



ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

ARCHIVIO ISTITUZIONALE DELLA RICERCA

Alma Mater Studiorum Università di Bologna Archivio istituzionale della ricerca

Pre-training inter-rater reliability of clinical instruments in an international psychosis research project

This is the final peer-reviewed author's accepted manuscript (postprint) of the following publication:

Published Version:

Berendsen S., Kapitein P., Schirmbeck F., van Tricht M.J., McGuire P., Morgan C., et al. (2021). Pre-training inter-rater reliability of clinical instruments in an international psychosis research project. *SCHIZOPHRENIA RESEARCH*, 230, 104-107 [10.1016/j.schres.2020.08.001].

Availability:

This version is available at: <https://hdl.handle.net/11585/810362> since: 2021-02-28

Published:

DOI: <http://doi.org/10.1016/j.schres.2020.08.001>

Terms of use:

Some rights reserved. The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>).
When citing, please refer to the published version.

(Article begins on next page)

This is the accepted manuscript of:

Berendsen S, Kapitein P, Schirmbeck F, van Tricht MJ, McGuire P, Morgan C, Gayer-Anderson C, Kempton MJ, Valmaggia L, Quattrone D, di Forti M, van der Gaag M, Kirkbride JB, Jongsma HE, Jones PB, Parellada M, Arango C, Arrojo M, Bernardo M, Sanjuán J, Santos JL, Szöke A, Tortelli A, Llorca PM, Tarricone I, Tripoli G, Ferraro L, La Cascia C, Lasalvia A, Tosato S, Menezes PR, Del-Ben CM, Nelson B, Riecher-Rössler A, Bressan R, Barrantes-Vidal N, Krebs MO, Nordentoft M, Ruhrmann S, Sachs G, Rutten BPF, van Os J, Velthorst E, de Haan L; EU-GEI High Risk Study. Pre-training inter-rater reliability of clinical instruments in an international psychosis research project. *Schizophr Res.* 2020 Nov 23:S0920-9964(20)30429-1

Final peer reviewed version available at : <https://doi.org/10.1016/j.schres.2020.08.001>

Rights / License:

The terms and conditions for the reuse of this version of the manuscript are specified in the publishing policy. For all terms of use and more information see the publisher's website.

This item was downloaded from IRIS Università di Bologna (<https://cris.unibo.it/>)

When citing, please refer to the published version.

Letter to the editor

Title: Pre-training inter-rater reliability of clinical instruments in a large international multi-center psychosis research project.

Running title: Pre-training inter-rater reliability.

Total word count: 998.

Authors: Steven Berendsen^{1*}, Pim Kapitein¹, Frederike Schirmbeck¹, Mirjam J. van Tricht¹, Philip McGuire², Craig Morgan³, Charlotte Gayer-Anderson³, Matthew J Kempton², Lucia Valmaggia⁴, Diego Quattrone², Marta di Forti², Mark van der Gaag⁵, James B. Kirkbride⁶, Hannah E Jongsma^{6,7}, Peter B Jones⁷, Maria Parellada⁸, Celso Arango⁸, Manuel Arrojo⁹, Miguel Bernardo¹⁰, Julio Sanjuán¹¹, José Luis Santos^{12,13,14}, Andrei Szöke¹⁵, Andrea Tortelli¹⁶, Pierre-Michel Llorca¹⁷, Ilaria Tarricone¹⁸, Giada Tripoli¹⁹, Laura Ferraro¹⁹, Caterina La Cascia¹⁹, Antonio Lasalvia²⁰, Sarah Tosato²⁰, Paulo Rossi Menezes²¹, Cristina Marta Del-Ben²², Barnaby Nelson^{23,24}, Anita Riecher-Rössler²⁵, Rodrigo Bressan²⁶, Neus Barrantes-Vidal^{27,28}, Marie-Odile Krebs²⁹, Merete Nordentoft⁴¹, Stephan Ruhrmann³⁰, Gabriele Sachs³¹, Bart P. F. Rutten³², Jim van Os^{2,32,33}, EU-GEI High Risk Study, Eva Velthorst^{1,34,35}, Lieuwe de Haan^{1,36}.

