo aprovado em 20/06/16

Artigo publicado no sistema em 30/06/16