EU-GEI High Risk Study Group not mentioned in main author list: Maria Calem², Stefania Tognin², Gemma Modinos², Sara Pisani², Tamar C. Kraan¹, Daniella S. van Dam¹, Nadine Burger⁴⁴, G Paul Amminger²³, Athena Politis²³, Joanne Goodall²³, Stefan Borgwardt³⁷, Erich Studerus³⁷, Ary Gadelha²⁶, Elisa Brietzke³⁸, Gracielle Asevedo³⁸, Elson Asevedo³⁸, Andre Zugman³⁸, Tecelli Domínguez-Martínez³⁹, Manel Monsonet²⁸, Lidia Hinojosa²⁸, Paula Cristóbal-Narváez²⁸, Anna Racioppi²⁸, Thomas R. Kwapil⁴⁰, Mathilde Kazes²⁹, Claire Daban²⁹, Julie Bourgin²⁹, Olivier Gay²⁹, Célia Mam-Lam-Fook²⁹, Dorte Nordholm⁴¹, Lasse Randers⁴¹, Kristine Krakauer⁴¹, Louise Birkedal Glenthøj⁴¹, Birte Glenthøj⁴², Dominika Gebhard³⁰, Julia Arnhold⁴³, Joachim Klosterkötter³⁰, Iris Lasser³¹, Bernadette Winklbaur³¹, Philippe A Delespaul³².

Acknowledgements

The European Network of National Schizophrenia Networks Studying Gene-Environment Interactions (EU-GEI) Project is funded by grant agreement HEALTH-F2-2010-241909 (Project EU-GEI) from the European Community's Seventh Framework Programme. Additional support was provided by a Medical Research Council Fellowship to M Kempton (grant MR/J008915/1). Further, we would like to thank EU-GEI WP2 Group not mentioned in main author list: Kathryn Hubbard, Stephanie Beards, Simona A. Stilo, Pedro Cuadrado, José Juan Rodríguez Solano, David Fraguas, Álvaro Andreu-Bernabeu, Gonzalo López, Bibiana Cabrera, Juan Nacher, Javier Costas, Mario Matteis, Marta Rapado-Castro, Emiliano González, Covadonga M. Díaz-Caneja, Emilio Sánchez, Manuel Durán-Cutilla, Nathalie Franke, Fabian Termorshuizen, Daniella van Dam, Elles Messchaart, Marion Leboyer⁴, Franck Schürhoff, Stéphane Jamain, Grégoire Baudin, Aziz Ferchiou, Baptiste Pignon, Jean-

Romain Richard, Thomas Charpeaud, Anne-Marie Tronche, Flora Frijda, Giovanna Marrazzo, Crocettarachele Sartorio, Fabio Seminerio, Camila Marcelino Loureiro, Rosana Shuhama, Mirella Ruggeri, Chiara Bonetto, Doriana Cristofalo, Domenico Berardi, Marco Seri, Elena Bonora, Anastasios Nougus, Giuseppe D'Andrea, Laura Ferraro, Ulrich Reininghaus, Enrique García Bernardo, Laura Roldán, Esther Lorente-Rovira, Ma Soledad Olmeda, Daniele La Barbera, Cristina M. Del-Ben, Lucia Sideli.

*Corresponding author: Steven Berendsen, MD, UMC Amsterdam, Location AMC. Meibergdreef 9, 1105 AZ Amsterdam. Telephone: +3120 566 9111, email: s.berendsen@amsterdamumc.nl.

¹ Department of Psychiatry, Amsterdam UMC, Amsterdam, the Netherlands.

² Department of Psychosis Studies, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, England.

³ Department of Health Service and Population Research, Institute of Psychiatry, King's College London, London, UK.

⁴ Department of Psychology, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

⁵ Amsterdam Public Mental Health Research Institute, Department of Clinical Psychology, Faculty of Behavioural and Movement Sciences, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands

⁶ PsyLife Group, Division of Psychiatry, UCL, London, England.

⁷ Department of Psychiatry, University of Cambridge, Cambridge, England

⁸ Department of Child and Adolescent Psychiatry, Institute of Psychiatry and Mental Health, Hospital General Universitario Gregorio Marañón, School of Medicine, Universidad Complutense, Instituto de Investigación Sanitaria Gregorio Marañón (IiSGM), Spanish Mental Health Research Network (CIBERSAM), Madrid, Spain

⁹ Department of Psychiatry, Instituto de Investigación Sanitaria (IDIS), Complejo Hospitalario Universitario de Santiago de Compostela, Santiago de Compostela, Spain

¹⁰ Barcelona Clinic Schizophrenia Unit, Neuroscience Institute, Hospital Clinic of Barcelona, University of Barcelona, Barcelona, Spanish Mental Health Research Network (CIBERSAM), Spain

¹¹ Department of Psychiatry, Hospital Clínico Universitario de Valencia, School of Medicine, Universidad de Valencia, Valencia, Spain

¹² Department of Psychiatry, University of Oviedo, Spanish Mental Health Research Network (CIBERSAM), Instituto de Investigación Sanitaria del Principado de Asturias (ISPA), Mental Health Services of Principado de Asturias (SESPA), Oviedo, Spain

¹³ Neurobiological Research Group, Institute of Technology, Universidad de Castilla-La Mancha, Cuenca, Spain

¹⁴ Department of Psychiatry, Hospital "Virgen de la Luz", Cuenca, Spain

¹⁵ INSERM, U955, Créteil, France

¹⁶ Etablissement Public de Santé Maison Blanche, Paris, France

¹⁷ Centre Hospitalier Universitaire de Clermont-Ferrand, Clermont-Ferrand, France

¹⁸ Department of Medical and Surgical Sciences, Psychiatry Unit, Alma Mater Studiorum Università di Bologna, Bologna, Italy

¹⁹ Department of Biomedicine, Neuroscience and advanced Diagnostics, University of Palermo, Palermo, Italy

²⁰ Section of Psychiatry, Department of Neuroscience, Biomedicine and Movement, University of Verona, Verona, Italy

²¹ Department of Preventive Medicine, Faculdade de Medicina, Universidade of São Paulo, São Paulo, Brazil

- ²²Ribeirão Preto Medical School, University of São Paulo, Brazil
- ²³Orygen, Parkville, Victoria, Australia
- ²⁴Centre for Youth Mental Health, The University of Melbourne, Parkville, Victoria, Australia
- ²⁵Medizinische Fakultät, Universität Basel, Basel, Switzerland
- ²⁶LiNC-Lab Interdisciplinar Neurociências Clínicas, Depto Psiquiatria, Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
- ²⁷Departament de Psicologia Clínica i de la Salut, Universitat Autònoma de Barcelona, Barcelona, Spain
- ²⁸Fundació Sanitària Sant Pere Claver, Spanish Mental Health Research Network (CIBERSAM), Spain
- ²⁹University of Paris, GHU-Paris, Sainte-Anne, C'JAAD, Inserm U1266, Institut de Psychiatrie (CNRS 3557), Paris, France
- ³⁰Department of Psychiatry and Psychotherapy, Faculty of Medicine and University Hospital, University of Cologne, Cologne, Germany
- ³¹Department of Psychiatry and Psychotherapy, Medical University of Vienna, Vienna, Austria
- ³²Department of Psychiatry and Neuropsychology, School for Mental Health and Neuroscience, Maastricht University Medical Centre, Maastricht, the Netherlands
- ³³Department of Psychiatry, Brain Centre Rudolf Magnus, University Medical Center Utrecht, Utrecht University, Utrecht, the Netherlands
- ³⁴Seaver Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, NY, USA
- ³⁵Department of Psychiatry, Icahn School of Medicine at Mount Sinai, NY, USA
- ³⁶Arkin Institute for Mental Health, Amsterdam, the Netherlands
- ³⁷University of Basel, Basel, Switzerland
- ³⁸Department of Psychiatry, Escola Paulista de Medicina, Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
- ³⁹CONACYT-Dirección de Investigaciones Epidemiológicas y Psicosociales, Instituto Nacional de Psiquiatría Ramón de la Fuente Muñiz, México
- ⁴⁰Department of Psychology, University of Illinois at Urbana-Champaign, IL, USA
- ⁴¹Mental Health Center Copenhagen and Center for Clinical Intervention and Neuropsychiatric Schizophrenia Research, CINS, Mental Health Center Glostrup, Mental Health Services in the Capital Region of Copenhagen, University of Copenhagen, Copenhagen, Denmark
- ⁴²Centre for Neuropsychiatric Schizophrenia Research (CNSR) & Centre for Clinical Intervention and Neuropsychiatric Schizophrenia Research (CINS), Mental Health Centre Glostrup, University of Copenhagen, Glostrup, Denmark
- ⁴³Psyberlin, Berlin, Germany
- ⁴⁴Parnassia Psychiatric Institute, Department of Psychosis Research, Zoutkeetsingel 40, 2512 HN The Hague, The Netherlands.

Dear Editor,

Inter-rater reliability (IRR) is an important component of methodology to establish valid results and prevent large measurement errors. However, only a minority of reports in psychiatric research present information concerning assessor training or reliability of applied instruments. For example, a recent study found that IRR coefficients and training procedures were strongly underreported in double-blind RCTs with antipsychotic medication[1].

IRR scores without training of raters are typically low, only four studies investigated pre-training IRR [2-5]. The authors reported that the IRR scores of the PANSS, HAM-D or GAF [abbreviations written out in the supplement] before training were generally moderate to poor, other observational instruments were not investigated. On the other hand, the authors reported significant improvement of the IRR after assessors were trained.

Selection of assessors based on their clinical backgrounds and assessment experience may also lead to improved pre-training IRR scores. However, merely three studies addressed the topic of assessor selection and pre-training reliability. The first study of Kobak et al. provided evidence that assessors with a PhD or medical degree showed significantly higher HAM-D clinical assessment skills necessary to conduct reliable assessments compared to assessors with lower educational degrees [6]. In contrast, Loevdahl et al. and Kollias et al. found no differences in pre-training reliability of the GAF or the CAARMS between psychiatrists, residents, psychologists and nurses [5, 7].

This raises the question whether acceptable IRR scores can be achieved without assessor training or selection. Therefore, we aimed to determine the pre-training IRR of seven observational instruments that capture different aspects of psychosis in a large international multi-center research project by scoring video-taped interviews. In addition, we investigated the effect of assessor characteristics on pre-training IRR scores.

Assessors of the large multi-center study EU-GEI were instructed to rate participants on seven instruments via an online training platform [8]. These instruments were chosen to measure predictors and outcome in psychosis. Ratings were based on videotaped assessments of interviews with actors playing the role of the patient. Demographic characteristics (age and gender), professional background (psychiatrists, psychologists, medical doctors or research assistants) and assessment experience (in months) of assessors were collected. The pre-training IRR of the following instruments were evaluated: CAARMS, SIS-R, LoTE, BQ, CECA, OPCRIT and GAF.

Pre-training IRR was calculated by Krippendorff's alpha (K-alpha) [9]. According to interpretation guidelines, K-alpha values of >0.8 were considered high, $0.67 - 0.8$ moderate, and <0.67 low [10]. For each K-alpha 95% confidence intervals were computed based on 10.000 bootstraps. Differences in age, assessment experience and IRR between different professional groups were analyzed for each assessment instrument by analysis of variance (ANOVA), followed by Bonferroni corrected pair-wise post-hoc comparisons.

**Table 1.*

In total 12 psychiatrists, 17 psychologists, 14 medical doctors and 13 research assistants participated in the online training platform. Mean age [30.18 years, $F=13.43$, $p<0.001$; see supplement table 1] and assessment experience ($F=5,76$, $p=0.002$; see supplement figure 1) were significantly higher for psychiatrists compared to medical doctors and research assistants, and at trend level compared to psychologists.

Observed pre-training IRR score was moderate for LoTE (K-alpha =0.67), low for GAF (K-alpha=0,45), BQ (K-alpha =0.47), SIS-R (K-alpha = 0.55), CAARMS (0,57), CECA (K-alpha =0.60) and OPCRIT (K-alpha =0.64).

IRR scores of subgroups are shown in Table 1. Overall mean IRR scores were significantly higher for psychiatrists compared to medical doctors ($F=3,905$, $p=0.0216$). Comparisons for separate instruments showed significantly higher IRR scores for psychiatrists, psychologists and research assistants compared to medical doctors on the OPCRIT ($F=18,38$, $p<0.001$), SIS-R ($F=20,66$, $p<0.001$), GAF ($F=12,53$, $p<0.001$) and CAARMS ($F=13,34$, $p<0.001$). Additionally, medical doctors and research assistants scored significantly higher IRR scores compared to psychiatrists and psychologists on the BQ ($F=16,75$, $p<0.001$). For detailed information on pair-wise comparisons of IRR scores between professionals and assessment experience see supplement figures 2a-2f.

Our study demonstrated that only one instrument showed moderate pre-training IRR, whereas the observed reliability scores of all other instruments were insufficient. Furthermore, medical doctors demonstrated significantly lower reliability scores compared to other professional subgroups in mean IRR ratings and several investigated instruments. These findings are important, in light of previous research which noted that rater training was strongly underreported and the impact of unreliability on study outcome [11, 12].

Our findings are in accordance with earlier results concerning insufficient pre-training IRR [2, 3, 5, 13]. Differences in mean IRR scores between professions could be explained by the significantly higher assessment experience of psychiatrists compared to the other professions. However, observed IRR scores of separate instruments were also different between psychologists and research assistants compared to medical doctors, while the latter two subgroups did not significantly differ in assessments experience. Our hypothesis concerning the latter variation is that research assistants and psychologist probably received more training in psychopathology scales such as the CAARMS or SIS-R during their general education, in comparison to medical doctors.

Our findings concerning differences between professionals seem to contrast with previous literature, which found no significant differences in pre-training IRR of GAF scores between psychiatrists and psychologists, compared to psychiatric nurses [5]. Similarly, another study concerning the CAARMS provided evidence that psychiatry residents produced almost similar IRR scores compared to psychiatrists and psychologists [7]. Possible explanations for these inconsistent findings could be that psychiatry residents have more experience with observational instruments and psychiatric diagnosis compared to medical doctors.

Of note, we evaluated *pre-training* IRR in this report. All included researchers achieved high IRR scores after training before permitted to perform assessments. However, we should acknowledge an important limitation of our study: we do not have data concerning previous training or clinical background of raters.

In conclusion, our study emphasizes the importance of rater training and assessor selection for research in psychiatry. Without rater training, reliability is generally insufficient. This has potentially major implications for the interpretation of study-results because of decreased power and higher placebo-response^{*see supplement} [14, 15]. Future research should focus on specific assessors characteristics that predict higher IRR scores after training. Finally, considering its importance, we propose training procedures and reliability coefficients should be reported in all studies.

References

1. Berendsen, S., et al., *Burying Our Heads in the Sand: The Neglected Importance of Reporting Inter-Rater Reliability in Antipsychotic Medication Trials*. Schizophr Bull, 2020.
2. Muller, M.J. and A. Dragicevic, *Standardized rater training for the Hamilton Depression Rating Scale (HAM-D-17) in psychiatric novices*. J Affect Disord, 2003. **77**(1): p. 65-9.
3. Muller, M.J. and H. Wetzel, *Improvement of inter-rater reliability of PANSS items and subscales by a standardized rater training*. Acta Psychiatr Scand, 1998. **98**(2): p. 135-9.
4. Rosen, J., et al., *Web-based training and interrater reliability testing for scoring the Hamilton Depression Rating Scale*. Psychiatry Res, 2008. **161**(1): p. 126-30.
5. Loevdahl, H. and S. Friis, *Routine evaluation of mental health: reliable information or worthless "guesstimates"?* Acta Psychiatr Scand, 1996. **93**(2): p. 125-8.

6. Kobak, K.A., et al., *A new approach to rater training and certification in a multicenter clinical trial*. J Clin Psychopharmacol, 2005. **25**(5): p. 407-12.
7. Kollias, C., et al., *Inter-rater reliability of the Greek version of CAARMS among two groups of mental health professionals*. Psychiatriki, 2015. **26**(3): p. 217-22.
8. van Os, J., et al., *Identifying gene-environment interactions in schizophrenia: contemporary challenges for integrated, large-scale investigations*. Schizophr Bull, 2014. **40**(4): p. 729-36.
9. Hayes, A.F. and K. Krippendorff, *Answering the Call for a Standard Reliability Measure for Coding Data*. . Communication Methods and Measures, 2007.
10. Krippendorff, *Agreement and Information in the Reliability of Coding*. Communication Methods and Measures, 2011.
11. Mulsant, B.H., et al., *Interrater reliability in clinical trials of depressive disorders*. Am J Psychiatry, 2002. **159**(9): p. 1598-600.
12. Kobak, K.A., et al., *Why do clinical trials fail? The problem of measurement error in clinical trials: time to test new paradigms?* J Clin Psychopharmacol, 2007. **27**(1): p. 1-5.
13. Vatnaland, T., et al., *Are GAF scores reliable in routine clinical use?* Acta Psychiatr Scand, 2007. **115**(4): p. 326-30.
14. Perkins, D.O., R.J. Wyatt, and J.J. Bartko, *Penny-wise and pound-foolish: the impact of measurement error on sample size requirements in clinical trials*. Biol Psychiatry, 2000. **47**(8): p. 762-6.
15. Kobak, K.A., et al., *Site versus centralized raters in a clinical depression trial: impact on patient selection and placebo response*. J Clin Psychopharmacol, 2010. **30**(2): p. 193-7.

Table 1. Omnibus test across all groups.							
	Psychiatrists (N=12)	Psychologist (N=17)	Medical doctors (N=14)	Research assistants (N=13)	F	dF	P-value
Mean IRR (SD)	0.67 (0.14)	0,60 (0.17)	0.43 (0.16)	0.64 (0.09)	3.905	3	0.0216
OPCRIT [95% CI]	0.81 [0.75-0.86]	0.73 [0.65-0.80]	0.44 [0.35-0.53]	0.68 [0.58-0.77]	16.02	3	<0.0001
SIS-R [95% CI]	0.75 [0.69-0.82]	0.66 [0.57-0.74]	0.32 [0.21-0.42]	0.74 [0.63-0.83]	22.92	3	<0.0001
LoTE [95% CI]	0.78 [0.62-0.91]	0.66 [0.47-0.82]	0.62 [0.40-0.79]	0.67 [0.47-0.85]	0.6553	3	0.5819
GAF [95% CI]	0.64 [0.54-0.73]	0.49 [0.39-0.56]	0.27 [0.16-0.37]	0.53 [0.38-0.63]	12.53	3	<0.0001
CECA [95% CI]	0.61 [0.58-0.78]	0.60 [0.55-0.64]	0.55 [0.47-0.64]	0.60 [0.54-0.68]	0.5124	3	0.6744
BQ [95% CI]	0.43 [0.27-0.59]	0.26 [0.15-0.37]	0.57 [0.44-0.69]	0.73 [0.63-0.81]	16.44	3	<0.0001
CAARMS [95% CI]	-	0.78 [0.72-0.84]	0.21 [-0.02-0.42]	0.53 [0.33-0.70]	13.82	2	<0.0001

